Time series analysis and forecasting with applications to climate science

Safia Amur Ali ALMarhoobi
School of Mathematics

Submitted in partial fulfillment of the requirements for the degree of *Doctor of Philosophy*

2022

Abstract

Singular spectrum analysis (SSA) is the popular tool for analysing and forecasting time series. SSA can be used for parametric estimation, forecasting and gap filling amongst many other tasks. SSA was used for the extraction of seasonality, simultaneous extraction of cycles with small and large periods and finding structure in short time series. This thesis aims to study the application of singular spectrum analysis which is supported by empirical evidence to further promote the value, effectiveness and applicability of strengthening SSA's quality in the field of time series analysis and forecasting.

We investigate the hourly, daily and monthly temperature and humidity time series collected at meteorological stations in Oman from 2009 to 2018. This data is provided by the Directorate General of Meteorology of Oman. Our investigation cover missing value imputation, splitting the hourly time series in the sum of several components corresponding to different frequencies and detection of trends. We investigated three methods of imputation: SSA-based iterative approach, regression methods and regression with lagging. We found that imputation by regression with lagging is a more reliable and reasonable method and provides natural results for filling gaps for any length of time series. We applied SSA to hourly time series for extracting the annual oscillations and the daily periodicities. SSA was able to extract these components very effectively. Moreover, we may use SSA for obtaining more refined decompositions with larger number of components and also for forecasting. We applied three commonly used tests for detecting

trends in time series: the Mann-Kendall test, Spearman's rho test and the Sen's innovative trend method test. We found that there are no monotonic trends in the annual oscillations and the daily periodicities over the period of ten years. Also we did not find trends in the monthly variability of daily periodicities.

We provide a statistical framework on studying which SSA forecasting algorithm is best on the example of real data representing monthly temperature and humidity in Oman. We demonstrated that the sensitivity of the root mean squared errors (RMSE) for retrospective forecasts is rather small to parameters the window length L and the number of singular values r. We shown that the efficiency of SSA forecasts with the automatic choice of parameters is rather high. We also found that SSA-R and SSA-V forecasts are more similar to each other with a slight dominance of SSA-V forecasts.

Last part of thesis focuses on the performance of the application of SSA to daily time series of humidity and temperature in Oman. We apply SSA forecasting algorithms: recurrent SSA (SSA-R) forecasting, recurrent SSA original (SSA-R (original)) forecasting and vector SSA (SSA-V) forecasting algorithms based on SSA with double projection and SSA without projection. We have also studied the effect of series length and choice of parameters on the performance of the aforementioned algorithms. The findings show that SSA with double projection improve the accuracy of short term forecast using smaller set of observations.

Acknowledgement

This thesis is the culmination of my academic journey. However, I have been afforded the want and resilience to get to this point by a great many people. Here, I would like to acknowledge a non-exhaustive list of those people.

Firstly, I offer my utmost gratitude to my supervisors, Dr Andrey Pepelyshev and Professor Anatoly Zhigljavsky, Your support, encouragement and guidance throughout this project have been invaluable. You have made a great team for me.

Secondly, clearly without the financial support I received from the national postgraduate scholarship programme in Sultanate of Oman, I was unable to fulfil one of my dreams. I am very grateful for this generous funding. I have also express many thanks to everyone who supporting me in ministry of higher education, research and innovation and from Cardiff University, School of Mathematics: to academic and research staff, to professional service staff and all other post graduate students.

On a more personal note, I would like to thank my family who have been a source of great inspiration and motivation throughout life. They never questioned my decisions and have supported me to the fullest. I would like to express my sincere gratitude to my husband "Abdullah" to my sons " Mohamed and Abdulamlik" and to my daughters "Alla and Israa".

I offer my thanks to my family, who have been a source of inspiration and support

during my life. Thank you to my parents, brothers, sisters, nieces and nephews. I must also thank many true friends who were always encouraging and supportive for their unconditional support in my PhD.

Dissemination of Work

Publications (3 Published and 1 in preparation)

1. 2021: ALMarhoobi, S. and Pepelyshev, A. Analysis of temperature and humidity in Oman using singular spectrum analysis. Communications in Statistics-Simulation and Computation, 1-14.;

2. 2021: ALMarhoobi, S. Study of temperature variations in Sultanate of Oman using Singular Spectrum Analysis. Nitag,OSAC Scientific Committee, 3th edition, 98-95;

3. 2021: ALMarhoobi, S. and Pepelyshev, A. Study of automatic choice of parameters for forecasting in singular spectrum analysis. Statistics and Its Interface.

In preparation

1. 2022: ALMarhoobi, S. and Mahmoudvand, R. Forecasting daily temperature and humidity in Oman by using Singular Spectrum Analysis.

Talks

1. Analysis of temperature and humidity in Oman using singular spectrum analysis. British Early Career Mathematicians' Colloquium 2020. 14-15 July 2020.

2. Imputation missing values by advance techniques. PGR talks, mathematics school of Cardiff university. 14th October 2020

3. Study of automatic choice of parameters for forecasting in singular spectrum analysis. 2021 UK National Student-SIAM Chapter Conference.

Contents

List of Figures

List of Tables

List of abbreviations

L	Window length
r	Number of singular values
ARIMA	Autoregressive integrated moving average
EOF	Empirical orthogonal function
ETS	Exponential smoothing
ITM	Innovation trend method
LRR	Linear recurrence relation
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MK	Mann Kendall test
MSSA	Multivariate singular specturm analysis
NLSA	Non-linear Laplacian spectral analysis
NN	recurrent Neural networks
NRMSE	Normalized Root Mean square error
PCA	Principal component analysis

RMSE	Root mean square error
SR	Spearman's rho test
SSA	Singular spectrum analysis
SSA-R	Recurrent SSA
SSA-RO	Recurrent original SSA
SSA-V	Vector SSA
SVD	Singular value decomposition

Chapter 1

Introduction

1.1 Time series analysis and forecasting

Time series analysis is a statistical technique that deals with data collected over time that covers essential information about a physical, biological, meteorological or economic system. The purpose of time series analysis is to know how the time series behaved in the past. This is helpful for predicting the system's future behaviors [31]. The univariate or multivariate time series can be analysed for different purposes such as gap filling and forecasting.

In recent years, several new advanced techniques have been used to analyse time series in order to predict future trends. One of these techniques is singular spectrum analysis (SSA), a powerful technique that can be used for smoothing, extracting trends, extracting periodicities, forecasting, filling in missing values, estimating signal parameters and detecting points of change. SSA is a non-parametric approach, there are no statistical assumptions about issues such as the stationarity of the series or the normality of the residuals [40, 48, 60, 89].

Many researches have conducted on a wide range of classical and advanced time series analysis techniques including the autoregressive integrated moving average

(ARIMA) model, exponential smoothing (ETS) and recurrent neural networks (RNN). Each technique has advantages and disadvantages. Parametric models are restricted by assumptions of stationarity and normality that are unlikely to hold in a real-world scenario, especially following recessions that make time series non-stationary [121].

Analysis of a time series provides an overview of the nature of data that can then be used in statistical forecasting methods to predict future values. Therefore, researchers constantly endeavor to provide a high level of accuracy in forecasting by using more efficient techniques.

Several literature reviews have reported an increase in the use of SSA to analyse climatic, meteorological and geophysical time series [10, 91, 113].

Meteorological time series is an exciting and complex challenge that can include non-normal forms of distribution, serial dependency, irregular spacing [94]. Variables such as temperature, relative humidity and precipitation are important factors that can be used to forecast yearly, monthly, daily and hourly meteorological conditions and play an important role in decision making. For example, increasing temperatures can reduce crop yields, while precipitation can increase yields to a level that roughly matches crop's rate of evapotranspiration [69]. These changes make a meteorological time series non-stationary. It is very important to develop a methodology that is dynamic enough to account for these changes, in order to produce accurate predictions. There have been several recent theoretical developments in SSA and related applications [24].

A Google Scholar search for studies that used SSA for the period from 2007 to 2021 shows that its use has increased. There are also publications about further improving SSA as an effective tool for modelling and forecasting. This research

study adds to the rich literature on the use of SSA supported with empirical evidence.

The structure of Chapter 1 is as follows. In Section 1.1.1, we state the research objectives of the thesis. In Section 1.2, we consider reasons for choosing SSA and outline the research objectives. In Section 1.3, we describe the structure of the thesis and its Chapters. In Section 1.4, we outline the novel contributions of the thesis.

1.1.1 Research objectives

This research aims to study application of SSA in the field of meteorological time series. This thesis leads to achieve several research objectives. These objectives represent the contributions to SSA and the field of time series analysis and forecasting.

1. Apply singular spectrum analysis.
2. Determine different ways of imputing missing values in a time series.
3. Extract annual oscillations and daily periodicities for time series of humidity and temperature.
4. Study of automatic choice of parameters for forecasting monthly time series of temperature and humidity.
5. Forecasting daily time series of temperature and humidity using SSA with double projection and SSA without projection at $1, 2, \ldots, 14$ and $1, 2, 3$ days ahead forecasts across the SSA forecasting algorithms.

We can achieve these objectives using empirical evidence from time series of temperature and humidity in Oman.

1.2 Motivation

1.2.1 Why Singular Spectrum Analysis?

This section considers the reasons for selecting SSA and introduces the topic of forecasting meteorological time series.

SSA is a non parametric method that is powerful for time series analysis. SSA can be used for parametric estimation, forecasting and filling gaps among many other tasks [40, 42, 60, 103]. SSA is also a very useful tool for extracting various signals from noisy observations [52]. SSA can be used to extract data about seasonality and to simultaneously extract both long and short cycles. It can identify structure in short time series. The core of SSA is its ability to decompose the original time series data into a small number of components by using singular value decomposition (SVD) of a trajectory matrix [45, 150].

In addition, SSA enables researchers to decompose time series to obtain a richer understanding of the underlying dynamics. Moreover, once the signal is extracted, SSA enables users to forecast a particular signal. For example, if we are only interested in the trend component, we have the option of extracting the trend from the data and then, forecasting the trend [65]. SSA seeks to filter the noise from a time series and reconstruct a less noisy signal, which is then, used to forecast future data points using the window length L and the number of singular values r [59, 66].

SSA uses non-parametric techniques to decompose time series into main components and reconstruct the series by leaving behind the random (noise) component [54]. SSA is an effective method for forecasting based on time series that are polluted with different types of noise [107].

SSA is a powerful method for analysis of both stationary and nonstationary time series. To assess the performance of SSA when applied to a nonstationary time series, we consider a real life example of US male unemployment from 1950 to 1980 (using example from [42]). The length N of the time series is 400, and we take L= 200,150,100,70,40 and 20 as window lengths. Figure 1.1 displays the first leading principal components and the corresponding contribution obtained from analysis. We observe that the reproduced trend changes from detailed to smooth with increasing window length, and the corresponding contribution percentages (trend line) are gradually decreased. With different window lengths, some statistical quantitative verification measures of forecasts can be calculated to evaluate their performance according to RMSE.

1.2.2 Why These Objectives?

The first objective seeks to apply and analyse time series and using that analysis as a basis for forecasting. The second objective is determining ways to impute any missing values in time series. We used SSA to find forecasts corresponding to missing values and then, combine the forecasted values to estimate missing values [88].

There are different types of missing data, including the temporary absence of observers, damaged monitoring equipment, or lack of financial resources [128]. The important pre processing procedure of data refinement should be performed

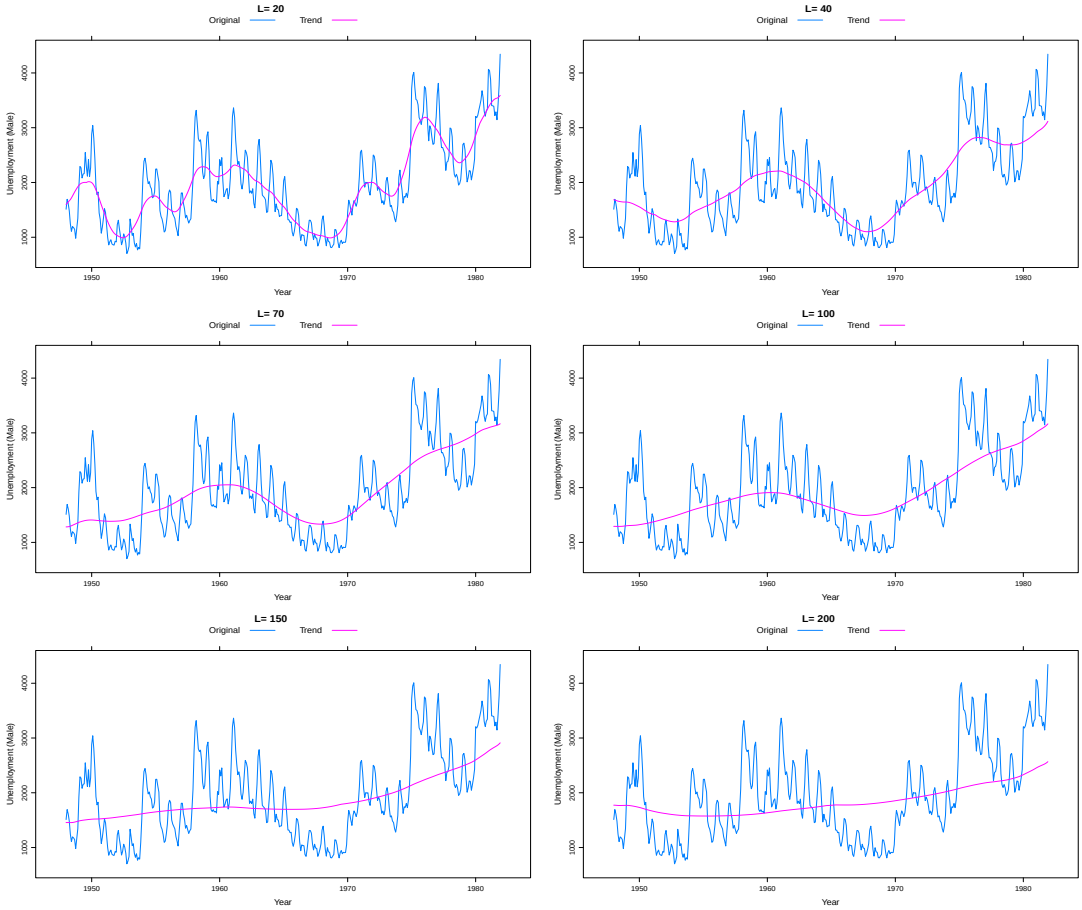

Figure 1.1: Non-stationarity time series and and 1st leading components with SSA.

and gap values are tackled either by eliminating the vectors containing missing samples or by simply using some mean [126].

Imputations of missing values are computationally intensive and some algorithms must be run multiple times to get satisfactory results and the run duration that is necessary increases as the volume of missing data increases [125]. Imputation using SSA-based iterative approach, multiple regression and regression with lagging have been considered and Chapter 3 discusses the various issues associated with imputing missing values.

The third objective is to extract annual oscillations and daily periodicities of temperature and humidity. I have focused on trend analysis by using the Mann-Kendall (MK) test, the Spearman's rho test (SR) and the Sen's innovative trend method (ITM) test by extracting annual oscillations and daily periodicities using SSA.

Researchers, government organizations, practitioners and forecasters publish annually, monthly, quarterly, daily, or hourly forecasts for a variety of variables. Such forecasts are generated using SSA which can produce more accurate results than some classical time series methods.

The fourth objective of this research is to study the automatic choice of parameters for forecasting using singular spectrum analysis. We study the sensitivity of the RMSE and investigate the reliability of the automatic choice of parameters for forecasting monthly temperature and humidity recorded at three meteorological stations in Oman.

SSA uses the window length to decompose data and the number of singular values to reconstruct and forecast. The number of singular values can determine the accuracy of its predictions; it is important to choose the right number. Selecting the window length is also important because a poorly chosen window length would lead to an inferior decomposition [87]. The RMSE is used to measure the accuracy and quality of SSA forecasts [57]. To achieve the fourth objective, we have compared SSA with classic forecasting methods such as ARIMA, ETS and RNN [59, 66].

The final objective of this research focuses on analysing how SSA with projection works for daily time series of humidity and temperature in Oman. We have used recurrent SSA (SSA-R), SSA-RO (original) and vector SSA (SSA-V) forecasting

algorithms based on SSA with double projections and SSA without projections.

1.3 Structure of the thesis

This thesis contains seven Chapters. A summary of each Chapter is given below.

- Chapter 2 describes SSA in detail, including SSA parameters, SSA algorithms for imputing missing data, SSA with projection and SSA forecasting. The Chapter also provides benchmarking forecast models and the RMSE.
- Chapter 3 presents the imputation techniques for filling gaps using SSA-based iterative approach, multiple regression and regression with lagging. The Chapter concludes with an investigation of the best method for imputing missing values for the time series of the temperature and humidity data.
- Chapter 4 describes three methods of trend analysis: the MK test, the SR test and the ITM test. In addition, it focuses on extracting annual oscillations and daily periodicities.
- Chapter 5 explores the automatic choice of parameters for forecasting in SSA. The SSA algorithm has two parameters: the window length L and the number of singular values r. Choice of parameters L and r is depending on both the structure of the time series and the forecasting aims. The Chapter also discusses the SSA forecasting algorithms.
- Chapter 6 explores daily time series of temperature and humidity data using SSA with double projection and SSA without projection. The RMSE of 1,2,...,14 day ahead forecasts and the RMSE of 1,2,3 days ahead forecasts

across several truncation points depending on the window length L and the number of singular values r.

- Chapter 7 summarizes the research presented in the thesis and establishes avenues for further work.

1.4 Novel contributions of the thesis

This section lists the novel contributions of each Chapter with a concise description of the research problem, the literature related to the problem and how this thesis addresses the problem.

Chapter 3 addresses the issue of how to deal with missing values using different approaches. It uses three methods to impute missing values: imputation by the SSA-based iterative approach, imputation by multiple regression and imputation by regression with lagging. Imputation by regression with lagging produces more reliable and reasonable method and provides natural results for filling gaps for any length.

Chapter 4 presents annual oscillations and daily periodicities of several variables of an hourly time series. It considers three trend tests which are the MK test, the SR test and the ITM test. Theses tests provide information for planners and policymakers who must take into account future changes in meteorological time series.

Chapters 5 and 6 make two major contributions to forecasting using the SSA forecasting algorithms. These two Chapters use the SSA forecasting algorithms for monthly and daily time series for temperature and humidity. Chapter 5 provides a statistical framework for studying the SSA forecasting algorithm and

demonstrates the sensitivity of the RMSE for retrospective forecasts to optimize the SSA parameters. Chapter 6 considers daily time series of temperature and humidity using SSA with double projection and SSA without projection and three forecasting algorithms: SSA-R, SSA-R (original) and SSA-V.

Chapter 2

Methodology

2.1 Introduction

In this Chapter, we are explaining the methodology of singular spectrum analysis (SSA) and introduce the details of the SSA algorithms.

This Chapter is structured as follows.

- In Section 2.2, we provide a general overview of SSA.
- In Section 2.3, we address the methodology of SSA and present the algorithms.
- In Section 2.4, we explore SSA with projection.
- In Section 2.5, we discuss how to choose certain parameters of the SSA algorithm.
- In Section 2.6, we discuss methods for forecasting.
- In Section 2.7, we present the algorithm of gap filling.
- In Section 2.9, we present the benchmark forecasting models.

- In Section 2.10, we focus on metrics used for assessing accuracy of forecasts.

- In Section 2.11, we provide a summary of this Chapter.

2.2 Singular spectrum analysis

SSA is a non-parametric, powerful method for time series analysis. It can be applied to many areas including parametric estimation, forecasting and gap-filling, see [40, 42, 60, 103].

SSA has become a popular time series analysis since its introduction in [15]. Note that the ideas of SSA were also independently developed in Russia (St.Petersburg, Moscow) [53] and in several groups in different areas in the world. Several papers discussing the methodological aspects and applications of SSA can be found in [7, 24, 32, 40, 42, 55, 59, 60, 65, 137, 150]. The first formal description of SSA can be attributed to [42, 43, 48].

SSA is a very useful tool for extracting various signals from noisy observations [52]. It has been used for the extraction of seasonality and simultaneous extraction of cycles of small and long periods, finding structure in a short time series [45]. SSA provides meaningful results in many research areas without imposing any restrictive assumptions on the data. The algorithm of SSA relies on the decomposition of the original time series data into the sum of a small number of components using a singular value decomposition (SVD) of a trajectory matrix [150]. SSA is able to filter a time series and then, reconstruct a less noisy series which can be used for forecasting [42, 64, 112].

SSA is also a non-parametric method that can be used without making any

assumptions on processed data [112]. SSA aims to decompose the trajectory matrix into a sum of elementary matrices of rank 1 and assume that the initial object is a sum of identifiable components such as seasonality or signal and noise. Then, the aim of SSA is to reconstruct these components. The possibility to reconstruct the components is called separability of the components. [41, 42, 48].

Basic SSA is a core version of SSA that consists of embedding a time series into the space of Hankel matrices and the subsequent decomposition into rank-one matrices through use of conventional singular value decomposition. By inverting the embedding procedure, SSA yields a decomposition of the original time series into the sum of components such as a trend, oscillatory components and noise [42, 60].

Many variations of SSA are available in the literature including: multivariate SSA [43, 97, 110], complex valued SSA [43] and non-linear laplacian spectral analysis (NLSA) [36].

Multivariate SSA, or MSSA, is a natural extension of SSA for analysing multivariate time series. MSSA, similarly to SSA, has many applications such as trend extraction, causality and forecasting. MSSA is especially popular to analyse and to forecast economic and financial time series with short and long series length [110]. In the MSSA module of the hybrid model the time series of energy consumption and meteorological factors are decomposed into independent components such as additive, trend, harmonic, and random components [97].

Any real-valued SSA variation can be transferred to the complex-valued case. Complex-valued SSA forecasting and parameter estimation are straightforward extensions of the corresponding techniques for the real-valued time series [42].

In [36], the authors discuss about NLSA for time series with intermittency and

low-frequency variability. Many processes in science develop multiscale temporal and spatial patterns, with complex underlying dynamics and time-dependent external forcings. Because of the importance in understanding and predicting these phenomena, extracting the salient modes of variability empirically from incomplete observations is a problem of wide contemporary interest. NLSA is a technique for analyzing high-dimensional, complex time series that exploits the geometrical relationships between the observed data points to recover features characteristic of strongly nonlinear dynamics which are not accessible to classical SSA.

NLSA is a technique for spatiotemporal data analysis which generalizes SSA to take into account the nonlinear manifold structure of complex datasets. Through such basis functions, determined efficiently via graph-theoretic algorithms, NLSA captures intermittency, rare events, and other nonlinear dynamical features which are not accessible through linear approaches such as SSA [35]. The key principle underlying NLSA is that the functions used to represent temporal patterns should exhibit a degree of smoothness on the nonlinear data manifold a constraint absent from classical SSA [37].

Applications of SSA range from physics, mathematics, economics, finance, meteorology and oceanography as well as social science and market research. In many applications, the components extracted by SSA can be identified as trends, periodic components or noise. Many scientific papers in the last two decades have used SSA as an effective tool for analysing and forecasting time series [5]. SSA can be applied in forecasting the time series that approximately satisfy the linear recurrent relation (LRR) [43, 48, 53, 67].

Let $x_1, x_2, \ldots, x_N$ be a time series. For a given window length L $(1 < L < N)$, we construct the L-lagged vectors $X_{(i)} = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}}$, $i = 1, 2, \ldots, K$, where

$K = N - L + 1$, and compose these vectors into the matrix

$$\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = \left[X_{(1)}, \ldots, X_{(K)}\right].$$

This matrix has size $L \times K$ and is often called 'trajectory matrix'. It is a Hankel matrix, which means that all the elements along the diagonal $i+j$ =const are equal. The singular value decomposition of the matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ yields a collection of L eigenvalues and eigenvectors. For a given integer r when $1 \leq r < L$ we create a group using r largest eigenvalues and corresponding eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. The chosen eigenvectors determine an r-dimensional subspace in R^L which is denoted as S_r. The L-dimensional data $X_{(1)}, \ldots, X_{(K)}$ is then projected onto this r-dimensional subspace S_r and the subsequent averaging over the diagonals give us some Hankel matrix $\widetilde{\mathbf{X}}$, which we consider as an SSA approximation to $\mathbf{X}$. With a proper choice of r and L, the time series corresponding to $\widetilde{\mathbf{X}}$ is often used as an estimator of a signal or a trend. The main guideline for selecting r and L is to take sufficiently large, say $L \approx \frac{N}{2}$ and, if we want to extract a periodic component with known period, to take the window length to be divisible by the period, while r is chosen on the base of relations between eigenvalues and the spectral properties of eigenvectors considered as time series, see [42, 5]. We discuss in more details how to choose the parameters of SSA in Section 2.5.

Figure 2.1 explains the generic scheme of the SSA family and the main concepts. SSA has four stages namely: embedding, SVD, eigentriple grouping, and diagonal averaging.

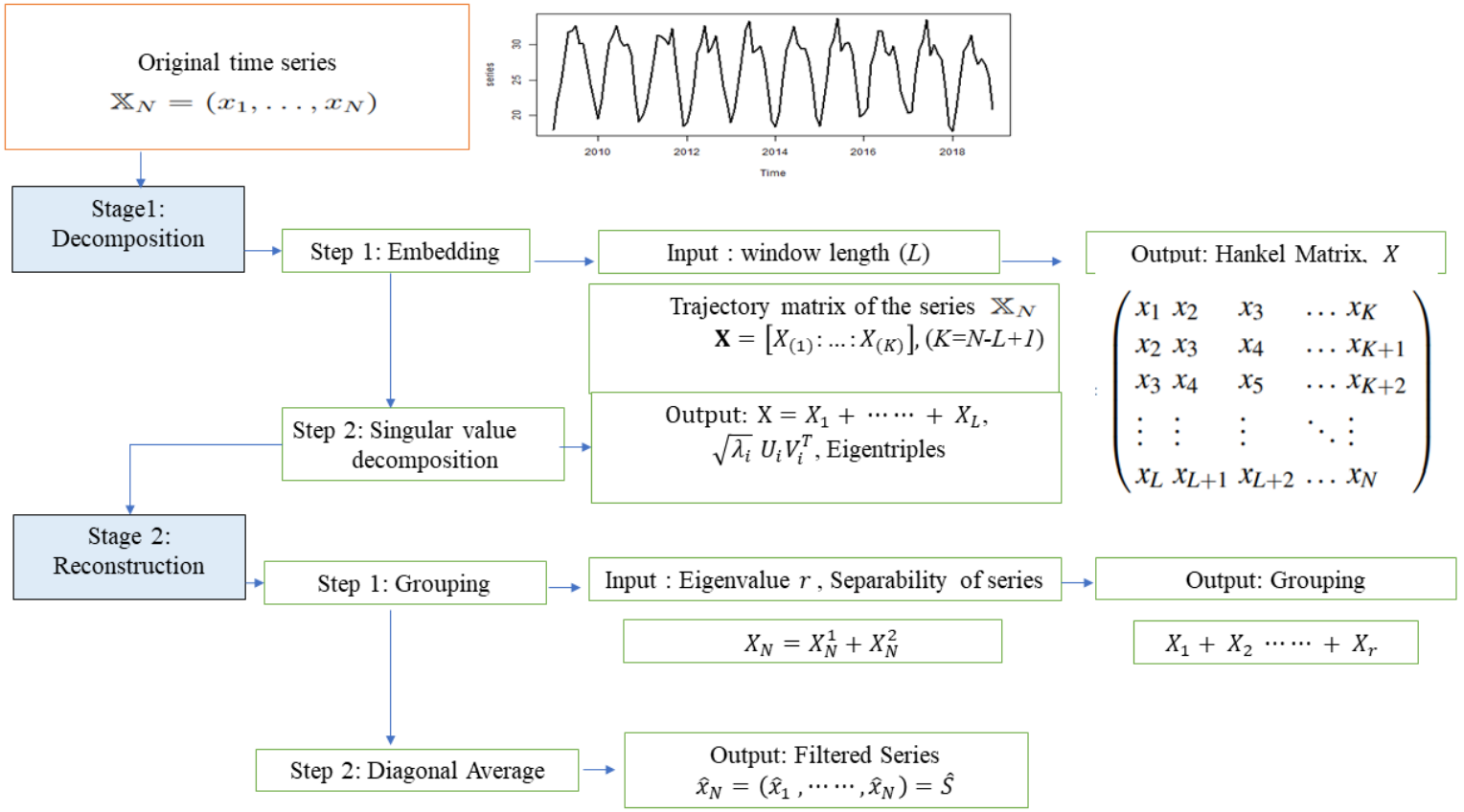

Figure 2.1: Generic scheme of the SSA family and the main concepts.

2.3 Basic SSA

Basic SSA is a variant of SSA that can be used for the analysis of one dimensional time series, where the decomposition into rank-one matrices can be performed into four steps. The theory of Basic SSA is detailed and explained in [42].

SSA decomposes the original time series into the sum of a small number of interpretable components, such as a slowly varying trend, oscillatory components and noise. SSA consists of two complementary stages: (1) the decomposition which includes embedding and singular value decomposition, and (2) the reconstruction which includes grouping and diagonal averaging [60].

Let us consider a noisy time series $\mathbb{X}_N$ with any time series length N as explained in Figure 2.1 such that $\mathbb{X}_N = (x_1, \ldots, x_N)$; the input $\mathbb{X}$, an ordered collection of N numbers

$$\mathbb{X}_N = S_N + \varepsilon_N = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}, \tag{2.1}$$

where S_N represents the signal of data and ε_N represents noise. In general, time series methods consider both the signal and noise while SSA has a different concept to separate the signal from noise. Thereafter, time series is the filtered, approximated signal component that is used to forecast future points, leaving aside the approximated ε_N. Note that the term "approximated" is used as in practice one is unable to extract the proper signal [120].

2.3.1 Method

This section explains the process of SSA in detail.

Stage 1: Decomposition

The decomposition stage is the first stage at SSA. We need to specify the window length L which is an integer $(1 < L < N)$, where N is the length of the time series.

1st step: Embedding

The embedding process is the first step in the decomposition stage. It is a mapping operation that transforms a one dimensional time series $\mathbb{X}_N = (x_1, x_2, \ldots, x_N)$ to transfer into a matrix, where N is the series length. We construct the L-lagged vectors

$$X_{(i)} = (x_i, \ldots, x_{i+L-1})^{\mathrm{T}}, \tag{2.2}$$

when $i = 1, 2, \ldots, K$, $K = N - L + 1$, T denotes transposition and compose these vectors into the matrix

$$\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = \left[X_{(1)}, \ldots, X_{(K)}\right]. \tag{2.3}$$

The series $\mathbb{X}$ is mapped to a sequence of L-lagged vector of size L, which form the trajectory matrix $\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = \mathcal{T}_{\mathrm{SSA}}(\mathbb{X}_N)$, where $\mathcal{T}$ is a linear map transforming $\mathbb{X}$ into an $L \times K$ matrix of certain structure. The trajectory matrix is the output from the embedding step which is called a Hankel matrix, where all the elements along the diagonal $i + j = const$ are constant. In one-dimensional real-valued time series, $\mathbf{X} = \left[X_{(1)}, \ldots, X_{(K)}\right]$ and $\mathcal{T} = \mathcal{T}_{\mathrm{SSA}}(\mathbb{X}_N)$ maps R^N to the space of Hankel matrices of size $L \times K$, with equal values on the anti-diagonals

$$\mathbf{X} = \left[X_{(1)}, \ldots, X_{(K)}\right] = (x_{ij})_{i,j=1}^{L,K} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{bmatrix}. \tag{2.4}$$

2nd step: Singular value decomposition

The second step of the decomposition stage is determining to get the singular values of the trajectory matrix $\mathbf{X}$. These singular values or eigenvalues contain a lot of information about the time series $\mathbb{X}_N$. Set $\mathbf{S}=\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and denote by $(\lambda_1,\ldots,\lambda_d)$ the positive *eigenvalues* of $\mathbf{S}$ taken in the decreasing order of magnitude $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ and $U_i,\ldots, U_d$ an orthonormal system of the eigenvectors of the matrix $\mathbf{S}$ corresponding to these eigenvalues; $V_i = \mathbf{X}^{\mathrm{T}}U_i/\sqrt{\lambda_i}$ are called factor vectors. At this step, we performe the SVD of the trajectory matrix:

$$\mathbf{X} = \Sigma_{i=1}^{d}\sqrt{\lambda_i}U_iV_i^{\mathrm{T}} = \mathbf{X}_1 + \ldots + \mathbf{X}_d. \tag{2.5}$$

The matrices $\mathbf{X}_i = \sqrt{\lambda_i}U_iV_i^{\mathrm{T}}$ in Equation (2.5) have rank 1; such matrices are called elementary matrices. The collection $\sqrt{\lambda_i}U_iV_i^{\mathrm{T}}$ consisting of the singular value $\sqrt{\lambda_i}$, the left singular vector U_i and the right singular vector V_i can be called i-th *eigentriple*. Note that $\lambda_i = \|\mathbf{X}_i\|_F^2$ and $\|\mathbf{X}\|_F^2 = \|\mathbf{X}_1\|_F^2 + \ldots + \|\mathbf{X}_d\|_F^2$. The contribution of i-th component $\mathbf{X}_i$ can thus be measured by $\lambda_i/\sum_j \lambda_j$. For real-world time series, $d = \mathrm{rank}\mathbf{X}$ is typically equal to $\min(L, K)$; that is, the trajectory matrix is of full rank.

Stage 2: Reconstruction

The second stage of SSA is the reconstruction which has only one parameter, the number of eigenvalues r. The reconstruction stage can be divided into two steps: grouping and diagonal averaging.

1st step: Grouping

Grouping is the first step in the reconstruction stage which involves splitting each element in matrices $\mathbf{X}_i$ into several groups (signal and noise) and reconstruct by summing the matrices within each group. Let $\boldsymbol{I} = \{i_1, \ldots, i_p\} \subset \{1, \ldots, d\}$ be a set of indices. Then, the resultant matrix $\mathbf{X}_{\boldsymbol{I}}$ corresponding to the group $\boldsymbol{I}$ is defined as $\mathbf{X}_{\boldsymbol{I}} = \mathbf{X}_{i_1} + \ldots + \mathbf{X}_{i_p}$.

Assume that a partition of the set of indices $\{1, \ldots, d\}$ into m disjoint subsets $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_m$ is specified. Then, the expansion (2.5) leads to the decomposition

$$\mathbf{X} = \mathbf{X}_{\boldsymbol{I}_1} + \ldots + \mathbf{X}_{\boldsymbol{I}_m}. \tag{2.6}$$

The procedure of choosing the sets $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_m$ is called the *eigentriple grouping*. If $m = d$ and $\boldsymbol{I}_j = \{j\}, j = 1, \ldots, d$, then, the corresponding grouping is called *elementary*. For a given group $\boldsymbol{I}$, the contribution of the component $\mathbf{X}_{\boldsymbol{I}}$ in (2.6) is measured by the share of the corresponding eigenvalues: $\sum_{i \in \boldsymbol{I}} \lambda_i / \sum_{i=1}^{d} \lambda_i$. If the original series contains signal and noise, one then, considers two groups of indices, $\boldsymbol{I}_1 = \{1, \ldots, r\}$ and $\boldsymbol{I}_2 = \{r+1, \ldots, L\}$ and associate the group $\boldsymbol{I} = \boldsymbol{I}_1$ with the signal component and the group $\boldsymbol{I}_2$ with noise.

The grouping is performed by analyzing the eigentriples, each group corresponds to an identifiable series component. The choice of several leading eigentriples corresponds to an optimal approximation of the time series, in accordance with the well-known optimality property of the SVD.

2nd step: Diagonal averaging

Diagonal averaging is the process of transferring a matrix to the form of a Hankel matrix which can be converted to a time series. Hankel matrices are an important family of matrices that play a fundamental role in diverse fields of study, such as computer science, engineering, mathematics and statistics. The Hankel matrix is a matrix with the same entries along the anti-diagonals [90]. This step is a linear operation which translates the original series' trajectory matrix onto the initial series itself to obtain the series' decomposition into several additive components [43, 66]. Diagonal averaging converts a matrix to the form of a Hankel matrix which can be subsequently transformed to a time series. If z_{ij} stands for an element of a matrix $\mathbf{Z}$, then, the k-th term of the resulting time series is obtained by averaging z_{ij} and $i + j = k + 1$. This procedure is known as Hankelization of the matrix $\mathbf{Z}$. The output of the Hankelization of a matrix $\mathbf{Z}$ is the Hankel matrix $\mathcal{H}\mathbf{Z}$, which is the trajectory matrix corresponding to the series obtained as a result of the diagonal averaging. In its turn, the Hankel matrix $\mathcal{H}\mathbf{Z}$ individually defines the series by relating the value in the diagonals to the values in the series.

In [42], the operator $\mathcal{T} = \mathcal{T}_{\mathrm{SSA}} : \mathrm{R}^N \to \mathcal{M}^{(H)}_{L,K}$ makes a correspondence between time series (collections of N numbers) and $\mathcal{M}^{(H)}_{L,K}$ the set of Hankel matrices of size $L \times K$. Since the correspondence defined by $\mathcal{T}$ is one-to-one, there exists the inverse $\mathcal{T}^{-1}$, which transfers any Hankel matrix of size $L \times K$ to a series of length N. Let us also introduce the projector $\mathit{\Pi}_{\mathcal{H}} : \mathrm{R}^{L\times K} \to \mathcal{M}^{(H)}_{L,K}$ into the space of Hankel matrices as the operator of hankelization.

$$(\mathit{\Pi}_{\mathcal{H}}\mathbf{Y})_{ij} = \sum_{(l,k)\in A_s} y_{lk}/w_s, \tag{2.7}$$

where $s = i + j - 1$, $A_s = \{(l, k) : l + k = s + 1, 1 \le l \le L, 1 \le k \le K\}$ and

$w_s = |A_s|$ denotes the number of elements in the set A_s. This corresponds to averaging the matrix elements over the 'anti-diagonals'. The weights w_s are equal to the number of series elements x_s in the trajectory matrix (2.4) and has a trapezoidal shape, decreasing towards both ends of the series. Any matrix $\mathbf{Y} \in \mathrm{R}^{L\times K}$ can be transferred to a series of length N by applying $\mathfrak{T}^{-1} \circ \Pi_{\mathcal{H}}$.

The diagonal averaging (2.7) applied to a resultant matrix $\mathbf{X}_{l_k}$ produces *reconstructed series* $\widetilde{\mathbb{X}}^{(k)} = (\widetilde{x}_1^{(k)}, \ldots, \widetilde{x}_N^{(k)}) = \mathfrak{T}_{\mathrm{SSA}}^{-1} \circ \Pi_{\mathcal{H}}(\mathbb{X}^{(k)})$. In this way, the initial series $(x_1, x_2, \ldots, x_N)$ is decomposed into a sum of m reconstructed series:

$$x_n = \sum_{k=1}^{m} \widetilde{x}_n^{(k)}, \;\; n = 1, \ldots, N. \tag{2.8}$$

The elementary grouping's reconstructed series are referred as an *elementary reconstructed series*. If the grouping is appropriate, we can decompose the data into identifiable series components. Signal plus noise or trend plus seasonality plus noise are mainly the two resulting decompositions as explained in [42, 43].

In this step, a suitable grouping leads to the decomposition in the expansion (2.6). This indicates that pairwise scalar products of distinct matrices are small, which corresponds to approximate separability. It should be noted that if L is large enough, the eigenvectors in a sense imitate the behavior of the corresponding time series components. In SSA the eigenvectors produced by slowly-varying series components are slowly-varying, the eigenvectors produced by a sine wave are sine waves with the same frequencies [41].

Separability measure

The so-called w-correlation matrix contains very helpful information that can be used for detection of separability and identification of groups. This matrix consists of weighted cosines of angles between the reconstructed time series components.

Let $w_n (n = 1, 2, \ldots, N)$ be the weights defined in Section 2.7: w_n is equal to the number of times the series element x_n appears in the trajectory matrix. Define the w-scalar product of time series of length N as $(\mathbb{Y}_N, \mathbb{Z}_N)_w = \sum_{n=1}^{N} w_n y_n z_n = \langle \mathbf{Y}, \mathbf{Z} \rangle_F$, where $\mathbf{Y}$ and $\mathbf{Z}$ are the L-trajectory matrices of the series $\mathbb{Y}_N$ and $\mathbb{Z}_N$) respectively. Define the so-called w-correlation between $\mathbb{Y}_N$ and $\mathbb{Z}_N$) as

$$\rho_w(\mathbb{Y}_N, \mathbb{Z}_N) = (\mathbb{Y}_N, \mathbb{Z}_N)_w / (\|\mathbb{Y}_N\|_{\mathbf{w}} \|\mathbb{Z}_N\|_{\mathbf{w}}) \tag{2.9}$$

Well separated components in 2.8 have weak (or zero) correlation whereas poorly separated components typically have high correlation. Therefore, looking at the matrix of w-correlations between elementary reconstructed series $\widetilde{\mathbb{X}}_N^{(i)}$ and $\widetilde{\mathbb{X}}_N^{(j)}$ one can find groups of correlated series components and use this information for the subsequent grouping. One of the main rules is: 'do not include highly correlated components into different groups'. The w-correlations can also be used for checking the grouped decomposition [42].

2.3.2 Trajectory spaces and signal subspaces

In [42], the authors discussed trajectory spaces and signal subspaces in details. Let $\mathbf{X}$ be the trajectory matrix corresponding to some object $\mathbb{X}$. The *column (row) trajectory space* of $\mathbf{X}$ is the linear subspace spanned by the columns (correspondingly, rows) of $\mathbf{X}$. The term 'trajectory space' usually means 'column trajectory space'. The column trajectory space is a subspace of R^L, while the row trajectory space is a subspace of R^K . In general, for real-world data the trajectory spaces coincide with the corresponding Euclidean spaces, since they are produced by a noisy data. However, in the 'signal plus noise' model, when the signal has rank-deficient trajectory matrix, the signal trajectory space can be called 'signal subspace'. Both column and row signal subspaces can be considered; note that the dimensions of the row and column subspaces coincide.

2.3.3 Algorithm

This section explains that the algorithms of Basic SSA have presented by writing down the algorithms at the original form as shown in [42, 43, Section 2.3]. Let us have a time series $\mathbb{X}_N = (x_1, \ldots, x_N)$ with the window length L ($L \leq \frac{N}{2}$) and $K = N - L + 1$.

For the decomposition stage, input data for the whole algorithm of SSA are the window length L and the way of grouping of the elementary components X_i. However, the rule for grouping is made after the decomposition step. Therefore, the grouping becomes the input data for the reconstruction stage. For this reason, we split the algorithm into two parts.

Algorithm 1 Basic SSA: decomposition [42]

***Input*:** Time series $\mathbb{X}$ of length N, window length L.

***Output*:** Decomposition of the trajectory matrix on elementary matrices $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, where $d = \text{rank}\mathbf{X}$ and $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ $(i = 1, \ldots, d)$.

1: Construct the trajectory matrix $\mathbf{X} = \mathfrak{T}_{\mathrm{SSA}}(\mathbb{X})$.

2: Compute the SVD $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.

For the reconstruction stage, inputs have a decomposition of the trajectory matrix into a sum of rank-one matrices and the split of the rank-one components into groups.

Algorithm 2 Reconstruction [42]

***Input*:** Decomposition $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, where $\mathbf{X}_i = \sigma_i U_i V_i^{\mathrm{T}}$ and $\left\|U_i\right\| = \left\|V_i\right\| = \mathbf{1}$; partition of indices: $\{1, \ldots, d\} = \bigsqcup_{j=1}^{m} \boldsymbol{I}_j$.

***Output*:** Decomposition of the time series $\mathbb{X}$ into identifiable components $\mathbb{X} = \mathbb{X}_1 + \ldots + \mathbb{X}_m$.

1: Construct the grouped matrix decomposition $\mathbf{X} = \mathbf{X}_{\boldsymbol{I}_1} + \ldots + \mathbf{X}_{\boldsymbol{I}_m}$, where $\mathbf{X}_{\boldsymbol{I}} = \sum_{i \in \boldsymbol{I}} \mathbf{X}_i$.

2: Compute $\mathbb{X} = \mathbb{X}_1 + \ldots + \mathbb{X}_m$, where $\mathbb{X}_i = \mathfrak{T}_{\mathrm{SSA}}^{-1} \circ \mathit{\Pi}_{\mathcal{H}}(\mathbf{X}_{\boldsymbol{I}_i})$.

2.4 SSA with projection

This section explains how SSA with projection works. If one of the series elements, such as the trend of a one-dimensional series, has a parametric model (linear in parameters), a projection on a suitable subspace is performed. The purpose of using SSA with projection is to make efficient use of available data or information about a series of components. One special case of SSA with projection is SSA with centering and SSA with double centering for extraction of both constant and linear trend [42].

SSA with projection is used to create a subspace from a supporting series and project the main series onto it for an arbitrary polynomial trend. SSA with projection involves preliminary projections of the series trajectory matrix's rows and columns to given subspaces, and it can extract polynomial trends much better than Basic SSA, particularly for linear trends.

2.4.1 Method

Let $\mathbb{X}$ is a time series of length N and L is the window length, $K = N - L + 1$ and $\mathbf{X}$ is the trajectory matrix. The general form of centering can be expressed in [42, Section 2.3].

1. Calculation of a special matrix $\mathbf{C}^{(\text{center})} = \mathbf{C(X)}$ based on a priori information.
2. Computation of $\mathbf{X}^\star = \mathbf{X} - \mathbf{C}^{(\text{center})}$.
3. Construction of the SVD $= \mathbf{X}^\star = \sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}$.

After calculation, we obtain the decomposition $\mathbf{X} = \mathbf{C}^{(\text{center})} + \sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}$. Denote $E_M = (1, \ldots, 1)^{\mathrm{T}} \in \mathrm{R}^M$ the M-vector of ones. There are three types of centering that have been considered in [42, 43]:

- Single row centering $\mathbf{C}_{\text{row}}^{(\text{center})}(\mathbf{X}) = (\mathbf{X}E_K/K)E_K^{\mathrm{T}}$ corresponds to averaging by rows; that is each element of a row of $\mathbf{C}_{\text{row}}^{(\text{center})}$ consists of the average of the corresponding row of the trajectory matrix.
- Single column centering when $\mathbf{C}_{\text{col}}^{(\text{center})}(\mathbf{X}) = E_L(\mathbf{X}^{\mathrm{T}}E_L/L)^{\mathrm{T}}$ corresponds to averaging by columns.
- Double centering when $\mathbf{C}_{\text{both}}^{(\text{center})} = \mathbf{C}_{\text{row}}^{(\text{center})} + \mathbf{C}_{\text{col}}^{(\text{center})}(\mathbf{X} - \mathbf{C}_{\text{col}}^{(\text{center})}(\mathbf{X}))$.

The single centering can be thought as a projection of rows or columns of $\mathbf{X}$ on span (E_K) or span (E_L) respectively, since $E_K E_K^{\mathrm{T}}$ and $E_L E_L^{\mathrm{T}}$ are exactly the matrices of the projection operators. As a result, in SSA, centering can be thought of as initial projection of the trajectory matrix on a particular subspace; the residual matrix can then, be expanded using the SVD or another decomposition.

By generalization to projections to arbitrary spaces as shown in [46], $\mathit{\Pi}_{\text{col}} : \mathrm{R}^L \to \mathcal{L}_{\text{col}}$ and $\mathit{\Pi}_{\text{row}} : \mathrm{R}^K \to \mathcal{L}_{\text{row}}$ be orthogonal projectors, where $\mathcal{L}_{\text{col}} \in \mathrm{R}^L$ is called the column projection space and $\mathcal{L}_{row} \in \mathrm{R}^K$ is called the row projection space. For any $\mathbf{Y} \in \mathrm{R}^{L \times t}$, denote $\mathit{\Pi}_{\text{col}}(\mathbf{Y})$ is the matrix consisting of the columns which result from projections of the columns of $\mathbf{Y}$. For any $\mathbf{Y} \in \mathrm{R}^{t \times K}$, denote $\mathit{\Pi}_{\text{row}}(\mathbf{Y})$ is the matrix consisting of the rows which result from projections of the rows of $\mathbf{Y}$.

In SSA with projection, the scheme of SSA with centering is extended to arbitrary projections; that $\mathbf{C} = \mathit{\Pi}_{\text{col}}(\mathbf{X})$ for the column projection, $\mathbf{C} = \mathit{\Pi}_{\text{row}}(\mathbf{X})$ for the row projection and $\mathbf{C} = \mathit{\Pi}_{\text{both}}(\mathbf{X})$ for the double projection, where

$$\begin{aligned} \Pi_{\text{both}}(\mathbf{X}) &= \Pi_{\text{row}}(\mathbf{X}) + \Pi_{\text{col}}(\mathbf{X} - \Pi_{\text{row}}(\mathbf{X})) \\ &= \Pi_{\text{col}}(\mathbf{X}) - \Pi_{\text{row}}(\mathbf{X} - \Pi_{\text{col}}(\mathbf{X})) \\ &= \Pi_{\text{row}}(\mathbf{X}) + \Pi_{\text{row}}(\mathbf{X}) - (\Pi_{\text{col}} \circ \Pi_{\text{row}})\mathbf{X}. \end{aligned} \tag{2.10}$$

If either the column or row basis is absent (that is the space for column or row projection consists of zero) then, we formally set the corresponding projector to be zero operator implying $\mathbf{C} = \Pi_{\text{both}}(\mathbf{X})$ in any mode.

The decomposition form which has provided by SSA with projection is

$$\mathbf{X} = \mathbf{C} + \sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}, \tag{2.11}$$

where $\mathbf{C} = \Pi_{\text{both}}(\mathbf{X})$ and $\sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}$ is the SVD of $\mathbf{X}^* = \mathbf{X} - \mathbf{C}$.

Without loss of generality, we assume that $\{P_i = 1, \ldots, p\}$ and $\{Q_i = 1, \ldots, q\}$ are orthonormal bases of $\mathcal{L}_{col}$ and $\mathcal{L}_{row}$. It is shown in [46, 42] that the expansion (2.11) can be represented as a sum of elementary matrices of rank 1. The matrix $\mathbf{C}$ can be considered as a sum of $q + p$ elementary matrices of the forms $\sigma_i^{(r)} \widetilde{P}_i Q_i^{\mathrm{T}}$, $i = 1, \ldots, q$ and $\sigma_i^{(c)} P_i \widetilde{Q}_i^{\mathrm{T}}$, $i = 1, \ldots, p$. The triples $\sigma_i^{(r)} \widetilde{P}_i Q_i^{\mathrm{T}}$ and $\sigma_i^{(c)} P_i \widetilde{Q}_i^{\mathrm{T}}$ have the same meaning as eigentriples. For double projection, its representation depends on the order of projections; this can be processed by the application of the row projector first. The decomposition equation can be transformed into a decomposition of $\mathbf{X}$ into a sum of $q + p + d^*$ elementary rank-one matrices, which are orthogonal with respect to the Frobenius norm $\|.\|$ by construction. As a consequence, the contribution of the projection term $\mathbf{C}$ into the decomposition is measured by $\|\mathbf{C}\|^2 / \|\mathbf{X}\|^2$. The reconstruction stage is exactly the same as in the Basic SSA method. Furthermore, before using SSA with projection, we must first

understand the behaviour of the time series.

For SSA with projection, a known series variable with a trajectory matrix $\mathbf{Y}$ must comply with projections that $\mathit{\Pi}_{\text{col}}\mathbf{Y} = \mathbf{Y}$ for column projections, $\mathit{\Pi}_{\text{row}}\mathbf{Y} = \mathbf{Y}$ for row projections, and $\mathit{\Pi}_{\text{both}}\mathbf{Y} = \mathbf{Y}$ for double projections. For column and row projections, this is true if the corresponding projection is performed on the column or row trajectory space of the known series component.

2.4.2 Algorithm

This sections considers the algorithm of SSA with projection. The difference with the reconstruction by Basic SSA is that the matrices $\mathbf{X}_i, i = 1, \ldots, q + p$, which produced by projections, should be included to the same group (number q of row-projection components, number p of column-projection components, grouping $\{1, \ldots, d\} = \bigsqcup_{j=1}^{m} \boldsymbol{I}_j$, which does not split the first $q + p$ projection components which should be at the same group).

Algorithm 3 SSA with projection: decomposition [42]

***Input*:** Time series $\mathbb{X}$ of length N, window length L, orthonormal basis of the column projection space $\{P_i = 1, \ldots, p\}$ and orthonormal basis of the row projection space $\{Q_i = 1, \ldots, q\}$. Either p or q can be zero.

***Output*:** Decomposition of the trajectory matrix on elementary matrices $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, $\mathbf{X}_i = \sigma_i U_i V_i^{\mathrm{T}}$ are either zero or rank-one matrices.

1: Construct the trajectory matrix $\mathbf{X}_i = \mathcal{T}_{\mathrm{SSA}}(\mathbb{X})$.

2: Subtract the row projection: $\mathbf{X}' = \mathbf{X} - \mathbf{C}_{\mathbf{row}}$, where $\mathbf{C}_{\mathrm{row}} = \mathit{\Pi}_{\mathrm{row}}(\mathbf{X}) = \sum_{i=1}^{q} \sigma_i^{(r)} \widetilde{P}_i Q_i^{\mathrm{T}}$, $\sigma_i^{(r)} = \|\mathbf{X} Q_i\| \, \widetilde{P}_i = \mathbf{X} Q_i / \sigma_i^{(r)}$ if $\sigma_i^{(r)} > 0$ otherwise, $\widetilde{P}_i$ is the zero vector.

3: Subtract the column projection $\mathbf{X}^\star = \mathbf{X}' - \mathbf{C}_{\mathbf{col}}$, where $\mathbf{C}_{\mathbf{col}} = \mathit{\Pi}_{\mathbf{col}}(\mathbf{X}') = \sum_{i=1}^{P} \sigma_i^{(c)} P_i \widetilde{Q}_i^{\mathrm{T}}$, $\sigma_i^{(c)} = \left\| \mathbf{X}'^{\mathrm{T}} P_i \right\|, \widetilde{Q}_i = \mathbf{X}'^{\mathrm{T}} P_i / \sigma_i^{(c)}$ if $\sigma_i^{(c)} > 0$; otherwise, $\widetilde{Q}_i$ is the zero vector.

4: Construct an SVD of the matrix $\mathbf{X}^\star$: $\mathbf{X}^\star = \sum_{i=1}^{d^\star} \mathbf{X}_i^\star$, where $\mathbf{X}_i^\star = \sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}$.

5: As a result, $\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i$, where $d = q + p + d^\star$, $\mathbf{X}_i = \sigma_i^{(r)} \widetilde{P}_i Q_i^{\mathrm{T}}$ for $i = 1, \ldots, q$, $\mathbf{X}_{i+q} = \sigma_i^{(c)} P_i \widetilde{Q}_i^{\mathrm{T}}$ for $i = 1, \ldots, p$, and $\mathbf{X}_{i+q} = \sum_{i=1}^{d^*} \sqrt{\lambda_i^\star} U_i^\star (V_i^\star)^{\mathrm{T}}$ for $i = 1, \ldots, d^\star$.

Algorithm 4 SSA with projection: reconstruction [42]

***Input*:** Decomposition $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d$, $\mathbf{X}_i = \sigma_i U_i V_i^{\mathrm{T}}$, number q of row-projection components, number p of column-projection components, grouping $\{1, \ldots, d\} = \bigsqcup_{j=1}^{m} \boldsymbol{I}_j$, which does not split the first $q+p$ projection components.

***Output*:** Decomposition of the time series $\mathbb{X}$ into identifiable components $\mathbb{X} = \mathbb{X}_1 + \ldots + \mathbb{X}_m$.

1: Construct the grouped matrix decomposition $\mathbf{X} = \mathbf{X}_{\boldsymbol{I}_1} + \ldots + \mathbf{X}_{\boldsymbol{I}_m}$, where $\mathbf{X}_{\boldsymbol{I}} = \sum_{i \in \boldsymbol{I}} \mathbf{X}_i$.

2: Compute $\mathbb{X} = \mathbb{X}_1 + \ldots + \mathbb{X}_m$, where $\mathbb{X}_i = \mathcal{T}_{\mathrm{SSA}}^{-1} \circ \mathit{\Pi}_{\mathcal{H}}(\mathbf{X}_{\boldsymbol{I}_i})$.

2.5 Choice of parameters

The SSA algorithm has two parameters: the window length L and the number of singular values r. The choice of parameters L and r depends on both the structure of the time series and the forecasting aims [42, 59, 66]. The window length L defines the number of columns of the Hankel matrix and an inappropriate choice of L leads to a poor decomposition, incomplete reconstruction, and non-accurate forecasting [39, 141, 87]. The parameter r should correspond to the rank of the signal [112].

The window length L and the number of singular values r determining the subspace L (for extraction of either trend or periodic components) is chosen automatically in [39]. Moreover in [62] the authors evaluate the theory of separability between the modulated signal and the noise components that lead to determine the optimal value of the window length in SSA.

The conditions of separability provide guidelines for selecting the window length L: it should be sufficiently large ($L \sim \frac{N}{2}$) and if we want to extract a periodic component with known period, then, the window lengths, which are divisible by the period, provide better separability. There are two parameters to choose: the window length L and the group of $\boldsymbol{I}$ indices which determine the subspace L. Optimal choice of these parameters should depend on the task we are using SSA for [150].

The window length L has a great importance to the reconstruction of the trajectory (Hankel) matrix of the measured time series with a limited length N, the improper choice of L would imply an inferior the decomposition and incomplete the reconstruction and misleading results in forecasting [87, 141]. Moreover, setting

the L parameter too large could leads to the noise mixing up with the signal, and choosing L too small opens up the risk of losing some parts of the signal to the noise it that should be precise on selecting L [42]. In [78] the authors discuss that L could be half of the time series length $\frac{1}{2}$, $\frac{1}{4}$ or $\frac{3}{4}$ depending on the length of time series and setting L much shorter to achieve optimal signal-noise separation and provides better SSA forecasts. By setting $L = \frac{N+1}{2}$ where N is the length of the series we attain the minimum value for the weighted correlation (w-correlation) statistic [62]. However, there are no specific rules of selecting L because it depends on the structure of time series and the purpose of analyzing the data [70].

The second parameter r is used in the reconstruction stage [30, 56]. In [59] the authors note that the value of r is used to increase the noise in the reconstructed series. Moreover, it is to consider parts of the signal that could be ignored by choosing r smaller than what is needed. In [30] the authors discuss the relationship between L and r and how they interact with one another to affect performance. As a result, it is important to make sure that the methods used to choose the two parameters result in sufficient signal-to-noise separation. The parameter r should correspond to the rank of the signal. As noted in [39, 59] the authors argue that the value of r needs to be greater than the true value since parts of the signal can be missed but the increase of noise in the reconstructed series can be small. Expert analysis of the singular values of the trajectory matrix, weighted correlation among the components of the time series and errors of the reconstruction and forecasting are the main tools for the choosing the optimal choice of r.

The first step of the SSA algorithm provides a Hankel trajectory matrix, which plays an important role in the SSA, as the other steps depend on its structure and the extracted eigenvalues obtained from Hankel matrix. This matrix also

depends on the window length L. Certainly, the choice of parameters depends on the available data and the analysis we have to perform [87]. In [47] and [87], the authors discuss that optimal window lengths for analysis and forecast can differ and parameters for reconstruction and forecasting are not the same generally. Using different window lengths in reconstruction and forecasting increases the precision of forecasting. Moreover, window length close to $N/2$ for a series with length N provides maximum separability and gives better reconstruction with a minimum RMSE. However, the optimal window length for forecasting depends on the model and whether noise is white or red. This means that the forecasting procedure in SSA needs some modifications to consider these issues.

For this point, we are discussing what exactly do if the length of time series is very large or very small. Large time series have a more complex structure than that of shorter ones. When considering Basic SSA and time series of small length, the selection of a large window length would mix the trend and periodic components of the series. On the other hand, the selection of a small window length would result in periodic components that are not separated from each other, and therefore these lengths are not suitable. For a relatively long series, approximate separability of the components is often achieved due to the theoretical concept of asymptotic separability which holds for a rather wide class of components. It is recommended that the window length be chosen as large as possible. Nevertheless, even in the case of long series it is recommended that L be chosen such that L/N is an integer.

For short time series, other forms of SSA can outperform the basic version. When selecting the window length, it is preferable to take into account the conditions for pure (nonasymptotic) separability, if one knows that the time series has a periodic component with an integer period N (for example, if this component is a seasonal component). Better to take the window length L proportional to that period [43].

2.5.1 Rank of the trajectory matrix

In [60], the authors have discussed the rank of trajectory matrix. Considering the trajectory matrix X of dimension $L \times K$, whose entities are defined by Equation 2.4, we can say that the maximum number of components that we can obtain in decomposing the corresponding time series is equal to $L \times K$. This number is the maximum rank of the trajectory matrix and it is easy to see that the maximum rank is attainable when $L = L_{\max}$, where

$$L_{\max} = \begin{cases} \frac{N+1}{2} & \text{if } N \text{ is odd,} \\ \frac{N}{2} + 1 & \text{if } N \text{ is even.} \end{cases} \tag{2.12}$$

Note that $L_{\max}$ is the median of $1, \ldots, N$ and $L_{\max}$ is the closet integer to the half of time series.

2.6 Forecasting

SSA can be used to forecast a time series. One of the main advantages of SSA is that may provide forecasts for each individual component of the time series or the deterministic/trending component without taking into consideration variability after the reconstruction [68]. In Chapter 5 and 6, we are considering application of SSA for forecasting.

The main parametric model of SSA is the linear recurrence relation (LRR) which the time series under consideration should approximately satisfy [42, Section 3.1.1.1]. SSA can approximately handle functions that are governed by the LRR [138] and satisfies LRR [112]:

$$x_{i+d} = \sum_{k=1}^{d} a_k x_{i+d-k}, \qquad 1 \le i \le N-d, \tag{2.13}$$

of some dimension d with the coefficients $a_1, \ldots, a_d$. In SSA The recurrent coefficients $a_k = (a_{d-1}, a_{d-2}, \ldots, a_k)$ and $k = 1, \ldots, d$. In the SSA decomposition, if the original time series $\mathbb{X}_N$ satisfies a LRR, then, for any N and L there are at most d nonzero singular values in the SVD of the trajectory matrix $\mathbf{X}$; hence, 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.

The two SSA forecasting algorithms SSA-R and SSA-V are described in detail in [43, Sec 2.1.1]. If the number of terms r in the SVD of the trajectory matrix $\mathbf{X}$ is smaller than the window length L, then, the series satisfies some LRR of some dimension $d < r$. Let us formally describe the forecasting algorithm under consideration.

Algorithm input:

(a). Time series $Y_N = (y_1, \ldots, y_N)$.

(b). Window length L, $1 < L, N$.

(c). Linear space $\mathcal{L}_r \subset \mathbf{R}^L$ of dimension $r < L$. It is assumed that $e_{\mathrm{L}} \notin \mathcal{L}_r$, where $e_{\mathrm{L}} = (0, 0, \ldots, 1) \in \mathbf{R}^L$.

(d). Number M of points to forecast for.

Notations

(a). $\mathbf{X} = [X_1, \ldots, X_K]$ is the trajectory matrix of the time series Y_N.

(b). $P_1, \ldots, P_r$ is an orthonormal basis in $\mathcal{L}_r$.

(c). $\widehat{\mathbf{X}} = [\widehat{X_1} : \ldots : \widehat{X_K}] = \sum_{i=1}^{r} P_i P_i^{\mathrm{T}} \mathbf{X}$. The vector $\widehat{X_i}$ is the orthogonal projection of X_i onto the space $\mathcal{L}_r$.

(d). $\widetilde{\mathbf{X}} = \mathcal{H}\widehat{\mathbf{X}} = [\widetilde{X_1} : \ldots : \widetilde{X_K}]$ is the result of the Hankellization of the matrix $\widehat{\mathbf{X}}$. The matrix $\widetilde{\mathbf{X}}$ is the trajectory matrix of some time series $\widetilde{Y}_N = (\widetilde{y}_1, \ldots, \widetilde{y}_N)$.

(e). For any vector $Y \in \mathbf{R}^L$ we denote by $Y_\Delta \in \mathbf{R}^{L-1}$ the vector consisting of the last $L-1$ components of the vector Y, while $Y^\Delta \in \mathbf{R}^{L-1}$ is the vector consisting of the first $L-1$ components of the vector Y.

(f). We set $v^2 = \pi_1^2 + \ldots + \pi_r^2$, where π_i is the last component of the vector P_i $(i = 1, \ldots, r)$.

(g). Assume that $e_{\mathrm{L}} \notin \mathcal{L}_r$. This implies that $\mathcal{L}_r$ is not a vertical space. Then, $v^2 < 1$. It can be proved that the last component y_L of any vector $Y = (y_1, \ldots, y_L)^{\mathrm{T}} \notin \mathcal{L}_r$ is a linear combination of the first components $y_1, \ldots, y_{L-1}$

$$y_L = a_1 y_{L-1} + \ldots + a_{L-1} y_1.$$

The SSA-R forecasting algorithm can be presented as shown in [42, 43, 112].

For a chosen the window length L, the signal subspace $\mathcal{S} \in \mathrm{R}^L$ and therefore, the min-norm LRR has order $L-1$. For each column vector P_i of $\mathbf{P}_r$, denote π_i the last coordinate of P_i,$\underline{P_i} \in \mathbf{R}^{L-1}$ the vector P_i with the last coordinate removed, and $v^2 = \sum_{i=1}^{r} \pi_i^2$. Then, the elements of the vector

$$\mathcal{R} = (a_{L-1}, \ldots, a_1) = \frac{1}{1 - v^2} \sum_{i=1}^{r} \pi_i \underline{P_i} \tag{2.14}$$

provide the coefficients of the $min - norm$ governing LRR as shown in [42, Sec 3.1.1.1]

$$s_n = \sum_{i=1}^{L-1} a_i s_{n-i}. \tag{2.15}$$

2.6.1 Recurrent forecasting

The recurrent SSA forecasting is performed by means of the min-norm LRR defined in (2.14). The SSA-R algorithm as formulated in [42, Section 3.2.1.2] is as follows.

1. The time series $\mathbb{Y}_{N+M} = (y_1, \ldots, y_{N+M})$ is defined by

$$y_i = \begin{cases} \widetilde{x}_i & \text{for } i = 1, \ldots, N, \\ \sum_{j=1}^{L-1} a_j y_{i-j} & \text{for } i = 1+N, \ldots, N+M, \end{cases} \tag{2.16}$$

   where $\tilde{x}_i$ is the reconstructed value of x_i and $a_1, \ldots, a_d$ is the coefficients.

2. The numbers $(y_{N+1}, \ldots, y_{N+M})$ form the M terms of the recurrent forecast.

Thus, SSA-R is performed by the direct use of the forecasting LRR with coefficients taken from $\mathcal{R} = (a_{L-1}, \ldots, a_1)$.

We define the linear operator $\mathcal{P}_{Rec} : \mathrm{R}^L \mapsto \mathrm{R}^L$ by the formula

$$\mathcal{P}_{\mathrm{R}ec} Z = \begin{pmatrix} \bar{Z} \\ \mathcal{R}^{\mathrm{T}} \bar{Z} \end{pmatrix}, \tag{2.17}$$

where $\bar{Z}$ consists of the last $L-1$ coordinates of Z. Set

$$Y_i = \begin{cases} \widetilde{X}_i & \text{for } i = 1, \ldots, K, \\ \mathcal{P}_{\mathrm{Rec}} Y_{i-1} & \text{for } i = K+1, \ldots, K+M, \end{cases} . \tag{2.18}$$

where $\widetilde{X}_i$ is the reconstructed columns of the trajectory matrix after grouping and filtering the noise components. The matrix $\mathbf{Y} = [Y_1 : \ldots : Y_{K+M}]$ is the trajectory matrix of the series $\mathbb{Y}_{N+M}$.

In recurrent forecasting, the original series can be taken instead of the reconstructed series as the initial data for the forecasting LRR. This may be sensible only if the leading components are chosen for forecasting. This option can reduce the bias caused by the reconstruction inaccuracy but the volatility of forecasts may increase [42].

The algorithm steps for the recurrent SSA forecasting can be found in [42, Section 3.2.2].

Algorithm 5 Recurrent SSA forecasting [42]

***Input*:** Time series $\mathbb{X}$ of length N, window length L, orthonomal system of vectors $(P_i)_{i=1}^r$, forecast horizon M.

***Output*:** Forecast values $(\widetilde{x}_{N+1}, \ldots, \widetilde{x}_{N+M})$.

1: Construct the vector $\mathcal{R} = (a_{L-1}, \ldots, a_1)^{\mathrm{T}}$ of the minimal sum of squared coefficients (the so-called min-norm LRR) to $\{\boldsymbol{P}_i, i \in \boldsymbol{I}\}$.

2: Construct the reconstructed matrix $\widehat{\mathbf{X}} = \mathbf{P}\mathbf{P}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{P} = [P_1 : \ldots : P_r]$, and the reconstructed series $\widetilde{\mathbb{X}} = (\widetilde{x}_1, \ldots, \widetilde{x}_N)$ by $\widetilde{\mathbb{X}} = \mathfrak{T}_{\mathrm{SSA}}^{-1} \circ \Pi_{\mathcal{H}}(\widehat{\mathbf{X}})$.

3: Calculate the forecast values by applying the min-norm LRR: $\widetilde{x}_n = \sum_{i=1}^{L-1} a_i \widetilde{x}_{n-i}, n = N+1, \ldots, N+M$.

2.6.2 Recurrent (original) forecasting

SSA-R (original) forecasting is a modified version of recurrent forecasting. By using recurrent forecasting, the aim is to create a new series that should continue the current series based on a given decomposition. To extract the missing (last) values of vectors, the algorithm sequentially projects the incomplete embedding vectors (from either the original or the reconstructed series) onto the subspace spanned by the selected eigentriples of the decomposition. The forecasting elements are created one by one in this way.

SSA-R (original) forecasting corresponds to application of the LRR formula to initial data taken from the original series [42].

This approach works as that the m-th step of the forecast is calculated by means of the LRR $y_{n+m} = \sum_{k=1}^{L-1} a_k y_{n+m-k}$, where the starting points $y_{n-(L-2)}, \ldots, y_n$ are taken from the initial (base="initial") time series.

2.6.3 Vector forecasting

Let $\mathfrak{L}_r = \mathbf{span}(P_i, i \in \boldsymbol{I})$ and $\widehat{X_i}$ be the projection of the lagged vector X_i on $\mathfrak{L}_r$. Consider the following matrix

$$\mathit{\Pi} = \underline{\mathbf{P}}\underline{\mathbf{P}}^{\mathrm{T}} + (1 - v^2)\mathcal{R}\mathcal{R}^{\mathrm{T}}, \tag{2.19}$$

where $\underline{\mathbf{P}} = \left[\underline{P_1} : \ldots : \underline{P_r}\right]$ and $\mathcal{R}$ is defined in Equation (2.14). The matrix $\mathit{\Pi}$ defines the linear operator that performs the orthogonal projection $\mathbf{R}^{L-1} \mapsto \underline{\mathfrak{L}_r}$, where $\underline{\mathfrak{L}_r} = \mathbf{span}(\underline{P_i}, i \in \boldsymbol{I})$. Then, define linear operator $\mathcal{P}_{\mathrm{Vec}} : \mathbf{R}^L \mapsto \mathfrak{L}_r$ by the

formula

$$\mathcal{P}_{\mathrm{V}ec} Z = \begin{pmatrix} \mathit{\Pi}\bar{Z} \\ \mathcal{R}^{\mathrm{T}}\bar{Z} \end{pmatrix}, \tag{2.20}$$

where $\bar{Z}$ consists of the last $L-1$ coordinates of Z.

The *vector forecasting method* can be formulated as follows.

1. Define the vectors

$$Y_i = \begin{cases} \widehat{X}_i & \text{for } i = 1, \ldots, K, \\ \mathcal{P}_{\mathrm{V}ec} Y_{i-1} & \text{for } i = K+1, \ldots, K+M+L-1. \end{cases} \tag{2.21}$$

2. By constructing the matrix $\mathbf{Y} = [Y_1 : \ldots : Y_{K+M+L-1}]$ and making its diagonal averaging we obtain the series $y_1, \ldots, y_{N+M+L-1}$.

3. The numbers $y_{N+1}, \ldots, y_{N+M}$ form the M terms of the vector forecast.

In recurrent forecasting, we perform diagonal averaging to obtain the reconstructed series and then apply the LRR. In the vector forecasting algorithm, these steps are applied in the reverse order. The current fast implementation of the vector forecasting makes the vector forecasting comparable with recurrent forecasting it terms of the computational cost, see [42, Section 3.2.2].

Algorithm 6 Vector SSA forecasting [42]

***Input*:** Time series $\mathbb{X}$ of length N, window length L, orthonomal system of vectors $(P_i)_{i=1}^r$, forecast horizon M.

***Output*:** Forecast values $(\widetilde{x}_{N+1}, \ldots, \widetilde{x}_{N+M})$.

1: Obtain the vector $\mathcal{R} = (a_{L-1}, \ldots, a_1)^{\mathrm{T}}$ of coefficients of the min-norm LRR to $\{\boldsymbol{P}_i, i \in \boldsymbol{I}\}$.

2: Calculate the matrix Π of projection.

3: Construct the reconstructed matrix $\widehat{\mathbf{X}} = \mathbf{P}\mathbf{P}^{\mathrm{T}}\mathbf{X}$, where $\mathbf{P} = [P_1 : \ldots : P_r]$.

4: Extend the reconstructed matrix $\widehat{\mathbf{X}} = (\widehat{X}_1, \ldots, \widehat{X}_K)$ by column vectors: $\widehat{X}_n = \mathcal{P}_{\mathrm{Vec}}\widehat{X}_{n-1}$ for $n = K+1, \ldots, K+M+L-1$, where $\mathcal{P}_{\mathrm{Vec}}$ is given in Equation (2.20) and uses Π and $\mathcal{R}$. Denote the extended matrix $\widehat{\mathbf{X}}_{\mathbf{ext}} \in \mathrm{R}^{L\times(K+M+L-1)}$.

5: Obtain the extended reconstructed series $\widetilde{\mathbb{X}}_{\mathbf{ext}} = (\widetilde{x}_1, \ldots, \widetilde{x}_{N+M+L-1})$ as $\widetilde{\mathbb{X}}_{\mathbf{ext}} = \mathcal{T}_{\mathrm{SSA}}^{-1} \circ \Pi_{\mathcal{H}}(\widehat{\mathbf{X}}_{\mathbf{ext}})$.

6: Return the forecast values $(\widetilde{x}_{N+1}, \ldots, \widetilde{x}_{N+M})$.

2.6.4 Theoretical comparison of SSA-R and SSA-V

In [60], the authors explain a theoretical comparison of SSA-R and SSA-V. We refer to the following lemma.

Lemma 1: *Considering the notations in SSA-R and SSA-V, the coefficient vector A and projection matrix Π satisfy the following equalities:*

$$A = (U^{\nabla}U^{\nabla *T})^{-}U^{\nabla}U_{\Delta}^{T}, \tag{2.22}$$

$$\Pi = U^{\nabla *T}(U^{\nabla}U^{\nabla *T})^{-}U^{\nabla}, \tag{2.23}$$

where $U^{\nabla} = \left[U_1^{\nabla}, \ldots, U_r^{\nabla}\right]$, $U_{\Delta}^{T} = [\pi_1, \ldots, \pi_r]$ and $\bar{Z}$ denotes the generalized inverse of matrix Z.

According to the Lemma 1, it can be concluded that both SSA-R and SSA-V use the same projection. Note that in SSA-R, we first perform diagonal averaging and then continue the series by LRR to obtain forecasts. However, in SSA-V, we first continue the columns by the projection matrix and then use diagonal averaging to obtain forecasts. Therefore, SSA-R allows the use of more previous data than SSA-V. In SSA-R, all entities under the main off-diagonal (the off-diagonal of a matrix running from the upper right entry) are used to obtain forecasts. There are $\frac{L(L-1)}{2}$ entities under the main off-diagonal.

In contrast, SSA-V uses only the last column which has L observations. This might be a reason why SSA-V is more robust than SSA-R. In general, the difference between SSA-R and SSA-V consists in the difference between the last column of the approximated trajectory matrix before and after diagonal averaging. If these are close to each other, then SSA-R and SSA-V perform equivalently; but if there is a significant difference one should not expect equivalent results.

2.7 Gap filling method

This section is devoted to the extension of the SSA forecasting algorithms for the analysis of time series with missing data.

2.7.1 Iterative approach

The SSA iterative gap filling algorithm were proposed in [42, 81]. The ideas of the iterative gap filling algorithm in [140] are as follows.

1. An inner-loop iteration is started by computing the leading empirical orthogonal function (EOF) of the centered, zero-padded data.
2. The inner-loop iteration is performed again on the new time series.
3. The principal component corresponding to that in EOF alone is used to obtain non-zero values in place of the missing points.
4. Correct the mean of the new time series.
5. Inner iteration has converged (proven mathematically in [12]).
6. An outer-loop iteration is performed by adding a second EOF for the reconstruction, and then, the inner iteration is repeated.
7. The embedding dimension (window width) and the number of the selected principal component are optimized by the cross-validation method.

2.7.2 Algorithm

For a collection $\mathbb{Y}$ and a set of indices P we denote by $\mathbb{Y}\Big|_p$ the part of the collection with the indices from P. Set $\mathcal{N} = \{1, \ldots, N\}$.

Algorithm 7 Iterative gap filling [42]

***Input*:** Time series $\mathbb{X}$ of length N containing gaps, set of indices of missing values P, window length L, version of SSA, series $\mathbb{G}$ of length N as the source of initial values for gaps, rank for the reconstruction r, stop criterion STOP.

***Output*:** Reconstructed series component $\widetilde{\mathbb{X}}$ with no gaps.

1: $k \leftarrow 0, \widetilde{\mathbb{G}}^{(k)}\,|_p = \mathbb{G}|_p\,, \boldsymbol{I} = \{1, \ldots, r\}$.

2: Set $\widetilde{\mathbb{X}}^{(k+1)}$ such that $\widetilde{\mathbb{X}}^{(k+1)}\Big|_{N\setminus p} = \mathbb{X}\Big|_{\mathcal{N}\setminus p}$ and that $\widetilde{\mathbb{X}}^{(k+1)}\Big|_p = \mathbb{G}^{(k)}\Big|_p$.

3: Apply the selected version of SSA with the chosen L and $\boldsymbol{I}$ to $\widetilde{\mathbb{X}}^{(k+1)}$ and obtain the reconstructed series $\mathbb{G}^{(k+1)}$.

4: $k \leftarrow k + 1$.

5: If not STOP, go to Step 2; else $\widetilde{\mathbb{X}} = \mathbb{G}^{(k)}$.

2.8 Comparing SSA and PCA

In this Section, we are considering the similarity and dissimilarity between SSA and principal component analysis (PCA).

The details of differences between SSA and PCA in [60]. Consider a data matrix as below

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np}, \end{bmatrix}. \tag{2.24}$$

where each of the n rows represents a sample, and each of the p columns gives a particular kind of feature. In order to use PCA, there are no restrictions on observations x_{ij}. But, n and p are fixed and p must be greater than 2. In contrast, univariate SSA start with a univariate vector $\mathbb{X}_N = (x_1, x_2, \dots, x_N)$ and produce data matrix

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_K \\ x_2 & x_3 & \dots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \dots & x_N, \end{bmatrix}. \tag{2.25}$$

where $k = N - L + 1$. This matrix is a Hankel matrix, means that there are restriction on the elements of the matrix. Unlike PCA where the number of rows and columns are fixed, the rows number in SSA, L, can be adjusted between 2 and $N/2$. This means that the subspaces in PCA is limited, whereas the subspaces in

SSA can be adjusted by varying the window length L. However, both SSA and PCA use SVD in their algorithms. PCA takes a data matrix as input, perform an operation and then output a resulting matrix. In contrast, SSA take a univariate vector, construct a trajectory matrix and then use a PCA process on this matrix and finally transform the results to a univariate vector.

2.9 Benchmark forecasting models

In Section 2.9, we provide a brief idea for the forecasting algorithms of autoregressive integrated moving average (ARIMA), exponential smoothing (ETS) and recurrent neural networks (RNN). Many studies are applied different models of forecasting [38]. Details on the selected benchmark models are presented in [34, 122, 14] and the algorithms of forecasting methods in [72].

2.9.1 Literature review

Section 2.9.1 provides a literature review of forecasting techniques used within the field of meteorology. Many authors emphasize the importance of forecasting in different fields of meteorological research and economics research. After an extensive review of current literature found that no single forecasting model outperforms all of them in all scenarios.

In [8], the authors examine forecasts by using meteorological variables. They compare the performance of non-causal methods: an ARIMA and RNN. They found that incorporating meteorological variables can increase predictive power. Furthermore, in [92], the authors discuss how SSA is used to reduce noise and

extract trend information from the original wind speed data, and how some models are utilized a comparison model to examine the proposed model's prediction performance. The application of SSA forecasting had superior performance in comparison to other methods. The authors discussed the two forecasting variations of SSA which are SSA-V and SSA-R, and recommend that SSA-V is more robust and provides better forecasts than SSA-R [33].

The importance of forecasting meteorological data, for example daily rainfall time series was demonstrated in [133]. The components such as non-linear trend, periodic components, noise and cyclic components were extracted from daily rainfall data. In addition, the authors forecasted the daily air temperature and precipitation time series in different sites from different climatic zones. In their forecasting methods, the authors used the autoregressive integrated moving average and the time series regression, including trend and seasonality components methodology with R software. After forecasting, models can capture the dynamics of the time series data and produce sensible forecasts [96]. In [99], the authors used automatic univariate time series forecasting methods to explore the predictability of monthly temperature and precipitation. Different forecasting methods are used including as Auto-Regressive Fractionally Integrated Moving Average (ARFIMA), ETS, ARMA, and Theta and Prophet methods. These methods are used for testing the performance of multi-step ahead forecasts.

In [75], the authors argue that forecasting of temperature is an important aspect of meteorology study and is very important in helping promote sustainable development. In [98], the authors use different models such as ARIMA combined with non-linear models like NN for getting accurate forecasting results. In [134], ARIMA is used to separate stationary and non stationary components from the climate data and forecasting the daily time series with reliability and accuracy.

In [11], the authors study the forecasting time series data by using various types of forecasting techniques such as ARIMA and NN, all of these methods are used for obtaining reliable forecasting results.

In [18, 118], the authors take a simple time series approach for modeling and forecasting daily average temperature and wind speed. Different models have been used such as ARIMA and RNN. However, they do not necessarily generate better forecasting results for all the forecasting time horizons studied. In [73, 91], SSA is used for forecasting hydrological time series. The authors also compare the performance of SSA forecasts with results from ARIMA, ETS and RNN.

2.9.2 Autoregressive integrated moving average

ARIMA is one of the most popular benchmark forecasting techniques which can be provided through the forecast package for R by using *auto.arima* function. A detailed description of the algorithm can be found in [70].

Autoregressive models are based on the idea that the current value of the series, x_t, can be explained as a function of p past values, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value [119].

According to [71] a non-seasonal ARIMA (p, d, q) process is given by

$$\phi(B)(1 - B^d)y_t = c + \theta(B)\varepsilon_t, \tag{2.26}$$

where ε_t is a white noise process with mean zero and variance σ^2, B is the backshift operator and $\phi(z)$ and $\theta(z)$ are polynomials of order p and q respectively. If $c \neq 0$, there is an implied polynomial of order d in the forecast function.

The seasonal ARIMA (p, d, q) $(P, D, Q)_m$ process is given by

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t, \tag{2.27}$$

where m is the seasonal frequency, $\Phi(z)$ and $\Theta(z)$ are the polynomials of orders P and Q respectively, each containing no roots inside the unit circle, and ε_t is white noise. If $c \neq 0$, there is an implied polynomial of order $d + D$ in the forecast function.

Not that , We use uppercase notation for the seasonal parts of the model (P, D, Q), and lowercase notation for the non-seasonal parts of the model (p, d, q).

In Chapter 5, we have used Auto Arima with order (3,0,3) for 12 month. Good models are obtained by minimizing the RMSE .

2.9.3 Exponential smoothing

ETS method is an automated forecast model that incorporates the exponential smoothing foundations and is given through the forecast package for R.

In ETS, the forecasts are made by considering weighted averages of past observations [76, 131]. A detailed description of ETS and $ets()$ function by using forecast package can be found in [70]. In brief, ETS model combines three components which are the error, trend and seasonal along with several possible options for selecting the best exponential smoothing model.

Figure 2.2 summarises the several ETS formulae that are evaluated in the forecast package to select the best model to fit the data. l_t represents the level of the series at time t, b_t denotes the slope, s_t denotes the seasonal component of the

series, and m is the length of seasonality; $\alpha, \beta, \gamma, \phi$ are smoothing parameters.

Figure 2.2 provides the different ETS formulas that have been evaluated in the forecast package to select the best possible model to fit time series.

Additive error models

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
A_d	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$y_t = \ell_{t-1}b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
M_d	$y_t = \ell_{t-1}b_{t-1}^{\phi} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1}^{\phi})$

Multiplicative error models

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A_d	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
M	$y_t = \ell_{t-1}b_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
M_d	$y_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Figure 2.2: State space equations for each of the models in ETS framework [70].

2.9.4 Recurrent Neural networks

Recurrent neural networks (RNN) is a class of artificial neural networks that can represent temporal dynamic behavior through feedback loops in neurons, have been utilized to model nonlin- ear dynamic systems and have been incorporated in the design of model predictive controllers (MPC) that optimize process performance based on RNN prediction result [145]. RNN are powerful models for time series are powerful models for sequential data (time series) [31], and they use the previous output to predict and they use the previous output to predict the next output. In this case, the networks themselves have repetitive loops. These loops, which are in the hidden neurons, allow the storing of previous input information for a while so that the system can predict future outputs. The hidden layer output is retransmitted t times to the hidden layer. The output of a recursive neuron is only sent to the next layer when the number of iterations is completed. In this case, the output is more comprehensive, and the previous information is kept for longer. Finally, the errors are returned backward to update the weights [9].

A simple RNN is essentially a collection of common neural networks arranged together, each of them transmitting a message to another. In other words, these networks have a memory that stores knowledge about the data seen, but their memory is short term and cannot maintain long-term time series. A simple recurrent network has only one internal memory—h_t—which is computed from:

$$h_t = g(Wx_t + U_f h_{t-1} + b), \tag{2.28}$$

where $g()$ denotes an activation function, U and W are flexible weight matrices of the h layer, b is a bias, and X is an input vector.

2.9.5 Long short-term memory

Long short-term memory (LSTM) is a kind of model or structure for time series that uses a special combination of hidden units, elementwise products, and sums between units to implement gates that control "memory cells" [9]. These cells are designed to retain information without modification for long periods. To predict the next step, the weight values on the network have to be updated, which requires the maintenance of information from the initial steps. A simple RNN can only learn a limited number of short-term relationships and it cannot learn long-term series. However, LSTM can learn these long-term dependencies properly, and LSTM has three gates: input, forget, and output. The forget gate is embedded to indicate how much the previous memory remembers and how much it has forgotten. For LSTM, the hidden state ht is computed as follows

$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_i), \tag{2.29}$$

$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f), \tag{2.30}$$

$$O_t = \sigma(W_O X_t + U_O h_{t-1} + b_O), \tag{2.31}$$

$$\widetilde{C}_t = \tanh(W_c X_t + U_c h_{t-1} + b_c), \tag{2.32}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t, \tag{2.33}$$

$$h_t = \tanh(C_t) * O_t, \tag{2.34}$$

where i_t, f_t and O_t are the input, forget, and output gates at time t, respectively; W_i, W_f, W_o and W_c are weights that map the hidden layer input to the three gates of input, forget, and output while U_i, U_f, U_o and U_c weights matrices map the hidden layer output to gates; b_i, b_f, b_o and bc are vectors. Moreover, C_t and h_t are the outcome of the cell and the outcome of the layer, respectively.

2.10 Metrics

This section considers the various metrics which are used to compare the forecasting results. The predictive performance of different models is estimated by comparing the observation and prediction [143]. The RMSE, the normalized root mean square error (NRMSE), the mean absolute percentage error (MAPE), the mean absolute error (MAE) are different metrics which are used to estimate the accuracy of forecasting models.

2.10.1 Root mean squared error

The RMSE is the most popular criterion to measure the error of forecasts [42, 59, 89, 142].

The RMSE of the $1, 2, \ldots, h$-step ahead forecasts with several truncation points is given by

$$\mathrm{RMSE}_{T_1}^{T_2} = \left(\frac{1}{(T_2 - T_1 + 1)h} \sum_{T=T_1}^{T_2} \sum_{j=1}^{h} (\tilde{y}_{T,j} - y_{T+j})^2 \right)^{1/2}, \tag{2.35}$$

where $\tilde{y}_{T,j}$ is the j-step ahead forecast of the truncated time series $y_1, \ldots, y_T$, the value y_{T+j} is the true value of the given time series at time $T + j$, T_1 and T_2 are the first and last truncation points, respectively.

For the task of forecasting a given time series, the automatic choice of parameters L and r relies on finding values of parameters minimising the RMSE of forecasts with desired forecasting horizonts, see [42, Sec 3.5.7] and [60, 33, 66, 105]. Further, the forecasting algorithm with automatically chosen parameters is applied to obtain future forecasts.

2.11 Chapter summary

This chapter provides a review of the SSA algorithm, the existing literature and discusses parameter selection. Then, Chapter 2 explains the algorithms for SSA with projection and SSA-based iterative approach. Three forecasting algorithms namely SSA-R, SSA-R (original) and SSA-V are introduced.

Benchmark forecasting models allow the comparison of SSA forecasting with some classic forecasting models ARIMA, ETS, and RNN. Finally, the RMSE is introduced as a measure of the accuracy of forecasting.

Chapter 3

Imputation of missing values

3.1 Introduction

This Chapter analyses different methods for addressing missing values in time series. To extract important information from a time series, the data samples must have no interruptions [17, 127].

Gaps in time series are compromise the quality of the information extracted from the original data. There are several reasons for these gaps, such as failures in measurement equipment, human failure, or technical problems such as failures in the hardware and software that store and manipulate this information. Extreme climates and economic problems can also cause gaps in time series [135].

Information without any missing values causes good results for decision makers and meets the needs of scientific researchers. However, an incomplete record leads to biased statistical results and significantly affects the quality of estimations [100].

In this study, meteorological time series contain missing data and we were faced with the problem of imputing missing values before running statistical procedures

on them. Missing data is one of the problems that frequently occurs in the data observation or data recording process [106]. Imputing missing value is a solution that gives reasonable results and several techniques for solving missing data problems are available. Choosing the wrong method can severely affect the forecasting process [77].

There are different methods for imputing missing values, such as mean, median, and mode imputation. However, simple methods are not good enough for handling missing values and could create biased results. Some argue that choosing the method for handling missing observations and damaged data can be more important than the choice of forecasting methods [77]. Advanced methods of imputing missing values fall into two categories. These methods are reconstruction and forecasting based techniques that have used to replace missing values with SSA filtering and forecast values [88, 109].

Three methods for filling gaps in time series are applied in this Chapter: the iterative approach using singular spectrum analysis (SSA) [42], the multiple regression method, and regression with lagging. The concept of filling in missing data is to a great extent similar to the concept of forecasting. The approach we consider consists of extending the structure of the extracted component to the gaps caused by missing data. In the particular situation where missing values are situated toward the end of a series, filling in gaps coincides with forecasting. Thus, the techniques created are capable of both filling in the missing values and forecasting [44]. Time series often have missing data that makes data analysis difficult and reduces the precision of the results. SSA-based methods are available for filling in the missing data [53].

When SSA is modified to permit missing data, it eliminates the need to screen, fill, and subdivide time series before using SSA and makes it possible to analyse longer

data lengths that are incomplete [114]. Model-based and parameter-dependent gap-filling methods can be used instead of relying entirely on raw data and using non parametric statistical methods. In models involving numerical boundary conditions or spectral estimation, missing data creates various problems [82].

This Chapter illustrates a variety of advanced methods for handling missing data and filling in gaps. Choosing the proper imputing techniques depends on the structure of the time series concerned. In addition, some variables have been imputed using simple methods such as imputation by zero. Imputing missing values can remove the obstacle of missing data and can often produce reasonable results.

The structure of this Chapter is as follows.

- In Section 3.1.1, we introduce the method of imputation that uses the SSA-based iterative approach.
- In Section 3.1.2, we discuss imputation using the multiple regression approach.
- In Section 3.1.3, we present the imputation by regression with lagging approach.
- In Section 3.2, we present the meteorological data of Oman.
- In Section 3.3, we discuss the results of imputation methods and analysis.
- In Section 3.4, we summarize Chapter 3.

3.1.1 Imputation using SSA-based iterative approach

One of the effective applications of SSA is imputation in time series. Several methods for imputation based on SSA have been designed for time series. The SSA iterative method can extract reliable information from some data observations with limitations such as noise and can begin to establish a prediction model based on obtained information that leads to good prospects of recovering missing data [42, 44, 77, 140].

Missing values are replaced with beginning values that are subsequently reconstructed until convergence is achieved, after which the most recently reconstructed values are considered to be imputed values [82].

Iterative SSA is a new method for filling in gaps in a wide range of geophysical records as example. These series could have gaps that are randomly distributed in time or they could contain patches of missing information. The accuracy and reliability of the method are determined by the pattern of missing data, the length of gaps in relation to the length of the dataset, and the percentage of variance captured by robust, oscillatory modes [82].

In [140], the authors discuss the main ideas of the SSA iterative interpolation process are: the leading empirical orthogonal function (EOF) of the centred, zero-padded data is computed to begin an inner-loop iteration. The algorithm is performed again on the new time series in which the principal component corresponding to that in EOF alone is used to obtain non-zero values in place of the missing points and correct the mean of the new time series. When inner iteration has converged mathematically, an outer-loop iteration is performed by adding a second EOF for reconstruction and the inner iteration is repeated. The process of the SSA iterative interpolation algorithm is available in [140, 40].

The SSA-based iterative approach to imputation in time series was proposed in [81]. This approach fills in missing values using some initial values and iteratively improves these values using SSA approximations. At each iteration, the recently computed values from the SSA approximation are inserted in place of missing entries. This approach works well when the missing entries are initially filled in using some reasonable values [12, 42, 82]. This approach is implemented in the function *igapfill* in the R package *Rssa* [83]. We applied this function with the window length $L = 120$ and the number of components r which minimizes the RMSE of retrospective forecasts [42, Sect. 3.5.7].

3.1.2 Imputation by multiple regression

The multiple regression method is used to replace incomplete values with possible solutions and create a complete dataset using multiple imputation methods [108].

In [22], the authors explain that the multiple regression method is an extension of the single imputation regression replacement method that support in order to reduce any bias. The multiple regression involves three steps: the first step is determining the imputation of missing data, the second step is running of an independent statistical analysis on the resulting individual datasets and the third step is the pooling the results of imputations.

In the regression imputation method, a linear relationship is assumed between different variables and the value of one variable changes in a linear way with the other variables. In this case, missing values are replaced by a linear regression function instead of replacing all missing data if the relationships are linear; otherwise the imputation values can be biased [19, 101]. Multiple regression imputation is applied to the data in the form of a matrix in which columns

correspond to variables and rows correspond to observations. Several subsets of complete observations are created using several regression models for variables with missing values. These models yield several predicted values and pooling those values gives a value for imputation [22]. In our research, we used multiple regression imputation by applying SPSS software and pooling five predicted values.

The first step of imputations is using values correlated to the target variable that used in a regression model to predict the values of the missing data. The second step is analyzing the imputed data and running the analysis simultaneously across each imputed dataset by specifying the data. The final step is pooling the results. Pooling generates a single output that incorporates the potential uncertainty that is inherited in the imputation process into the standard errors. In [22], the authors suggest three to ten imputations that are important to produce results and that incorporate enough variation in the prediction process.

In this study, we performed five imputations to produce suitable results. SPSS software can specify which data is imputed, automatically performs the simultaneous estimation. SPSS enables us to choose a theory-based model to make predictions across each of the imputed datasets. Because the software used five rounds of imputation to generate values for the missing data, it created five distinct datasets and estimated a theory-based model for five times. After SPSS executes the analysis, it pools the results and presents a report of the pooled output. The pooling process incorporates the uncertainty from imputed data into the estimates of the standard errors. The results can be interpreted as they would be for data that does not include imputed values. The algorithm depends on the linear regression as shown in [95, 129, 136].

The multiple linear regression model for pairs (x_j, Y_j), $j = 1, \ldots, n$, with intercept

β_0 and slope β_1 is as follows

$$Y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + \ldots + \beta_j x_j + \epsilon_j. \tag{3.1}$$

Data gaps in time series may occur in two ways: all variables may have missing values, including the dependent variable, or at least one variable may have complete data.

In equation (3.1), assume that the dependent variable y has no missing values $X = \{X_i : i = 1, \ldots, n\}$ is the set of all predicted values, X^{Miss} is a set of variables that have missing values, $\left\{X^{Miss} : X^{Miss} \subset X\right\}$ and X^{Comp} is a set of variables without any missing values, $\left\{X^{Comp} : X^{Comp} \subset X\right\}$. The variable from X^{Miss} that can be chosen as the dependent variable in the first regression iteration is selected under specific conditions: it is highly correlated with the variable y and the number of all observations that contain missing values in both the candidate variable and the y variable is predetermined.

This variable is the dependent variable; y and X^{Comp} can be independent in the regression equation. If the model is designed to impute missing values in that variable, then, the imputed variable X^{Imp} can be inserted as an independent variable and the independent variables became: y: X^{Comp} and X^{Miss}_{Imp}. Another variable from X^{Miss} can be chosen to be dependent variable, and the model can be designed again to impute missing values in the variable of interest. This procedure has been repeated until all missing values have been imputed [95]. For imputation purposes, independent variables that are highly correlated with a dependent variable with missing observations can be modelled to obtain highly plausible imputations [1].

3.1.3 Imputation by regression with lagging

In this subsection, we describe an imputation procedure using regression with lagging and in [5], the authors describe imputation by regression with lagging. Recall that our data in the next section is several time series observed at six locations of Oman and it is natural that these time series are correlated. Since the distances between some locations are quite big, changes in meteorological characteristics may occur with some lag. Also, time series may have specific periodic patterns which are related to the locations of meteorological stations. Thus, we consider the following model for a time series at the i-th location

$$y_i(t) = \beta_0 + \beta_i P(t - \lfloor t \rfloor) + \sum_{j \neq i}^{K} \beta_j y_j(t - L_j) + \varepsilon(t), \tag{3.2}$$

where $P(t)$ is the annual pattern calculated by taking the average across several years, $\lfloor t \rfloor$ stands for the integer part operation, and $\beta_0, \beta_1 \ldots, \beta_K, L_1, \ldots, L_K$ are parameters to be estimated, $j = 1, 2, 3, \ldots, K$ and K is a number of other time series without missing values.

We assume that the model (3.2) holds for short time intervals $[t_1, t_2]$ and parameters may depend on time. Since time series are correlated, there exists a problem of collinearity and therefore, we estimate the parameters such that the parameters $\beta_1 \ldots, \beta_K$ are positive. The positive constraint is needed to avoid overfitting in the presence of multicollinearity.To avoid the computational burden in the global estimation problem, we estimate parameters $L_1, \ldots, L_K$ independently. Specifically, we choose L_j to maximize the cross-correlation between $y_i(t)$ and $y_j(t)$.

There are some observations at time t that are likely to be correlated with observations at times $(t-1)$, $(t-2)$, and so on. Lagged variables have been

generated and included as predictors to capture the relationship between past and current values. Regardless of the number of predictors that were used, including lagged variables from time $(t-L)$ improves the algorithm's performance. However, including more lagged variables $(t-L-1)$ and $(t-L+1)$ shows minimal improvements in accuracy [20].

3.2 Meteorological data from Oman

Climate data are typically processed and analyzed at different low-resolution levels such as hourly, daily, weekly, monthly, and yearly. However, the analysis of high-resolution data offers a greater ability to understand the behavior of data variability and trends in nature and to detect small changes. Climate studies require complete time series data. When data is missing in climate time series, imputation must be undertaken [1].

Finding changes in climate characteristics is attracting many researchers but not much is known about climate change in Oman. The authors of [3] provide a good picture about the past climate in Oman since 1961 and simulate future climate projections that have harmful consequences such as the increase of the minimum temperature and the decrease of rainfall. In [4], the authors found trends in temperature and rainfall records at the Saiq meteorological station in Oman over the period 1979–2012. In [16], the authors discuss that the climate trends in mountain oases of northern Oman over the last three decades are reported.

Oman is a country with an arid and semi-arid climate, where nature and agriculture are very sensitive to climate changes and water is an important limiting resource [3, 117, 130]. Meteorological trends serve as indicators of climate change and

should be used by policymakers to maintain the ecosystem in the wealthy state.

We investigate the hourly temperature and humidity time series collected at six meteorological stations in Oman from 2009 to 2018. This data is provided by the Directorate General of Meteorology of Oman and has not been studied in the literature.

We consider hourly temperature (measured in centigrade), humidity (measured in %), and precipitation (measured in mm) collected at six meteorological stations in the Sultanate of Oman from 2009 to 2018. These stations are located in Khasab Airport (K), Masirah (MA), Muscat International airport (MU), Saiq (SQ), Salalah (SA), and Thumrait (TH) and are shown in Figure 3.1. We use hourly time series for imputing any missing values (24 hours in a day) for 10 years, which are 86400 sample of data. Since we have some missing values, working with hourly time series is preferred.

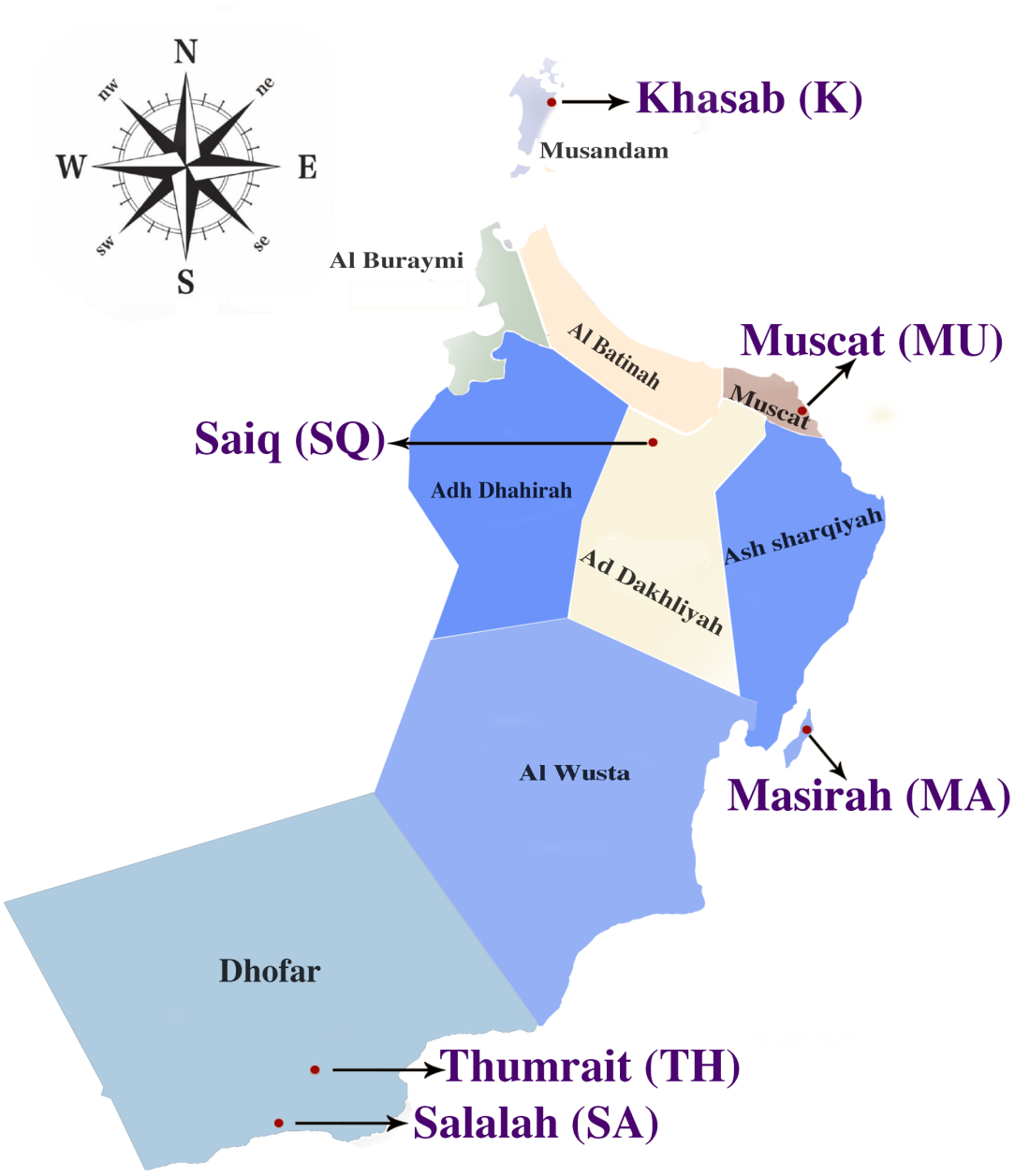

Figure 3.1: Locations of meteorological stations in Sultanate of Oman.

Let us briefly highlight the major climate and geographic specifics of Oman, see [3] for a more precise climatic description. Oman is located on the southeastern corner of the Arabian Peninsula in southwest Asia. The climate in Oman is hot and dry from May to the end of October and has mild winters, except for the south of the country. Also, the south of Oman is affected by a monsoon climate from June to September.

The individual specifics of six meteorological stations are as follows.

- Khasab (K) is located in Musandam Governorate which is a mountainous Omani peninsula and has wet summers and rainy, cold winters.
- Masirah Island (MA) is the largest island in Oman and has hot summers and warm winters.
- Muscat (MU) is a capital of Sultanate of Oman. The city lies on the Arabian Sea along the Gulf of Oman. It has very hot summers and warm winters.
- Saiq (SQ) is located in the mountain of Al Jabal Al Akhdar city. It is one of the highest points in Oman and eastern Arabia. Temperature drops during winter to below zero Celsius, with snow falling, and rises in the summer to typically 22 degrees Celsius.
- Salalah (SA) is the capital and largest city of the southern Oman governorate of Dhofar. It is very cloudy and foggy from July to August with little rain fall.
- Thumrait (TH) is a town of the Dhofar Governorate in southern Oman. It has mild summers and warm winters. Rainfall occurs from February to April, as well as June to August due to the monsoon.

3.2.1 Data cleaning

Data cleaning refers to review data for finding possible errors, incomplete information, and outliers and then, to fix the errors or problems identified. Cleaning the dataset before beginning to fill in gaps is the first step before starting any analysis. Data cleaning also involves detecting and removing errors and inconsistencies from collected data. Previous research has used cleaned data primarily for the purpose of analysing structured data [25]. There are several different ways to clean dataset such as replacing missing values and remove any error in original data.

We began by removing errors in the variable that tracked relative humidity. We replaced values of less than 5% for relative humidity with 5% and values that were greater than 100% with 100%.

Table 3.1: Descriptive statistics of the time series for precipitation, humidity and temperature meteorological stations in Oman, 2009–2018.

Series	Precipitation			Humidity			Temperature		
Station	Min	Max	Missing	Min	Max	Missing	Min	Max	Missing
K	0	38.60	1965	11	93	1961	11.1	49.9	1957
MA	0	33.20	2557	11	100	2552	11.80	46.40	2544
MU	0	37	30037	11	100	2942	12.70	48.40	29407
SQ1	0	38	21684	11	98	24267	1.30	35.70	22163
SQ2	0	28.60	54813	11	100	53505	0.10	34.20	53569
SA	0	25.60	30508	11	98	29761	11	94	1544
TH	0	11	1531	11	94	1544	5	45.50	1530

Table 3.1 presents a summary of the original time series dataset, which includes the maximum and minimum values and number of missing values for precipitation, humidity and temperature from each station in Oman for a period of ten years.

3.2.2 Hourly time series

In this part, we have illustrated the time series of temperature, humidity and precipitation which collected from six different stations in Oman.

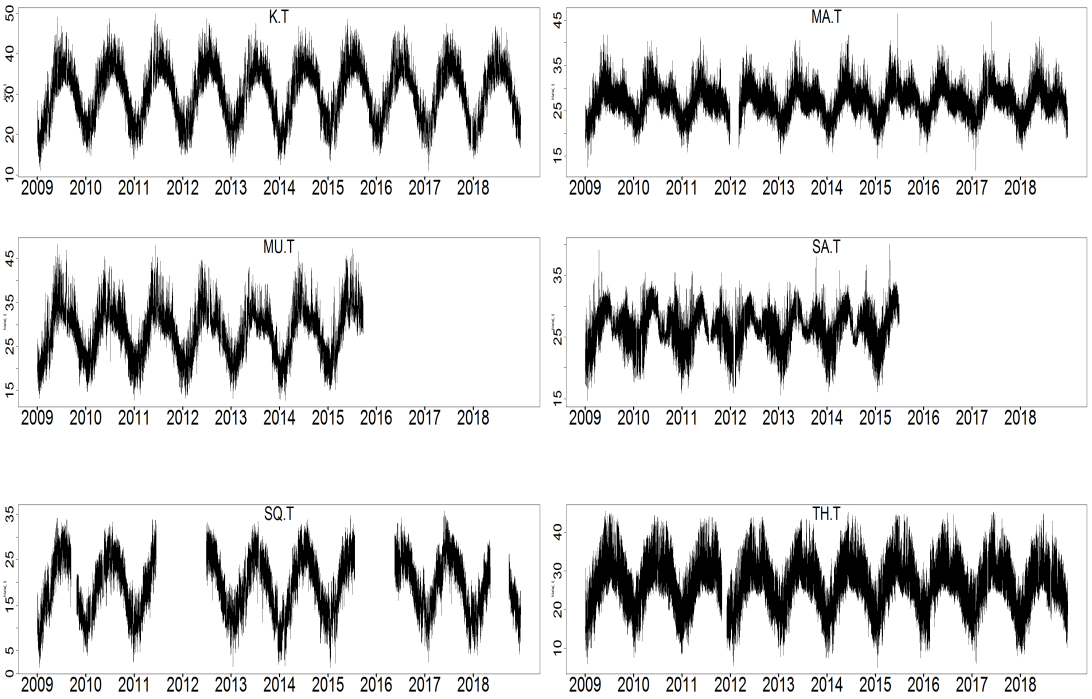

Figure 3.2: Hourly temperature at six meteorological stations.

We show hourly time series of temperature in Figure 3.2. We can see that the lowest temperature is observed at the station SQ and the highest temperature is recorded at the station K. The annual temperature pattern has a sinusoidal

shape in stations K, MU, SQ and TH and a two-mode shape in stations MA and SA. We depict hourly time series of humidity in Figure 3.3, we can observe that humidity is very volatile. The annual pattern is clearly visible for the station SA and slightly visible for stations MA and MU. The high humidity is more often observed at the station MA since this place is close to sea area.

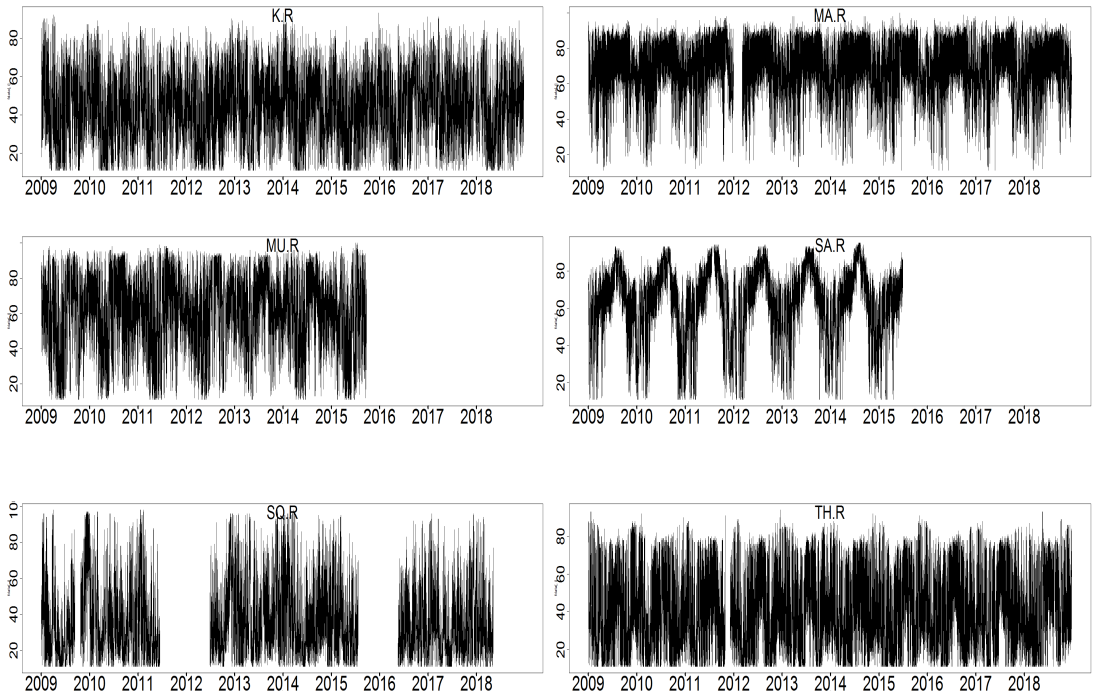

Figure 3.3: Hourly humidity at six meteorological stations.

Precipitation in Oman occurs very rarely and therefore, graphs of hourly time series of precipitation would not be appropriate. As a result, in Figure 3.4, we show the cumulative precipitation for each year to have a cyclic trend over few years. We can see that precipitation usually occurs in winter months and the larger precipitation was observed in stations SQ and K, while the lower precipitation is recorded in stations TH and SA. The station TH is remarkable due to very long periods with no precipitation.

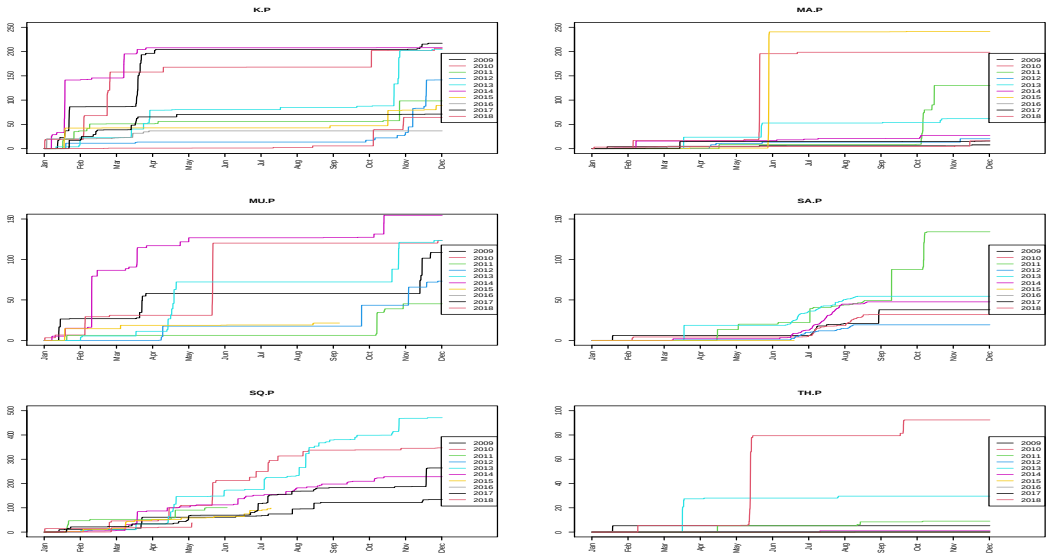

Figure 3.4: Cumulative precipitation at six meteorological stations.

3.2.3 Imputation by zero

Precipitation (mm) series have huge gaps and most of the missing values represent zero. In this case, we prefer to replace the missing data with zero because the most frequent and most common values in precipitation are zero, is as explained in detail in [86, 26]. In the zero-imputation strategy, we replace missing values with zero [84].

The following graphs are examples of imputation values by zero for three locations.

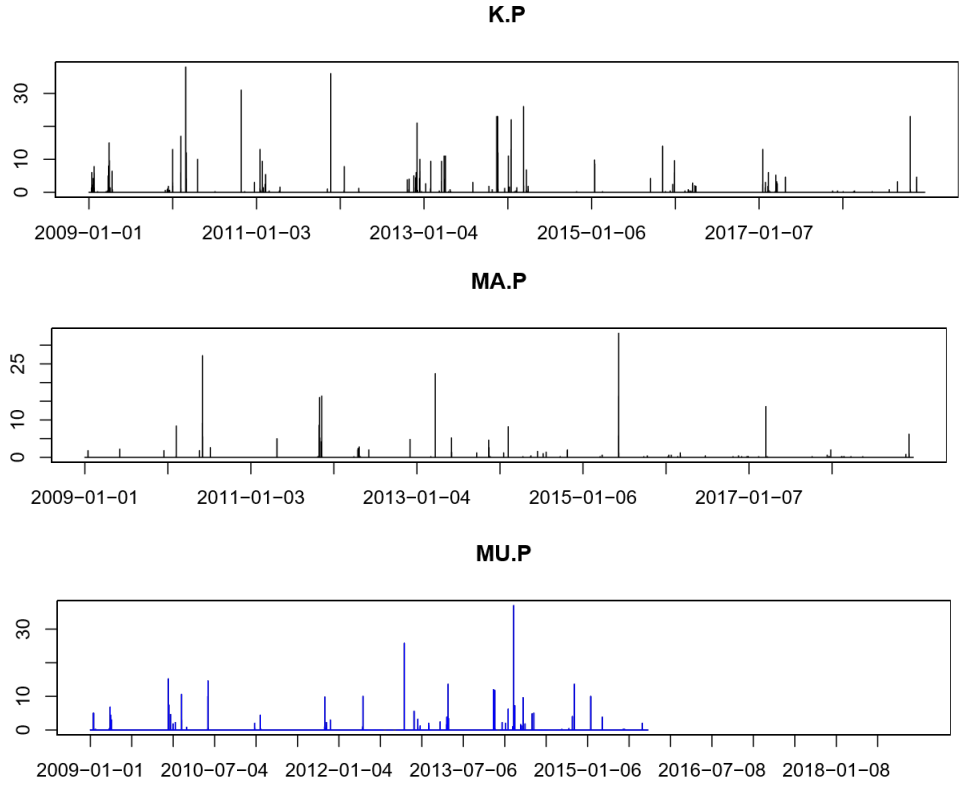

Figure 3.5: Imputation by zero for precipitation at stations K, MA and MU.

Figure 3.5 illustrates that for all stations, most values of precipitation are zero for the period 2009 to 2018. We can conclude from our time series that more than 95% of precipitation is zero for all stations.

3.3 Result of imputation methods

This section considers applications of three different imputation methods for hourly time series from 2009 to 2018.

Meteorological data in our study contain instrumental errors and missing values which are shown in Figures 3.6–3.13. To run traditional algorithms of time series analysis, we have to perform imputation. The simplest methods such as mean, median and mode imputation are not reasonable because with have data in form of time series. In this section, we study three methods of replacing missing values: (i) the SSA-based iterative approach, (ii) the regression method, and (iii) regression with lagging. Note that, the imputed values are shown in the red line and non-missing values are shown in the black line.

We can see that the SSA-based iterative approach works well for imputing short gaps, up to a couple of days, see, e.g., Figure 3.7. If missing values create a long gap then, the imputed values look almost like a clear periodic wave with a linear trend. For example, in Figure 3.6 a periodic wave has the form of sinusoid which is an aggregation of neighbour daily cycles. In contrast, imputation by regression with lagging yields more realistic values than the SSA-based iterative approach and imputation with multiple regression.

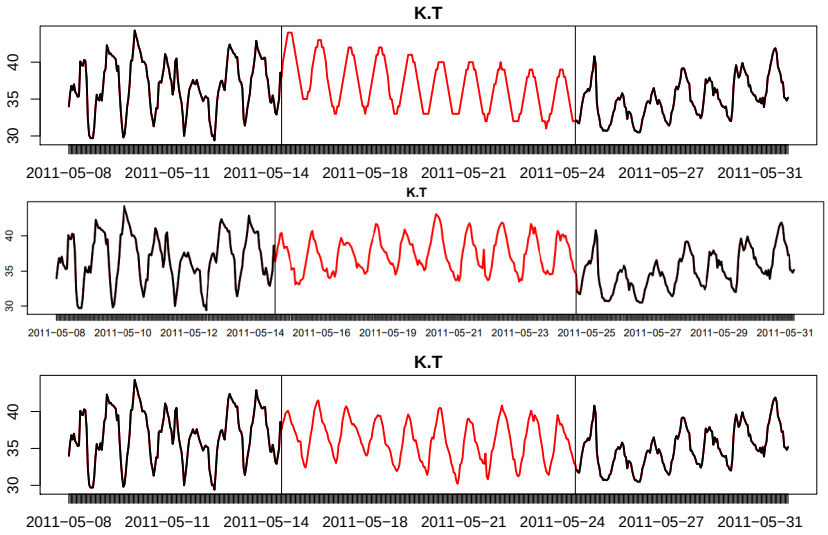

Figure 3.6: Temperature at the station K with non-missing (black) and imputed (red) values in May 2011. Top: The SSA-based iterative approach.Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

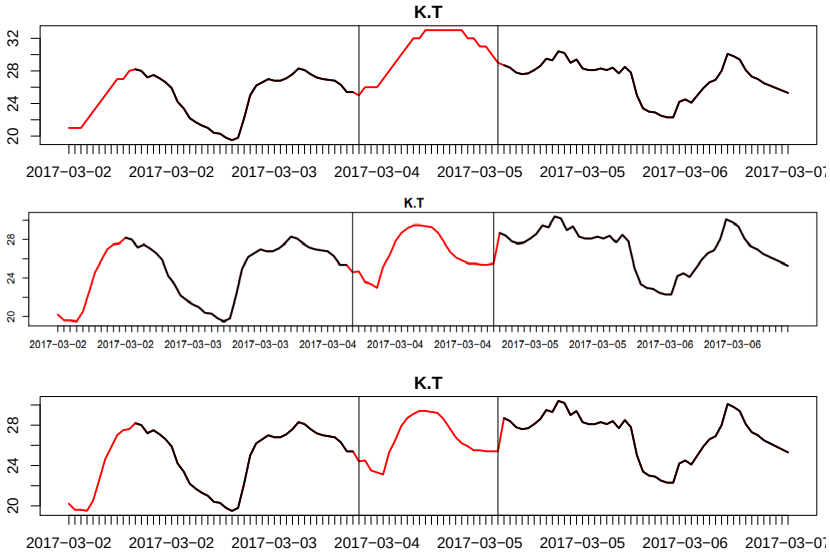

Figure 3.7: Temperature at the station K with non-missing (black) and imputed (red) values in March 2017. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

In Figure 3.8, we observe that the missing values imputed by the SSA-based iterative approach do not contain the day-to-day variations in the temperature amplitude and thus, this method cannot be recommended for filling in long gaps into time series with unstable structure. We can observe that the imputation with multiple regression and regression with lagging are more realistic than SSA-based iterative approach.

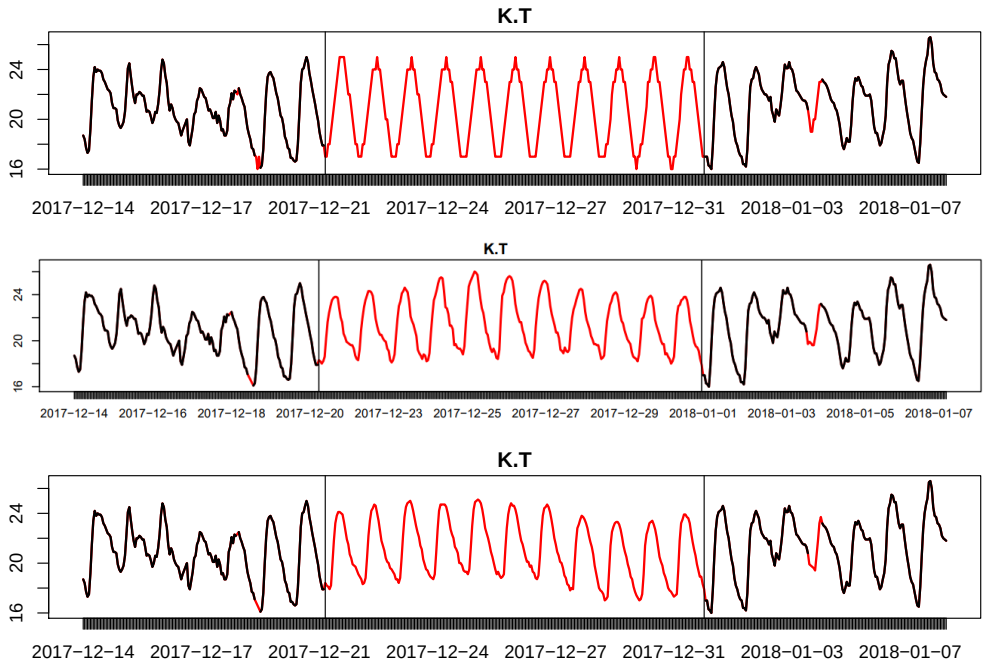

Figure 3.8: Temperature at the station K with non-missing (black) and imputed (red) values around New Year 2018. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

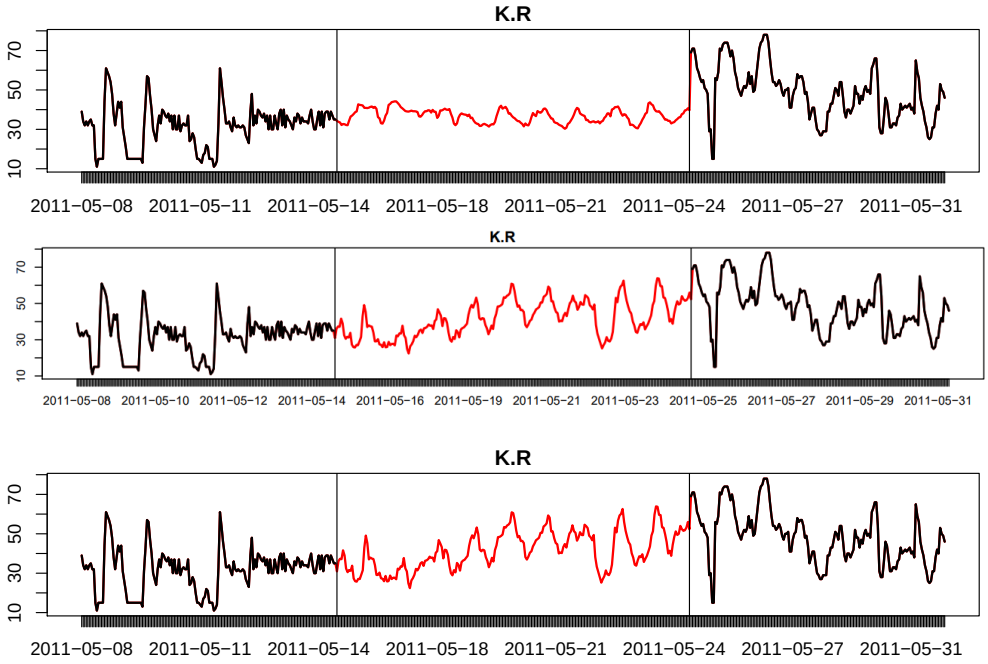

Figure 3.9: Humidity at the station K with non-missing (black) and imputed (red) values in May 2011. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

In Figure 3.9, we see that humidity is almost constant during two days before a gap with missing values and has non-regular fluctuations after the gap. Multiple regression imputation fills the gap by daily oscillations of very small magnitude. Imputation by multiple regression and regression with lagging fill by a non-regular daily oscillation of small amplitude at the top part and large amplitude at the bottom part of the gap. Thus, imputation by multiple regression and regression with lagging looks more natural in the station K.

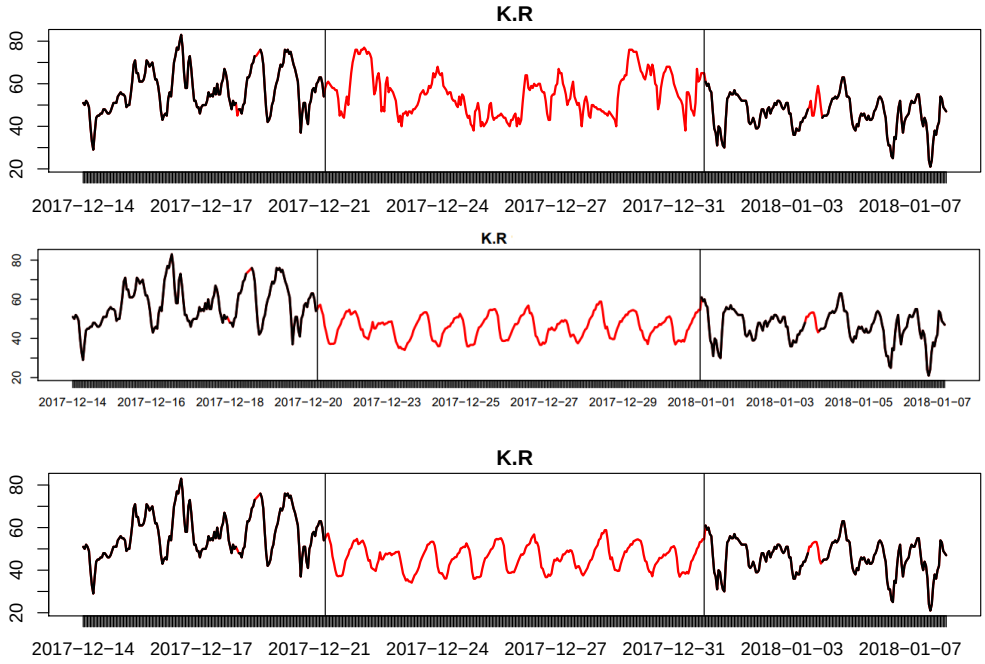

Figure 3.10: Humidity at the station K with non-missing (black) and imputed (red) values around New Year 2018. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

In Figure 3.10, we observe that humidity has non-regular oscillations around a gap with missing values. The SSA-based iterative approach fills by values with disturbed daily oscillations. But imputation by multiple regression and regression with lagging produce imputed values with clear daily oscillations. The former methods are more reliable.

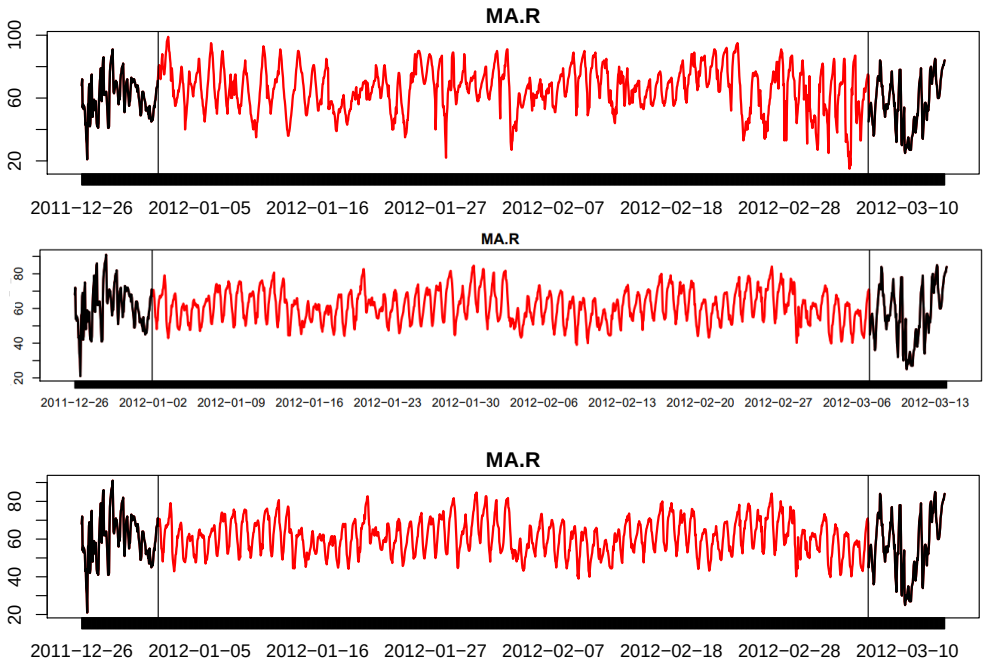

Figure 3.11: Humidity at the station MA with non-missing (black) and imputed (red) values around New Year 2012. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

In Figure 3.11, we depict humidity time series with a big gap of missing values. Here both the SSA-based iterative approach, multiple regression imputation and imputation by regression with lagging yield reasonable results.

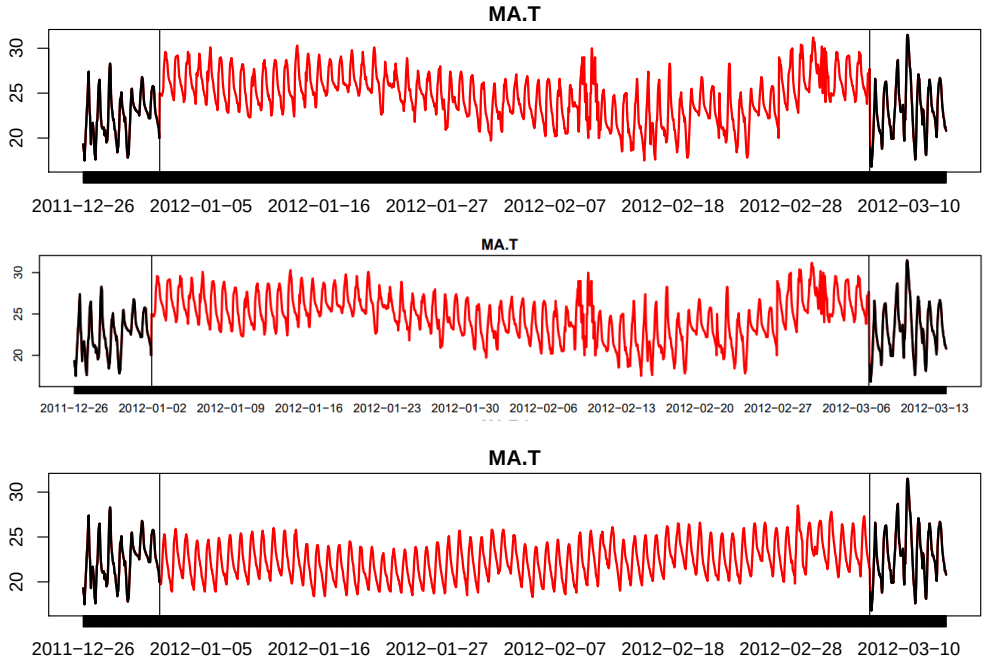

Figure 3.12: Temperature at the station MA with non-missing (black) and imputed (red) values around New Year 2012. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

In Figure 3.12, we depict temperature time series around the same big gap as shown in Figure 3.11. The SSA-based iterative approach and multiple regression imputation produce daily oscillations with rather large trend. But imputation by regression with lagging gives imputed values with a reasonable trend.

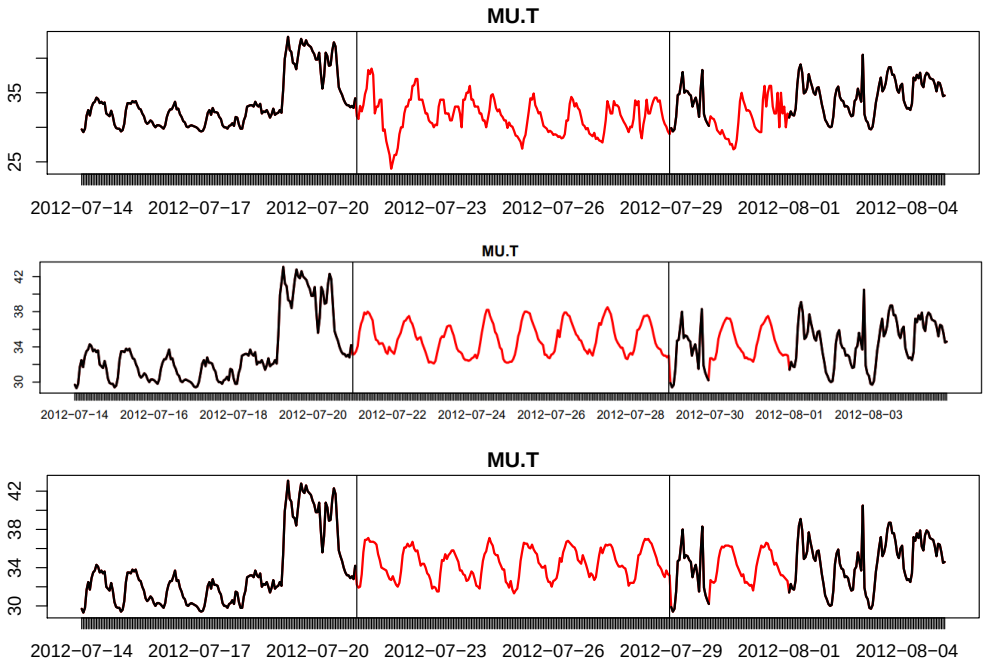

Figure 3.13: Temperature at the station MU with non-missing (black) and imputed (red) values around June 2012. Top: The SSA-based iterative approach. Middle: Multiple regression imputation. Bottom: Imputation by regression with lagging.

Figure 3.13 shows temperature time series at the station MU.The SSA-based iterative approach gives rather non-regular missing values. However, imputation by multiple regression yields more sensible daily oscillations. For short gaps of missing values, both multiple regression imputation and imputation by regression with lagging give similar results, which are not shown here due to lack of space.

Finally in Figure 3.14, we compare imputation methods using temperature time series with artificial gaps. We can observe that imputation by regression with lagging is more accurate than imputation by the SSA-based iterative approach.

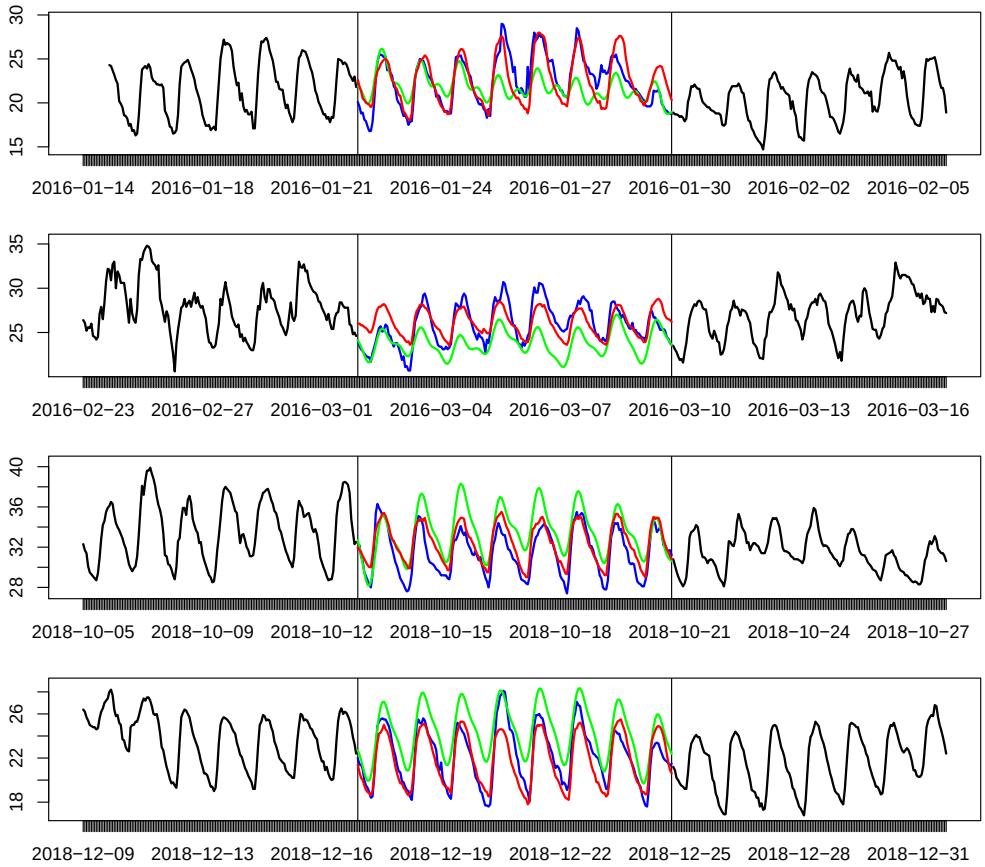

Figure 3.14: Temperature at the station K with non-missing values (black curve). The blue curve corresponds to the observed values which are artificially missed. The green curve corresponds to values obtain by the SSA-based iterative approach. The red curve corresponds to values obtained by imputation by regression with lagging.

Imputation missing values are not limited to methods which we discuss in this Chapter; there are different ways and other techniques that are used for imputation. One of these methods is imputation-based forecasting algorithms. The impute

based forecasting algorithm works by decomposing the input time series with missing values into a matrix as a preliminary step to quantify the characteristics of missing observations. This matrix is defined as the 'missing profile matrix' and includes relevant information for the impute algorithm. The missing profile matrix includes three parameters for each patch, where a patch is defined as a continuous block of missing observations with a potential minimum size of one. The first parameter is the starting index of each missing patch in the time series; the second parameter is the ending index, and the third parameter is the over-all patch size. The parameters in the matrix are used as reference points for fast indexing of the missing patches, as in [13].

3.4 Chapter summary

We analyzed hourly time series of temperature and humidity from six meteorological stations in Oman from 2009 to 2018. Dealing with imputation missing values in the dataset is an important step in the data analysis and improve the quality of the data by exploiting all variables. We have described and evaluated an imputation procedure that can be used to impute missing values in a variety of complex data structures involving many types of variables. We investigated three methods of imputation: SSA-based iterative approach, regression methods and regression with lagging. We found that imputation by regression with lagging is a more reliable and reasonable method and provides natural results for filling gaps for any length in meteorological time series for this study. Choosing an appropriate imputation method is dependent on characteristics of the dataset to evaluate. We can use an imputation by forecasting in a further study.

Chapter 4

Extracting annual oscillations and daily periodicities

4.1 Introduction

This Chapter explores ways to extract annual oscillations and daily periodicities with SSA for meteorological time series from 2009 to 2018.There are many methods for analysing trends and we are focusing on three trend tests based on many propitiates. These tests are the Mann-Kendall (MK) test, the Spearman's rho test (SR) and the Sen's innovative trend method (ITM) test. The main reason for choosing these tests is because of they do not require the time series to be normally distributed; they are robust to missing values, easy to calculate, are easy to analysis large data. Section 4.3 discusses trend tests in details.

We also used SSA to extract components such as the annual and daily oscillations in the hourly time series.

The pattern of oscillations and the periodicities of time series becomes are important for studying the influences of time series. In [133], the authors discuss how they used SSA to extract the components of rainfall such as time period, trends,

and cycles and then, used that data to forecast daily time series [149].

The authors of [31] used SSA for extracting the oscillations and the periodicities. Extraction oscillations have also been used with annual precipitation series to extract the trend and period components of annual, monthly, and hourly time series [91].

Trend analysis is used to detect trend, if the data trend increases, decreases, or exhibits no trend over time, detecting the trend is a complex process that has different characteristics [79]. The descriptive statistics of our meteorological time series that we provided in Chapter 3 show variability over time that can be cyclical in terms of seasons, trends, or other variations.

In Chapter 4, we have applied trend tests by using the time series of the temperature and humidity. We have analysis trends for individual years using the MK test and have extracted daily trends using SSA [132].

The structure of this Chapter is as follows.

- In Section 4.2, we discuss the annual oscillations for hourly time series of temperature and humidity.
- In Section 4.3, we apply the trend analysis.
- In Section 4.4, we explore the variability of daily periodicity.
- In Section 4.5, we present the summary and conclusion of the Chapter.

4.2 Annual oscillations of temperature and humidity

The annual oscillations for hourly time series can be viewed as a trend (a slowly changing component) and therefore, can be extracted as the first component of the SSA decomposition with a small window length [42]. We selected the window of length $L = 5 \times 24 = 120$ which corresponds to the number of hours in five days.

In Figure 4.1, we depict the hourly time series of the annual oscillations of temperature for six stations. We can see a small variation in temperature from year to year and random fluctuations from week to week. We also see that the annual oscillation has a sinusoidal pattern at stations K, MU, and SQ and two-saw shape pattern at stations MA, SA and TH.

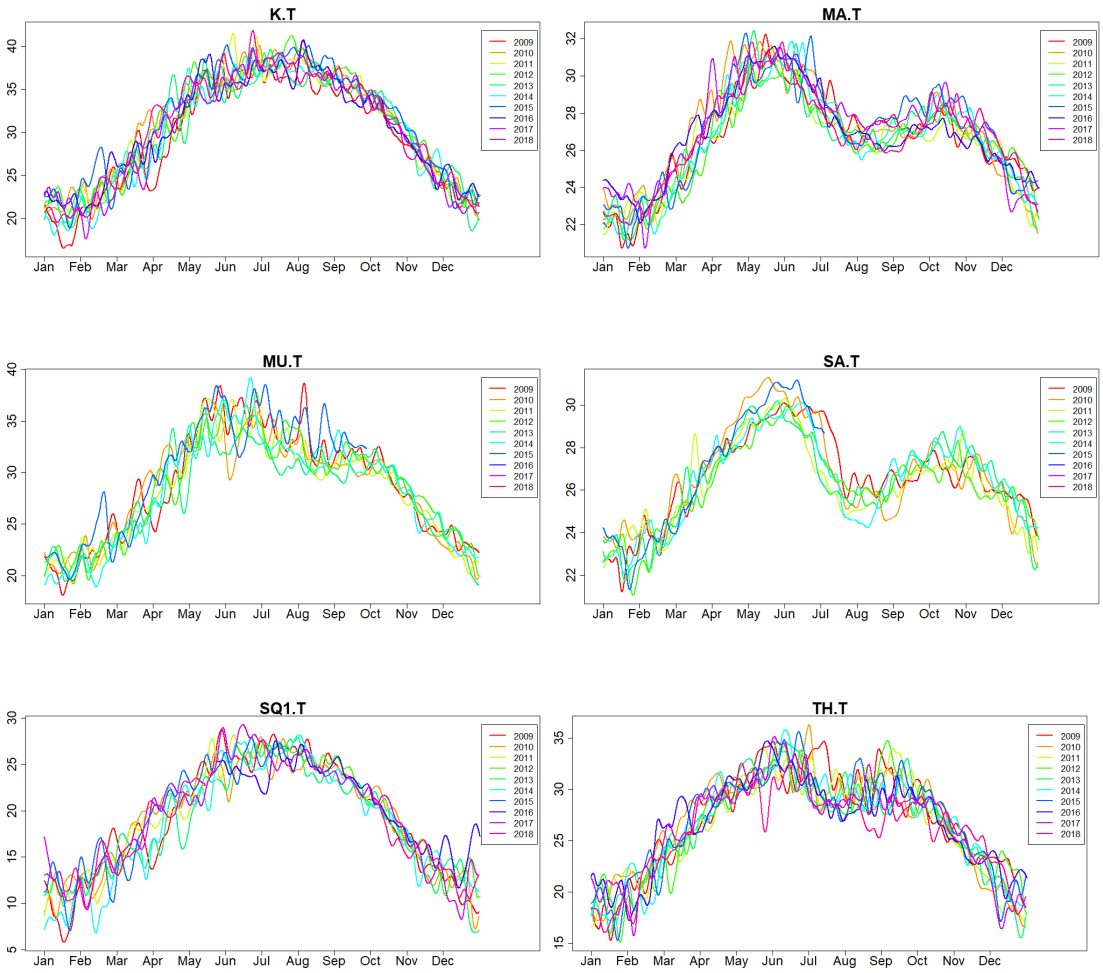

Figure 4.1: Annual oscillations of temperature at six stations.

In Figure 4.2, we depict annual oscillations in humidity. We can see very large random variations in humidity from one week to another except summer months at stations MA and SA. The lowest humidity is observed at the station SQ during almost all year, and the highest humidity occurs in July–August at the station SA.

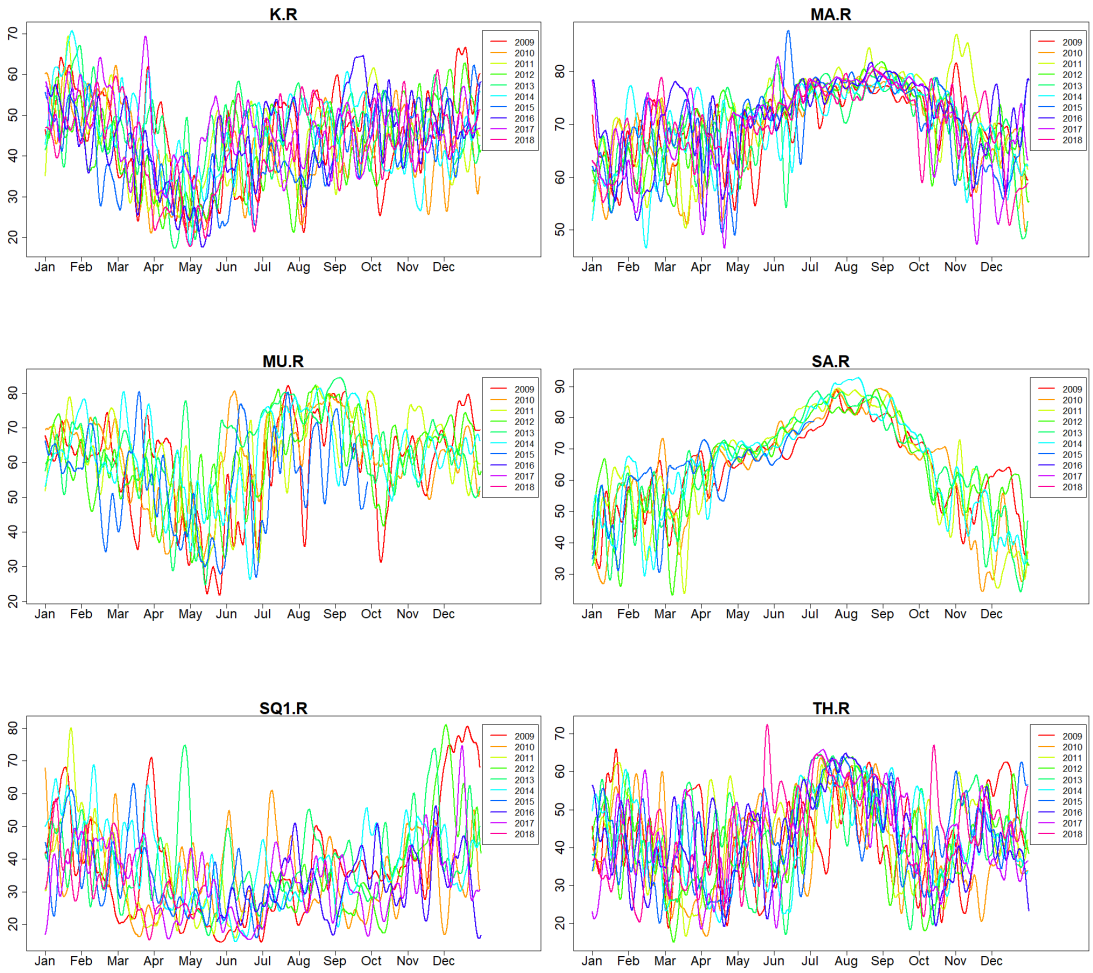

Figure 4.2: Annual oscillations in humidity at six stations.

4.3 Trend tests

Trend tests are used to investigate whether a trend in data points moves upward, downward, or is static [29]. The main purpose of trend analysis is to understand present and past climatic changes so that future forecasts can be more useful for decision makers. There are several methods for investigating trends. In this study,

we have used different methods for detecting a trend, see [80, 123].

To detect trends in temperature and humidity, we used the MK test, the SR test and the ITM test [115, 116, 117, 147]. The ITM is less restrictive than the others and is already widely applied to various meteorological time series. For example, the ITM test showed trends in monthly stream flows in northern regions of Turkey during 1964–2007, see [80], trends in monthly rainfall in a region of Ethiopia during 1980–2016 see [29], and trends in annual temperature in China during 1960–2015, see [21].

These tests are widely used to identify monotonic patterns in weather, climate, and hydrology and to measure the importance of hydrometeorological time series patterns [21, 23, 29, 80, 123]. We apply three tests individually to sequences of length 10 generated for each hour of year across 10 years. The first two tests should be applied under assumptions of independence and normality which can be assumed to be satisfied for our data. Also, the power of the tests largely depends on the length of sequences which is rather short in our study.

4.3.1 The Mann-Kendall trend test

The non parametric MK test is widely used to identify monotonic patterns in weather, climate, or hydrological data sequences and to measure the importance of hydrometeorological time series patterns [23, 49, 124].

The MK test has been applied to different kinds of data such as annual, monthly, and seasonal time series in climate time series. It is suitable for situations where the trend may be assumed to be a monotonic and normal distribution or where there is no trend. The significance of the MK test is that it can be used for

non-normal data such as seasonal data or where values are missing or have been censored. It also is an asymptotically efficient estimator [28, 93]. The MK test is calculated as shown in [6, 149].

The MK test is based on the statistic

$$S_{MK} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{sgn}(x_j - x_i), \tag{4.1}$$

where

$$\text{sgn}(x_j - x_i) = \begin{cases} +1 & if \ (x_j - x_i) > 0, \\ 0 & if \ (x_j - x_i) = 0, \\ -1 & if \ (x_j - x_i) < 0, \end{cases} \tag{4.2}$$

N is the length of a time series $x_1, \ldots, x_N$ and $\text{sgn}(\cdot)$ is the sign function.

The variance of S_{MK} is estimated as

$$Var(S_{MK}) = \frac{N(N-1)(2N+5) - \Sigma_{i=1}^{P} t_i(t_i-1)(2t_i+5)}{18}, \tag{4.3}$$

where P is the number of tied groups, the summary sign Σ indicates the summation over all tied groups, and t_i is the number of data values in tied group. A tied group is a set of total of time series having the same value.

The standardized test statistic Z for the MK test can be computed by

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & if \ \ S>0, \\ 0 & if \ \ S=0, \\ \frac{S+1}{\sqrt{Var(S)}} & if \ \ S<0. \end{cases} \tag{4.4}$$

Positive values for Z_{MK} indicate increasing trends while negative valued for Z_{MK} indicate decreasing trends. Testing trends is done at the specified significance level. When $|Z_{MK}| > Z_{1-\alpha/2}$, the null hypothesis was rejected and a significant trend existed in the time series. $Z_{1-\alpha/2}$ was obtained from the standard normal distribution table. In this study, we used a significance levels of $\alpha = 0.05$. Our null hypothesis H_0 was that there was no monotonic trend in the series. The alternate hypothesis H_a was that a trend exists and this trend can that could be positive, negative, or non-null. We used a program developed in R to generate the algorithm of the non parametric MK test.

We performed the MK test for each year of temperature and humidity series at six stations at a confidence level of 95%. We obtained the MK test statistic $-Z$ for each month of the year for the period 2009–2018. The result shows that no months had a significant trend at a confidence level of 95%.

To investigate for the presence of trends in the annual oscillations we use the MK test. We apply this test individually to sequences of length 10 generated for each hour of year across 10 years. Since the presence of a trend for one particular hour is not important, we combine several 24 p-values within each day to one p-value by taking the maximum.

We show the combined p-values of the annual oscillation of temperature and humidity in Figures 4.3 and 4.4 respectively for six stations. Also there are many

missing values in Nov-Dec for locations MU and SA and therefore, p-values for this period were not computed. These figures show that it could be just few days in year with p-values smaller than 0.1. Thus we conclude that there is no monotonic changes in temperature and humidity over the period 2009-2018.

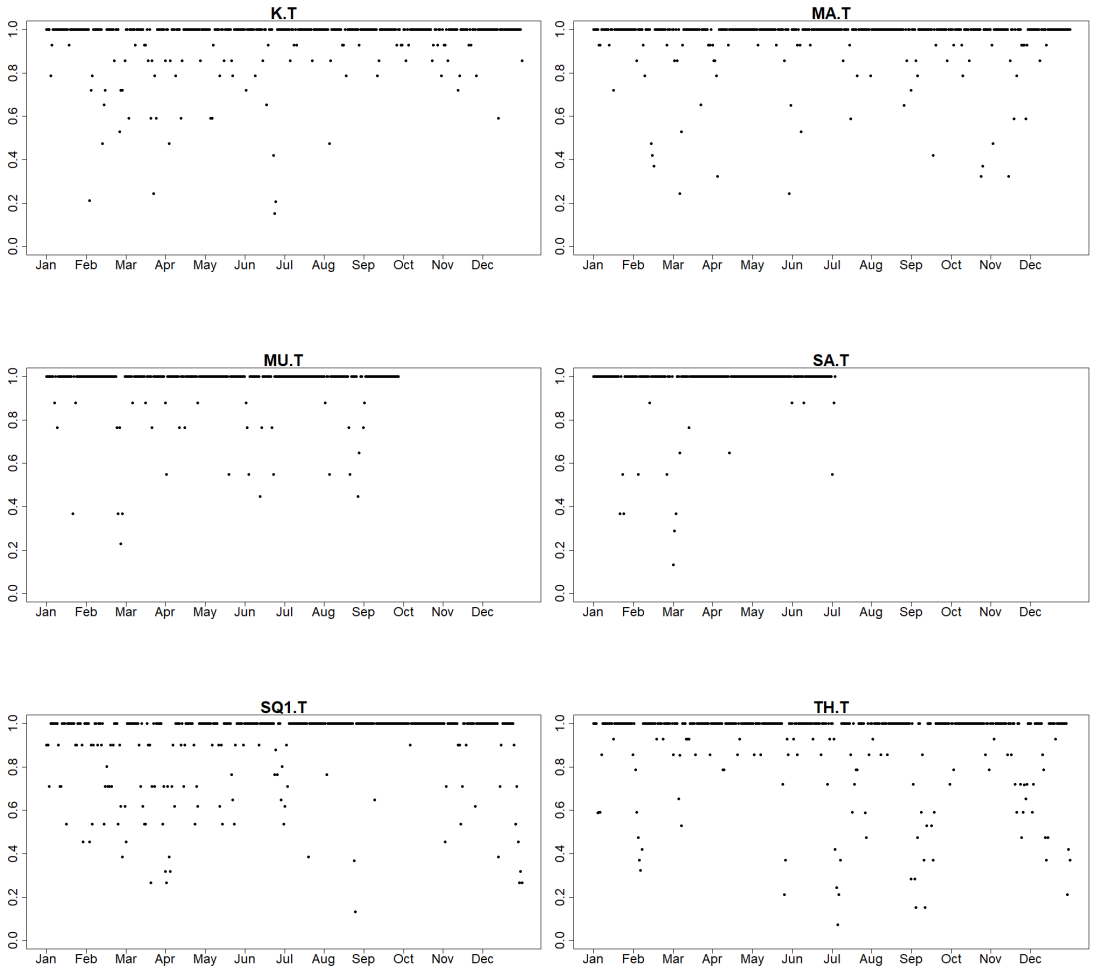

Figure 4.3: The combined p-values of the MK test for the annual oscillation of temperature in 6 locations.

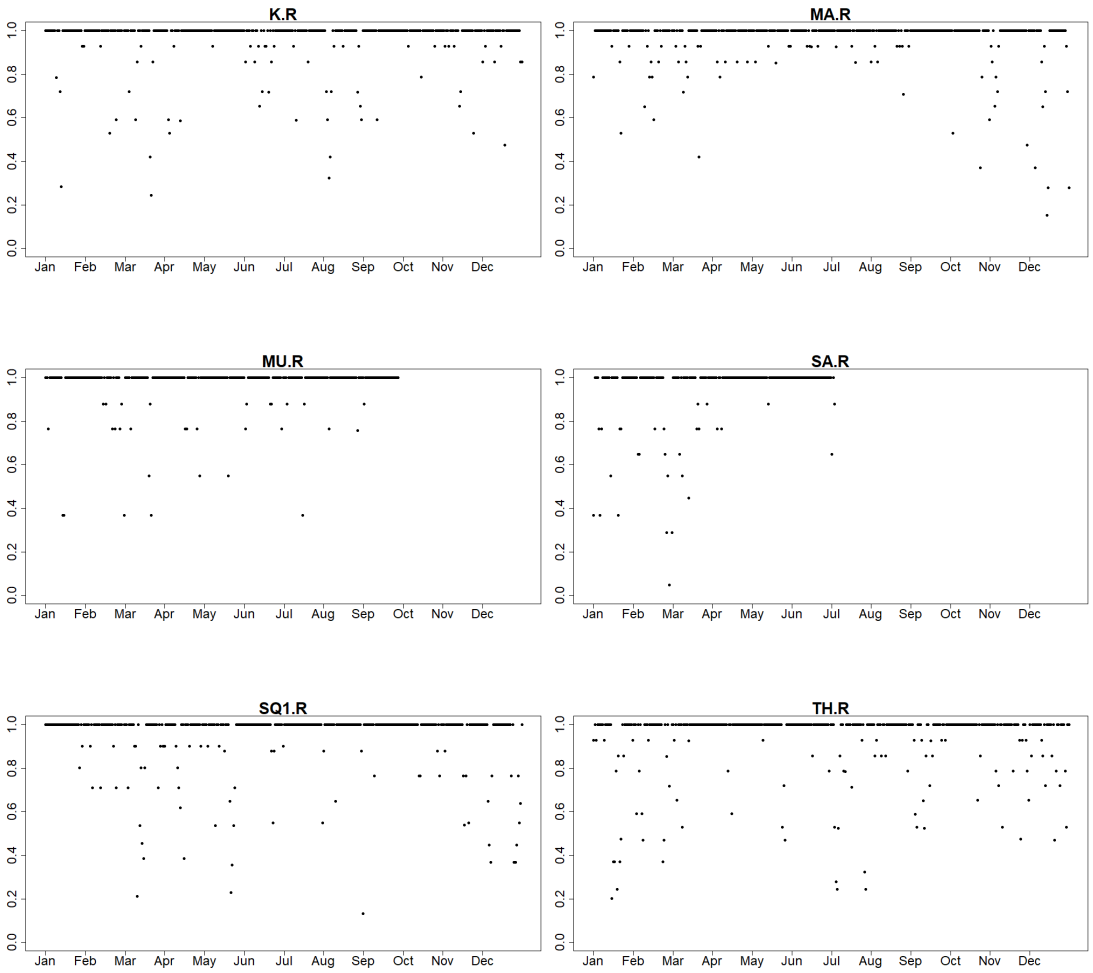

Figure 4.4: The combined p-values of the MK test for the annual oscillation of humidity in 6 locations.

4.3.2 Innovative trend method

The ITM test is fundamental for understanding how the relationship between ecosystem dynamics and climate change directly affects temperature and humidity, especially in arid and semi-arid environments, where water is an important limiting resource not just because it is scarce but also because its presence is intermittent and unpredictable [85, 117].

We used the ITM test to detect the trend in our long-term time series, a method that was proposed in [115]. This method splits the time series into two equivalent portions and sorts each portion in ascending order. The first portion of the time series is placed on the $\mathrm{X} - axis$ and the second is placed on the $\mathrm{Y} - axis$ [80].

The two halves are placed on a coordinate system. The first segment $(X_i : i = 1, 2, 3, \ldots, n/2)$ is presented in the horizontal axis while the second segment $(X_j : j = n/2 + 1, n/2 + 2, \ldots, n)$ is presented in vertical axis in the Cartesian coordinate system [29].

If a scatter plot of the time series shows a straight line at a 45°, that indicates that there is no trend. When the data points accumulate about the 45° line, that indicates an increasing trend and when they accumulate below that line, that indicates a decreasing trend. The difference in mean values between X_i and X_j provides the trend magnitude of the data series. This method can show the trend in plots of sub series data [124].

The ITM test provides a clear picture of the preliminary analysis of any trend detection study. In [116], the authors described the application of trend analysis in detail. Figure 4.5 illustrates the ITM test.

For the estimation of the trend, the S_{ITM} statistic is computed [6] as follows

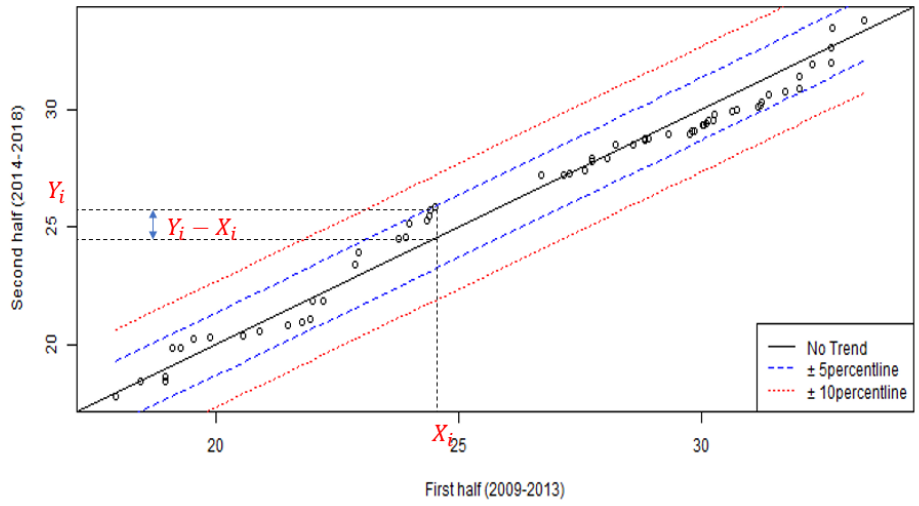

Figure 4.5: Illustration of innovative trend analysis.

$$S_{ITM} = \frac{2(\bar{X}_i - \bar{X}_j)}{N}, \tag{4.5}$$

where S_{ITM} is the base of the slope that results from the ITM test, N is is the length of a time series and $\bar{X}_i$ and $\bar{X}_j$ are the mean value of the first and second halves of the series, respectively.

The null hypothesis H_0 of no significant trend cannot be rejected if the calculated slope value s, remains below a critical value, s_{sci}, $s < s_{sci}$. The alternative hypothesis H_a of the presence of a significant trend in time series is accept if $s > s_{sci}$.

The trend indicator is given in [29] as follows and the trend indicator of ITM test is multiplied by 10 to make the scale similar with the other two tests

$$\Phi = \frac{1}{N}\sum_{i=1}^{N}\frac{10(X_j - X_i)}{u}, \tag{4.6}$$

where Φ is the trend indicator, n is is the length of a time series in the subseries,

X_i is the data series in the first half of the subseries class, X_j is the data series in the second half of the subseries class, and u is the mean of the data series in the first half of the subseries class. A positive value of Φ indicates an increasing trend while a negative value of Φ indicates a decreasing trend. However, when the scatter points closest cluster around the 1:1 straight line, it implies the non-existence of significant trend.

For monthly temperature and humidity time series of a length of 120, we depict the ITM diagnostic shows in Figures 4.6 and 4.7. Figures 4.6 and 4.7 compare the empirical distributions of the first and second halves of time series, (10 years in 12 months) which are equally divided into two segments, one from 2009 to 2013 and the other from 2014 to 2018.

In Figures 4.6 and 4.7, the trends are inconsistent which are showing both increasing and decreasing trends across stations K, MA and TH and data falling between 5% and 10% with a significance levels. The first half of the series is plotted on the horizontal axis and the second half on the vertical axis leading to the graph with a $1:1$ (45°) straight-line on it. If scatter points are above (or under) the $1:1$ line then there is a monotonic increasing (or decreasing) trend on the parent time series.

The ITM test is applied to the annual mean temperature and humidity series, which are shown in Figures 4.6 and 4.7 from 2009 to 2018. The results are for three stations: K, MA and TH; while other stations in this study have gaps and huge missing values; the ITM does not work with missing values. The results in Figure 4.6 represent the ITM test for annual mean temperature at the stations K, MA and TH from 2009 to 2018. At the station K, the Sen's slope is 0.0016 and 99% confidence interval of $(-0.0018, 0.0018)$ which indicates that there is no

significant trend with respect to 10% relative band. Similarly, the ITM test has shown that there has not significant trend with respect to 10% relative band at the station MA with Sen's slope 0.0047 at confidence interval $(-0.0009, 0.0009)$. The station TH indicates that there has not significant trend with respect to 10% relative band with Sen's slope -0.0024 at confidence interval $(-0.00242, 0.00242)$.

Figure 4.7 shows the ITM test for annual mean humidity at the stations K, MA and TH from 2009 to 2018. Figure 4.7 presents a very small positive trend at the station K with medium and high (> 55) values that shows monotonic increasing trend but very close to 10% limit with the slope of 0.007 at confidence interval $(-0.0048, 0.0048)$. The ITM test shows that there is no significant trend with respect to 10% relative band at the the station MA, where the slope is -0.0018 at confidence interval $(-0.0028, 0.0028)$. The station TH exhibits a small positive trend, the slope was 0.0136 at 99% confidence interval.

For monthly temperature and humidity time series of length 120 (12 month in 10 years), we depict the ITM test diagnostic in Figure 4.8 which compares the empirical distributions of the first and second halves of time series [116]. In Figure 4.8, the trends are small increasing at the K station for the time series of humidity and at the MA station for the time series of temperature. While other stations are between the range of 10% level of significance and fall within the 10% range from the $1:1$ line. We can observe that the ITM test diagnostic shows the absence of the trend in annual oscillations of monthly temperature and humidity at all stations.

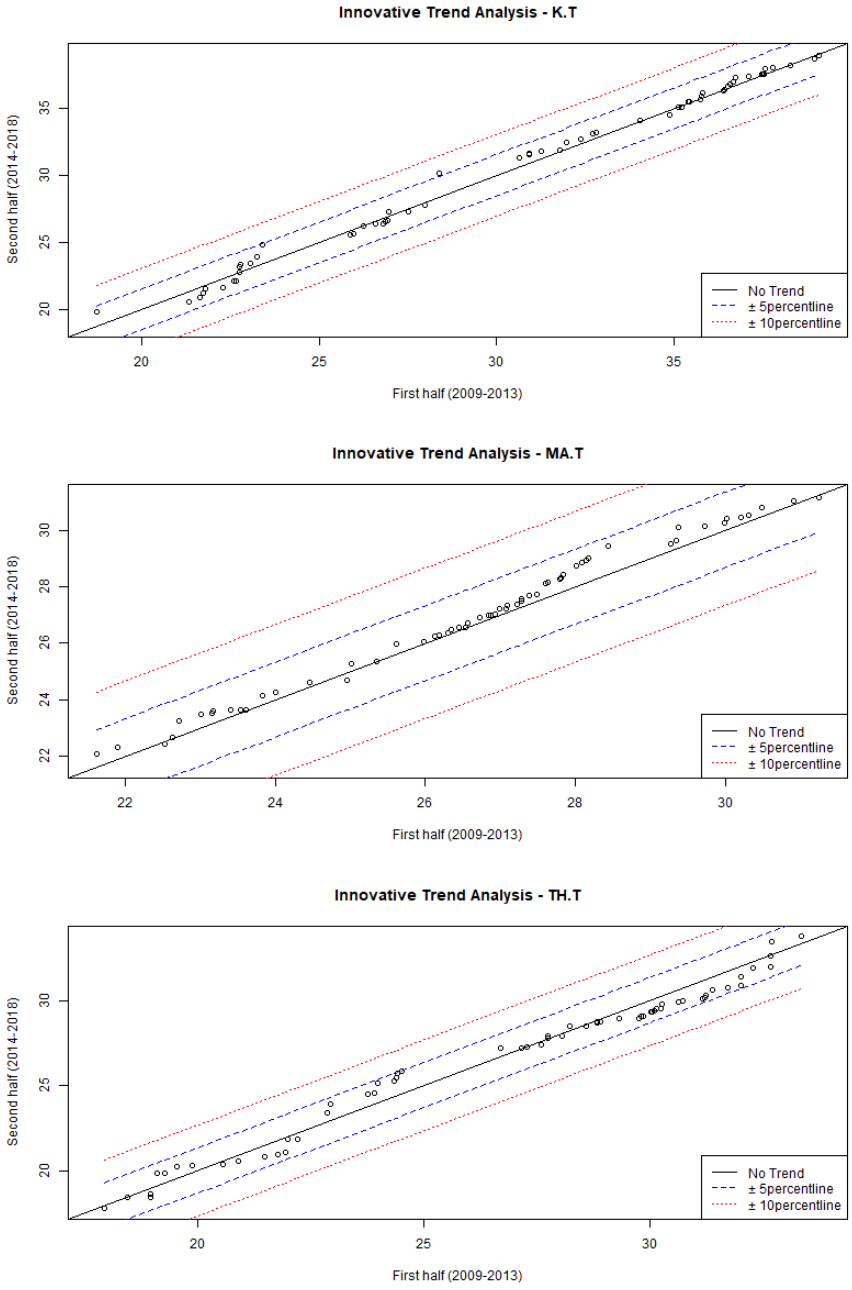

Figure 4.6: Results of the ITM test for annual mean temperature at the stations K, MA and TH from 2009 to 2018.

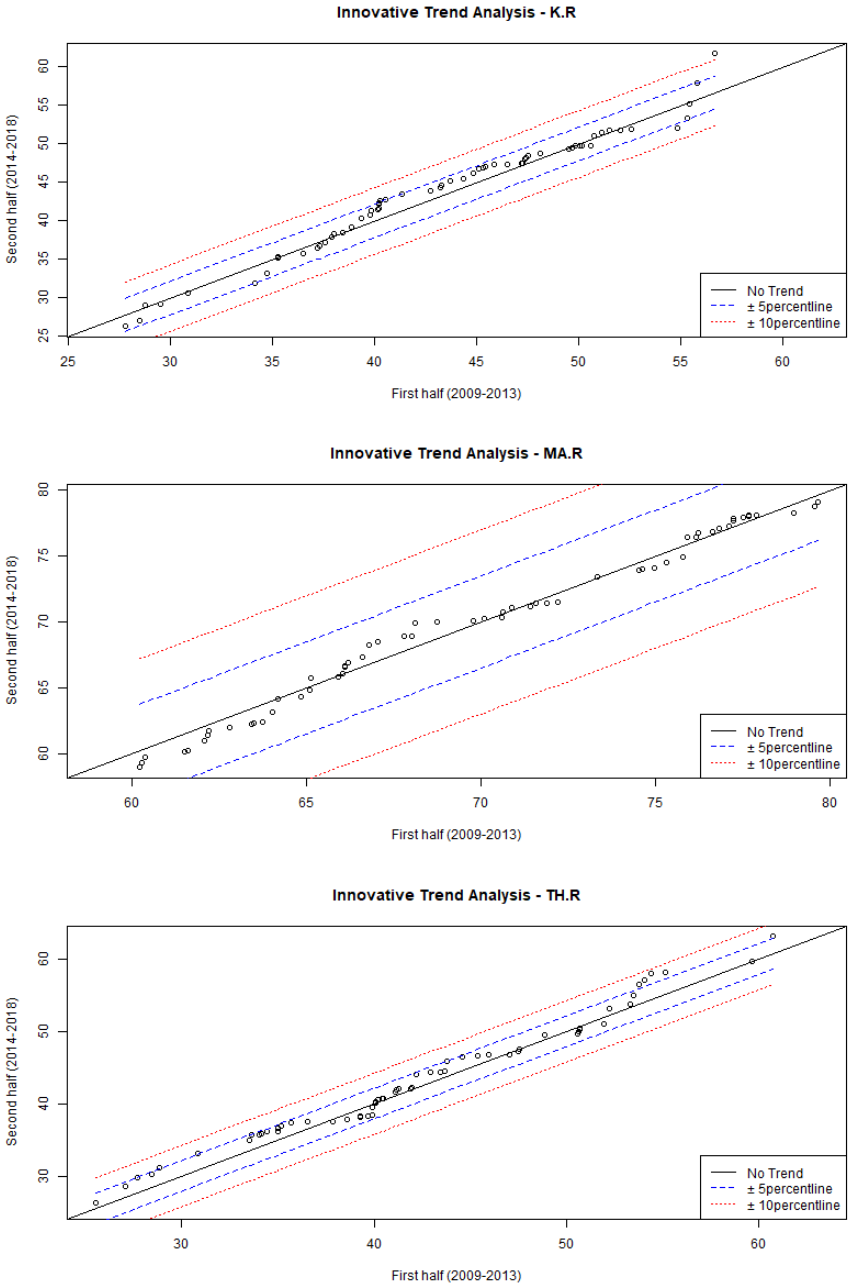

Figure 4.7: Results of the ITM test for annual mean humidity at the stations K, MA and TH from 2009 to 2018.

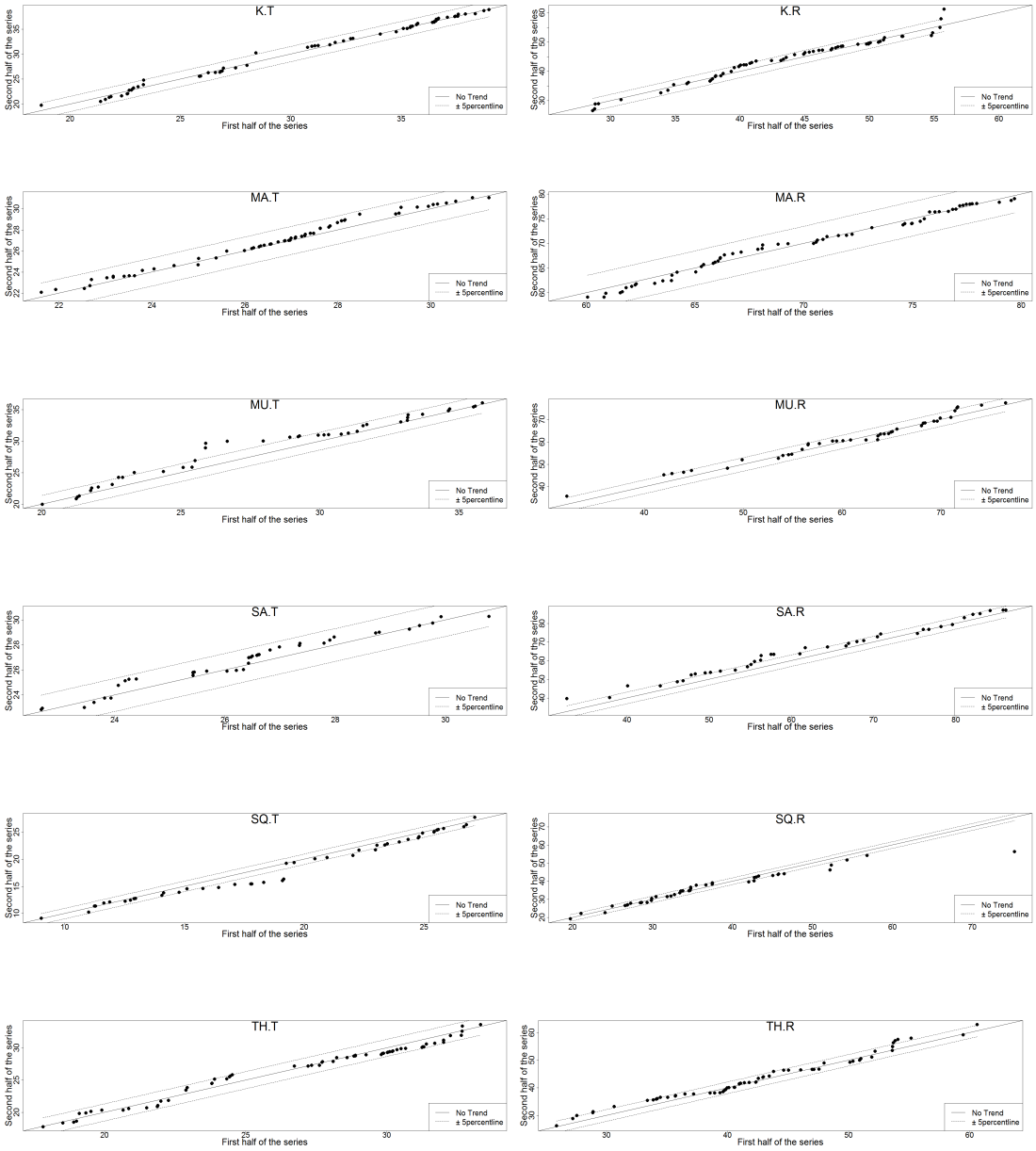

Figure 4.8: The ITM test diagnostic of monthly time series of temperature (left) and humidity (right) in 6 locations.

4.3.3 Spearman's rho test

The SR test is one of the rank-based non parametric statistical tests that can be used to detect a monotonic trend in time series [147]. It is also a simple method that has uniform power for both linear and non-linear trends [117, 130].

In the SR test, the null hypothesis H_0 is that all the data in the time series are independent and that there is no monotonic trend and the alternative hypothesis H_1 is that increasing or decreasing trends exist. The Spearman's rank correlation coefficient statistic S_{SR} and the standardized test statistic Z_{SR} are expressed in [117, 130].

The SR test is based on the Pearson correlation coefficient between ranks and can be computed by

$$S_{SR} = 1 - \frac{6}{N(N^2-1)} \sum_{i=1}^{N} (R_i - i)^2. \tag{4.7}$$

The standardized statistics Z_{SR} are defined as

$$Z_{SR} = S_{SR} \sqrt{\frac{N-2}{1-(S_{SR})^2}}, \tag{4.8}$$

where R_i is the rank of ith element of a time series and N is the length of a time series $x_1, \ldots, x_N$. Positive values of Z_{SR} indicate upward trends, while negative values of Z_{SR} indicate downward trends in the time series. When $|Z_{SR}| > t_{N-1,1-\frac{\alpha}{2}}$, the null hypothesis is rejected and a significant trend exists in the time series. $t_{N-1,1-\frac{\alpha}{2}}$ is the critical value at 5% significance level.

We have applied the SR test for the temperature and humidity data series for six stations using time-series data with a length $L = 120$ months for 2009 to 2018. The trend tests revealed no statistically significant trends at stations K, MU, and

TH with respective p-values of $0.508, 0.776$ and 0.612 for the temperature series and $0.4540, 0.501$ and 0.492 for the humidity series. Stations MA, SQ, and SA have missing values for more than one year, which created a short time series. These stations had a small positive trend at 5% significance level.

4.3.4 Comparison of the trend tests

Trend tests have been used to detect monotonic trends in time series. Using these three methods, we can conclude that there were no monotonic changes in temperature and humidity in Oman over the period 2009–2018. The MK test is not suitable for data with periodicities such as seasonal effects and is not preferred for short time series. In contrast, the SR test is suitable for work with short time series. The ITM test does not depend on any particular assumption about distribution, serial correlation, or seasonal cycles and is a simple test to understand and calculate. In addition, it is a method that can provide critical information and predictions to decision makers and can be used in different climate change scenarios [2, 6, 21].

Note that we performed multiple tests, and therefore, some p-values can be below than 0.05 by chance. We found that the empirical distribution of 8760 p-values for each test is rather uniform for all three tests. Since the presence of a significant p-value for one particular hour is not important, we study the allocation of significant p-value along days in a year. Specifically, in Figures 4.9 and 4.10, we show the cumulative number of significant p-value for six stations. Since there are many missing values in Nov-Dec for locations MU and SA, p-values for this period were not computed. These figures show several jumps, meaning that there are few days in year with many significant p-values.

Since the height of these jumps is not sufficiently large, we concluded that there are no monotonic trend in the annual oscillations of temperature and humidity over the period 2009–2018 in Oman.

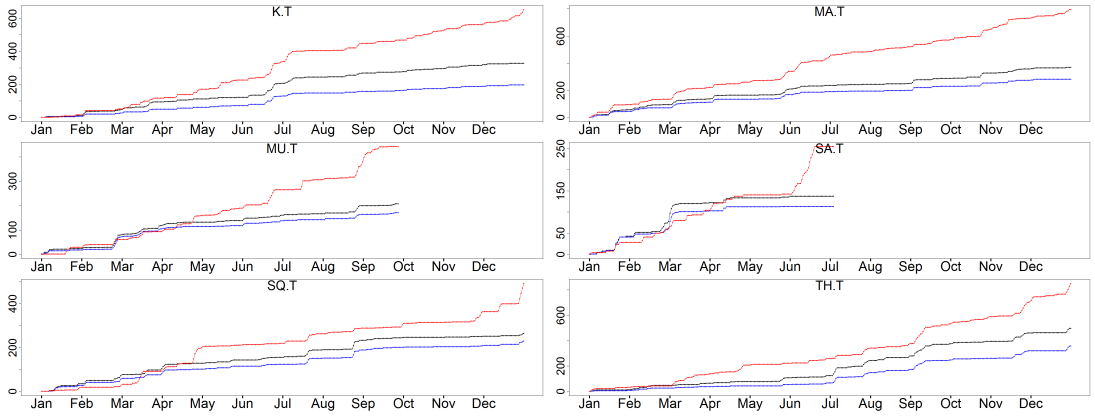

Figure 4.9: The cumulative numbers of significant p-values of the MK test (black), the SR test (blue), the ITM test (red) for the annual oscillation of temperature in 6 locations.

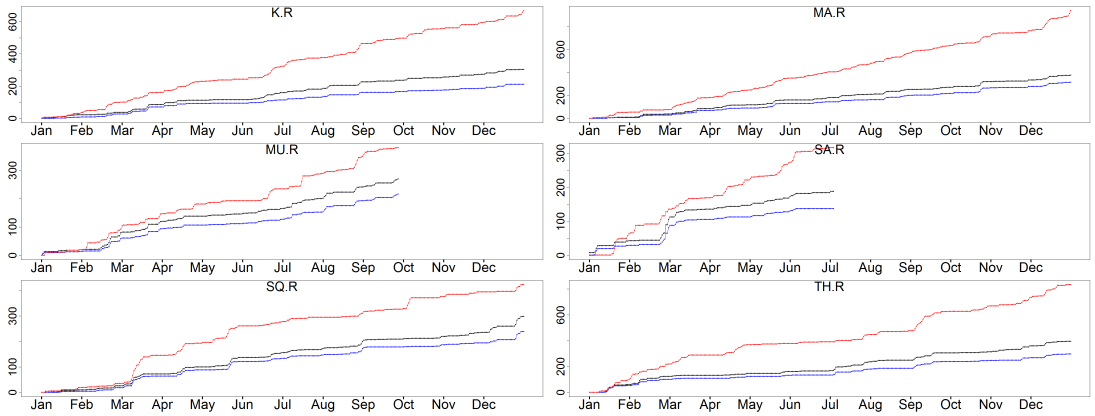

Figure 4.10: The cumulative numbers of significant p-values of the MK test (black), the SR test (blue), the ITM test (red) for the annual oscillation of humidity in 6 locations.

4.4 Daily periodicities

Detection of periodicities components of natural processes is critical because it can help in understanding the different processes [139, 146]. Furthermore, the presence of periodicity complicates the efficient modeling of time series using stochastic modeling techniques [151]. Sometimes, the presence of noise in the data makes it extremely difficult to detect periodicity components in time series [27]. SSA is an efficient way for detecting periodicities in the time series compared to other time series techniques [144].

If we want to extract a periodic component with known period, then, the window lengths, which are divisible by the period, provide better separability. If we choose a few leading eigentriples, then SSA with small L performs smoothing of the series as a filter of order $2L - 1$, the choice of the window length is important. Therefore, the result is usually stable with respect to small changes in the values of L. If the time series has a complex structure, then the so-called Sequential SSA is recommended. Sequential SSA consists of two stages; at the first stage, trend is extracted with a small window length and then, periodic components are detected and extracted from the residual with $L = N/2$ [48]. Additionally, the periodic components are chosen by the choice of eigenvectors with desired frequencies.

For extracting the daily oscillation, we apply SSA with $L = 24$. Specifically, we obtain the daily oscillation by substraction of the leading component of the SSA decomposition from the analysed time series. We show the daily oscillation of temperature for several locations in July in 4.11. We can see that the the daily periodicities is very similar from year to other year and from day to day and depends on month except for the station MU. However, the shape of periodicities depends on the stations. Also, we applied the MK test, the SR test and the ITM

test to values of the daily periodicities with one year increment and can conclude that there is no monotonic trend in the daily periodicities of temperature.

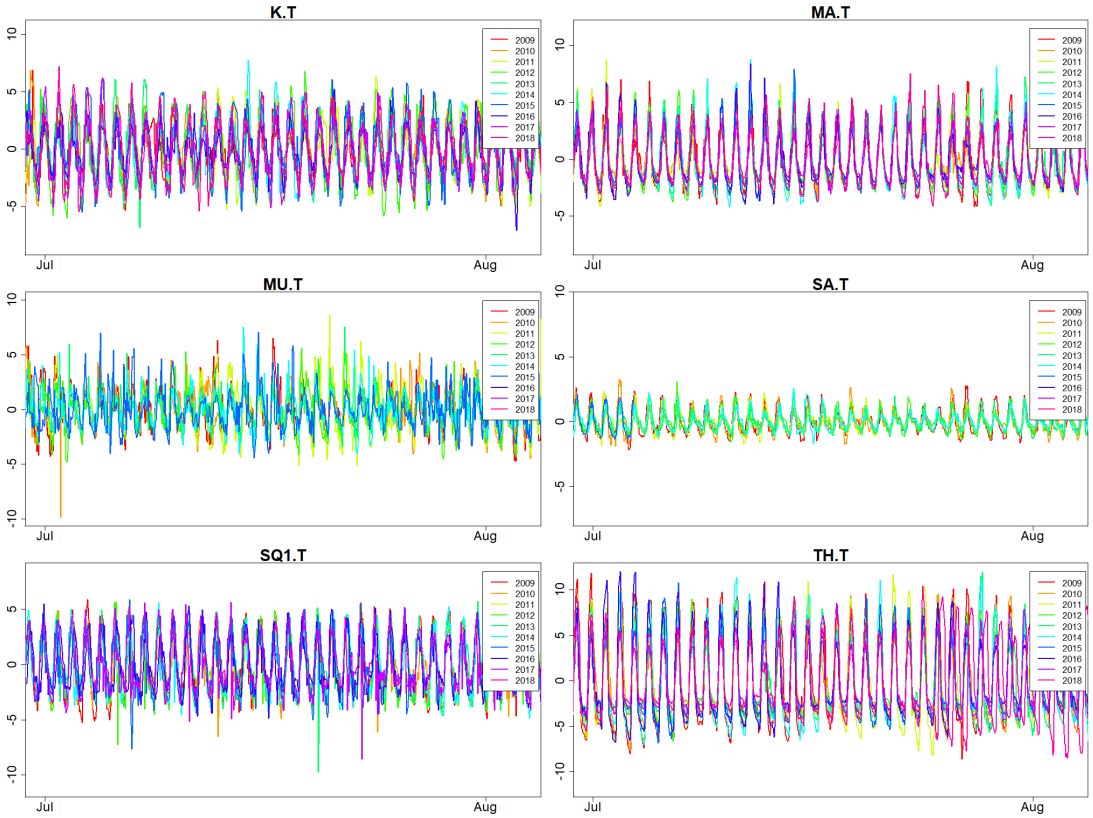

Figure 4.11: The daily periodicities of temperature in July at six stations.

In Figure 4.12, we depict the daily periodicities of humidity in July for six stations. We can see that the daily periodicities of humidity has a very small amplitude at the station SA, sharply replicates from year to year and from day to day within July at the stations MA and TH and very volatile at stations K, MU and SQ.

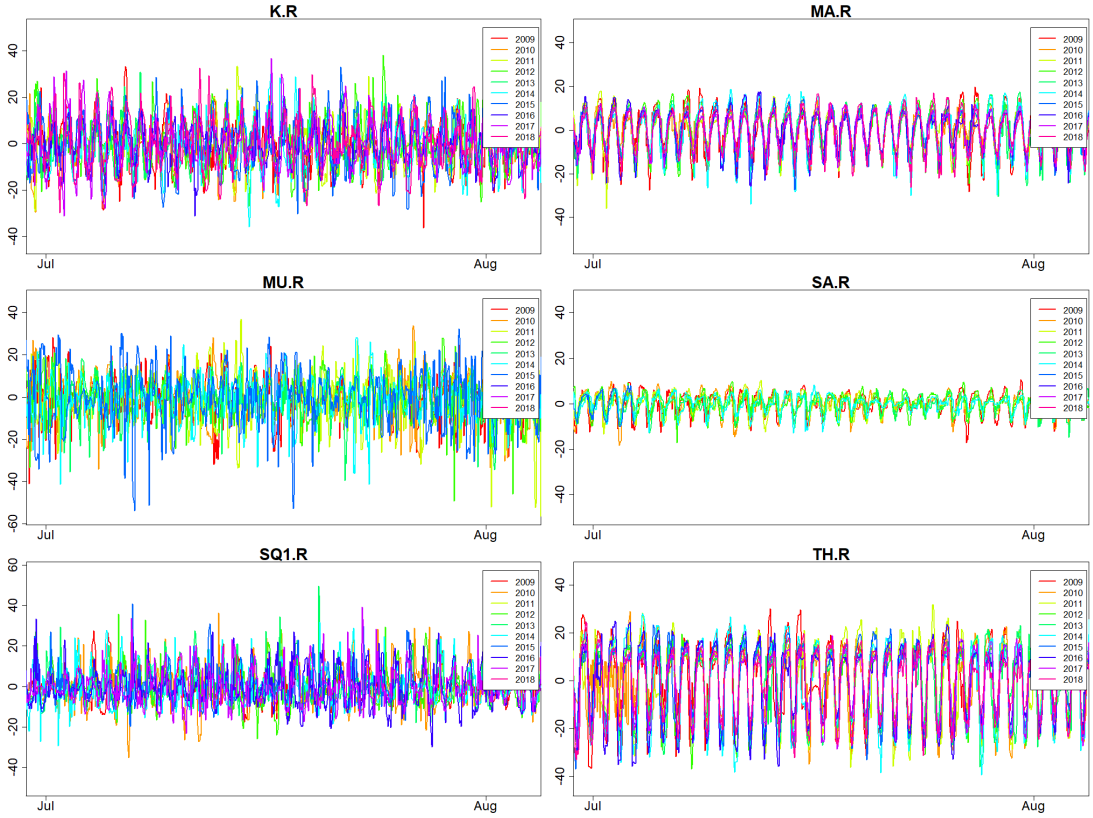

Figure 4.12: The daily periodic of humidity in July at six stations.

For studying variability of the daily periodicities along year, we consider the standard deviation of hourly time series of each month, specifically, we compute

$$s_m = \sqrt{\sum_{t=h_1(m)}^{h_2(m)} x_t^2},$$

where $h_1(m)$ is the first hour of the m-th month, $h_2(m)$ is the last hour of the m-th month.

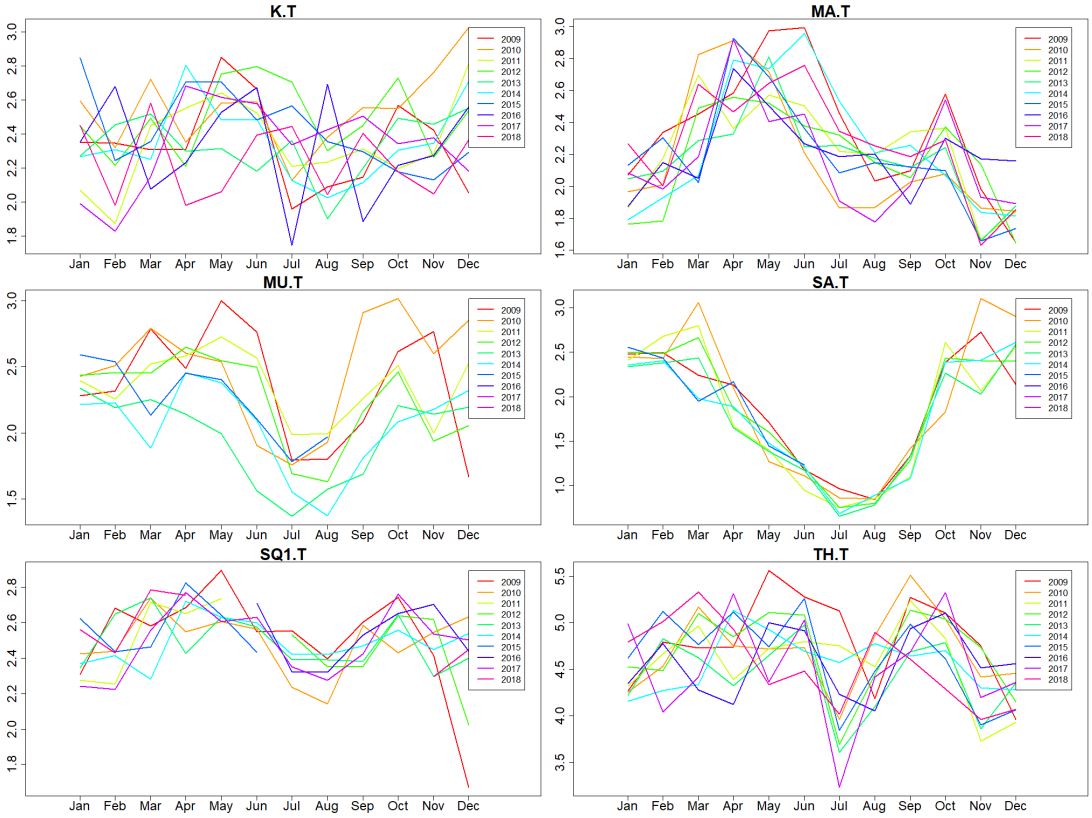

Figure 4.13: The monthly standard deviation of the daily periodicities of temperature.

In Figure 4.13, we can see that variability of the daily periodicities of temperature does not depend on the month at stations K, SQ and TH but it strongly depends on month at stations MA, MU and SA. The MK test does not detect monotonous trends in monthly variability of the daily periodicities of temperature from year to year.

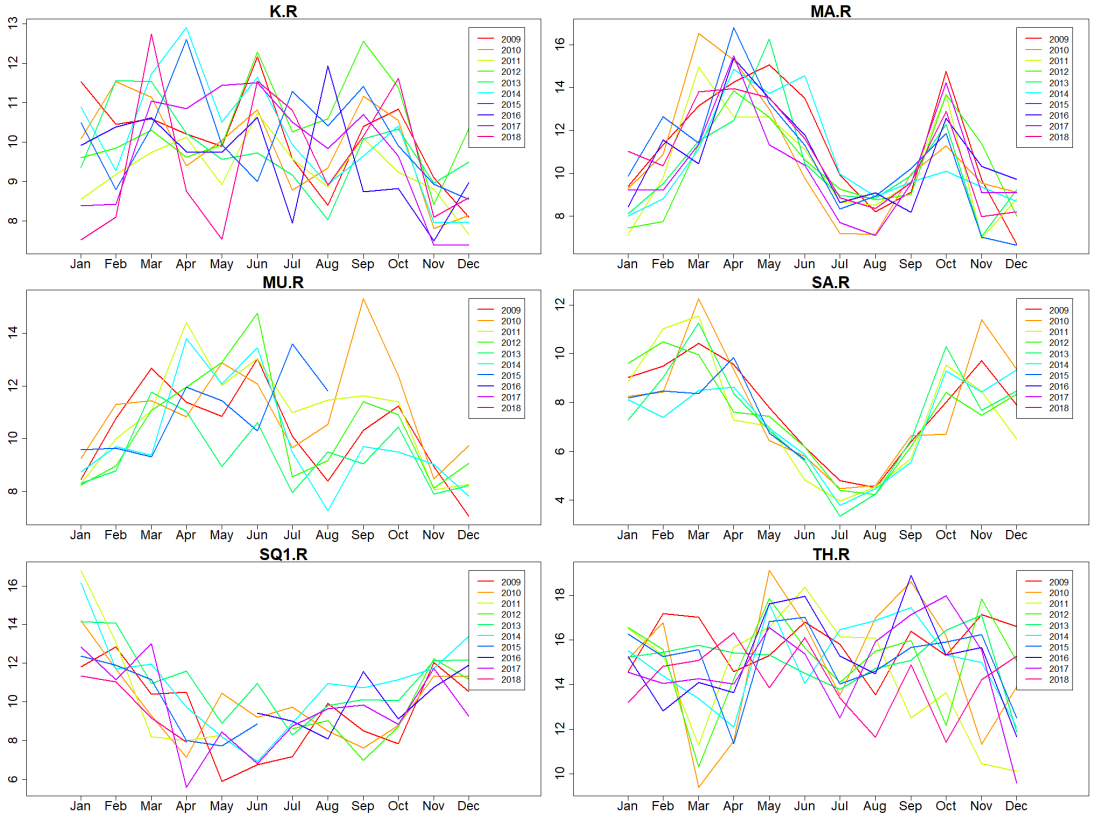

Figure 4.14: The monthly standard deviation of the daily periodicities of humidity.

In Figure 4.14, we can see that variability of the daily periodicities of the humidity is almost independent of the month for stations K and TH but it strongly depends on the month at stations MA and SA. Trends in monthly variability of the daily periodicities of humidity from year to year were not found by the MK test, the SR test and the ITM test.

4.5 Chapter summary

We applied SSA to hourly time series for extracting the annual oscillations and the daily periodicities. SSA was able to extract these components very efficiently. Moreover, we may use SSA for obtaining more refined decompositions with a larger number of components and also for forecasting.

We applied three commonly used tests for detecting trends in time series: the MK test, the SR test and the ITM test. We found that there are no monotonic trends in the annual oscillations and the daily periodicities over ten years. Also, we did not find trends in the monthly variability of daily periodicities.

The developments of this Chapter have been started to respond to an inquiry from different scientific institutes in Oman and can contribute to the field of meteorological research.

Chapter 5

Forecasting monthly temperature and humidity

5.1 Introduction

Forecasting the time series of temperature and humidity are critical elements of climate analysis and they can have significant economic and climatic impact. Chapter 5 demonstrates the accuracy of forecasting time series of temperature and humidity by using SSA forecasting algorithms. SSA provides accurate results compared to other methods in many practical problems [53]. Additionally, SSA forecasting algorithms have been successfully applied in fields like meteorology and economics among others [74, 152]. SSA is a very useful tool for extracting various signals from noisy observations [52, 63]. A modification of SSA was used in [45] to find structures in short time series by extracting seasonality and simultaneously extracting cycles of small and long periods.

Many time series forecasting methods are based on the analysis of historical data [96]. They assume that past patterns in the data can be used to forecast future events [96]. Chapter 5 considers the monthly time series of humidity and

temperature from 2009 to 2018. A full description of the practical aspect of SSA and forecasting algorithms along with some criteria for selecting SSA parameters are described in Chapter 2. We used the stationary seasonal ARIMA model, ETS and RNN for fitting and forecasting our monthly time series. The SSA algorithm has two parameters: the window length L and the number of singular values r. In Chapter 2, we provide a description of choosing SSA parameters. For Basic SSA, see Chapter 1 and [43, Sec 2.1], the general guideline for selecting r and L is to take sufficiently large, say $L \approx \frac{N}{2}$ see [42]. There are two main SSA forecasting algorithms: SSA-R forecasting and SSA-V forecasting, which both depend on two parameters L and r as shown in [42, Sec 3.2.1.2] and [42, Sec 3.2.1.3]. These parameters can be chosen by an expert looking at the signal structure or using the automatic choice based on the RMSE of retrospective forecasts [42]. For assessing the quality of the automatic choice, we firstly analyze the sensitivity of the RMSE on parameters and then, perform a study of reliability of the automatic choice for forecasting monthly temperature and humidity recorded at three meteorological stations in Oman.

Chapter 5 is structured as follows.

- In Section 5.2, we describe the features of the time series of temperature and humidity at three meteorological stations in Oman and some application of Basic SSA.
- In Section 5.3, we report the dependence of the RMSE on parameters.
- In Section 5.4, we study the automatic choice of parameters.
- In Section 5.5, we discuss the choice of parameters for SSA-V and SSA-R forecasting algorithms.
- In Section 5.6, we offer concluding remarks.

5.2 Monthly time series

In Section 5.2, we demonstrate the accuracy of forecasting algorithms for monthly time series of temperature (measured in Centigrade) and humidity (measured in %), which were provided by the Directorate General of Meteorology of Oman. The data was collected from Jan 2009 to Dec 2018 at three meteorological stations in the Sultanate of Oman: the Khasab Airport (K), the Masirah (MA) and the Thumrait (TH) stations.

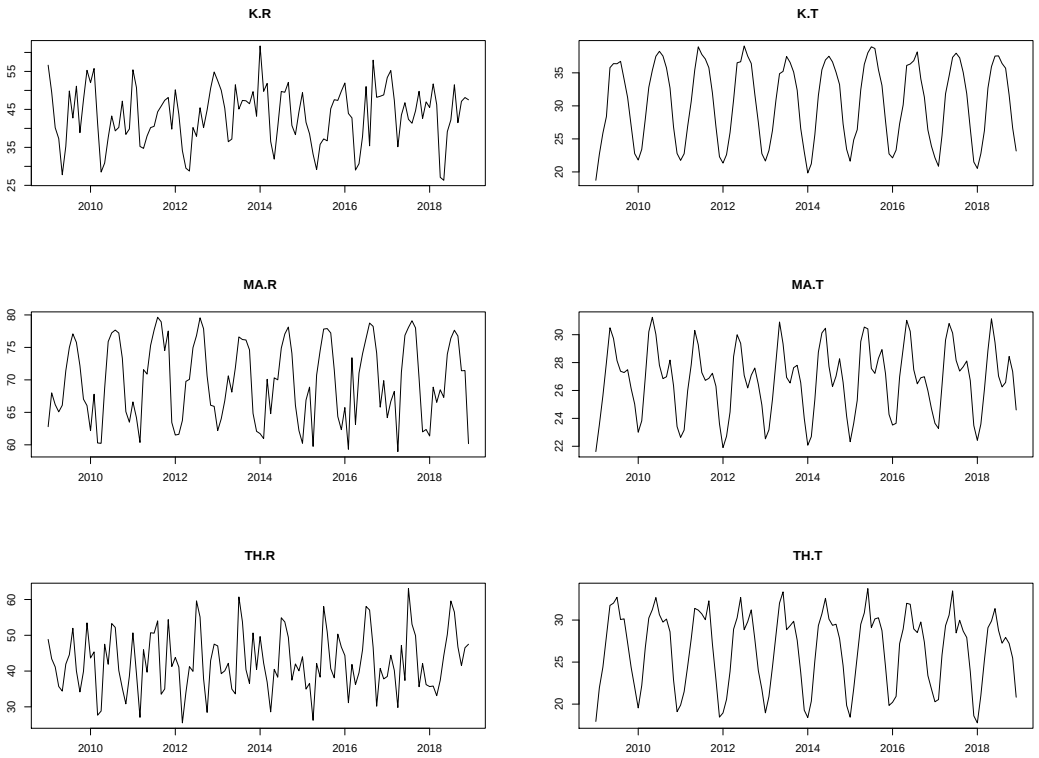

Figure 5.1: Monthly humidity (left) and temperature (right) at the stations K, MA and TH from 2009 to 2018.

Table 5.1: Descriptive statistics for the monthly time series of humidity and temperature from 2009-2018.

Series	Station	Mean	Median
Humidity	K	43.56	44.36
	MA	69.87	70.09
	TH	42.64	41.40
Temperature	K	30.54	31.81
	MA	26.77	26.98
	TH	26.59	27.93

Figure 5.1 demonstrate the monthly time series of humidity and temperature at the stations K, MA and TH from 2009 to 2018. In Figure 5.1, we depict all time series which do not exhibit trends as shown in Chapter 3. We can observe that the annual pattern of temperature is rather stable from year to year. However, the temperature at the station K has a simple sinusoidal shape but the annual pattern of the temperature at the stations MA and TH is more complicated. We can also see that humidity is very volatile and the annual pattern is more visible only at the station MA. Therefore, the humidity at the station MA is much larger than humidity at the stations K and TH.

Table 5.1 presents the brief descriptive statistics of the time series of humidity and temperature for three stations. It provides values of mean and median of time series. The range of average values for the time series of humidity is between 40 and 70 and the average for the time series of temperature is between 25 and 31.

5.2.1 Temperature time series at the station TH

Section 5.2.1 demonstrates the use of Basic SSA when applied to the time series of temperature at the station TH over the period 2009 to 2018 as shown in Figure 5.2. This illustrates the capability of SSA at removing various components from a time series such as trend, oscillation, noise and forecasting.

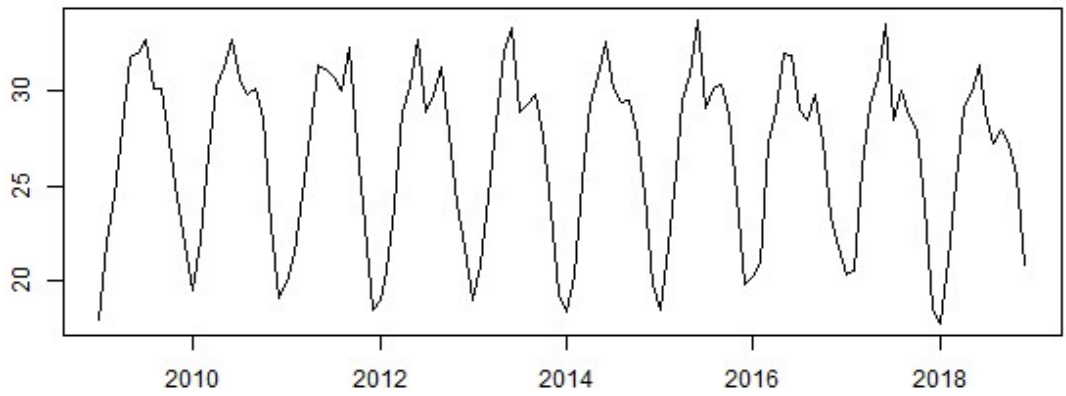

Figure 5.2: Monthly the time series of the temperature at the station TH from 2009 to 2018.

In the decomposition stage, the window length L is the only parameter that requires specification. The window length L should be large enough but not greater than $\frac{N}{2}$ [43, 53]. If the time series (of length $N = 120$ months in this case) has a periodic or seasonal component, then it is beneficial in terms of SSA separability to take a window length proportional to that period. For this example, the window length has been set to $L = 60$. Figure 5.3 contains the result of the decomposition which is used for the extraction of the trend and seasonality.

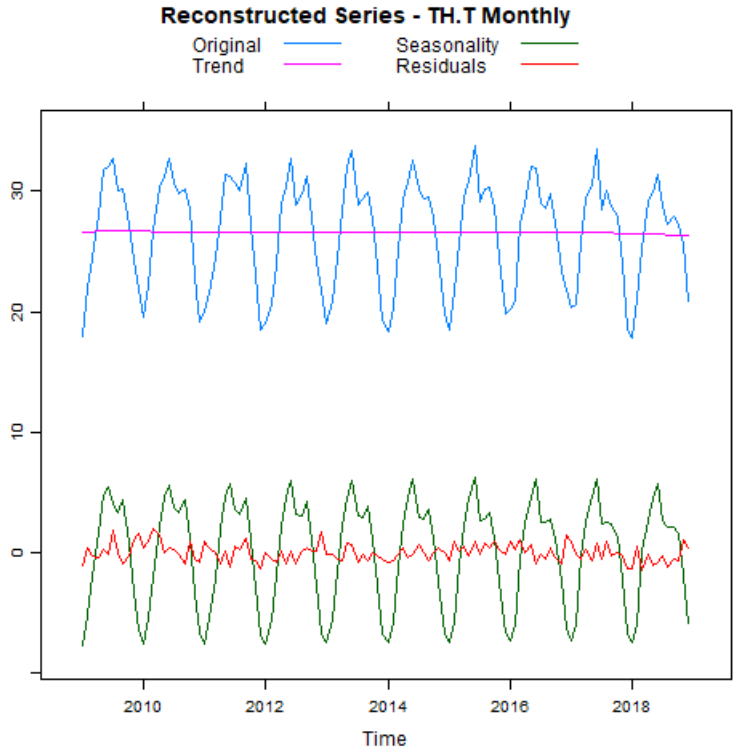

Figure 5.3: Decomposition for the time series of the temperature at the station TH.

Figure 5.4 displays the plots the reconstructed components. It describe eight most significant intial reconstructed components of the orginal time series. After taking a quick look, the first reconstructed component is relate to slow motion component (the trend behaviour) while the remainder of reconstructed component are connected to fluctuating components. Figure 5.4 shows that the first eigenvector is slowly varying and on the basis of recommendations of [42], we include ET1 into the trend group. The components of eigenvectors represent the structure of a sub-series of the original series.

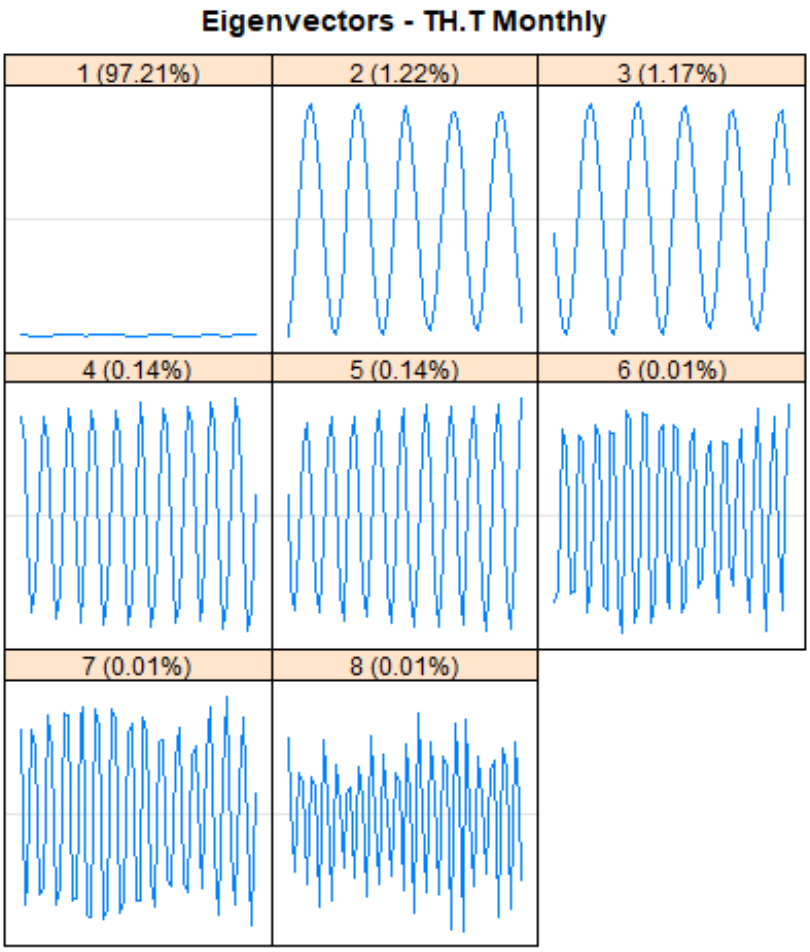

Figure 5.4: 1D graphs of eigenvectors for the time series of the temperature at the station TH.

Figure 5.5 shows 2D-scatterplots of eigenvectors for the time series of the temperature at the station TH and the eigenvector pairs $2-3$, $4-5$, $5-6$, $6-7$, $8-9$ are produced by modulated sine waves since the corresponding 2D-scatterplots of eigenvectors resemble regular polygons. We make this observation based on the following properties: a sine wave has rank 2 and produces two eigentriples, which are sine waves with the same frequency and have a phase shift exactly or

approximately equal to $\frac{\pi}{2}$, due to the orthogonality of eigenvectors. By counting the numbers of polygon vertices in Figure 5.5, the periods of the sine-waves can be determined as $12, 4, 6, 2.4, 3$. Figure 5.5 indicates the number of vertices for the five pairs listed. There is no significant trends and high noise in the monthly time series of the temperature at the station TH.

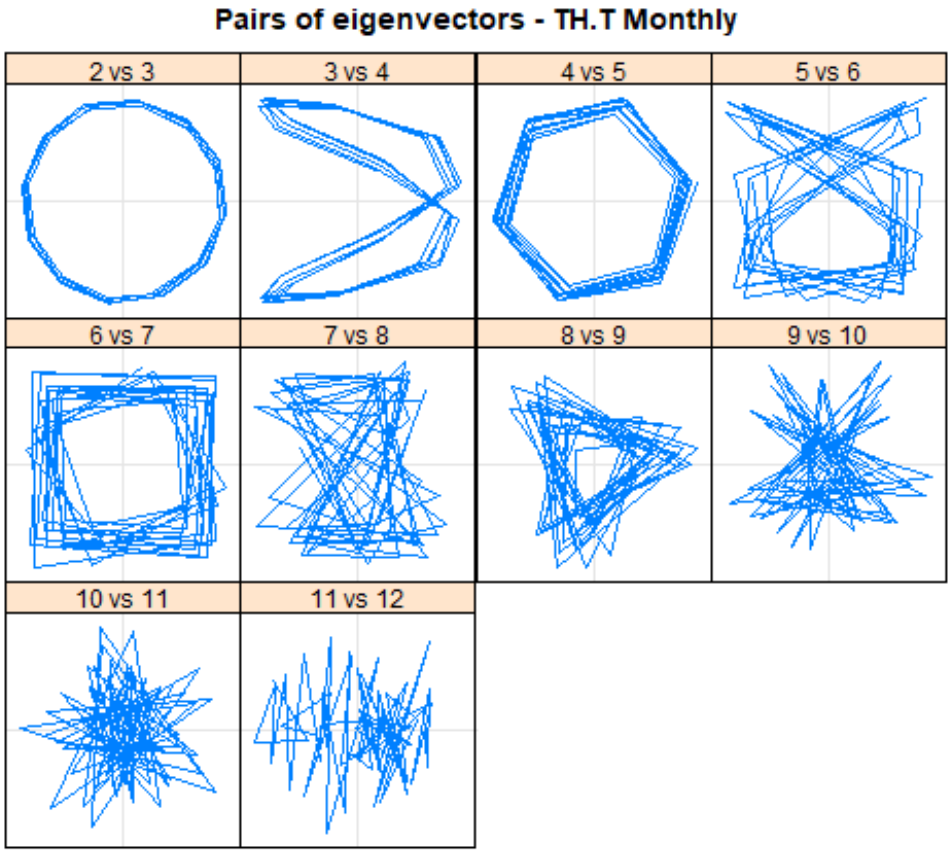

Figure 5.5: 2D scatterplots of eigenvectors for the time series of the temperature at the station TH.

The matrix of absolute values of w-correlations is depicted in grayscale in Figure 5.6

(white color corresponds to zero and the black color corresponds to the absolute values equal to 1). Large w-correlations between reconstructed components suggested that the components should be grouped and lead to the same component in the SSA decomposition. It confirms that the indicated pairs are separated between themselves and from the trend component since the w-correlations between the pairs are small, while w-correlations between the components from the same pair are very large. From Figure 5.6, we reconstruct the series using eigenvalues $1-10$ and classify that the remaining eigenvalues correspond to noise components.

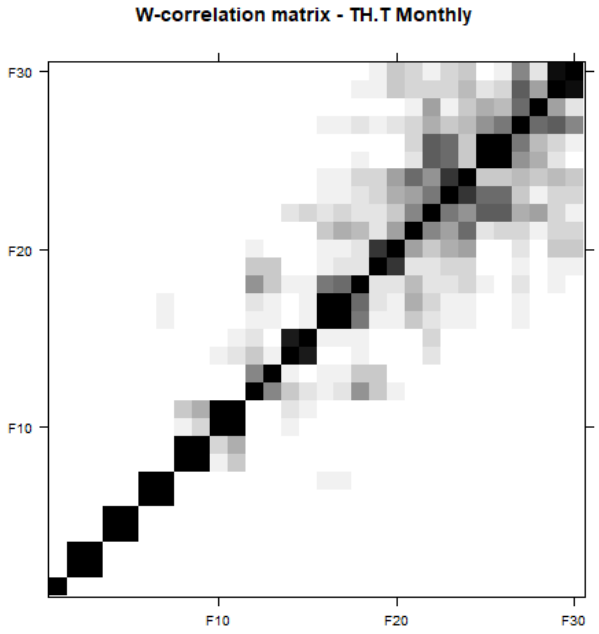

Figure 5.6: Weighted correlations for the time series of the temperature at the station TH.

Figure 5.6 provides helpful information for detection of separability and identification of groups. It indicates that well-separated components have weak correlation while poorly separated components have high correlation. If the correlations are high, these components are well separated from a block of the remaining compo-

nents; otherwise, if the correlations are messy, these reconstructed components are possibly considered noise components.

Reconstruction is the second stage of SSA which includes two separate steps: grouping and diagonal averaging. The grouping step is for identifying signal component and noise and the diagonal averaging step is for using grouped eigentriples to reconstruct the new series without noise. Figure 5.7 displays four reconstructed modulated sine waves and shows that several sine waves have harmonic amplitudes.

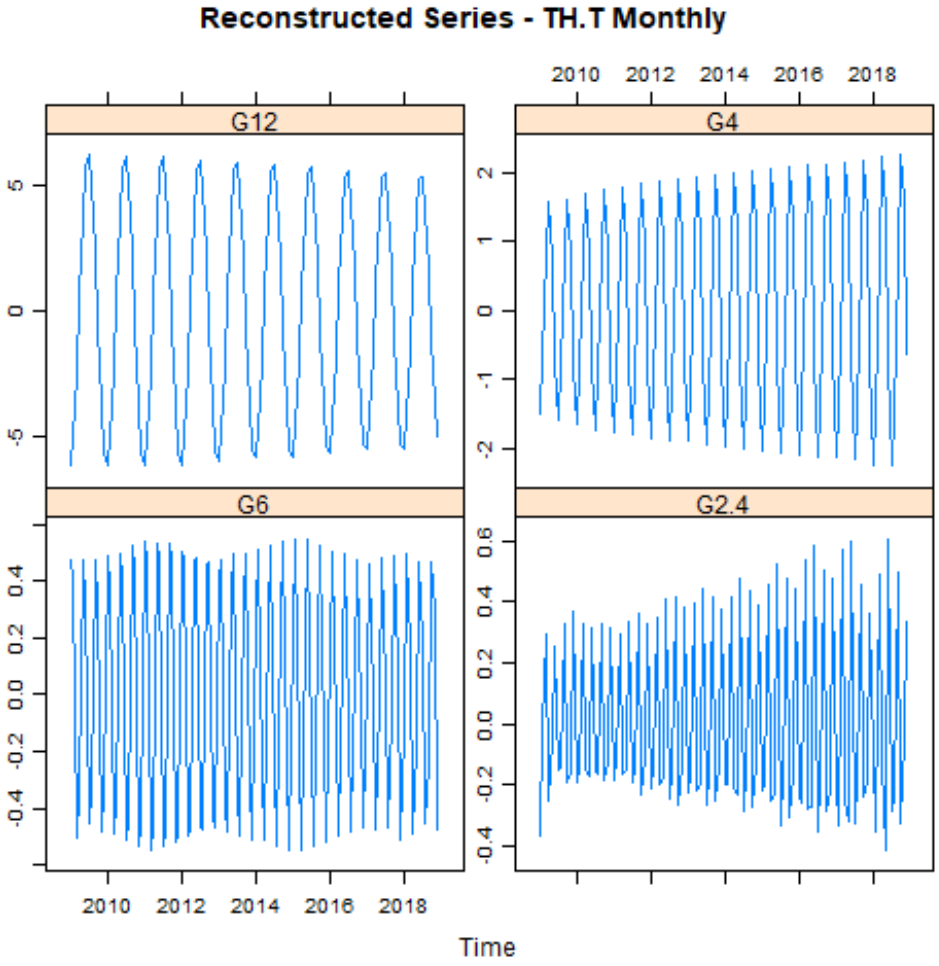

Figure 5.7: Reconstructed sine waves for the time series of the temperature at the station TH.

In the grouping step, trend is a component that does not contain any oscillatory components. Practically, in this case it shows that the annual oscillations for hourly time series can be viewed as a trend (a slowly changing component) and hourly time series of the annual oscillations of temperature, as mentioned in previous Chapters. A small variation of temperature from year to year and random fluctuations from week to week can be seen and the annual oscillation has the two-saw shape in the station TH.

The second component is the harmonic component and is difficult to identify and separate from other oscillatory components of the time seires. The last one is noise and grouping of the eigentriples do not appear to contain elements of trend and oscillations is a natural way of extracting noise. Diagonal averaging is the last step of SSA technique. Three components as groups are considered: the trend, harmonic component and noise.

5.3 Dependence of the RMSE on parameters

Let us study how the RMSE of 1,2,...,12-step ahead forecasts across several truncation points depends on the parameters L and r. We take Jan 2017 as the first truncation point and Dec 2017 as the last truncation point.

The functions *SSA-V.Forecasting* and *SSA-R.Forecasting* are provided for implementing vector forecasting and recurrent forecasting, respectively, in R [42, Section 3.5.7].

Retrospective forecasts, as shown in [102], are performed as follows. Retrospective forecasting is accomplished by truncating the series and forecasting values at the points that have been temporarily removed. These forecasts can be used to assess the forecasts' quality by comparing them to the observed values of the time series.

Choice of L. In Section 5.2, we study the characteristics of the time series of temperature and humidity. If the structure of the sequence is consistent, large values of L can be used, of the order $L \cong 60$, small values should be avoided, of the order $L \cong 12$. However, we do not make any assumption that the structure of the series is stable or not. In this case, selecting large values of L would make SSA lack flexibility and on the other hand, using very small of L leads to noise and high sensitivity. The value of L should be somewhere in between $16 \leq L \leq 48$.

Choice of r. we have to decide what the proper grouping is and how to find the proper groups for For low frequency sinusoid and a high frequency sinusoid in SSA parameters of the eigentriples. In other words, we need to identify an eigentriple corresponding to the related time series component. Since each eigentriple consists of an eigenvector (left singular vector), a factor vector (right singular vector) and

a singular value, this is to be achieved using only the information contained in these vectors (considered as time series) and the singular values [43].

The value of r should be determined by the type of forecast we want to make [103]. Our choice is $4 \leq r \leq 15$ and whatever the rule for selecting r, we must be cautious that some r values are too small, implying that some of the signal has been lost, whereas other r values are too high, implying noise [103, 102].

To summarise, we select parameters $16 \leq L \leq 48$ and $4 \leq r \leq 15$ for all series. We made a decision regardless of the series' structure as described in Section 5.2.

The parameter r should correspond to the rank of the signal. As noted in [39, 59] the authors argue that the value of r needs to be greater than what it should be since parts of the signal can be missed but the increase of noise in the reconstructed series can be small.

The optimal L and r for obtaining forecasts for our time series are the parameters L and r that correspond to the lowest value of RMSE. As a result, we search for the best combination of L and r, which represents the best decomposition and reconstruction options for model [8].

In Tables 5.2–5.7, we show the RMSE for six time series using two SSA forecasting algorithms with the wide range of parameters L and r. The lowest values for the RMSE for each forecasting algorithm are highlighted. We can see that the RMSE does not monotonically depends on parameters but tendencies are rather clear. Specifically, the RMSE is larger for very small r because such small values of r are smaller than the signal rank. The RMSE also become larger for large r especially for noisy time series because large r is greater than the signal rank and forecasting becomes unstable. We can observe that the RMSE for small $L = 16$ is larger because the structure of time series is not captured well. For fixed L

and r, the RMSE of two forecasting algorithms are close each other with a slight dominance of SSA-V forecasting.

In Table 5.2, we show the RMSE for humidity at the station K. We see that the lowest RMSE is 5.343 for SSA-R forecasting and 5.365 for SSA-V forecasting and it is attained at $L = 36$ and $r = 5$ for both algorithms. However, for other values of L and r SSA-V forecasting is usually slightly better. Overall, the RMSE is weakly depending on L and r when $L \geq 24$.

Table 5.2: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for humidity at the station K.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	6.382	7.013	6.808	6.780	6.132	5.950	6.354	6.177
5	6.395	7.071	6.210	6.284	**5.343**	**5.365**	5.590	5.386
6	9.346	9.819	7.800	8.003	6.859	6.415	5.861	5.898
7	10.057	7.987	6.026	5.831	6.642	6.652	6.651	6.334
8	11.924	9.461	5.890	5.934	5.629	6.032	6.986	6.460
9	12.086	10.538	6.145	6.396	6.683	6.142	6.687	6.276
10	11.874	10.375	7.532	6.313	6.996	6.126	6.662	6.098
11	15.015	11.339	6.503	5.980	7.115	6.153	6.622	6.092
12	14.935	10.494	6.888	6.085	6.987	6.114	6.721	6.019
13	12.106	10.279	7.293	5.920	6.641	6.142	6.741	5.983
14	37.023	20.195	7.753	6.390	6.762	6.822	6.772	6.631
15			8.228	6.324	6.671	6.722	6.669	6.549

In Table 5.3, we display the RMSE for temperature at the station K and we see that the lowest RMSE is 0.853 attained at $L = 48$ and $r = 5$ for both SSA-R and SSA-V forecasting algorithms. We observe that the RMSE is quite small for $r = 5$ and any L because temperature at the station K has a simple sinusoidal shape. The RMSE for $r = 4$ is larger than the RMSE for $r = 6$ indicating the acurracy of forecasting is dropping faster when taking r to be smaller than the signal rank.

Table 5.3: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for temperature at the station K.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	1.026	1.240	1.250	1.273	1.206	1.297	1.170	1.243
5	0.882	0.972	0.948	0.991	0.867	0.963	**0.853**	**0.853**
6	1.064	1.082	0.931	1.016	0.875	0.939	0.864	0.950
7	1.198	1.131	1.263	1.079	1.145	1.126	1.069	1.076
8	1.378	1.154	1.373	1.066	1.179	1.142	1.103	1.069
9	2.095	1.206	1.337	1.136	1.240	1.154	1.028	1.065
10	2.869	1.161	1.399	1.132	1.315	1.159	1.037	1.066
11	3.225	1.449	1.482	1.214	1.359	1.151	1.024	1.000
12	4.106	1.564	1.500	1.225	1.416	1.154	1.024	1.003
13	4.916	1.998	1.544	1.221	1.434	1.150	1.077	1.042
14	7.291	10.167	1.722	1.144	1.477	1.175	1.104	1.029
15			1.804	1.218	1.557	1.190	1.133	1.009

Table 5.4 contains the RMSE for humidity at the station MA, the smallest RMSE is 2.917 attained at $L = 36$ and $r = 4$ for SSA-R forecasting and 3.145 attained at $L = 48$ and $r = 6$ for SSA-V forecasting. Overall, the RMSE is quite small for $r = 4$ or $L = 48$ indicating that the signal rank is 4.

Table 5.4: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for humidity at the station MA.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	3.121	3.828	3.398	3.417	**2.917**	3.351	2.962	3.337
5	3.636	3.837	3.213	3.332	5.343	5.365	3.171	3.205
6	3.895	3.872	3.548	3.234	6.859	6.415	3.258	**3.145**
7	4.648	4.464	4.110	4.016	6.642	6.652	3.620	3.286
8	4.687	4.343	4.089	3.849	5.629	6.032	3.560	3.192
9	5.105	4.840	4.195	3.907	6.683	6.142	3.564	3.265
10	5.200	4.774	4.130	3.908	6.996	6.126	3.654	3.269
11	5.611	4.326	4.541	3.748	7.115	6.153	3.809	3.217
12	8.314	5.810	5.636	4.237	6.987	6.114	4.078	3.201
13	9.686	5.576	7.154	4.300	6.641	6.142	4.256	3.189
14	11.565	6.998	7.725	4.355	6.762	6.822	4.415	3.237
15			8.331	4.452	6.671	6.722	4.678	3.372

In Table 5.5, we present the RMSE for the temperature at the station MA and we see that the lowest RMSE is 0.561 attained at $L = 24$ and $r = 5$ for SSA-R forecasting and 0.564 attained at $L = 36$ and $r = 9$ for SSA-V forecasting. In general, the RMSE is quite small for $r = 5$ and any L showing the signal rank is likely to be 5.

Table 5.5: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for temperature at the station MA.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	1.039	1.953	1.512	1.599	1.412	1.608	1.402	1.643
5	0.632	0.732	**0.561**	0.614	0.564	0.595	0.592	0.572
6	0.811	0.886	0.567	0.644	0.609	0.596	0.731	0.731
7	0.821	0.828	0.957	0.992	0.578	0.584	0.841	0.741
8	0.827	0.925	0.915	0.910	0.563	0.584	0.814	0.722
9	1.037	0.961	0.888	1.007	0.570	**0.564**	0.821	0.713
10	0.980	0.921	0.770	0.788	0.643	0.763	0.853	0.747
11	0.905	0.874	0.796	0.751	0.885	0.725	0.882	0.759
12	0.935	0.859	0.864	0.850	1.122	0.724	0.913	0.797
13	1.013	0.893	0.892	0.851	1.129	0.712	1.013	0.839
14	1.055	0.910	0.922	0.916	1.108	0.672	1.042	0.866
15	1.517	1.327	1.175	0.862	1.123	0.665	1.062	0.886

In Table 5.6, we show the RMSE for humidity at the station TH and we observe that the lowest RMSE is 4.441 attained at $L = 36$ and $r = 15$ for SSA-R forecasting and 4.445 attained at $L = 36$ and $r = 11$ for SSA-V forecasting. We see that the RMSE is quite small for any $r \geq 5$ and $L \in \{24, 36\}$. Note that humidity at the station TH is the most volatile in our study.

Table 5.6: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for humidity at the station TH.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	6.979	9.040	6.383	8.319	6.744	8.214	7.077	8.217
5	6.099	6.671	4.529	4.871	4.582	4.486	4.979	5.009
6	6.223	6.490	4.871	4.937	4.875	4.583	5.638	5.256
7	5.972	6.715	4.897	4.854	4.767	4.531	5.383	5.365
8	5.996	6.236	4.943	5.203	4.818	4.888	5.372	5.415
9	5.799	6.247	4.881	5.091	4.810	4.715	5.422	5.278
10	5.873	6.527	4.742	4.886	4.537	4.442	5.256	5.165
11	6.244	6.323	4.769	4.863	4.553	**4.441**	5.215	5.123
12	6.680	7.439	4.797	4.791	4.603	4.498	5.301	5.265
13	8.299	6.245	4.869	4.792	4.495	4.555	5.331	5.318
14	8.953	8.393	5.189	4.842	4.484	4.562	5.357	5.344
15			5.216	4.891	**4.445**	4.529	5.333	5.255

In Table 5.7, we present the RMSE for the temperature at the station TH and we observe that the lowest RMSE is 1.149 attained at $L = 36$ and $r = 7$ for the SSA-R forecasting and 1.092 attained at $L = 48$ and $r = 9$ for the SSA-V forecasting. Overall, the RMSE is small for $r = 9$ and $L \geq 48$.

Table 5.7: The RMSE of $1, 2, \ldots, 12$-month ahead forecasts using SSA-R and SSA-V forecasting algorithms for temperature at the station TH.

	$L = 16$		$L = 24$		$L = 36$		$L = 48$	
r	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V	SSA-R	SSA-V
4	1.494	2.110	1.813	2.059	1.777	2.018	1.777	2.018
5	1.206	1.278	1.193	1.233	1.206	1.236	1.258	1.182
6	1.315	1.285	1.266	1.234	1.175	1.245	1.235	1.191
7	1.373	1.403	1.247	1.292	**1.149**	1.212	1.190	1.114
8	1.422	1.377	1.341	1.277	1.180	1.205	1.207	1.107
9	1.517	1.409	1.533	1.272	1.277	1.190	1.295	**1.092**
10	1.597	1.388	1.575	1.271	1.353	1.202	1.325	1.104
11	1.885	1.435	1.639	1.422	1.448	1.288	1.312	1.148
12	1.862	1.765	1.612	1.414	1.479	1.339	1.372	1.225
13	5.358	1.799	1.713	1.419	1.392	1.297	1.473	1.354
14	4.076	1.901	1.988	1.542	1.434	1.297	1.465	1.311
15	3.416	3.413	2.025	1.455	1.490	1.284	1.468	1.292

In Figures 5.8 and 5.9, we depict humidity and temperature at three stations together with $1, 2, \ldots, 12$-month ahead SSA-R and SSA-V forecasts from 12 truncations points. We can observe that SSA-R and SSA-V forecasts are very stable and close each other. Note that humidity is quite volatile and therefore, forecasts are not close to the observed values. In contrast, temperature has the clear annual pattern and consequently SSA forecasts are much more accurate.

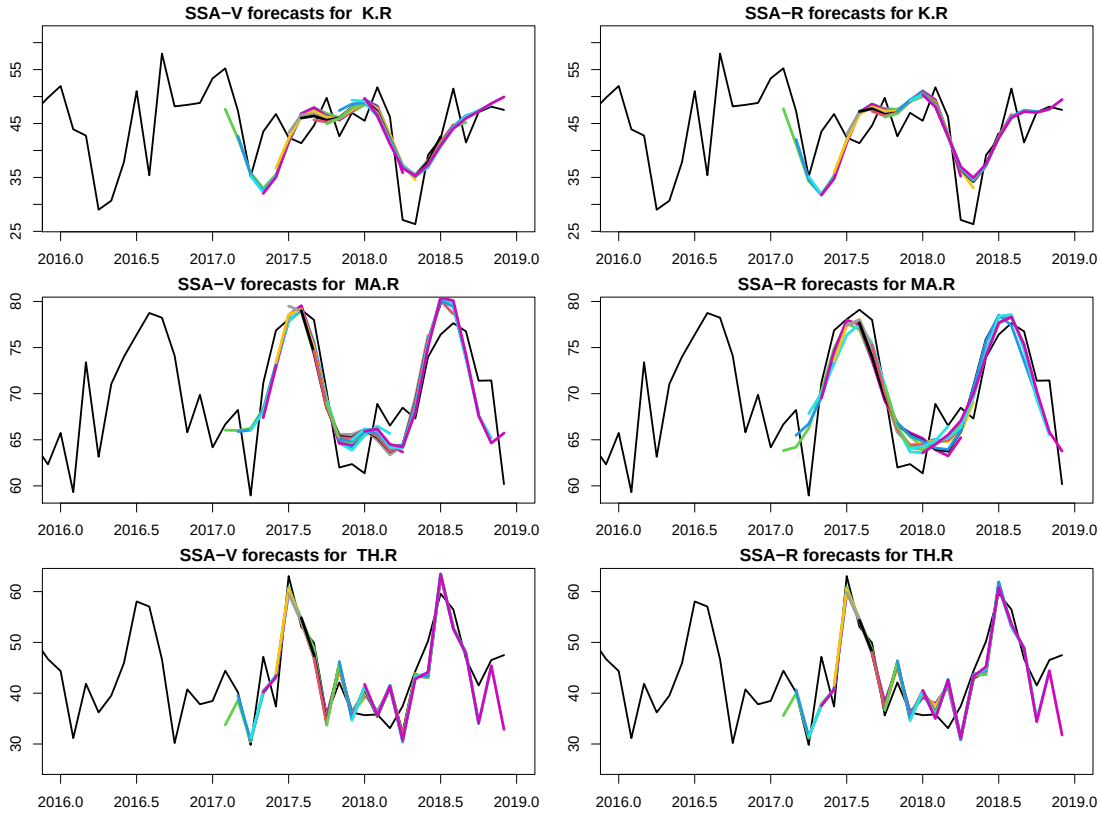

Figure 5.8: Humidity (black) at the stations K, MA and TH with $1, 2, \ldots, 12$-month ahead SSA-R and SSA-V forecasts (colored) with parameters L and r providing the smallest RMSE.

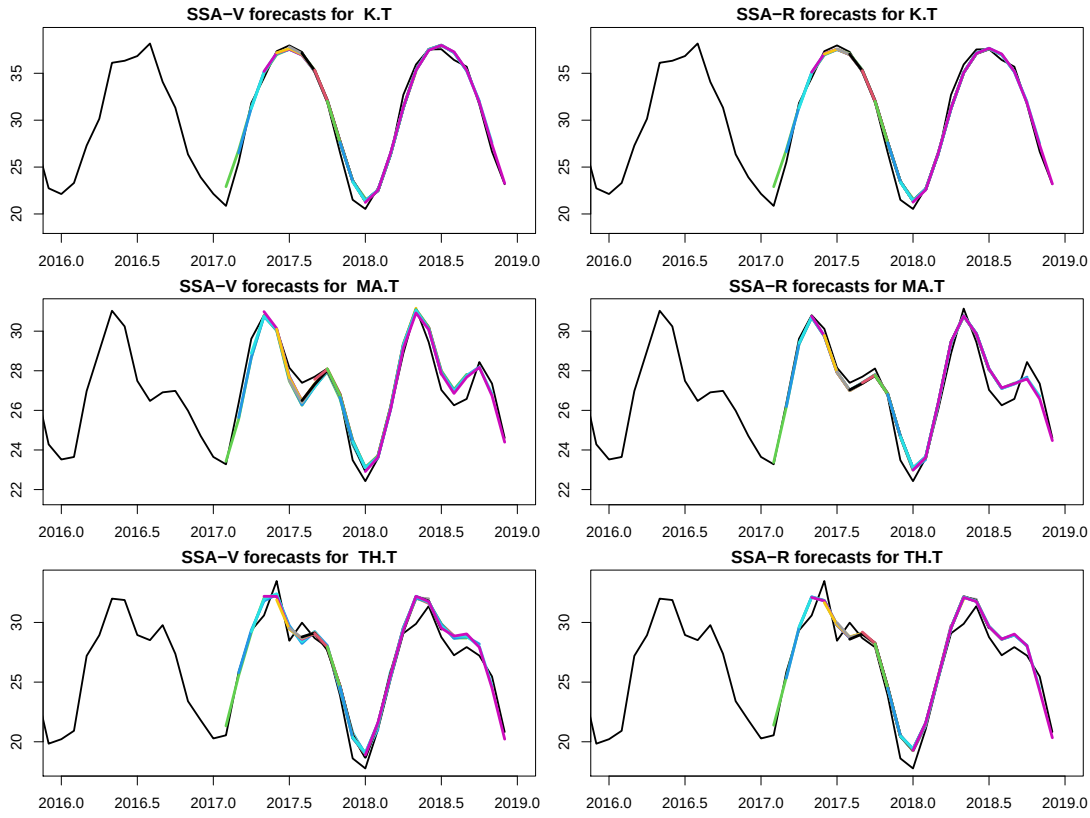

Figure 5.9: Temperature (black) at the stations K, MA and TH with $1, 2, \ldots, 12$-month ahead SSA-R and SSA-V forecasts (colored) with parameters L and r providing the smallest RMSE.

5.4 Study of the automatic choice of parameters

For proper assessing the accuracy of forecasts with the automatic choice of parameters, we consider the time series from Jan 2018 to Dec 2018 as future values and perform the automatic choice of parameters for time series from Jan 2009 to Dec 2017. This automatic choice is based on minimizing the $\text{RMSE}_{Jan2016}^{Dec2016}$ for $1, 2, \ldots,$ 12-month ahead forecasts with truncation points from Jan 2016 to Dec 2016 and reported in Table 5.8. Also we compute the accuracy of future forecasts by means of the $\text{RMSE}_{Jan2017}^{Dec2017}$ with chosen parameters and the efficiency of future forecasts as the ratio of the smallest $\text{RMSE}_{Jan2017}^{Dec2017}$ across different parameters to the $\text{RMSE}_{Jan2017}^{Dec2017}$ with chosen parameters.

We can see that the efficiency of forecasting with the automatic choice is around 90% which is rather high. Overall, SSA-V forecasts are little more accurate than SSA-R forecasts with the automatic choice. We can observe that the $\text{RMSE}_{Jan2017}^{Dec2017}$ is close to the $\text{RMSE}_{Jan2016}^{Dec2016}$ confirming that the structure of time series has not changed much.

By comparing SSA with other forecasting methods as in Figure 5.8, we can observe that SSA is producing forecasting results when compared to ARIMA, ETS and NN for $1, 2, \ldots,$ 12-month ahead forecasts with truncation points from Jan 2017 to Dec 2017.

In Table 5.8, we also present the RMSE for ARIMA forecasting with automatic parameters implemented in the function `auto.arima` from the R package `forecast`. Specifically, we used the stationary seasonal ARIMA model for fitting and forecasting our monthly data. In addition, we also add ETS and RNN forecasting as presented in the functions `ets` and `nnetar` respectively. We can see that SSA-R

and SSA-V forecasting is better than ARIMA forecasting of humidity at the station TH and temperature at all three stations. Again that SSA-R and SSA-V forecasting is better than ETS and RNN forecasting of humidity and temperature at all three stations.

Table 5.8: The automatic choice of parameters L and r based on the $\text{RMSE}^{Dec2016}_{Jan2016}$ for 1,2,...,12-month ahead forecasts of humidity and temperature for the stations K, MA, and TH, the $\text{RMSE}^{Dec2017}_{Jan2017}$ with chosen parameters and its efficiency. The last column contains the $\text{RMSE}^{Dec2017}_{Jan2017}$ for ARIMA, ETS and RNN forecasting with automatic parameters.

		SSA-R					SSA-V					ARIMA	ETS	RNN
Series	Station	L	r	$\text{RMSE}^{Dec2016}_{Jan2016}$	$\text{RMSE}^{Dec2017}_{Jan2017}$	Eff	L	r	$\text{RMSE}^{Dec2016}_{Jan2016}$	$\text{RMSE}^{Dec2017}_{Jan2017}$	Eff	$\text{RMSE}^{Dec2017}_{Jan2017}$	$\text{RMSE}^{Dec2017}_{Jan2017}$	$\text{RMSE}^{Dec2017}_{Jan2017}$
Humidity	K	36	5	6.441	5.479	1	36	13	5.906	6.192	0.875	5.783	6.021	6.408
Temperature	K	36	7	1.001	1.163	0.77	24	8	0.981	1.110	0.876	1.485	0.769	1.065
Humidity	MA	24	6	3.452	2.801	0.846	24	9	3.320	3.831	0.825	3.909	2.801	3.947
Temperature	MA	24	5	0.758	0.555	1	24	5	0.702	0.603	0.976	1.540	0.564	0.8
Humidity	TH	36	10	6.169	4.351	0.982	24	13	5.786	4.666	0.914	5.253	5.667	5.262
Temperature	TH	36	5	1.115	1.218	0.946	24	6	1.112	1.250	0.944	2.091	1.118	1.34

5.5 Discussion

We apply the RMSE to study the accuracy of automatic and dependence of the RMSE on parameters for SSA-V and SSA-R forecasts. To make the forecasts we use parameters that are $L = 16, 24, 36$ and 48 and $r = 4$ to $r = 15$. The selection of parameters yield the optimal pair (L, r) for SSA-R and SSA-V with retrospective 12-month ahead forecasts. For automated SSA-R and SSA-V, we use the R program's functions to perform forecasting of automatic algorithms to get optimal pair (L, r) which correspond to the minimal RMSE.

5.5.1 Parameters effects

The window length L and the number of leading components r are needed for SSA signal forecasting. For a given time series, there are many methods for determining L and r. The selection of parameter the window length L in SSA is significant because it is dependent on the time series structure and the forecasting analysis target [59, 66]. Some previous studies suggested selecting $L = \frac{N}{4}$; the window length L should not be larger than $\frac{N}{2}$, [42]. In [43], the authors suggested L should not exceed half of the length given time series. The first part of choosing the parameter is selecting a value of L appropriate to seasonal fluctuations, for example, by analysing the periodogram to check if there are any strong signals. Then, an analysis of paired eigenvectors enables the differentiation of signal from noise. Depending on the length of the time series, one can pick the required number of eigenvalues r and during the reconstruction stage, we consider the difference to be noise.

In Section 5.3, we discussed the RMSE of $1, 2, \ldots,$ 12-month ahead forecasts using

SSA-R and SSA-V forecasting algorithms for monthly time series of the humidity and temperature at the stations K, MA and TH. The length N of the monthly time series is 120, and we take $5N/2 = 48$, $3N/10 = 36$, $N/5 = 24$ and $N/7.5 = 16$ as window lengths. With consideration of these selected window lengths and the SVD of the trajectory matrix, several bunch components are obtained and ordered according to their contributions to the decomposition. It can be found that the window length of $L = 36$ can achieve a good result and that can be used for forecasting.

In [59], the authors explain that choosing a r greater than what is actually required results in noise being included in the reconstructed signal. In [62, 63, 78], the authors present two other approaches to the selection of L and r. We applied different parameter values for L and r until we got optimal values for pair (L, r) by checking the lowest RMSE. Our automated approach focuses on the minimization of forecasting errors (RMSE) during the validation (training) phase by considering all possibilities of L and r as optimal SSA choices for forecasting. The selection of L and r are optimised to obtain the best possible forecast from a statistical perspective.

We also find that the automatic choice of parameters for SSA-V and SSA-R forecasting algorithms is depending on the window length L and the number of singular values r. In [72], the authors explain how automatic time series forecasting works with the R package forecast.

We believe it is important to briefly discuss the computational complexity of the two approaches to parameter selection. Both depends on SSA choices of L and r at the decomposition and reconstruction stages, which makes them very similar in terms of computation. There is no significant difference in computational complexity; both methods would take the same amount of time to produce

forecasts [33]. When dealing with outliers or large shocks in the series, the vector forecast is usually more reliable than the recurrent one [61].

5.5.2 Comparison of SSA, ARIMA, ETS and RNN

A real time series dose not satisfy any model [42]. Seasonal ARIMA, ETS and RNN methods correspond to concrete model families and the frequency of the periodic component should be specified. In [42, Sec 3.5.8], the authors discussed the idea of criteria of using ARIMA, ETS and RNN methods that can be used in some measure of correspondence between the model and the time series and then, adjust it by the number of parameters in the model. In Section 5.4, the study of automatic choice of parameters discussed in detail and the RMSE is calculated.

To compare accuracy of the methods, we compute the accuracy of future forecasts by means of the $\mathrm{RMSE}_{Jan2017}^{Dec2017}$ with chosen parameters and the efficiency of future forecasts as the ratio of the smallest $\mathrm{RMSE}_{Jan2017}^{Dec2017}$ across different parameters to the $\mathrm{RMSE}_{Jan2017}^{Dec2017}$ with chosen parameters. We used the stationary seasonal ARIMA , ETS and RNN models for fitting and forecasting monthly time series.

Comparing SSA forecasting with ARIMA, ETS, and RNN, in general, we can observe that SSA-V and SSA-R forecasts have the best result for forecasting for the temperature and humidity at three stations. We conclude SSA works well with different series and outperforms other forecasting techniques.

In Table 5.8, we can observe that SSA forecasting has the lowest value of the RMSE compare to ARIMA, ETS, and RNN which indicate a good model for forecasting based on monthly time series.

5.5.3 Future Forecast

We create 12-month ahead forecasts employing the 5 methods. Since the last observation available is in December 2018, then the forecasts must start in January 2019 and end in December 2019. These 12-month ahead forecasts for stations K, MA and TH can be seen in Figure 5.10, each method represented by their own distinct colour.

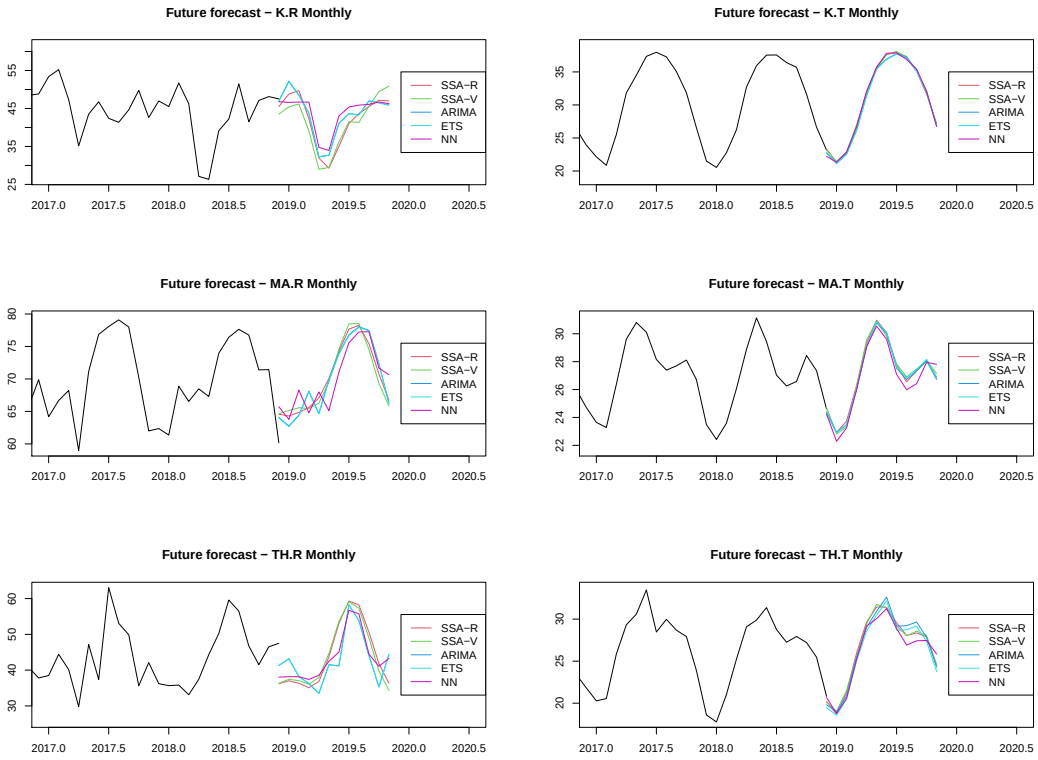

Figure 5.10: Future forecast of 1, 2,..., 12-month ahead forecasts using five forecasting algorithms for humidity (left) and temperature (right) at stations K, MA and TH.

We can see that the methods mostly produce diverse forecasts for humidity and

temperature, where all five forecasts agree. As discussed, the simple structure ensures that any forecasts generated have very minimal fluctuation. We can be confident for the 12-month ahead forecasts of monthly time series are accurate and similar to the true future observations for all methods used. From the information retrieved from analysing the retrospective forecasts, it is reasonable to assume that the 12-month ahead forecast for the time series of humidity is very volatile. It is more likely to follow the forecasts that created by SSA-R and SSA-V than any of the other methods, although there are considerable differences between the forecasts generated.

5.6 Chapter summary

In Chapter 5, by using real data representing monthly temperature and humidity in Oman, we have provided a statistical framework for studying which SSA forecasting algorithm is best. We demonstrated that the sensitivity of the RMSE for retrospective forecasts is rather small to the parameters L and r. We have shown that the efficiency of SSA forecasts with the automatic choice of parameters is rather high. We also found that SSA-R and SSA-V forecasts are similar to each other with a slight dominance of SSA-V forecasts.

Comparing SSA forecasting with ARIMA, ETS and RNN has been performed. The evidence from monthly time series shows that SSA can provide a powerful tool for forecasting the monthly temperature and humidity and that it outperforms the competing models. We believe that the findings presented in Chapter 5 help to increase the confidence of researchers to recognise and apply the SSA forecasting algorithms with the automatic choice of parameters.

Chapter 6

Forecasting daily temperature and humidity

In Chapter 6, we focus on the performance of various SSA forecasting algorithms when applied to daily time series of humidity and temperature in Oman. We apply recurrent SSA (SSA-R), SSA-R (original), and vector SSA (SSA-V) forecasting algorithms based on SSA with double projection and SSA without projection. We also study the effect of series length and the choice of SSA parameters on the performance of the aforementioned algorithms.

6.1 Introduction

Forecasts of daily temperature and humidity can help to support long term planning and decision making. Forecasting temperature and humidity may help deal with incidences of drought owing to global warming [51]. Consequently, scientific efforts to develop forecasting algorithms has intensified. A thorough review on the applications of SSA show that there are no general rules for finding the best algorithms. This means that, we have to look for the best model among a variety of models using some criteria. We note also that forecasts of

daily temperature and humidity depend not only on the method but also on the parameters, data and the lengths of time series.

The structure of Chapter 6 is as follows.

- In Section 6.2, we discuss several past studies on daily forecasting.
- In Section 6.3, we apply three forecasting algorithms by using SSA-R, SSA-R (original) and SSA-V in two variants: SSA with double projection and SSA without projection.
  - In Section 6.3.1, we introduce daily time series of temperature and humidity at three meteorological stations in Oman.
  - In Section 6.3.2, we present the optimal tuple $(\textbf{method}, N, L, r)$ and the RMSE of $1, 2, \ldots, 14$-day ahead forecasts and $1, 2, 3$-days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection.
  - In Section 6.3.3, we discuss the numerical study of the optimal tuple.
- In Section 6.4, we provide a summary of Chapter 6.

6.2 Literature review

Daily time series analysis and forecasting are of utmost importance in industry and manufacturing for the principal reason that forecasting models have the potential to improve planning and decision-making. The precise methodology of the forecasting technique, the accuracy of these predictions, and the ability to correctly predict the course of future changes play a central role [58].

Researchers have studied the problem of forecasting daily temperature and humidity especially in Oman, a country characterised by an arid semi-arid climate. In Oman, the daily forecasts facilitate planning the allocation of agricultural and the associated agrarian economy [91]. It follows the forecasting of meteorological time series with a high degree accuracy is highly important [148].

In [104], the authors discuss the forecasting performance for several methods which are affected by the increase of the intermittence of the data, as well as by the coefficient of variation. Their analysis shows that, as the values of these two factors increase, the respective accuracy of all methods decreases. It was also found that some forecasting methods performed better for longer forecasting horizons. The study aims to identify the main determinants of forecasting accuracy, by investigating several popular forecasting methods and seven time series features (seasonality, trend, cycle, randomness, number of observations, inter demand interval and coefficient of variation) and as well as one strategic decision (the forecasting horizon).

In a similar vein, in [111], the authors observe that resorting to a large window length has the potential to produce a good model fit, yet the approach is unlikely to produce a parsimonious forecasting model.

6.3 Applications

This section considers applications for daily time series by using SSA-R, SSA-R (original) and SSA-V in two forms which are SSA with double projection and SSA without projection. The RMSE of the retrospective forecast used to assess the accuracy.

6.3.1 Daily time series

This section consider the accuracy of the forecasting algorithms for daily time series of temperature (measured in deg C) and humidity (measured in %), with the data provided by the Directorate General of Meteorology of Oman. The data was collected from January 2009 to December 2018 at three meteorological stations in the Sultanate of Oman: the Khasab Airport (K), the Masirah (MA) and the Thumrait (TH).

In Figure 6.1, the graphs depict all time series which do not exhibit trends as explained in [5] and Chapter 4. We can observe that the pattern of temperature has found to be rather stable. In particular, temperature at the station K has a simple sinusoidal shape, while the pattern of temperature at stations MA and TH is more complicated. We can also see that humidity is highly volatile and the pattern is visible at the station MA. Note that humidity at the station MA is much larger than humidity at stations K and TH.

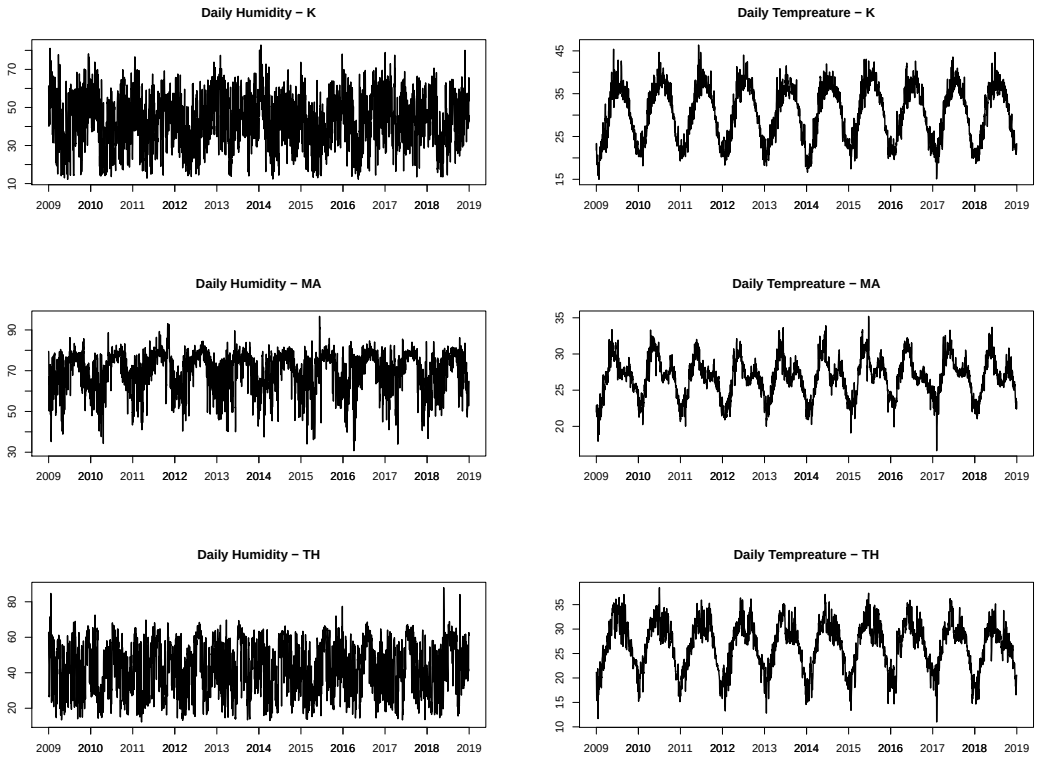

Figure 6.1: Daily humidity (left) and temperature (right) at stations K, MA and TH from 2009 to 2018.

In our daily time series of temperature and humidity, we prefer to forecast 14 and 3 days ahead since the uncertainty of weather forecasts increases in proportion to the length of the forecasting horizon. We chose two different horizons: short and medium horizons. The short range of forecasting is 3 days ahead and the medium range is 14 day ahead [50]. For this reason, we work to demonstrate the optimal tuple for 1,2,...,14 day ahead forecasts and 1,2,3 days ahead forecasts. The optimal values of L and r which are used in this chapter, it is the best values based on the lowest value of RMSE.

6.3.2 The optimal tuple for 14 day and 3 days ahead forecasts

In this section, we compare three SSA forecasting algorithms: SSA-R, SSA-R (original) and SSA-V by using two methods which are SSA with double projection and SSA without projection. Section 6.3.2 studies how the RMSE of 1,2,...,14 day ahead forecasts and examines the RMSE of 1,2,3 days ahead forecasts across several truncation points depends on the following parameters: the window length L and the number of singular values r.

This section presents the optimal tuple (**method**, N, L, r). It has been selected by getting the optimal pairs (L, r) with the lowest RMSE on retrospective 14-day and 3-days ahead forecasts for stations K, MA and TH using the three forecasting algorithms. The optimal tuple contains as the method element either SSA with double projection or SSA without projection. It further contains the number of time series length, the window length L and the number of singular values r.

For analysing and diagnostic of the forecasting accuracy, we have used different N. For SSA with double projection and SSA without projection, we use the following range for the parameters in Table 6.1.

Table 6.1: Range of the parameters.

T_0	N	L	r
2892	30	$\{2, 3, \ldots, 16\}$	$\{2, 3, \ldots, 15\}$
	90	$\{16, 20, 24, 30\}$	$\{4, 5, \ldots, 15\}$
	180	$\{16, 20, 24, 30\}$	$\{4, 5, \ldots, 15\}$
	360	$\{100, 101, \ldots, 150\}$	$\{1, 2, \ldots, 18\}$
	All	$\{90, 180, 360, 500\}$	$\{4, 5, \ldots, 18\}$

Then, we consider the optimal pair (L, r) based on the lowest RMSE.

Expert analysis of the singular values of the trajectory matrix, weighted correlation among the components of the time series and errors of the reconstruction and forecasting are the main tools for choosing the optimal parameters. Here, we consider the last one and use the RMSE by the following description.

The RMSE is the most popular criterion to measure the error of forecasts, see e.g. [42, 59, 89, 142]. The RMSE of the $1, 2, \ldots, h$-step ahead forecasts with several truncation is given by

$$\text{RMSE} = \left(\frac{1}{h(m+1)} \sum_{i=0}^{m} \sum_{j=1}^{h} (\tilde{y}_{\mathrm{T}_i,j} - y_{\mathrm{T}_{i+j}})^2 \right)^{1/2}, \tag{6.1}$$

where $\mathrm{T}_i = \mathrm{T}_0 + ih$; $i = 0, 1, \ldots, m$ and m is the largest integer satisfy in condition $\mathrm{T}_0 + (m+1)h \leq \mathrm{T}_{\max}$ ($\mathrm{T}_{\max}$ denote the index of last observation) and $\tilde{y}_{\mathrm{T}_i,j}$ is the j-step ahead forecast using the truncated time series $\{y_{\mathrm{T}_i-N+1}, \ldots, y_{\mathrm{T}_i}\}$ for $N \leq \mathrm{T}_i$. However, in this study, we consider two approaches: (1) N is a fixed number like as 30 and (2) N is vary by going ahead and particularly $N = \mathrm{T}_i$. We call the second approach by $N =$ all. The first approach uses less observations than the second approach and it is worth to see if it is enough for modelling. Table 6.2 describes the discrepancy between two approaches. We note that the notations in columns 3 and 4 are the same but they are not the same as the models come from different sources.

Table 6.2: Description of modelling and forecasting.

Approach	Observation for modelling	h-steps ahead forecasts	MSE
Recent observation	$y_{\mathrm{T}_0-N+1}, \ldots, y_{\mathrm{T}_0}$	$\tilde{y}_{\mathrm{T}_0,1}, \ldots, \tilde{y}_{\mathrm{T}_0,h}$	$\frac{1}{h}\sum_{j=1}^{h}\left(\tilde{y}_{\mathrm{T}_0,j} - y_{\mathrm{T}_0+j}\right)^2$
All observation	$y_1, y_2, \ldots, y_{\mathrm{T}_0}$		
Recent observation	$y_{\mathrm{T}_0-N+1+h}, \ldots, y_{\mathrm{T}_0+h}$	$\tilde{y}_{\mathrm{T}_0+h,1}, \ldots, \tilde{y}_{\mathrm{T}_0+h,h}$	$\frac{1}{h}\sum_{j=1}^{h}\left(\tilde{y}_{\mathrm{T}_0+h,j} - y_{\mathrm{T}_0+h+j}\right)^2$
All observation	$y_1, y_2, \ldots, y_{\mathrm{T}_0+h}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Recent observation	$y_{\mathrm{T}_0-N+1+mh}, \ldots, y_{\mathrm{T}_0+mh}$	$\tilde{y}_{\mathrm{T}_0+mh,1}, \ldots, \tilde{y}_{\mathrm{T}_0+mh,h}$	$\frac{1}{h}\sum_{j=1}^{h}\left(\tilde{y}_{\mathrm{T}_0+mh,j} - y_{\mathrm{T}_0+mh+j}\right)^2$
All observation	$y_1, y_2, \ldots, y_{\mathrm{T}_0+mh}$		

6.3.3 Numerical study of the optimal tuple

The optimal tuple for 14 day and 3 days ahead forecasts are presented in Section 6.3.3.

Table 6.3 and Table 6.4 exhibit the RMSE of $1, 2, \ldots, 14$-day ahead forecasts and the RMSE of $1, 2, 3$-days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity and temperature at stations K, MA and TH. Bold values in Tables 6.3 and 6.4 are the lowest RMSE in terms of the factors: method, h, the algorithms and the station. This means that we find the lowest RMSE in terms of N. For instance, if we employ SSA with double projection on the observations of time series of humidity at the station K to produce 3 steps ahead forecasting, we found the lowest five values of the RMSE $\{12.806, 13.178, 12.769, 12.822, 12.639\}$ which give the lowest RMSE equal to 12.639 (see Table 6.3).

Table 6.3: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts and the RMSE of $1, 2, 3$-days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at stations K, MA and TH.

Method	N	h	SSA-R			SSA-R (original)			SSA-V		
			K	MA	TH	K	MA	TH	K	MA	TH
SSA with double projection	30	3	12.806	8.717	14.736	12.739	**8.252**	14.518	12.687	8.627	14.644
		14	16.57	10.639	17.795	15.799	10.639	18.164	16.225	10.562	17.614
	90	3	13.178	9.544	14.825	12.505	9.129	14.286	12.769	9.642	14.301
		14	14.685	10.007	16.442	14.584	9.903	16.100	14.017	9.694	15.223
	180	3	12.769	9.004	**14.361**	**12.463**	8.665	**13.838**	12.809	9.386	**13.929**
		14	13.681	10.007	14.416	13.943	9.903	14.540	14.136	9.694	14.724
	360	3	12.822	9.16	14.752	12.534	9.125	14.234	13.234	9.079	14.279
		14	13.567	8.778	14.840	13.024	8.515	14.759	13.159	8.529	14.514
	ALL	3	**12.639**	**8.331**	14.570	12.464	8.314	14.693	**12.465**	**8.306**	14.791
		14	**12.347**	**8.182**	**13.236**	**12.199**	**8.211**	**13.225**	**12.336**	**8.342**	**13.217**
SSA without projection	30	3	14.192	8.697	14.757	13.796	**8.253**	14.482	13.483	8.609	14.754
		14	21.035	21.57	24.218	20.992	21.137	21.526	20.899	21.476	41.632
	90	3	13.369	10.371	15.524	12.818	9.626	15.017	12.993	10.119	15.285
		14	15.001	11.022	16.423	14.901	11.609	15.984	14.025	11.784	15.231
	180	3	12.884	8.916	14.101	12.863	8.490	13.949	13.297	8.848	14.5001
		14	13.657	9.857	14.551	12.599	10.166	14.960	13.417	10.140	14.872
	360	3	13.173	8.333	14.786	**12.599**	8.270	14.008	13.417	**8.265**	14.5001
		14	14.065	8.778	15.980	13.746	8.515	15.085	14.237	8.529	16.203
	ALL	3	**12.465**	**8.321**	**13.87**	12.923	8.343	**13.840**	**12.727**	8.342	**13.850**
		14	**12.055**	**8.166**	**13.216**	**12.054**	**8.210**	**13.248**	**12.039**	**8.180**	**13.215**

Table 6.4: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts and the RMSE of $1, 2, 3$-days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at stations K, MA and TH.

Method	N	h	SSA-R			SSA-R (original)			SSA-V		
			K	MA	TH	K	MA	TH	K	MA	TH
SSA with double projection	30	3	**2.049**	1.077	**1.989**	**2.022**	**1.032**	**1.924**	**2.043**	1.071	**1.983**
		14	2.34	1.287	2.807	2.307	1.322	2.661	2.300	1.273	2.779
	90	3	2.163	1.223	2.226	2.069	1.149	2.115	2.199	1.189	2.259
		14	2.572	1.535	2.790	2.550	1.565	2.758	2.560	1.537	2.924
	180	3	2.156	1.171	2.112	2.106	1.083	2.049	2.157	1.153	2.178
		14	2.555	1.420	2.357	2.603	1.468	2.363	2.786	1.463	2.454
	360	3	2.306	1.245	2.550	2.1134	1.113	2.222	2.1037	1.090	2.2258
		14	2.654	1.405	3.017	2.429	1.342	2.760	2.426	1.363	2.821
	ALL	3	2.160	**1.059**	2.1045	2.1134	1.038	2.048	2.1037	**1.037**	2.0815
		14	**1.965**	**1.069**	**2.115**	**1.9305**	**1.044**	**2.144**	**1.927**	**1.057**	**2.123**
SSA without projection	30	3	2.330	1.077	1.990	2.205	**1.033**	**1.926**	2.286	1.071	**1.981**
		14	4.634	2.329	3.053	4.700	2.304	2.806	7.188	4.390	4.073
	90	3	2.178	1.266	2.164	2.084	1.193	2.109	2.233	1.192	2.221
		14	4.630	1.561	2.824	2.684	1.591	2.815	2.610	1.552	2.930
	180	3	2.199	1.218	2.125	2.107	1.156	2.101	2.825	1.154	2.182
		14	2.608	1.425	2.365	2.669	1.471	2.367	2.825	1.466	2.486
	360	3	2.254	1.214	2.300	**2.0588**	1.321	2.116	2.825	1.252	2.16
		14	2.672	1.204	3.003	2.426	1.255	2.722	2.635	1.252	3.054
	ALL	3	**2.131**	**1.068**	**2.054**	2.120	1.051	2.006	**2.120**	**1.058**	2.014
		14	**1.956**	**1.069**	**2.114**	**1.925**	**1.044**	**2.146**	**1.923**	**1.043**	**2.125**

In order to find the best combination of the factors N, method and algorithm, we specified combinations that provided the lowest RMSE for forecasting as shown in Tables 6.3 and 6.4. The optimal tuple of the SSA forecasting algorithms by applying SSA with double projection and SSA without projection for humidity and temperature at stations K, MA and TH are summarised in Table 6.5 and 6.6. Tables 6.5 and 6.6 present that the optimal tuple of short term forecasting using smaller observations could be fine when we apply SSA with double projection. It

shows also that SSA-R (original) for short term forecasting is superior to SSA-R and SSA-V. Nevertheless, conclusions for $h = 14$ are completely different; we observe that the set of all observations provide more accurate medium term forecasts than the truncated set. In addition, SSA with double projection does not improve the accuracy of long term forecasts and SSA-V and SSA-R are superior to SSA-R (original).

Table 6.5: The optimal tuple for making $1, 2, \ldots, 14$-day ahead forecasts of humidity and temperature at stations K, MA and TH.

		The optimal tuple			
Series	Station	Method	N	L	r
Humidity	K	SSA-V forecasting without projection	All	360	8
	MA	SSA-R forecasting without projection	All	360	8
	TH	SSA-V forecasting without projection	All	360	8
Temperature	K	SSA-V forecasting without projection	All	360	8
	MA	SSA-V forecasting without projection	All	360	6
	TH	SSA-R forecasting without projection	All	360	8

Table 6.6: The optimal tuple for making $1, 2, 3$-days ahead forecasts of humidity and temperature at stations K, MA and TH.

		The optimal tuple			
Series	Station	Method	N	L	r
Humidity	K	SSA-R (original) forecasting with double projection	180	30	4
	MA	SSA-R (original) forecasting with double projection	30	3	2
	TH	SSA-R (original) forecasting with double projection	180	24	4
Temperature	K	SSA-R (original) forecasting with double projection	30	3	2
	MA	SSA-R (original) forecasting with double projection	30	3	2
	TH	SSA-R (original) forecasting with double projection	30	3	2

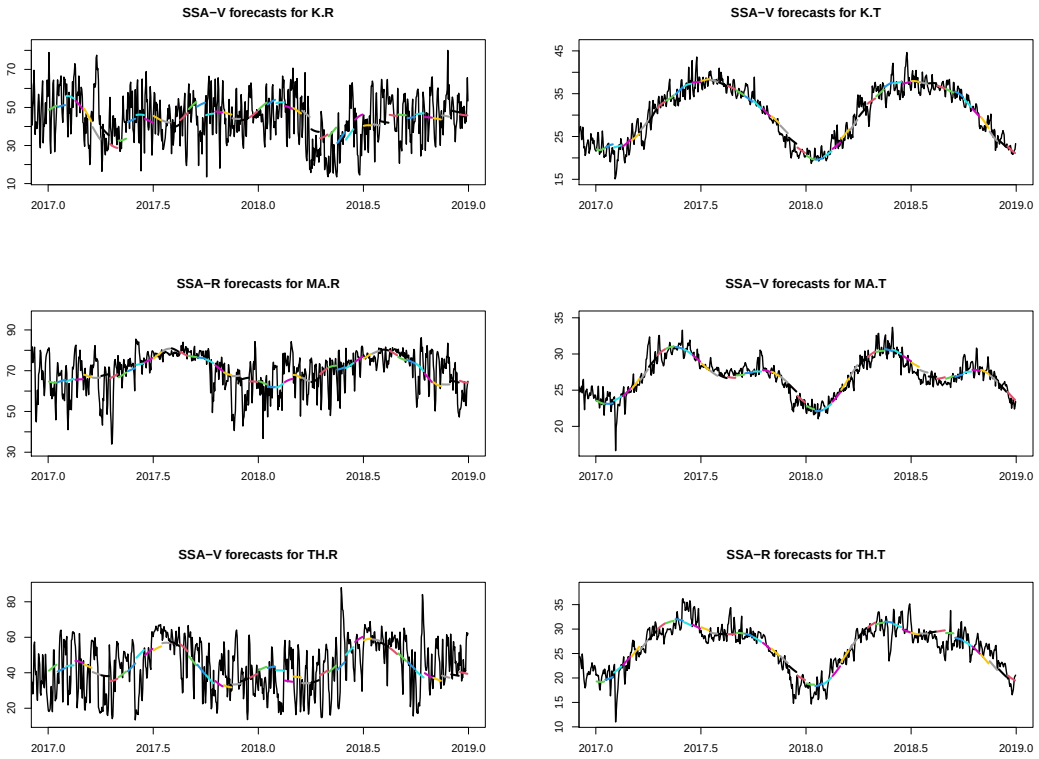

Figure 6.2: Humidity (left) and temperature (right) at stations K, MA and TH with $1, 2, \ldots, 14$-day ahead SSA-R and SSA-V forecasts (colored) with parameters L and r providing the smallest RMSE.

In Figure 6.2, we depict humidity and temperature at all three stations with $1, 2, \ldots, 14$-day ahead SSA-R and SSA-V forecasts obtained across several truncation points. We can see that SSA-R and SSA-V forecasts for SSA without projection are very stable and close to each other. Note that humidity is quite volatile and therefore, forecasts are not close to the observed values. By contrast, temperature has a clear daily pattern and consequently SSA forecasts for temperature are far more accurate.

In Figure 6.3, we jointly represent humidity and temperature at all three stations

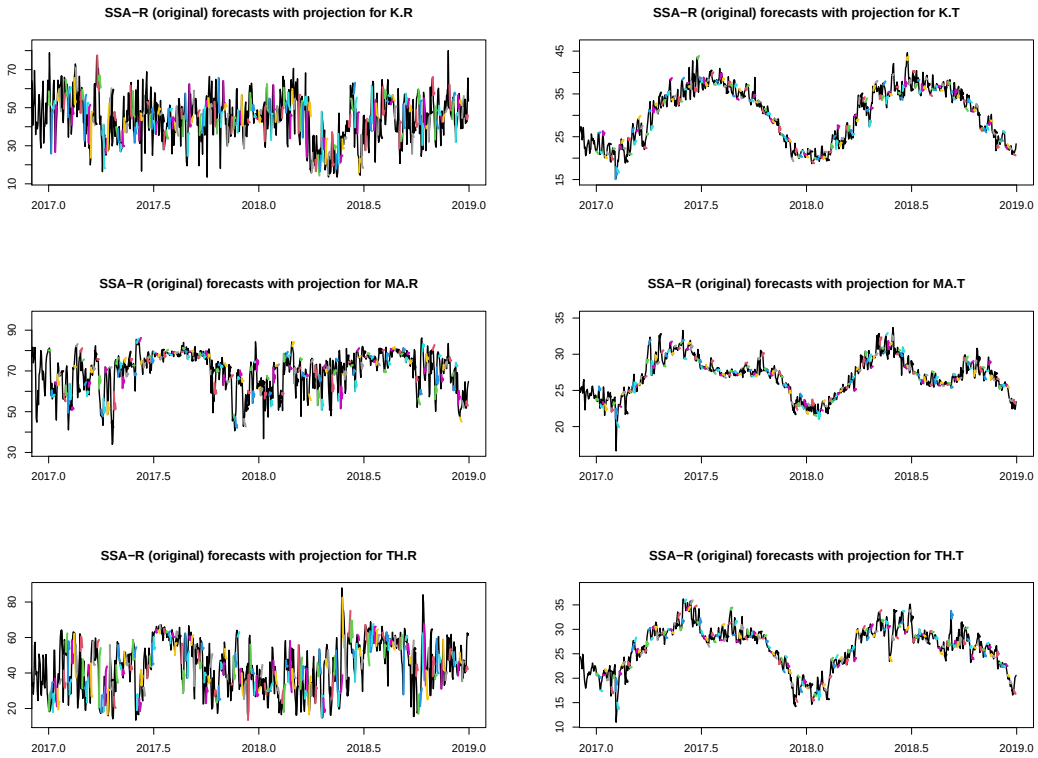

Figure 6.3: Humidity (left) and temperature (right) at stations K, MA and TH with 1,2,3-days ahead SSA-R (original) forecasts (colored) with parameters L and r providing the smallest RMSE.

together with $1, 2, 3$ days ahead for SSA-R (original) forecasts across several truncation points. We can see that SSA-R (original) forecast with projection is very stable. Note that humidity and temperature form clear daily patterns and consequently SSA forecasts are much more accurate.

6.4 Chapter summary

The accuracy of three forecasting algorithms (SSA-R, SSA-R (original) and SSA-V) has been analysed when forecasting daily temperature and humidity in Oman. SSA with double projection and without projection have been considered.

In general, SSA-R (original) with projection outperform SSA-R and SSA-V for $1, 2, 3$ days ahead forecasts. The performance of forecasting by using SSA with double projection and SSA without projection for daily time series depends on the forecast horizon, the window length L and the number of singular values r.

For the time series of humidity, notice that SSA-V forecasts have a negligibly smaller value of the RMSE than SSA-V and SSA-R (original) forecasts at stations K and TH while SSA-R forecasts have smaller values of the RMSE than SSA-V and SSA-R (original) forecasts at station MA by using forecasting SSA without projection for all period series with $1, 2, \ldots, 14$-day ahead forecasts. At the same time, by using $1, 2, 3$ days ahead forecasts, SSA-R (original) forecasts have the lowest RMSE with SSA projection at stations K, MA and TH when $N = 30$ and $N = 180$ days.

For the time series of temperature, SSA-V forecasts have a marginally smaller value of the RMSE than SSA-R and SSA-R (original) forecasts at the station K while SSA-R forecasts have the lowest RMSE compared to SSA-V and SSA-R

(original) forecasts at stations MA and TH by using forecasting SSA without projection for all period series. By using $1, 2, 3$ days ahead forecasts, SSA-R (original) forecasts provide the lowest RMSE with SSA projection at stations K, MA and TH when $N = 30$ days.

The accuracy of forecasting algorithms and the sensitivity of the RMSE to parameters L and r for retrospective forecasts is rather small. SSA-R and SSA-V forecasts have greater similarity to each other with a slight dominance of SSA-V forecasts.

By studying the accuracy of forecasting algorithms for two different forecast horizons, we hope to provide important empirical tools to allow decision-makers to make informed decisions based on better analysis of daily time series.

Chapter 7

Conclusion

This chapter has summarised and reflected on the work reported in this thesis. The summary here is brief since each chapter concludes with a detailed summary.

7.1 Research summary

Chapter 1 described the importance of time series and forecasting and then, outlined the research objectives. It introduced five research objectives and discussed the reasons for using SSA. It also described the structure of the thesis and the novel contribution.

Chapter 2 presented SSA methodology, including details about using SSA and choice of parameters. It also covers the SSA algorithms that uses for imputing missing values, SSA with projection and the use of SSA for forecasting. It described three benchmarking forecast models (ARIMA, EST and NN). The chapter also considered using the RMSE to measure the error of forecasts.

Chapter 3 studied hourly time series of temperature and humidity from six meteorological stations in Oman for the period from 2009 to 2018. The chapter covers three methods for imputing missing values: imputation using the SSA-based

iterative approach, imputation by multiple regression and imputation with lagging. We argued that regression with lagging produces more reliable imputations and provides a natural result for filling in gaps of any length.

Chapter 4 described an hourly time series for extracting annual oscillations and daily periodicities. It presented three tests for detecting trends in time series: the MK test, the SR test and the ITM test. We concluded that there are no monotonic trends in annual oscillations and daily periodicities over the ten year period from 2009 to 2018. In addition, we did not find any trends in the monthly variability of daily periodicities.

Chapter 5 provided a statistical framework for studying SSA forecasting algorithms. These algorithms have two parameters that should be chosen either by the researcher or by using automatic choice based on the RMSE of retrospective forecasts. We demonstrated the sensitivity of the RMSE for retrospective forecasts. We used recurrent SSA and vector SSA forecasts for monthly temperature and humidity in Oman from 2009 to 2018.

Chapter 6 investigated forecasting accuracy using daily time series of humidity and temperature by using SSA-R, SSA-R (original) and SSA-V in two forms which are SSA with double projection and SSA without projection. We studied how the RMSE of 1,2,. . . ,14 and 1,2,3 days ahead forecasts across several truncation points depend on the window length L and the number of singular values r.

7.2 Contributions

This thesis has made novel contributions in each of its five research topics: the SSA algorithms, imputing missing values, extracting annual oscillations and daily

periodicities, monthly forecasting and daily forecasting. This section summarizes these contributions.

Chapter 3 discussed three methods for imputing missing values: SSA-based iterative approach, multiple imputation by regression and regression with lagging. We demonstrated methods of imputation using meteorological data of Oman from 2009 to 2018. We argued that regression with lagging is the best method for imputing missing values because it provides a natural method of filling gaps of any length and produces more reliable results than other methods.

Chapter 4 used three tests for detecting a trend in the time series. Extracting the annual oscillations and the daily periodicities over ten years contributes to the meteorological research.

Chapters 5 and 6 investigated automatic choice of SSA parameters for forecasting. Chapters 5 and 6 provide data that help researchers understand which method of selecting SSA parameters is more suitable, the researcher's experience or automatic choice based on the RMSEs of retrospective forecasts. Chapters 5 and 6 considered the SSA algorithms to forecast monthly and daily temperature and humidity and considered different factors that contribute to decisions about SSA parameters, such as the structure of a time series and the length of a forecast.

The thesis supports the Directorate General of Meteorology in Oman by providing updated analysis of meteorological time series for the period 2010-2019.

Temperature, relative humidity and precipitation are the main types of meteorological data that impact economic capital and human populations in Oman. We have developed an understandable, innovative approach that uses statistical computation methods and meteorological time series. SSA is a faster and more precise model for forecasting time series. Development of this research can contribute

to the national capacities of Oman in the field of meteorological research and could raise awareness of the issues and challenges of climate change. Oman must focus on factors such as sustaining the ecosystem, building eco house projects and developing water resources as it contends with climate change. Research on various indicators of climate change can provide scientific facts for decision-makers that, in turn, can help them plan and formulate regulations and policies.

7.3 Future research directions

Each part of this thesis has given rise to interesting questions and research directions that would complement this study. One of the most important challenges is how the results of this research can be translated into practical recommendations for decision makers.

future work, we can explore two extensions of SSA which are complex SSA and multivariate SSA (MSSA) when performing tasks such as smoothing, change point detection and forecasting of complex values. In addition, MSSA may be very useful for analyzing several series with common structure. MSSA may also be used for establishing a causality between two series, for example the time series of humidity and temperature.

Analysing the association with meteorological factors might be helpful in understanding the behaviour of temperature and humidity. We can apply long time series for a period of 50 years and we can apply non-linear laplacian spectral analysis (NLSA) for time series with intermittency and low-frequency variability.

Appendix A

The optimal tuple for 14 day ahead forecast

Table A.1: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station K.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$14.685_{(L=20,r=5)}$	$14.584_{(L=20,r=4)}$	$14.017_{(L=30,r=5)}$
	180	$13.681_{(L=24,r=4)}$	$13.943_{(L=30,r=4)}$	$14.136_{(L=20,r=4)}$
SSA without projection	90	$15.001_{(L=30,r=4)}$	$14.901_{(L=20,r=4)}$	$14.025_{(L=30,r=4)}$
	180	$13.657_{(L=24,r=4)}$	$14.069_{(L=20,r=4)}$	$14.503_{(L=24,r=4)}$
	All	$12.055_{(L=360,r=10)}$	$12.054_{(L=360,r=8)}$	$12.039_{(L=360,r=8)}$
The lowest RMSE	$12.039_{(L=360,r=8)}$			

Table A.1 shows the RMSE for humidity at the station K. We observe that the lowest RMSE is 12.039 for SSA-V forecasting without projection and it is attained at $L = 360$ and $r = 8$ when using a period of all-time series. We notice that SSA-V forecasts have a slightly smaller RMSE than SSA-R and SSA-R (original)

forecasts by 0.13% and 0.12% respectively. Overall, the RMSE is dependent on the L and r parameters.

Table A.2: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station K.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$2.572_{(L=20,r=5)}$	$2.550_{(L=30,r=4)}$	$2.560_{(L=20,r=5)}$
	180	$2.555_{(L=24,r=5)}$	$2.603_{(L=16,r=4)}$	$2.786_{(L=24,r=5)}$
SSA without projection	90	$2.630_{(L=30,r=4)}$	$2.684_{(L=20,r=4)}$	$2.610_{(L=20,r=4)}$
	180	$2.608_{(L=24,r=4)}$	$2.669_{(L=20,r=4)}$	$2.825_{(L=24,r=4)}$
	All	$1.956_{(L=360,r=8)}$	$1.925_{(L=360,r=8)}$	$1.923_{(L=360,r=8)}$
The lowest RMSE	$1.923_{(L=360,r=8)}$			

Table A.2 presents the RMSE for temperature at the station K. We observe that the lowest RMSE is 1.923 attained at $L = 360$ and $r = 8$ for SSA-V forecasting without projection. We can see that SSA-V forecasts have a negligibly smaller RMSE than SSA-R and SSA-R (original) forecasts by 0.1% and 1.68% respectively when $N=$ all time series. Overall, the RMSE is significantly smaller for $r = 8$ and $L = 360$ indicating that the signal rank is 8.

Table A.3: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station MA.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$10.007_{(L=16,r=4)}$	$9.903_{(L=30,r=4)}$	$9.694_{(L=30,r=4)}$
	180	$9.4491_{(L=16,r=4)}$	$9.5238_{(L=16,r=4)}$	$10.147_{(L=24,r=9)}$
SSA without projection	90	$11.022_{(L=20,r=5)}$	$11.609_{(L=20,r=5)}$	$11.784_{(L=20,r=7)}$
	180	$9.857_{(L=16,r=6)}$	$10.166_{(L=16,r=4)}$	$10.140_{(L=24,r=8)}$
	All	$8.166_{(L=360,r=8)}$	$8.210_{(L=360,r=8)}$	$8.180_{(L=360,r=6)}$
The lowest RMSE	$8.166_{(L=360,r=8)}$			

Table A.3 exhibits the RMSE for humidity at the station MA. We note that the lowest RMSE is 8.166 attained at $L = 360$ and $r = 8$ for SSA-R forecasting without projection. We observe that SSA-R forecasts have a slightly smaller RMSE than SSA-V and SSA-R (original) forecasts by 0.17% and 0.53% respectively when $N=$ all time series. In general, the RMSE is significantly smaller for $r = 8$ and L is larger for using all the period series compared to using the other times.

Table A.4: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station MA.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$11.535_{(L=20,r=5)}$	$1.565_{(L=20,r=5)}$	$1.537_{(L=20,r=5)}$
	180	$1.420_{(L=16,r=7)}$	$1.468_{(L=16,r=7)}$	$1.463_{(L=16,r=7)}$
SSA without projection	90	$1.561_{(L=20,r=4)}$	$1.591_{(L=20,r=4)}$	$1.552_{(L=20,r=4)}$
	180	$1.425_{(L=16,r=6)}$	$1.471_{(L=16,r=6)}$	$1.466_{(L=16,r=6)}$
	All	$1.069_{(L=360,r=8)}$	$1.044_{(L=360,r=6)}$	$1.043_{(L=360,r=6)}$
The lowest RMSE	$1.043_{(L=360,r=6)}$			

In Table A.4 presents the RMSE for temperature at the station MA. We observe that the lowest RMSE is 1.043 attained at $L = 360$ and $r = 6$ for SSA-V forecasting without projection. We observe that SSA-R forecasts have a marginally smaller the RMSE than SSA-V and SSA-R (original) forecasts by 2.43% and 0.09% respectively when $N=$ all time series. We see that the RMSE is quite small for $r = 6$ and $L = 360$ for SSA-V and SSA-R (original). Note that temperature at the station MA has two mode shapes.

Table A.5: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station TH.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$16.442_{(L=30,r=5)}$	$16.100_{(L=30,r=5)}$	$15.223_{(L=30,r=5)}$
	180	$14.416_{(L=20,r=5)}$	$14.540_{(L=20,r=4)}$	$14.724_{(L=20,r=4)}$
SSA without projection	90	$16.423_{(L=30,r=4)}$	$15.984_{(L=30,r=4)}$	$15.231_{(L=30,r=4)}$
	180	$14.551_{(L=20,r=4)}$	$14.960_{(L=20,r=4)}$	$14.872_{(L=20,r=4)}$
	All	$13.237_{(L=360,r=6)}$	$13.248_{(L=360,r=4)}$	$13.215_{(L=360,r=6)}$
The lowest RMSE	$13.215_{(L=360,r=6)}$			

Table A.5 exhibits the RMSE for humidity at the station TH. We observe that the lowest RMSE is 13.215 attained at $L = 360$ and $r = 6$ for SSA-V forecasting without projection. Note that SSA-V forecasts have a negligibly smaller RMSE than SSA-V and SSA-R (original) forecasts by 0.17% and 0.25% respectively when $N=$ all time series. Overall, the RMSE is small for $r = 6$ and $L = 360$ for SSA-V.

Table A.6: The RMSE of $1, 2, \ldots, 14$-day ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station TH.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	90	$2.790_{(L=16,r=7)}$	$2.758_{(L=30,r=4)}$	$2.924_{(L=20,r=5)}$
	180	$2.357_{(L=20,r=9)}$	$2.363_{(L=20,r=9)}$	$2.454_{(L=16,r=11)}$
SSA without projection	90	$2.824_{(L=30,r=4)}$	$2.815_{(L=30,r=4)}$	$2.930_{(L=30,r=4)}$
	180	$2.365_{(L=20,r=8)}$	$2.367_{(L=20,r=8)}$	$2.486_{(L=20,r=8)}$
	All	$2.114_{(L=360,r=6)}$	$2.146_{(L=360,r=4)}$	$2.125_{(L=360,r=6)}$
The lowest RMSE	$2.114_{(L=360,r=6)}$			

Table A.6 shows the RMSE for temperature at the station TH. We can see that the lowest RMSE is 2.1144 attained at $L = 360$ and $r = 6$ for SSA-R forecasting without projection. We observe that SSA-R forecasts have a marginally smaller RMSE than SSA-V and SSA-R (original) forecasts by 0.5% and 1.49% respectively when $N=$ all time series.

Appendix B

The optimal tuple for 3 days ahead forecast

Table B.1: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station K.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$14.792_{(L=3,r=2)}$	$14.623_{(L=3,r=2)}$	$14.899_{(L=3,r=2)}$
	30	$12.806_{(L=5,r=2)}$	$12.739_{(L=7,r=2)}$	$12.687_{(L=6,r=2)}$
	90	$13.178_{(L=16,r=4)}$	$12.505_{(L=30,r=4)}$	$12.769_{(L=30,r=4)}$
	180	$12.769_{(L=16,r=8)}$	$12.463_{(L=30,r=4)}$	$12.809_{(L=30,r=4)}$
SSA without projection	10	$14.885_{(L=3,r=1)}$	$14.796_{(L=3,r=1)}$	$14.982_{(L=3,r=1)}$
	30	$14.192_{(L=11,r=2)}$	$13.796_{(L=9,r=2)}$	$13.483_{(L=12,r=2)}$
	90	$13.369_{(L=24,r=4)}$	$12.818_{(L=30,r=4)}$	$12.993_{(L=24,r=4)}$
	180	$12.884_{(L=16,r=6)}$	$12.863_{(L=24,r=4)}$	$13.297_{(L=24,r=4)}$
The lowest RMSE	$12.463_{(L=30,r=4)}$			

Table B.1 shows the RMSE for humidity at the station K. We can see that the lowest RMSE is 12.463 for SSA-R (original) forecasting with projection and is attained at $L = 30$ and $r = 4$ when $N = 180$ days. We notice that SSA-R (original) forecasts have a slightly smaller RMSE than SSA-R and SSA-V forecasts. Overall, the RMSE is dependent on L and r.

Table B.2: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station K.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$2.299_{(L=3,r=2)}$	$2.242_{(L=3,r=2)}$	$2.305_{(L=3,r=2)}$
	30	$2.049_{(L=3,r=2)}$	$2.022_{(L=3,r=2)}$	$2.043_{(L=3,r=2)}$
	90	$2.163_{(L=16,r=6)}$	$2.069_{(L=20,r=4)}$	$2.199_{(L=30,r=4)}$
	180	$2.156_{(L=16,r=4)}$	$2.106_{(L=20,r=4)}$	$2.157_{(L=16,r=4)}$
SSA without projection	10	$2.035_{(L=3,r=1)}$	$2.253_{(L=3,r=1)}$	$2.314_{(L=3,r=1)}$
	30	$2.330_{(L=9,r=2)}$	$2.205_{(L=9,r=2)}$	$2.286_{(L=9,r=2)}$
	90	$2.178_{(L=16,r=6)}$	$2.084_{(L=24,r=4)}$	$2.233_{(L=24,r=4)}$
	180	$2.199_{(L=20,r=6)}$	$2.107_{(L=20,r=4)}$	$2.198_{(L=20,r=4)}$
The lowest RMSE	$2.022_{(L=3,r=2)}$			

Table B.2 displays the RMSE for temperature at the station K. We observe that the lowest RMSE is 2.022 attained at $L = 3$ and $r = 2$ for SSA-R (original) forecasting with projection. We can see that SSA-R (original) forecasts have a negligibly smaller RMSE than SSA-R and SSA-V forecasts at $N = 30$ days series. Overall, the RMSE is quite small indicating that the signal rank is 2.

Table B.3: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station MA.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$10.331_{(L=3,r=2)}$	$9.627_{(L=3,r=2)}$	$10.268_{(L=3,r=2)}$
	30	$8.717_{(L=3,r=2)}$	$8.252_{(L=3,r=2)}$	$8.627_{(L=3,r=2)}$
	90	$9.544_{(L=16,r=4)}$	$9.129_{(L=20,r=4)}$	$9.462_{(L=16,r=4)}$
	180	$9.044_{(L=20,r=6)}$	$8.665_{(L=24,r=4)}$	$9.286_{(L=16,r=4)}$
SSA without projection	10	$10.384_{(L=3,r=1)}$	$9.686_{(L=3,r=1)}$	$10.321_{(L=3,r=1)}$
	30	$8.697_{(L=3,r=1)}$	$8.253_{(L=3,r=1)}$	$8.609_{(L=3,r=1)}$
	90	$10.371_{(L=16,r=4)}$	$9.626_{(L=16,r=4)}$	$10.119_{(L=16,r=4)}$
	180	$8.619_{(L=24,r=6)}$	$8.490_{(L=16,r=4)}$	$8.848_{(L=24,r=6)}$
The lowest RMSE	$8.252_{(L=3,r=2)}$			

Table B.3 represents the RMSE for humidity at the station MA. We note that the lowest RMSE is 8.252 attained at $L = 3$ and $r = 2$ for SSA-R (original) forecasting with projection. We observe that SSA-R forecasts have a slightly smaller RMSE than SSA-R and SSA-V forecasts at $N = 30$ days. In general, the RMSE is smaller for $r = 3$ and $L = 3$.

Table B.4: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station MA.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$1.232_{(L=3,r=2)}$	$1.166_{(L=3,r=2)}$	$1.228_{(L=3,r=2)}$
	30	$1.077_{(L=3,r=2)}$	$1.032_{(L=3,r=2)}$	$1.071_{(L=3,r=2)}$
	90	$1.223_{(L=16,r=4)}$	$1.149_{(L=16,r=4)}$	$1.184_{(L=16,r=4)}$
	180	$1.171_{(L=16,r=4)}$	$1.083_{(L=16,r=4)}$	$1.153_{(L=16,r=4)}$
SSA without projection	10	$1.232_{(L=3,r=1)}$	$1.196_{(L=3,r=1)}$	$1.229_{(L=3,r=1)}$
	30	$1.077_{(L=3,r=1)}$	$1.033_{(L=3,r=1)}$	$1.071_{(L=3,r=1)}$
	90	$1.266_{(L=24,r=4)}$	$1.193_{(L=24,r=4)}$	$1.182_{(L=24,r=4)}$
	180	$1.218_{(L=16,r=4)}$	$1.156_{(L=30,r=4)}$	$1.154_{(L=20,r=4)}$
The lowest RMSE	$1.032_{(L=3,r=2)}$			

Table B.4 displays the RMSE for temperature at the station MA. We observe that the lowest RMSE is 1.032 attained at $L = 3$ and $r = 2$ for SSA-R (original) forecasting with projection at $N = 30$ days. We observe that the RMSE is fairly small for $r = 2$ and $L = 3$ by contrast with other methods of forecasting.

Table B.5: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for humidity at the station TH.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$17.809_{(L=3,r=2)}$	$16.838_{(L=3,r=2)}$	$17.829_{(L=3,r=2)}$
	30	$14.736_{(L=4,r=2)}$	$14.518_{(L=3,r=2)}$	$14.644_{(L=4,r=2)}$
	90	$14.825_{(L=24,r=4)}$	$14.286_{(L=24,r=4)}$	$14.301_{(L=24,r=4)}$
	180	$14.361_{(L=30,r=6)}$	$13.838_{(L=24,r=4)}$	$13.929_{(L=24,r=4)}$
SSA without projection	10	$18.210_{(L=3,r=1)}$	$17.155_{(L=3,r=1)}$	$18.473_{(L=3,r=1)}$
	30	$14.757_{(L=3,r=1)}$	$14.482_{(L=3,r=1)}$	$14.754_{(L=4,r=1)}$
	90	$15.524_{(L=16,r=4)}$	$15.017_{(L=30,r=4)}$	$15.285_{(L=30,r=4)}$
	180	$14.101_{(L=16,r=4)}$	$13.949_{(L=30,r=4)}$	$14.190_{(L=30,r=4)}$
The lowest RMSE	$13.838_{(L=24,r=4)}$			

Table B.5 presents the RMSE for humidity at the station TH. We can observe that the lowest RMSE is 13.838 attained at $L = 24$ and $r = 4$ for SSA-R (original) forecasting with projection. It is worth noting that SSA-V forecasts have a negligibly smaller RMSE than SSA-V forecasts at $N = 180$. Overall, the RMSE is small for SSA-R (original) when $r = 4$ and $L = 24$.

Table B.6: The RMSE of $1, 2, 3$ days ahead forecasts using SSA-R, SSA-R (original) and SSA-V forecasting algorithms by applying SSA with double projection and SSA without projection for temperature at the station TH.

	N	SSA-R	SSA-R (original)	SSA-V
SSA with double projection	10	$2.372_{(L=3,r=2)}$	$2.237_{(L=3,r=2)}$	$2.370_{(L=3,r=2)}$
	30	$1.989_{(L=3,r=2)}$	$1.924_{(L=3,r=2)}$	$1.983_{(L=4,r=2)}$
	90	$2.226_{(L=16,r=4)}$	$2.115_{(L=16,r=4)}$	$2.259_{(L=16,r=4)}$
	180	$2.112_{(L=16,r=8)}$	$2.049_{(L=16,r=4)}$	$2.178_{(L=24,r=4)}$
SSA without projection	10	$2.399_{(L=3,r=1)}$	$2.262_{(L=3,r=1)}$	$2.387_{(L=3,r=1)}$
	30	$1.990_{(L=3,r=1)}$	$1.926_{(L=3,r=1)}$	$1.981_{(L=4,r=1)}$
	90	$2.164_{(L=16,r=4)}$	$2.109_{(L=24,r=4)}$	$2.221_{(L=16,r=4)}$
	180	$2.125_{(L=16,r=4)}$	$2.101_{(L=30,r=4)}$	$2.182_{(L=24,r=6)}$
The lowest RMSE	$1.924_{(L=3,r=2)}$			

Table B.6 shows the RMSE for temperature at the station TH. We can see that the lowest RMSE is 1.924 attained at $L = 3$ and $r = 2$ for SSA-R (original) forecasting with projection at $N = 30$.

Bibliography

[1] Afrifa, E. Y., Mueller, U. A., Taylor, S., and Fisher, A. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, 27(1):e1873.

[2] Al Balasmeh, O., Babbar, R., and Karmaker, T. (2019). Trend analysis and ARIMA modeling for forecasting precipitation pattern in wadi shueib catchment area in jordan. *Arabian Journal of Geosciences*, 12(2):27.

[3] Al Charabi, Y. and Al-Yahyai, S. (2013). Projection of future changes in rainfall and temperature patterns in Oman. *J. Earth Sci. Clim. Change*, 4(5):1–8.

[4] Al-Kalbani, M. S., John, C., Martin, F., et al. (2015). Recent trends in temperature and precipitation in al jabal al akhdar, sultanate of Oman, and the implications for future climate change. *Journal of Earth Science & Climatic Change*, 6(8):1.

[5] AL Marhoobi, S. and Pepelyshev, A. (2021). Analysis of temperature and humidity in Oman using singular spectrum analysis. *Communications in Statistics-Simulation and Computation*, pages 1–14.

[6] Ali, R., Kuriqi, A., Abubaker, S., and Kisi, O. (2019). Long-term trends and seasonality detection of the observed flow in Yangtze River using Mann-Kendall and Sen's innovative trend method. *Water*, 11(9):1855.

[7] Allen, M. R. and Smith, L. A. (1996). Monte carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of climate*, 9(12):3373–3404.

[8] Álvarez-Díaz, M. and Rosselló-Nadal, J. (2010). Forecasting british tourist arrivals in the balearic islands using meteorological variables. *Tourism Economics*, 16(1):153–168.

[9] Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., and Chau, K.-W. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water*, 12(5):1500.

[10] Ataei, M., Lohmann, B., Khaki-Sedigh, A., and Lucas, C. (2004). Model based method for estimating an attractor dimension from uni/multivariate chaotic time series with application to bremen climatic dynamics. *Chaos, Solitons & Fractals*, 19(5):1131–1139.

[11] Babu, C. N. and Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. *Applied Soft Computing*, 23:27–38.

[12] Beckers, J.-M. and Rixen, M. (2003). Eof calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and oceanic technology*, 20(12):1839–1856.

[13] Bokde, N., Beck, M. W., Álvarez, F. M., and Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern recognition letters*, 116:88–96.

[14] Brockwell, P. J. and Davis, R. A. (2009). Time series: Theory and methods, (springer series in statistics).

[15] Broomhead, D. S. and King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236.

[16] Buerkert, A., Fernandez, E., Tietjen, B., and Luedeling, E. (2020). Revisiting climate change effects on winter chill in mountain oases of northern Oman. *Climatic Change*, 162(3):1399–1417.

[17] Bulhoes, J. S., Assis, A. O., Martins, C. L., Furriel, G. P., Silva, B. C., Rodrigues, L., Reis, M. R., Calheiros, D. F., Oliveira, M. D., and Calixto, W. P. (2017). Gap filling in time series: A new methodology applying spectral analysis and system identification. In *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, pages 1–7. IEEE.

[18] Campbell, S. D. and Diebold, F. X. (2005). Weather forecasting for weather derivatives. *Journal of the American Statistical Association*, 100(469):6–16.

[19] Chhabra, G., Vashisht, V., and Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*, 10(19):1–7.

[20] Chong, A., Lam, K. P., Xu, W., Karaguzel, O. T., and Mo, Y. (2016). Imputation of missing values in building sensor data. *ASHRAE and IBPSA-USA SimBuild*, pages 407–414.

[21] Cui, L., Wang, L., Lai, Z., Tian, Q., Liu, W., and Li, J. (2017). Innovative trend analysis of annual and seasonal air temperature and rainfall in the Yangtze River Basin, China during 1960–2015. *Journal of Atmospheric and Solar-Terrestrial Physics*, 164:48–59.

[22] Curley, C., Krause, R. M., Feiock, R., and Hawkins, C. V. (2019). Dealing with missing data: A comparative exploration of approaches using the integrated city sustainability database. *Urban affairs review*, 55(2):591–615.

[23] Da Silva, R. M., Santos, C. A., Moreira, M., Corte-Real, J., Silva, V. C., and Medeiros, I. C. (2015). Rainfall and river flow trends using Mann–Kendall and Sen's slope estimator statistical tests in the cobres river basin. *Natural Hazards*, 77(2):1205–1221.

[24] Danilov, D. and Zhigljavsky, A. (1997). Principal components of time series: the 'caterpillar'method. *St. Petersburg: University of St. Petersburg*, pages 1–307.

[25] Deshpande, P., Rasin, A., Tchoua, R., Furst, J., Raicu, D., and Antani, S. (2020). Enhancing recall using data cleaning for biomedical big data. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 265–270. IEEE.

[26] Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

[27] Elfeky, M. G., Aref, W. G., and Elmagarmid, A. K. (2005). Warp: time warping for periodicity detection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.

[28] Fu, G., Chen, S., Liu, C., and Shepard, D. (2004). Hydro-climatic trends of the Yellow River basin for the last 50 years. *Climatic Change*, 65(1-2):149–178.

[29] Gedefaw, M., Yan, D., Wang, H., Qin, T., Girma, A., Abiyu, A., and Batsuren, D. (2018). Innovative trend analysis of annual and seasonal rainfall variability in amhara regional state, ethiopia. *Atmosphere*, 9(9):326.

[30] Ghanati, R., Hafizi, M. K., Mahmoudvand, R., and Fallahsafari, M. (2016). Filtering and parameter estimation of surface-nmr data using singular spectrum analysis. *Journal of Applied Geophysics*, 130:118–130.

[31] Ghil, M., Allen, M., Dettinger, M., Ide, K., Kondrashov, D., Mann, M., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., et al. (2002). Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1):3–1.

[32] Ghil, M. and Taricco, C. (1997). Advanced spectral-analysis methods. In *Past and present variability of the solar-terrestrial system: measurement, data analysis and theoretical models*, pages 137–159. IOS Press.

[33] Ghodsi, M., Hassani, H., Rahmani, D., and Silva, E. S. (2018). Vector and recurrent singular spectrum analysis: which is better at forecasting? *Journal of Applied Statistics*, 45(10):1872–1899.

[34] Ghodsi, Z., Silva, E. S., and Hassani, H. (2015). Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. *Genomics, proteomics & bioinformatics*, 13(3):183–191.

[35] Giannakis, D. and Majda, A. J. (2012a). Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear laplacian spectral analysis. *Geophysical Research Letters*, 39(10).

[36] Giannakis, D. and Majda, A. J. (2012b). Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences*, 109(7):2222–2227.

[37] Giannakis, D. and Majda, A. J. (2013). Nonlinear laplacian spectral analysis: capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(3):180–194.

[38] Gilleland, E. and Roux, G. (2015). A new approach to testing forecast predictive accuracy. *Meteorological Applications*, 22(3):534–543.

[39] Golyandina, N. (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *arXiv preprint arXiv:1005.4374*.

[40] Golyandina, N. and Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*, 71:934–954.

[41] Golyandina, N., Korobeynikov, A., Shlemov, A., and Usevich, K. (2013). Multivariate and 2d extensions of singular spectrum analysis with the Rssa package. *arXiv preprint arXiv:1309.5050*.

[42] Golyandina, N., Korobeynikov, A., and Zhigljavsky, A. (2018). *Singular spectrum analysis with R*. Springer.

[43] Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. CRC press.

[44] Golyandina, N. and Osipov, E. (2007). The "caterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical planning and Inference*, 137(8):2642–2653.

[45] Golyandina, N. and Shlemov, A. (2013). Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series. *arXiv preprint arXiv:1308.4022*.

[46] Golyandina, N. and Shlemov, A. (2017). Semi-nonparametric singular spectrum analysis with projection. *arXiv preprint arXiv:1507.05286*.

[47] Golyandina, N. and Stepanov, D. (2005). SSA-based approaches to analysis and forecast of multidimensional time series. In *proceedings of the 5th St. Petersburg workshop on simulation*, volume 293, page 298. St. Petersburg State University St. Petersburg, Russia.

[48] Golyandina, N. and Zhigljavsky, A. (2020). Basic SSA. In *Singular Spectrum Analysis for Time Series*, pages 21–90. Springer.

[49] Güçlü, Y. S. (2020). Improved visualization for trend analysis by comparing with classical Mann-Kendall test and ITA. *Journal of Hydrology*, 584:124674.

[50] Gustin, M., McLeod, R. S., and Lomas, K. J. (2018). Forecasting indoor temperatures during heatwaves using time series models. *Building and Environment*, 143:727–739.

[51] Hammer, G. L., Cooper, M., and Reynolds, M. P. (2021). Plant production in water-limited environments.

[52] Harmouche, J., Fourer, D., Auger, F., Borgnat, P., and Flandrin, P. (2017). The sliding singular spectrum analysis: A data-driven nonstationary signal decomposition tool. *IEEE Transactions on Signal Processing*, 66(1):251–263.

[53] Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

[54] Hassani, H. (2010a). A brief introduction to singular spectrum analysis. *Paper in pdf version available at: www. ssa. cf. ac. uk/a_brief_introduction_to_ssa. pdf.*

[55] Hassani, H. (2010b). A brief introduction to singular spectrum analysis. *Optimal decisions in statistics and data analysis.*

[56] Hassani, H., Heravi, S., and Zhigljavsky, A. (2009). Forecasting european industrial production with singular spectrum analysis. *International journal of forecasting*, 25(1):103–118.

[57] Hassani, H., Heravi, S., and Zhigljavsky, A. (2013a). Forecasting uk industrial production with multivariate singular spectrum analysis. *Journal of Forecasting*, 32(5):395–408.

[58] Hassani, H., Kalantari, M., and Yarmohammadi, M. (2017a). An improved SSA forecasting result based on a filtered recurrent forecasting algorithm. *Comptes Rendus Mathematique*, 355(9):1026–1036.

[59] Hassani, H. and Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, 1(01):55–83.

[60] Hassani, H. and Mahmoudvand, R. (2018). *Singular Spectrum Analysis: Using R*. Springer.

[61] Hassani, H., Mahmoudvand, R., Omer, H. N., and Silva, E. S. (2014). A preliminary investigation into the effect of outlier (s) on singular spectrum analysis. *Fluctuation and Noise Letters*, 13(04):1450029.

[62] Hassani, H., Mahmoudvand, R., and Zokaei, M. (2011). Separability and window length in singular spectrum analysis. *Comptes rendus mathematique*, 349(17-18):987–990.

[63] Hassani, H., Mahmoudvand, R., Zokaei, M., and Ghodsi, M. (2012). On the separability between signal and noise in singular spectrum analysis. *Fluctuation and Noise Letters*, 11(02):1250014.

[64] Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., and Gupta, R. (2017b). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, 63:112–127.

[65] Hassani, H., Silva, E. S., Gupta, R., and Das, S. (2018). Predicting global temperature anomaly: A definitive investigation using an ensemble of twelve competing forecasting models. *Physica A: Statistical Mechanics and its Applications*, 509:121–139.

[66] Hassani, H., Silva, E. S., Gupta, R., and Segnon, M. K. (2015). Forecasting the price of gold. *Applied Economics*, 47(39):4141–4152.

[67] Hassani, H., Soofi, A. S., and Zhigljavsky, A. (2013b). Predicting inflation dynamics with singular spectrum analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3):743–760.

[68] Hassani, H. and Thomakos, D. (2010). A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, 3(3):377–397.

[69] Hoffman, A. L., Kemanian, A. R., and Forest, C. E. (2018). Analysis of climate signals in the crop yield record of sub-saharan africa. *Global change biology*, 24(1):143–157.

[70] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[71] Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of statistical software*, 27(1):1–22.

[72] Hyndman, R. J., Khandakar, Y., et al. (2007). *Automatic time series for forecasting: the forecast package for R*. Number 6/07. Monash University, Department of Econometrics and Business Statistics.

[73] Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–454.

[74] Iqelan, B. M. (2017). A singular spectrum analysis technique to electricity consumption forecasting. *Int. Journal of Engineering Research and Application*, 9.

[75] Ise, T. and Oba, Y. (2019). Forecasting climatic trends using neural networks: An experimental study using global historical data. *Frontiers in Robotics and AI*, 6:32.

[76] Jain, G. and Mallick, B. (2017). A study of time series models ARIMA and ETS. *Available at SSRN 2898968*.

[77] Kalantari, M., Yarmohammadi, M., Hassani, H., and Silva, E. S. (2018). Time series imputation via Norm-based singular spectrum analysis. *Fluctuation and Noise Letters*, 17(02):1850017.

[78] Khan, M. A. R. and Poskitt, D. S. (2013). A note on window length selection in singular spectrum analysis. *Australian & New Zealand Journal of Statistics*, 55(2):87–108.

[79] Kisi, O. and Ay, M. (2016). Reply to the comments on "comparison of Mann–Kendall and innovative trend method for water quality parameters of the Kizilirmak River, Turkey" by Kisi, o. and ay, m.[j. hydrol. 513 (2014) 362–375] and "an innovative method for trend analysis of monthly pan evaporations" by kisi, o.[j. hydrol. 527 (2015) 1123–1129]. *Journal of Hydrology*, 538:883–884.

[80] Kisi, O., Santos, G., Augusto, C., Marques, D., Richarde, and Zounemat Kermani, M. (2018). Trend analysis of monthly streamflows using Şen's innovative trend method. *Geofizika*, 35(1):53–68.

[81] Kondrashov, D., Denton, R., Shprits, Y., and Singer, H. (2014). Reconstruction of gaps in the past history of solar wind parameters. *Geophysical Research Letters*, 41(8):2702–2707.

[82] Kondrashov, D. and Ghil, M. (2006). Spatio-temporal filling of missing points in geophysical data sets.

[83] Korobeynikov, A., Shlemov, A., Usevich, K., and Golyandina, N. (2015). RSSA:a collection of methods for singular spectrum analysis.

[84] Lipton, Z. C., Kale, D. C., and Wetzel, R. (2016). Modeling missing data in clinical time series with RNNS. *arXiv preprint arXiv:1606.04130*.

[85] Liu, Q., Yang, Z., and Cui, B. (2008). Spatial and temporal variability of annual precipitation during 1961–2006 in Yellow River Basin, China. *Journal of hydrology*, 361(3-4):330–338.

[86] Luo, Y., Cai, X., Zhang, Y., Xu, J., et al. (2018). Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1596–1607.

[87] Mahmoudvand, R., Najari, N., and Zokaei, M. (2013). On the optimal parameters for reconstruction and forecasting in singular spectrum analysis. *Communications in Statistics-Simulation and Computation*, 42(4):860–870.

[88] Mahmoudvand, R. and Rodrigues, P. C. (2016). Missing value imputation in time series using singular spectrum analysis. *International Journal of Energy and Statistics*, 4(01):1650005.

[89] Mahmoudvand, R., Rodrigues, P. C., and Yarmohammadi, M. (2019). Forecasting daily exchange rates: A comparison between SSA and MSSA. *REVSTAT*, 17(4):599—616.

[90] Mahmoudvand, R. and Zokaei, M. (2012). On the singular values of the hankel matrix with application in singular spectrum analysis. *Chilean Journal of Statistics*, 3(1):43–56.

[91] Marques, C., Ferreira, J., Rocha, A., Castanheira, J., Melo-Gonçalves, P., Vaz, N., and Dias, J. (2006). Singular spectrum analysis and forecasting of

hydrological time series. *Physics and Chemistry of the Earth, Parts A/B/C*, 31(18):1172–1179.

[92] Mi, X., Liu, H., and Li, Y. (2019). Wind speed prediction model using singular spectrum analysis, empirical mode decomposition and convolutional support vector machine. *Energy Conversion and Management*, 180:196–205.

[93] Modarres, R. and Sarhadi, A. (2009). Rainfall trends analysis of Iran in the last half of the twentieth century. *Journal of Geophysical Research: Atmospheres*, 114(D3).

[94] Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

[95] Mostafa, S. M. (2019). Imputing missing values using cumulative linear regression. *CAAI Transactions on Intelligence Technology*, 4(3):182–200.

[96] Murat, M., Malinowska, I., Gos, M., and Krzyszczak, J. (2018). Forecasting daily meteorological time series using arima and regression models. *International agrophysics*, 32(2).

[97] Nadtoka, I. and Vyalkova, S. (2019). Hybrid speculation model of energy consumption based on multivariate singular spectrum analysis and neural fuzzy network. In *2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pages 1–5. IEEE.

[98] Panigrahi, S. and Behera, H. S. (2017). A hybrid ets–ann model for time series forecasting. *Engineering Applications of Artificial Intelligence*, 66:49–59.

[99] Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2018). Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica*, 66(4):807–831.

[100] Peña, M., Ortega, P., and Orellana, M. (2019). A novel imputation method for missing values in air pollutant time series data. In *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE.

[101] Peng, L. and Lei, L. (2005). A review of missing data treatment methods. *Intelligent Information Management Systems and Technologies*, 1(3):412–419.

[102] Pepelyshev, A. and Zhigljavsky, A. (2010). Assessing the stability of long-horizon SSA forecasting. *Statistics and its Interface*, 3(3):321–327.

[103] Pepelyshev, A. and Zhigljavsky, A. (2017). SSA analysis and forecasting of records for earth temperature and ice extents. *Statistics and Its Interface*, 10(1):151–163.

[104] Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, 237(1):152–163.

[105] Plaza, E. G. and López, P. N. (2017). Surface roughness monitoring by singular spectrum analysis of vibration signals. *Mechanical Systems and Signal Processing*, 84:516–530.

[106] Pratama, I., Permanasari, A. E., Ardiyanto, I., and Indrayani, R. (2016). A review of missing values handling methods on time series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. IEEE.

[107] Rekapalli, R. and Tiwari, R. (2015). A short note on the application of singular spectrum analysis for geophysical data processing. *J. Ind. Geophys. Union*, 19(1):77–85.

[108] Rezvan, P. H., Lee, K. J., and Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, 15(1):30.

[109] Rodrigues, P. C. and De Carvalho, M. (2013). Spectral modeling of time series with missing data. *Applied Mathematical Modelling*, 37(7):4676–4684.

[110] Rodrigues, P. C. and Mahmoudvand, R. (2018). The benefits of multivariate singular spectrum analysis over the univariate version. *Journal of the Franklin Institute*, 355(1):544–564.

[111] Rodrigues, P. C. and Mahmoudvand, R. (2020). A new approach for the vector forecast algorithm in singular spectrum analysis. *Communications in Statistics-Simulation and Computation*, 49(3):591–605.

[112] Sanei, S. and Hassani, H. (2015). *Singular spectrum analysis of biomedical signals*. CRC press.

[113] Schmidt, O. T., Mengaldo, G., Balsamo, G., and Wedi, N. P. (2019). Spectral empirical orthogonal function analysis of weather and climate data. *Monthly Weather Review*, 147(8):2979–2995.

[114] Schoellhamer, D. H. (2001). Singular spectrum analysis for time series with missing data. *Geophysical research letters*, 28(16):3187–3190.

[115] Şen, Z. (2012). Innovative trend analysis methodology. *Journal of Hydrologic Engineering*, 17(9):1042–1046.

[116] Şen, Z. (2017). Innovative trend significance test and applications. *Theoretical and applied climatology*, 127(3-4):939–947.

[117] Shadmani, M., Marofi, S., and Roknian, M. (2012). Trend analysis in reference evapotranspiration using Mann-Kendall and Spearman's Rho tests in arid regions of Iran. *Water resources management*, 26(1):211–224.

[118] Shi, J., Guo, J., and Zheng, S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews*, 16(5):3471–3480.

[119] Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

[120] Silva, A. E. D. S. (2016). *Theoretical advancements and applications in singular spectrum analysis.* PhD thesis, Bournemouth University.

[121] Silva, E., Hassani, H., et al. (2015). On the use of singular spectrum analysis for forecasting us trade before, during and after the 2008 recession. *International Economics*, 141:34–49.

[122] Silva, E. S., Hassani, H., Heravi, S., and Huang, X. (2019). Forecasting tourism demand with denoised neural networks. *Annals of Tourism Research*, 74:134–154.

[123] Silva, R., Santos, C., Silva, A., and Neto, R. (2020). Spatial distribution and estimation of rainfall trends and erosivity in the Epitácio Pessoa reservoir catchment, Paraíba, Brazil. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, pages 1–21.

[124] Sonali, P. and Kumar, D. N. (2013). Review of trend detection methods and their application to detect temperature changes in India. *Journal of Hydrology*, 476:212–227.

[125] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338.

[126] Taie Semiromi, M. and Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. *Hydrological Sciences Journal*, 64(14):1711–1726.

[127] Tardivo, G. and Berti, A. (2012). A dynamic method for gap filling in daily temperature datasets. *Journal of Applied Meteorology and Climatology*, 51(6):1079–1086.

[128] Teegavarapu, R. S., Tufail, M., and Ormsbee, L. (2009). Optimal functional forms for estimation of missing precipitation data. *Journal of hydrology*, 374(1-2):106–115.

[129] Templ, M., Alfons, A., Kowarik, A., and Prantner, B. (2019). Vim: Visualization and imputation of missing values, 2011a. *URL http://CRAN. R-project. org/package= VIM. R package version*, 3(0).

[130] Tonkaz, T., Çetin, M., and Tülücü, K. (2007). The impact of water resources development projects on water vapor pressure trends in a semi-arid region, Turkey. *Climatic change*, 82(1-2):195–209.

[131] Tran, Q. T., Hao, L., and Trinh, Q. K. (2019). A comprehensive research on exponential smoothing methods in modeling and forecasting cellular traffic. *Concurrency and Computation: Practice and Experience*, page e5602.

[132] Unnikrishnan, P. and Jothiprakash, V. (2015). Extraction of nonlinear rain-

fall trends using singular spectrum analysis. *Journal of Hydrologic Engineering*, 20(12):05015007.

[133] Unnikrishnan, P. and Jothiprakash, V. (2018). Daily rainfall forecasting for one year in a single run using singular spectrum analysis. *Journal of Hydrology*, 561:609–621.

[134] Unnikrishnan, P. and Jothiprakash, V. (2020). Hybrid SSA-ARIMA-ANN model for forecasting daily rainfall. *Water Resources Management*, 34(11):3609–3623.

[135] Ustoorikar, K. and Deo, M. (2008). Filling up gaps in wave data with genetic programming. *Marine Structures*, 21(2-3):177–195.

[136] van der Loo, M. (2017). Simputation: simple imputation. *R package version 0.2*, 2.

[137] Vautard, R., Yiou, P., and Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4):95–126.

[138] Viljoen, H. and Nel, D. (2010). Common singular spectrum analysis of several time series. *Journal of Statistical Planning and Inference*, 140(1):260–267.

[139] Vlachos, M., Yu, P., and Castelli, V. (2005). On periodicity detection and structural periodic similarity. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 449–460. SIAM.

[140] Wang, H. Z., Zhang, R., Liu, W., Wang, G. H., and Jin, B. G. (2008). Improved interpolation method based on singular spectrum analysis iteration and its application to missing data recovery. *Applied Mathematics and Mechanics*, 29(10):1351–1361.

[141] Wang, R., Ma, H., Liu, G., and Zuo, D. (2015). Selection of window length for singular spectrum analysis. *Journal of the Franklin Institute*, 352(4):1541–1560.

[142] Wang, W. and Lu, Y. (2018). Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In *IOP Conference Series: Materials Science and Engineering*, volume 324, page 012049.

[143] Wang, Z.-Y., Qiu, J., and Li, F.-F. (2018). Hybrid models combining EMD/EEMD and ARIMA for long-term streamflow forecasting. *Water*, 10(7):853.

[144] Wilks, D. S. (2011). Principal component (eof) analysis. In *International Geophysics*, volume 100, pages 519–562. Elsevier.

[145] Wu, Z., Rincon, D., and Christofides, P. D. (2020). Process structure-based recurrent neural network modeling for model predictive control of nonlinear processes. *Journal of Process Control*, 89:74–84.

[146] Yaseen, Z. M., Kisi, O., and Demir, V. (2016). Enhancing long-term streamflow forecasting and predicting using periodicity data component: application of artificial intelligence. *Water resources management*, 30(12):4125–4151.

[147] Yue, S., Pilon, P., and Cavadias, G. (2002). Power of the Mann–Kendall and Spearman's Rho tests for detecting monotonic trends in hydrological series. *Journal of hydrology*, 259(1-4):254–271.

[148] Zahroh, S., Hidayat, Y., Pontoh, R. S., Santoso, A., Sukono, F., and Bon, A. (2019). Modeling and forecasting daily temperature in bandung. In *Proceedings of the International Conference on Industrial Engineering and Operations Management Riyadh, Saudi Arabia*, pages 406–12.

[149] Zeleňáková, M., Jothıprakash, V., Purcz, P., Unnıkrıshnan, P., and Hlavatá, H. (2017). Investigation of precipitation trends in eastern Slovakia using singular spectrum analysis.

[150] Zhigljavsky, A. (2010). Singular spectrum analysis for time series: Introduction to this special issue. *Statistics and its Interface*, 3(3):255–258.

[151] Zhu, L., Wang, Y., and Fan, Q. (2014). MODWT-ARMA model for time series prediction. *Applied Mathematical Modelling*, 38(5-6):1859–1865.

[152] Zokaei, M., Mahmoudvand, R., and Najari, N. (2011). Comparison of singular spectrum analysis and ARIMA models.