"Science Stays True Here"
Journal of Mathematics and Statistical Science, 1-12 | Science Signpost Publishing

On Spatial Dependence in Multivariate Singular Spectrum Analysis

Richard Awichi

Abstract

In this paper, I present a method for utilizing the usually intrinsic spatial information in spatial data sets to improve the quality of temporal predictions within the framework of singular spectrum analysis (SSA) techniques. The SSA-based techniques constitute a model free approach to time series analysis and ordinarily, SSA can be applied to any time series with a notable structure. Indeed it has a wide area of application including social sciences, medical sciences, finance, environmental sciences, mathematics, dynamical systems and economics. SSA has two broad aims:

i) To make a decomposition of the original series into a sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and a structure-less noise.

ii) To reconstruct the decomposed series for further analysis in the absence of the noise component.

Multivariate singular spectrum analysis (MSSA) is an extension of SSA to multivariate statistics and takes advantage of the delay procedure to obtain a similar formulation as SSA though with larger matrices for multivariate data. In situations where spatial data is an important focus of investigation, it is not uncommon to have attributes whose values change with space and time and an accurate prediction is thus important. The usual question asked is whether the intrinsic location parameters in spatial data can improve data analysis of such data sets. The proposed method is based on the inverse distance technique and is exemplified on climate data from Upper Austria for the period Jan 1994 to Dec 2009.

Results show that the proposed technique of incorporating spatial dependence into MSSA analysis leads to improved quality of statistical inference.

Keywords: time series analysis, MSSA, inverse distance weighting, spatial dependence.

1. Introduction

Singular Spectrum Analysis (SSA), a well developed tool for time series analysis, is a model free approach to time series analysis, as opposed to model based time series analysis with several restrictive

assumptions, see for example [5]. The beginning of SSA is usually attributed to [6]. Ordinarily, SSA can be applied to any time series with a notable structure, see [11].

SSA has two broad aims:

i) To make a decomposition of the original series into a sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and a structureless noise.

ii) To reconstruct the decomposed series for further analysis in the absence of the noise component.

SSA is implemented through a sequence of steps and the following are the steps in brief, detailed exposition and background theory can be found in [11]. The first step is the *Embedding* step in which the time series $F_N = (f_1, \cdots, f_N)$ is transformed into a multidimensional data matrix **X**, called the trajectory matrix using an embedding operator: $\mathcal{T}(F) \longrightarrow \mathbf{X}$. The single most important parameter in this step is the window length L. The second step is the *Singular value decomposition* (SVD) step in which the trajectory matrix is factorized into a sum of elementary matrices using the nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$. The *Grouping* step is the third step where the elementary matrices are split further through the procedure known as *eigentriple grouping*. The final step is the *Diagonal averaging* or also commonly known as *Hankelization*. This step transfers the sum of the elementary matrices after eigentriple grouping back to the time series. It is in a way the reverse of step one.

Of importance in SSA analysis is the concept of *separability*. Separability entails how well the (additive) components of the time series can be separated from each other to allow further analysis to be meaningfully done. A time series may comprise trend (slowly varying component), periodic or quasi periodic components and noise. These may be generalized into signal and noise components. SSA decomposition of the series F_N can only be successful if the resulting additive components of the seriesare approximately separable from each other and this enhances the quality of the decomposition, see [11].

The notion of *weak* separability applies to orthogonality of the rows (and columns) of the matrices whereas *strong* separability imposes a further condition of distinct eigenvalues of the trajectory matrix.

Multivariate Singular Spectrum Analysis (MSSA) is a direct extension of SSA to multivariate analysis and takes advantage of the (delay) embedding procedure to obtain a similar formulation as SSA, albeit with larger matrices for multidimensional time series. It has previously been successfully applied to

the study of climate fields, see [19]. When geographical coordinates of the data gathering sites are included as part of the data, then we talk of spatial data. In this paper, we propose a technique of harnessing this location information to improve statistical inference of spatial data within the framework of MSSA.

Section 2 is devoted to reviewing the basics of MSSA. Section 3 contains the discussion on the inverse distance technique while in Section 4, we discuss the results and conclusions appear in Section 5.

2. Multivariate Singular Spectrum Analysis, MSSA

MSSA is an extension of the univariate SSA to the multidimensional time series. The main aim of MSSA is to extract signal from the multivariate time series leaving out the residual (noise) so as to perform further analysis, see [18], [13]. MSSA has been successfully applied to many different series: in Climatology [8], in Economics [18], [14] and in medical sciences [9] to mention just a few.

2.1. Stages of MSSA

MSSA, like SSA comprises two stages, namely: Decomposition and Reconstruction, each of which has two steps: Embedding and SVD for stage one and Grouping and Diagonal Averaging for stage two. Here we discuss mainly the embedding step which is the major difference from the SSA steps.

2.1.1. Embedding

Let $y_{N_i} = (y_{N_i}^{(1)}, \cdots, y_{N_i}^{(s)})$ be an s-variate time series, L_i the window length, $X^{(i)}$ the trajectory matrix of the one dimensional time series $\{y_{N_i}^{(i)}\}, (i = 1, \cdots, s)$.

As in [10], [18] and [14], the trajectory matrix X of the multivariate series is given as;

$$X = [X^{(1)} : \cdots : X^{(q)} : \cdots : X^{(s)}] \tag{1}$$

Each $X^{(i)}$ is given as;

$$X^{(i)} = \begin{pmatrix} y_1^{(i)} & y_2^{(i)} & \cdots & y_{K_i}^{(i)} \\ y_2^{(i)} & y_3^{(i)} & \cdots & y_{K_i+1}^{(i)} \\ y_3^{(i)} & y_4^{(i)} & \cdots & y_{K_i+2}^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ y_{L_i}^{(i)} & y_{L_i+1}^{(i)} & \cdots & y_{N_i}^{(i)} \end{pmatrix}$$

3. Inverse Distance Weighting, IDW

For spatial data sets, there is always the intrinsic (geographic) information–the location attribute embedded in every recording. Whether it is in environmental sciences, economics, agriculture, climatology, geology or any other fields where spatial data is frequently encountered, the desire to harness this embedded information for purposes of analysis cannot be over emphasized. Data mining is an automated search for knowledge hidden in large collections of data set attributes. In environmental science and other areas where space-time behaviour is an important focus of investigation, it is not uncommon to have attributes whose values often change with space and time. This leads to spatial dependence which subsequently influences data analysis, see [17] and therefore a technique to incorporate spatial information into the analysis of such data sets is desirable.

Data close together in space and time usually exhibit higher dependencies than those that are farther apart, see [7]. This dependency is thus inversely proportional to the distance of separation between any two data gathering sites. Amethod that utilizes this inverse proportionality in the distance of separation as a technique of incorporating spatial dependence into the analysis is the inverse distance weighting (IDW). Inverse distance weighting assigns bigger weights at near points and smaller weights at distant locations. There are several ways of computing the inverse distance weights to be used in the analysis, see [16]. Here, we present the inverse distance technique first introduced in [3] and developed further in [1].

The multivariate data set $y = \{y_{ij}\}$ is an $s \times N$ matrix and the inverse distance is given as $w_{ij} = 1/d_{ij}$ where d_{ij} is the Euclidean distance between locations i and j. For missing values in the data, a new weight is calculated by excluding the corresponding distance measure from the w_{ij} s. To include spatial information into the analysis based on the model free MSSA framework, we premultiply the data set by the row-normalized spatial weight matrix, $W = \{w_{ij}\}$ to yield the spatially weighted averages Wy.

4. Application

The proposed technique was applied to climate data from several recording sites in Upper Austria. The data was provided by the Zentralanstalt für Metereologie in Austria and is described in more detail in [16]. This data set contains climatic data measured at 37 stations irregularly placed over the region provided from http://www.zamg.ac.at/fix/klima/oe71-00/klima2000/klimadaten_oesterreich_1971_frame1.htm. Here, we have (incomplete) monthly data from Jan 1994 to Dec 2009 on average temperature and

total rainfall. Due to some missing observations, however, not all of the stations could be effectively used. A map of the region with the respective locations of the measurement stations and the contours for the rainfall data is displayed in Figure 1.

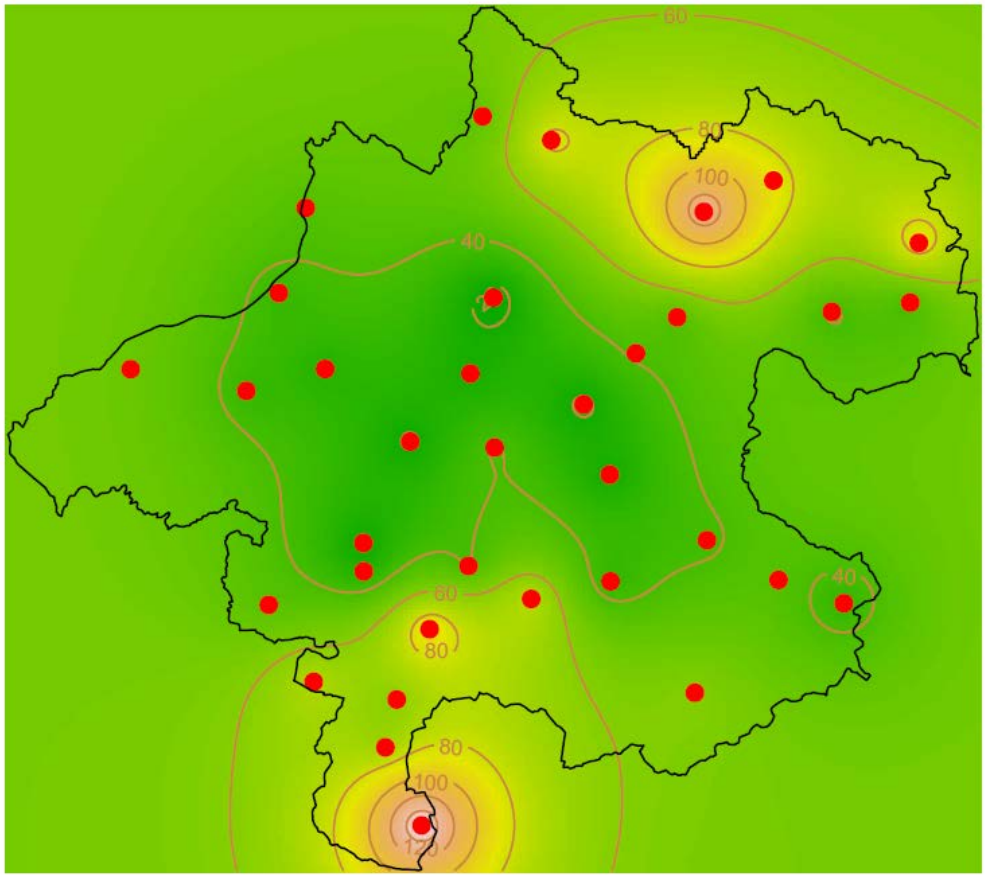

Figure 1. The Sampling Locations of the Climatic Data Set Within Upper Austria

For purposes of this application, we used the rainfall data from 11 locations that have no missing information, hence the series length is 192 (monthly recordings from Jan 1994 to Dec 2009). The rest of the sites have missing data to varying degrees of missingness. The data was also preprocessed by log-transformation, see [12].

To determine the effect of the spatial dependence on the data, we pooled together the data at different levels. The pooling was done by conditioning data from a particular site, y_i on data from the rest of the sites and likewise for the spatially weighted averages. To assess the accuracy, we calculated the root mean square errors, RMSE, i.e. $R(y_i \mid y)$ and $R(y_i \mid Wy)$. The RMSE $R(y_i)$ of the single unweighted series is referred to as the default RMSE in this paper. If $R(y_i \mid y) < R(y_i)$ (or $R(y_i \mid Wy) < R(y_i)$), then the proposed technique leads to improved quality of the results, otherwise it is worse than results without pooling. The analyses were done using [15]. For comparative purposes, we computed the ratio of

the root mean square error, which is given by $RRMSE = \frac{\sum RMSE}{\sum RMSE(y_i)}$. For more information on how to perform spatial analysis in R, see [4].

4.1. Results

The results are presented for both in-sample and out-of-sample analyses. We report findings for the unweighted and the spatially weighted for the set of entire sites. The other set of results fall between these two.

4.1.1. In-Sample

Figure 2 shows the time series graph of one of the sites, Freistadt which is typical of all the other sites. One of the basic capabilities of SSA-based techniques is shown in this graph, i.e. to separate the original series into its (additive) components.

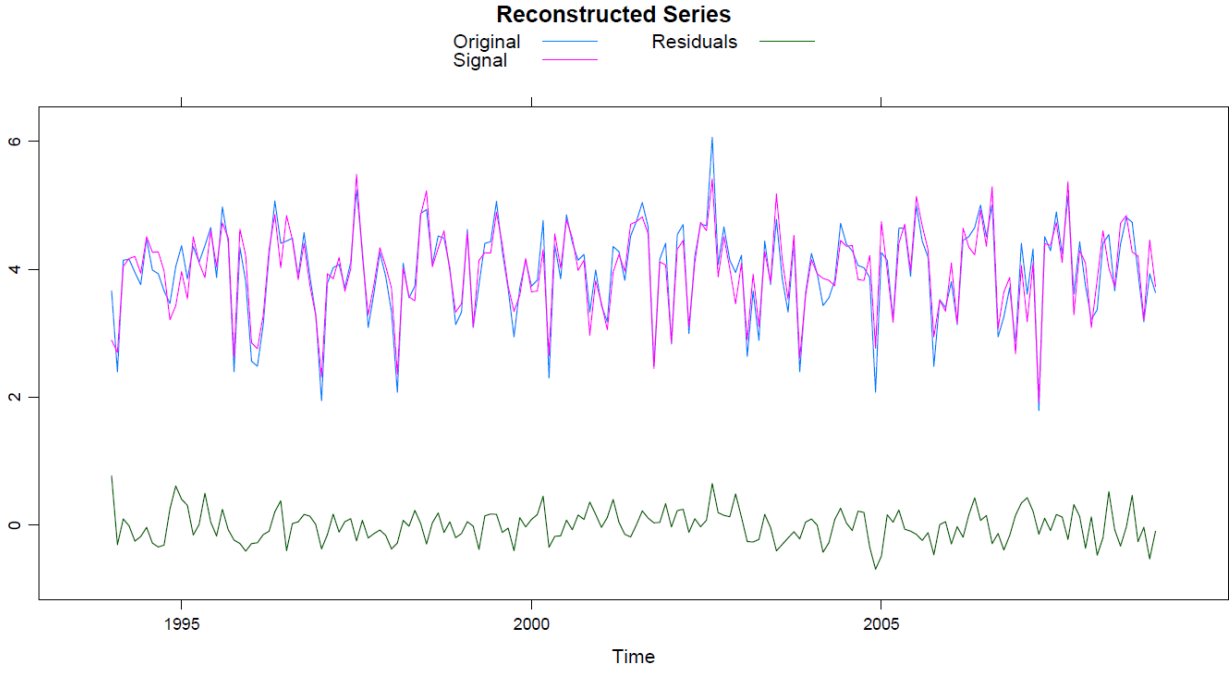

Figure 2. Time Series of one Selected Site

Table 1 shows the RMSE values corresponding to the different window lengths. For MSSA, $L \leq \frac{sN}{s+1}$, see [10]. A good choice of the window length ensures proper separability.

Table 1. RMSE Values, R($y_i \mid y$) for Different L

Site	Default RMSE L=96	RMSE Values corresponding to different Window Lengths				
		L=96	L=144	L=156	L=168	L=180
1	0.2661	0.2833	0.2197	0.2084	0.1786	0.1351
2	0.2466	0.2662	0.1865	0.1630	0.1514	0.1419
3	0.2503	0.2912	0.2352	0.2112	0.1803	0.1267
4	0.2413	0.3087	0.2298	0.1998	0.1825	0.1651
5	0.2394	0.3794	0.3709	0.2478	0.1863	0.0741
6	0.2349	0.2624	0.2008	0.1931	0.1765	0.1475
7	0.2065	0.2515	0.1722	0.1474	0.1456	0.1389
8	0.2083	0.2516	0.1834	0.1530	0.1263	0.1020
9	0.2017	0.2423	0.1821	0.1613	0.1281	0.1222
10	0.2037	0.2574	0.2149	0.1903	0.1626	0.1020
11	0.2010	0.2490	0.2053	0.1916	0.1484	0.0851

Due to pooling of the data, the RMSE values progressively decrease towards their minimum at the optimal window length.

4.1.2. Spatially Weighted Conditioning

Figure 3 shows the Reconstructed series graphs of three selected sites being compared with their weighted counterparts.

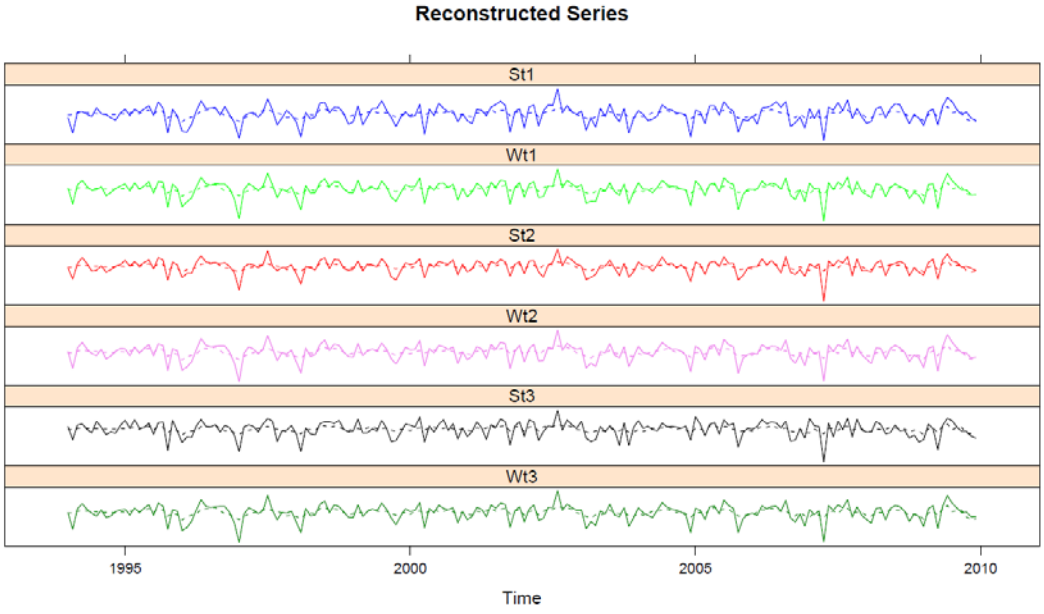

Figure 3. Reconstructed Series of Three Selected Sites Compared with their Weighted Counterparts

The graphs show a close similarity in the unweighted and the spatially weighted data sets implying that the 'temporal structure' of the series is not affected by the spatial weighting as the intra-seasonal variability remains in the spatially weighted series.

In Table 2 are the RMSE values for the spatially weighted pooling for the set of entire sites and in Table 3 we have computed the RRMSE as described earlier as a measure of the performance at different window lengths.

Table 2. RMSE Values, R($y_i \mid Wy$) for Different L: Spatially Weighted for all Sites

Site	Default RMSE	RMSE Values for different Window Lengths				
	L=96	L=96	L=144	L=156	L=168	L=180
1	0.2661	0.2841	0.2152	0.1778	0.0935	0.0062
2	0.2466	0.2606	0.1771	0.1521	0.0836	0.0091
3	0.2503	0.2838	0.2291	0.1783	0.0861	0.0062
4	0.2413	0.3071	0.2302	0.1578	0.0979	0.0046
5	0.2394	0.4137	0.3892	0.2440	0.0720	0.0031
6	0.2349	0.2565	0.1904	0.1587	0.0829	0.0081
7	0.2065	0.2450	0.1590	0.1281	0.0956	0.0081
8	0.2083	0.2403	0.1633	0.1194	0.0799	0.0094
9	0.2017	0.2373	0.1661	0.1198	0.0841	0.0100
10	0.2037	0.2416	0.1876	0.1483	0.0811	0.0112
11	0.2010	0.2365	0.1830	0.1417	0.0764	0.0153

Table 3. Differences in RMSE Values

Site	Default RMSE	Differences from default RMSE			
	L=96	R($y_i \mid y$)	R($y_i \mid Wy_i$)	R($y_i \mid Wy$)	R($y_i \mid Wy$) $_{(L=96)}$
1	0.266111	0.131060	0.233964	0.259907	-0.017994
2	0.246588	0.104707	0.219042	0.237449	-0.013963
3	0.250342	0.123672	0.220715	0.244135	-0.033471
4	0.241345	0.076241	0.210099	0.236794	-0.065711
5	0.239399	0.165341	0.210293	0.236306	-0.174275
6	0.234904	0.087389	0.202348	0.226827	-0.021546
7	0.206510	0.067623	0.184227	0.198417	-0.038519
8	0.208275	0.106287	0.185549	0.198860	-0.032012
9	0.201736	0.079528	0.178254	0.191761	-0.035604
10	0.203661	0.101666	0.179011	0.192459	-0.037911
11	0.201021	0.115954	0.177240	0.185754	-0.035474
		Ratios of RMSE (RRMSE)			
		0.536192	0.119664	0.036489	1.202602

Further details about in-sample prediction can be found in [2].

4.1.3. Out-of-Sample

For the out-of-sample analysis, we used the data up to Dec 2008 implying that the window length reduced to 180. The data for the final year, Jan-Dec 2009, was used to compare with the predicted values for the same period for different forecast procedures. To assess the effect of the spatial weighting, we computed both RMSE and the mean absolute percentage deviation, MAPD for the different forecast steps ($MAPD = \frac{1}{M}\sum_{t=1}^{M}\left(\frac{|y_t - \hat{y}_t|}{|y_t|}\right)$).

Table 4. Out-of-Sample Forecasts

Forecast for 2009	Actual Value	Weighted Forecast	Unweighted Forecast	Weighted		Unweighted	
				MAPD	RMSE	MAPD	RMSE
Jan	3.044522	2.939382	5.208310	M=1:			
Feb	4.174387	2.770864	4.487440	0.0345	0.1051	0.7107	2.1638
Mar	4.369448	3.915668	3.894846	M=3:			
Apr	3.091042	3.860014	4.259645				
May	4.615121	4.376068	3.735367	0.1582	0.8538	0.2981	1.2917
Jun	5.347108	3.931153	4.622641	M=6:			
Jul	4.997212	5.077728	3.817844				
Aug	4.317488	4.812715	3.637416	0.1733	0.8982	0.2664	1.1306
Sept	3.931826	4.607082	4.456604	M=12:			
Oct	4.127134	4.629895	3.005752				
Nov	3.637586	3.189317	2.980330				
Dec	3.433987	4.079065	2.969817	0.1480	0.7313	0.2261	0.9876

Table 4 shows the results for the selected site. The spatially weighted forecasting conditioning on Wy outperforms the default forecast at all levels of forecast steps. This performance can be checked against the RMSE and MAPD values for the spatially weighted and the unweighted forecasts.

A rolling forecast was undertaken for the spatially weighted series for the same selected site. Results of the comparison with the year-long forecast is shown in Table 5.

Table 5. Comparison of Rolling and Year Long Predictions

Actual y_t	Year-long $\hat{y}_t$	Rolling $\hat{y}_t$	Year-long $\lvert y_t - \hat{y}_t \rvert$	Rolling $\lvert y_t - \hat{y}_t \rvert$
3.044522	2.939382	2.939382	0.105140	0.105140
4.174387	2.770864	2.559900	1.403523	1.614487
4.369448	3.915668	3.380577	0.453780	0.988871
3.091042	3.860014	3.468199	0.768972	0.377157
4.615121	4.376068	3.696763	0.239053	0.918357
5.347108	3.931153	3.308346	1.415954	2.038761
4.997212	5.077728	4.753443	0.080516	0.243769
4.317488	4.812715	4.575633	0.495227	0.258145
3.931826	4.607082	5.613441	0.675257	1.681616
4.127134	4.629895	7.590778	0.502761	3.463644
3.637586	3.189317	6.127708	0.448269	2.490121
3.433987	4.079065	6.221942	0.645078	2.787955
		MAPD	0.1480	0.3518
		RMSE	0.7313	1.7716

5. Conclusion

Using IDW to incorporate spatial dependence into the analysis of spatial data within the framework of MSSA time series analysis leads to improved quality of statistical analysis. This can be seen from Table 1, where only qualitative neighbourhood effects have been included into the analysis, while in Table 2, spatial lag (weight) matrix was used to incorporate spatial dependence into the analysis and from Table 3 where the RRMSE value is smallest for the spatially weighted set. Table 6 gives the comparison of the accuracy measures for the entire set of sites. The year-long prediction outperforms the rolling forecast potentially due to the seasonality within the original time series.

We therefore highly recommend incorporation of spatial dependence (via spatial weight matrix) into the model free time series analysis within the framework of the SSA-based techniques. For prediction, the inherent characteristics or features of the time series under investigation should be taken into consideration before deciding upon the method to use.

Table 6. Comparison of Prediction Measures

Site		MAPD		RMSE	
N ^{o}	Name	Weighted	Default	Weighted	Default
1	Freistadt	0.1480	0.2261	0.7313	0.9876
2	Linz/Stadt	0.2018	0.2632	1.0416	1.2058
3	Reichenau	0.2161	0.2233	1.1048	1.0576
4	Wolfsegg	0.2144	0.2606	1.1479	1.2745
5	Wels	0.1979	0.2369	0.9044	1.0685
6	Hoersching	0.1674	0.1573	0.8566	0.7227
7	Kremsmuenster2	0.1904	0.2282	1.0005	1.1317
8	Mondsee	0.1188	0.1757	0.7226	0.9917
9	Gmunden	0.1473	0.2284	0.7541	1.1337
10	Bad Goisern	0.1712	0.2105	0.9427	1.0822
11	Bad Ischl	0.1830	0.1569	0.9915	0.8793

Acknowledgment

The author is greatly indebted to Prof. W. G. Müller for his invaluable guidance and provision of the data set.

References

[1]. Awichi, R. O. (2015). Spatiotemporal Predictions using an MSSA Approach. *Unpublished PhD Thesis*; IFAS, Johannes Kepler University, Linz, Austria.

[2]. Awichi, R. O. and Müller, W. G. (2015). In-Sample Spatio-temporal Predictions by Multivariate Singular Spectrum Analysis. *Procedia Environmental Sciences* 26, 19-23.

[3]. Awichi, R. O. and Müller, W. G. (2013). Improving SSA Predictions by Inverse Distance Weighting. *Revstat Statistical Journal*, Vol.11(1), 105-119.

[4]. Bivand, R. S., Pebesma, E. J. and Gòmez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Second Edition, Springer.

[5]. Brockwell, P.J. and Davis, R.A. (2010). *Introduction to Time Series and Forecasting*. Springer, New York.

[6]. Broomhead, D.S. and King, G.P. (1986). Extracting Qualitative Dynamics from Experimental Data. *Physica D*, Vol 20, 217-236.

[7]. Cressie N. A (1993). Statistics for Spatial Data. Revised Edition, Wiley.

[8]. Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F. and Yiou, P. (2002). Advanced Spectoral Methods for Climatic Time Series. *Reviews of Geophysics*, Vol 40(1), 1-41.

[9]. Ghodsi, M., Hassani, H. and Sanei, S. (2010). Extracting fetal heart signal from noisy maternal ECG by multivariate singular spectrum analysis. *Statistics and Its Interface*, Vol.3, 399-411.

[10]. Golyandina, N., Korobeynikov, A., Shlemov, A. and Usevich, K. (2013). Multivariate and 2D Extensions of SSA with Rssa Package. *arXiv:1309.5050V1[stat.ML]*.

[11]. Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall; CRC, New York-London.

[12]. Golyandina, N. and Zhigljavsky, A. (2013). Singular Spectrum Analysis for Time Series. Springer.

[13]. Hassani H. and R. Mahmoudvand (2013). Multivariate Singular Spectrum Analysis: A general view and New Vector Forecasting Approach. *International Journal of Energy and Statistics*, Vol. 1(1), 55-83.

[14]. Hassani, H., S. Heravi and Zhigljavsky, A. (2013). Forecasting European Industrial Production with MSSA. *Journal of Forecasting*, Vol. 32(5), 395-408.

[15]. Korobeynikov, A., Shlemov, A., Usevich, K. and Golyandina, N. (2015). Rssa: A collection of methods for singular spectrum analysis http://CRAN.R-project.org/package=Rssa. R package version 0.13.

[16]. Mateu, J. and Müller, W. G. (Eds.) (2012). *Spatio-temporal design: Advances in efficient data acquisition; (Statistics in Practice)*. Wiley.

[17]. Müller, W. G. (2007). *Collecting Spatial Data*. Third Edition, Springer.

[18]. Patterson, K., Hassani, H., Heravi, S. and Zhigljavsky, A. (2011). Multivariate Singular Spectrum Analysis for Forecasting Revisions to Real-time Data. *Journal of Applied Statistics*, 38:10, 2183-2211.

[19]. Raynaud, S. Yiou, P. Kleeman, R. and Speich, S. (2005). Using MSSA to Determine Explicitly the Oscillatory Dynamics of Weakly Nonlinear Climate Systems. *Journal of Nonlinear Processes in Geophysics* Vol. 12, 807-815.

Published: Volume 2018, Issue 1 / January 25, 2018