Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 19 – 23

Procedia
Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

In-Sample Spatio-temporal Predictions by Multivariate Singular Spectrum Analysis

Richard O. Awichi[a], Werner G. Müller[b]*

[a] Uganda Martyrs University, Nkozi, Uganda
[b] Johannes Kepler University, Linz, Austria

In this paper, we present a method of utilizing spatial information, usually intrinsic in spatial data sets, to improve the quality of temporal predictions within the framework of singular spectrum analysis (SSA) techniques. Those constitute a model free approach to time series analysis and 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.

The aim of SSA is twofold: 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 to make a forecast in the absence of the noise component. It has two stages each with two steps. The first stage is the Decomposition Stage with steps comprising: (i) embedding – a usual procedure in time series analysis; the result of which is a called a trajectory matrix in SSA literature; (ii) singular value decomposition step that computes and arranges eigenvalues of this matrix for further analysis. The second stage is the Reconstruction Stage: it comprises two steps of (i) grouping the eigenvalues of the SVD step and (ii) diagonal averaging that works on the grouped matrices to realize a series close to the original series and this series is then used for further analysis.

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 environmental sciences and other areas 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. Spatial dependence subsequently influences data analysis. The usual question asked is whether the location parameters can be of use in the analysis of such data sets. We present a method that can be used to harness the location attributes to enhance prediction of spatial data sets using an MSSA approach.

This technique is applied to climate data recordings (particularly rainfall data) from Upper Austria.

* Corresponding author. Tel.: +43-732-2468-6801; fax: +43-732-2468-6800.
E-mail address: werner.mueller@jku.at

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.016

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: Series Analysis; MSSA; Inverse Distance Weighting; Spatial Dependence ;

1. Introduction

SSA, a recently popularized tool for time series analysis, is a model free approach to the analysis of time series, as opposed to model based time series analyses with several restrictive assumptions, see for example, [3] . The beginning of SSA is usually attributed to [4]. Literally, any time series with a notable structure can provide an application of SSA. It can be applied to many areas, see [8,5].

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 so as to make a forecast in the absence of the noise component.

The following are the steps of SSA in brief. The first step is the embedding step in which the time series is transformed into a trajectory matrix **X** using an embedding operator: which maps the time series into the multidimensional data matrix **X**. The single most important parameter in this step is the window length, L, which defines the amount of lagging. The second step is the singular value decomposition step in which the trajectory matrix is factorized into a sum of elementary matrices using the nonzero eigenvalues of . The grouping step is the third step where the elementary matrices are split further through the procedure of *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*. This requires that for any two or more time series components to be separated from one another, the corresponding columns (and rows) of the components must be orthogonal. This is then referred to as *weak* separability. For *strong* separability, an additional condition of unique eigenvalues of the matrix is imposed, see [8,10] for details.

MSSA is a direct extension of SSA to multivariate analysis and has been applied before to climate studies, [16]. Kriging and inverse distance techniques are ways in which spatial information can be included in the predictions of SSA and subsequently MSSA. Here as in [1], we explore the inverse distance technique.

2. Multivariate Singular Spectrum Analysis, MSSA

This is an extension of SSA to multidimensional data in which the trajectory of the **s** variate time series is given as . The matrix **X** is of order , see [15]. The aim of MSSA is also to decompose the initial time series into additive and interpretable components and later to reconstruct the decomposed series for further analyses. MSSA also comprises two stages of decomposition and reconstruction, each with two steps. The algorithms of MSSA are an extension of those of SSA except for the first step which is rather different from the first step of the univariate SSA. For emphasis, we present a brief discussion of the first two steps of MSSA algorithms below, the other two steps are a somewhat direct extension of the corresponding steps of SSA.

Let denote the general representation of a multivariate time series of the same series length N, otherwise for different series lengths, we can use to denote the time series.

Step 1: Embedding

For a fixed L the window length, the embedding procedure produces K = N - L + 1 lagged vectors, of the trajectory matrix **X** given as, ; this is a block Hankel matrix as opposed to the single Hankel matrix of the univariate step. Each of the s blocks of K columns corresponds to the trajectory matrix for a particular vintage. For simplicity, we have assumed here that N and K are uniform, otherwise if they different for each block, then the individual trajectory matrices are stacked horizontally (or indeed vertically) and may have different column dimensions. The trajectory space is the linear space spanned by the lagged vectors.

Step 2: Singular Value Decomposition, SVD

This is basically similar to step 2 of SSA, though it is also possible to use Principal Component Analysis, PCA instead of the SVD at this step. PCA extracts the orthogonal components of the initial series to achieve a reduced dimensionality of the trajectory matrix, see [15]. The SVD matrices are much larger and depend on whether the individual trajectory matrices are stacked horizontally or vertically and thus different conditions on L and K. This step represents the trajectory matrix **X** as, .
The major aim of the first two steps is to achieve *separability* of the components in the decomposition of the series. This makes the selection of L critical: it should not be too small since then not all the components will be captured, neither should it be too big because it becomes rather difficult to trace the behaviour of the series as some noise components may be included in the selection, see [6]. It should just be large enough to capture the essential behaviour of the time series. It is known that for univariate SSA, , whereas for MSSA, . The degree of separability can be assessed empirically by means of the corresponding w-correlation coefficients or from a correlation coefficient, see [9].

Forecasting and Separability of the MSSA technique are direct extension of the analysis given for SSA. Environmental Science is an area where space-time behaviour is an important focus of study. Many times several unobservable factors affect the analysis of the data so collected, [14]. This may lead to spatial dependence. To include factors due to spatial dependence into the analysis, one may use kriging or a function of the (Euclidean) distance between the points of analysis - the inverse distance weighting technique. With geo-referenced data, a question usually arises whether location information can be used to improve statistical analysis within the framework of MSSA.

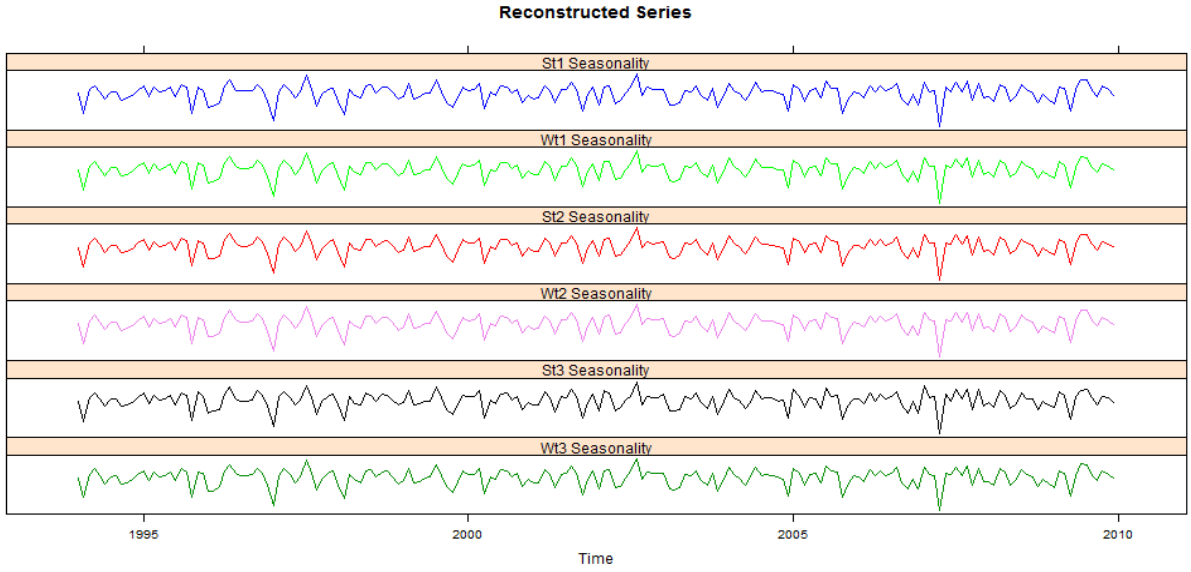

Fig. 1. Three selected series compared with their spatially weighted counterparts.

3. MSSA with Inverse Distance Technique

Spatial data analysis is usually required for planning purposes and decision making. It includes information on geographical locations of data points and the analysis is aimed at improving the quality of 'local' accuracy. Inverse distance weighting assigns bigger weights at near points. There are several ways of computing the inverse distance weights to be used in the analysis, see [13]. Here, we use the Euclidean distance approach introduced in [1] and we have extended it to multivariate data. The multivariate data set **y** = is an matrix and the inverse distance is given as where is the Euclidean distance between the locations. For missing values in the data, a new weight is calculated by excluding the corresponding distance measure from the s. The nearer the locations, the stronger the dependence and vice versa. Eventually the spatial weight matrix **W** is row normalized to conserve the location attributes, see [12]. The data set is then premultiplied by this normalized weight matrix to yield the spatially weighted averages **Wy**. These calculations were done in R using the packages *dist* and *apply*, see [17]. For more information on how to do spatial data analysis in R, see [2]. In our case, we used rainfall data from 11 sites without missing data thus N=192. We first pooled the unweighted data for the 11 locations by conditioning each site data on all the others. For purposes of comparison, we report the root mean square errors, RMSE, for the corresponding in-sample predictions as for a particular row of **y**, using all the other .

We then pooled the spatially weighted data by conditioning in pairs each site data with its spatial average, i.e. and finally by conditioning over the entire set of locations, i.e for the corresponding particular values of . The RMSE for the unweighted series, is referred to as the default RMSE here. The analysis was done in R using the package Rssa, see [11,7,9]. The bigger the RMSE difference from the default RMSE, the better the technique. However, for better overall analysis, we also computed the ratio of the root mean square errors, RRMSE which is given by , respectively. This ratio tends towards zero since the RMSE is much smaller than the default RMSE for the optimal window length.

4. Results

The data is provided by the Zentralanstalt für Metereologie in Austria and is described in more detail in [13]. 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 .

A map of the region with the respective locations of the measurement stations and contours of the rainfall values is displayed in Figure 2 .

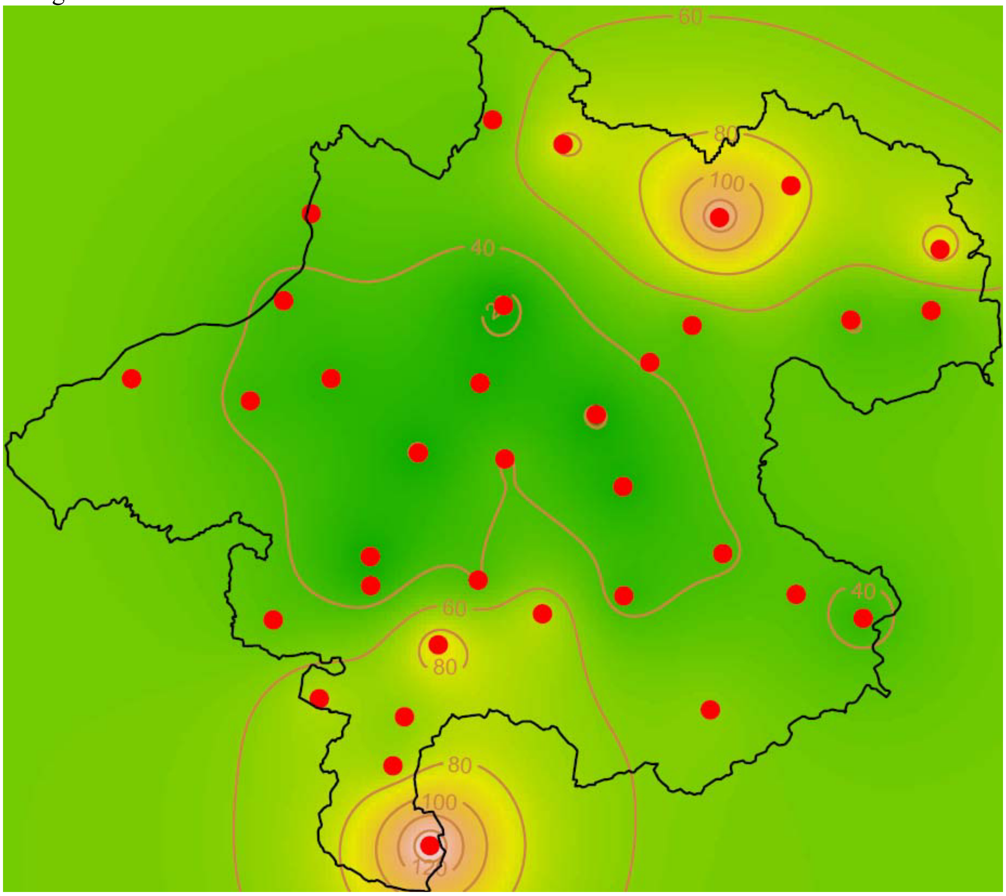

Fig. 2. The sampling locations of the climatic data set within the region of Upper-Austria

Table 1 now gives the differences of RMSE values from the default RMSE. The RMSE values were obtained for the optimum = 180, though other values of L were also tried. All other results fall within unweighted pooling and the spatially weighted with the 11 sites. The last column shows the values for L=96, the ideal value of L for the univariate case. As can be seen, this gives the worst results.

Table 1. Differences from default RMSE

Site				(L=96)
1	0.131060	0.233964	0.259907	-0.017994
2	0.104707	0.219042	0.237449	-0.013963
3	0.123672	0.220715	0.244135	-0.033471
4	0.076241	0.210099	0.236794	-0.065711
5	0.165341	0.210293	0.236306	-0.174275
6	0.087389	0.202348	0.226827	-0.021546
7	0.067623	0.184227	0.198417	-0.038519
8	0.106287	0.185549	0.198860	-0.032012
9	0.079528	0.178254	0.191761	-0.035604
10	0.101666	0.179011	0.192459	-0.037911
11	0.115954	0.177240	0.185754	-0.035474
RRMSE	0.5362	0.1197	0.0365	1.2026

4. Conclusions

From Table 1, it can be concluded that inclusion of spatial information into the MSSA analysis of time series improves the quality of results. The proposed method of including spatial information into the analysis is promising at least for in-sample analysis. Further research is being done for the out-of-sample performance.

References

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

[2] Bivand R. S, Pebesma E.J and V. Gòmez-Rubio (2008). *Applied Spatial Data Analysis with R*, Springer.

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

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

[5] Elsner, J. B. and Tsonis, A.A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenum.

[6] Golyandina, N. (2010). On the Choice of Parameters in Singular Spectrum Analysis and related space-based methods. *Statistics and Its Interface* Vol.3 , 259–279.

[7] Golyandina, N. and Korobeynikov, A. (2014). Basic Singular Spectrum Analysis and Forecasting with R. *Computational Statistics and Data Analysis* Vol 71, 934–954.

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

[9] Golyandina, N., A. Korobeynikov, A. Shlemov and K. Usevich (2013). Multivariate and 2D Extensions of SSA with Rssa Package. *arXiv.1309.5050v1*.

[10] Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison, *Journal of Data Science* Vol.5, 239–257.

[11] Korobeynikov A, Shlemov A., Usevich K and Golyandina N. (2014). *Rssa: A collection of methods for singular spectrum analysis www.CRAN.R-project.org/package=Rssa*. R package version 0.11.

[12] LeSage James and R. K Pace (2009). Introduction to Spatial Econometrics, Chapman & Hall; CRC.

[13] Mateu, J. and Müller, W. G. (Eds.) (2012). *Spatio-temporal Design: Advances in Efficient Data Acquisition; (Statistics in Practice)*, Wiley.

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

[15] 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.

[16] 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.

[17] R Core Team (2012). R: A language and environment for statistical computing, R Foundation for statistical computing, Vienna, Austria (www.r-project.org).