Journal of Environmental Statistics
August 2013, Volume 5, Issue 2. http://www.jenvstat.org

A spatio-temporal model based on the SVD to analyze daily average temperature across the Sicily region

Rossella Onorati
Dip. Scienze Statistiche
e Matematiche "S. Vianelli"
University of Palermo, Italy

Paul Sampson
Department of Statistics
University of Washington
USA

Peter Guttorp
Department of Statistics
University of Washington
USA

Abstract

A common problem in the analysis of space-time data is to compress a large dataset in order to extract the underlying trends. Empirical orthogonal function (EOF) analysis is a useful tool for examining both the temporal and the spatial variation in atmospherical and physical processes, and a convenient method of performing this is the Singular Value Decomposition (SVD). Many spatio-temporal models for measurements $Z(t, \mathbf{s})$ at location $\mathbf{s}$ at time t, can be written as a sum of a systematic component and a residual component: $\mathbf{Z} = \mathbf{M} + \mathbf{E}$, where $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{E}$ are all $T \times N$ matrices. Our approach permits modeling of incomplete data matrices using an EM-like iterative algorithm for the SVD. We model the trend, $\mathbf{M}$, by linear combinations of smooth temporal basis functions derived from left (temporal) singular vectors of $\mathbf{Z}$ with the dimension of the model chosen by cross-validation. We further decompose by SVD the spatio-temporal residual matrix $\mathbf{E}$ computed as residuals from regressions at each site (column) of the observations on smoothed temporal basis functions. Finally we fit an autoregressive model to the columns (time series) of residuals from the SVD of $\mathbf{E}$. Our aim is to illustrate a simple model to characterize trends and model the variability in large spatio-temporal data matrices. The methodology is demonstrated with 30 years of daily temperature data from Sicily; we obtain a good fit and a compact description of the spatio-temporal variability using just a few smoothed singular vectors.

Keywords: SVD, spatio-temporal model.

1. Introduction

Spatio-temporal processes such as environmental processes are complicated. They arise from multiple factors, biological and physical, and interactions of various processes that occur at various scales in space and time. A model, in its simplified abstract view of the complex reality, becomes very useful in this context. There are several approaches to modeling, and we believe that a simple hierarchical modeling approach is helpful. We present a model based on Empirical Orthogonal Functions (EOFs) computed by SVD.

Empirical orthogonal functions are not a tool specific to the environmental sciences, but are widely used in a number of scientific domains under different names: proper orthogonal decomposition (POD) or Karhunen-Loève decompositions are mostly used in a functional framework, whereas the matrix versions (working on discrete datasets) are called proper orthogonal modes (POM), principal component analysis (PCA), factor analysis, or empirical orthogonal function analysis (Beckers and Rixen 2003).

We use the SVD decomposition to compute the EOFs, which decomposes a time-space matrix $Z_{[T\times N]}$, for T indicating time and N indicating the sites, of real-valued data and *rank* k as a product of three matrices:

$$\mathbf{Z} = \mathbf{UDV}' \tag{1}$$

where $\mathbf{U}$ is a $T \times k$ orthogonal matrix with columns $\mathbf{u}_j$, $\mathbf{V}$ is a $N \times k$ orthogonal matrix with columns $\mathbf{v}_j$ and $\mathbf{D}$ is a $k \times k$ diagonal matrix with diagonal entries $\mathbf{d}_j$ typically taken to be a decreasing sequence of nonnegative numbers, for $j = 1, ..., k$, where $k = \min(N, T)$. In the application below $T > N$, so $k = N$. For the purposes of this first SVD, the $\mathbf{u}_j$ and $\mathbf{v}_j$ represent a set of temporal and spatial *empirical orthogonal functions* (EOF), respectively.

The appeal of the singular value decomposition is partly due to its interpretation as a multiplicative model based on row and column factors. Letting the elements of the temporal (left) singular vectors $\mathbf{u}_j$ be written as $u_j(t)$ and the elements of the spatial (right) singular vectors $\mathbf{v}_j$ be denoted by $v_j(\mathbf{s})$, the spatio-temporally indexed datum $Z(t, \mathbf{s})$, can be written as a sum of products of spatial and temporal variables: $Z(t, \mathbf{s}) = \sum_{j=1}^{N} d_j u_j(t) v_j(\mathbf{s})$. If it suffices to approximate $\mathbf{Z}$ by just its first EOF, then we are saying that $Z(t, s) \approx d_1 u_j(t) v_1(\mathbf{s})$, i.e. the spatiotemporal process can be approximated by a separable process (Banerjee, Carlin, Gelfand, and Raton 2004).

An important thing to point out is the direct relation between PCA and SVD in the case where principal components are calculated from the covariance matrix. In fact centering each column of the $Z(t, \mathbf{s})$ matrix by the vector of column averages, then $\mathbf{Z}'\mathbf{Z} = \mathbf{VD}^2\mathbf{V}'$ is a $T \times T$ matrix that is proportional to the sample spatial covariance matrix. Then the diagonalization of $\mathbf{Z}'\mathbf{Z}$ yields $\mathbf{V}'$, which also constitutes the principal components of columns variables; the $\mathbf{v}_j$ are the same as the principal components of the column variables, i.e. sites, and the eigenvalues of $\mathbf{Z}'\mathbf{Z}$ are equivalent to d_j^2 (Wall, Rechtsteiner, and Rocha 2003). The SVD of $\mathbf{Z}$ provides a one-step method for computing all the components of the eigenvalue problem, without having to compute and store large

covariance matrices. It produces two orthonormal bases, one defined by right singular vectors and the other by left singular vectors, that span the space of variation over time and the space of spatial variation over the sites, respectively. Then according to our interest in the relations between observations over time or in the variation across sites we may express the signal of interest by linear combinations of right singular vectors or by left singular vectors.

Our analysis involves two SVD computations and a temporal correlation model. We first model temporal trend, **M**, by linear combinations of smoothed orthogonal temporal basis functions, derived from the singular value decomposition of **Z**, with the number of smooth basis functions chosen by a cross-validation criterion to explain spatially varying seasonality and long-term trend. We then decompose, again using SVD, the spatio-temporal matrix **E** of residuals from the linear models fitted at each site to the temporal trend basis functions from the first SVD fit. This second SVD represents short scale temporal processes that are also correlated in space. We finally fit autoregressive models to the columns of residuals from this second SVD. Our aim is to illustrate a simple model to describe the spatio-temporal trends and residual variability in a large spatio-temporal data matrix. The methodology was applied to a spatiotemporal dataset of daily temperature averages observed at fifty monitoring sites on the island of Sicily over thirty years, $1965 - 1994$. The analysis shows a good fit and provides a compact description of the spatio-temporal variability using just a few smoothed singular vectors. In the next section we show the dataset under study; section 3 presents the proposed model and a simulation study; section 4 illustrates the application of the model to the above-mentioned dataset, and conclusions follow in section 5.

2. The data and the study area

We analyze a spatiotemporal dataset of daily temperature averages observed at fifty monitoring sites, belonging to the Sicilian Regional Hydrographic Service, from 1965 to 1994 in Sicily, the largest island in the Mediterranean with a total area of about $25,000 km^2$ extending in latitude between 36° and 38° north and longitude between 12° and 15° east. Despite the geographic situation of a very complex area, with unique morphological aspects, you can divide the Sicilian territory into three distinct regions: the northern, the southern and eastern sides. Sicily is characterized by a mesothermic humid sub-tropical climate with dry summers. Its surface is mostly covered by plains; a mountain chain runs along the North-East side of the island and an active volcano is at the lower end of the chain on the East coast; the central and south-western is mainly hilly.

Monitoring sites spanning the area show very hight temporal long memory and very high spatial correlations ranging from 0.86 to 1. A few of the sites have data missing for long periods and in general about 10% of the entire dataset is missing.

3. Model formulation

Denoting by N the number of sites $\mathbf{s}_1, ..., \mathbf{s}_N$ and by T the number of observations for each site, it is common in the space-time modeling literature to decompose the measurement at location $\mathbf{s}$ at time t into the sum of a systematic trend component and residuals

$$Z(t, \mathbf{s}) = M(t, \mathbf{s}) + E(t, \mathbf{s}) \tag{2}$$

where $M(t, \mathbf{s})$ denotes the mean structure and $E(t, \mathbf{s})$ denotes the residual process. The allocation of total spatial and temporal variation between these two components is partly the decision of the modelers based on their understanding of the problem and partly a statistical decision based on model fit and parsimony. A flexible model is useful to provide such discretion. Here, we use a nonparametric approach to model the spatio-temporal mean field of a random field, an approach first introduced in Fuentes, Guttorp, and Sampson (2007). This approach starts with decomposing the matrix $\mathbf{Z}$ using SVD and, applying *cubic smoothing splines* to some of the left singular vectors, decomposes the trend as follows:

$$M(t, \mathbf{s}) = \beta_{s_0} + \sum_{j=1}^{J} \beta_{s_j} f_j[u_j(t)] \tag{3}$$

where:

- β_{s_0} is the long term average of the site s;
- J is the number of left singular vectors in the singular value decomposition of $Z = UDV'$ chosen by Cross-Validation across the sites;
- $f_j[u_j(t)]$ are orthogonal temporal basis functions (*cubic smoothing splines*);
- $\boldsymbol{\beta}_{N \times J}$ is the matrix of trend coefficients estimated by multiple linear regression on the smoothed basis functions f. Because of the spatially varying nature of these coefficients could be modeled as spatial random fields.

Following the above methodology we are able to account for the very strong seasonal cycle of atmospheric temperature and long-term trends. Moreover since the phase and amplitude of the seasonal cycle (not necessary sinusoidal) vary as a function of space, the parameters describing the seasonal cycle, β_{s_j}, are allowed to vary as a function of latitude and longitude; they represent a spatial field underlying the process of interest. Also, since the numerical SVD algorithms requires a full matrix Z we apply an "EM-like" iterative algorithm for the SVD as in Fuentes *et al.* (2007).

We may further decompose the spatiotemporal residual matrix E as a sum of a mean zero spatiotemporal process, $w(.)$ and a temporally correlated noise process, $\epsilon(.)$: $w(t, \mathbf{s}) +$

$\epsilon(t, \mathbf{s})$. We derive this from a second SVD,

$$E(t, \mathbf{s}) = \sum_{k=1}^{K} d_k^{(2)} \mathbf{u}_k^{(2)}(t) \mathbf{v}_k'^{(2)}(\mathbf{s}) + \epsilon(t, \mathbf{s}) \tag{4}$$

- $\mathbf{u}_k^{(2)}(t)\mathbf{v}_k'^{(2)}(\mathbf{s})$ and $d_k^{(2)}$ are the kth EOF and singular value in the singular value decomposition of E and it is a mean-zero spatio temporal process;
- $\epsilon(t, \mathbf{s})$ is a matrix of time series each modeled as $AR(L)$,

$$\epsilon(t, \mathbf{s}) = \sum_{l=1}^{L} \phi_l(\mathbf{s}) \epsilon(s, t-l) + \eta(s, t), \tag{5}$$

  for $s = 1, ..., N$ and $\eta \sim N(0, \sigma_\eta)$.

The second decomposition, (4), yields another reduction in dimension, representing E as a sum of K products of a spatial process and a temporal process. This second SVD is useful to reduce the autocorrelation of residuals; also, it reduces the residual sum of squares when we compare this model with a model with no second SVD but just with more smoothed basis functions. What is left then is only a set of low order autoregressive models for each site.

The idea behind this final model is that each site has an individual story with its own pattern over years and geographic location, but at the same time they lie in the same area and they share common weather. After all the main structure has been removed, each site should follow an individual residual model that allows for filtering out most of the short-range temporal correlation typical of meteorologically driven data such as air pollution data (Meiring, Guttorp, and Sampson 1998).

3.1. Choosing the number of components

The choice of the number of temporal basis functions, the dimension of the model, is often based on a heuristic approach. The magnitude of any one singular value is indicative of its importance in explaining the data. Therefore *scree plots* (Cattell 1966) illustrating the fraction of total variance in the data explained by each principal component $d_j^2 / \sum_{j=1}^{N} d_j^2$, have been proposed as a graphical method to decide on the number of important components.

In our particular case of modeling through SVD, we may take into consideration the typical effect of the SVD on covariance matrices, namely to concentrate most of the explained variation in the first component. To partially solve this problem, higher order of smoothing of the temporal basis can be used, and we note that the more we smooth these basis functions, the less structure we catch in the systematic trend. This increased smoothing leaves more variability in the residual matrix E (as the simulation section

shows). Also, a stronger autocorrelation structure of the left singular vectors obtained after the second SVD is highlighted.

Our method of component selection uses the Bayesian Information Criterion, BIC (also known as the Schwarz Criterion), computed for cross-validation fits leaving out a subset or just one monitoring site at a time. In addition to this criterion, the amount of variability explained by components is also taken into consideration for the final choice. Starting from a first guess of the number of components, say j, we leave out a random subset of sites, then we follow these steps:

- Compute the SVD;
- Smooth the J left-singular vectors using *cubic smoothing spline* with smoothing parameter chosen by generalized Cross Validation (GCV);
- Compute the trend prediction of every site in the left out cross validation group by least squares regression on the smoothed trend components and evaluate the BIC criterion.

After we get a vector of BIC values we redo the previous steps with $j + 1$ components, getting again a vector of BIC values. We run again those steps varying the number of components and the number of left out sites, in this way we will get sites showing better fits according to BIC with more components and some sites with fewer components. Therefore, we choose the number of components with better BIC values for the most sites in these cross-validated fits.

3.2. Simulation study

To test the capability of the model to reproduce an original signal we carried out a simulation study on 1000 synthetic realizations of a spatial process with time evolution. These 1000 matrices have dimension $T \times N$ where $T = 10957$ and $N = 50$ and they have been generated with a structure similar to the real data set illustrated in the next section, i.e. as a systematic trend with three principal components, one empirical orthogonal function for the mean zero process noise component $w(t, \mathbf{s})$ of the decomposition of equation (4), and a matrix of time series for columns, each of them generated from AR(2) models. To assure that we are generating data from three clearly distinguishable components, we specify the first three singular values equal to 5000, 500 and 50 and zero for the higher orders while the left and right singular vectors are taken from the Sicily data analyzed in the next section. Also the first EOF elements, $U^{(2)}$ and $V'^{(2)}$ of equation (4), are generated as $AR(3)$ and Gaussian random field, respectively, with autoregressive parameters and partial sill and range values estimated from the Sicily data (partial sill and range are maximum likelihood estimates). We use the real data to estimate the autoregressive parameters for the generation of the columns of the final residual matrix, ϵ. That is, we consider the scenario described by

$$\mathbf{Z} = \mathbf{UDV}' + \mathbf{E} = \mathbf{UDV}' + \mathbf{U}^{(2)}\mathbf{V}'^{(2)} + \boldsymbol{\epsilon} \tag{6}$$

where:

$\mathbf{U}$ is a matrix of left singular vectors 10957×50;

$\mathbf{D}$ is a matrix of eigenvalues with the first three diagonal entries equal to $5000, 500, 50$ and zero elsewhere;

$\mathbf{V}$ is a matrix of right singular vectors 50×50;

$\mathbf{U}^{(2)}$ is a vector of length 10957 from an AR(3) with $\rho_1 = 1.275$ $\rho_2 = -0.529$ and $\rho_3 = 0.072$;

$\mathbf{V}'^{(2)}$ is a Gaussian random field with Exponential covariance function and parameters $\sigma^2 = 0.0009$ and $\phi = 0.32$;

$\boldsymbol{\epsilon}_{[T\times N]}$ is a matrix where each column are generated from $AR(2)$ with $\rho_1 \in (0.408, 0.774)$ and $\rho_2 \in (-0.066, 0.216)$ for $s = 1, ..., N$;

In these synthetic datasets, we know the true signal (components or singular vectors) and the noise, so that we can make all necessary statistical comparisons and error analyses of the true and estimated components. We consider a simulation study similar to the one presented by Beckers and Rixen (2003) and for each generated matrix we consider the following three matrices:

- $\mathbf{Z_t} = \mathbf{UDV'}$, a matrix representing the *true* expectation value we want to reproduce, for fixed value of U, D, and V' in the scenario;
- $\mathbf{Z_p} = \mathbf{UDV'}+\mathbf{E}$, a matrix that, for each replication, represents the true expectation value *perturbed* with a simulated noise component;
- $\mathbf{Z_a}$, a matrix representing what we really observe, obtained from $\mathbf{Z_p}$ removing randomly 20% of the values.

For each of these matrices we compute their estimation through SVD and evaluate the error after the first SVD in terms of mean square error. In particular the mean square error, MSE, between the absolute values of the true left and right singular vectors (the ones from Sicily data we used to make the scenario (6)) and the left and right singular vectors estimated from the SVD in the case of missing data:

$$MSE_u = \frac{\Sigma_{i=1}^{T}(|u|_{i,k} - |\tilde{u}|_{i,k})^2}{T} \qquad MSE_v = \frac{\Sigma_{j=1}^{N}(|v|_{j,k} - |\tilde{v}|_{j,k})^2}{N} \tag{7}$$

and pairwise mean square error of every element (absolute values) of the singular decomposition estimated from each of the three matrices previously shown, where we indicate

the comparison of the SVD elements between Z_t and Z_p by (tp), between Z_a and Z_a by (at) and between Z_a and Z_t by (at):

$$MSE_{u,(tp)} = \frac{\Sigma_{i=1}^{T}(|\tilde{u}|_{ik,(t)} - |\tilde{u}|_{ik,(p)})^2}{T} \text{ same for } MSE_{u,(ap)} \text{ and } MSE_{u,(at)} \quad (8)$$

$$MSE_{v,(tp)} = \frac{\Sigma_{j=1}^{N}(|\tilde{v}|_{jk,(t)} - |\tilde{v}|_{jk,(p)})^2}{N} \text{ same for } MSE_{v,(ap)} \text{ and } MSE_{v,(at)} \quad (9)$$

$$MSE_{d,(tp)} = \Sigma_{j=1}^{N}(\tilde{d}_{j,(t)} - \tilde{d}_{j,(p)})^2 \quad \text{same for } MSE_{d,(ap)} \text{ and } MSE_{d,(at)} \quad (10)$$

for $i = 1, ..., T$, $j = 1, ..., N$ and $k = 1, 2, 3$.

The consideration of absolute values in these MSEs is due to the fact that simulations result in randomly signed singular vectors, creating a bimodal mean square error distribution unless we consider only the absolute values of the components of the vectors.

If we focus on a only one realization of this scenario we observe the following first six estimated singular values vectors: $(4999.59, 511.02, 118.91, 110.92, 109.88, 109.77)$ and $(4999.64, 510.57, 118.18, 111.54, 110.67, 110.53)$ for the Za and Zp matrices, respectively; while the following vectors: $(0.00000004, 0.000004, 0.00006, 0.0002, 0.00009, 0.0001)$, $(0.00000007, 0.000005, 0.0014, 0.02, 0.02, 0.02)$ and $(0.12, 121.8, 4749.1, 12302.4, 12073.4, 12048.7)$ for $MSE_{u,(at)}$, $MSE_{v,(at)}$ and $MSE_{d,(at)}$, respectively.

Table 1: Mean values of the first six estimated ds for the three matrices Z_{tp}, Z_{ap} and Z_{at} over 1000 simulated datasets.

	Average of estimated singular values					
	$\bar{d}_1$	$\bar{d}_2$	$\bar{d}_3$	$\bar{d}_4$	$\bar{d}_5$	$\bar{d}_6$
$\tilde{d}_{Z_t}$	4999.932	499.980	49.997	~ 0	~ 0	~ 0
$\tilde{d}_{Z_p}$	5002.554	526.623	198.638	195.176	192.246	189.695
$\tilde{d}_{Z_a}$	5002.650	528.059	203.029	193.514	190.214	187.197

For a global point of view of the singular values analysis in this scenario we look at the 1000 simulations and, as Table 1 illustrates, we note the noise added to the data increases all singular values compared to the unperturbed field, and this means that some of the noise structure has been interpreted as part of the smooth signal. Table 1 also shows that the third component is more affected by the noise than the first two, particularly for the matrices $\mathbf{Z}_p$ and $\mathbf{Z}_a$, as we would expect from its relatively small singular value. The singular values obtained from the $\mathbf{Z}_a$ are of the same magnitude as those from $\mathbf{Z}_p$. The magnitude of the second and third smoothed components going from $\mathbf{Z}_p$ to $\mathbf{Z}_a$ is overestimated (from 526.62 to 528.06 and from 198.64 to 203.03), while for the higher smoothed components is underestimated (from 195.18 to 193.51, from 192.25 to 190.21 and from 189.69 to 187.2). It seems that the amount of missing data specified here results in little influence the estimation of the singular values and the number of components to choose.

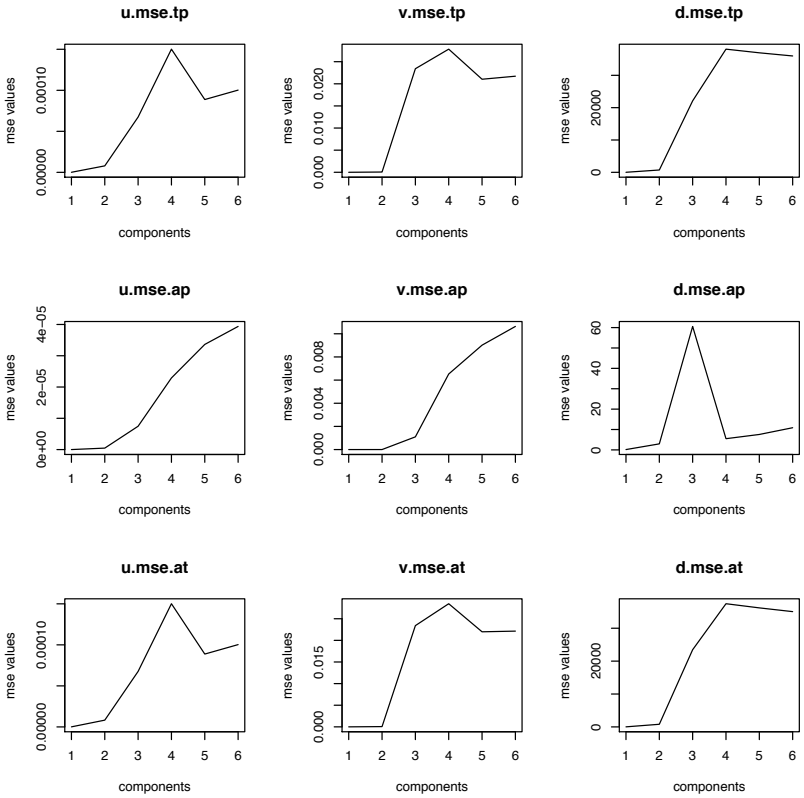

Figure 1: MSE average for the first 6 components computed over 1000 simulated matrices generated as three principal components. We note that the MSE increases from the third component for all the three SVD components in all the pairwise comparison between the matrices Z_{tp}, Z_{ap}, Z_{at}.

Figure 1 shows the average MSE for the first 6 components computed over 1000 simulated matrices generated with three principal components, and we see that the MSE increases, as expected, from the third component on for all three SVD components in all the pairwise comparisons between the matrices $\mathbf{Z}_{tp}, \mathbf{Z}_{ap}, \mathbf{Z}_{at}$. Even if we generate observations from a model with three clearly distinct components, the fact that the third singular value (equal to 50) is closer to the higher order singular values (all zero) than to the first and the second (5000, 500) makes the third component difficult to estimate when the systematic trend is perturbed.

Figure 2 illustrates that the MSE between the true and the estimated left and right singular vectors, according to (7), has almost a uniform distribution, despite that we would expect an asymmetrical shape for the mean square error empirical distribution. This is an effect of bias introduced from the stochastic part, and is very evident for the temporal elements of the SVD, the $\mathbf{u}_j$.

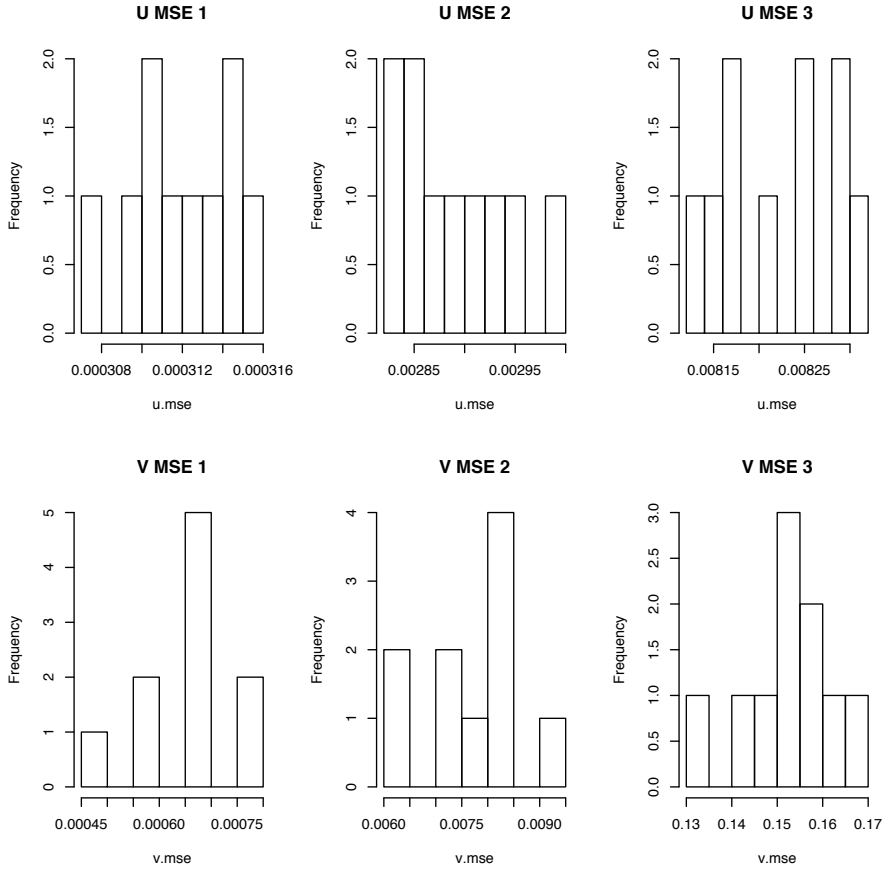

Figure 2: MSE histograms of the first 3 true components and the estimated ones.

Figure 3, according to (8), (9) and (10), shows the MSE histograms for the first 3 components computed over 1000 simulated matrices generated with three principal components, in the comparison between $\mathbf{Z}_t$ and $\mathbf{Z}_a$. Most of them show positive asymmetry as expected for the empirical distribution of the mean square error.

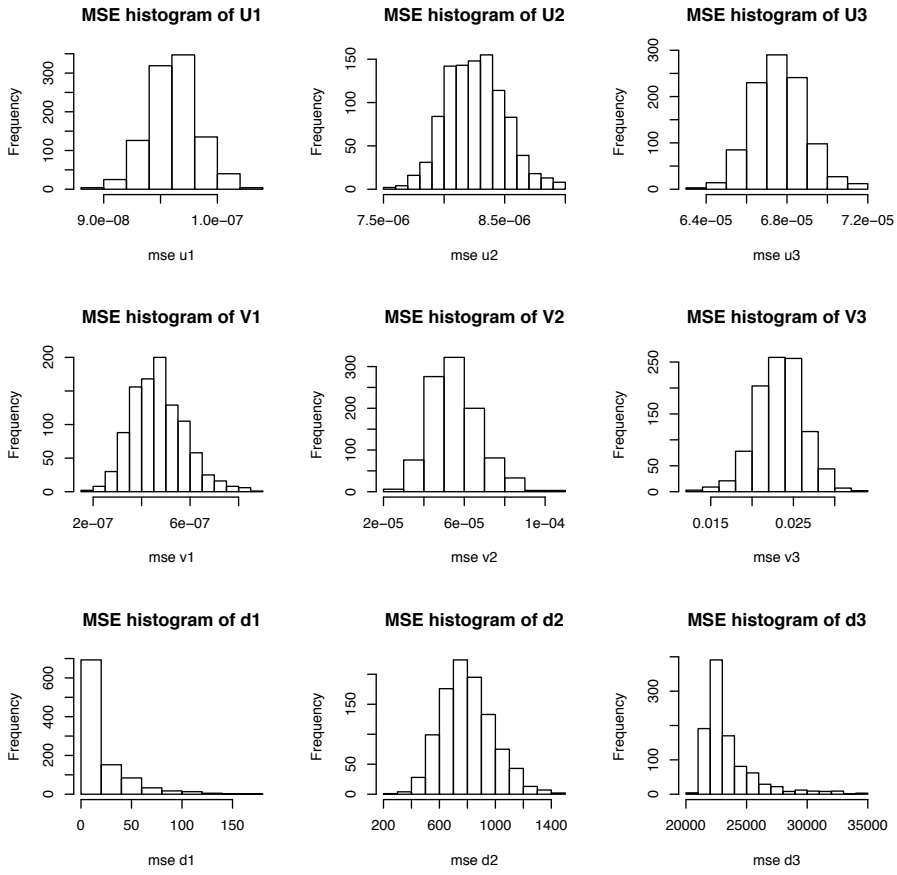

Figure 3: MSE histograms of the first 3 components of $\mathbf{Z}_a$ and $\mathbf{Z}_t$ computed over 1000 simulated matrices generated with three principal components.

An analogous simulation study has been run varying both the parameters and the covariance function for generating the Gaussian random field $\mathbf{V}'^{(2)}$. We considered $\sigma = 0.1, 0.5, 1, 10, 100$, $\phi = 0.1, 1, 10, 100, 1000$ and exponential, squared exponential and spherical families. The results (not shown) indicate that the estimated spatial structure does not seem to change according to increasing values of these parameters for each of the three families considered. Also, they hardly change between the different families. In terms of MSE evaluation, there are very subtle differences between the three cases of generating the Gaussian random field $\mathbf{V}'^{(2)}$ which are reasonable, since this affects only the small scale variability. The consideration of different temporal structures, for example considering an AR(1) instead of an AR(2), does not change things.

Finally, we show the importance of temporal smoothing in the first step in order to catch the seasonality, and the effect of smoothing on the dimensionality of the stochastic part. Figure 4 compares the average over the 1000 matrices of the singular values obtained after this second SVD. We can see there is a striking difference in term of dimensional reduction according to whether we smooth or not. Not smoothing the first three left singular vectors leaves us unable to identify a low-rank model for its further stochastic residual matrix, as opposed to the smoothed case where the choice of one EOF is clear.

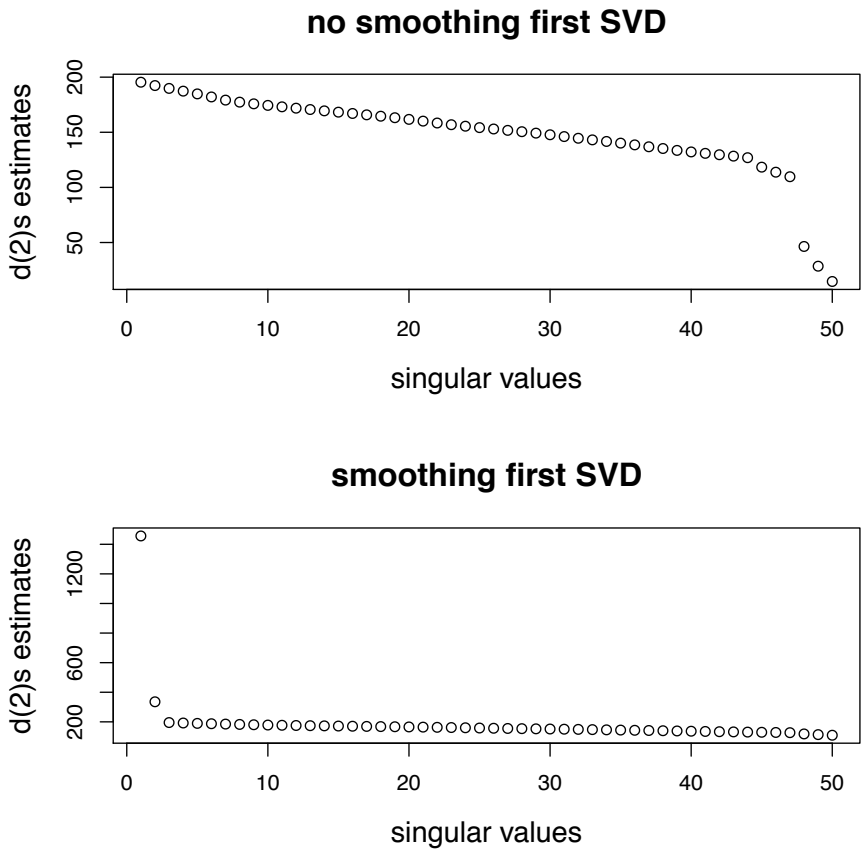

Figure 4: Singular values from the second SVD computed on the stochastic part. The top graph shows that in the case we do not smooth the first three components the second SVD identifies 47 EOFs; instead, the bottom graph shows that in the smoothing case only one EOF is identified.

In order to give a complete idea of our methodology, we have considered and analyzed one more scenario. This has two main components, a spatial part for the mean zero process $w(t, s)$ of the noise part decomposition and an error matrix. In particular, we set the singular values for those two components equal to 5000 and all the rest to 0. We

considere just 10 sites instead of 50, but retain 10957 observations. We run simulations varying both the covariance functions for generating the GRF, $\mathbf{V}^{(2)}$, (Gaussian, Squared Exponential and Spherical) and the matrix of final error, $\epsilon(t, s)$ (as a matrix of AR(2) for columns and as white noise).

The results show that we can recover the singular values. In particular, in the case of missing values ($\mathbf{Z}_a$) the first value is overestimated and the second is underestimated, the third has a value around 600 instead of zero and all the rest go to zero. This happens for both the AR(2) or white noise error matrix with slight differences in numeric values. Furthermore, if we repeat such estimation for the three covariance families, we note that in case of white noise error they are the same for all the three families, while in case of an AR(2) error matrix, the estimation in the case of Spherical covariance structure, show some differences from the other two families. In both white noise and AR(2) matrix error, the residual sum of squares from the three families have similar values and the spatial components estimated after the second SVD, $\mathbf{V}^{(2)}$, in the three families are different only in the fourth decimal place.

In conclusion, concerning the estimation of the singular values, the two components estimation shows similar results to the simulation done with three clear components, namely bigger values for the first component and smaller for the second, and also increasing values for the singular value set to zero. Further there is no strong effect of the three covariance families either for the estimating the singular values or the spatial pattern.

4. Sicilian temperatures

We developed this model in order to analyze the spatiotemporal dataset of daily temperature averages observed at fifty monitoring sites in Sicily. The analysis of these data led us to consider three temporal basis functions ($N = 3$). We smooth these three SVD components using a natural spline with smoothing parameter chosen by Generalized Cross Validation and number of knots equal to 211 according the default setting of *smooth.spline* function in R. Figure 5 shows a dominant first function representing the dominant year to year seasonal pattern of temperature with highest values during the summer months; it represents most of the variability present in the data. The initial (unsmoothed) temporal singular vector explains 93.6% of the variation in the data while the smoothed temporal basis functions explains 82.6% of the total variation. The second and third functions explain very little of the remaining variability, around 0.6% each. Some of this residual variability appears to have structure and long term trend, particularly the third, as we can see from the peaks around the early years and 1983 for the second component and the last years for the third one. The latter is the type of signal we would expect to see from warming of the climate in this region.

We use the smoothed components to estimate the $\boldsymbol{\beta}$ matrix of trend coefficients by a linear regression fitted at each monitoring site. Figure 6 illustrates the fit using these three components on a sample of three monitoring sites. The fit appears good with

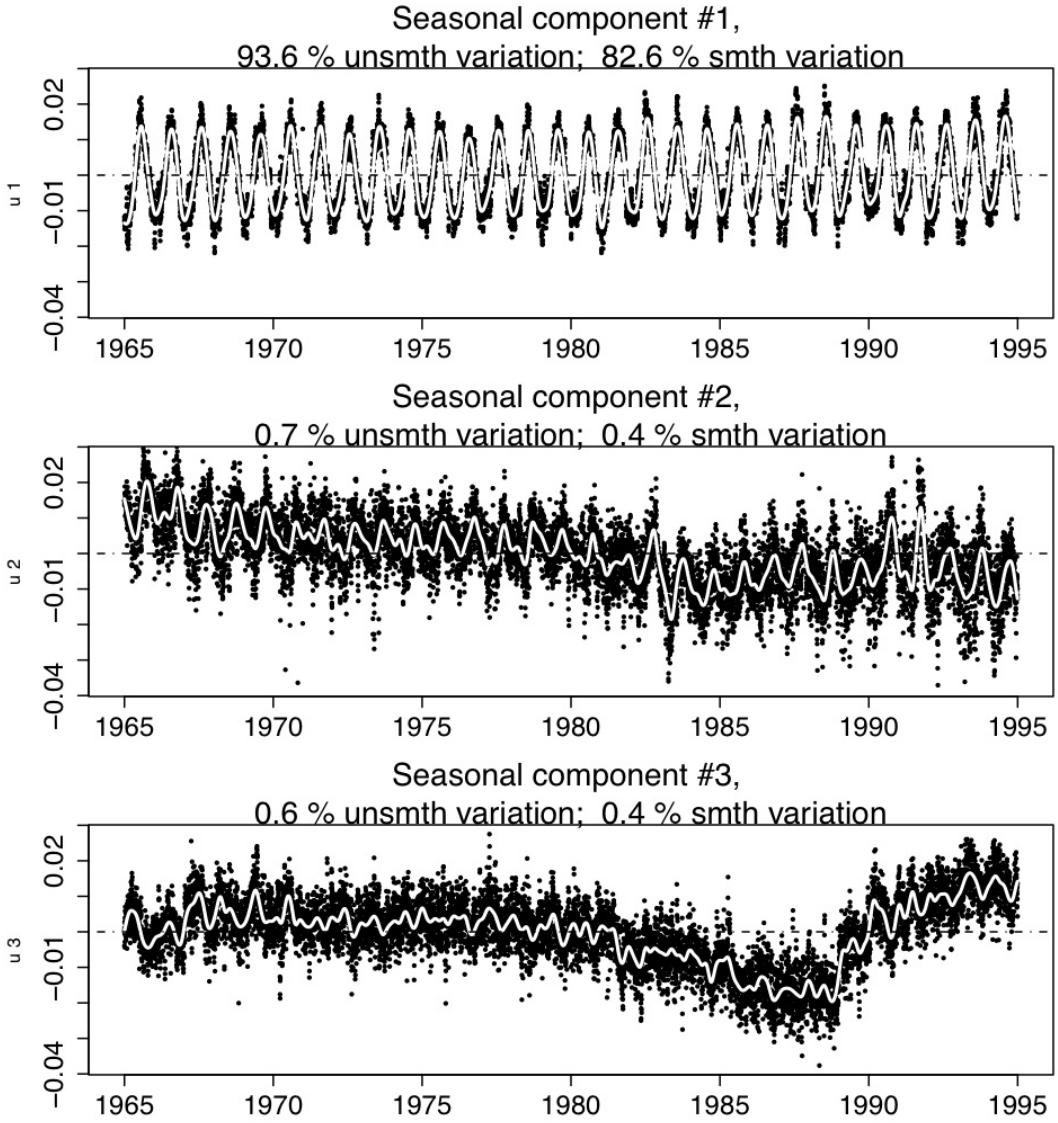

Figure 5: Temporal basis functions (black) with their smoothed versions (grey).

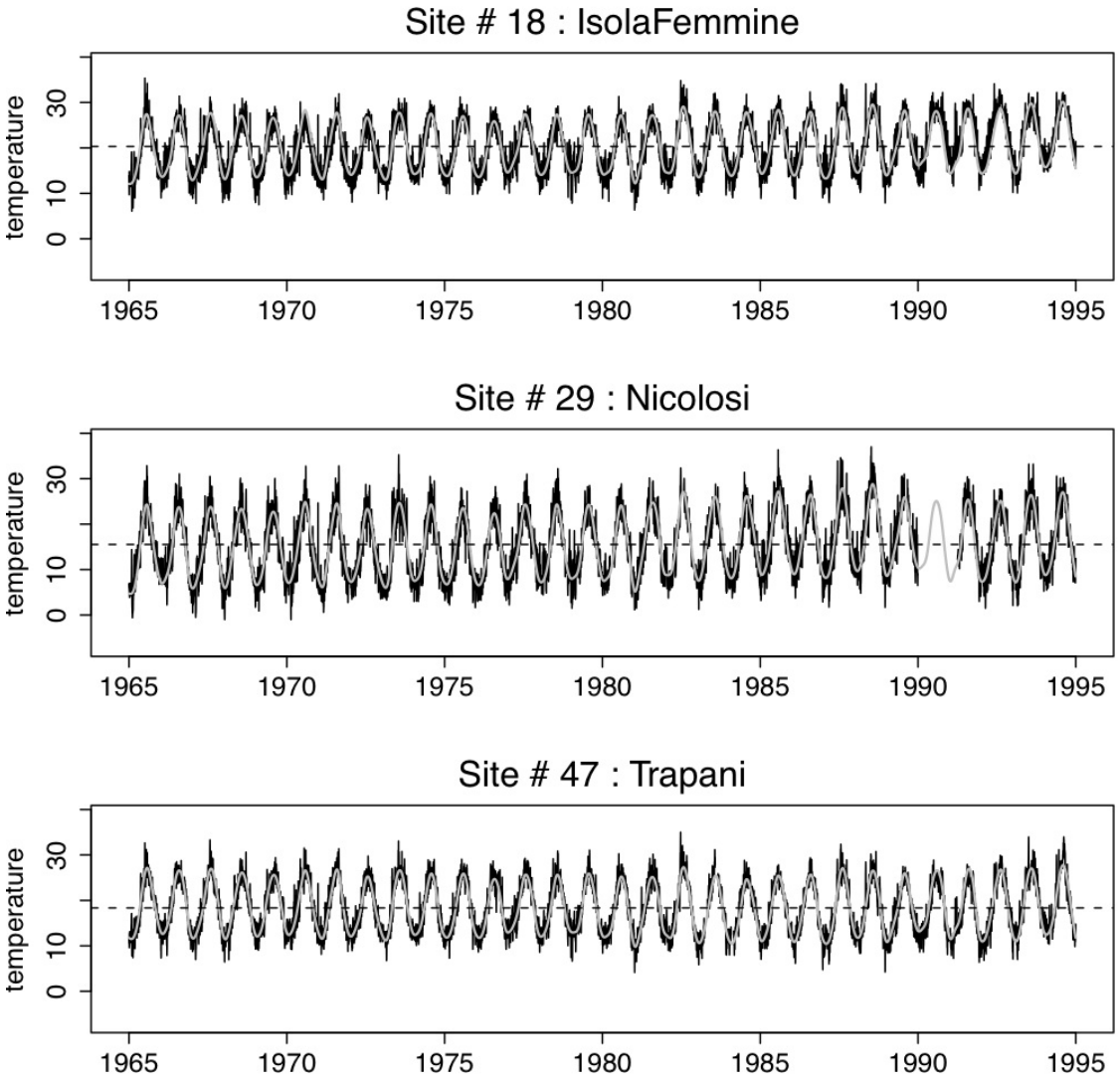

Figure 6: The smoothed trend in grey color calculated for three of the fifty sites (dots representing raw data).

no systematic errors. The four β_i represent the spatially varying trend. Plotting their values on a map (Figure 7) we see some spatial structure present, in particular systematic differences between sites around the coast and sites inland. The long term trend, β_0 highlights a temperature average over 30 years higher around the coast. The other β_i show an opposite sign of variability between inland and coast sites with extremely positive and negative values characterizing amplitude of the temporal basis function at these sites. An extreme positive value for β_2 occurs for the site located in a small island for which the first years of data are missing, and an extreme negative value corresponds to a site located inland, for which the data in the years $1983 - 84$ were taken between two periods of missing data. Generally β_2 has more negative values for sites inland, characterized by a semiarid climate. β_3 exhibits almost constant spatial pattern across the region, with very few extremely positive or negative values for inland and coast sites, respectively.

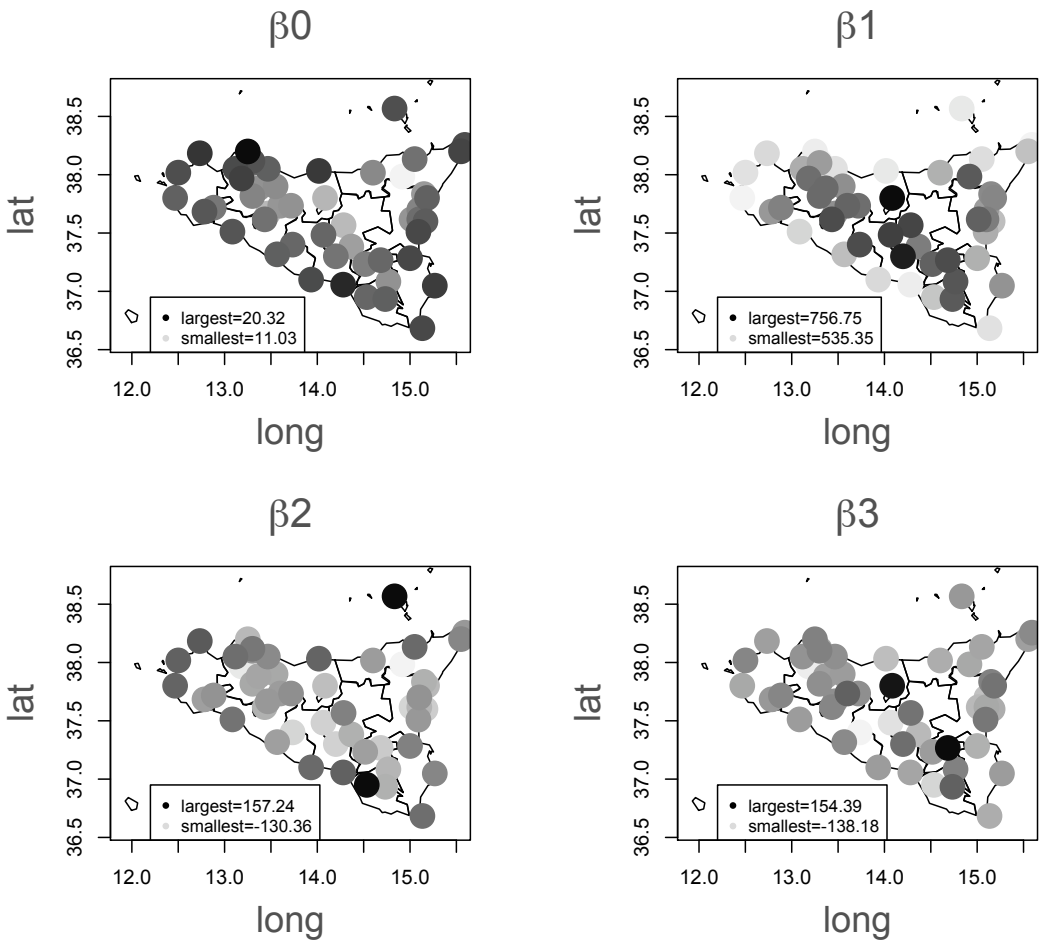

Figure 7: Trend coefficients using a grey scale. Note that the grey scale changes from plot to plot.

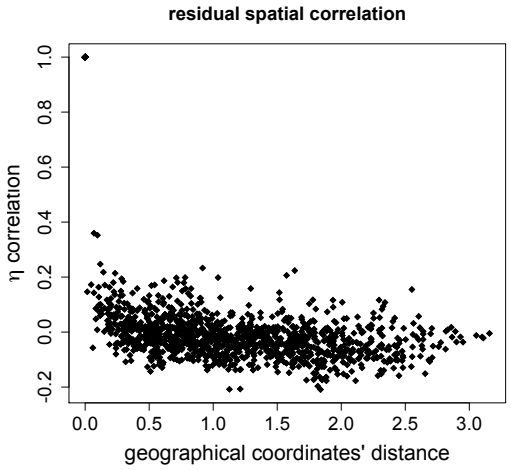

Figure 8: The final residual spatial correlation as a function of geographic distance. Highest correation value is equal to 0.36 and it occurs at a very short distance.

The spatiotemporal residual matrix $\mathbf{E}$ is calculated from the fitted linear models. The SVD decomposition applied to this matrix led us to consider only one component (K=1) because of the relatively high value of the first singular value compared to the rest. The first component explains 59.7% of variability in the residual matrix. The temporal process, $\mathbf{u}^{(2)}$, shows some indication of long memory, but partial autocorrelation coefficients for lags larger then 3 are not significantly different from zero. The spatial processes, $\mathbf{v}^{(2)}$, takes values in the range $(-0.203, -0.078)$ showing little spatial structure.

After removing spatial and temporal structures, an AR(2) process ($L = 2$) seems to explain the remaining pattern for most of the locations, the autoregressive coefficient ϕ_1 takes values in $(0.398, 0.777)$ and ϕ_2 takes values in $(-0.065, 0.198)$.

Because the different altitude of the sites could suggest the need for a nonstationary correlation model we looked at the correlation between the sites against the distance before and after applying the model. The only sites with moderate correlation ($cor \sim 0.36$) from the final residuals matrix, $\eta(t, \mathbf{s})$, are in the middle North of the island and around the coast; these are not the sites with the highest altitude. All the other sites have correlation close to zero (see Figure 8) so it does not seem likely that altitude is leading to nonstationarity. The MSE between data and fitted values equals 103.53 for this dataset.

5. Conclusion and Discussions

The attractiveness of this model is its simple form describing both spatial and temporal patterns. In contrast with more complex temporally dynamic models, our model adopts

the simple perspective of a spatiotemporal process as a sum of products of temporal and spatial processes. It is also able to accommodate large incomplete spatiotemporal data matrices showing good capability to contend with time as well as space. The simulation study gives evidence of good potential to reproduce the signal of interest when the components are well defined, and that missing data do not cause serious bias. Our interpretation of the simulation results is that our methodology is a robust technique, being able to recover the truth even in cases where the assumptions are not quite right. The application of our methodology to a dataset of daily temperatures from Sicily over 30 years shows a good fit, and is a compact description of the spatio-temporal variability using just a few smoothed singular vectors and empirical orthogonal functions reducing the spatial correlation.

Another useful property of this model is its quite fast evaluation. The computation of the model fitted to the Sicily data (10957 daily temperatures across 50 monitoring sites), that consists of three smoothed left singular vectors, one EOF, and autoregressive models of order 2 estimated from the residuals, takes only 14.937 sec on 2.4 GHz Intel Core 2 Duo.

This model is a low dimensional representation of spatio-temporal variability and describes the underlying patterns with just an AR(2) residual temporal process, that seems appropriate for short time scale variation in meteorological data.

References

Banerjee S, Carlin BP, Gelfand AE, Raton B (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall CRC.

Beckers J, Rixen M (2003). "EOF Calculations and Data Filling from Incomplete Oceanographic Datasets." *Journal of Atmospheric and Oceanic Technology*, **20**(12), 1839–1856.

Cattell R (1966). "The scree test for the number of factors." *Multivariate Behavioral Research*, **1**, 245–76.

Fuentes M, Guttorp P, Sampson PD (2007). *Using Transforms to Analyze Space-time Processes*, chapter in Statistical Methods for Spatio-Temporal Systems, pp. 77–149. Taylor& Francis Group.

Meiring W, Guttorp P, Sampson PD (1998). "Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model." *Environmental and Ecological Statistics*, **5**, 197–222.

Wall ME, Rechtsteiner A, Rocha LM (2003). *A Practical Approach to Microarray Data Analysis*, chapter in Singular value decomposition and principal component analysis. Kluwer Academc Publishers.

Affiliation:

Rossella Onorati
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", University of Palermo
Palermo, 90128
E-mail: rossellaonorati@gmail.com

Paul D. Sampson
Department of Statistics, University of Washington
Seattle, WA 98195-4322
E-mail: pds@u.washington.edu
URL: http://www.stat.washington.edu/pds/

Peter Guttorp
Department of Statistics, University of Washington
Seattle, WA 98195-4322
E-mail: peter@u.washington.edu
URL: http://www.stat.washington.edu/peter/

Journal of Environmental Statistics
Volume 5, Issue 2
August 2013

http://www.jenvstat.org
Submitted: 2012-03-12
Accepted: 2013-06-27