

Dimensionality Reduction for large Spatio-Temporal datasets based on SVD

Rossella Onorati ¹, Paul D. Sampson ², Peter Guttorp ³

Dept. of Statistical and Mathematical Science “S. Vianelli”, e-mail:

onorati@dssm.unipa.it, Dept. of Statistics, e-mail: pds@stat.washington.edu, Dept. of Statistics, e-mail: peter@stat.washington.edu

Abstract: Many models for spatio-temporal measurements $Z(s, t)$ can be written as a sum of a systematic component and a residual component: $Z = M + E$. The approach presented here incorporates two Singular Value Decompositions (SVD). The first SVD is applied to the space-time data matrix Z with cross-validation to choose the number of smoothed singular vectors to use as temporal basis functions for modelling spatially varying temporal trend in the matrix M . The second SVD is applied to the spatio-temporal matrix E of residuals from the trend models fitted at each site; it represents spatially correlated short time scale temporal processes. The remaining stochastic structure is explained by simple autoregressive models fit to the final residuals. The procedure is applied to 30 years of daily temperature data from Sicily.

Keywords: Spatio-temporal processes, SVD

1. Introduction

A common problem in the analysis of space-time datasets is to extract the underlying trends and characterize spatio-temporal correlation patterns. Empirical orthogonal function (EOF) analysis is a useful tool for examining the spatio-temporal variation in atmospheric and physical process and an efficient method to define and compute EOFs is the Singular Value Decomposition (SVD). The SVD decomposes a matrix $Z_{[T \times N]}$ as a product of three matrices:

$$Z = UDV' \quad (1)$$

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

The appeal of the singular value decomposition is partly due to its interpretation as a multiplicative model based on row and columns factors. Letting the elements of the temporal singular vectors $\mathbf{u}_j$ be written as $u_j(t)$ and the elements of the spatial singular vectors $\mathbf{v}_j$ be denoted by $v_j(s)$, the spatio-temporal indexed datum $Z(t, s)$, can be written as a sum of products of spatial and temporal variables: $Z(t, s) = \sum_{j=1}^N d_j u_j(t) v_j(s)$. If it suffices to approximate Z by just its first EOF, then we are saying that $Z(t, s) \approx d_1 u_1(t) v_1(s)$, i.e. the spatiotemporal process can be approximated by a separable process (Banerjee *et al.*, 2004).

Our analysis involves two SVD computations and a temporal correlation model. We first model temporal trend, $\mathbf{M}$, by linear combinations of smoothed orthogonal temporal basis

functions derived from the singular value decomposition of $\mathbf{Z}$ 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 $\mathbf{E}$ of residuals from the linear models fitted at each site to the temporal trend basis functions from the first SVD fit to represent 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 proposed model; section 3 presents the application of the model, and conclusions follow in section 4.

2. Model formulation

Denoting by N the number of sites $s_1, \dots, s_N$ and by T 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 $\mathbf{t}$ into the sum of a systematic trend component and residuals

$$Z(t, s) = M(t, s) + E(t, s) \quad (2)$$

where $M(t, s)$ denotes the mean structure and $E(t, s)$ denotes the residual process. We decompose the trend as

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

where $f_j[u_j(t)]$ are orthogonal temporal basis functions (computed as smooth *splines*). J is the number of left singular vectors in the singular value decomposition $Z = UDV'$ and is chosen by cross-validation leaving out one monitoring site at a time. $\beta_{N \times J}$ is a matrix of trend coefficients, estimated by multiple linear regression on the smoothed basis functions f . Because of the spatially varying nature of these coefficients we model them as spatial random fields, $\beta_j \sim N(\mu, \sum_{\beta_j})$, with spatial covariance matrix $\sum_{\beta_j}$ (Fuentes *et al.*, 2007). We may further decompose the spatiotemporal residual matrix $\mathbf{E}$ as a sum of a mean zero spatiotemporal process and a temporally correlated noise process: $w(y, s) + \epsilon(t, s)$. We derive this from a second SVD,

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

where $\mathbf{u}_k^{(2)} \mathbf{v}_k'^{(2)}$ and $d_k^{(2)}$ are the k th EOF and singular value in the singular value decomposition of E . It is a mean-zero spatio-temporal process. $\epsilon(t, s)$ is a column of time series each modeled as autoregressive of order L ,

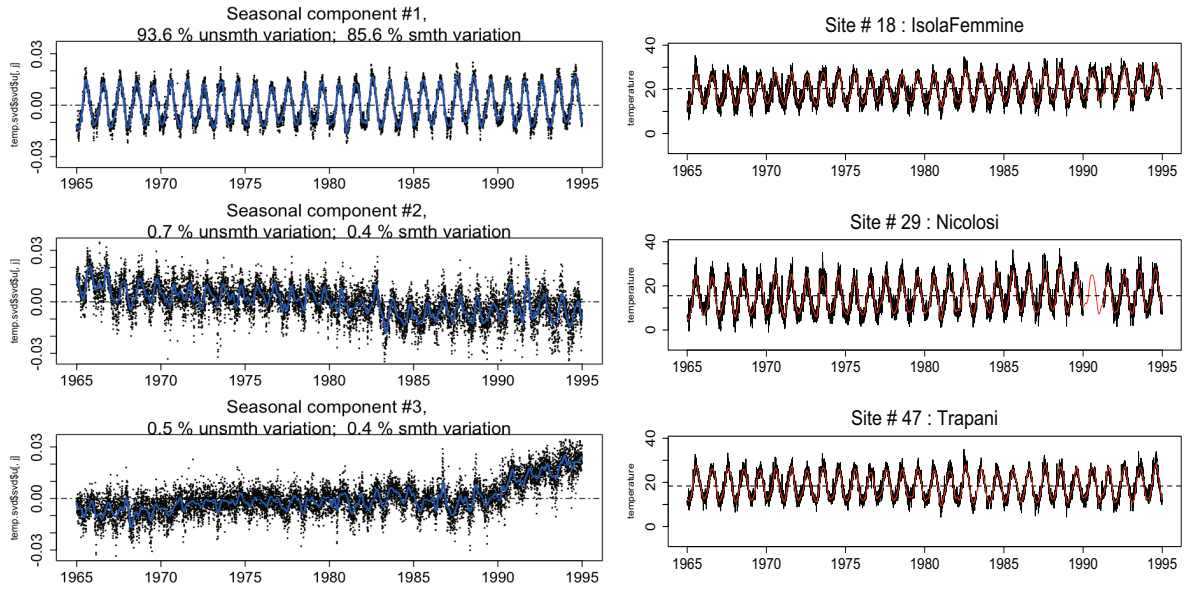
$$\epsilon(s, t) = \sum_{l=1}^L \phi_l(s) \epsilon(s, t-l) + \eta(s, t), \quad (5)$$

for $s = 1, \dots, N$ and $\eta \sim WN(0, \sigma_\eta)$. 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; what is left then is only an autoregressive model.

3. Application

We developed this model for a spatiotemporal dataset of daily temperature averages observed at fifty monitoring sites from 1965 to 1994 in Sicily, the largest island in the Mediterranean. 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 analysis of these data led us to consider three temporal basis functions ($N = 3$). The left top panel of Figure 1 shows that the first function represents 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 explain 85.6% of the total variation. The second and third functions explain very little of the variability left, around 0.5% each, some of these has some structure and long term trend. These are the type of signals we would expect to see from warming of the climate in this region.

Figure 1: This figure shows the three temporal basis functions with the smooths (left) and the smoothed trend calculated for three of the fifty sites (right; dots representing raw data).



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. We use these smoothed components to estimate the β matrix of trend coefficients by a linear regression fitted at each monitoring site. The right panels of Figure 1 illustrate the fit using these three component on a sample of three monitoring sites. The fit appears good with no systematic errors. The four β s represent the spatially varying trend. Plotting the values of β_1 , β_2 and β_3 on a map (Figure 2) we see some spatial structure present, in particular systematic differences between sites around the coast and sites inland.

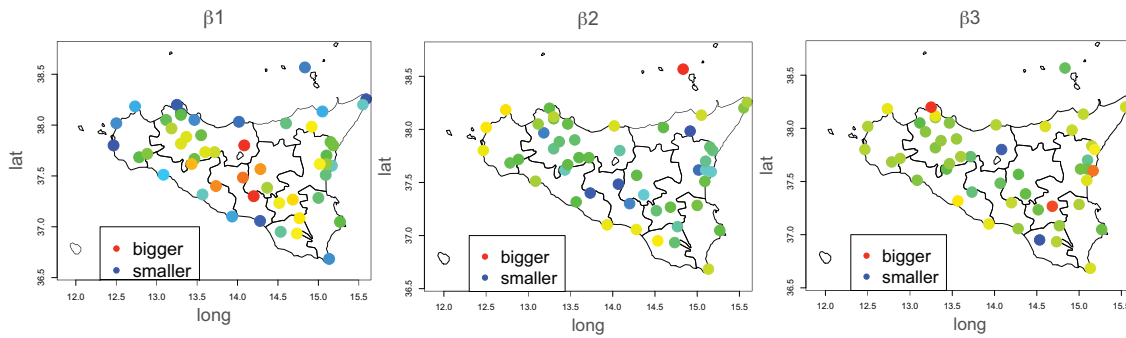
The spatiotemporal residual matrix E is calculated from the fitted linear models. The SVD decomposition applied to this matrix led us to consider only one component ($K=1$) because of the relatively high value of the first singular value compared with the rest. The

first component explains 59.7% of variability in the residual matrix; the temporal process, $u^{(2)}$, show some long memory and partial autocorrelation coefficients for lag larger then 2 are not significantly different from zero, while the spatial processes, $v^{(2)}$, assumes values in the range $(-0.2027, -0.0778)$ showing few 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 $(0.3981, 0.777)$ 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 from the final residuals matrix, $\eta(t, s)$, are in the middle North of the island and close to the volcano; these are not the sites with the highest altitude. All the other sites have correlation close to zero so it does not seem likely that altitude is leading to nonstationarity.

Figure 2: Trend coefficients's values in a color scale of red, yellow, green, blue with $\beta_1 \in (532.94, 745.48)$, $\beta_2 \in (-133.57, 141.35)$ and $\beta_3 \in (-173.54, 105.29)$.



4. Conclusion

The attractiveness of this model consists in its simple form describing both spatial and temporal patterns. In contrast with more complex temporally dynamic models, our model follows the intuitive idea to look at the a spatiotemporal process as products of temporal and spatial processes. It is a low dimensional representation of spatio-temporal variability and seems able to describe the underlying patterns with just an AR(2) residual temporal process that seems appropriate for short time scale variation in meteorological data. The final residuals show little spatial correlation.

References

- Banerjee S., Carlin B.P., Gelfand A.E. and Raton B. (2004) *Hierarchical modeling and analysis for spatial data*, Chapman & Hall CRC.
- Fuentes M., Guttorp P. and Sampson P.D. (2007) *Statistical Methods for Spatio-Temporal Systems*, Taylor& Francis Group, chapter Using Transforms to Analyze Space-time Processes, 77–149.
- Wikle C.K. and Cressie N. (1999) A dimension-reduced approach to space-time kalman filtering, *Biometrika*, 86, 4, 815–829.