Reduced rank covariances for the analysis of environmental data [1]

Orietta Nicolis [2]
Dept. of Information Technology and Mathematical Methods, e-mail:
orietta.nicolis@unibg.it
Doug Nychka [3]
Institute for Mathematics Applied to Geosciences, e-mail: nychka@ucar.edu

Abstract: In this work we propose a Monte Carlo estimator for non stationary covariances of large incomplete lattice data and irregularly distributed observations. In particular, following the multiresolution approach introduced by Nychka *et al.* (2003) and Matsuo *et al.* (2008), we estimate a spatial reduced rank covariance starting from a Matérn model and using the Wendland basis function in its decomposition. The basic idea is to estimate the covariance on a lower resolution grid starting from a stationary model and use the multiresolution property for evaluating the variance of the full data. Since this method doesn't need to compute the wavelet coefficients, it is very fast in estimating covariance in large dataset. The spatial forecasting performances of the method has been described through a simulation study. Finally, the method has been applied to aerosol optical thickness (AOT) satellite data, observed in Northen Italy, for the estimation of missing values and to ozone concentrations for spatial prediction.

Keywords: Monte-Carlo estimator, multi-resolution basis, non-stationary covariance, large data sets.

1. Introduction

The analysis of many geophysical and environmental problems requires the application of interpolation techniques based on the estimation of covariance matrices. Due to the non stationary nature of the data and to the large size of the dataset it is often difficult to use classical covariance models. In this work we proposed a wavelet-based non parametric estimator for computing the covariance matrices of massive data.
Let $\mathbf{y}$ be the m data points of the field on a fine grid and Σ the $(m \times m)$ covariance matrix among grid points. By the multiresolution approach (Nychka *et al.* (2003)), a spatial covariance matrix Σ can be decomposed as

$$\Sigma = WDW^T = WHH^TW^T \tag{1}$$

where W is a matrix of basis functions evaluated on the grid, D is the matrix of coefficients, H is a square root of D, and the apex T denotes transposition. Unlike the eigenvector/eigenvalue decomposition of a matrix, W need not be orthogonal and D need not be diagonal. Since for massive data sets Σ may be very large, some authors (Nychka *et al.*

[1] The present paper is partially supported by PRIN project 2006.

[2] Address of correspondence: University of Bergamo, viale Marconi, 5, 24055 Dalmine, Italy.

[3] Address of correspondence: National Center for Atmospheric Research, Boulder, CO (USA).

(2003)) suggest an alternative way of building the covariance by specifying the basis functions and a matrix H. The basic idea of this work is to estimate in a iterative way the matrix H on a lower resolution grid starting from a stationary model for Σ. The evaluation of the wavelet basis on a fine grid in Eq. (1) provides a reduced rank covariance matrix.

The method can be used for the estimation of covariance functions of irregularly distributed data points and lattice data with many missing values.

In this paper, the multiresolution method based on the reduced rank covariance is applied to two environmental data sets: the AOT satellite data (Nicolis *et al.* (2008)) and to daily ozone concentrations (Nychka (2005)).

Next Section discusses the multiresolution approach for the analysis of observational data. Section 3 describes the Reduced Rank Covariance (RCC) algorithm for the estimation of conditional variance in large data sets. Section 4 shows some simulation results. Applications to satellite and ozone data are described in Section 5. Section 6 presents conclusions and further developments.

2. Modelling observational data

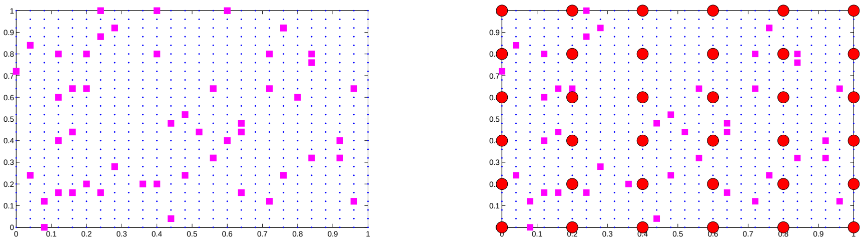
In many geophysical applications the spatial fields are observed over time and one can exploit temporal replication to estimate sample covariances. In this section we focus on this case and also for gridded data with the goal of deriving a estimator that scale to large problems.

Suppose that the point observations $\mathbf{y}$ are samples of a centered Gaussian random field on a fine grid and are composed of the observations at irregularly distributed locations $\mathbf{y}_o$, and the missing observations $\mathbf{y}_m$. In other words, we assume that the grid is fine enough in resolution so that any observation can registered on the grid points (as in Figure 1 (a)). Hence,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_o \\ \mathbf{y}_m \end{pmatrix} \sim MN(0, \Sigma) \tag{2}$$

where $\Sigma = WDW^T$ is the covariance fine gridded data described in Eq. (1), D is a non diagonal matrix and W is a matrix of non-orthogonal scaling and wavelet functions.

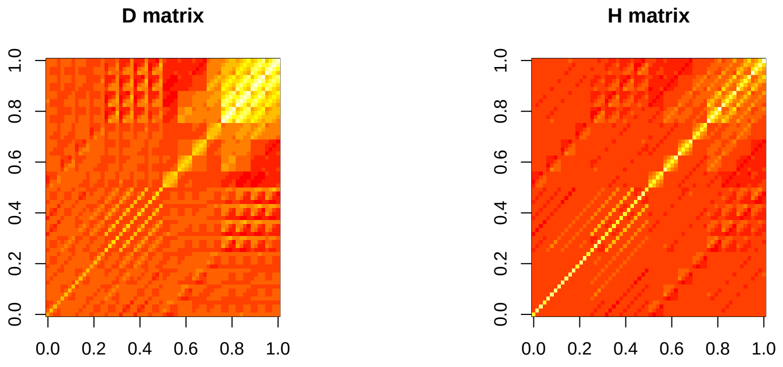
Figure 1: *Gridded data: irregularly distributed data (squares); missing data (blue circles) and knots (red circles).*



(a) (b)

Although the non-orthogonality property of wavelet basis provide off diagonal coefficients in the matrix D, the localized support of these functions ensures that many covariance terms in D will be close to zero, reducing the computational complexity in the interpolation problems of surfaces. An important class of compactly supported basis function, used in the applications of this work, is the Wendland family proposed by Wendland (1995).

Figure 2: *Example of D (a) and H matrix (b) on a 8×8 grid data (b) using wavelet-based non stationary covariance.*



(a) (b)

The Wendland functions are also implemented in the R statistical language (http://www.r-project.org) in the fields package (Nychka (2005)). Figure 2 provides an example for the matrices D and H, obtained by Eq.(1),

$$D = W^{-1} \Sigma W^{-T}, \tag{3}$$

where Σ is the covariance resulting from the fitting of a Matérn model to a regular grid, W is a matrix whose columns contain Wendland functions, and W^{-T} is the transpose of the inverse of W.

The observational model (2) can be written as $\mathbf{z}_o = K\mathbf{y} + \varepsilon$ where ε is a multivariate normal $MN(0, \sigma^2 I)$, $\mathbf{z}_o$ is a vector of m observations, and $\mathbf{y}$ is the underlying spatial field on the grid. The matrix K denotes an incidence matrix of ones and zeroes with a single one in each row indicating the position of each observation with respect to the grid. The conditional distribution of $\mathbf{y}$ given $\mathbf{z}_o$ is Gaussian with mean

$$\Sigma_{o,m} (\Sigma_{o,o})^{-1} \mathbf{z}_o \tag{4}$$

and variance

$$\Sigma_{m,m} - \Sigma_{m,o} (\Sigma_{o,o})^{-1} \Sigma_{o,m} \tag{5}$$

where $\Sigma_{o,m} = W_o H H^T W_m^T$ is the cross-covariance between observed and missing data, $\Sigma_{o,o} = W_o H H^T W_o^T + \sigma^2 I$ is covariance of observed data and $\Sigma_{m,m} = W_m H H^T W_m^T$ is the covariance of missing data. The matrices W_o and W_m are wavelet basis evaluated at the observed and missing data, respectively.

For a chosen multiresolution basis and a sparse matrix H there are fast recursive algorithms for computing the covariance Σ. Motsuo et al. (2008) proposed a method that allows for sparse covariance matrices for the basis coefficients. However the evaluation of wavelet coefficients can be slow for large data sets.

3. The Reduced Rank Covariance (RRC) method

In this section we propose an estimation method for Σ based on the evaluation of a reduced rank matrices. We denote by "*knots*" the spatial points on a lower resolution grid $\mathcal{G}$ of size $(g \times g)$, where $g \leq m$. The idea is to estimate the matrix H on the grid of knots starting from a stationary model for Σ and using the Monte Carlo simulation for providing an estimator for the conditional covariance. A flexible model of stationary covariance is the Matérn covariance given by

$$C(h) = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu} \frac{h}{\theta} \right) \mathcal{K}_\nu \left(2\sqrt{\nu} \frac{h}{\theta} \right), \quad \theta > 0, \nu > 0$$

where h is the distance, θ is the spatial range and $\mathcal{K}_\nu(\cdot)$ is the Bessel function of the second kind whose order of differentiability is ν (smoothing parameter). Since W is fixed for a chosen basis, the estimation procedure for the conditional covariance is given by the estimation of the matrix H after a sequence of approximations. Following this approach the covariance in Eq. 1 can be approximated as

$$\Sigma \approx W \tilde{H}_g \tilde{H}_g^T W^T, \tag{6}$$

where $\tilde{H}_g$ is an estimate of the matrix H on the grid $\mathcal{G}$.
The RRC estimation algorithm can be described by the Monte Carlo EM algorithm in the following steps.

1. Find Kriging prediction on the grid $\mathcal{G}$:

   $$\hat{\mathbf{y}}_g = \Sigma_{o,g} (\Sigma_{o,o})^{-1} \mathbf{z}_o,$$

   where $\tilde{H}_g = (W_g^{-1} \Sigma_{g,g} W_g^T)^{1/2}$ and $\Sigma_{g,g}$ is stationary covariance model (es. Matern).
2. Generate synthetic data: $\mathbf{z}_o^s = K \mathbf{y}_g^s + \varepsilon$ where $\mathbf{y}_g^s = W_g \tilde{H}_g a$ with $a \sim N(0,1)$.
3. Compute Kriging errors:

   $$\mathbf{u}^* = \mathbf{y}_g^s - \hat{\mathbf{y}}_g^s,$$

   where $\hat{\mathbf{y}}_g^s = \Sigma_{o,g} (\Sigma_{o,o})^{-1} \mathbf{z}_o^s$.
4. Find conditional field $\mathbf{y}_m | \mathbf{z}_o$:

   $$\hat{\mathbf{y}}_u = \hat{\mathbf{y}}_g + \mathbf{u}^*.$$

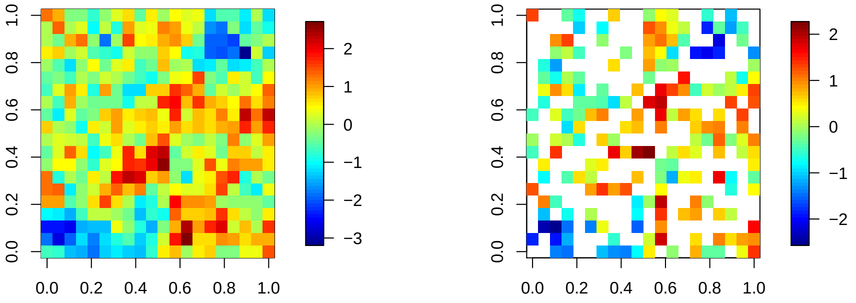
5. Compute the conditional covariance on T replications, $\Sigma_u = COV(\hat{\mathbf{y}}_u)$ and use the new $\tilde{H}_g$ in the step 1.

Performing this several times will give an ensemble of fields and, of course, finding the sample covariance across the ensemble provides a Monte Carlo based estimate of the conditional covariance.

4. Simulation study

The purpose of this study is to investigate the forecasting ability of the proposed RRC method in two different contexts: (i) approximation of stationary covariance models and (ii) estimation of non-stationary covariances.

Figure 3: *Simulated Gaussian random field on a* 20×20 *grid with Matèrn covariance* ($\theta = 0.1$ *and* $\nu = 0.5$) *without (a) and with (b) 50% of missing values.*

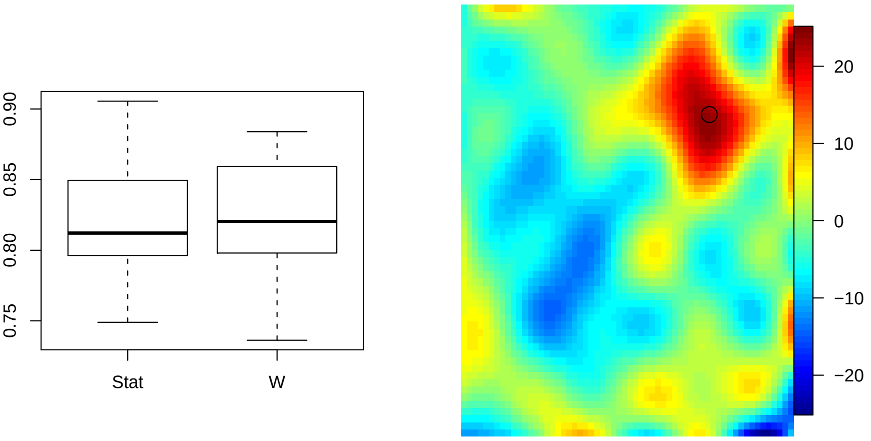


(a)

(b)

In order to study the properties of approximation of the RRC method, we simulated $n = 20$ Gaussian random fields on a 20×20 grid using a Matèrn model with parameters $\theta = 0.1$ and $\nu = 0.5$. In order to generate the missing data we removed randomly 50% of the simulated data. An example of simulated random field with missing data is shown in Figure 3.

Figure 4: *(a) RMSE of the 50% of missing values. The estimates are obtained using a Matèrn model (Stat) and RRC method (W) with five iterations; (b) covariance between the point indicated by black circle and the rest of grid points.*



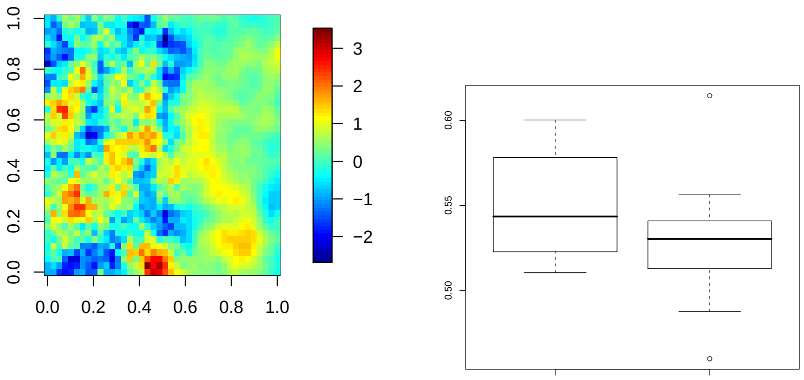
(a)

(b)

For each simulated random field we estimated the missing data using the RRC method on a grid of 8×8 knots and then we computed the root mean square errors on the predictions. The parameters of the Matèrn model used in the step 1. of the algorithm has been chosen by cross validation.

Figure 4 (a) compares the RMSE for each simulated random field for the Matèrn model and the non-stationary wavelet-based covariance model (RRC). The similarity of the two boxplots indicates a good approximation of the proposed method to the stationary model. The covariance between a specific point and the rest of grid points shown in Figure 4 (b) highlight the higher correlation between neighboring points.

Figure 5: *Non-stationary random field simulated on a* 40×40 *grid with Matèrn covariance (*$\theta = 0.1$ *and* $\nu = 0.5$*) (a) and RMSE results for 50% of missing values using a Matérn model and a RRC method on a grid of* 8×8 *knots.*



(a)

(b)

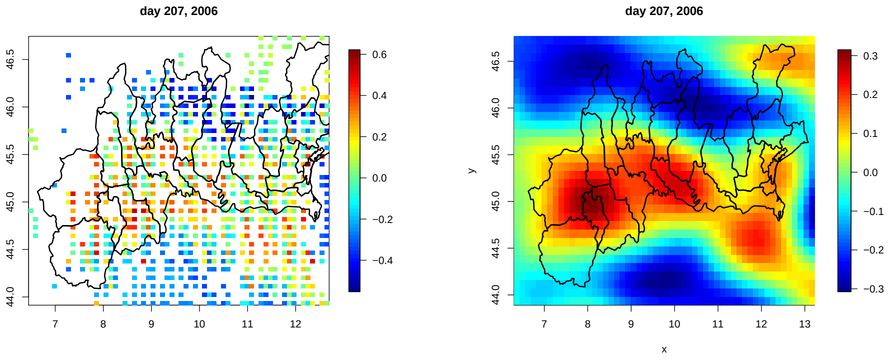
In order to estimate non-stationary covariance by RRC method we generated $n = 20$ synthetic non-stationary data on a grid of 40×40 points with 50% of missing values. These spatial data has been obtained from a mixture of two dependent stationary random fields as in Matsuo *et al.* (2008) with Matèrn covariances ($\Sigma_1(\nu = 1, \theta = 0.125)$ and $\Sigma_2(\nu = 0.5, \theta = 0.1)$), and a weight function $w(\mathbf{u}) = \Phi((u_x - 0.5)/.15)$ where $\Phi(\cdot)$ is the normal cumulative distribution function and u_x is the horizontal coordinate. Figure 5 (a) shows an example of simulated non-stationary random field. The RRC method with a grid of 8×8 knots has been applied for forecasting 50% of the missing values. In this case the RMSE results (Figure 5 (b)) indicates that RRC method provides better estimates than a stationary model.

5. Applications

5.1 Satellite data

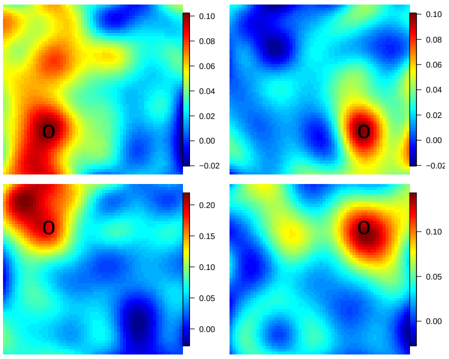
Satellite data are very important in the analysis of environmental data as they cover wide monitoring areas and can sometimes be easily downloaded from specialized Internet websites. However, their statistical analysis often requires special techniques to cope with large data sets or to treat other exogenous variables that affect the satellite measurements in the atmosphere. The satellite aerosol optical thickness (AOT) data is an example. The study of these measurements is particularly important to assess the amount of fine particulate matters in the air and the consequent risk to human health. However, the meteorolog-

Figure 6: *Satellite AOT measurements (a) and kriging predictions (b) using RRC method on a grid of* 8×8 *knots.*



(a) (b)

Figure 7: *Non-stationary covariance* $W\tilde{H}_g\tilde{H}_g^TW^T$ *obtained after five iterations of MC simulations. Each panel shows the covariance between the point indicated by black circle and the rest of grid points.*



ical conditions determine the AOT retrieval since the availability of the data is restricted to cloud-free conditions. Indeed, the cloudy coverage causes a large number of missing data, sometimes making problematic the application of traditional correlation models.
In this work we consider the column aerosol optical thickness (AOT)[4] data derived from the Moderate Resolution Imaging SpectroRadiometer (MODIS) on the Terra/Aqua satellites in Northern Italy for the period April 1st - June 30th, 2006 (Nicolis *et al.* (2008)). The Terra satellite crosses Europe near 10:30 local solar time (morning orbit), while Aqua crosses Europe near 13:30 local solar time (afternoon orbit). Hence, at least two observations of any place in Europe are obtained per day during daylight hours. These data are based on analyzing 20×20 pixels at 500 m resolution and reported at 10×10 km^2 resolution. The reflectivity measured at 2.1 μm at the top-of-atmosphere is used to infer surface reflectivity at that wavelength. Fig.6 (a) shows the daily average of AOT data on July 26, 2006 in Northen Italy measured on a grid of 54×32 locations. Figure 7 shows the non-stationary structure of the estimated conditional covariance, $W\tilde{H}_g\tilde{H}_g^TW^T$, for four

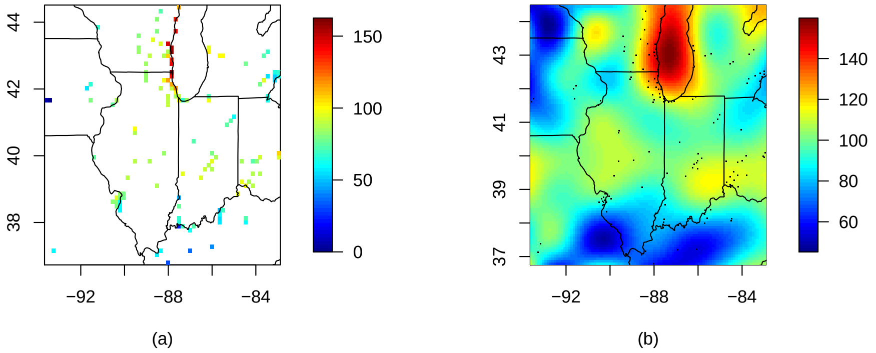
[4] AOT measurements can be downloaded from the NASA web page http://disc.sci.gsfc.nasa.gov/.

points obtained after five iterations of MC simulations using a grid of 8×8 knots. It is important to note that the correlation structure is slightly different for the four graphs which indicates that the model is able to capture the non-stationarity of data.
The Kriging prediction for the day July 26, is shown in Figure 6 (b). These results indicate that higher estimates of AOT values are in the Western part of Northern Italy around the cities of Turin and Milan. Similar results were found by Fassó *et al.* (2007) and **?** for the analysis of fine particulate matters (PM_10) in Northen Italy.

5.2 Ozone data

In order to apply the RRC method to irregular data, we considered ozone concentrations, included in `Fields` package of `R` software (Nychka (2005)).

Figure 8: *(a) Daily ozone concentrations on June 18, 1987; (b) Kriging map using the RRC method on a grid of* 8×8 *knots.*



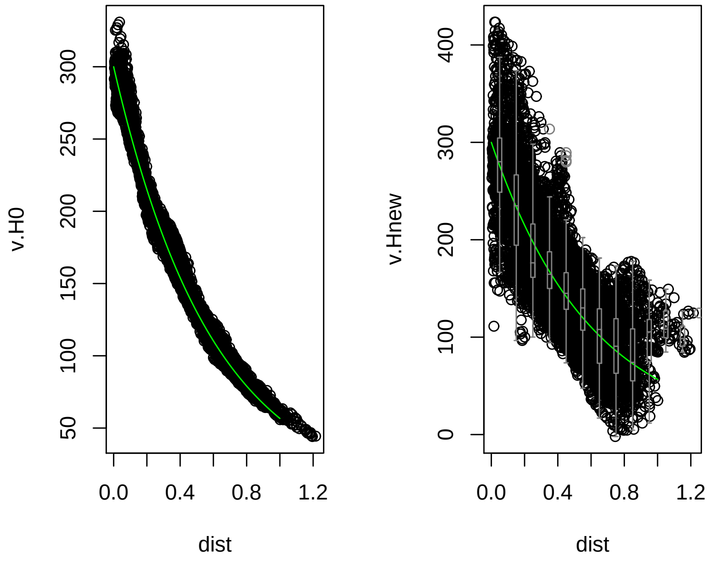
(a) (b)

The database consists of daily 8-hour average ozone concentration measured in parts per billion (PPB) for 153 sites in the Midwestern US over the period June 3,1987 through August 31, 1987 (89 days). Many of these station have incomplete records both in time and space. Figure 8 shows ozone concentrations on July 19, 1987. The application of RRC method allows to estimate the ozone concentrations on a fine grid with resolution (100×100) using a lower resolution grid of 8×8 knots. Figure 9 shows the non-stationary of the estimated reduced rank covariance.

6. Conclusions and further developments

We have proposed a practical method for approximating stationary covariance models and estimating non-stationary covariances. The method based on empirical estimation the reduced rank matrix provides an efficient tool for handling large data sets.
Although this method is still in its preliminary stages, the results from the simulation study are very encouraging. We believe that RRC method can be used for a preliminary analysis

Figure 9: *Covariances vs distances: (a) Reduced rank covariance after one iteration (circles) and Matérn covariance (green line); (b) Reduced rank covariance after five iteration (circles) and Matérn covariance (green line). Boxplot in (b) indicates the distribution of the estimated reduced rank covariance.*



of the spatial covariance structure and can be developed for the covariance estimation of space time models. We also intend to find a parametrization for the RRC matrix and using an EM algorithm for the estimation of the model with missing data.

References

Fassó A., Cameletti M. and Nicolis O. (2007) Air quality monitoring using heterogeneous networks, *Environmetrics*, 18.

Kwong M.K. and Tang P.T.P. (1994) W-matrices, nonorthogonal multiresolution analysis, and finite signals of arbitrary length, *Argonne National Laboratory Technical Report*.

Matsuo T., Nychka D. and Paul D. (2008) Nonstationary covariance modeling for incomplete data: Smoothed monte carlo em approach, *In review*.

Nicolis O., Fassò A. and Mannarini G. (2008) Aot calibration by spatio-temporal model in northern italy, *Spatio-Temporal Modelling (METMA 5), 24-26 September, Alghero, Sardinia*.

Nychka D. (2005) Fields: Tools for spatial data, *National Center for Atmospheric Research, Boulder, CO*.

Nychka D., Wikle C. and Royle J.A. (2003) Multiresolution models for nonstationary spatial covariance functions, *Statistical Modeling*, 2, 315–332.

Sahu S. and Nicolis O. (2008) An evaluation of european air pollution regulations for particulate matter monitored from a heterogeneous network, *Environmetrics*, 20, 943–

961.

Wendland H. (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree, *Adv. Comput. Math.*, 4, 389–396.