SpatioTemporal: An R Package for Spatio-Temporal Modelling of Air-Pollution

Johan Lindström
Lund University & University of Washington

Adam Szpiro
University of Washington

Paul D. Sampson
University of Washington

Silas Bergen
University of Washington

Lianne Sheppard
University of Washington

Abstract

Modelling of Gaussian spatio-temporal processes provide ample opportunity for different model formulations, however two principal directions have emerged. The data can be modelled either as a set of spatially varying temporal basis functions or as spatial fields evolving in time. This package provides maximum-likelihood estimation and cross-validation tools for the first case. Development of the package was motivated by the need to provide accurate spatio-temporal predictions of ambient air pollution at small spatial scales for a health effects study.

The package provides tools for extracting temporal basis functions from the data. It handles incomplete and highly unbalanced spatio-temporal sampling designs and allows for a flexible set of covariates and covariance structures to capture the spatial variability in the temporal basis functions and in the spatio-temporal residuals. Further, the package provides bias corrected predictions for log-transformed data, cross-validation tools and rudimentary MCMC-routines to asses the modelfit.

Here we describe the package, providing a brief summary of the theory, but focusing our attention on an example illustrating how the package can be used for model fitting and cross-validation analysis; the example is based on data included in the package.

Keywords: Spatio-temporal modelling, likelihood based estimation, cross-validation, air pollution, NO_x smooth EOFs, log-Gaussian process, unbalanced data.

1. Introduction

R (R Core Team 2013) package **SpatioTemporal** provides functions for fitting and evaluation of a class of Gaussian spatio-temporal processes that are based on spatially varying temporal basis functions, with spatially correlated residuals.

Development of **SpatioTemporal** was motivated by the need, in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air), for accurate predictions of ambient air pollution. MESA Air is a cohort study funded by the Environmental Protection Agency (EPA) with the aim of assessing the relationship between chronic exposure to air pollution and the progression of sub-clinical cardiovascular disease (Kaufman *et al.* 2012). A primary focus of the MESA Air study is the development of accurate predictions of ambient air pollution (Bild

et al. 2002; Kaufman *et al.* 2012) — primarily gaseous oxides of nitrogen (NO_x) and particulate matter with aerodynamic diameter less than 2.5 μm ($PM_{2.5}$) — at the home locations of study participants in six major US metropolitan areas: Los Angeles, CA; New York, NY; Chicago, IL; Minneapolis-St. Paul, MN; Winston-Salem, NC; and Baltimore, MD.

To fullfill the prediction needs of MESA Air the spatio-temporal model implemented in this package has been developed in a series of papers (Szpiro *et al.* 2010; Sampson *et al.* 2011; Lindström *et al.* 2011, 2013). The model is based on the notion of spatially varying smooth temporal basis functions (see e.g. Sec. 3 in Fuentes *et al.* 2006), and represents one of many different ways that spatio-temporal dependencies can be modelled.

Several general overviews of statistical modeling approaches for spatially and spatio-temporally correlated data exist (Banerjee *et al.* 2004; Cressie and Wikle 2011), including non-separable spatio-temporal covariance functions (Gneiting and Guttorp 2010) and dynamic model formulations (Gamerman 2010). There are also several methods developed specifically for the modelling of air pollution data (Smith *et al.* 2003; Sahu *et al.* 2006; Calder 2008; Fanshawe *et al.* 2008; Paciorek *et al.* 2009; De Iaco and Posa 2012). Additionally, other R packages that handle spatio-temporal models and data are summarised in the relevant task view (`http://CRAN.R-project.org/view=SpatioTemporal`) on the Comprehensive R Archive Network (CRAN) `http://CRAN.R-project.org`

The main purposes of this paper are to: 1) introduce **SpatioTemporal** to R users, 2) present details regarding the model and its implementation necessary for users to fully understand and analys output from the package, 3) demonstrate the use of **SpatioTemporal** by analysing an example data-set, and 4) provide an outlook of future features that may be added to the package.

This paper starts with an outline of the model and theory in Section 2, including details regarding the smooth temporal basis functions (Section 2.1), estimation (Section 2.2), prediction (Section 2.3), and cross-validation (Section 2.4). This is followed by a description of key package features and assumptions in Section 3 and an example analysis of Los Angeles NO_x data in Section 4. Section 5 concludes with an outlook towards possible future features.

Results in this paper were obtained using version 1.1.7 of **SpatioTemporal**. The current version of the package can be obtained from CRAN at `http://CRAN.R-project.org/package=SpatioTemporal`.

A longer vignette providing detailed descriptions of function outputs and elaborating on features not covered here (e.g. simulation) is included as `vignette("ST_tutorial", package="SpatioTemporal`

2. Model and Theory

We are interested in models of the form

$$y(s,t) = \mu(s,t) + \nu(s,t), \tag{1}$$

where $y(s,t)$ denotes the spatio-temporal observations, $\mu(s,t)$ is the structured mean field, and $\nu(s,t)$ is the space-time residual field.

The mean field is modelled as

$$\mu(s,t) = \sum_{l=1}^{L} \gamma_l \mathcal{M}_l(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t), \tag{2}$$

where the $\mathcal{M}_l(s,t)$ are spatio-temporal covariates; γ_l are coefficients for the spatio-temporal covariates; $\{f_i(t)\}_{i=1}^m$ is a set of (smooth) temporal basis functions, with $f_1(t) \equiv 1$; and the $\beta_i(s)$ are spatially varying coefficients for the temporal functions.

The $\beta_i(s)$-coefficients in (2) are treated as spatial fields with a universal kriging structure, allowing the temporal structure to vary between locations:

$$\beta_i(s) \in \mathsf{N}\left(X_i\alpha_i, \Sigma_{\beta_i}(\theta_i)\right) \quad \text{for } i = 1, \ldots, m, \tag{3}$$

where X_i are $n \times p_i$ design matrices, α_i are $p_i \times 1$ matrices of regression coefficients, and $\Sigma_{\beta_i}(\theta_i)$ are $n \times n$ covariance matrices. The X_i matrices often contain geographical covariates and we dentote this component a "land use" regression (LUR). This structure allows for different covariates and covariance structures in the each of the $\beta_i(s)$ fields; the fields are assumed to be apriori independent of each other.

The residual space-time field, $\nu(s,t)$, is assumed to be independent in time with stationary, parametric spatial covariance

$$\nu(s,t) \in \mathsf{N}\left(0, \underbrace{\begin{bmatrix} \Sigma_\nu^1(\theta_\nu) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_\nu^T(\theta_\nu) \end{bmatrix}}_{\Sigma_\nu(\theta_\nu)}\right), \tag{4}$$

or

$$\nu(s,t) \in \mathsf{N}\left(0, \Sigma_\nu^t(\theta_\nu)\right) \text{ for } t = 1, \ldots, T \qquad \text{and} \qquad \nu(s_1,t_1) \perp \nu(s_2,t_2),\ t_1 \neq t_2.$$

Here the size of each block matrix, $\Sigma_\nu^t(\theta_\nu)$, is the number of observations, n_t, at each time-point. Conceptually the ν-field consists of a correlated component, ν^*, and an uncorrelated nugget-effect comprising small-scale variability and measurement errors, ν_{nugget}, i.e.:

$$\nu(s,t) = \nu^*(s,t) + \nu_{\text{nugget}}(s,t) \qquad \text{and} \qquad \Sigma_\nu = \Sigma_\nu^* + \Sigma_{\nu,\text{nugget}}, \tag{5}$$

where $\Sigma_{\nu,\text{nugget}}$ is a diagonal matrix.

The temporal independence in (4) is based on the assumption that the temporal basis, $\{f_i(t)\}_{i=1}^m$, in (2) accounts for the temporal correlation in data. A summary of the notation can be found in Table 1.

To simplify the model we introduce a sparse $N \times mn$-matrix $F = (f_{st,is'})$ with elements

$$f_{st,is'} = \begin{cases} f_i(t) & s = s', \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

along with the $N \times 1$-vectors $Y = y(s,t)$, $V = \nu(s,t)$, and $\mathcal{M}_l(s,t)$ by stacking the elements into single vectors varying first s and then t, i.e. (assuming that the corresponding observations exist)

$$Y = \begin{bmatrix} y(s_1,1) & y(s_2,1) & \cdots & y(s_1,2) & y(s_2,2) & \cdots & y(s_n,T) \end{bmatrix}^\top.$$

Table 1: Important notation and symbols

Symbol	Meaning
$y(s,t)$	Spatio-temporal observations.
$y_u(s,t)$	Spatio-temporal process at the un-observed locations/times.
$y^*(s,t)$	Smoothed version of the spatio-temporal process (i.e. excl. nugget).
$z(s,t)$	The log-Gaussian process, $z(s,t) = \exp(y(s,y))$.
$\mu(s,t)$	Mean field part of $y(s,t)$.
$\nu(s,t)$	Space-time residual part of $y(s,t)$.
$f_i(t)$	Smooth temporal basis functions.
$\beta_i(s)$	Spatially varying regression coefficients, weighing the i:th temporal basis differently at each location.
X_i	Land use regression (LUR) basis functions for the spatially varying regression coefficients in $\beta_i(s)$.
α_i	Regression coefficients for the i:th LUR-basis.
$\mathcal{M}_l(s,t)$	Spatio-temporally varying covariates.
γ_l	Regression coefficient for the spatio-temporally varying covariates.
N	No. of observations.
T	No. of observed time-points.
n	No. of observed locations.
n_t	No. of observations at time t ($N = \sum_{t=1}^{T} n_t$ and $n_t \leq n\ \forall t$.
m	No. of temporal basis functions (incl. intercept).
L	No. of spatio-temporal model outputs.
p_i	No. of LUR-basis functions for the i:th temporal-basis.

The unknown regression and covariance parameters of the model are are collected into column vectors

$$\boldsymbol{\gamma} = \begin{bmatrix}\gamma_1 & \cdots & \gamma_L\end{bmatrix}^\top, \qquad \boldsymbol{\alpha} = \begin{bmatrix}\alpha_1^\top & \cdots & \alpha_m^\top\end{bmatrix}^\top, \qquad \theta_B = \{\theta_i\}_{i=1}^m, \qquad \Psi = \{\theta_B, \theta_\nu\},$$

and all spatio-temporal covariates are gathered in a $N \times L$-matrix, $\mathcal{M} = \begin{bmatrix}\mathcal{M}_1 & \cdots & \mathcal{M}_L\end{bmatrix}$. Components of the β_i-fields are assembled into block matrices as

$$B = \begin{bmatrix}\beta_1(s) \\ \vdots \\ \beta_m(s)\end{bmatrix}, \qquad X = \begin{bmatrix}X_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & X_m\end{bmatrix}, \qquad \Sigma_B(\theta_B) = \begin{bmatrix}\Sigma_1(\theta_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_m(\theta_m)\end{bmatrix}. \tag{7}$$

Using these matrices (1) can be written as

$$Y = \mathcal{M}\boldsymbol{\gamma} + FB + V, \quad \text{where} \quad B \in \mathsf{N}\left(X\boldsymbol{\alpha}, \Sigma_B(\theta_B)\right) \quad \text{and} \quad V \in \mathsf{N}\left(0, \Sigma_\nu(\theta_\nu)\right). \tag{8}$$

Since (8) is a linear combinations of independent Gaussians we introduce the matrices

$$\widetilde{X} = \begin{bmatrix}\mathcal{M} & FX\end{bmatrix} \qquad \text{and} \qquad \widetilde{\Sigma}(\Psi) = \Sigma_\nu(\theta_\nu) + F\Sigma_B(\theta_B)F^\top, \tag{9}$$

and write the distribution of Y as

$$[Y|\Psi, \boldsymbol{\gamma}, \boldsymbol{\alpha}] \in \mathsf{N}\left(\widetilde{X}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\alpha}\end{bmatrix}, \widetilde{\Sigma}(\Psi)\right). \tag{10}$$

Having defined the model the following Sections discuss: 2.1) specification of the (smooth) temporal basis functions, 2.2) parameter estimation, 2.3) prediction, and 2.4) model evaluation through cross-validation.

2.1. Smooth Temporal Functions

The objective of the smooth temporal basis functions, $f_i(t)$, is to capture the temporal variability in the data. These functions can either be specified as *deterministic functions*, or obtained as *smoothed singular vectors* (See Fuentes *et al.* 2006, for details.).

To derive the $m-1$ smoothed singular vectors ($m-1$ since $f_1(t) \equiv 1$) we first construct the $T \times n$ data matrix

$$D(t,s) = \begin{cases} y(t,s), & \text{if the observation } y(t,s) \text{ exists,} \\ \text{NA}, & \text{otherwise,} \end{cases} \tag{11}$$

and fill in missing observations using the algorithm described by Fuentes *et al.* (2006):

Step 0 Centre and scale each column (to mean zero, variance one) and compute the mean of all available observations for each time-point, $u_1(t)$. Missing values in $D(t,s)$ are then imputed using fitted values from a linear regression where each column of $D(t,s)$ is regressed onto u_1. For this step to be well defined the data matrix must have *at least one* observation in each *row and column*.

Step 1. Compute the SVD (singular value decomposition) of the new data matrix with the missing values imputed.

Step 2. Do regression of each column of the new data matrix on the first $m-1$ orthogonal basis functions from Step 1. The missing values are then replaced by the fitted values of this regression.

Step 3. Repeat from Step 1 until convergence; convergence being measured by the change in the imputed values between iterations.

Having imputed the missing values in $D(t,s)$ we then use splines to smooth the leading $m-1$ singular vectors, i.e. the $m-1$ first columns of U in the SVD: $D = USV^\top$.

Cross-validation can be used to determine the number of smooth temporal basis functions needed to capture the temporal variability in data. In a cross-validation the j^{th} column of $D(t,s)$ is held out, smooth temporal functions are computed for the reduced matrix as above, and the functions are evaluated by how well they explain the held out j^{th} column of $D(t,s)$. Repeating for all columns in $D(t,s)$ we obtain a set of regression statistics describing how well the left out columns are explained by smooth temporal functions based on the remaining columns. The computed statistics — mean squared errors (MSE), R^2, AIC (Akaike information criterion), and BIC (Bayesian information criterion) – together with correlation analysis of temporal residuals (recall the assumption of temporal independence in (4)) are

used to determine a suitable number of temporal basis functions; an example is provided in Section 4.3.

2.2. Parameter Estimation

Parameter estimates are obtained by maximising the log-likelihood of (10)

$$\begin{aligned} 2l(\Psi, \alpha, \gamma|Y) = &- N\log(2\pi) - \log\left|\widetilde{\Sigma}(\Psi)\right| \\ &- \left(Y - \widetilde{X}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\alpha}\end{bmatrix}\right)^\top \widetilde{\Sigma}^{-1}(\Psi)\left(Y - \widetilde{X}\begin{bmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\alpha}\end{bmatrix}\right). \end{aligned} \tag{12}$$

Here $\widetilde{\Sigma}$ is a dense $N \times N$-matrix and to obtain a computationaly feasable solution that utlizes the block diagonal structure of Σ_ν and Σ_B Lindström *et al.* (2011, 2013) simplified the log-likelihood by application of various matrix identities, including the Woodbury identity (Thm. 18.2.8 Harville 1997)

$$\widetilde{\Sigma}^{-1} = \Sigma_\nu^{-1} - \Sigma_\nu^{-1}F\left(\Sigma_B^{-1} + F^\top\Sigma_\nu^{-1}F\right)^{-1}F^\top\Sigma_\nu^{-1}, \tag{13}$$

and replaced $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ with the generalised least squares estimates

$$\begin{bmatrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{bmatrix} = \left(\widetilde{X}^\top\widetilde{\Sigma}^{-1}\widetilde{X}\right)^{-1}\left(\widetilde{X}^\top\widetilde{\Sigma}^{-1}Y\right). \tag{14}$$

Introducing the matrices

$$\Sigma_{B|Y}^{-1} = \Sigma_B^{-1} + F^\top\Sigma_\nu^{-1}F, \tag{15a}$$

$$\Sigma_{\alpha|Y}^{-1} = X^\top\Sigma_B^{-1}X - X^\top\Sigma_B^{-1}\Sigma_{B|Y}\Sigma_B^{-1}X, \tag{15b}$$

$$\begin{aligned} \widehat{\Sigma} = &\Sigma_\nu^{-1} - \Sigma_\nu^{-1}F\Sigma_{B|Y}F^\top\Sigma_\nu^{-1} \\ &- \Sigma_\nu^{-1}F\Sigma_{B|Y}\Sigma_B^{-1}X\Sigma_{\alpha|Y}X^\top\Sigma_B^{-1}\Sigma_{B|Y}F^\top\Sigma_\nu^{-1}, \end{aligned} \tag{15c}$$

and using (14) the log-likelihood (12) can be replaced by the corresponding profile or restricted maximum log-likelihood (REML):

$$\begin{aligned} 2l_{\text{PROF}}(\Psi|Y) = &- \log|\Sigma_\nu(\theta_\nu)| - \log|\Sigma_B(\theta_B)| - \log\left|\Sigma_{B|Y}^{-1}(\Psi)\right| - Y^\top\widehat{\Sigma}(\Psi)Y \\ &+ Y^\top\widehat{\Sigma}(\Psi)\mathcal{M}\left(\mathcal{M}^\top\widehat{\Sigma}(\Psi)\mathcal{M}\right)^{-1}\mathcal{M}^\top\widehat{\Sigma}(\Psi)Y + \text{const.} \end{aligned} \tag{16}$$

or

$$2l_{\text{REML}}(\Psi|Y) = 2l_{\text{PROF}}(\Psi|Y) - \log\left|\mathcal{M}^\top\widehat{\Sigma}(\Psi)\mathcal{M}\right| - \log\left|\Sigma_{\alpha|Y}^{-1}(\Psi)\right| \tag{17}$$

respectively. Parameter estimates are obtained by maximisation of either (16) or (17) as

$$\widehat{\Psi}_{\text{PROF}} = \underset{\Psi}{\text{argmax}}\ l_{\text{PROF}}(\Psi|Y), \qquad \text{or} \qquad \widehat{\Psi}_{\text{REML}} = \underset{\Psi}{\text{argmax}}\ l_{\text{REML}}(\Psi|Y). \tag{18}$$

2.3. Predictions

Given the structure of the model (1) with a mean component (2) and β-fields (3) the contribution to any predictions of unobserved Y's can be decomposed in to parts due to the regression model: $\mathcal{M}\gamma + FX\alpha$, the β-fields: $\mathcal{M}\gamma + FB$, and the full predictions. These different predictions are illustrated in Figure 1, using data from the example in Section 4. The different predicitons play an important part in model evaluation by highlighting at which level of the model different featurs of the data are captured.

First, given observations, Y, and estimates of the covariance parameters, Ψ, the regression coefficients are given by (14) with variances

$$\mathsf{VAR}\left(\left[\begin{matrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{matrix}\right]\middle| Y, \Psi\right) = \left(\widetilde{X}^\top \widetilde{\Sigma}^{-1} \widetilde{X}\right)^{-1}.$$

Before providing predictions for β and Y some notation is needed. Let $\mathcal{M}_u$, X_u and F_u denote spatio-temporal covariates, geographic covariates, and temporal basis funcions at unobserved locations/times. Further, B_u denotes the collection of β-fields at the unobserved locations, $\Sigma_{B,uo}$ and $\Sigma_{\nu,uo}$ are the cross-covariance matrices between observed and unobserved points, and $\Sigma_{B,uu}$ and $\Sigma_{\nu,uu}$ are the covariance matrices for unobserved points. Using this notation relevant variations on the matrices in (9) are

$$\widetilde{X}_u = \begin{bmatrix}\mathcal{M}_u & F_u X_u\end{bmatrix} \qquad \text{and} \qquad \widetilde{\Sigma}_{uo} = \Sigma_{\nu,uo} + F_u \Sigma_{B,uo} F^\top. \tag{19}$$

Prediction of β-fields

Treating (8) as a hierarchical model straight forward but tedious calculations give predictions for the β-fields as

$$\mathsf{E}\left(B_u | Y, \Psi\right) = X_u \widehat{\boldsymbol{\alpha}} + \Sigma_{B,uo} F^\top \widetilde{\Sigma}^{-1} \left(Y - \widetilde{X}\left[\begin{matrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{matrix}\right]\right), \tag{20}$$

with variance

$$\begin{aligned}\mathsf{VAR}\left(B_u | Y, \Psi, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) =& \Sigma_{B,uu} - \Sigma_{B,uo}\Sigma_B^{-1}\Sigma_{B,ou} + \Sigma_{B,uo}\Sigma_B^{-1}\Sigma_{B|Y}\Sigma_B^{-1}\Sigma_{B,ou}\\ =& \Sigma_{B,uu} - \Sigma_{B,uo} F^\top \Sigma_\nu^{-1} F \Sigma_{B|Y} \Sigma_B^{-1} \Sigma_{B,ou}.\end{aligned} \tag{21}$$

Here, the first two terms in the top line of (21) contain the standard spatial prediction uncertainty for β, the last term is the added uncertainty from estimating β at the observed locations given Y. The variance in (21) is conditional on both regression and covariance parameters, adding the uncertainty in the regression coefficients the variance becomes

$$\begin{aligned}\mathsf{VAR}\left(B_u | Y, \Psi\right) =& \mathsf{VAR}\left(B_u | Y, \Psi, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) + \left(\begin{bmatrix}0 & X_u\end{bmatrix} - \Sigma_{B,uo} F^\top \widetilde{\Sigma}^{-1} \widetilde{X}\right) \cdot\\ & \left(\widetilde{X}^\top \widetilde{\Sigma}^{-1} \widetilde{X}\right)^{-1} \left(\begin{bmatrix}0 & X_u\end{bmatrix} - \Sigma_{B,uo} F^\top \widetilde{\Sigma}^{-1} \widetilde{X}\right)^\top.\end{aligned} \tag{22}$$

For β-fields at observed locations (20) and (21) simplify to

$$\begin{aligned}\mathsf{E}\left(B | Y, \Psi\right) =& X\widehat{\boldsymbol{\alpha}} + \Sigma_{B|Y} F^\top \Sigma_\nu^{-1} \left(Y - \widetilde{X}\left[\begin{matrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{matrix}\right]\right),\\ \mathsf{VAR}\left(B | Y, \Psi, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) =& \Sigma_{B|Y}.\end{aligned}$$

Prediction of Y

The model (10) is multivariate Normal and the full predictions of unobserved Y's are in principal standard kriging estimates. For the predictions we are primarily interested in the smooth underlying field, denoted Y^*; this can also been interpreted as smoothing over the nugget in (5) (Cressie 1993, Ch. 3.2.1). The predictions of Y^* and Y *differ only at observed locations*.

Using (9) and (19) predictions for Y_u^* are

$$\mathsf{E}\left(Y_u^*|Y,\Psi\right) = \widetilde{X}_u \begin{bmatrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{bmatrix} + \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\left(Y - \widetilde{X}\begin{bmatrix}\widehat{\boldsymbol{\gamma}}\\ \widehat{\boldsymbol{\alpha}}\end{bmatrix}\right), \tag{23}$$

with variances

$$\mathsf{VAR}\left(Y_u^*|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right) = \widetilde{\Sigma}_{uu}^* - \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{\Sigma}_{ou}^* \tag{24a}$$

$$\mathsf{VAR}\left(Y_u|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right) = \widetilde{\Sigma}_{uu} - \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{\Sigma}_{ou}^* \tag{24b}$$

Here $\widetilde{\Sigma}_{uo}^*$ is the cross-covariance matrix *excluding* the nugget in ν, cf. (5); this distinction is only relevant when Y_u^* includes observed points. For the prediction variances, (24a) gives the uncertainty in the prediction of the underlying y^*-field, while (24b) gives the uncertainty for a new observations at this point; the difference is similar to that between confidence and prediction intervals in regression (Ch. 11.3.5 in Casella and Berger 2002) and is of importance for the cross-validation. As for (21), (24) is conditional on both regression and covariance parameters, accounting for uncertainty in the regression coefficients gives

$$\begin{aligned}\mathsf{VAR}\left(Y_u^*|Y,\Psi\right) =& \mathsf{VAR}\left(Y_u^*|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right) + \\ & \left(\widetilde{X}_u - \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{X}\right)\left(\widetilde{X}^\top\widetilde{\Sigma}^{-1}\widetilde{X}\right)^{-1}\left(\widetilde{X}_u - \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{X}\right)^\top\end{aligned} \tag{25}$$

for (24a) and similarly for (24b).

For unobserved time-points it should be noted that the lack of temporal correlation in ν implies that predictions of $y(s,t_u)$ are identical to the contribution from the β-fields in (20) (see Figure 1),

$$\mathsf{E}\left(y(s,t_u)|Y,\Psi\right) = \mathcal{M}_u\widehat{\boldsymbol{\gamma}} + F_u\mathsf{E}\left(B_u|Y,\Psi\right).$$

The prediciton variance for unobserved time-points,

$$\mathsf{VAR}\left(y(s,t_u)|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right) = F_u\,\mathsf{VAR}\left(B_u|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right)F_u^\top + \Sigma_{\nu,uu},$$

will typically be much larger than for observed time-points due to the added uncertainty from the completely unknown $\nu(s,t_u)$-field.

Temporal averages

A primary interest in MESA Air is the health effects of chronic exposure to air pollution. Thus we are interested in the long term average exposure at each location, $\overline{y}(s) = \left(\sum_t y(s,t)\right)/T$.

Predictions and uncertainties of temporal averages are given by

$$\mathsf{E}\left(\overline{y^*}(s)\middle|Y,\Psi\right) = \frac{1}{T}\sum_{t=1}^{T}\mathsf{E}\left(y^*(s,t)|Y,\Psi\right), \tag{26a}$$

$$\mathsf{VAR}\left(\overline{y^*}(s)\middle|Y,\Psi\right) = \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\mathsf{COV}\left(y^*(s,t_1),y^*(s,t_2)|Y,\Psi\right), \tag{26b}$$

where $\mathsf{COV}\left(y^*(s,t_1),y^*(s,t_2)|Y,\Psi\right)$ is the matrix form of (25).

log-*Gaussian fields*

Transformation of data is commonly used to facilitate the modelling of non-Gaussian data under Gaussian assumptions (Tukey 1957; Box and Cox 1964). The log-transformation has been successfully applied to environmental data, including, but not limited to, air-pollution (e.g. PM_{10}, $PM_{2.5}$ and NO_x; see Paciorek *et al.* 2009; Szpiro *et al.* 2010; Sampson *et al.* 2011) and precipitation (Damian *et al.* 2003). For log-transformed data exact expressions exist for both bias-corrected estimates and their associated mean squared prediction errors (MSPE) (Cressie 1993, 2006; De Oliveira 2006). This stands in contrast to the approximate δ-method described in Ch. 3.2.2 of Cressie (1993) for general trans-Gaussian Kriging.

To accomondate log-transformed data both point and temporal average predictions for log-Gaussian processes, as described below, are implemented in **SpatioTemporal**. Figure 1 illustrates the difference between predictions of the log-Gaussian process (original data) and of the Gaussian process (transformed data).

If $y(s,t)$ is a Gaussian random process (10) then the corresponding log-Gaussian random process (i.e. the original, untransformed, data) is given by $z(s,t) = \exp\left(y(s,t)\right)$ with expectation

$$\mu_z(s,t) = \mathsf{E}\left(z(s,t)\right) = \exp\left(\mathsf{E}\left(y(s,t)\right) + \frac{\mathsf{VAR}\left(y(s,t)\right)}{2}\right).$$

Assuming that both covariance (Ψ) and regression ($\boldsymbol{\alpha},\boldsymbol{\gamma}$) parameters are known the best unbiased predictor of an unobserved part of the log-Gaussian field is (Cressie 1993, Ch. 3.2.2)

$$\widehat{Z_u^*} = \exp\left(\mathsf{E}\left(Y_u^*|Y,\Psi\right) + \frac{\mathsf{VAR}\left(Y_u^*|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right)}{2}\right). \tag{27}$$

Here we are, just as in (23), interested in the underlying smooth field excluding the nugget (see Appendix A in Cressie 2006); to obtain a predictor, $\widehat{Z_u}$, that includes the nugget, the variance from (24a) is replaced by the variance from (24b) in (27). The MSPE of $\widehat{Z_u^*}$ is,

$$\begin{aligned}\mathsf{MSPE}\left(\widehat{Z_u^*}\right) &= \mu_z(s_u,t_u)^2\left(\exp\left(\widetilde{\Sigma}_{uu}^*\right) - \exp\left(\widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{\Sigma}_{ou}^*\right)\right)\\ &\approx \widehat{Z_u^*}^2\exp\left(\widetilde{\Sigma}_{uu}^*\right)\left[1-\exp\left(-\mathsf{VAR}\left(Y_u^*|Y,\Psi,\boldsymbol{\alpha},\boldsymbol{\gamma}\right)\right)\right],\end{aligned} \tag{28}$$

where the approximation follows since $\widehat{Z_u^*}$ is an unbiased estimator of $\mu_z(s,t)$. Prediction and confidence intervals for $z(s,t)$ are obtained by transformation of the corresponding intervals obtained for $y(s,t)$.

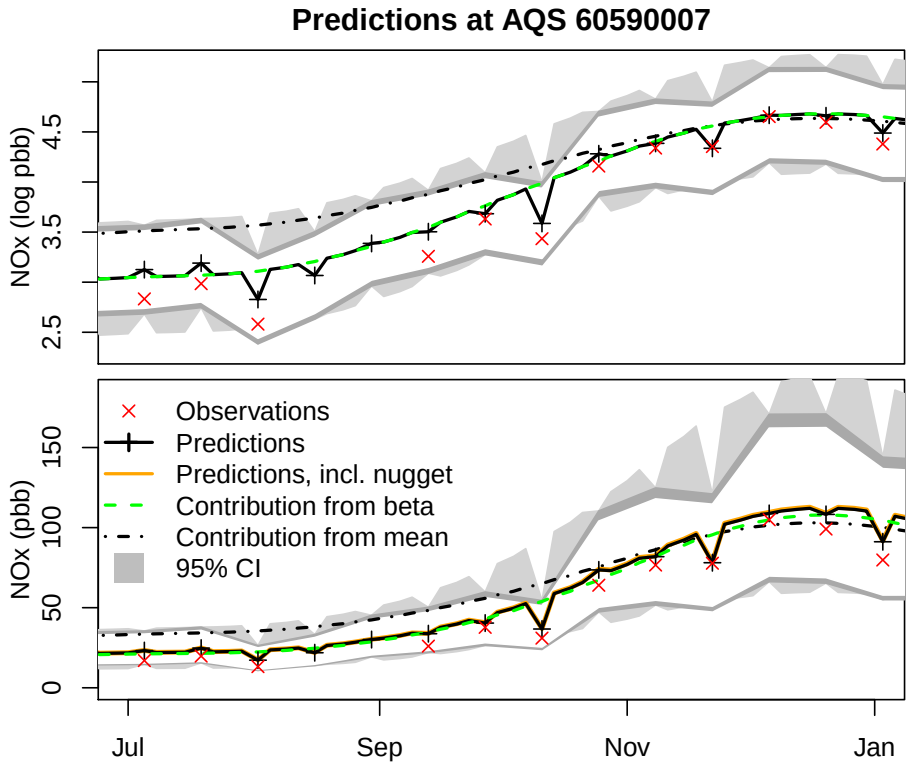

Figure 1: Example of predictions on the Gaussian (top) and on the original scale (bottom) at an AQS site (60590007) given data at all other locations. The black line gives the predictions ($\mathsf{E}\left(Y_u^*|Y\right)$ or $\widehat{Z_u^*}$); the orange line is $\widehat{Z_u}$, i.e. predictions incl. the nugget (lower pane only); the dashed green line gives the contribution from the β-fields, $\mathcal{M}\widehat{\boldsymbol{\gamma}} + F\mathsf{E}\left(B_u|Y\right)$; and the dotted line is the contribution from the mean, $\mathcal{M}\widehat{\boldsymbol{\gamma}} + FX\widehat{\boldsymbol{\alpha}}$. For the lower pane, the last two are simply transformed as $\exp\left[\cdot\right]$ without bias correction. Observations occur every 14-days (red $\times$), predictions at these time-points (black $+$) are close to the observations while observations at un-observed time-points co-incide with (top), or are close to (bottom), the dashed green line. Three different 95% intervals are given: a confidence interval for observed time-points (white), prediction interval for observed time-points (dark grey), and prediction interval for the additional un-observed time-points (light grey).

The distinction between predictions with and without nugget is much more important for the log-Gaussian case (27) than for the Gaussian case (23). In general $\widehat{Z_u^*} \neq \widehat{Z_u}$, since

$$\mathsf{VAR}\left(Y_u^*|Y, \Psi, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right) \neq \mathsf{VAR}\left(Y_u|Y, \Psi, \boldsymbol{\alpha}, \boldsymbol{\gamma}\right), \text{ whereas } \mathsf{E}\left(Y_u^*|Y, \Psi\right) \neq \mathsf{E}\left(Y_u|Y, \Psi\right)$$

only at observed locations. The difference between $\widehat{Z_u^*}$ and $\widehat{Z_u}$ is important when comparing predictions to left-out observations in cross-validation.

For the case with unknown regression parameters two possible predictors exist, either the unbiased predictor $\widehat{\widehat{Z_u^*}}^{\text{UB}}$ (De Oliveira 2006; Cressie 2006) or the biased, minimum mean squared

error predictor $\widehat{\widehat{Z}}_u^{*\,\mathrm{ME}}$ (De Oliveira 2006),

$$\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}} = \exp\left(\mathsf{E}\left(Y_u^*|Y,\Psi\right) + \frac{\mathsf{VAR}\left(Y_u^*|Y,\Psi\right)}{2} - \Lambda_z\right), \tag{29a}$$

$$\widehat{\widehat{Z}}_u^{*\,\mathrm{ME}} = \exp\left(\mathsf{E}\left(Y_u^*|Y,\Psi\right) + \frac{\mathsf{VAR}\left(Y_u^*|Y,\Psi\right)}{2} - 2\Lambda_z\right). \tag{29b}$$

Here Λ_z is the Lagrange multiplier for the Kriging predictor in (23),

$$\Lambda_z = \left(\widetilde{X}_u - \widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{X}\right)\left(\widetilde{X}^\top\widetilde{\Sigma}^{-1}\widetilde{X}\right)^{-1}\widetilde{X}_u^\top. \tag{30}$$

The MSPE of the predictors in (29) are

$$\mathsf{MSPE}\left(\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}}\right) \approx \left(\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}}\right)^2 \exp\left(\widetilde{\Sigma}_{uu}^*\right)\left[1 - \exp\Big(-\mathsf{VAR}\left(Y_u^*|Y,\Psi\right)\Big)\left(2e^{\Lambda_z} - e^{2\Lambda_z}\right)\right], \tag{31a}$$

$$\mathsf{MSPE}\left(\widehat{\widehat{Z}}_u^{*\,\mathrm{ME}}\right) \approx \left(\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}}\right)^2 \exp\left(\widetilde{\Sigma}_{uu}^*\right)\left[1 - \exp\Big(-\mathsf{VAR}\left(Y_u^*|Y,\Psi\right)\Big)\right], \tag{31b}$$

where we have used that $\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}}$ is an unbiased estimator of $\mu_z(s,t)$. We note that

$$\mathsf{MSPE}\left(\widehat{\widehat{Z}}_u^{*\,\mathrm{ME}}\right) \leq \mathsf{MSPE}\left(\widehat{\widehat{Z}}_u^{*\,\mathrm{UB}}\right)$$

in accordance with De Oliveira (2006)

Following the discussion of block prediction for log-Gaussian processes in De Oliveira (2006) and Cressie (2006), estimates of temporal averages are computed as

$$\widehat{\widehat{z^*}}(s) = \frac{1}{T}\sum_{t=1}^{T}\widehat{z}_u^*(s,t), \tag{32}$$

and similarly for $\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s)$ and $\widehat{\widehat{z^*}}^{\,\mathrm{ME}}(s)$. The MSPE for these predictors are

$$\begin{aligned}\mathsf{MSPE}\left(\widehat{\widehat{z^*}}(s)\right) \approx & \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\widehat{z^*}(s,t_1)\widehat{z^*}(s,t_2) \\ & \left[\exp\left(\left[\widetilde{\Sigma}_{uu}^*\right]_{st_1,st_2}\right) - \exp\left(\left[\widetilde{\Sigma}_{uo}^*\widetilde{\Sigma}^{-1}\widetilde{\Sigma}_{ou}^*\right]_{st_1,st_2}\right)\right]\end{aligned} \tag{33a}$$

$$\begin{aligned}\mathsf{MSPE}\left(\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s)\right) \approx & \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s,t_1)\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s,t_2)\exp\left(\left[\widetilde{\Sigma}_{uu}^*\right]_{st_1,st_2}\right) \\ & \left[1 - \exp\Big(-\mathsf{COV}\left(y^*(s,t_1),y^*(s,t_2)|Y,\Psi\right)\Big)\right. \\ & \left.\left(e^{\left[\Lambda_z\right]_{st_1,st_2}} + e^{\left[\Lambda_z^\top\right]_{st_1,st_2}} - e^{\left[\Lambda_z\right]_{st_1,st_2}+\left[\Lambda_z^\top\right]_{st_1,st_2}}\right)\right]\end{aligned} \tag{33b}$$

$$\begin{aligned}\mathsf{MSPE}\left(\widehat{\widehat{z^*}}^{\,\mathrm{ME}}(s)\right) \approx & \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s,t_1)\widehat{\widehat{z^*}}^{\,\mathrm{UB}}(s,t_2)\exp\left(\left[\widetilde{\Sigma}_{uu}^*\right]_{st_1,st_2}\right) \\ & \left[1 - \exp\Big(-\mathsf{COV}\left(y^*(s,t_1),y^*(s,t_2)|Y,\Psi\right)\Big)\right]\end{aligned} \tag{33c}$$

Here $\left[\bullet\right]_{st_1,st_2}$ denotes elements in the $\bullet$-matrix; to make Λ_z a $T \times T$-matrix $\widetilde{X}_u$ and $\widetilde{\Sigma}^*_{uo}$ in (30) should now contain covariates for all times over which we are averaging.

2.4. Cross-Validation

The model's predictive accuracy can be assessed through cross-validation. In Lindström *et al.* (2013) a cross-validation setup is presented; the setup, implemented in this package, handles the highly unbalanced sampling design and the MESA Air study's interest in predicting long term average exposures.

Dividing the observed locations into groups, n-fold cross-validation, consisting of parameter estimation and predictions, is performed as usual (Hastie *et al.* 2001, Ch. 7.10). Given the predictions and prediction variances, coverage of 95% prediction intervals, root mean squared error (RMSE) and cross-validated R^2's,

$$R^2 = \max\left(0, 1 - \frac{\text{RMSE}^2}{\mathsf{VAR}(y(s,t))}\right), \tag{34}$$

can be computed.

When assessing predictions of temporal averages at each location, the possibility of missing observations and the resulting missmatch between averages over available observations and averages over all predictions must be considered. To account for this, the averages in (26) are adjusted to include only observed time-points

$$\overline{y}(s) = \sum_{t \in \{\tau\,:\,\exists y(s,\tau)\}} \frac{y(s,t)}{\|\{\tau\,:\,\exists y(s,\tau)\}\|}. \tag{35}$$

For locations with only a few observations a large part of the R^2 may be attributable to the temporal variability. In Lindström *et al.* (2013) the temporal and spatial effects were separated by replacing $\mathsf{VAR}(y(s,t))$ in (34) with the MSE of a reference model. Suggested reference models were: 1) the spatial average at each time-point based on observations from all locations with good temporal coverage; 2) the observation from the closest available location with good temporal coverage; 3) smooth temporal trends fitted to data from the closest location with good temporal coverage. The resulting R^2's represent the improvement in predictions provided by this model, compared to central location or nearest neighbour schemes commonly used in epidemiology studies (Pope *et al.* 1995; Miller *et al.* 2007).

3. Package Features

The R-package **SpatioTemporal** includes functions for estimation (`estimate.STmodel` and `MCMC.STmodel`), prediction (`predict.STmodel`), cross-validation (`estimateCV.STmodel` and `predictCV.STmodel`), and simulation (`simulate.STmodel`) of the model described in Section 2. In addition to these functions the package also contains functions for construction (`createSTdata` and `createSTmodel`) of objects encapsulating model definition and data; plotting and evaluation of both data and results; and functions (see Section 4.3) that compute and evaluate the smooth temporal basis functions described in Section 2.1.

To reduce computational times matrix identities, such as (13), have been used when appropriate and functions from the **Matrix**-package (Bates and Maechler 2013) as well as some custom C-functions have been utilized for sparse and block matrix computations.

3.1. Key Assumptions

By construction, the model contains several key assumptions, including: 1) All temporal structure is captured by the smooth temporal basis functions, 2) Spatial dependencies (in the coefficients of the temporal functions) can be described using stationary universal Kriging 3) The residual $\nu(s,t)$-field is homoscedastic and independent in time; this will often require roughly equidistant temporal sampling, with occasional missing time points included for prediction but treated as unobserved. 4) When computing temporal averages the addition of many unobserved times-points will result in an average that tends to

$$\overline{y}(s) = \frac{1}{T}\sum_{i=1}^{m} \beta_i(s) \int_0^T f_i(t)\,\mathrm{d}t,$$

eliminating any contribution from the $\nu(s,t)$-field; this is an effect of the assumption of temporal independence in the $\nu(s,t)$-field.

4. Example: Analysis of Los Angeles NO_x Data

An example analysis of a small NO_x data set from Los Angeles is used to illustrate features of the **SpatioTemporal**-package. The data, which is included as an example in the pacakge, is a subset of data available to the MESA Air study; a detailed description of the full dataset can be found in Cohen *et al.* (2009).

Following a short description of the data (Section 4.1), we illustrate how to: 4.2) collect data into the `S3`-structure used by the package, 4.3) construct and evaluate the smooth temporal basis functions, 4.4) specify the covariates and covariances structures of the model in (3) and (4), 4.5) estimate parameters and do predictions, and 4.6) evaluate the model using cross-validation.

4.1. Data

NO_x Observations

The data used in this example consists of *log-transformed* 2-week average NO_x concentrations (ppb) in Los Angelse; observed at 20 locations from the national AQS (Air Quality System) network of regulatory monitors as well as at 5 locations from the supplementary MESA Air monitoring.

The national AQS network of regulatory monitors consists of a modest number of fixed sites that measure ambient concentrations of several different air pollutants including NO_x. The MESA Air supplementary monitoring consists of three sub-campaigns, (see Cohen *et al.* 2009, for details): "fixed sites", "home outdoor", and "community snapshot". Only data from the "fixed sites" have been included in this tutorial; this campaign consisted of five fixed site monitors that provided 2-week averages during the entire MESA Air monitoring period. To allow for comparison of the different monitoring protocols, one of the MESA Air fixed sites in coastal Los Angeles was colocated with an existing AQS monitor.

Covariates

To model the NO_x data, and to predict at unobserved locations, a number of geographic and

spatio-temporal covariates will be utilized. Geographic covariates used in this example are: 1) distance to major roads, i.e. census feature class code A1–A3 (distances truncated to be $\geq$10m and log-transformed); 2) distance to the closest road, i.e. the minimum of distances in 1) above; 3) distance to coast (truncated to be $\leq$15km); and 4) average population density in a 2 km buffer. Here census feature class code A1 roads refer to interstates and other limited access highways; A2 are primary roads without limited access; and A3 are secondary roads, e.g. state highways (see pp. 3–27 in US Census Bureau 2002). For details on the variable selection process that lead to these covariates as well as a more complete list of the covariates available to MESA Air the reader is referred to Mercer *et al.* (2011).

In addition to the geographic covariates a spatio-temporal covariate in the form of output from a deterministic air pollution model is also available. The spatio-temporal covariate is the output from a slightly modified version of Caline3QHC (EPA 1992; Wilton *et al.* 2010; MESA Air Data Team 2010). Caline is a line dispersion model for air pollution. Given locations of major (road) sources and local meteorology Caline uses a Gaussian model dispersion to predict how nonreactive pollutants travel with the wind away from sources; providing hourly estimates of air pollution at distinct points. The hourly contributions from Caline have been averaged to produce a 2-week average spatio-temporal covariate. The Caline predictions in this tutorial only includes air pollution due to traffic on major roads (A1, A2, and large A3).

4.2. Creating an `STdata`-Object

To get started we load the package, along with a few additional packages and the data used in this example:

```
> library(SpatioTemporal)
> library(Matrix)
> library(plotrix)
> library(maps)
> data(mesa.data.raw, package="SpatioTemporal")
```

Here the raw data contains observations along with geographic and spatio-temporal covariates.

The first step is to create an `STdata` S3-object from the raw data. This object collects observations, covariates and temporal trends and is used as input to several of the functions in **SpatioTemporal**.

```
> mesa.data <- createSTdata(obs=mesa.data.raw$obs, covars=mesa.data.raw$X,
+     SpatioTemporal=list(lax.conc.1500=mesa.data.raw$lax.conc.1500))
```

Here the observations, `mesa.data.raw$obs`, can be given either as a (number of time-points) – by – (number of locations) matrix, where the location and time of each observation is given by row- and columnames of the matrix and missing observations are denote by `NA`; or as a `data.frame` with fields `date`, `ID` and `obs`.

Geographic covariates and locations of the observation are given as a `data.frame`, `mesa.data.raw$X`, and matched to the observations through either 1) a field named `ID`, or 2) the rownames of the `data.frame`. All observations *must be associated* with a row in `mesa.data.raw$X`; additional rows giving covariates for unobserved locations at which we want predictions may be included in `mesa.data.raw$X`. If a field `type` exists in `mesa.data.raw$X` it is used to denote which type of monitoring system each location belongs to. In this example:

```
> table(mesa.data.raw$X$type)

  AQS FIXED
   20     5
```

The spatio-temporal covariates can be given either as a `list` of matrices or as a 3D-`array`. The row- and column names should provide the dates and locations of the spatio-temporal covariates.

The `STdata` object has `print`, `summary`, `plot`, `qqnorm` and `scatterPlot` methods that can be used for exploratory analysis of the data. For example `print(mesa.data)` provides an overview of observations and covariates while the plot functions can be used to illustrate which space-time locations that have been observed, investigate the Gaussianity of our data, or study the dependence between observations and covariates, see Figure 2.

```
> layout(matrix(c(1,2,1,3), 2, 2))
> par(mar=c(2.3,3.3,2,1), mgp=c(2,1,0))
> plot(mesa.data, "loc", main="Occurrence of Observations", xlab="",
+      ylab="Location", col=c("black", "red"), legend.loc=NULL)
> par(mar=c(3.3,3.3,2,1))
> qqnorm(mesa.data, line=1)
> scatterPlot(mesa.data, covar="km.to.coast", xlab="Distance to coast",
+             ylab="NOx (log ppb)", pch=19, cex=.25,
+             smooth.args=list(span=4/5,degree=2))
```

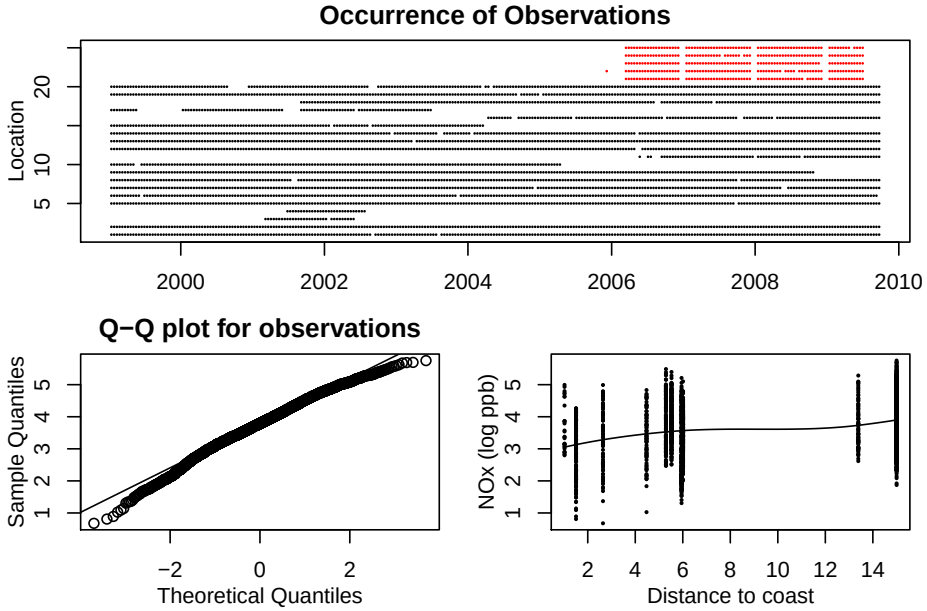

Figure 2: Counterclockwise from the top: Space-time locations of our observations divided into AQS (black) and MESA fixed (red) locations, `qqnorm`-plot for all observations, and dependence between observations and distance to coast.

4.3. Temporal Basis Functions

Having collected the data, we are now ready to evaluate the temporal structure and specify temporal basis functions. The algorithms described in Section 2.1 are implemented in `SVDmiss`, `SVDsmooth`, and `SVDsmoothCV`; `calcSmoothTrends` and `updateTrend` provide tools for altering the smooth temporal basis functions of `STdata`- and `STmodel`-objects.

To estimate the smooth temporal functions we first need to construct a data-matrix (11)

```
> D <- createDataMatrix(mesa.data)
```

For highly unbalanced measurement designs a `subset` option exists. This can be used to restrict the data matrix to only contain observations from some locations.

Determining the Number of Basis Functions

Here the temporal functions will be based on all available data. To determine a suitable number of basis functions the cross-validation described in Section 2.1 is run, evaluting 0 to 4 smooth basis functions (i.e. $m = 1, \ldots, 5$, since $f_1(t) \equiv 1$).

```
> SVD.cv <- SVDsmoothCV(D, 0:4)
```

The output of `SVDsmoothCV` consists of the cross-validated MSE, R^2, AIC, and BIC computed for each column using `summary.lm` and `extractAIC`. Averaging over all columns/cross-validation groups we have

```
> print(SVD.cv)

Result of SVDsmoothCV, average of CV-statistics:
                 MSE        R2       AIC       BIC
n.basis.0 0.37587597 0.0000000 -273.2316 -269.8684
n.basis.1 0.06979783 0.7583608 -643.9370 -637.2106
n.basis.2 0.05334179 0.8225333 -691.6487 -681.5591
n.basis.3 0.05229519 0.8284688 -695.0773 -681.6245
n.basis.4 0.04988794 0.8370008 -704.5244 -687.7084
```

with a graphical summary obtained from (see Figure 3)

```
> plot(SVD.cv)
```

As would be expected in any regression scenario, increasing the number of basis functions increases R^2 and decreases the MSE. All four statistics flatten out noticable after 2 basis functions, indicating that 2 basis functions is likely to provide the most efficient description of the temporal variability. The lack of auto-correlation when fitting the temporal basis to data at each location (Figure 5) shows that 2 basis functions are sufficient to capture the temporal structure.

We now use the `updateTrend` function to add the smooth temporal basis functions to the `STdata`-object

```
> mesa.data <- updateTrend(mesa.data, n.basis=2)
```

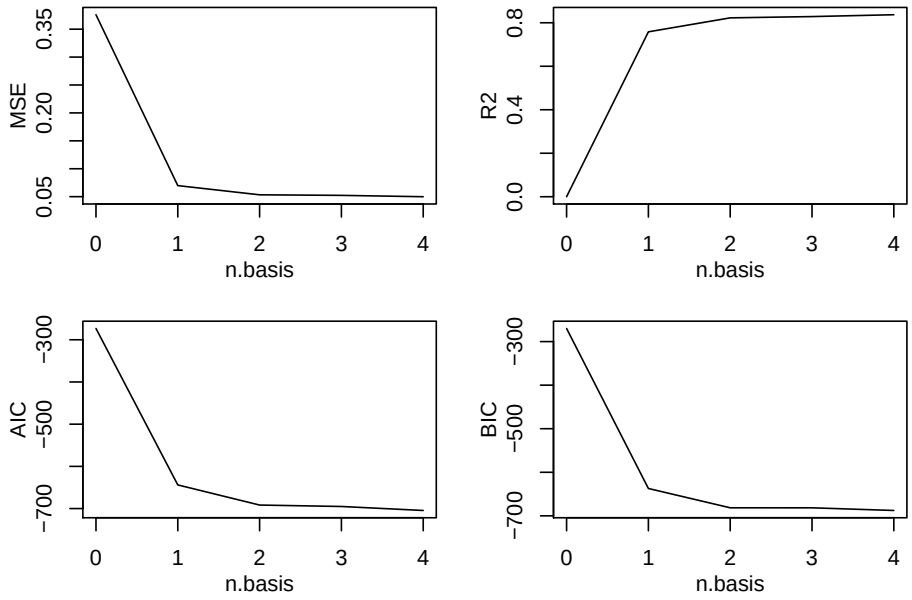

Figure 3: Cross-validation results for different numbers of smooth temporal trends

Alternatively `calcSmoothTrends` can be used to compute both the basis functions based on all data and those obtained when excluding each column in the data-matrix

```
> smooth.trend <- calcSmoothTrends(mesa.data, n.basis=2, cv=TRUE)
```

This allows for a sensitivity analysis of the temporal basis functions. Here we illustrate this by the fit at one location (Figure 4), but the different temporal basis functions could be carried through the entire analysis.

```
> mesa.data.cv <- vector("list",  length(smooth.trend$trend.fnc.cv))
> for(i in 1:length(mesa.data.cv)){
+   suppressMessages(mesa.data.cv[[i]] <- updateTrend(mesa.data,
+                    fnc=smooth.trend$trend.fnc.cv[[i]]))
+ }
> plot(mesa.data, main="Possible temporal trends",
+      xlab="", ylab="NOx (log ppb)", pch=c(19,NA), cex=.25)
> for(i in 1:length(mesa.data.cv)){
+   plot(mesa.data.cv[[i]], add=TRUE, col=i, pch=NA, lty=c(NA,2))
+ }
```

Evaluating the Basis Functions

The `plot.STdata` function can now be used to evaluate how well the temporal basis functions capture the temporal structure, see Figure 5. `plot.STdata` fits a linear regression of observations for a particular location of the smooth basis functions, and plots fitted values, residuals, auto-correlation, or partial auto-correlation functions. Here the two temporal basis func-

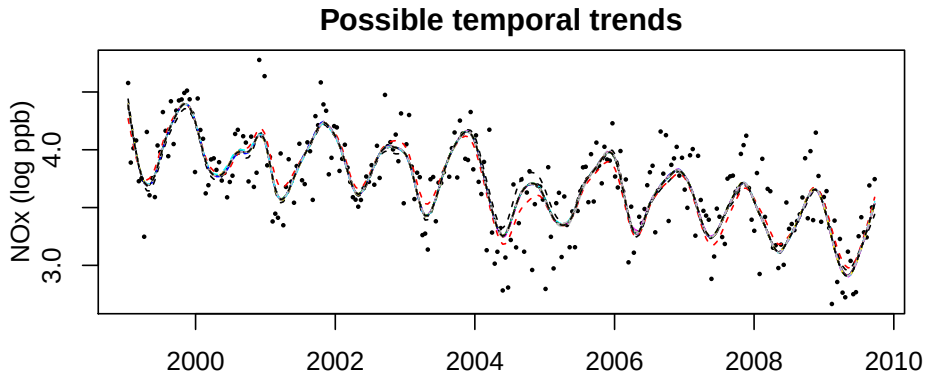

Figure 4: Cross-validated temporal basis functions. The solid line is the regression fit of 2 basis functions, based on all data, to observations at AQS site 60370002, the dashed lines are the corresponding fits of basis functions obtained when excluding one column at a time from the data-matrix.

tions capture the temporal variability in the data, as illustrated by residuals and correlation functions.

```
> par(mar=c(3.3,3.3,1.5,1), mgp=c(2,1,0))
> layout(matrix(c(1,1,2,2,3,4), 3, 2, byrow=TRUE))
> plot(mesa.data, "obs", ID="60370113",
+      xlab="", ylab="NOx (log ppb)",
+      main="Temporal trend 60370113")
> plot(mesa.data, "res", ID="60370113",
+      xlab="", ylab="NOx (log ppb)")
> plot(mesa.data, "acf", ID="60370113")
> plot(mesa.data, "pacf", ID="60370113")
```

Deterministic Basis Functions

In lieu of the SVD based basis functions we could use a set of deterministic temporal functions, e.g. $f_1(t) = 1$, $f_2(t) = 2\pi t/365$, $f_3(t) = \sin(2\pi t/365)$, and $f_4(t) = \cos(2\pi t/365)$. Specifying these and comparing to the data driven smooths extrated above (see Figure 6) we note that, for this data, three deterministic basis functions achieve a slightly worse fit than the two functions based on smoothed SVDs.

```
> mesa.data.fnc <- updateTrend(mesa.data, fnc=function(x){
+   x = 2*pi*as.numeric(x)/365;
+   return( cbind(x, sin(x), cos(x)) )})
> par(mfrow=c(2,1), mar=c(2.3,3.3,1.5,1), mgp=c(2,1,0))
> for(i in c("60370016","60371103")){
+   plot(mesa.data, ID=i, pch=c(19,NA), cex=.25, xlab="",
+        ylab="NOx (log ppb)", main=paste("AQS site",i))
+   plot(mesa.data.fnc, ID=i, add=TRUE, col=2, pch=NA)
+ }
```

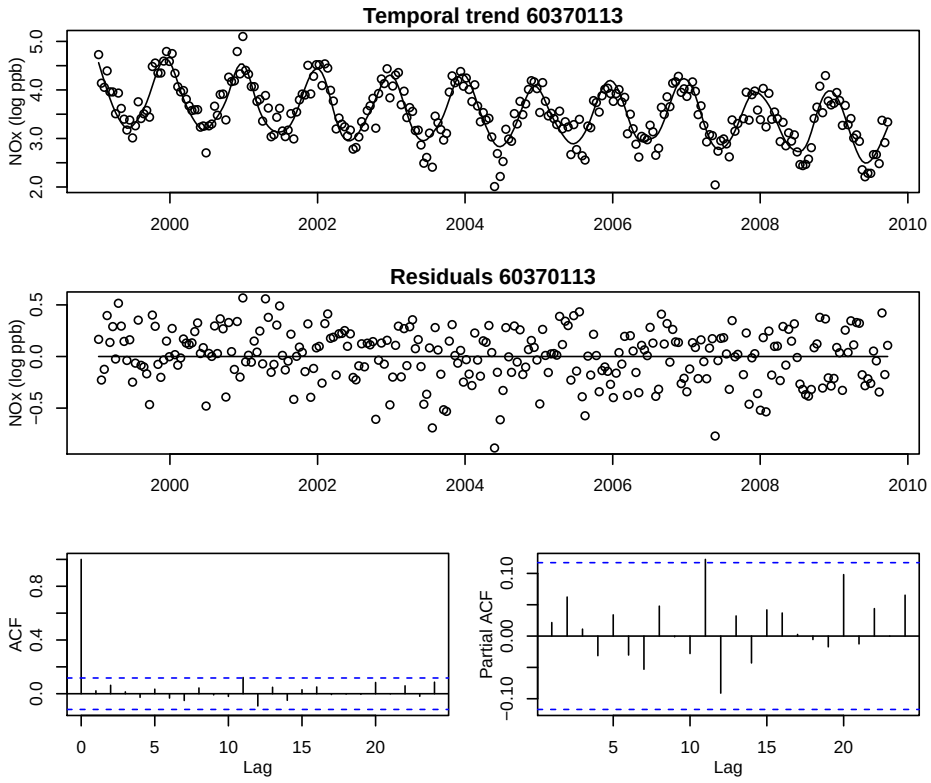

Figure 5: The smooth temporal trends fitted to data at site 60370113. Fitted trends, residuals, auto-correlation, and partial auto-correlation functions are shown.

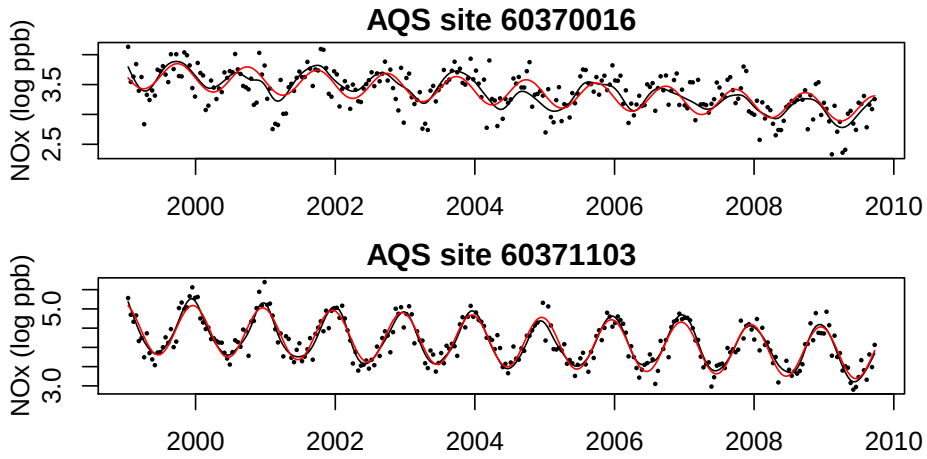

Figure 6: Comparing the fit of 2 smooth (black) and 3 deterministic (red) basis functions at two locations.

4.4. Specifying the Model

Having constructed a `STdata`-object with suitable temporal basis functions we are now ready to create a `STmodel`-object that can be used by the estimation and prediction functions of

SpatioTemporal. A `STmodel`-object is created through `createSTmodel`, by adding covariance and covariate specifications for the β- and ν-fields to a `STdata`-object.

Suitable covariates and covariance structures for the β-fields in (3) can be determined by considering the empirical estimates of these fields obtained by regressing the observations at each location on the temporal basis functions. The resulting regression coefficients can be analysed using standard geo-statistical software, e.g. provided by the R-packages **geoR** (Ribeiro and Diggle 2001) or **fields** (Furrer *et al.* 2012), to determine suitable mean and covariance models (see also Mercer *et al.* 2011).

Here we briefly illustrate the point by computing the regression coefficents and comparing them to a few possible covariates (Figure 7).

```
> beta.lm <- estimateBetaFields(mesa.data)
> par(mfrow=c(1,2), mar=c(3.3,2.3,1.5,1), mgp=c(2,1,0))
> plotCI(mesa.data$covars$log10.m.to.a1, beta.lm$beta[,1],
+        uiw=1.96*beta.lm$beta.sd[,1], ylab="", xlab="Distance to A1-road",
+        main="Beta-field for f1(t)")
> plotCI(mesa.data$covars$km.to.coast, beta.lm$beta[,2],
+        uiw=1.96*beta.lm$beta.sd[,2], ylab="", xlab="Distance to coast",
+        main="Beta-field for f2(t)")
```

To simplify the analysis of unbalanced measurement designs `estimateBetaFields` provides a `subset` option that restricts the evaluated locations.

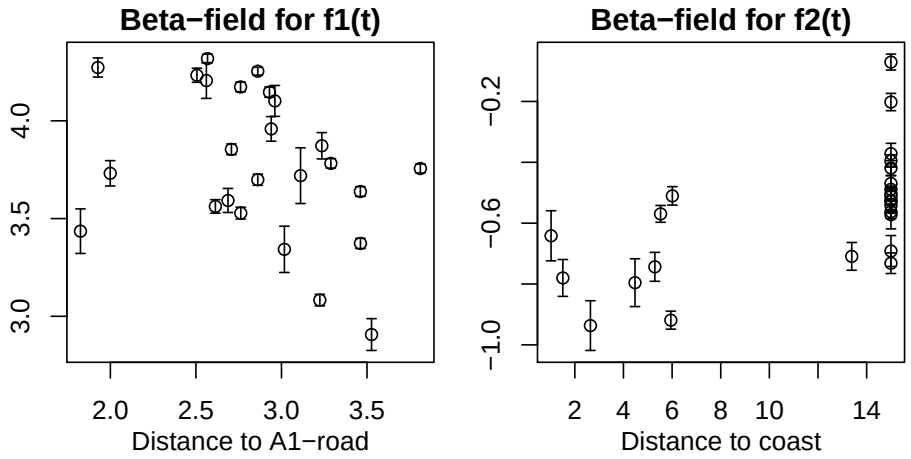

Figure 7: Regression estimates of $\beta_1(s)$ and $\beta_2(s)$ at each location, with 95% confidence intervals, as a function of the distance to A1 roads or the distance to coast.

Partially based on Figure 7 we specify different mean models for the three β-fields, but use an exponential covariance function without nugget for all fields; it is possible to specify different covariance functions for each β-field. Possible covariance functions are describe by the functions `namesCovFuns` and `parsCovFuns`, while `updateCovf` can be used to alter the covariance structure of an existing `STmodel`.

```
> LUR <- list(~log10.m.to.a1+s2000.pop.div.10000+km.to.coast,
```

```
+              ~km.to.coast, ~km.to.coast)
> cov.beta <- list(covf="exp", nugget=FALSE)
```

For the ν-field we use an exponential covariance with a nugget that can differ between AQS and MESA fixed sites.

```
> cov.nu <- list(covf="exp", nugget=~type, random.effect=FALSE)
```

In this specification the log-nugget will be given by a linear model

$$\log \sigma^2_{\text{nugget}}(s) = \theta_{\nu,\text{const}} + \theta_{\nu,\text{type}} \mathbf{1}(s \in \text{MESA fixed sites}). \tag{36}$$

The `random.effect=FALSE` option specifies that we do not want a random mean for each $\nu(\cdot, t)$ field. A random effect in time provides a way of capturing temporally uncorrelated large scale deviations from the smooth temporal basis functions. Using both a regression model for log-nugget (36) and a temporal random effect the full spatio-temporal covariance for the ν-field would be

$$r_\nu(s_1, t_1; s_2, t_2) = \begin{cases} \sigma^2_{\text{nugget}}(s_1) + g(0) + \tau^2, & \text{if } s_1 = s_2,\ t_1 = t_2, \\ g(s_1 - s_2) + \tau^2, & \text{if } s_1 \neq s_2,\ t_1 = t_2, \\ 0, & \text{if } t_1 \neq t_2, \end{cases} \tag{37}$$

where $\sigma^2_{\text{nugget}}(s)$ is a (spatially varying) nugget, $g(s)$ is a stationary covariance function, and τ^2 is the variance of a temporal random effect.

The final step is to specify coordinates for the observations and create the `STmodel`-object:

```
> locations <- list(coords=c("x","y"), long.lat=c("long","lat"),
+                   others="type")
> mesa.model <- createSTmodel(mesa.data, LUR=LUR, ST="lax.conc.1500",
+                             cov.beta=cov.beta, cov.nu=cov.nu,
+                             locations=locations)
```

In `locations` the element `coords` specifies which components of `mesa.data$covars` to use when computing distances between observation locations; the elements `long.lat` and `others` give additional fields that are carried over from `mesa.data$covars` to `mesa.model$locations`, e.g. data that could be useful for plotting. All variables referenced in the `LUR` and `cov.nu$nugget` formulas, and in the `locations`-list must be found in `mesa.data$covars`.

Most of the `S3`-methods available for `STdata`-objects also exists for `STmodel`-objects.

4.5. Parameter Estimation and Prediciton

Parameters of the spatio-temporal model (10) can by obtained by maximising (16) through `estimate.STmodel`. Given (estimated) parameters predictions are obtained from `predict.STmodel`.

Estimation

To avoid potential numerical optimisation issues, the estimation function allows for multiple starting points, returning all optima found. The functions `loglikeSTdim` and `loglikeSTnames` gives the number of parameters (and other model dimension) and the names, i.e. expected order, of the parameters. Using this information a two column matrix, where each column represents a different optimisation starting point, is constructed

```
> dim <- loglikeSTdim(mesa.model)
> x.init <- cbind(c( rep(2, dim$nparam.cov-1), 0),
+                 c( rep(c(1,-3), dim$m+1), -3, 0))
> rownames(x.init) <- loglikeSTnames(mesa.model, all=FALSE)
```

Model parameters are then estimated through

```
> est.mesa.model <- estimate(mesa.model, x.init, type="p", hessian.all=TRUE)
```

where `type` allows us to use either the f(ull), p(rofile) (16), or r(eml) (17) log-likelihoods. Since the estimation takes time (10–15 minutes, depending on the computer) the package contains precomputed results which we load and examine.

```
> data(est.mesa.model, package="SpatioTemporal")
> print(est.mesa.model)


Optimisation for STmodel with 2 starting points.
  Results: 2 converged, 0 not converged, 0 failed.
  Best result for starting point 1, optimisation has converged

No fixed parameters.

Estimated parameters for all starting point(s):
                                        [,1]          [,2]
gamma.lax.conc.1500              0.0008975546  0.0009008652
alpha.const.(Intercept)          3.7402698246  3.7406058853
alpha.const.log10.m.to.a1       -0.2021288763 -0.2022633497
alpha.const.s2000.pop.div.10000  0.0402182221  0.0401923686
alpha.const.km.to.coast          0.0374363255  0.0374629689
alpha.V1.(Intercept)            -0.7429226257 -0.7411611359
alpha.V1.km.to.coast             0.0174017754  0.0172645957
alpha.V2.(Intercept)            -0.1292573245 -0.1281826343
alpha.V2.km.to.coast             0.0155467684  0.0154883000
log.range.const.exp              2.4245453422  2.4205527321
log.sill.const.exp              -2.7522202309 -2.7568455241
log.range.V1.exp                 2.9175969067  2.9484089871
log.sill.V1.exp                 -3.5231738064 -3.5086290882
log.range.V2.exp                 1.7816565861  1.8062154010
log.sill.V2.exp                 -4.6812486307 -4.6730245593
nu.log.range.exp                 4.3836003026  4.3839197279
nu.log.sill.exp                 -3.2126038230 -3.2123312163
nu.log.nugget.(Intercept).exp   -4.4121007058 -4.4118479775
nu.log.nugget.typeFIXED.exp      0.6766505805  0.6748262774

Function value(s):
[1] 5748.563 5748.561
```

The estimation results indicate that the optimisation for both starting points have converged, starting point 1 being marginally better, with very similar parameter estimates and function values.

Plots of the estimated parameters together with approximate 95%-confidence intervals computed using the Hessian, i.e. the observed information matrix (Casella and Berger 2002, Ch. 10), can be seen in Figure 10 below; this Figure also includes parameter estimates from the cross-validation. Confidence intervals for covariance parameters of the residual ν-field are much smaller than the corresponding intervals for the covariance parameters characterising the β-fields. This is due to us having *only one replicate* of each β-field — in principal given by the regression of observations on the smooth temporal basis functions — but $T = 280$ replicates of the residual ν-field, *one for each timepoint* — in principal given by the residuals from the regression.

Predictions

Given estimated parameters, predictions can be computed for the Gaussian model using (23); for log-Gaussian processes using (27) or (29); and for the temporal averages using (26a) or (32).

```
> pred <- predict(mesa.model, est.mesa.model, LTA=TRUE, type="p")
> pred.log <- predict(mesa.model, est.mesa.model, LTA=TRUE,
+                     transform="unbiased", type="p")
```

Here the `type` option controls the computation of variances, with `type="p"` (or `type="f"`) the regression parameters are assumed known resulting in variances/MSPEs according to (24) or (28); for `type="r"` regression parameters are assumed unknown and variances/MSPEs are given by (25) or (31). Items computed include predictions of $y(s,t)$, contributions to the predictions from different parts of the model (see Figure 1 above), prediction variances (or MSPE), and predictions of the β-fields; `print.predictSTmodel` can be used to display all computed components.

The latent β-fields (20) can be compared to the empirical, regression based, estimates presented in Figure 7. Comparing these two estimates (see Figure 8) we see that they are close, with the largest discrepancies occuring for β's associated with the second temporal trend, $f_3(t)$. However, the uncertainty for these β's are substantial.

```
> par(mfrow=c(2,2), mar=c(3.3,3.3,1.5,1), mgp=c(2,1,0), pty="s")
> for(i in 1:3){
+   plotCI(x=beta.lm$beta[,i], y=pred$beta$EX[,i],
+          uiw=1.96*beta.lm$beta.sd[,i], err="x",
+          pch=NA, sfrac=0.005,
+          main=paste("Beta-field for f", i, "(t)", sep=""),
+          xlab="Empirical estimate",
+          ylab="Spatio-Temporal Model")
+   plotCI(x=beta.lm$beta[,i], y=pred$beta$EX[,i],
+          uiw=1.96*sqrt(pred$beta$VX[,i]),
+          add=TRUE, pch=NA, sfrac=0.005)
+   abline(0, 1, col="grey")
+ }
```

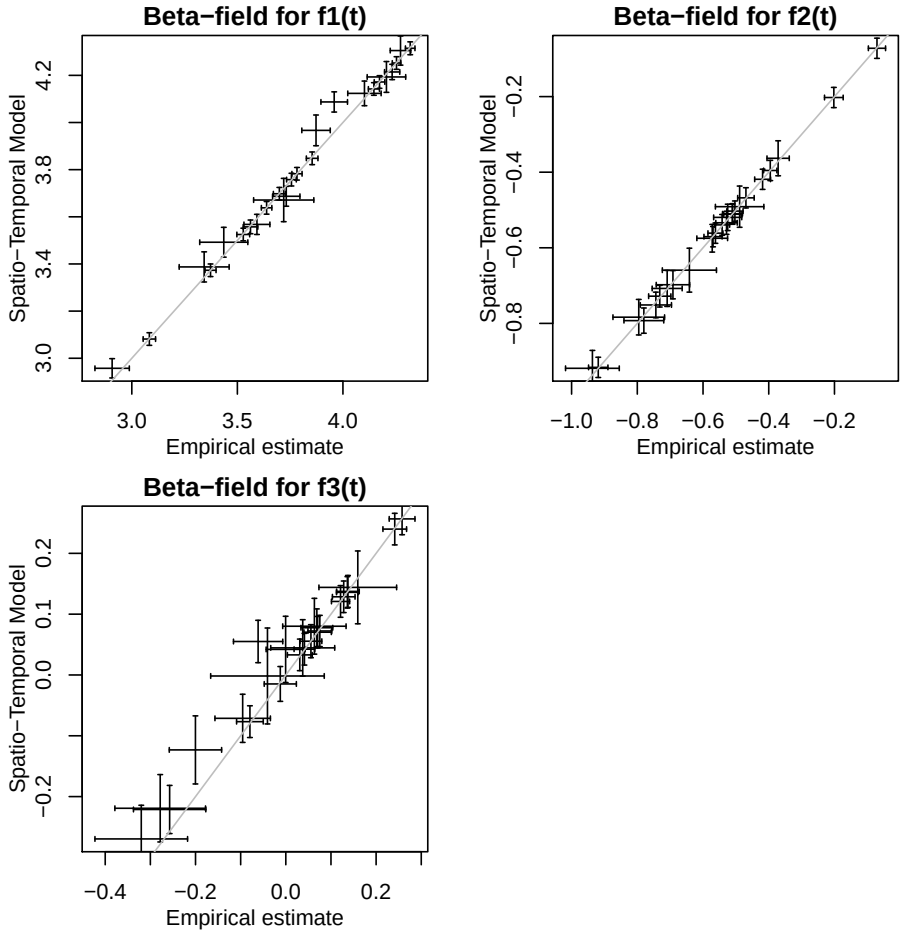

Figure 8: Comparison of the regression based and full (20) estimates of the β–field, including the uncertainties in both estimates.

The predictions also include the temporal averages at each location. In Figure 9 these predictions are compared to the temporal averages of observations at each site. A minor issue (see also Section 2.4) with this comparison is the temporal miss-match between predictions (averaged over 10-years) and the observations (averaged over 1 to 10 years depending on available data).

```
> par(mfrow=c(1,2), mar=c(3.3,3.3,1.5,1), mgp=c(2,1,0))
> with(pred$LTA, plotCI(colMeans(D, na.rm=TRUE), EX, uiw=1.96*sqrt(VX.pred),
+                       xlim=c(2.9,4.4), ylim=c(2.9,4.4),
+                       ylab="Predictions", xlab="Observations",
+                       main="Average NOx (log ppb)"))
> abline(0, 1, col="grey")
> with(pred.log$LTA, plotCI(colMeans(exp(D), na.rm=TRUE),
+                           EX, uiw=1.96*sqrt(VX.pred),
+                           xlim=c(25,95), ylim=c(25,95),
+                           ylab="Predictions", xlab="Observations",
+                           main="Average NOx (ppb)"))
```

```
> abline(0, 1, col="grey")
```

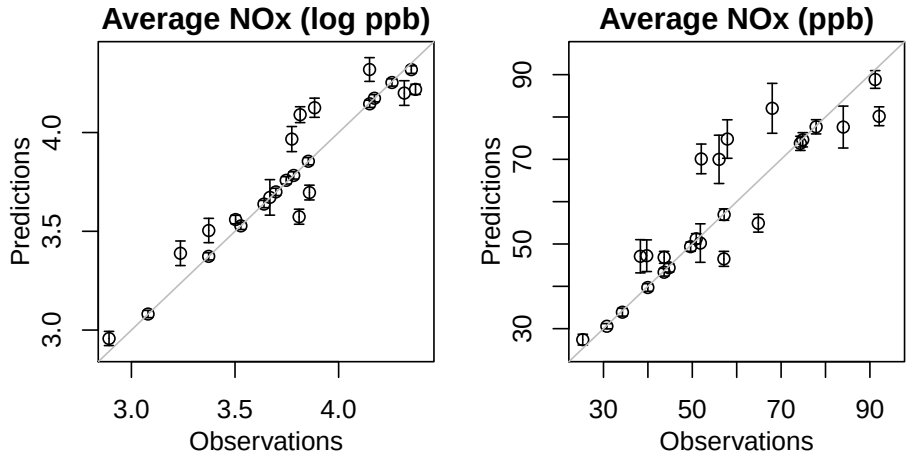

Figure 9: Observed and predicted temporal averages at each location, both for the Gaussian (left) and log-Gaussian (right) processes. The difference between observed and predicted values is (partially) due to temporally incomplete observations, i.e. the predicted average is over the *entire* 10-year period, regardless of when we have observations.

4.6. Cross-validation and Model Evaluation

The model's predictive ability can be evaluated through cross-validataion (Lindström *et al.* 2013). The first step is to define a vector that splits the observations into CV-groups

```
> Ind.cv <- createCV(mesa.model, groups=10, min.dist=.1)
```

For the i^{th} CV-group observations for which Ind.cv $= i$ are predicted using parameter estimates and conditional expectations based on the observations where Ind.cv $\neq i$. Since the primary focus when constructing the package was spatial prediction the number of locations in each group will be roughly even

```
> ID.cv <- sapply(split(mesa.model$obs$ID, Ind.cv), unique)
> print( sapply(ID.cv, length) )

 1  2  3  4  5  6  7  8  9 10
 3  3  3  2  2  2  2  2  2  4
```

although the number of observations in each grup may differ substantially

```
> table(Ind.cv)

Ind.cv
  1   2   3   4   5   6   7   8   9  10
438 389 811 556 546 165 228 487 160 797
```

Having four locations in the 10^{th} group is due to the fact that AQS site `60371103` is colocated with MESA Air site `L001`.

```
> print(mesa.model$D.beta[ID.cv[[10]],ID.cv[[10]]])
```

```
            60371103 60371601 60371701        L001
60371103  0.00000000 16.36527 43.75141  0.08363892
60371601 16.36527030  0.00000 29.06447 16.44669372
60371701 43.75141252 29.06447  0.00000 43.83429713
L001      0.08363892 16.44669 43.83430  0.00000000
```

As seen from the distance matrix above, the two stations are only 0.084 km apart (less than `min.dist=.1`) causing `createCV` to treat them as "one" location.

For cases where the primary interest is in temporal predictions suitable `Ind.cv` vectors could be construted with the help of elements in `mesa.model$obs`.

Estimation and Prediciton

For the estimation part of the cross-validation a set of starting point(s) is needed; here we use the optimum found in the optimisation and the starting point, of the two possible, which lead to the optimum.

```
> x.init <- coef(est.mesa.model, pars="cov")[,c("par","init")]
```

As before the estimation is computationally expensive

```
> est.cv.mesa <- estimateCV(mesa.model, x.init, Ind.cv)
```

and we load precomputed results.

```
> data(est.cv.mesa, package="SpatioTemporal")
```

Studying results from `print(est.cv.mesa)` (not shown) it can be seen that all 10 of the estimations have converged, and comparing these estimates (available through `coef(est.cv.mesa)`) to those obtained for the entire data set shows reasonable agrement, both with regards to values and uncertainties (Figure 10).

```
> par(mfrow=c(1,1), mar=c(13.5,2.5,.5,.5), las=2)
> with(coef(est.mesa.model, pars="all"),
+      plotCI((1:length(par))+.3, par, uiw=1.96*sd,
+             col=2, xlab="", xaxt="n", ylab=""))
> boxplot(est.cv.mesa, "all", boxwex=.4, col="grey", add=TRUE)
```

Given parameter estimates the prediction part of the cross-validation can be carried out for either Gaussian or log-Gaussian models. In both cases temporal averages are computed based on only the observed time-points, as in (35).

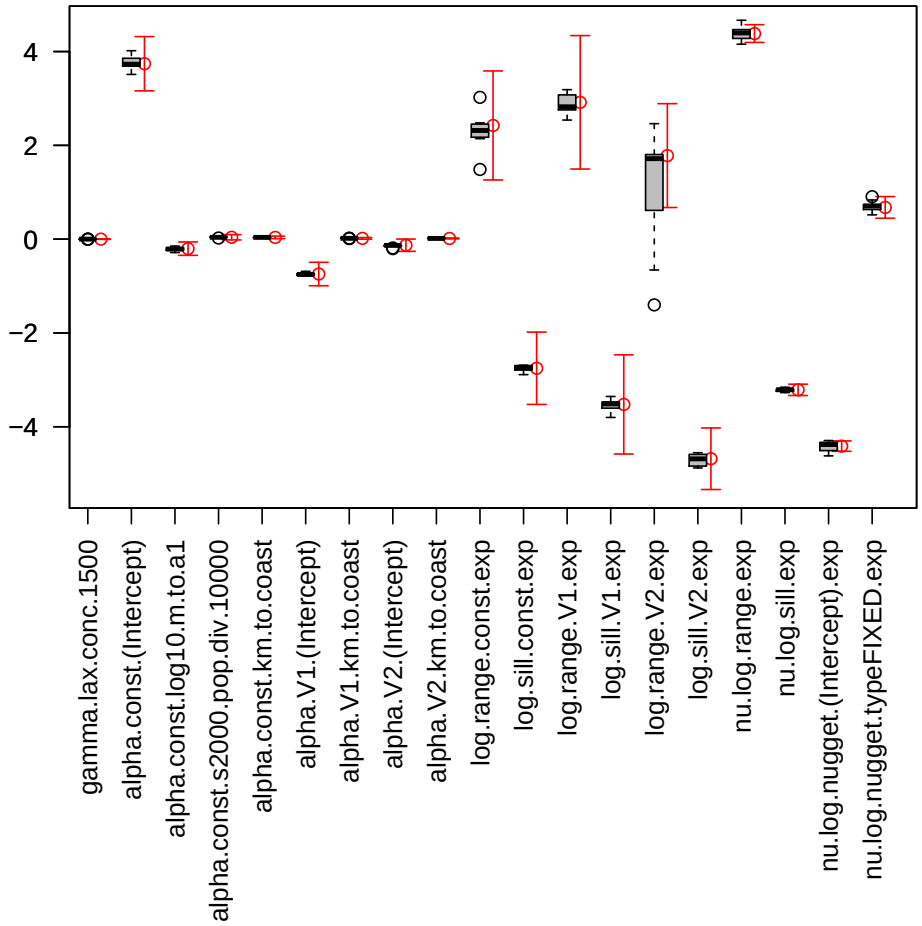

Figure 10: Estimated parameters and approximate 95% confidence intervals (red) compared to parameter estimates from 10-fold cross-validation (box-plots).

```
> ##pred.cv.mesa <- predictCV(mesa.model, est.cv.mesa, LTA=TRUE)
> data(pred.cv.mesa, package="SpatioTemporal")
> pred.cv.mesa.log <- predictCV(mesa.model, est.cv.mesa,
+                               LTA=TRUE, transform="unbiased")
```

As for `estimateCV` the computations take time and precomputed results are included in `data(CV.mesa.model)` which we loaded previously.

Alternatively the function `computeLTA` can be applied to the result from `predictCV`, however this will only compute the temporal averages in (35) and will *not* include the variances from (26b) or (33a)-(33c).

Model Evaluation

The `summary.predCVSTmodel` function computes RMSE, R^2, and coverage of prediction intervals for the cross-validation. For the log-Gaussian data we have:

```
> summary(pred.cv.mesa.log)
```

```
Cross-validation predictions for STmodel with 10 CV-groups.
  Predictions for 4577 observations.
  Temporal averages for 25 locations.

RMSE:
           EX.mu EX.mu.beta       EX  EX.pred
obs     25.10440   20.38305 16.28674 16.24542
average 16.66251   11.49040 10.64955 10.58167

R2:
            EX.mu EX.mu.beta        EX   EX.pred
obs     0.6047337  0.7394276 0.8336362 0.8344794
average 0.1813137  0.6106795 0.6655743 0.6698244

Coverage of 95% prediction intervals:
          EX.pred
obs     0.9224383
average 0.8800000
```

The results show reasonable models fits with R^2 values of 0.834 and 0.67 for observations and temporal averages respectively. It should also be noted that although the mean component,

$$\exp\left(\mathcal{M}\widehat{\boldsymbol{\gamma}} + FX\widehat{\boldsymbol{\alpha}}\right),$$

accounts for a substantial part of the R^2 (0.6) for the observations, it captures very little of the variability in the temporal averages ($R^2 \approx 0.18$). The 88% coverage of a 95% confidence interval for the temporal averages might seem a tad low, but one should remember that it translates to 22 of 25 locations.

To seperate temporal and spatial predictive ability for locations with few observations (not an issue here), the reference models described in Lindström *et al.* (2013) can be computed by `predictNaive`, and including these in the `summary`-call above gives adjusted R^2-values. Further `summary(pred.cv.mesa.log, by.date=TRUE)` provides individual cross-validation statistics for each time-point. The result from `predCV` contains, in `pred.cv.mesa.log$pred.obs`, a `data.frame` with observations, predictions, variances, etc. that can be used to compute additional cross-validation statistics.

To visualise the cross-validation results, several different plots are available (Figure 11); both for the original (ppb, log-Gaussian) and the transformed (log ppb, Gaussian) data. Studying time-series at single locations, both the predictions and 95% prediction intervals agree well with left-out observations. Plotting all predictions against observations also shows reasonable agreement, although some locations exhibit small, but consistent, biases. The predicted temporal averages (35) match the observations, but with rather large prediction intervals. The width of the intervals is caused by several locations only having a few years of data to average over. The normalised residuals $(y - \mathsf{E}(y))/\sqrt{\mathsf{VAR}(y)}$ are approximately $\mathsf{N}(0, 1)$, with no strong seasonal differences, indicating that our distributional assumptions are reasonable. Finally plots of residuals agains covariates show little unexplained, residual structure.

```
> par(mar=c(3.3,3.3,1.5,1), mgp=c(2,1,0))
> layout(matrix(c(1,1,2,2,3,4,5,6), 4, 2, byrow=TRUE))
> plot(pred.cv.mesa, ID="60371601", xlab="", ylab="NOx (log ppb)",
+      main="Predictions for 60371601", lty=c(1,NA), lwd=2,
+      pch=c(NA,19), cex=.75)
> plot(pred.cv.mesa, ID="60371601", pred.type="EX.mu",
+      lty=4, lwd=2, col="blue", add=TRUE)
> plot(pred.cv.mesa, ID="60371601", pred.type="EX.mu.beta",
+      lty=2, lwd=2, col="green", add=TRUE)
> plot(pred.cv.mesa.log, ID="60371601", xlab="", ylab="NOx (ppb)",
+      main="Predictions for 60371601", pred.type="EX.pred",
+      lty=c(1,NA), lwd=2, pch=c(NA,19), cex=.75)
> plot(pred.cv.mesa.log, ID="60371601", pred.type="EX.mu",
+      lty=4, lwd=2, col="blue", add=TRUE)
> plot(pred.cv.mesa.log, ID="60371601", pred.type="EX.mu.beta",
+      lty=2, lwd=2, col="green", add=TRUE)
> legend("topright", c("Observations", "Predictions",
+                      "Contribution from beta",
+                      "Contribution from mean",
+                      "95% CI"), bty="n",
+        lty=c(NA,1,2,4,NA), lwd=c(NA,2,2,2,NA),
+        pch=c(19,NA,NA,NA,15), pt.cex=c(.75,NA,NA,NA,2.5),
+        col=c("red", "black", "green", "blue", "grey"))
> plot(pred.cv.mesa, "obs", ID="all", pch=c(19,NA), cex=.25, lty=c(NA,2),
+      col=c("ID", "black", "grey"), xlab="Observations",
+      ylab="Predictions", main="Cross-validation NOx (log ppb)")
> with(pred.cv.mesa.log$pred.LTA, plotCI(obs, EX.pred, uiw=1.96*sqrt(VX.pred),
+                                  xlab="Observations", ylab="Predictions",
+                                  main="Temporal average NOx (ppb)"))
> abline(0, 1, col="grey")
> I.season <- as.factor(as.POSIXlt(pred.cv.mesa$pred.obs$date)$mon+1)
> levels(I.season) <- c(rep("Winter",2), rep("Spring",3),
+                       rep("Summer",3), rep("Fall",3), "Winter")
> qqnorm(pred.cv.mesa, norm=TRUE, main="Normalised residuals",
+        col=I.season)
> legend("bottomright", legend=as.character(levels(I.season)),
+        pch=1, col=1:nlevels(I.season), bty="n")
> scatterPlot(pred.cv.mesa, STdata=mesa.model, covar="log10.m.to.a1",
+             group=I.season, col=c(2:5,1), type="res",
+             xlab="Distance to A1-Road", ylab="Residuals",
+             main="Residuals (log ppb)")
```

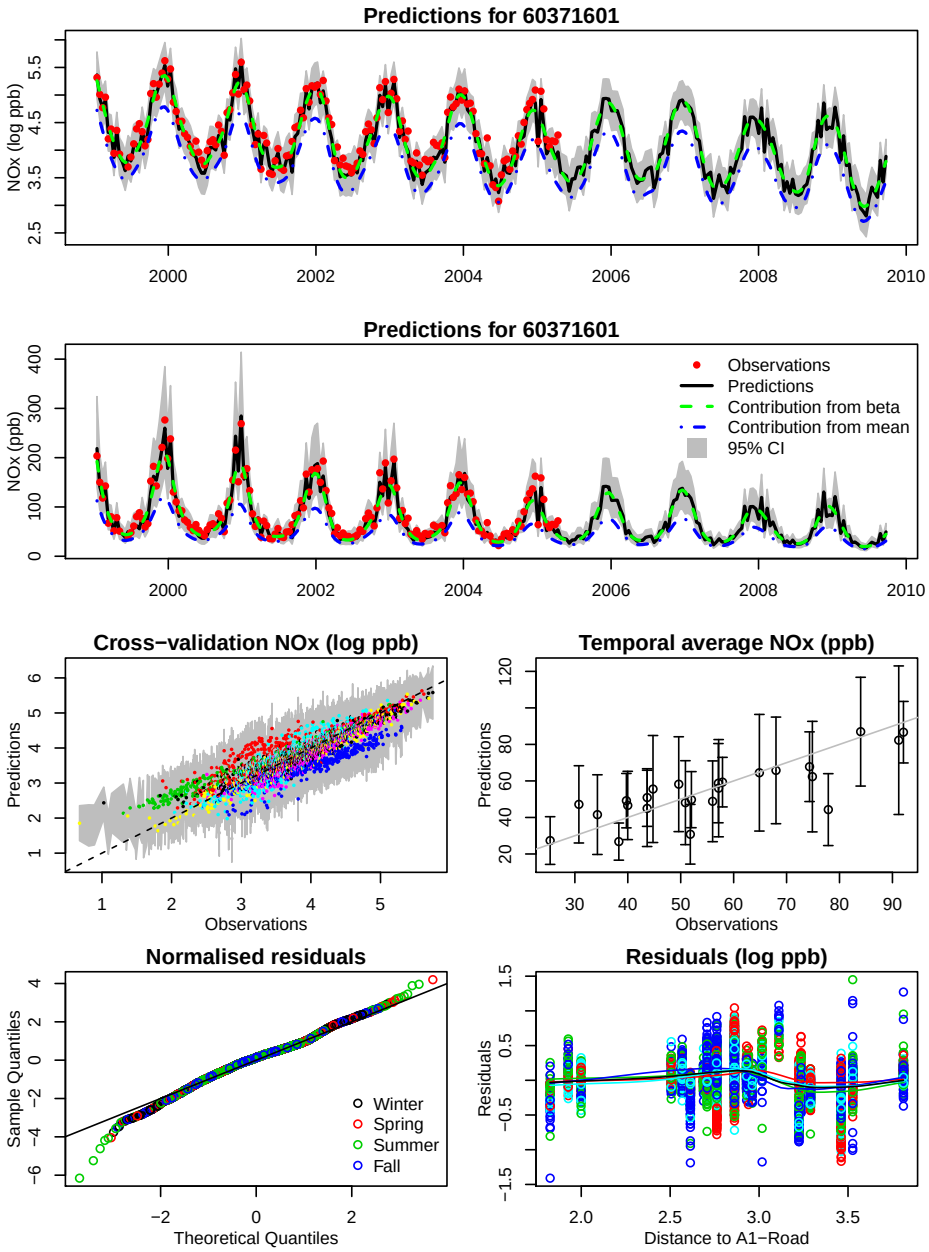

Figure 11: The two top panes shows predictions for a left-out location, in log and original scale (cf. Figure 1). In the left pane of the third row predictions have been plotted against observations; the points are coloured by location and grouping of data from single locations can be seen. On the right predictions and observations of temporal averages (35) are compared. In the last row, the left pane shows a QQ-plot for normalised residuals for the Gaussian model, coloured by season; the solid line gives the theoretical behaviour of $\mathsf{N}(0,1)$ residuals. On the right, residuals (coloured by season) are plotted as a function of distance to A1-roads; smooths, for each season and all data, have been added to help identify any remaining structure.

5. Outlook

Having introduced the R-package **SpatioTemporal** and illustrated the main features of the package, we now discuss one possible future extension of the model.

The computational feasibility of the model implemented here relies on the simplification of (12) using, mainly, the matrix identity in (13). This simplification will only lead to reduced computational cost if Σ_ν has a structure, e.g. block-diagonal, that allows for fast computations of Σ_ν^{-1}. To obtain this block-diagonal structure an assumption of temporal independence in the residual $\nu(s,t)$-field is required. In the future we hope to relax this assumption, allowing for a temporally correlated $\nu(s,t)$-field. Computational feasibility can hopefully be retained by obtaining a sparse Σ_ν matrix through tapering (Furrer *et al.* 2006).

Acknowledgements

Data used in the example has been provided by **the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air)**. Details regarding the data can be found in Cohen *et al.* (2009); Wilton *et al.* (2010).

Although this tutorial and development of the package described there in has been funded wholly or in part by the United States Environmental Protection Agency through **assistance agreement CR-834077101-0** and **grant RD831697** to the University of Washington, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Travel for Johan Lindström has been paid by **STINT Grant IG2005-2047**.

Additional funding was provided by grants to the University of Washington from the **National Institute of Environmental Health Sciences (P50 ES015915)** and the **Health Effects Institute (4749-RFA05-1A/06-10)**.

References

Banerjee S, Carlin BP, Gelfand AE (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC.

Bates D, Maechler M (2013). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.0-11, URL http://CRAN.R-project.org/package=Matrix.

Bild DE, Bluemke DA, Burke GL, R D, Diez Roux AV, Folsom AR, Greenland P, Jacob Jr DR, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP (2002). "Multi-Ethnic Study of Atherosclerosis: Objectives and Design." *American Journal of Epidemiology*, **156**(9), 871–881.

Box G, Cox D (1964). "An Analysis of Transformations." *J. Roy. Statist. Soc. Ser. B*, **26**(2), 211–252.

Calder CA (2008). "A Dynamic Process Convolution Approach to Modeling Ambient Particulate Matter Concentrations." *Environmetrics*, **19**(1), 39–48.

Casella G, Berger RL (2002). *Statistical Inference*. Second edition. Duxbury.

Cohen MA, Adar SD, Allen RW, Avol E, Curl CL, Gould T, Hardie D, Ho A, Kinney P, Larson TV, Sampson PD, Sheppard L, Stukovsky KD, Swan SS, Liu LJS, Kaufman JD (2009). "Approach to Estimating Participant Pollutant Exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air)." *Environmental Science & Technology*, **43**(13), 4687–4693.

Cressie N (1993). *Statistics for Spatial Data*. Revised edition. John Wiley & Sons Ltd.

Cressie N (2006). "Block Kriging for Lognormal Spatial Processes." *Math. Geol.*, **38**(4), 413–443.

Cressie N, Wikle CK (2011). *Statistics for Spatio-Temporal Data*. Wiley.

Damian D, Sampson PD, Guttorp P (2003). "Variance Modeling for Nonstationary Processes with Temporal Replications." *J. Geophys. Res.*, **108**(D24), 8778.

De Iaco S, Posa D (2012). "Predicting Spatio-Temporal Random Fields: Some Computational Aspects." *Comput. and Geosci.*, **41**(0), 12–24.

De Oliveira V (2006). "On Optimal Point and Block Prediction in Log-Gaussian Random Fields." *Scand. J. Statist.*, **33**(3), 523–540.

EPA (1992). "User's Guide to CAL3QHC Version 2.0: A Modeling Methodology for Predicting Pollutant Concentrations near Roadway Intersections." *Technical Report EPA-454/R-92-006*, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA.

Fanshawe TR, Diggle PJ, Rushton S, Sanderson R, Lurz PWW, Glinianaia SV, Pearce MS, Parker L, Charlton M, Pless-Mulloli T (2008). "Modelling Spatio-Temporal Variation in Exposure to Particulate Matter: A Two-Stage Approach." *Environmetrics*, **19**(6), 549–566.

Fuentes M, Guttorp P, Sampson PD (2006). "Using transforms to analyze space-time processes." In B Finkenstadt, L Held, V Isham (eds.), *Statistical Methods for Spatio-Temporal Systems*, pp. 77–150. CRC/Chapman and Hall.

Furrer R, Genton MG, Nychka D (2006). "Covariance Tapering for Interpolation of Large Spatial Datasets." *J. Comput. Graph. Statist.*, **15**(3), 502–523.

Furrer R, Nychka D, Sain S (2012). *fields: Tools for spatial data*. R package version 6.7, URL http://CRAN.R-project.org/package=fields.

Gamerman D (2010). "Dynamic Spatial Models Including Spatial Time Series." In AE Gelfand, P Diggle, P Guttorp, M Fuentes (eds.), *Handbook of Spatial Statistics*, pp. 437–448. Chapman & Hall/CRC.

Gneiting T, Guttorp P (2010). "Continuous Parameter Spatio-Temporal Processes." In AE Gelfand, P Diggle, P Guttorp, M Fuentes (eds.), *Handbook of Spatial Statistics*, pp. 427–436. Chapman & Hall/CRC.

Harville DA (1997). *Matrix Algebra From a Statistician's Perspective*. first edition. Springer.

Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Kaufman JD, Adar SD, Allen RW, Barr RG, Budoff MJ, Burke GL, Casillas AM, Cohen MA, Curl CL, Daviglus ML, Roux AVD, Jacobs DR, Kronmal RA, Larson TV, Liu SLJ, Lumley T, Navas-Acien A, O'Leary DH, Rotter JI, Sampson PD, Sheppard L, Siscovick DS, Stein JH, Szpiro AA, Tracy RP (2012). "Prospective Study of Particulate Air Pollution Exposures, Subclinical Atherosclerosis, and Clinical Cardiovascular Disease: The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air)." *Am. J. Epidemiology*, **176**(9), 825–837.

Lindström J, Szpiro AA, Sampson PD, Oron A, Richards M, Larson T, Sheppard L (2013). "A Flexible Spatio-Temporal Model for Air Pollution with Spatial and Spatio-Temporal Covariates." *Under revision for Environmental and Ecological Statistics*, **TBD**, ?–?

Lindström J, Szpiro AA, Sampson PD, Sheppard L, Oron A, Richards M, Larson T (2011). "A flexible spatio-temporal model for air pollution: Allowing for spatio-temporal covariates." *Technical Report Working Paper 370*, UW Biostatistics Working Paper Series. URL http://www.bepress.com/uwbiostat/paper370.

Mercer LD, Szpiro AA, Sheppard L, Lindström J, Adar SD, Allen RW, Avol EL, Oron AP, Larson T, Liu LJS, Kaufman JD (2011). "Comparing Universal Kriging and Land-Use Regression for Predicting Concentrations of Gaseous Oxides of Nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air)." *Atmo. Environ.*, **45**(26), 4412–4420.

MESA Air Data Team (2010). "Documentation of MESA Air Implementation of the Caline3QHCR Model." *Technical report*, University of Washington, Seattle, WA, USA.

Miller KA, Sicovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD (2007). "Long-Term Exposure to Air Pollution and Incidence of Cardiovascular Events in Women." *N. Engl. J. Med.*, **356**(5), 447–458.

Paciorek CP, Yanosky JD, Puett RC, Laden F, Suh HH (2009). "Practical Large-Scale Spatio-Temporal Modeling of Particulate Matter Concentrations." *Ann. Appl. Statist.*, **3**(1), 370–397.

Pope CA, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath Jr CW (1995). "Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults." *Am. J. Respir. Crit. Care med.*, **151**, 669–674.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Ribeiro PJJ, Diggle PJ (2001). "geoR: A Package for Geostatistical Analysis." *R-NEWS*, **1**(2), 15–18. URL http://CRAN.R-project.org/package=geoR.

Sahu SK, Gelfand AE, Holland D (2006). "Spatio-Temporal Modeling of Fine Particulate Matter." *J. Agric. Bio. and Environ. Statist.*, **11**(1), 61–86.

Sampson PD, Szpiro AA, Sheppard L, Lindström J, Kaufman JD (2011). "Pragmatic Estimation of a Spatio-temporal Air Quality Model with Irregular Monitoring Data." *Atmo. Environ.*, **45**(36), 6593–6606.

Smith RL, Kolenikov S, Cox LH (2003). "Spatio-Temporal Modeling of PM2.5 Data with Missing Values." *J. Geophys. Res.*, **108**(D24), 9004.

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar S, Kaufman J (2010). "Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies." *Environmetrics*, **21**(6), 606–631.

Tukey JW (1957). "On the Comparative Anatomy of Transformations." *Ann. Math. Statist.*, **28**(3), 602–632.

US Census Bureau (2002). "UA Census 2000 TIGER/Line Files Technical Documentation." *Technical report*, U.S. Census Bureau – Washington, DC. URL https://www.census.gov/geo/www/tiger/tigerua/ua2ktgr.pdf.

Wilton D, Szpiro AA, Gould T, Larson T (2010). "Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA." *Sci. Total Environ.*, **408**(5), 1120–1130.

Affiliation:

Johan Lindström
Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden
E-mail: johanl@maths.lth.se
URL: http://www.maths.lth.se/~johanl