PRELIMINARY RESULTS ON TAPERING MULTIVARIATE SPATIO TEMPORAL MODELS FOR EXPOSURE TO AIRBORNE MULTIPOLLUTANTS IN EUROPE

Alessandro Fassó [1] , Francesco Finazzi [1] and Ferdinand Ndongo[1]

ABSTRACT: The population exposure distribution is important for assessing the human risk related to air pollution. Since it requires multipollutant concentrations over a fine spatial grid at daily level, spatio-temporal modelling is required to interpolate ground level monitoring network data. The same authors recently introduced a trans-Gaussian model able to model multipollutant concentration over Europe in high resolution and, thanks to this, able to estimate the population exposure distribution, uncertainty included. Since modelling multivariate daily data at continental level may be a challenging computationl issue, we assess here the consequences of using the so-called one-taper tapering. It is concluded that for the data at hand and software considered, one-taper tapering is not a viable solution as deteriorates inference without a relevant reduction of computation time.

KEYWORDS: Air quality; EM algorithm, Large datasets; Matérn correlation function; Wendland function

1 Data

The European exposure data used here are daily data of airborne pollutants over Europe from background monitoring stations. In particular, we consider seven pollutants, namely benzene (C_6H_6), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matters $(PM_{10}$ and $PM_{2.5})$ and sulphur dioxide (SO_2), during years $(2009-2011)$. The network is heterogeneous with extensive missing data, as shown in Table 1, where we have more than 9×10^6 response variable observations with and more than 2×10^6 missing data. Covariates include Meteo (Wind speed, Pressure and Air Temperature), Land Use, Elevation, Population and Weekend effects.

[1] Università degli Studi di Bergamo, (e-mail: alessandro.fasso@unibg.it)

	C_6H_6	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
Stations	344	591	1978	1800	1837	748	1340
Missing (%)	47.0	29.5	21.5	15.5	23.2	32.2	26.1

Table 1. *Number of stations and missing data rates by pollutant type.*

2 D-STEM model

For jointly modelling the above seven pollutants, Fassò *et al.*, 2015 use a trans-Gaussian hierarchical multivariate spatio-temporal model. To see this, let $\mathbf{y}(\mathbf{s},t)$ be the $q-$dimensional vector of the pollutants concentrations $y_1,...,y_q$ observed at spatial locations $\mathbf{s} \in \mathcal{R}$ and time $t = 1,2,....$ Assuming a log link function, we suppose that $\xi = \log(\mathbf{y})$ is a $q-$dimensional linear spatio-temporal process as defined in Finazzi and Fassò (2014), Section 5. In particular, ξ_i $i = 1,...,7$ are supposed to be locally linearly related to above covariates here denoted by x_{ij}. In fact, the following D-STEM model is used

$$\xi_i(\mathbf{s},t) \qquad = \eta_i(\mathbf{s},t) + \varepsilon_i(\mathbf{s},t) \tag{1}$$

$$\eta_i(\mathbf{s},t) \quad = \omega_{i0}(\mathbf{s},t) + \textstyle\sum_{j=1}^{b} \omega_{ij}(\mathbf{s},t)\, x_{ij}(\mathbf{s},t). \tag{2}$$

for $i = 1,...,7$. In (1), the random variables ε_i are Gaussian distributed with zero mean and variance σ_i^2, uncorrelated over space, time and pollutants. In (2), ω_{i0} is a spatio-temporal stochastic trend and ω_{ij} is the stochastic coefficient of x_{ij}. In particular $\omega' s$ are given by the following model

$$\omega_{ij}(\mathbf{s},t) = \beta_{ij} + z_{ij}(t) + \alpha_{ij} w_{ij}(\mathbf{s},t) \tag{3}$$

where β_{ij} is a constant, $\mathbf{z}(t) = \left(z_{11}(t),...,z_{1b}(t),...,z_{q1}(t),...,z_{qb}(t)\right)'$ has a vector autoregressive model $\mathbf{z}(t) = \mathbf{G}\mathbf{z}(t-1) + \eta(t)$, where $\mathbf{G}$ is a stable $p \times p$ transition matrix and $\eta(t) \sim N_p(\mathbf{0}, \Sigma_\eta)$, $\Sigma_\eta > 0$.

Moreover, in equation (3), $\mathbf{w}_j(\mathbf{s},t) = (w_{1j}(\mathbf{s},t),...,w_{qj}(\mathbf{s},t))'$, $j = 1,...,b$, are zero-mean and unit variance independent Gaussian processes, uncorrelated over time but correlated over space. In particular, each $\mathbf{w}_j(\mathbf{s},t)$, for fixed t, is linear cregionalization model of order 1 with correlation matrix $\mathbf{V}$ and Matern correlation function $\rho(\|\mathbf{s}-\mathbf{s}'\|;\theta_j,\nu)$ parametrized by θ_j and ν.

It follows that, the model parameter vector is $\psi = \left\{\beta, \sigma^2, \alpha, \mathbf{G}, \sigma_\eta, \theta, \mathbf{v}\right\}$, where $\sigma^2 = \left(\sigma_1^2,...,\sigma_q^2\right)'$ are the error measurement variances, $\beta = \left(\beta_{11},...,\beta_{1b},...,\beta_{q1},...,\beta_{qb}\right)'$, $\alpha = \left(\alpha_{11},...,\alpha_{1b},...,\alpha_{q1},...,\alpha_{qb}\right)'$, σ_η is the $p(p+1)/2$ dimensional vector of unique elements of Σ_η, $\theta = (\theta_1,...,\theta_b)'$

and $\mathbf{v}$ is the $bq(q-1)/2$ dimensional vector obtained by stacking the unique and non-diagonal elements of $\mathbf{V}_1,...,\mathbf{V}_b$. The trans-Gaussian transformation is omitted here being inessential for the results.

3 Tapering assessment

Estimation of ψ in model $(1)-(2)$ can be efficiently performed using the EM algorithm of Fassò & Finazzi, 2011 and D-STEM software discussed in Finazzi & Fassò, 2014. Nonetheless the data of section 1 come from $S = 8'638$ sensors and $T = 1'095$ days. Hence the EM algorithm, in order to handle missing data, must solve a linear system $8'638 \times 8'638$ for each iteration and each day.

In the last years compactly supported spatial covariances and tapering techniques are deserving more and more attention, especially in the multivariate case. See e.g. Masry, 2014 and Furrer *et al.*, 2015. Hence the computationally efficient one-taper tapering is considered here in order to assess the computational gain and the precision loss. In practice in the EM computation for model $(1)-(2)$ the spatial variance-covariance matrix is computed substituting the Matérn correlation function ρ with

$$\rho(\left\|\mathbf{s}-\mathbf{s}'\right\|;\theta_j,\nu)\cdot\Phi\left(\frac{\|\mathbf{s}-\mathbf{s}'\|}{\phi}\right),$$

where $\Phi\left(\frac{\|\mathbf{s}-\mathbf{s}'\|}{\phi}\right)$ is the compactly supported radial Wendland function (Wendland & Mathematik, 1995), with ϕ the tapering range. Note that, unlike the so called two-tapers tapering of Kaufman *et al.*, 2008 , the one-taper may give biased estimate of θ in case the tapering range ϕ is small. However, the former has a substantially heavier computational burden while the one-taper approach gives better kriging performance, which is an important aim for exposure distribution estimation.

In order to develop a cost/benefit analysis of the one-taper tapering, we estimated and validate seven univariate models, one for each pollutant and for various tapering ranges ϕ. Moreover, due to computational burden, we performed a limited number of iterations for the seven-variate model.

From the result table, not reported here for shortage of space reasons, we see that a shrinkage of the tapering range from 500km to 100km induces a bias in the estimate of θ and a loss in performance as measured by $R_i^2 = 1-MSE(\hat{y}_i)/Var(y_i), i = 1,...,7$, computed in validation dataset as discussed in Fassò *et al.*, 2015. As expected, this is especially true when θ is large, e.g.

for PM_{10}. On the other side computing time is larger when sparsity is not very high and the gain for $\phi = 100 km$ is miserable. This is due to the fact that algorithms for sparse matrices are efficient when sparsity is high, e.g. above 95%. Additionally, each EM iteration involved in the space-time model estimation is based on multiple and complex operations (including the Kalman smoother) and not all of them exploit the matrix sparsity. The gain in using sparse matrices is thus different from the results given in Vetter *et al.*, 2014 and Furrer *et al.*, 2006, where only spatial models are considered.

References

FASSÒ, A., & FINAZZI, F. 2011. Maximum Likelihood Estimation of the Dynamic Coregionalization Model with Heterotopic Data. *Environmetrics*, **22**(6), 735–748.

FASSÒ, A., FINAZZI, F., & NDONGO, F. 2015. Multivariate spatio temporal models for exposure to airborne multipollutants in Europe. *submitted.*

FINAZZI, F., & FASSÒ, A. 2014. D-STEM: A Software for the Analysis and Mapping of Environmental Space-Time Variables. *Journal of Statistical Software*, **62**(6), 1–25.

FURRER, R., GENTON, M.G., & NYCHKA, D. 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.

FURRER, R., BACHOC, F., & J., DU. 2015. Asymptotic properties of multivariate tapering for estimation and prediction. *arXiv:1506.01833.*

KAUFMAN, C.G., SCHERVISH, M.J., & NYCHKA, D.W. 2008. Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. *Journal of the American Statistical Association*, **103**(484), 1545–1555.

MASRY, E. 2014. Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment*, **4**, 1249–1263.

STEIN, M.L. 2013. Statistical Properties of Covariance Tapers. *Journal of Computational and Graphical Statistics*, **222**(4), 866–885.

VETTER, P., SCHMID, W., & SCHWARZE, R. 2014. Efficient Approximation of the Spatial Covariance Function for Large Datasets - Analysis of Atmospheric CO2 Concentrations. *Journal of the American Statistical Association*, **6**(3), 1–36.

WENDLAND, H., & MATHEMATIK, A. 1995. Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree. *Advances in Computational Mathematics*, **4**, 389–396.