Semiparametric Approaches for Modelling Air Pollution Processes

O. Bodnar[1] and W. Schmid[1,*]

[1] *Department of Statistics, European University Viadrina, PO Box 1786, 15207 Frankfurt (Oder), Germany; obodnar@euv-frankfurt-o.de, schmid@euv-frankfurt-o.de*
[*] *Corresponding author*

Abstract. *In this paper we introduce a semiparametric spatio-temporal process for modelling the PM_{10} concentration. Estimators for the parameters are obtained by making use of the unscented Kalman filter of Julier et al. (2000) and the generalized EM algorithm. The nonparametric component is estimated by a Nadaraya-Watson type estimator. Moreover, an interpolation procedure is proposed which is based on a local approach in the sense of Bodnar and Schmid (2010). In an empirical study this model is applied to describe the PM_{10} content of the Berlin-Brandenburg region in Germany.*

Keywords. *Spatio-temporal processes; Semiparametric model; Non-linear Kalman filter; Kriging; Environmental statistics.*

1 Introduction

In the last years spatio-temporal processes have been intensively discussed in literature (e.g., Stroud et al. (2001), Le and Zidek (2006)). They have turned out to be extremely useful for modelling environmental processes. Nowadays we find a variety of applications to different types of processes like, e.g., atmospheric pollutant concentrations, precipitation fields and surface winds.

Fassò and Cameletti (2009) introduced a very general spatio-temporal process for modeling the concentration of PM_{10}. It can be presented as a state-space model. In this model it is assumed that there is a linear relationship between the PM_{10} concentration and the geographical and meteorological covariates. In Bodnar and Schmid (2010) a similar model is considered and it is used to calculate the locally weighted scatterplot smoothing (LOESS) kriging predictor. Their approach is based on the idea to find a balance between a local and a more global method. This means that necessarily not all available measurement stations are used to interpolate the process at an arbitrary position. Criteria for finding an

optimal set of included stations are discussed in the paper.

In the present paper we introduce a semiparametric spatio-temporal model. It can be considered as an extension of the models discussed in Fassò and Cameletti (2009) and Bodnar and Schmid (2010). This model is more flexible because it is not assumed that there is a linear relationship between the concentration and the geographical and meteorological covariates. It has further desirable properties. For instance, it is guaranteed that the concentration process is always positive and the present concentration depends directly on previous values. The price of the generalization is a process which is more difficult to handle. Because both the state and the space equations are non-linear, the standard Kalman filter cannot be used for estimating the model parameters. For that reason we make use of the unscented Kalman Filter suggested by Julier et al. (2000). This method is combined with the EM algorithm and a Nadaraya-Watson type estimator for the nonparametric component to get estimators of the process parameters. Furthermore we consider the problem of interpolating the process at arbitrary locations. This is done in a similar way as described in Bodnar and Schmid (2010). Our results are applied to model the PM_{10} concentration of the Berlin-Brandenburg region in Germany.

2 A Semiparametric Spatial-Temporal Process

Let $Z_t(\mathbf{s})$ denote an observed univariate spatio-temporal process at the geographical location $\mathbf{s}$ at time t. The network data at time point t at the geographical locations $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_n$ are written as $\mathbf{Z}_t = (Z_t(\mathbf{s}_0), Z_t(\mathbf{s}_1), ..., Z_t(\mathbf{s}_n))'$. In the present paper we introduce an extension of the general spatio-temporal model of Fassò and Cameletti (2009) and Bodnar and Schmid (2010). The model is given by

$$Z_t(\mathbf{s}) = U_t(\mathbf{s}) + \varepsilon_t \tag{1}$$

$$U_t(\mathbf{s}) = \tilde{Y}_t + \mu(\mathbf{X}_t(\mathbf{s}); \beta) + \sum_{i=1}^{m} \alpha_i U_{t-i}(\mathbf{s}) + \omega_t(\mathbf{s}) \tag{2}$$

$$log(\tilde{Y}_t) = \mu_0 + g(log(\tilde{Y}_{t-1}) - \mu_0) + \eta_t \tag{3}$$

for $t \in \{1, 2, ..., T\}$ with $E(log(\tilde{Y}_0)) = \mu_0$.

In equation (1) it is described how the observed spatio-temporal process $\{Z_t(\mathbf{s})\}$ is related to the unobservable "true" spatio-temporal process $\{U_t(\mathbf{s})\}$. The variables $\{\varepsilon_t\}$ denote a measurement error process with no spatial component. $\{\varepsilon_t\}$ is assumed to be independent and identically distributed with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2$.

In the second equation the unobservable "true" spatio-temporal process is modeled as a function $\mu = \mu(\mathbf{X}_t(\mathbf{s}); \beta)$ of the covariates $\mathbf{X}_t(\mathbf{s})$ at time t and at site $\mathbf{s}$, preceeding values of the "true" process, a process $\{\tilde{Y}_t\}$, and a spatial noise process $\{\omega_t(\mathbf{s})\}$. It is assumed to have mean zero and a covariance function given by

$$Cov(\omega_t(\mathbf{s}), \omega_t(\mathbf{s}')) = \sigma_\omega^2 \Gamma(||\mathbf{s} - \mathbf{s}'||; \theta)$$

where θ is a parameter and $||\mathbf{s} - \mathbf{s}'||$ denotes the Euclidean distance between the sites $\mathbf{s}$ and $\mathbf{s}'$. The symbol $C_\theta(.)$ stands for the covariogramm which is assumed to be isotropic. Moreover, it is assumed to be independent over time.

The process $\{\tilde{Y}_t\}$ is explained in (3). It stands for the space-constant temporal process of the analyzed region. Because the process is always taking positive values it is modeled via the logarithm in (3). Let $\sigma_0^2 = Var(log(\tilde{Y}_0))$. The error process $\{\eta_t\}$ is assumed to be independent and identically distributed with

mean 0 and variance σ_η^2. Finally, it is assumed that all of the three error processes $\{\varepsilon_t\}$, $\{\omega_t(\mathbf{s})\}$, and $\{\eta_t\}$ are mutually independent.

In equation (2) the function μ stands for a regression function describing the influence of the covariates. The most general approach would be to assume that there is an arbitrary relationship μ between the variables. Then it would be necessary to estimate μ by a d-dimensional nonparametric estimator which is not quite easy. For that reason we prefer a semiparametric method which reduces the dimension of the estimation problem and thus its complexity. The most common semiparametric models (cf. Härdle et al. (2004)) are the additive model with $\mu(X_{t1},...,X_{td}) = \sum_i^d g_i(X_{it})$ where the functions $g_i(.)$ should be estimated nonparametrically, the partial linear model with $\mu(X_{t1},...,X_{td}) = \tilde{\mathbf{X}}_t'\tilde{\beta} + \breve{g}(\breve{\mathbf{X}}_t)$, $\tilde{\mathbf{X}}_t = (X_{t1},...,X_{t\tilde{d}})'$ and $\breve{\mathbf{X}}_t = (X_{t\tilde{d}+1},...,X_{td})'$, where the function $\breve{g}(.)$ is estimated nonparametrically and the vector $\tilde{\beta}$ is calculated by using a parametric method and finally the single index model with $\mu(X_{t1},...,X_{td}) = g(\mathbf{X}_t'\beta)$ where $g(.)$ is estimated non-parametrically and β parametrically. In the following the single index model is applied for describing the concentration of PM_{10} as a function of the geographical and meteorological covariates. This means that we put $\mu(\mathbf{X}_t(\mathbf{s});\beta) = \mu(\mathbf{X}_t(\mathbf{s})'\beta)$.

Let $Y_t = log(\tilde{Y}_t) - \mu_0$. Rewriting (1) to (3) we get

$$Z_t(\mathbf{s}) = \nu exp(Y_t) + \mu(\mathbf{X}_t(\mathbf{s})'\beta) + \sum_{i=1}^m \alpha_i Z_{t-i}(\mathbf{s}) + e_t(\mathbf{s}) \quad (4)$$

$$Y_t = gY_{t-1} + \eta_t \quad (5)$$

where $e_t = \omega_t(\mathbf{s}) + \varepsilon_t - \sum_{i=1}^m \alpha_i \varepsilon_{t-i}$ and $\nu = exp(\mu_0)$. Note that the equations (4) and (5) define a non-linear state-space model (see, e.g., Julier et al. (2000)).

3 Estimation and Empirical Illustration

The quasi maximum likelihood estimator of $\Psi = (\beta, \alpha, \sigma_\omega^2, g, \sigma_\eta^2, \nu, \sigma_0^2, \theta, \gamma)'$ with $\alpha = (\alpha_1,...,\alpha_m)'$ is calculated using an iterative procedure and the generalized EM algorithm (e.g., McLachlan and Krishnan (1997)). Using data of the l-nearest monitoring stations the quasi log-likelihood function is expressed as

$$\begin{aligned} log\,L(\Psi;\mathbf{z}_1,..,\mathbf{z}_T) = &-\frac{lT}{2}\log(2\pi) - \frac{T}{2}\log(|\Sigma_e|) - \frac{1}{2}\sum_{t=1}^T (\mathbf{z}_t(\mathbf{s}) - \nu exp(Y_t)\mathbf{1} - \mu(\mathbf{X}_t(\mathbf{s})'\beta) - \sum_{i=1}^m \alpha_i \mathbf{z}_{t-i}(\mathbf{s}))' \\ \times\ & \Sigma_e^{-1}(\mathbf{z}_t(\mathbf{s}) - \nu exp(Y_t)\mathbf{1} - \mu(\mathbf{X}_t(\mathbf{s})'\beta) - \sum_{i=1}^m \alpha_i \mathbf{z}_{t-i}(\mathbf{s})) \\ - \ & \frac{T}{2}(\log(\sigma_\eta^2) - \frac{1}{2T\sigma_\eta^2}\sum_{t=1}^T (Y_t - gY_{t-1})^2) - \frac{1}{2}\log(\sigma_0^2) - \frac{Y_0^2}{2\sigma_0^2}, \end{aligned} \quad (6)$$

where $\Sigma_e = \sigma_\omega^2 (\Gamma(||\mathbf{s}_i - \mathbf{s}_j||;\theta))_{i,j=0,...,l}$ with

$$\Gamma(h) = \begin{cases} 1+\gamma & \text{for} \quad h = 0 \\ C_\theta(h) & \text{for} \quad h > 0 \end{cases}$$

and $\gamma = \sigma_\varepsilon^2/\sigma_\omega^2(1 + \sum_{i=1}^m \alpha_i^2)$.

For maximizing (6) we use the EM algorithm. At each step of iteration $k = 1,2,...$ the EM algorithm consists of an expectation (E) and a maximization (M) step. In the E-step we calculate the conditional expectation of the log-likelihood function given the data $\mathbf{Z} = \{\mathbf{z}_1,..,\mathbf{z}_T\}$ and the vector of estimated

parameters in the previous step $\hat{\Psi}^{(k)}$, i.e.

$$\begin{aligned} Q(\Psi) &= -2E_{\mathbf{Z},\Psi}(logL(\Psi;\mathbf{z}_1,..,\mathbf{z}_T)) = lT\log(2\pi) + T\log(|\Sigma_e|) \\ &+E_{\mathbf{Z},\Psi}(\sum_{t=1}^{T}(\mathbf{z}_t(\mathbf{s}) - \nu exp(Y_t)\mathbf{1} - \mu(\mathbf{X}_t(\mathbf{s})'\beta) - \sum_{i=1}^{m}\alpha_i\mathbf{z}_{t-i}(\mathbf{s}))'\Sigma_e^{-1}(\mathbf{z}_t(\mathbf{s}) - \nu exp(Y_t)\mathbf{1} - \mu(\mathbf{X}_t(\mathbf{s})'\beta) - \sum_{i=1}^{m}\alpha_i\mathbf{z}_{t-i}(\mathbf{s}))) \\ &+ T(\log(\sigma_\eta^2) + \frac{1}{T\sigma_\eta^2}\sum_{t=1}^{T}E_{\mathbf{Z},\Psi}((Y_t - gY_{t-1})^2)) + \log(\sigma_0^2) + \frac{E_{\mathbf{Z},\Psi}(Y_0^2)}{\sigma_0^2} \end{aligned}$$

with $E_{\mathbf{Z}}(Y_t) = y_t^T$, $Var_{\mathbf{Z}}(Y_t) = P_t^T$, and $Cov_{\mathbf{Z}}(Y_t, Y_{t-1}) = P_{t,t-1}^T$. The quantities y_t^T, P_t^T, and $P_{t,t-1}^T$ are calculated recursively using $\hat{\Psi}^{(k-1)}$. For the calculation of these quantities we use the unscented Kalman filter. Since $\mu(.)$ is an unknown function we replace $\mu(.)$ by the Nadaraya-Watson type estimator (see, e.g., Härdle et al. (2004)) given by

$$\hat{\mu}(\mathbf{X}_t(\mathbf{s}_i)'\beta) = \frac{\sum_{(\tilde{i},\tilde{t})\neq(i,t)} Z_{\tilde{i}}(\mathbf{s}_{\tilde{i}})K_h((\mathbf{X}_{\tilde{i}}(\mathbf{s}_{\tilde{i}}) - \mathbf{X}_t(\mathbf{s}_i))'\beta)I_{\chi(t,\mathbf{s}_i)}(\mathbf{X}_{\tilde{i}}(\mathbf{s}_{\tilde{i}}))}{\sum_{(\tilde{i},\tilde{t})\neq(i,t)} K_h((\mathbf{X}_{\tilde{i}}(\mathbf{s}_{\tilde{i}}) - \mathbf{X}_t(\mathbf{s}_i))'\beta)I_{\chi(t,\mathbf{s}_i)}(\mathbf{X}_{\tilde{i}}(\mathbf{s}_{\tilde{i}}))},$$

where $K_h(.)$ is a kernel function and $I_{\chi(t,\mathbf{s}_i)}(.)$ is the indicator function of the set $\chi(t,\mathbf{s}_i) = \{\mathbf{X} : ||\mathbf{X} - \mathbf{X}_t(\mathbf{s}_i)|| \leq h\}$. For computing $\hat{\Psi}^{(k+1)}$ the Newton-Raphson algorithm is used to minimize $Q(\Psi)$.

In an empirical study the suggested model is applied to model the PM10 concentration in the Berlin-Brandenburg region of Germany. For interpolating the values of the PM_{10} concentration at positions where no monitoring station is available, the conditional kriging predictor is derived for the model (1)-(3). Eleven covariates are taken into account, namely the type of the station, the weekend effect, the height above sea level, the temperature, the atmospheric pressure, the wind direction, the wind power, the wind velocity, the cloud cover, the sunshine duration, and the precipitation. Cross-validation is applied for identifying the number of the neighboring stations used for interpolating the PM_{10} concentration and the most relevant covariates.

References

[1] Bodnar, O. and Schmid, W. (2010) Nonlinear locally weighted kriging prediction for spatio-temporal environmental processes. To appear in *Environmetrics*.

[2] Fassò, A. and Cameletti, M. (2009) The EM algorithm in a distributed computing environment for modelling environmental space-time data. *Environmental Modelling & Software* **24**, 1027–1035.

[3] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004) *Nonparametric and semiparametric models*. Springer. Berlin, Heidelberg.

[4] Julier, S., Uhlmann, J. and Durrant-White, H.F. (2000) A new method for nonlinear transformation of means and covariancesin filters and estimators. *IEEE Transactions on Automatic Control* **45**, 477–482.

[5] Le, N.D. and Zidek, J.V. (2006) *Statistical analysis of environmental space-time processes*. Springer. New York.

[6] McLachlan, G.J. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley. New York.

[7] Stroud, J., Muller, P. and Sansò, B. (2001) Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B.* **63**, 673–689.