Spatio-Temporal Expectile Regression Models

Elmar Spiegel
University of Goettingen

Abstract

Spatio-temporal models are becoming increasingly popular in recent regression research. However, they usually rely on the assumption of a specific parametric distribution and homoscedastic error terms. In this paper we propose to apply semiparametric expectile regression to model spatio-temporal effects. Besides the removal of the assumption of a specific distribution and homoscedasticity, with expectile regression the whole distribution of the response is estimated and not just the mean. For the use of expectiles we interpret them as weighted means and estimate them by standard tools of the least squares regression. The spatio-temporal effect is set up as a three-dimensional interaction between time and space based on P-splines. Thus, the model can be split up into main effects and interactions. The method is presented with the analysis of spatio-temporal variation of temperatures in Germany from 1980 to 2014.

Keywords: generalized additive model, expectile regression, spatio-temporal model, p-spline, main effects

1 Introduction

In applied science data are often recorded at several locations and multiple time points. Although in basic reports the dimensions are reduced and only aggregated values are published (see for example World Meteorological Organization, 2017). Simple statistical models estimate the temporal and the spatial effects as additive models (see for example Fahrmeir et al., 2004). Hence, the impact of time and space is estimated independently and variations of the spatial effect depending on time are not considered. Thus, statisticians developed several approaches to incorporate both time and space jointly and in interaction, the so-called *spatio-temporal* models. Since these models are rather complex and computationally demanding Cressie and Wikle (2011) called them the "next frontier". In their book they explain ideas how to estimate spatio-temporal Kriging models. Other publications that tackled the "frontier" use P-splines as introduced by Eilers and Marx (1996). A first step towards spatio-temporal models was the introduction of two-dimensional P-splines as tensor products by Eilers and Marx (2003). The extension of the two-dimensional case to the three-dimensional case, the spatio-temporal model, is straightforward. However, Wood (2006) was among the first to discuss the three-dimensional splines in detail. Furthermore, he developed an alternative penalization which is based on a theoretic understanding of smoothness in the larger dimensions. The separation into main effects and interaction effects is important for the practical use of spatio-temporal models. Therefore, Wood (2006), Lee and Durbán (2009), Lee and Durbán (2011) and Wood et al. (2013) developed several approaches mostly relying on the representation of P-splines as mixed models (see Fahrmeir et al., 2004, for example). Several other papers deal with interactions of smooth effects, like Gu (2002), where tensor products of smoothing splines are discussed. However, in Gu (2002) the penalty term remains the integral of the second derivative, such that the estimation bases on more complicated techniques like reproducing Kernel-Hilbert Spaces.

Contrarily, P-splines as introduced in Eilers and Marx (1996) are easier to calculate due to their approximation of the penalty as differences of the coefficients.

In most of the articles a Gaussian distribution for the error terms is assumed and only some refer to other distributions of the exponential family. Even if the error distribution was specified correctly, the assumption of homoscedasticitiy of the errors might still not be fulfilled. If we suppose that the reasons for heteroscedasticity are measured by some covariates generalized additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005; Stasinopoulos et al., 2017) could be applied. These models have separate regression predictors assigned to each parameter of the distribution. Umlauf et al. (2016) model the spatio-temporal distribution of rain in Austria with help of the Bayesian version of GAMLSS (Klein et al., 2015). Therefore, besides the mean they also model the variance parameters of the normal distribution with some spatio-temporal trend. This model can be used to show that the variance increases for specific regions or times. In the GAMLSS framework a large variety of distributions is available. So the data can be modeled quite flexibly. However, the model always depends on the correct choice of the distribution and the link functions. Due to the complex design of the predictors these choices are non-trivial (Rigby et al., 2013). An alternative that also deals with heteroscedasticity is expectile regression as introduced by Newey and Powell (1987). With expectiles we do not assume a specific distribution and account for heteroscedasticity by putting more or less emphasis on specific parts of the distribution. Therefore, this model is very flexible and omits the specification of a parametric distribution. Basically, expectile regression is a weighted least squares regression, where the weights depend on the observations and the fitted values (for details see Section 2). Quantile regression is a similar alternative to model effects beyond the mean without distributional assumption. Since quantiles are defined as generalization of the median, while expectiles are a generalization of the mean, they are easier to interpret, but harder to estimate, in particular in smoothed settings.

Thus, we will use expectile regression to analyze the spatio-temporal trend of temperature in Germany. We estimate the distribution of temperatures depending on time and location as in previous spatio-temporal models. Based on expectile regression we further determine, where especially cold winters occur and which areas have relatively hot summers. Additionally to the detection of increased variance in some regions we may also specify the direction of the divergence from the mean.

In the remainder of this article we start with a brief introduction to expectile regression in Section 2. Afterwards, we recapture the ideas of semiparametric models, including spatio-temporal models, in Section 3. In Section 4 we summarize a small simulation study on the smoothing parameter selection in semiparametric expectile regression with interactions. As an example the spatio-temporal analysis of temperatures in Germany is displayed in Section 5. Finally, we conclude with a discussion in Section 6.

2 Expectile Regression

The classical linear model is based on the assumption of homoscedasticity. If this assumption is violated several possibilities to model the covariate effects are possible. As discussed in the introduction we will apply expectile regression as introduced by Newey and Powell (1987) in this article. Theoretically an expectile e_τ for some given density function g is defined as

$$e_\tau = \frac{(1-\tau)G(e_\tau) + \tau(\mu - G(e_\tau))}{(1-\tau)F(e_\tau) + \tau(1 - F(e_\tau))}$$

where μ is the ordinary mean and $F(x) = \int_{-\infty}^{x} g(u)\ du$ the cumulative distribution function and $G(x) = \int_{-\infty}^{x} u\ g(u)\ du$ the partial moment function.

In a standard expectile regression model the density g is unknown so we use connections to the classical least squares estimation, since expectiles are a generalization of the mean. In ordinary least squares models the squared residuals should be minimized. Taking the derivative of the least squares equation with respect to the coefficients and setting it to zero solves the problem. Furthermore, rearranging the derivative with respect to the intercept, shows that the sum of the residuals must be 0. Thus, the solution will be the predictor, where the sum of the residuals above and below are equal. Hence, the center of gravity is estimated. In expectile regression the emphasis is now put on outer parts of the distribution to detect variation of the effects beyond the mean. Therefore, in the least squares equation a weight $w_\tau(y_i)$ is included such that observations y_i below the fitted effect $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau$ get another weight than the observations above

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} w_\tau(y_i) \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau\right)^2, \tag{1}$$

where the weights are defined as

$$w_\tau(y_i) = \begin{cases} \tau & \text{if} \quad y_i \geq \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau \\ 1-\tau & \text{if} \quad y_i < \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau \end{cases}.$$

Thereby the predictor $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau$ depends on the specified asymmetry parameter τ. Based on this definition the 50% expectile is the ordinary mean. The fitted values then define the weighted center of gravity. As discussed before in classical linear regression the error terms $\varepsilon_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$ are assumed to be identical and independent normal distributed ($\varepsilon_i \sim N(0, \sigma^2)$). Contrarily in expectile regression we do not assume any distribution for the error terms $\varepsilon_{i,\tau} = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau$, nor do we assume identical distributed error terms. The only constraint is that the expectiles of the error terms themselves are 0, given the estimated expectile $\hat{e}_{i,\tau}$

$$\mathbb{E}\left(w_\tau(\varepsilon_{i,\tau})\left(\varepsilon_{i,\tau} - \hat{e}_{i,\tau}\right)^2\right) = 0.$$

Since Equation (1) can be written in matrix notation as

$$\hat{\boldsymbol{\beta}}_\tau = \left(\boldsymbol{X}^\top \boldsymbol{W}_\tau \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W}_\tau \boldsymbol{y}$$

where $\boldsymbol{W}_\tau$ is the diagonal matrix of weights $w_\tau(y_i)$, it is obvious that for the estimation of $\hat{\boldsymbol{\beta}}_\tau$ the standard weighted least squares techniques can be applied. The only problem is that $\boldsymbol{W}_\tau$ depends on $\hat{\boldsymbol{\beta}}_\tau$ and reverse. So the optimization of $\hat{\boldsymbol{\beta}}_\tau$ and $w_\tau(y_i)$ is done iteratively with an algorithm called least asymmetric weighted squares (LAWS, Newey and Powell, 1987), where we start with a classical linear model with equal weights for all observations. Afterwards, the new weights are estimated and a new weighted linear model is fitted. The estimation of weights and coefficients is iterated until the weights remain unchanged.

Additionally to the pure estimated effects their uncertainty is usually of special interest. Sobotka et al. (2013) showed that in expectile regression the estimated coefficients $\hat{\boldsymbol{\beta}}_\tau$ follow a normal distribution

$$\hat{\boldsymbol{\beta}}_\tau \sim N(\boldsymbol{\beta}_\tau, Var(\boldsymbol{\beta}_\tau))$$

where $Var(\boldsymbol{\beta}_\tau)$ has to be estimated appropriately. This approach is not restricted to linear effects. It can be adopted for smooth effects, similarly as in Marra and Wood (2012). However, in the setting of spatio-temporal models the number of observations is usually huge (in our example we have more than 4.3e6 observations) therefore the confidence intervals will be very small. Thus, they are neglected in the following.

Alternatively, quantile regression (Koenker and Bassett, 1978) can be used to estimate models beyond the mean. Therefore, in Equation (1) the l_2-norm is exchanged with the l_1-norm. In quantile regression the fitted effects represent the line where the ratio of numbers of observations below and above is the wanted fraction τ, while in expectile regression the fitted values give the weighted center of gravity, so the line where the sum of the weighted distances below and above is the given fraction is τ (Yao and Tong, 1996). So quantile regression only checks how many observations are below and above the fitted values. The distance between the fitted values and the observations is not taken into account. Since expectiles account also for the values of the distances it uses more information. Moreover, applying quantile regression (Koenker and Bassett, 1978; Koenker, 2005) instead of expectile regression in this setting would hardly be possible, due to the smooth three-dimensional interactions of space and time, which will be applied to model spatio-temporal effects. Estimating those would be computationally quite burdensome for the linear programming routines on which quantile regression relies. This might be the reason why we have not found any publication on spatio-temporal quantile regression. Although quantiles might be easier to interpret we rely on expectiles due to the limits of quantile regression in smoothed settings.

3 Additive Models

3.1 Basis Functions

In standard models the effect for each covariate is defined to be linear, or a polynomial of the original variable. This is often not sufficient to cover the true underlying effect, which results in biased estimates. Hastie and Tibshirani (1986) however introduced the class of generalized additive models (GAM), where the effect per covariate is defined as a smooth curve. One possibility to specify the smooth curve $f(x_1)$ for a covariate x_1 is to build a set of basis functions $B_{j_1}(x_1)$, $j_1 = 1, \dots, J_1$ and scale them with an estimated parameter γ_{j_1}. The resulting sum is the smooth curve

$$f(x_{i1}) = \sum_{j_1=1}^{J_1} B_{j_1}(x_{i1})\gamma_{j_1} = \boldsymbol{B}_{i1}^\top \boldsymbol{\gamma}.$$

The function can be written in matrix notation with $\boldsymbol{B}_{i1} = (B_1(x_{i1}), \dots, B_{J_1}(x_{i1}))^\top$. Since $B_{j_1}(x_1)$ can be treated as a new covariate, the coefficients $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{J_1})^\top$ are estimated based on the usual least squares approach. Several smooth effects can also be included in a model additively (for a detailed introduction into splines see Wood, 2017). Different functions define proper basis functions, including B-splines (de Boor, 1978) and thin plate splines (Duchon, 1977).

Additionally to the additive model of smooth one-dimensional effects, the smooth interaction between two covariates x_1, x_2 is regularly wanted. In the analysis of weather data for example the spatial effect should be an interaction between the north-south and east-west effect. Therefore, we would like to have a smooth interaction surface between both effects. This means we would like to define a function $f(x_1, x_2)$ for this surface.

Based on the ideas of splines from above we define the interacting surface as

$$f(x_{i1}, x_{i2}) = \sum_{j=1}^{J} B_j(x_{i1}, x_{i2})\gamma_j$$

where $B_j(x_1, x_2)$ is a two-dimensional basis function. One possibility to define $B_j(x_1, x_2)$ is to reduce the two-dimensional basis function $B_j(x_1, x_2)$ to be a product of two one-dimensional basis functions $B_{j_1}(x_1)$ and $B_{j_2}(x_2)$ in direction of x_1 and x_2 respectively (Eilers and Marx, 2003). Thus, we get the two-dimensional surface as

$$f(x_{i1}, x_{i2}) = \sum_{j=1}^{J} B_j(x_{i1}, x_{i2})\gamma_j = \sum_{j_1=1}^{J_1}\sum_{j_2=1}^{J_2} B_{j_1}(x_{i1})B_{j_2}(x_{i2})\gamma_{j_1,j_2} = (\boldsymbol{B}_{i1} \otimes \boldsymbol{B}_{i2})^\top \boldsymbol{\gamma}$$

where $\boldsymbol{\gamma} = (\gamma_{1,1}, \gamma_{1,2}, \ldots, \gamma_{2,1}, \ldots)^\top$ is the new vector of coefficients with appropriate ordering. Similar to the one-dimensional case we can write this in matrix notation with $\otimes$ the (row-wise) Kronecker product of the one-dimensional matrices. More details on two-dimensional surfaces can be found in Fahrmeir et al. (2013) and Wood (2017).

3.2 Penalization

In basic spline regression, as defined in Section 3.1, the number and location of the basis functions need to be optimized. This is challenging for one-dimensional splines and nearly impossible in higher dimensions. Consequently Eilers and Marx (1996) introduced a technique called P-splines for one-dimensional smooth functions where they start with a high number of basis functions but restrict the curves to be smooth. Therefore, they penalize the wiggliness of the curves by adding a penalty term to the least squares argument such that not only the optimal model fit is a criterion, but also the smoothness of the function. Additionally, this method has the advantage that the locations and the number of basis function do not need to be optimized anymore. Overall, smoothness is defined as the integrated second derivative of the function. For the one-dimensional case with only one covariate this results in

$$\lambda_1 \int \left(f''(x_1)\right)^2 dx_1$$

where λ_1 is a scalar parameter which indicates the influence of the smoothness penalty on the penalized least squares criterion

$$\sum_{i=1}^{n} (y_i - f(x_{i1}))^2 + \lambda_1 \int \left(f''(x_1)\right)^2 dx_1.$$

Estimating the second derivative of the unknown function is challenging. However, for B-splines Eilers and Marx (1996) showed that the integral can be approximated by the sum of the coefficients second order differences $\lambda_1 \sum_{j_1=3}^{J_1} (\gamma_{j_1} - 2\gamma_{j_1-1} + \gamma_{j_1-2})^2$. By applying $\boldsymbol{K}_1$ a penalty matrix that maps $\boldsymbol{\gamma}$ to the penalization term, the penalized least squares criterion shrinks to

$$\sum_{i=1}^{n} (y_i - \boldsymbol{B}_i^\top \boldsymbol{\gamma})^2 + \lambda_1 \boldsymbol{\gamma}^\top \boldsymbol{K}_1 \boldsymbol{\gamma}$$

which can be estimated by standard routines. Additionally more covariates can be included to the model either as linear effects or as smooth effects, so the semiparametric predictor is defined as

$$\eta_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \dots.$$

Each of the smooth effects needs a penalty for controlling the wiggliness. Thus, the penalized least squares criterion of the additive model is given as

$$\sum_{i=1}^{n} (y_i - \eta_i)^2 + \lambda_1 \boldsymbol{\gamma}_1^\top \boldsymbol{K}_1 \boldsymbol{\gamma}_1 + \lambda_2 \boldsymbol{\gamma}_2^\top \boldsymbol{K}_2 \boldsymbol{\gamma}_2 + \dots,$$

where $\boldsymbol{\gamma}_1$ are coefficients corresponding to the first smooth effect and so on. To get the best model fit the smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)^\top$ have to be optimized. Therefore, either goodness-of-fit criteria like the generalized cross-validation criterion (GCV) or methods based on the Schall algorithm (Schall, 1991) can be applied (see Wood, 2017, for details).

Based on the one-dimensional penalties the two-dimensional splines can also be penalized. Eilers and Marx (2003) suggest to use the sum of the squared differences of the coefficients in each direction to obtain a valid penalty. In detail they propose to apply the joint penalty

$$\begin{aligned} P &= \lambda_1 \sum_{j_2=1}^{J_2} \sum_{j_1=3}^{J_1} (\gamma_{j_1,j_2} - 2\gamma_{j_1-1,j_2} + \gamma_{j_1-2,j_2})^2 + \lambda_2 \sum_{j_1=1}^{J_1} \sum_{j_2=3}^{J_2} (\gamma_{j_1,j_2} - 2\gamma_{j_1,j_2-1} + \gamma_{j_1,j_2-2})^2 \\ &= \boldsymbol{\gamma}^\top (\lambda_1 \boldsymbol{K}_1 \otimes \boldsymbol{I}_{J_2} + \lambda_2 \boldsymbol{I}_{J_1} \otimes \boldsymbol{K}_2) \boldsymbol{\gamma} \end{aligned}$$

to reduce the surfaces wiggliness. Therefore, $\boldsymbol{K}_k$ are the penalty matrices in each direction $k = 1, 2$ and $\boldsymbol{I}_{J_k}$ are unit matrices of dimension of the number of basis function in the other direction. More details on this idea can also be found in Fahrmeir et al. (2013). Beside this approach Wood (2006) proposes to define the penalty as

$$\tilde{P} = \int_{x_1 x_2} \tilde{\lambda}_1 \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \tilde{\lambda}_2 \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2.$$

This definition turns out to be similar to the definition from Eilers and Marx (2003), but a reparameterization of the penalty and the basis functions has to be applied. Furthermore, using $\tilde{P}$ instead of P often results in smaller MSE. However, the reparameterization is numerically instable in our example such that we skip it here and refer to Wood (2006) for further details.

3.3 Spatio-Temporal Models

Correspondingly to the interaction between two covariates we can use the above strategy to build interactions in any dimensions (compare Wood, 2006). An application of a three-dimensional interaction is the temporal variation of a spatial effect. This is the so called spatio-temporal model. Therefore, we build the three-dimensional smooth surface based on the one-dimensional basis functions as

$$f(\text{time}_i, \text{lon}_i, \text{lat}_i) = (\boldsymbol{B}_{\text{time},i} \otimes \boldsymbol{B}_{\text{lon},i} \otimes \boldsymbol{B}_{\text{lat},i})^\top \boldsymbol{\gamma} \tag{2}$$

with *lon* the longitudinal and *lat* the latitudinal coordinate of the observation station. Moreover the penalty term is then defined as

$$P = \boldsymbol{\gamma}^\top (\lambda_{\text{time}} \boldsymbol{K}_{\text{time}} \otimes \boldsymbol{I}_{J_{\text{lon}}} \otimes \boldsymbol{I}_{J_{\text{lat}}} + \lambda_{\text{lon}} \boldsymbol{I}_{J_{\text{time}}} \otimes \boldsymbol{K}_{\text{lon}} \otimes \boldsymbol{I}_{J_{\text{lat}}} + \lambda_{\text{lat}} \boldsymbol{I}_{J_{\text{time}}} \otimes \boldsymbol{I}_{J_{\text{lon}}} \otimes \boldsymbol{K}_{\text{lat}}) \boldsymbol{\gamma}$$

with penalty matrices $\boldsymbol{K}_k$ as introduced above. However, in spatio-temporal models the need of a separation into the interacting surface and the main effects increases, since we would like to be able to interpret the main time effect and the main spatial effect separately from their interaction. Therefore, some identification constraints have to be applied.

In basic one-dimensional P-splines identifiability is not fulfilled since each smooth function could be shifted on the y-axis and the shift would be absorbed by the intercept or another smooth function. Therefore, the identifiability constraint $\sum_i f_k(x_{ik}) = 0$ is included. One approach to achieve this constraint is to transform the basis functions and the penalty matrices with a QR-decomposition of the vector of column means of $\boldsymbol{B}_k$ (see Wood, 2017, Section 5.4.1 and Section 1.8.1 for details). In a higher dimensional setting this procedure could be adopted, such that first the multi-dimensional basis functions (based on the unconstrained marginal basis functions) and penalties are built and afterward the QR-decomposition is applied to guaranty identifiability. This would result in the standard multi-dimensional interaction surface without separate main effects.

To determine the separate main effects the order of the QR-decomposition and the Kronecker product is switched. Thus, the multi-dimensional basis function is built as Kronecker product of the one-dimensional basis functions with transformation based on the QR-decomposition. To get valid estimates the marginal basis functions must be added to the design matrix. Therefore, the design matrix of the spatio-temporal manifold with separate main effects has the following structure

$$\left[\mathbf{1} : \breve{\boldsymbol{B}}_{\text{time}} \square \breve{\boldsymbol{B}}_{\text{lon}} \square \breve{\boldsymbol{B}}_{\text{lat}} : \breve{\boldsymbol{B}}_{\text{time}} \square \breve{\boldsymbol{B}}_{\text{lon}} : \breve{\boldsymbol{B}}_{\text{time}} \square \breve{\boldsymbol{B}}_{\text{lat}} : \breve{\boldsymbol{B}}_{\text{lon}} \square \breve{\boldsymbol{B}}_{\text{lat}} : \breve{\boldsymbol{B}}_{\text{time}} : \breve{\boldsymbol{B}}_{\text{lon}} : \breve{\boldsymbol{B}}_{\text{lat}}\right] \quad (3)$$

where $\breve{\boldsymbol{B}}_k$ are the marginal basis functions including the transformation based on the QR-decomposition. Here $\square$ is the row-wise Kronecker product, for matrices as used in Lee and Durbán (2011) and defined in Eilers et al. (2006). Based on this definition $\breve{\boldsymbol{B}}_{\text{time}}\boldsymbol{\gamma}_{\text{time}}$ build the main effect for time and $\breve{\boldsymbol{B}}_{\text{lon}} \square \breve{\boldsymbol{B}}_{\text{lat}}\boldsymbol{\gamma}_{\text{lon}\times\text{lat}} + \breve{\boldsymbol{B}}_{\text{lon}}\boldsymbol{\gamma}_{\text{lon}} + \breve{\boldsymbol{B}}_{\text{lat}}\boldsymbol{\gamma}_{\text{lat}}$ the main effect for space with corresponding subvectors of $\boldsymbol{\gamma}$. Since the transformed basis functions $\breve{\boldsymbol{B}}_k$ have one column less than $\boldsymbol{B}_k$ the dimensions of Equation (2) and Equation (3) coincide. Similarly the penalty matrix is now build as a block diagonal matrix

$$\boldsymbol{K}_{\boldsymbol{\lambda}} = \begin{pmatrix} 0 & & & & & & & \\ & \boldsymbol{A}_1 & & & & & & \\ & & \boldsymbol{A}_2 & & & \mathbf{0} & & \\ & & & \boldsymbol{A}_3 & & & & \\ & & & & \boldsymbol{A}_4 & & & \\ & & \mathbf{0} & & & \boldsymbol{A}_5 & & \\ & & & & & & \boldsymbol{A}_6 & \\ & & & & & & & \boldsymbol{A}_7 \end{pmatrix}$$

with values

$$\begin{aligned}
\boldsymbol{A}_1 &= \lambda_1 \breve{\boldsymbol{K}}_{\text{time}} \otimes \breve{\boldsymbol{I}}_{\text{lon}} \otimes \breve{\boldsymbol{I}}_{\text{lat}} + \lambda_2 \breve{\boldsymbol{I}}_{\text{time}} \otimes \breve{\boldsymbol{K}}_{\text{lon}} \otimes \breve{\boldsymbol{I}}_{\text{lat}} + \lambda_3 \breve{\boldsymbol{I}}_{\text{time}} \otimes \breve{\boldsymbol{I}}_{\text{lon}} \otimes \breve{\boldsymbol{K}}_{\text{lat}} \\
\boldsymbol{A}_2 &= \lambda_4 \breve{\boldsymbol{K}}_{\text{time}} \otimes \breve{\boldsymbol{I}}_{\text{lon}} + \lambda_5 \breve{\boldsymbol{I}}_{\text{time}} \otimes \breve{\boldsymbol{K}}_{\text{lon}} \\
\boldsymbol{A}_3 &= \lambda_6 \breve{\boldsymbol{K}}_{\text{time}} \otimes \breve{\boldsymbol{I}}_{\text{lat}} + \lambda_7 \breve{\boldsymbol{I}}_{\text{time}} \otimes \breve{\boldsymbol{K}}_{\text{lat}} \\
\boldsymbol{A}_4 &= \lambda_8 \breve{\boldsymbol{K}}_{\text{lon}} \otimes \breve{\boldsymbol{I}}_{\text{lat}} + \lambda_9 \breve{\boldsymbol{I}}_{\text{lon}} \otimes \breve{\boldsymbol{K}}_{\text{lat}} \\
\boldsymbol{A}_5 &= \lambda_{10} \breve{\boldsymbol{K}}_{\text{time}} \\
\boldsymbol{A}_6 &= \lambda_{11} \breve{\boldsymbol{K}}_{\text{lon}} \\
\boldsymbol{A}_7 &= \lambda_{12} \breve{\boldsymbol{K}}_{\text{lat}}
\end{aligned}$$

where $\breve{\boldsymbol{K}}_k$ are the penalty matrices of the marginal basis functions including the transformation based on the QR-decomposition. Therefore, $\breve{\boldsymbol{I}}_k$ are of one-dimension less than $\boldsymbol{I}_k$. Based on this decomposition we can determine the main effects separately from their interactions (Wood, 2017, p. 232). A similar idea has also been introduced by Lee and Durbán (2011) and Wood et al. (2013), but there not the QR-decomposition, but the mixed model representation of the marginal splines is used.

However, by applying the main effects decomposition we have to optimize 12 instead of 3 smoothing parameters $\boldsymbol{\lambda}$. An alternative would be to assume isotropy in the spatial effects. Then we would assume (i) that the manifold is defined as

$$\left[\mathbf{1} : \breve{\boldsymbol{B}}_{\text{time}} \square \breve{\boldsymbol{B}}_{\text{spat}} : \breve{\boldsymbol{B}}_{\text{spat}} : \breve{\boldsymbol{B}}_{\text{time}}\right] \tag{4}$$

where $\breve{\boldsymbol{B}}_{\text{spat}}$ is based on a two-dimensional isotropic basis function; (ii) that the penalty changes accordingly and is now dependent only on 4 smoothing parameters. In the isotropic setting different scaling of the wiggliness in longitudinal and latitudinal directions are not possible anymore. Therefore, isotropy might not be correct in meteorological data. Furthermore, we will discuss expectile regression later as we would like to reduce any assumption on the distribution. Thus, we retain the formerly discussed anisotropic model.

So far we estimated the spatial trend based on the longitudinal and latitudinal coordinates of the location. In many data sets, however, the location is only measured in regional grids like ZIP-code areas or states. Then the spatial surfaces based on the categorical covariates can be estimated as smooth effects, if neighboring regions have similar coefficients. This is achieved by some penalized regression, where the penalty is defined by joint borders of the regions. Generally this approach is motivated by Gaussian Markov random fields (GMRF) (Rue and Held, 2005). There the variation over time is interpreted as random walk, which is also described in terms of neighborhood structures. So the spatio-temporal model is built similarly as in Equation (4), where the interaction of time and space bases on the Kronecker products of the main effects and the penalties. Alternatively, the centroids of the regions can be used as standardized location of the observations, thus the three-dimensional P-splines can be applied (see for example Ugarte et al., 2010). Based on similarities of GMRF and P-splines as penalized models the spatio-temporal model could also be defined as an interaction of a P-spline and a GMRF.

3.4 Cyclic Splines

Additionally to the separation into the main effects, another condition has to be discussed. In our definition of spatio-temporal models we use multiple years to estimate the seasonal effect. Thus, the seasonal effect should be identical in multiple years. It is defined as a function from January to December. A general variation between the years, for example due to climate change, can be modeled as additional term. Alternatively, the seasonal effect could be varying for each year, such that we would get a temporal effect for each day in the observation period. However, predictions and interpretations of the spatio-temporal effect are more complex with the latter definition and each effect relies on less data. Thus, we propose to use the first idea, i.e. a single seasonal effect for all years. In order to get a valid curve for the seasonal effect we have to ensure that there is a smooth transition between December and January. Therefore, we implement conditions for the curve to be equal on the left and the right end of the parameter space up to the second derivative. For B-splines this could be done quite easily. In standard B-splines of degree 3 the first three and the last three basis functions usually are truncated at the edge of the parameter space. For cyclic B-splines those truncated basis function are now

defined to coincide appropriately. Moreover the neighborhood structure which is used for the penalization is now changed such that the originally truncated basis functions have neighbors on both sides (de Boor, 1978; Wood, 2017).

3.5 Semiparametric Expectile Regression

In Section 2 we defined expectile regression for linear predictors. In order to gain more flexibility semiparametric predictors are useful. Therefore, Schnabel and Eilers (2009) and Sobotka and Kneib (2012) introduced semiparametric expectile regression with the penalized least asymmetric weighted squares criterion

$$\sum_{i=1}^{n} w_\tau(y_i) \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\gamma}_{\tau,\boldsymbol{\lambda}}\right)^2 + \boldsymbol{\gamma}_{\tau,\boldsymbol{\lambda}}^\top \boldsymbol{K}_{\boldsymbol{\lambda}}\, \boldsymbol{\gamma}_{\tau,\boldsymbol{\lambda}}. \tag{5}$$

Here $\boldsymbol{x}_i^\top \boldsymbol{\gamma}_{\tau,\boldsymbol{\lambda}}$ is the semiparametric predictor of linear effects and smooth functions dependent on the smoothing parameters and the asymmetry τ. $\boldsymbol{K}_{\boldsymbol{\lambda}}$ is the penalty matrix including the smoothing parameters, which are dependent on the asymmetry. Due to the technical equivalence between weighted linear regression and expectile regression this is straightforward. So spatio-temporal models can also be applied in expectile regression. In the estimation the critical point is the optimization of the smoothing parameters $\boldsymbol{\lambda}$. It has to be done from outside the LAWS algorithm otherwise the iteration does not always converge.

In the application we use the power (memory and speed) of the `mgcv` package of Wood (2017) in the statistical programming language `R` (R Core Team, 2017) to estimate the spatio-temporal model as weighted least squares model. The `bam` function of the `mgcv` package was optimized to reduce the memory demand, by avoiding to calculate the design matrix and applying other smart tricks (see Wood et al., 2015, 2017, for details). In our example this function reduced the memory demand from more than 40GB to 5GB. The exact code for estimating expectile regression with spatio-temporal effects is attached in the supplementary material. Basically in the inner loop we fix the smoothing parameters and estimate the LAWS algorithm, with help of the function `bam`. Then we apply standard numerical optimization routines to optimize the smoothing parameters from outside. As criterion for the optimization the asymmetric generalized cross-validation criterion (GCV)

$$\frac{n \sum_{i=1}^{n} w_\tau(y_i) \left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}}_{\tau,\boldsymbol{\lambda}}\right)^2}{\left(\operatorname{trace}(\boldsymbol{I} - \boldsymbol{H})\right)^2}$$

proved good properties in Schnabel and Eilers (2009). Moreover, it is straightforward to perform the classical optimization of smoothing parameters in the linear model (see Wood, 2017, for example). In the above formula $\boldsymbol{I}$ is a unit matrix of dimension $n \times n$ and $\boldsymbol{H} = \boldsymbol{W}_\tau^{1/2} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{W}_\tau \boldsymbol{X} + \boldsymbol{K}_{\boldsymbol{\lambda}})^{-1} \boldsymbol{X}^\top \boldsymbol{W}_\tau^{1/2}$ is the hat matrix. More details on the estimation of semiparametric expectile regression in general are presented in Sobotka and Kneib (2012).

Due to the possibility to write P-spline as mixed models the Schall algorithm (Schall, 1991) for selecting smoothing parameters was introduced to semiparametric expectile regression by Schnabel and Eilers (2009). However, the Schall algorithm only allows for one smoothing parameter per smooth term. Thus, the Schall algorithm is not applicable in spatio-temporal models. Alternatively, the generalized Fellner-Schall algorithm of Wood and Fasiolo (2017) can be adopted to semiparametric expectile regression, by interpreting expectile regression as weighted linear regression. Moreover, the generalized Fellner-Schall algorithm allows for the estimation of smoothing parameters in interaction settings.

In the generalized Fellner-Schall algorithm the smoothing parameters are optimized iteratively with the model fit. So the model is estimated given some smoothing parameters $\boldsymbol{\lambda}$. Thus, $\boldsymbol{W}_\tau$ and $\hat{\boldsymbol{\gamma}}_{\tau,\boldsymbol{\lambda}}$ are the respective weights and the estimated coefficients of the expectile regression given the current smoothing parameter $\boldsymbol{\lambda}$. Afterwards new smoothing parameters are fitted as

$$\lambda_k^* = \phi \frac{\text{tr}(\boldsymbol{K}_{\boldsymbol{\lambda}}^- \boldsymbol{S}_k) - \text{tr}((\boldsymbol{X}^\top \boldsymbol{W}_\tau \boldsymbol{X} + \boldsymbol{K}_{\boldsymbol{\lambda}})^{-1} \boldsymbol{S}_k)}{\hat{\boldsymbol{\gamma}}_{\tau,\boldsymbol{\lambda}}^\top \boldsymbol{S}_k \hat{\boldsymbol{\gamma}}_{\tau,\boldsymbol{\lambda}}} \lambda_k$$

where ϕ is a scaling parameter based on the variance and $\boldsymbol{K}_{\boldsymbol{\lambda}}^-$ is the Moore-Penrose pseudoinverse of the full penalty matrix $\boldsymbol{K}_{\boldsymbol{\lambda}}$ including the current smoothing parameters $\boldsymbol{\lambda}$. Furthermore, $\boldsymbol{S}_k$ is similarly to $\boldsymbol{K}_k$, respectively $\breve{\boldsymbol{K}}_k$ the penalty matrix corresponding to one smoothing parameter λ_k. However, $\boldsymbol{S}_k$ is filled with zeros to have the same dimensions as $\boldsymbol{K}_{\boldsymbol{\lambda}}$. Nevertheless $\boldsymbol{S}_k$ does not include λ_k. In the estimation of new smoothing parameters some kind of step halving is established to ensure a better model fit in each iteration. Via step halving the new smoothing parameters are calculated as

$$\boldsymbol{\lambda}^{\text{new}} = \frac{\boldsymbol{\lambda}^* - \boldsymbol{\lambda}}{2^p} + \boldsymbol{\lambda}$$

where $p = 0, 1, 2, \ldots$ is the minimal integer such that the goodness-of-fit decreases. Therefore, either the penalized LAWS criterion as defined in Equation (5), or the GCV can be applied. The generalized Fellner-Schall algorithm has some numerical drawbacks. First, estimating $\boldsymbol{K}_{\boldsymbol{\lambda}}^-$ is only possible, if the smoothing parameters are in an appropriate range, otherwise the pseudoinverse will vanish. Therefore, we have to restrict the possible smoothing parameters and fix them, if they reach the limit. Second, calculating $(\boldsymbol{X}^\top \boldsymbol{W}_\tau \boldsymbol{X} + \boldsymbol{K}_{\boldsymbol{\lambda}})^{-1}$ is computationally burdensome, but it is estimated anyway for the standard confidence intervals (see for example Marra and Wood, 2012). Thus, these estimates can be reused.

Alternatively the SAP-algorithm of Rodríguez-Álvarez et al. (2015) could be extended to expectile regression. Though, the SAP-algorithm uses the mixed models representation of the splines, which we would like to avoid here. Moreover, it includes calculations based on the full design matrix, which is computationally demanding in our example. So we choose the generalized Fellner-Schall algorithm and the GCV optimization due to their compatibility with the output of the `mgcv` package.

4 Simulation Study

Applying the GCV to select smoothing parameters in semiparametric expectile regression with interactions is straightforward. Nevertheless the generalized Fellner-Schall algorithm has, to our knowledge, never been used in expectile regression before. Therefore, we provide in this section a small simulation study to compare the goodness-of-fit of both approaches. Here the goodness-of-fit is calculated as (predicted) mean weighted squared error (P)MWSE

$$\text{(P)MWSE} = \sum_{i=1}^{n} w_\tau(y_i) \left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\gamma}}_\tau\right)^2.$$

The covariates x_1, x_2 are simulated based on the standard uniform distribution ($x_1, x_2 \sim U(0,1)$). As distribution of the error terms a Gaussian distribution is assumed ($\varepsilon_i \sim N(0, \sigma_i^2)$). However, for the variance either homoscedasticity ($\sigma_i = 2 \ \forall i$) or heteroscedasticity ($\sigma_i = \frac{x_{i1}+1}{1.5} + \frac{x_{i2}+1}{1.5}$) is applied. As covariate effects we use two different

functions, similarly as in Wood (2006):

$$f_1(x_1, x_2) = 1.2\pi \left(1.2e^{-(x_1-0.2)^2/0.3^2-(x_2-0.3)^2/0.4^2} + 0.8e^{-(x_1-0.7)^2/0.3^2-(x_2-0.8)^2/0.4^2}\right)$$
$$f_2(x_1, x_2) = 2\sin(\pi\ x_1) + \exp(2x_2).$$

Finally the response is defined as $y_i = f_j(x_{i1}, x_{i2}) + \varepsilon_i$. For each replication a data set with 5000 observations is simulated to fit the model and for all replications a data set with 10000 observations was used to estimate the predictive goodness-of-fit. Overall 100 independent replications of expectile models with $\tau \in (0.1, 0.5, 0.9)$ are applied. For the estimation the model was specified with separation of main effects as

$$y \sim f(x_{i1}) + f(x_{i2}) + f(x_{i1}, x_{i2})$$

with 15 B-spline basis functions in each direction. Since the smoothing parameters in the Fellner-Schall algorithm have to be sized moderately, we restrict them to be larger than 1e-5 and smaller than 1e5. For improved comparability this restriction is also applied in the GCV approach. The resulting predictive mean weighted squared errors are displayed in Figure 1. The mean weighted squared error is slightly better for the optimization via GCV, while the predictive mean weighted squared error is better for the optimization via the generalized Fellner-Schall algorithm. Though the results are quite similar. Further analysis shows that the generalized Fellner-Schall algorithm is more dependent on the starting values.

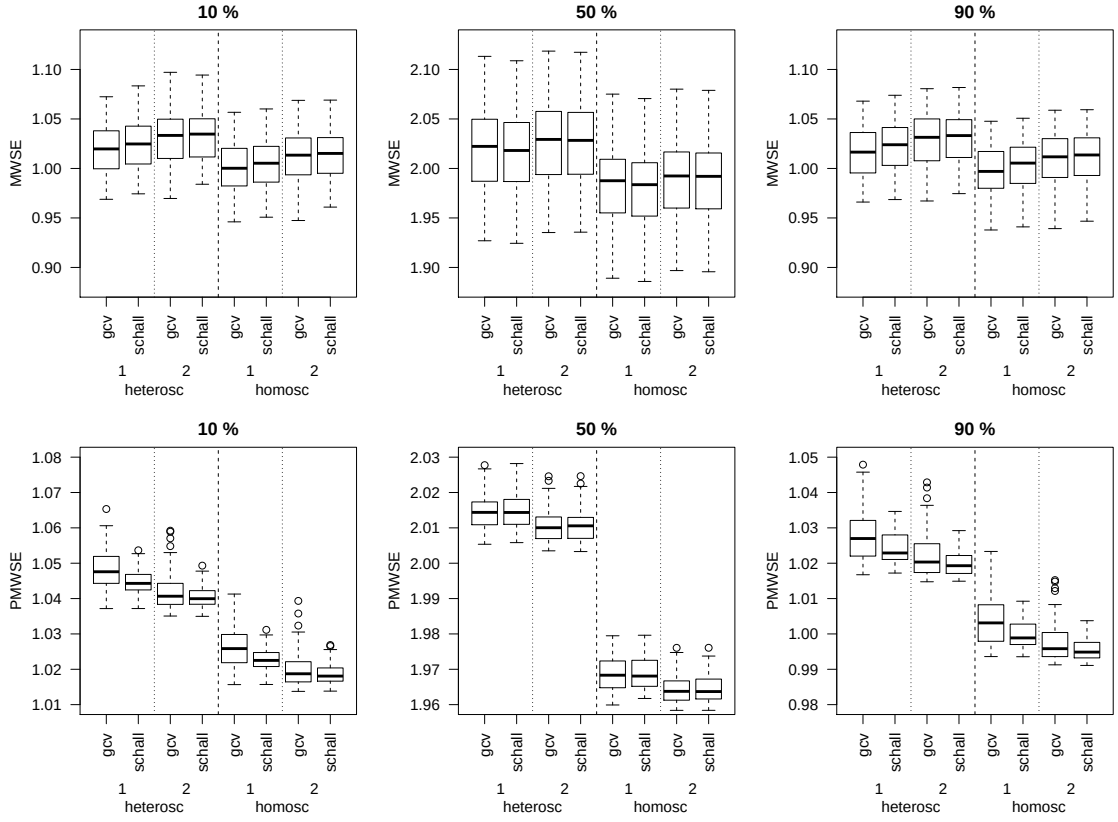

Figure 1: MWSE and PMWSE of the simulation study. On the left side of each plot the models with heteroscedastic errors are displayed and homoscedastic errors are on the right. For each data setting (1 or 2) the smoothing parameter selections are placed next to each other.

5 Spatio-Temporal Analysis of Temperatures in Germany

The distribution of the temperatures in Germany motivated us for a spatio-temporal estimation beyond the mean. Therefore, we use data from the German weather service DWD (2017). The response variable is the daily mean temperature, since we would like to analyze the variation within years, the aggregation on the average per day does not impair the expectile regression. In this example we use all stations with at least 24 years of observations in the period 1980 to 2014. Furthermore, stations above 900m are included if they have at least 3650 values, but the station on top of the "Zugspitze" is excluded, due to fitting problems based on the large gap in the elevation scale. So finally we use data of 374 stations, whose locations are visualized in Figure 2. On the right side of this figure the marginal frequencies of the daily mean temperature from 1980 to 2014 are plotted jointly with a Gaussian density of appropriate mean and standard deviation. Even if the marginal density fits well with the Gaussian density we model the spatio-temporal distribution of the daily temperatures with spatio-temporal expectile regression in the following. Doing so we get not only information on the general temperature pattern in Germany in different seasons of the year, but also information on areas, where at distinct time points the spectrum of temperatures is wider or smaller.

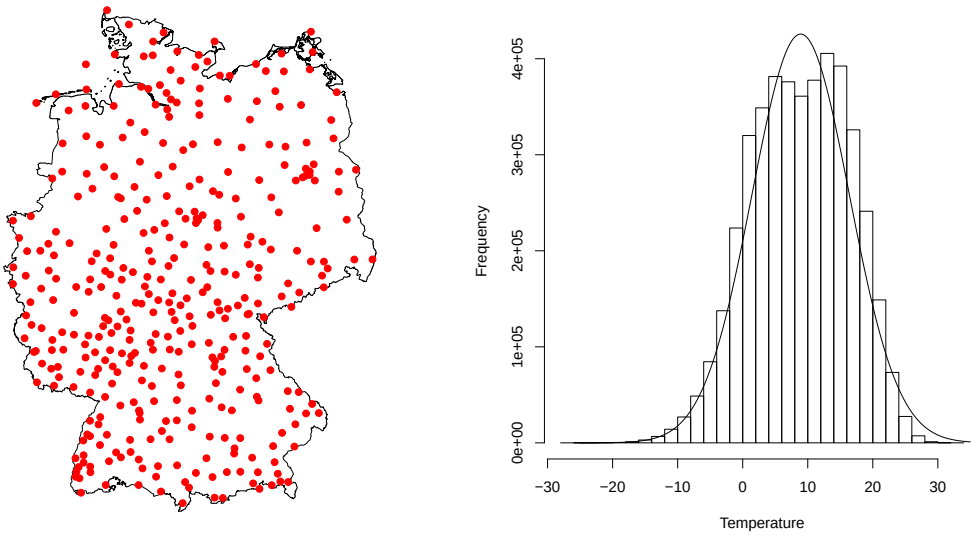

Figure 2: Location of the observation stations and marginal density of daily mean temperature.

To model the spatial and temporal variation of the temperatures we apply the following model

$$\begin{aligned}\text{temperature} = \beta_{0,\tau} &+ f_\tau(\text{elevation}) + f_\tau(\text{year}) + f_\tau(\text{day}) + f_\tau(\text{lon}) + f_\tau(\text{lat}) \\ &+ f_\tau(\text{day},\text{lon}) + f_\tau(\text{day},\text{lat}) + f_\tau(\text{lon},\text{lat}) + f_\tau(\text{day},\text{lon},\text{lat})\end{aligned}$$

for each asymmetry parameter $\tau \in (10\%, 50\%, 90\%)$ separately. Thereby *elevation* is the altitude above sea level of the observation station, while *lon*gitude and *lat*itude specify the location. With *day*, the day of the year is meant. The main effects $f_\tau(\text{lon})$, $f_\tau(\text{lat})$, $f_\tau(\text{day},\text{lon})$, $f_\tau(\text{day},\text{lat})$ are just included to get a valid design matrix and are only interpreted jointly with $f_\tau(\text{lon},\text{lat})$ and $f_\tau(\text{day},\text{lon},\text{lat})$, respectively.

For the estimation we apply penalized B-splines of degree 3 and 15 basis functions for the spline of the year. Moreover, the spatio-temporal effect has 15 basis functions for the daily effect and 6 respectively 9 basis functions for the one-dimensional spatial marginals. The spatial effect has only few basis functions in each direction to obtain reasonable

computational times. For the estimation of the elevation effect only 7 basis functions are applied to avoid arbitrary results due to gaps in the parameter space. The smoothing parameters are optimized via GCV. The results for optimizing the smoothing parameters via the generalized Fellner-Schall algorithm are similar.

In Figure 3 the estimated main effects for elevation and day are displayed. There we see some variation between the mean effects and the effects at the outer parts of the distribution. While for the low areas the curves are parallel, which induces homoscedasticity, the 10% and 90% expectile curves diverge from the mean for higher altitudes. The difference looks rather small, but we are talking about $2°C$ difference between the 10% and the 90% expectile. Thus, heteroscedasticity occurs in these areas.

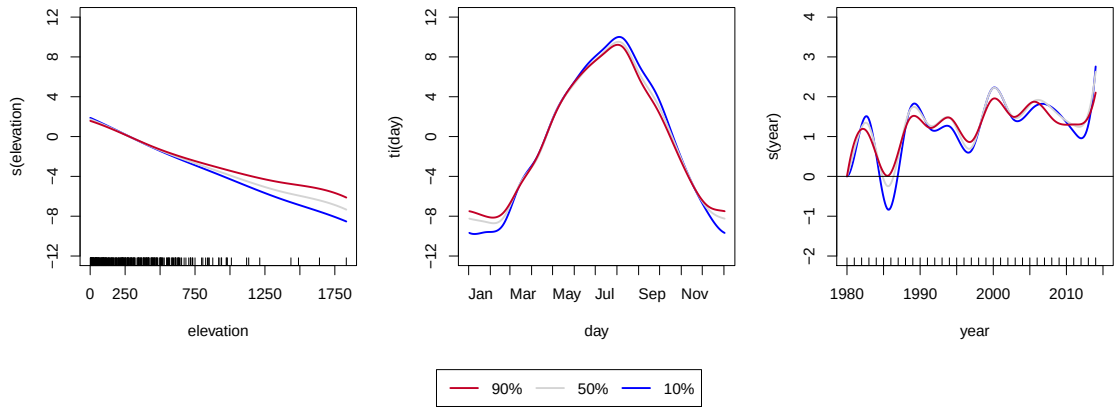

Figure 3: Main effects for day, elevation and year (without intercept).

For the main effect of the day in the year we can detect, that the lower expectile has a greater range $[-9.8°C; 10.0°C]$ than the upper expectile $[-8.1°C; 9.2°C]$ (both excluding the intercept of $4.8°C$ respectively $11.1°C$). This means in winter the lower temperatures are often lower than expected and in summer the low temperatures are higher than expected. Thus, the variance in winter is higher than in summer. Furthermore, the high temperatures in summer are not as high as they should be, if the underlying process is a homoscedastic Gaussian distribution. Overall the visualized expectiles look like there might be crossing expectiles, which should not be the case by theory, but we plotted the curves without intercept, as all other plots, to get a better view on the differences in the shape of the curves.

By including the parameter *year* in the model we control for varying effects in specific years. Additionally we can check if we find some impact of the climate change in this rather small example. The estimated curve for the trend per year is also plotted in Figure 3 on the right. There we detect some small general increase in the temperatures, beyond the natural fluctuation.

The main spatial effect for the whole year, as displayed in Figure 4, shows that in the northeast the higher temperatures (90% expectile) are not as high as expected by the mean regression. In general some twist in the spatial effect between the different parts of the distribution is visible. While the temperature decrease from southwest to northeast for the bottom part of the distribution it decreases more from south to north for the upper part.

To get a better impression on how the spatial effect varies with time we plotted in Figure 5 the temporal effect of four German cities. Their locations are indicated in the spatial effect maps. Out of this figure we conclude that in Cologne the winters are a lot

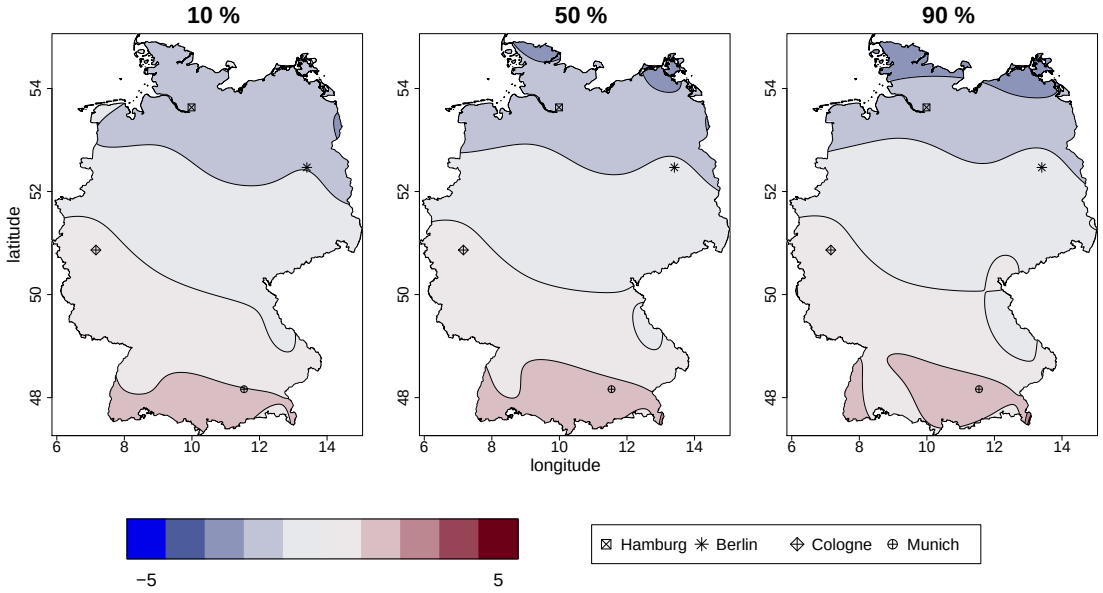

Figure 4: Main spatial effect without elevation and intercept.

warmer than in Munich or Berlin, while the summers are colder in Hamburg and warmer in Berlin. This differentiation is valid for all parts of the distribution, but the amplitude of the temperatures is smaller for the upper part of the distribution than for the lower part.

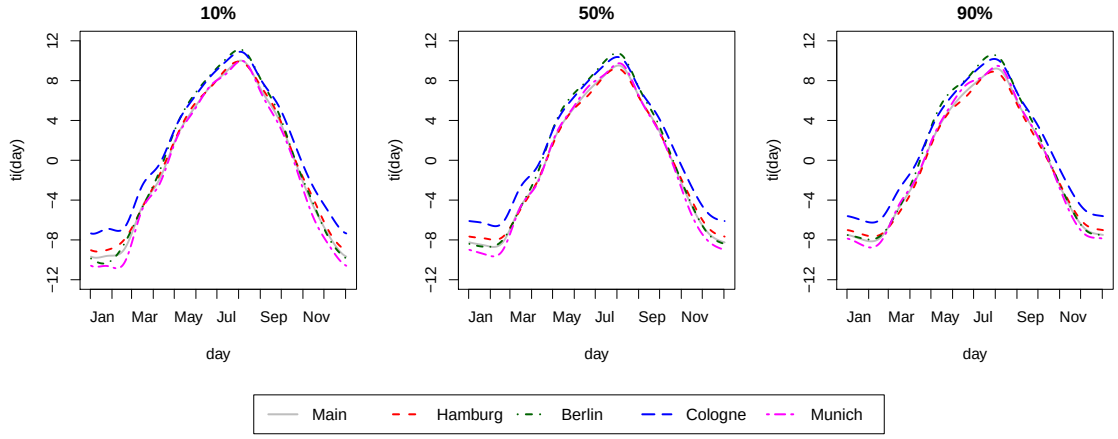

Figure 5: Effects of day in the year for four German cities, including elevation, but without intercept.

We see similar results from the spatial effect of January 31st and July 31st as displayed in Figure 6. The colors of both rows follow the same legend, while they are different from Figure 4. However, in Figures 4 and 6 the steps between the colors are $1°C$. From Figure 6 we can conclude two things. The most obvious one is that the spatial effect in January is different from July, since in January it gets colder from west to east, while in summer the north is colder than the south. On the other hand the variation from the mean is visible. So is the coast of the Baltic Sea a lot warmer in cold winters than expected for this location. Furthermore, the coldest winters are detected east of Berlin, while there is a rather constant effect for this area at the 90% expectile for January 31st. In general

the variation of temperatures on January 31st is larger for the lower expectile than for the upper expectile. Moreover, in northern Germany the variation for cold summers (10% expectile) is rather low, while there is a clear cooling towards the sea for warm summers. Similar patterns can be found for the effect with elevation as displayed in Figure 7 in the Appendix.

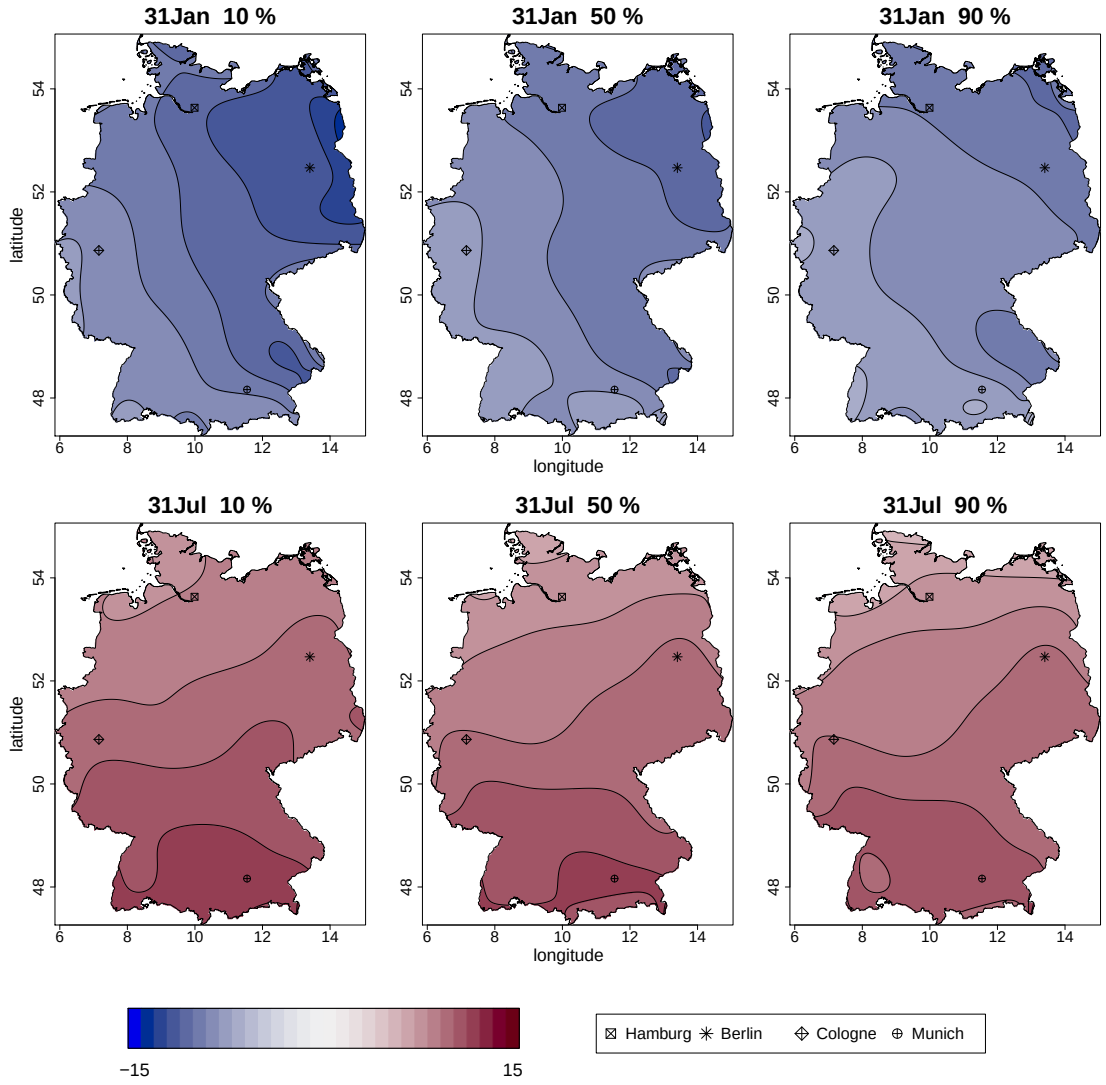

Figure 6: Forecast of the spatial effect for January 31 and July 31, excluding elevation and intercept.

6 Conclusion

In this paper we present spatio-temporal effects for expectile regression. Spatio-temporal modeling with interaction terms of P-splines is an established method. However, usually the data are analyzed with some pre-specified parametric distribution and the errors are assumed to be homoscedastic. Contrarily, in expectile regression no assumption on the distribution of the data is applied. Moreover, expectile regression is able to take heteroscedasticity of the data into account. Based on the idea of weighted least squares spatio-temporal expectile regression can be applied with help of the tools of standard least squares regression. So it is a natural extension of the standard approaches. Furthermore,

expectile regression can be used to check whether the homoscedasticity assumption is valid. Therefore, we check if all effects are equal for a grid of asymmetry parameters. This is similar to the test of Newey and Powell (1987). Our analysis showed that the assumption of homoscedasticity is not necessarily fulfilled for the temperature and the application of expectile regression is necessary. Thus, the effect of elevation varies for the different parts of the distribution and there are some regions where the spatial effect of the 10% and the 90% expectile varies from the mean effect.

Alternatively to the analysis of temperatures in Germany the amount of rain may also have effects beyond the mean, as Umlauf et al. (2016) analyzed for Austria. However, there we have to take care of the large number of days without any rain. Umlauf et al. (2016) do this by using a censored normal distribution. Another possibility would be to apply a hurdle model (Mullahy, 1986). Moreover, the hurdle model could then be generalized to effects beyond the mean with help of expectile regression. Nevertheless, then all expectiles than must have positive values. This could be achieved, for example, by including a link function around the classical expectile model. However, a fixed link function would impose a distributional assumption which is undesired in expectile regression. Thus, the link function should be estimated jointly with the covariate effects, but this is beyond the scope of this paper and left for further research. Modeling binary data with flexible response function was introduced in Spiegel et al. (2017), were we modified the approach of Muggeo and Ferrara (2008) to also include smooth effects in the predictor.

Another example where the temporal variation of spatial effects beyond the mean would be interesting is the analysis of the development of undernutrition in developing countries. Here *USAID* presents health data of children in many developing countries on a yearly basis. However, the location of the children is only measured on district levels, such that the models based on three-dimensional P-splines are not directly applicable. Though the interaction of the spatial effect, estimated with a GMRF, and the time could be applied. Then changes in the spatial distribution of undernutrition depending on the year could be estimated. This is a natural case for expectile regression, since it would be interesting to check if the effects for the undernourished children vary from the mean effect of the healthy children.

Acknowledgment

I acknowledge financial support by the German Research Foundation (DFG), grant KN 922/4-2.

References

Cressie, N. and C. K. Wikle (2011). *Statistics for spatio-temporal data*. Hoboken: John Wiley & Sons.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer Verlag.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*, 85–100.

DWD (2017). Climate Data Center. *Deutscher Wetterdienst*. `ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/daily/kl/historical/`.

Eilers, P. H. C., I. D. Currie and M. Durbán (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis 50*(1), 61–76.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Eilers, P. H. C. and B. D. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems 66*(2), 159–174.

Fahrmeir, L., T. Kneib and S. Lang (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14*, 715–745.

Fahrmeir, L., T. Kneib, S. Lang and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Berlin: Springer.

Hastie, T. and R. Tibshirani (1986). Generalized Additive Models. *Statistical Science 1*(3), 297–310.

Klein, N., T. Kneib and S. Lang (2015). Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association 110*(509), 405–419.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society 46*(1), 33–50.

Lee, D.-J. and M. Durbán (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics & Data Analysis 53*(8), 2968–2979.

Lee, D.-J. and M. Durbán (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling 11*(1), 49–69.

Marra, G. and S. N. Wood (2012). Coverage Properties of Confidence Intervals for Generalized Additive Model Components. *Scandinavian Journal of Statistics 39*(1), 53–74.

Muggeo, V. M. and G. Ferrara (2008). Fitting generalized linear models with unspecified link function: A P-spline approach. *Computational Statistics & Data Analysis 52*(5), 2529–2537.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*(3), 341–365.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society 55*(4), 819–847.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `http://www.R-project.org`.

Rigby, R., D. Stasinopoulos and V. Voudouris (2013). Discussion: A comparison of GAMLSS with quantile regression. *Statistical Modelling 13*(4), 335–348.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Rodríguez-Álvarez, M. X., D.-J. Lee, T. Kneib, M. Durbán and P. H. C. Eilers (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing 25*(5), 941–957.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: Theory and Applications*. CRC Press.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Schnabel, S. K. and P. H. C. Eilers (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis 53*(12), 4168–4177.

Sobotka, F., G. Kauermann, L. Schulze-Waltrup and T. Kneib (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing 23*(2), 135–148.

Sobotka, F. and T. Kneib (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis 56*(4), 755–767.

Spiegel, E., T. Kneib and F. Sobotka (2017). Generalized Additive Models with Flexible Response Functions. *Manuscript Submitted for Publication*.

Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris and F. De Bastiani (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Ugarte, M., T. Goicoa and A. Militino (2010). Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics 21*(3-4), 270–289.

Umlauf, N., N. Klein, A. Zeileis and M. Koehler (2016). bamlss: Bayesian additive models for location scale and shape (and beyond). *Unpublished manuscript*. `http://EconPapers.repec.org/RePEc:inn:wpaper:2017-05`.

Wood, S. N. (2006). Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics 62*(4), 1025–1036.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). CRC Press.

Wood, S. N. and M. Fasiolo (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*.

Wood, S. N., Y. Goude and S. Shaw (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 64*(1), 139–155.

Wood, S. N., Z. Li, G. Shaddick and N. H. Augustin (2017). Generalized Additive Models for Gigadata: Modeling the UK Black Smoke Network Daily Data. *Journal of the American Statistical Association*, 1–12.

Wood, S. N., F. Scheipl and J. J. Faraway (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing 23*(3), 341–360.

World Meteorological Organization (2017). WMO Statement on the State of the Global Climate in 2016. *WMO 1189*, 1–24.

Yao, Q. and H. Tong (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics 6*(2-3), 273–292.

Appendix

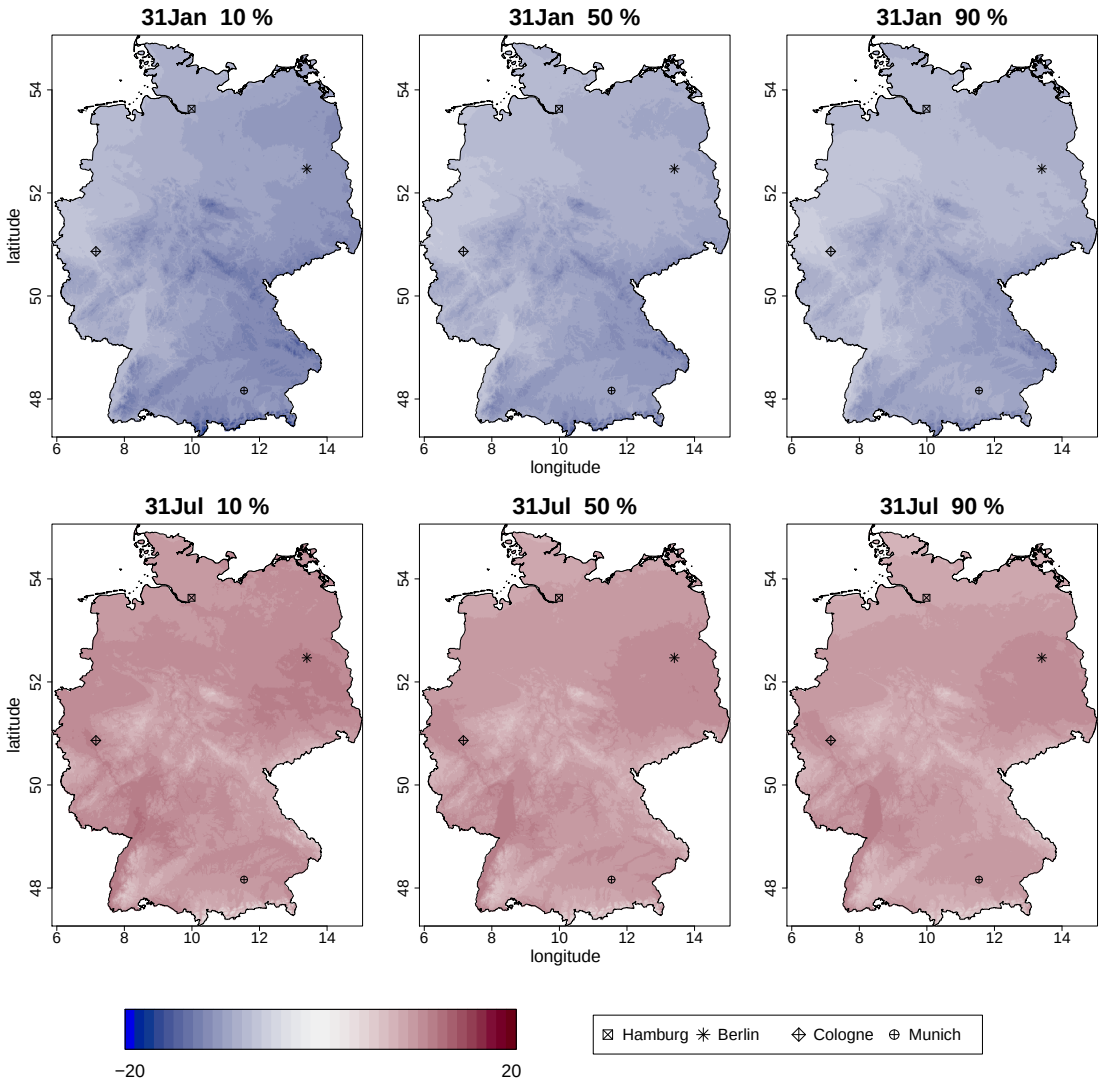

Figure 7: Forecast of the spatial effect for January 31 and July 31, including elevation, but without intercept.