Noname manuscript No.
(will be inserted by the editor)

Heavy tailed spatial autocorrelation models

A. Kreuzer · T. Erhardt · T. Nagler · C. Czado

Received: date / Accepted: date

Abstract Appropriate models for spatially autocorrelated data account for the fact that observations are not independent. A popular model in this context is the simultaneous autoregressive (SAR) model that allows to model the spatial dependency structure of a response variable and the influence of covariates on this variable. This spatial regression model assumes that the error follows a normal distribution. Since this assumption cannot always be met, it is necessary to extend this model to other error distributions. We propose the extension to the t-distribution, the tSAR model, which can be used if we observe heavy tails in the fitted residuals of the SAR model. In addition, we provide a variance estimate that considers the spatial structure of a variable which helps us to specify inputs for our models. An extended simulation study shows that the proposed estimators of the tSAR model are performing well and in an application to fire danger we see that the tSAR model is a notable improvement compared to the SAR model.

Keywords Simultaneous autoregressive model · Spatial dependence · Fire danger · Heavy tails

1 Introduction

“Coincidence of value similarity with locational similarity" is how Anselin and Bera (1998) loosely describe spatial autocorrelation. For illustration we show the Burning Index, a measure for fire danger, for different locations in the US (Figure 1). We observe

A. Kreuzer
Department of Mathematics, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany
Tel.: +49-89-289-17425
E-mail: a.kreuzer@tum.de

T. Erhardt, T. Nagler, C. Czado
Department of Mathematics, Technische Universität München, Boltzmanstraße 3, 85748 Garching, Germany

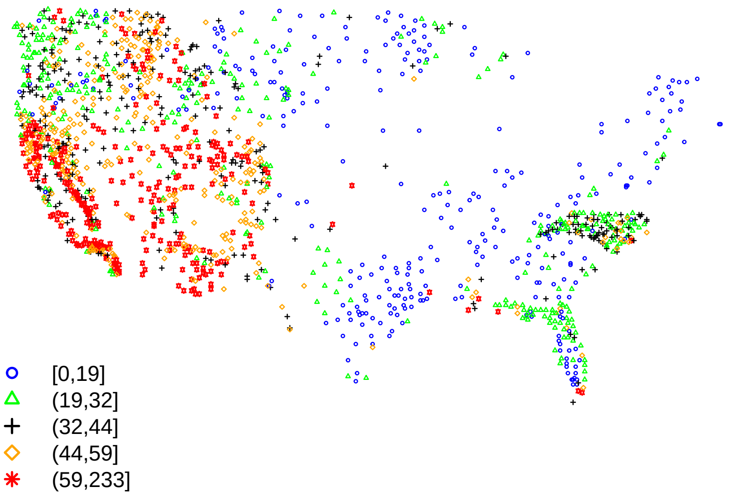

Fig. 1 Spatial distribution of the Burning Index visualized on the map. We have one observation for every location. The cutpoints of the symbol key are the 20%, 40%, 60% and 80% quantile of the variable.

that similar values cluster together, indicating (positive) spatial autocorrelation. Spatial autocorrelation occurs in many different types of data, for example in climate (fire danger, droughts) or economics (unemployment) data. This is why statistical methods that can deal with spatial autocorrelation are of high interest. A first contribution to this field was made by Whittle (1954) who provided a framework for stochastic processes on the plane. Whittle introduced autoregressive models in two dimensions. Following this idea, Ord (1975) proposed the simultaneous autoregressive (SAR) model. This model not only allows us to capture the spatial dependency structure of a response variable but also the influence of covariates on this variable. This property of the SAR model makes it very attractive and led to extensions. Pace and Barry (1997) studied how sparse spatial weight matrices can speed up the estimation procedure and De Oliveira and Song (2008) provide a Bayesian framework for the SAR model.

This work was motivated by an attempt to investigate the influence of weather conditions on fire danger in the continental US while accounting for spatial dependency. Data are obtained from the Wildland Fire Assessment System (WFAS). WFAS generates maps for observed and forecasted weather, fuel moisture and fire danger in the US. The SAR model is based on the assumption that the error follows a normal distribution, an assumption that cannot always be met. In our fitted model we observed residuals having heavier tails than the normal distribution. This is why we propose an extension of the SAR model to allow for a t-distributed error. We call this the tSAR model (Section 3.1). We show how parameters of the tSAR model can be estimated and how the fitted model can be used for prediction (Sections 3.2 and 3.3). Furthermore, we provide a spatially varying variance estimate which serves as input to our models (Section 3.4). In a simulation study (Section 4), we show that our proposed estimators for the tSAR model are reasonable and the application (Section 5) shows that the model fit can improve on the standard SAR model.

2 The SAR model

We recall some basic concepts related to the SAR model. First, we need to be able to determine how certain locations are related to each other, i.e., if there is a link between them and, if so, how strong the connection is. This is usually encoded in a proximity matrix (cf.,Waller and Gotway (2004) p.224 ff.). For n spatial locations $l_1, ..., l_n$, the *proximity matrix* is a $n \times n$ matrix where entry (i, j) indicates if and how strong location l_i is connected to location l_j. A value of zero means that there is no connection from l_i to l_j. The diagonal of the proximity matrix is set to zero such that a location is not connected to itself. Since this matrix does not need to be symmetric, we need to distinguish between a connection from l_i to l_j and a connection from l_j to l_i.

For a given proximity matrix $\boldsymbol{W}$ with entries w_{ij}, we can introduce the neighbors of location l_i which are all locations l_j such that $w_{ij} \neq 0$. We denote the *set of neighbors of location i* by N_i, i.e.,

$$N_i := \{j \in \{1, \dots n\} | w_{ij} \neq 0\}.$$

We now provide two possible choices of proximity matrices. In both cases we measure the strength of a connection by the inverse distance between the two corresponding locations. We will use the great circle distance (cf.,Banerjee (2005)) since our locations are specified as longitude/latitude pairs. For the first example we consider $N_i(k)$ the set of the k nearest neighbors of l_i, i.e., the k locations (excluding l_i) which have the smallest distance to l_i. Let d_{ij} denote the distance between l_i and l_j. For given k, entry (i, j) of the *non-standardized nearest neighbors based proximity matrix* $\tilde{\boldsymbol{W}}$ is then given by

$$\tilde{w}_{ij} := \begin{cases} \frac{1}{d_{ij}}, & \text{if } l_j \in N_i(k) \\ 0, & \text{else} \end{cases},$$

and entry (i, j) of the *row-standardized nearest neighbors based proximity matrix* $\boldsymbol{W}$ is defined by

$$w_{ij} := \frac{\tilde{w}_{ij}}{\tilde{w}_{i.}},$$

where $\tilde{w}_{i.} = \sum_{j=1}^{n} \tilde{w}_{ij}$ is the sum of the i-th row of the non-standardized nearest neighbors based proximity matrix $\tilde{\boldsymbol{W}}$. By defining the proximity matrix in this way, we ensure that each location has the same number of neighbors. This is no longer the case if we use a radius to determine the set of neighbors.

For a given radius r, entry (i, j) of the *non-standardized radius based proximity matrix* $\tilde{\boldsymbol{W}}$ is given by

$$\tilde{w}_{ij} := \begin{cases} \frac{1}{d_{ij}}, & \text{if } i \neq j \text{ and } d_{ij} \leq r \\ 0, & \text{else} \end{cases}.$$

As above, entry (i, j) of the *row-standardized radius based proximity matrix* $\boldsymbol{W}$ is then defined by

$$w_{ij} := \frac{\tilde{w}_{ij}}{\tilde{w}_{i.}},$$

where $\tilde{w}_{i.} = \sum_{j=1}^{n} \tilde{w}_{ij}$. A property of radius based proximity matrices is that they are symmetric which is not necessarily the case for nearest neighbors based proximity matrices. In the following, we will always consider row-standardized proximity matrices and refer to them as *nearest neighbors matrices* and *radius matrices*. This standardization allows us to consider a sum of values weighted with the corresponding entry of the proximity matrix as a weighted average, as we will see in the SAR model.

In the following we recall the classical SAR model. By $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \boldsymbol{S})$ we denote that the random vector $\boldsymbol{Z}$ follows a n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{S}$.

Definition 1 (The simultaneous autoregressive (SAR) model) Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ be a n-dimensional random vector and $x_{i1}, \ldots, x_{ip}$ for $i = 1, \ldots, n$ associated (fixed) covariates. Let $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ be a matrix whose i-th row is given by $\boldsymbol{x}_i^T$, $\boldsymbol{x}_i := (1, x_{i1}, \ldots, x_{ip})^T$. Then the *simultaneous autoregressive (SAR) model* is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\lambda \in \mathbb{R}$ is the spatial dependence parameter, $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is the proximity matrix and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ the unknown regression coefficient. For the error vector we assume $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 \boldsymbol{\Sigma}_\epsilon)$ with a positive scalar σ and a diagonal matrix $\boldsymbol{\Sigma}_\epsilon \in \mathbb{R}^{n\times n}$ with positive diagonal entries.

So the components of $\boldsymbol{\epsilon}$ are independent. In our application we need to allow for different error variances per location, i.e., the diagonal elements of $\boldsymbol{\Sigma}_\epsilon$ are different. Furthermore we require the matrix $(\mathbf{I}_n - \lambda \boldsymbol{W})$ to be a full rank matrix in order to ensure that the model is well defined. Here $\mathbf{I}_n$ denotes the n-dimensional identity matrix.

Writing Equation (1) component wise yields

$$Y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \lambda \sum_{j\in N_i} w_{ij}(Y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j) + \epsilon_i \text{ for } i = 1, \ldots, n, \tag{2}$$

where $N_i = \{j | w_{ij} \neq 0\}$ is the set of neighbors of the i-th location as introduced above. As we consider row-standardized proximity matrices, the *spatial component* $\lambda \sum_{j\in N_i} w_{ij}(Y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j)$ can be seen as a weighted average of the deviations of the *linear component* $\boldsymbol{X}\boldsymbol{\beta}$ from the response in the corresponding neighborhood. In the following, we always assume the proximity matrix $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_\epsilon$ to be known.

2.1 Parameter estimation

We briefly sketch how parameters of the SAR model are estimated since we want to approach parameter estimation for the tSAR model in similar way. We follow Waller and Gotway (2004) (p. 365 ff.) who estimate the parameters by maximizing the likelihood. This requires to derive the likelihood function.

Since $(\mathbf{I}_n - \lambda \boldsymbol{W})$ has full rank, we can express Equation (1) as

$$\boldsymbol{Y} = (\mathbf{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{\epsilon} + \boldsymbol{X\beta}, \tag{3}$$

and we see that $\boldsymbol{Y}$ (as a full rank linear transformation of a normal random variable) is normally distributed with mean vector

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X\beta},$$

and covariance matrix

$$\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{\Sigma}_Y(\lambda), \tag{4}$$

where $\boldsymbol{\Sigma}_Y(\lambda) := (\mathbf{I}_n - \lambda \boldsymbol{W})^{-1}\boldsymbol{\Sigma}_\epsilon(\mathbf{I}_n - \lambda \boldsymbol{W}^T)^{-1}$.

Knowing the distribution of $\boldsymbol{Y}$, the likelihood function for $(\boldsymbol{\beta}, \sigma, \lambda)$ for given data $\boldsymbol{y}$ is given by

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \lambda) :=& (2\pi)^{-\frac{n}{2}} \det[\sigma^2 \boldsymbol{\Sigma}_Y(\lambda)]^{-\frac{1}{2}} \cdot \\ & \cdot \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T \frac{1}{\sigma^2}\boldsymbol{\Sigma}_Y(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \right]. \end{aligned}$$

Instead of maximizing the likelihood function, we minimize the negative log-likelihood given by

$$\begin{aligned} \ell(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \lambda) :=& -\log\left[L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \lambda)\right] \\ =& \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left\{\det[\boldsymbol{\Sigma}_Y(\lambda)]\right\} + \\ & + \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}_Y(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}). \end{aligned} \tag{5}$$

Estimation of $\boldsymbol{\beta}$

First we take the derivative of $\ell(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \lambda)$ with respect to $\boldsymbol{\beta}$ and set it to zero. Solving for $\boldsymbol{\beta}$ yields the (on λ dependent) estimate

$$\hat{\boldsymbol{\beta}}(\lambda) = \left[\boldsymbol{X}^T \boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{y}, \tag{6}$$

which is independent of σ. For fixed λ, this is the generalized least squares estimator for $\boldsymbol{\beta}$ (cf.,Kariya and Kurata (2004) p. 35).

Estimation of σ

We proceed in the same way for σ^2 and obtain the (on $\boldsymbol{\beta}$ and λ dependent) estimate

$$\hat{\sigma}^2(\boldsymbol{\beta}, \lambda) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}_Y(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{7}$$

The estimate for σ is given by its positive square root, i.e.,

$$\hat{\sigma}(\boldsymbol{\beta}, \lambda) = \sqrt{\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}_Y(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})}.$$

Estimation of λ

There is no closed form solution for λ. So we focus on the negative profile log-likelihood given by

$$\frac{n}{2}\log(2\pi)+\frac{n}{2}\log\left[\hat{\sigma}(\hat{\boldsymbol{\beta}}(\lambda),\lambda)^2\right]+\frac{1}{2}\log\left\{\det[\boldsymbol{\Sigma}_Y(\lambda)]\right\}+ \\ +\frac{\left[y-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\right]^T\boldsymbol{\Sigma}_Y(\lambda)\left[y-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)\right]}{2\hat{\sigma}(\hat{\boldsymbol{\beta}}(\lambda),\lambda)^2},$$

which is obtained by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\lambda)$ and σ by $\hat{\sigma}(\hat{\boldsymbol{\beta}}(\lambda),\lambda)$ in the negative log-likelihood function (5). This one dimensional nonlinear minimization problem can be solved by appropriate optimization algorithms and yields $\hat{\lambda}$, the estimate of λ. The estimation procedure is implemented in the `R` package `spdep` (see Bivand and Piras (2015)). For optimization, the `R` function `optimize` which is a combination of golden section search and successive parabolic interpolation (see Brent (1973)) is used. The final estimate of $\boldsymbol{\beta}$ is then given by $\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}(\hat{\lambda})$ and the final estimate of σ is given by $\hat{\sigma}=\hat{\sigma}(\hat{\boldsymbol{\beta}},\hat{\lambda})$.

2.2 Prediction and residuals

From Equation (2) it follows that the conditional expectation of $\boldsymbol{Y}$ at spatial location i, given the values of all other spatial locations, is

$$\begin{aligned}\mathrm{E}(Y_i|\boldsymbol{Y}_{-i}=\boldsymbol{y}_{-i})&=\mathrm{E}(Y_i|Y_j=y_j,j\in N_i)\\&=\boldsymbol{\beta}^T\boldsymbol{x}_i+\lambda\sum_{j\in N_i}w_{ij}(y_j-\boldsymbol{\beta}^T\boldsymbol{x}_j),\end{aligned}$$

where $z_{-i}=\{z_1,\dots z_n\}\setminus\{z_i\}$ for a $n-$dimensional vector z. So we define the *i-th local prediction of* Y, where the neighbors' values are observed, by

$$\hat{y}_{i|N_i}:=\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i+\hat{\lambda}\sum_{j\in N_i}w_{ij}(y_j-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j),$$

and the corresponding *vector of local predictions* is defined by

$$\hat{\boldsymbol{y}}_{|N}:=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\lambda}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Based on the prediction we can define the *i-th local residual* as

$$\hat{\epsilon}_i:=y_i-\hat{y}_{i|N_i}.$$

Since the local residual is the only type of residual we consider, we also refer to it just as the *i-th residual*. The *i-th standardized residual* is given by

$$\tilde{\epsilon}_i:=\frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(\Sigma_\epsilon)_{ii}}},$$

since $\mathrm{Var}(\epsilon_i) = \sigma^2(\Sigma_\epsilon)_{ii}$. From a good fit we expect the standardized residuals to be approximately identically and independent standard normally distributed.

Furthermore, an estimate for the *standard error of* $\hat{\beta}_i$ is provided by

$$\hat{\mathrm{se}}(\hat{\beta}_i) := \hat{\sigma}\sqrt{\left(\left[X^T\Sigma_Y(\hat{\lambda})^{-1}X\right]^{-1}\right)_{ii}},$$

since

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}(\lambda)) &= \mathrm{Var}\left(\left[\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{Y}\right)\\ &= \left[\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\,\mathrm{Var}(\boldsymbol{Y})\cdot\\ &\quad\cdot\left\{\left[\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\right\}^T\\ &=\sigma^2\left[\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1}.\end{aligned}\tag{8}$$

This can be used to test the significance of β_i. For fixed λ, $\hat{\boldsymbol{\beta}}$ is normally distributed (as a linear transformation of the normally distributed vector $\boldsymbol{Y}$). We use the following test for the significance of β_i with significance level α, null hypothesis $H_0 : \beta_i = 0$ and alternative $H_1 : \beta_i \neq 0$. We reject H_0 if

$$\left|\frac{\hat{\beta}_i}{\hat{\mathrm{se}}(\hat{\beta}_i)}\right| > \Phi^{-1}(1-\frac{\alpha}{2}),$$

where $\Phi^{-1}(1-\frac{\alpha}{2})$ denotes the $1-\frac{\alpha}{2}$ quantile of the $N(0,1)$ distribution. But we need to use this test with caution because the standard error was estimated with the assumption that λ was known. Thus the standard error is too small since we do not account for the variation in λ.

3 The tSAR model

The tSAR model is a way of extending the SAR model to allow for a Student t error distribution. We replace the assumption that the error vector is normally distributed by the assumption that the components of the error vector are univariate t-distributed. This allows for heavier tailed errors in our model.

3.1 Model definition

We say that the one dimensional random variable X follows a *t-distribution* with mean μ ($\mu \in \mathbb{R}$), scale parameter s^2 ($s \in \mathbb{R}, s > 0$) and ν ($\nu \in \mathbb{N}$) degrees of freedom if X has the density

$$t(x|\mu, s^2, \nu) := \Gamma\left(\frac{\nu+1}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^{-1}\frac{1}{\sqrt{\nu\pi s^2}}\left[1+\frac{(x-\mu)^2}{\nu s^2}\right]^{-\frac{\nu+1}{2}},$$

where $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function. We write $X \sim t(\mu, s^2, \nu)$. Furthermore, we denote by $Sc(X)$ the scale parameter of X. According to Kotz and Nadarajah (2004) (p. 10 ff.) it holds that

$$\mathrm{E}(X) = \mu, \tag{9}$$

and

$$\mathrm{Var}(X) = \frac{\nu}{\nu - 2}s^2, \tag{10}$$

for $\nu > 2$.

Definition 2 (tSAR model) In the *tSAR model* we assume that

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \lambda \boldsymbol{W}(Y - \boldsymbol{X\beta}) + \boldsymbol{\epsilon},$$

with $\epsilon_i \sim t(0, \sigma^2(\Sigma_\epsilon)_{ii}, \nu)$ with a positive scalar σ and a diagonal matrix $\boldsymbol{\Sigma_\epsilon} \in \mathbb{R}^{n\times n}$ with positive diagonal entries and $\nu > 2$ degrees of freedom. Furthermore, we assume that the components of the vector $\boldsymbol{\epsilon}$ are independent. $\boldsymbol{X}, \lambda, \boldsymbol{W}$ and $\boldsymbol{\beta}$ are defined as in Definition 1.

3.2 Parameter estimation

As in the SAR model, we estimate parameters by maximizing the likelihood while assuming $\boldsymbol{W}, \boldsymbol{\Sigma}_\epsilon$ and the degrees of freedom ν to be known. We start with deriving the likelihood function. Since

$$\boldsymbol{\epsilon} = (\mathbf{I}_n - \lambda \boldsymbol{W})(\boldsymbol{Y} - \boldsymbol{X\beta}),$$

the components of the vector

$$\boldsymbol{Z} := \boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})(\boldsymbol{Y} - \boldsymbol{X\beta}),$$

where $\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}$ is a diagonal matrix with i-th diagonal entry $(\Sigma_\epsilon)_{ii}^{-\frac{1}{2}}$, are identically and independent $t(0, \sigma^2, \nu)$ distributed. So the density $f_{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ is the product of its marginal densities. Furthermore, we have that

$$\boldsymbol{Y} = (\mathbf{I}_n - \lambda \boldsymbol{W})^{-1}\sqrt{\boldsymbol{\Sigma}_\epsilon}\boldsymbol{Z} + \boldsymbol{X\beta}.$$

We obtain the density of $\boldsymbol{Y}$ by density transformation.

$$f_Y(\boldsymbol{y}) = \left|\det\left[\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})\right]\right| \prod_{i=1}^{n} t\left(\left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X\beta})\right)_i |0, \sigma^2, \nu\right),$$

where $f_{\boldsymbol{Z}}$ is the density of $\boldsymbol{Z}$. Hence the negative log-likelihood of data $\boldsymbol{y}$ given the model parameters $(\boldsymbol{\beta}, \sigma, \lambda)$ is

$$\begin{aligned}\ell(\boldsymbol{y}|\boldsymbol{\beta}, \lambda, \sigma) := &- \log\left\{\left|\det\left[\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})\right]\right|\right\} \\ &- \sum_{i=1}^{n} \log\left[t\left(\left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X\beta})\right)_i |0, \sigma^2, \nu\right)\right].\end{aligned} \tag{11}$$

Unfortunately we can not proceed as before (i.e., take the derivatives with respect to $\boldsymbol{\beta}$ and σ, set them to zero and solve analytically for the parameters) due to the more complex form of the likelihood function. For illustration of this problem we write down the derivative with respect to $\boldsymbol{\beta}$.

$$\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{y}|\boldsymbol{\beta},\lambda,\sigma) = \mathrm{c} + \frac{\nu+1}{2}\sum_{i=1}^{n}\frac{1}{\left[1+\frac{m_i(\boldsymbol{\beta})^2}{\sigma^2\nu}\right]}\frac{2m_i(\boldsymbol{\beta})}{\sigma^2\nu}\left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda\boldsymbol{W})\boldsymbol{X}\right)_i,$$

where c is a constant independent of $\boldsymbol{\beta}$, $m_i(\boldsymbol{\beta}) = \left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda\boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X\beta})\right)_i$ and $\left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda\boldsymbol{W})\boldsymbol{X}\right)_i$ is the i-th row of $\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda\boldsymbol{W})\boldsymbol{X}$. If we set this equation to zero, we can not solve it analytically for $\boldsymbol{\beta}$. Numerical optimization for all parameters would be computationally very complex since $\boldsymbol{\beta}$ is often high dimensional. Therefore we suggest to estimate $\boldsymbol{\beta}$ and σ as explained in the following.

Estimation of $\boldsymbol{\beta}$

A simple analytic estimator for $\boldsymbol{\beta}$ is the on λ dependent generalized least squares estimator, i.e.,

$$\hat{\boldsymbol{\beta}}(\lambda) = \left[\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{Y},$$

as in the SAR model. For fixed λ, this is the best linear unbiased estimator according to the Gauß Markov Theorem (cf., Kariya and Kurata (2004) p. 34).

Estimation of σ

For σ we suggest the following estimate dependent on $\boldsymbol{\beta}$ and λ,

$$\hat{\sigma}^2(\boldsymbol{\beta},\lambda) = \frac{\nu-2}{\nu}\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X\beta})^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}),$$

since $\hat{\sigma}^2(\boldsymbol{\beta},\lambda)$ can be written as

$$\begin{aligned}\hat{\sigma}^2(\hat{\boldsymbol{\beta}},\hat{\lambda}) &= \frac{\nu-2}{\nu}\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_Y(\hat{\lambda})^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\\ &= \frac{\nu-2}{\nu}\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\mathbf{I}_n - \hat{\lambda}\boldsymbol{W}^T)\boldsymbol{\Sigma}_\epsilon^{-1}(\mathbf{I}_n - \hat{\lambda}\boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\\ &= \frac{\nu-2}{\nu}\frac{1}{n}\hat{\boldsymbol{\epsilon}}^T\boldsymbol{\Sigma}_\epsilon^{-1}\hat{\boldsymbol{\epsilon}}\\ &= \frac{\nu-2}{\nu}\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\epsilon}_i^2}{(\Sigma_\epsilon)_{ii}},\end{aligned}$$

where we used the definition of the prediction vector $\hat{\boldsymbol{y}}_{|N} := \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\lambda}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ and the residual $\hat{\boldsymbol{\epsilon}} := \boldsymbol{y} - \hat{\boldsymbol{y}}_{|N}$ to express $\hat{\boldsymbol{\epsilon}}$ as

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \boldsymbol{y} - \hat{\boldsymbol{y}}_{|N}\\ &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\lambda}\boldsymbol{W}\boldsymbol{y} + \hat{\lambda}\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}}\\ &= (\mathbf{I}_n - \hat{\lambda}\boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).\end{aligned}$$

The quantity $\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\epsilon}_i^2}{(\Sigma_\epsilon)_{ii}}$ is an estimate of the variance of $\epsilon_i/\sqrt{(\Sigma_\epsilon)_{ii}}$ and so $\hat{\sigma}^2(\hat{\boldsymbol{\beta}}, \hat{\lambda})$ is an estimate of the scale parameter.

Estimation of λ

For the estimation of λ we proceed as in the SAR model, i.e., we obtain the negative profile log-likelihood by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\lambda)$ and σ by $\hat{\sigma}(\hat{\boldsymbol{\beta}}(\lambda), \lambda)$ in the negative log-likelihood function (11). Then $\hat{\lambda}$ is defined as the minimum of the negative profile log-likelihood which is found numerically. As before we set $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\lambda})$ and $\hat{\sigma} = \hat{\sigma}(\hat{\boldsymbol{\beta}}, \hat{\lambda})$.

3.3 Prediction and residuals

The *vector of local predictions* $\hat{\boldsymbol{y}}_{|N}$ and the *residual vector* $\hat{\boldsymbol{\epsilon}}$ are defined as for the SAR model, i.e.,

$$\hat{\boldsymbol{y}}_{|N} := \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\lambda}\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

and

$$\hat{\epsilon}_i := y_i - \hat{y}_{i|N_i}.$$

Since $Sc(\epsilon_i) = \sigma^2(\Sigma_\epsilon)_{ii}$, we define the *i-th standardized residual* by

$$\tilde{\epsilon}_i := \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(\Sigma_\epsilon)_{ii}}}.$$

As in (3), we can write $\boldsymbol{Y}$ as

$$\boldsymbol{Y} = (\mathbf{I}_n - \lambda\boldsymbol{W})^{-1}\boldsymbol{\epsilon} + \boldsymbol{X}\boldsymbol{\beta},$$

and obtain similarly to Equation (4),

$$\operatorname{Var}(\boldsymbol{Y}) = \frac{\nu}{\nu - 2}\sigma^2\boldsymbol{\Sigma}_Y(\lambda),$$

where $\boldsymbol{\Sigma}_Y(\lambda) := (\mathbf{I}_n - \lambda\boldsymbol{W})^{-1}\boldsymbol{\Sigma}_\epsilon(\mathbf{I}_n - \lambda\boldsymbol{W}^T)^{-1}$. Therefore we get similar to (8)

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}(\lambda)) = \frac{\nu}{\nu - 2}\sigma^2(\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\lambda)^{-1}\boldsymbol{X})^{-1},$$

and thus we estimate the *standard error of $\hat{\beta}_i$* by

$$\hat{\operatorname{se}}(\hat{\beta}_i) := \sqrt{\frac{\nu}{\nu - 2}\hat{\sigma}^2((\boldsymbol{X}^T\boldsymbol{\Sigma}_Y(\hat{\lambda})^{-1}\boldsymbol{X})^{-1})_{ii}}.$$

3.4 Specifying the matrix $\boldsymbol{\Sigma}_\epsilon$

To estimate a SAR or tSAR model we need to specify the matrix $\boldsymbol{\Sigma}_\epsilon$ which is proportional to the covariance matrix of the error vector. One possibility would be to choose this equal to the identity matrix which leads to all locations having the same variance. But we also want to account for different error variances. Therefore we provide a variance estimate which uses the restriction to a neighborhood. We define the *local empirical variance* of the spatial variable Z at location i, Z_i, with respect to the proximity matrix $\boldsymbol{W}$ as follows

$$\hat{\sigma}^2_W(Z_i) := \frac{1}{|N_i| - 1} \sum_{j \in N_i} (z_j - \bar{z}_{N_i})^2, \tag{12}$$

where the z_j are observations of Z, $N_i = \{j | w_{ij} \neq 0\}$ is the neighborhood of location i induced by $\boldsymbol{W}$, $|N_i|$ is the cardinality of the set N_i and $\bar{z}_{N_i} = \frac{1}{|N_i|} \sum_{j \in N_i} z_j$. The corresponding *local empirical variance matrix* is a diagonal matrix with i-th diagonal entry equal to $\hat{\sigma}^2_W(Z_i)$. For a SAR or tSAR model with response $\boldsymbol{Y}$, covariates $x_1, \ldots, x_p$ and proximity matrix $\boldsymbol{W}$, we propose to specify $\boldsymbol{\Sigma}_\epsilon$ in the following way.

1. We fit a linear regression model with response variable $\boldsymbol{Y}$ and covariates $x_1, \ldots, x_p$, i.e., we assume
$$y_i = (x_{1i}, \ldots, x_{pi})\boldsymbol{\beta}_{lm} + \epsilon_{lm,i}$$
with $\epsilon_{lm,i} \sim N(0, \sigma^2)$, $\boldsymbol{\beta}_{lm} \in \mathbb{R}^p, \sigma \in \mathbb{R}_+$. We obtain $\hat{\boldsymbol{\beta}}_{lm}$, the estimate of $\boldsymbol{\beta}_{lm}$ by least squares estimation. The i-th residual r_i is given by
$$r_i = y_i - (x_{1i}, \ldots, x_{pi})\hat{\boldsymbol{\beta}}_{lm}.$$
2. Then we set $\boldsymbol{\Sigma}_\epsilon$ equal to the local empirical variance matrix of the residual vector $\boldsymbol{r}$ with respect to $\boldsymbol{W}$. So $\boldsymbol{\Sigma}_\epsilon$ is a diagonal matrix with $i-$th diagonal entry $(\Sigma_\epsilon)_{ii} = \hat{\sigma}^2_W(r_i)$, where $\hat{\sigma}^2_W(\cdot)$ is generally defined in (12). We call this the *local regression variance matrix of* $\boldsymbol{Y}$ *with respect to* $\boldsymbol{W}$.

4 Simulation study

In this section we study if the proposed estimators of the tSAR model behave in a reasonable way and how they compare to the estimators of the already existing SAR model.

We simulate from a tSAR model in the following way.

1. (number of locations n) We specify the number of locations n as 250 or 1500.
2. (proximity matrix $\boldsymbol{W}$) We randomly select n different longitude/latitude values of the WFAS data set introduced in Section 5.1 to determine locations and corresponding neighborhoods. We set the proximity matrix $\boldsymbol{W}$ equal to a nearest neighbors matrix with $k = 30$ neighbors.

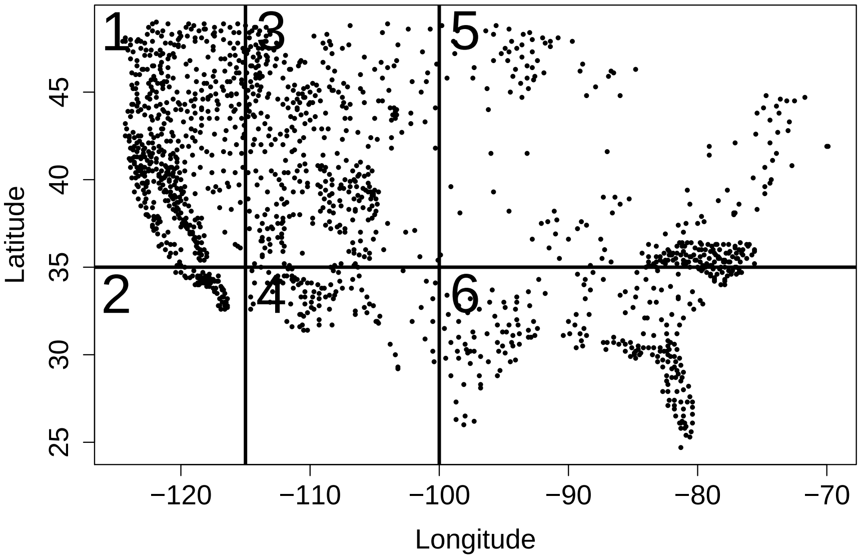

Fig. 2 Locations of the weather stations of the WFAS data and the 6 regions used in the simulation study visualized on the map.

3. (covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_7$) We obtain the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_7$ by sampling n times independently from the following distributions:

$$\begin{aligned} \boldsymbol{x}_1, \ldots, \boldsymbol{x}_5 &: \text{standard normal} \\ \boldsymbol{x}_6 &: \text{bernoulli with } p = 0.3 \\ \boldsymbol{x}_7 &: \text{bernoulli with } p = 0.7 \end{aligned} \tag{13}$$

4. (degrees of freedom ν) We specify the degrees of freedom ν as 4 or 20.
5. (simulation of $\boldsymbol{\epsilon}$) To account for a varying variance, we define 6 regions (see Figure 2) with corresponding $s_1 = 4, s_2 = 0.6, s_3 = 5, s_4 = 0.3, s_5 = 4, s_6 = 6$ and simulate independently for $i = 1, \ldots, n$: If location i belongs to region j simulate ϵ_i from $t(0, s_j^2, \nu)$.
6. (coefficients $\boldsymbol{\beta}$) We set

$$\beta_0 = 3,\ \beta_1 = 10,\ \beta_2 = 4,\ \beta_3 = 5,\ \beta_4 = 2,\ \beta_5 = 8,\ \beta_6 = 1,\ \beta_7 = 3.$$

7. (spatial parameter λ) We specify λ as 0.4 or 0.8.
8. (response $\boldsymbol{y}$) According to the assumptions of the tSAR model we set

$$\boldsymbol{y} = (\mathbf{I}_n - \lambda \boldsymbol{W})^{-1} \boldsymbol{\epsilon} + \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_7)^T$ and $\boldsymbol{X} = (\mathbf{1}^T, \boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_7^T)$.

Table 1 Different models estimated in the simulation study.

model		$\boldsymbol{\Sigma}_\epsilon$
1	SAR	$\mathbf{I}_n$
2	tSAR	$\mathbf{I}_n$
3	SAR	local regression variance matrix
4	tSAR	local regression variance matrix
5	SAR	true
6	tSAR	true

Choosing between SAR and tSAR and the different choices for $\boldsymbol{\Sigma}_\epsilon$ leads to 6 different models (see Table 1) that are estimated from the simulated data.

In the tSAR model we have one additional parameter ν, the degrees of freedom, which was assumed to be known in Section 3. Instead of specifying this parameter we use numerical optimization to obtain an estimate for it. We use the `R` function `optimize` with high tolerance (tolerance = 1) to speed up computation. Here we allow ν to be a real parameter between 3 and 20.

Note that in the SAR model σ is the standard deviation of $\epsilon_i/\sqrt{(\Sigma_\epsilon)_{ii}}$, whereas in the tSAR model σ is the square root of the scale parameter of $\epsilon_i/\sqrt{(\Sigma_\epsilon)_{ii}}$ and the standard deviation is given by $\sqrt{\frac{\nu}{\nu-2}}\sigma$. For easier comparison we introduce *s, the standard deviation of* $\epsilon_i/\sqrt{(\Sigma_\epsilon)_{ii}}$, and define its estimate $\hat{s}$, depending on the model, by

$$\begin{aligned} \hat{s} &:= \hat{\sigma} && \text{if the SAR model is used,} \\ \hat{s} &:= \sqrt{\frac{\nu}{\nu-2}}\ \hat{\sigma} && \text{if the tSAR model is used.} \end{aligned} \tag{14}$$

The results of the simulation study are shown in Table 2. To evaluate the estimates we use the root mean squared error which is given by

$$\operatorname{RMSE}(\hat{\boldsymbol{\theta}}) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\frac{1}{r}\sum_{i=1}^{r}(\theta_j - \hat{\theta}_{ji})^2}, \tag{15}$$

where r is the number of replications (in our case $r = 500$), θ_j is the j-th component of the p-dimensional vector $\boldsymbol{\theta}$ and $\hat{\theta}_{ji}$ its estimate in the i-th replication.

First we analyze the results with respect to the number of locations n. We compare models that only differ in the choice of this parameter. One usually expects that the root mean squared error decreases as the number of stations increases. We observe this behavior for all parameters in cases where $\boldsymbol{\Sigma}_\epsilon$ is the true value or the local regression variance matrix. If $\boldsymbol{\Sigma}_\epsilon$ is the identity matrix this does not hold for the parameter s. The parameter s scales $\boldsymbol{\Sigma}_\epsilon$ and if $\boldsymbol{\Sigma}_\epsilon$ is specified incorrectly we cannot expect a reasonable estimate for s. Furthermore, the results show that the choice of $\boldsymbol{\Sigma}_\epsilon$ has an influence on the estimates for $\boldsymbol{\beta}$. Comparing models that only differ in the choice of $\boldsymbol{\Sigma}_\epsilon$, the best estimates are obtained when $\boldsymbol{\Sigma}_\epsilon$ is the true value, the second best when $\boldsymbol{\Sigma}_\epsilon$ is the local regression variance matrix and the worst when $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}_n$. There is a notable difference between RMSE($\hat{\boldsymbol{\beta}}$) in cases where $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}_n$ compared to cases where $\boldsymbol{\Sigma}_\epsilon$ is equal to

Table 2 Results of the simulation study. The first two columns specify the number of locations and the model. Other columns show the root mean squared error, the average log-likelihood (ll), the true parameter and the average of estimated parameters. For $\boldsymbol{\beta}$ we average over its components.

		RMSE												
n	model	$\hat{\boldsymbol{\beta}}$	$\hat{\lambda}$	$\hat{s}$	$\hat{\nu}$	ll	$\hat{\boldsymbol{\beta}}$	$\boldsymbol{\beta}$	$\hat{\lambda}$	λ	$\hat{s}$	s	$\hat{\nu}$	ν
250	1	0.428	0.086	2.961		-722	4.49	4.5	0.31	0.4	4.38	1.41		
250	2	0.425	0.040	2.965	0.41	-660	4.49	4.5	0.36	0.4	4.38	1.41	3.59	4
250	3	0.106	0.062	0.390		-567	4.50	4.5	0.34	0.4	1.02	1.41		
250	4	0.107	0.047	0.389	0.41	-521	4.50	4.5	0.35	0.4	1.03	1.41	3.59	4
250	5	0.069	0.032	0.032		-474	4.50	4.5	0.37	0.4	1.38	1.41		
250	6	0.069	0.035	0.029	2.17	-457	4.50	4.5	0.36	0.4	1.39	1.41	4.98	4
250	1	0.651	0.054	2.972		-728	4.51	4.5	0.75	0.8	4.39	1.41		
250	2	0.652	0.035	2.970	0.41	-665	4.51	4.5	0.77	0.8	4.38	1.41	3.59	4
250	3	0.165	0.049	0.474		-576	4.50	4.5	0.75	0.8	0.94	1.41		
250	4	0.164	0.034	0.478	0.41	-529	4.50	4.5	0.77	0.8	0.94	1.41	3.59	4
250	5	0.113	0.022	0.035		-480	4.50	4.5	0.78	0.8	1.38	1.41		
250	6	0.120	0.020	0.025	2.34	-464	4.50	4.5	0.78	0.8	1.39	1.41	5.05	4
250	1	0.323	0.067	2.220		-652	4.49	4.5	0.33	0.4	3.27	1.05		
250	2	0.321	0.037	2.223	16.41	-607	4.49	4.5	0.36	0.4	3.28	1.05	3.59	20
250	3	0.073	0.048	0.101		-485	4.50	4.5	0.35	0.4	0.95	1.05		
250	4	0.074	0.043	0.098	16.14	-465	4.50	4.5	0.36	0.4	0.96	1.05	3.87	20
250	5	0.052	0.032	0.022		-402	4.50	4.5	0.37	0.4	1.03	1.05		
250	6	0.053	0.036	0.020	5.67	-402	4.50	4.5	0.36	0.4	1.03	1.05	16.41	20
250	1	0.462	0.047	2.219		-657	4.50	4.5	0.75	0.8	3.27	1.05		
250	2	0.460	0.031	2.218	16.41	-613	4.50	4.5	0.77	0.8	3.27	1.05	3.59	20
250	3	0.111	0.045	0.170		-499	4.50	4.5	0.76	0.8	0.88	1.05		
250	4	0.113	0.036	0.169	16.25	-476	4.50	4.5	0.76	0.8	0.88	1.05	3.76	20
250	5	0.080	0.021	0.022		-408	4.50	4.5	0.78	0.8	1.03	1.05		
250	6	0.084	0.022	0.015	5.14	-409	4.50	4.5	0.78	0.8	1.04	1.05	16.82	20
1500	1	0.175	0.009	3.213		-4430	4.50	4.5	0.39	0.4	4.63	1.41		
1500	2	0.175	0.005	3.214	0.41	-4028	4.50	4.5	0.39	0.4	4.63	1.41	3.59	4
1500	3	0.034	0.012	0.385		-3162	4.50	4.5	0.39	0.4	1.03	1.41		
1500	4	0.034	0.009	0.386	0.39	-2986	4.50	4.5	0.39	0.4	1.03	1.41	3.64	4
1500	5	0.029	0.004	0.009		-2981	4.50	4.5	0.40	0.4	1.40	1.41		
1500	6	0.029	0.008	0.008	0.51	-2865	4.50	4.5	0.39	0.4	1.41	1.41	4.18	4
1500	1	0.274	0.012	3.237		-4464	4.50	4.5	0.79	0.8	4.65	1.41		
1500	2	0.274	0.009	3.239	0.41	-4060	4.50	4.5	0.79	0.8	4.65	1.41	3.59	4
1500	3	0.051	0.010	0.446		-3205	4.50	4.5	0.79	0.8	0.97	1.41		
1500	4	0.052	0.008	0.445	0.40	-3027	4.50	4.5	0.79	0.8	0.97	1.41	3.63	4
1500	5	0.043	0.005	0.001		-3021	4.50	4.5	0.80	0.8	1.42	1.41		
1500	6	0.044	0.006	0.006	0.50	-2902	4.50	4.5	0.79	0.8	1.42	1.41	4.15	4
1500	1	0.134	0.012	2.407		-3997	4.50	4.5	0.39	0.4	3.46	1.05		
1500	2	0.134	0.005	2.407	16.41	-3722	4.50	4.5	0.39	0.4	3.46	1.05	3.59	20
1500	3	0.024	0.015	0.081		-2694	4.50	4.5	0.38	0.4	0.97	1.05		
1500	4	0.024	0.014	0.082	11.97	-2669	4.50	4.5	0.39	0.4	0.97	1.05	8.18	20
1500	5	0.022	0.005	0.003		-2547	4.50	4.5	0.40	0.4	1.05	1.05		
1500	6	0.022	0.009	0.002	3.67	-2545	4.50	4.5	0.39	0.4	1.05	1.05	17.56	20
1500	1	0.205	0.005	2.406		-4028	4.51	4.5	0.79	0.8	3.46	1.05		
1500	2	0.204	0.003	2.408	16.41	-3753	4.51	4.5	0.80	0.8	3.46	1.05	3.59	20
1500	3	0.036	0.009	0.136		-2737	4.50	4.5	0.79	0.8	0.92	1.05		
1500	4	0.036	0.008	0.135	12.42	-2713	4.50	4.5	0.79	0.8	0.92	1.05	7.69	20
1500	5	0.032	0.003	0.003		-2578	4.50	4.5	0.80	0.8	1.05	1.05		
1500	6	0.032	0.004	0.001	3.70	-2580	4.50	4.5	0.80	0.8	1.05	1.05	17.58	20

the local regression variance matrix. This shows that introducing the local regression variance matrix brings a notable improvement for estimating $\boldsymbol{\beta}$ compared to the trivial choice $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}_n$. For λ, reasonable estimates are provided in all cases whereas the best estimates are usually obtained when $\boldsymbol{\Sigma}_\epsilon$ is the true value. We see that the choice of $\boldsymbol{\Sigma}_\epsilon$ also has influence on the estimates for s, where the influence is similar as for $\boldsymbol{\beta}$, i.e., the best estimates are obtained when $\boldsymbol{\Sigma}_\epsilon$ is the true value, the second best when $\boldsymbol{\Sigma}_\epsilon$ is equal to the local regression variance matrix and the worst when $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}_n$. The differences in RMSE($\hat{s}$) are rather big since it is difficult to estimate s, which scales $\boldsymbol{\Sigma}_\epsilon$, if $\boldsymbol{\Sigma}_\epsilon$ is not specified correctly. Analysing the estimation of ν we observe that big values of RMSE($\hat{\nu}$) are obtained in cases where $\nu = 20$ and $\boldsymbol{\Sigma}_\epsilon$ is not equal to the true value. In these cases ν was estimated too low. Specifying $\boldsymbol{\Sigma}_\epsilon$ incorrectly causes that the variance of the residuals is estimated too low or too high for some of them which then causes that a t-distribution with lower degrees of freedom provides a better fit. Evaluating the overall fit with the log-likelihood and comparing models that only differ in the choice of one parameter we see that the choice between SAR and tSAR and the choice of $\boldsymbol{\Sigma}_\epsilon$ has influence. The tSAR model leads to higher likelihood values when $\nu = 4$ or mostly similar values when $\nu = 20$. For $\boldsymbol{\Sigma}_\epsilon$ the highest likelihood values are obtained when $\boldsymbol{\Sigma}_\epsilon$ is the true value, the second highest when $\boldsymbol{\Sigma}_\epsilon$ is equal to the local regression variance matrix and the lowest when $\boldsymbol{\Sigma}_\epsilon = \mathbf{I}_n$.

5 Application

We use the two models, SAR and tSAR, to fit data to assess the risk of fire danger in the US.

5.1 Data description

The data is obtained from the Wildland Fire Assessment System (WFAS) and contains the following variables observed at 1542 stations on the 23rd of June 2015.

- $Elev$ = Elevation in feet divided by 100
- Lat = Latitude
- $Long$ = Longitude
- Tmp = Temperature in Fahrenheit
- RH = Relative humidity in percent
- $Wind$ = Wind speed (10 min avg wind) in mi/h
- PPT = 24h precipitation in inches
- BI = Burning Index calculated according to the National Fire Danger Rating System (cf., National Wildfire Coordinating Group (2002)) (number related to the contribution of fire behavior to the effort of containing a fire. It is expressed as a numeric value closely related to the flame length in feet multiplied by 10.)

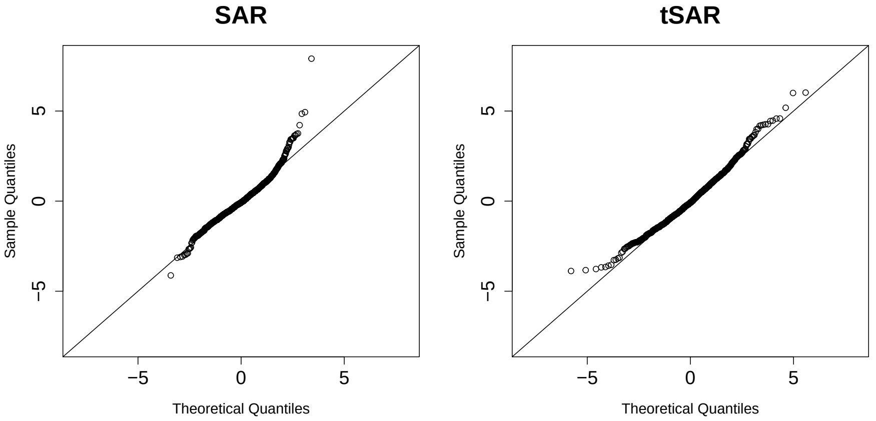

Fig. 3 qq-plots for a SAR and a tSAR model. We plot the quantiles of the standard normal distribution against the quantiles of the standardized residuals of the SAR model and the quantiles of the t-distribution with mean zero, scale parameter 1 and 6 degrees of freedom against the quantiles of the standardized residuals of the tSAR model with $\nu = 6$.

5.2 Model fitting

We consider the Burning Index BI as response variable and the other variables as covariates. These covariates can be measured using simple weather station technology. For our approach, there is no expert knowledge required compared to the calculation of the Burning Index according to the National Fire Danger Rating System.

Fitting several SAR and tSAR models, we observed misbehavior in the residuals. The residuals did not follow the desired normal or t-distribution. Figure 3 illustrates this problem for one case where we fit one SAR and one tSAR model with $\nu = 6$ degrees of freedom. We use BI as response and all other variables as covariates. As proximity matrix $\boldsymbol{W}$ we choose a nearest neighbors matrix with $k = 30$ neighbors and for $\boldsymbol{\Sigma}_\epsilon$ we use the local regression variance matrix of Y with respect to $\boldsymbol{W}$. To deal with this problem and to further improve our fit, we now consider Box-Cox transformations of the response variable (cf., Box and Cox (1964)) for SAR models.

We show how Box-Cox transformations that were developed for linear regression models can be used for SAR and tSAR models. We are given $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, an observation of the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. For $l \in \mathbb{R}$ and $m \in \mathbb{R}$ such that $Y_i > -m$ for all $i = 1, \ldots, n$, the Box-Cox transformed variable $Y_i^{m,l}$ is given by

$$Y_i^{m,l} = \begin{cases} \frac{(Y_i+m)^l-1}{l}, & \text{if } l \neq 0 \\ \log(Y_i + m), & \text{else} \end{cases} . \tag{16}$$

We consider l and m fixed and assume that $\boldsymbol{Y}^{m,l}$ is distributed according to a SAR or tSAR model with parameters $\boldsymbol{\theta}^{m,l} = (\boldsymbol{\beta}^{m,l}, \sigma^{m,l}, \lambda^{m,l})$. We denote its log-likelihood by $\ell_S(\boldsymbol{y}^{m,l}|\boldsymbol{\theta}^{m,l})$ where the observation $y_i^{m,l}$ of $Y_i^{m,l}$ is obtained by applying the same transformation (16) on the observation y_i. The density of $\boldsymbol{Y}$ can be obtained using the density transformation rule. The log-likelihood of $\boldsymbol{\theta}^{m,l}$ with respect to the observations $y_1, \ldots, y_n$ is then given by

$$\ell(\boldsymbol{y}|\boldsymbol{\theta}^{m,l}) = \sum_{i=1}^{n}(l-1)\log(y_i + m) + \ell_S(\boldsymbol{y}^{m,l}|\boldsymbol{\theta}^{m,l}).$$

The log-likelihood is a sum of two components where the first component is independent of $\boldsymbol{\theta}^{m,l}$, and therefore not needed for the maximization with regard to $\boldsymbol{\theta}^{m,l}$. So we need to maximize the second component which we know how to do since it is the log-likelihood of a SAR or tSAR model. Knowing the log-likelihood, the corresponding BIC is

$$\mathrm{BIC}(\boldsymbol{y}, \boldsymbol{\theta}^{m,l}) = -2\ell(\boldsymbol{y}|\boldsymbol{\theta}^{m,l}) + \dim(\boldsymbol{\theta}^{m,l})\log(n),$$

which can be used for selection among different models corresponding to different m and l values.

For fitting SAR models we use a step wise procedure where we adjust the Box-Cox transformation parameter and eliminate a non-significant covariate in each step. The procedure (Algorithm 1) for a given variable $\boldsymbol{R}$, parameter m and proximity matrix $\boldsymbol{W}$ is shown in the following. The available covariates are denoted by $x_1, \ldots, x_p$. Algorithm 1 is applied to the response variable $\boldsymbol{BI}$ with $m = 10$ and different choices of the proximity matrix $\boldsymbol{W}$. Instead of iterating over different values for m we choose one value, 10, to reduce computational time. For the proximity matrix we use nearest neighbors matrices with $k = 10, 20, 30, 40, 50$ neighbors and radius matrices with radius $r = 350, 500$. So we obtain 7 different models corresponding to different proximity matrices.

Algorithm 1 Step wise procedure for SAR models

```
$\mathcal{X} \leftarrow \{x_1, \ldots, x_p\}$
$maxp \leftarrow 1$
$c \leftarrow \{\}$
while $maxp > 0.05$ do
    $\mathcal{X} = \mathcal{X} \setminus \{c\}$
    for $l = -2, -1, -1/2, -1/3, 0, 1/3, 1/2, 1, 2$ do
        $\boldsymbol{Y} \leftarrow \begin{cases} \frac{(\boldsymbol{R}+m)^l - 1}{l}, & \text{if } l \neq 0 \\ \log(\boldsymbol{R} + m), & \text{else} \end{cases}$ componentwise
        $mod_l \leftarrow$ fitted SAR model with response variable $\boldsymbol{Y}$, covariates $\mathcal{X}$, proximity matrix $\boldsymbol{W}$ and
        $\boldsymbol{\Sigma}_\epsilon$ is the local regression variance matrix of $\boldsymbol{Y}$ with respect to $\boldsymbol{W}$.
    end for
    $mod \leftarrow$ model with lowest BIC among $\{mod_l | l = -2, -1, -1/2, -1/3, 0, 1/3, 1/2, 1, 2\}$
    $maxp \leftarrow$ maximum of the p-values of the tests for significance of the coefficients in model $mod$
    $c \leftarrow$ covariate corresponding to $maxp$
end while
```

After fitting SAR models using the procedure just described, we fit tSAR models. We proceed in the following way. For a certain proximity matrix $\boldsymbol{W}$ we take the same covariates and transformation as in the corresponding just fitted SAR model and fit a tSAR model where we optimize the degrees of freedom parameter ν numerically. For

Table 3 BIC for different models for transformed BI and the value of the transformation parameter l. “nnx" means that a nearest neighbors matrix with x neighbors was used and “rx" means that a radius matrix with radius x was used.

model type	nn10	nn20	nn30	nn40	nn50	r350	r500
SAR	12624.72	12488.34	12480.19	12499.37	12539.93	12617.03	12737.47
tSAR	12424.78	**12400.30**	12418.51	12438.61	12470.69	12561.87	12684.05
l	1/3	1/3	1/3	1/3	1/3	1/3	1/3

Table 4 Parameter estimates, estimated standard errors and their quotient for the model with the best BIC of the Box-Cox transformed Burning Index ($m = 10, l = \frac{1}{3}$).

	estimate	$\widehat{\text{se}}$	estimate/$\widehat{\text{se}}$
Intercept	7.72	1.14	6.76
Elev	0.01	0.00	2.85
Lat	-0.07	0.03	-2.44
Long	-0.03	0.01	-2.56
RH	-0.04	0.00	-12.39
Wind	0.16	0.01	20.97
PPT	-0.80	0.13	-6.34
λ	0.85		
σ	0.84		
ν	6.34		

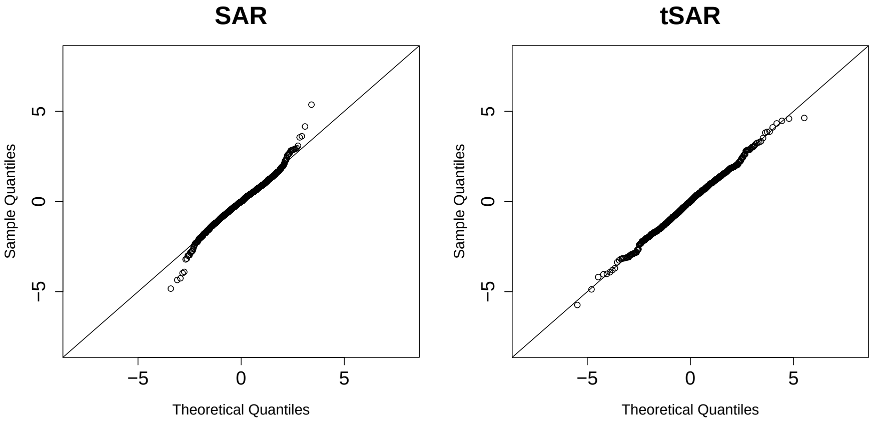

Fig. 4 qq-plots for the SAR and tSAR model with the best BIC value.

the matrix $\mathbf{\Sigma}_\epsilon$ we use as before the local regression variance matrix of the transformed response variable with respect to $\boldsymbol{W}$. Table 3 shows the BIC values of the models. If we consider only nearest neighbors matrices, we see that the BIC of the worst tSAR model is still lower than the BIC of the best SAR model. The best model is a tSAR model where the proximity matrix is a nearest neighbors matrix with $k = 20$ neighbors. Estimates for this model are given in Table 4.

In Figure 4 we check if the residuals of the best tSAR model have the distribution as expected. As the data points do not deviate far from the $x = y$ line, our fitted model seems to be appropriate. For comparison we also show this plot for the SAR model with the lowest BIC. We see that the tSAR model is not only preferred in terms of BIC.

5.3 Out of sample prediction

Now we perform out of sample predictions. This allows us to predict the Burning Index at locations where only the covariates are available. To do so, we need to relate a random variable at a location which was not part of the sample to $\boldsymbol{Y}$, the vector of random variables in the sample. For an out of sample random variable Y_o at location l_o we assume that

$$Y_o = \boldsymbol{\beta}^T \boldsymbol{x}_o + \lambda \sum_{j \in N_o} w_{oj}(Y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j) + \epsilon_o,$$

where w_{oj} relates location l_o to l_j for $j = 1...n$ such that $\sum_{j=1}^n w_{oj} = 1$ to stay consistent with the row-standardized proximity matrix. We will choose w_{oj} similar to how we chose the entries of the proximity matrix. If $\boldsymbol{W}$ is a k nearest neighbors matrix, w_{oj} is the inverse distance between location l_o and l_j times a standardization constant, if location l_j is among the k nearest neighbors of l_o and zero else. N_o is the neighborhood of location l_o defined as in Section 2. For the error we assume $\epsilon_o \sim N(0, \sigma^2 \Sigma_o)$ in the case of a SAR model or $\epsilon_o \sim t(0, \sigma^2 \Sigma_o, \nu)$ in the case of a tSAR model. Similar to the SAR and tSAR model, Σ_o is assumed to be known. We specify Σ_o similar to how we specified $\boldsymbol{\Sigma}_\epsilon$. If $\boldsymbol{\Sigma}_\epsilon$ is the local regression variance matrix of $\boldsymbol{Y}$, the diagonal entries of $\boldsymbol{\Sigma}_\epsilon$ were calculated with linear regression residuals $r_1, \ldots r_n$. Σ_o is then the empirical variance of $\{r_j | j \in N_o\}$.

With this assumption the expectation of Y_o given $\boldsymbol{Y}$ is given by

$$\mathrm{E}(Y_o|\boldsymbol{Y}) = \mathrm{E}(Y_o|Y_j = y_j, j \in N_o) = \boldsymbol{\beta}^T \boldsymbol{x}_0 + \lambda \sum_{j \in N_o} w_{oj}(y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j),$$

where $\boldsymbol{\beta}$, σ and ν are the parameters of the SAR or tSAR model for $\boldsymbol{Y}$. So we define the local prediction of Y_o, where the neighbors' values are observed, by

$$\hat{y}_{o|N_o} := \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_o + \hat{\lambda} \sum_{j \in N_o} w_{oj}(y_j - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j),$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ are the estimates of the model for $\boldsymbol{Y}$. In addition to the prediction we provide confidence intervals. The $1 - \alpha$ confidence interval is given by

$$\mathrm{CI}(1-\alpha) = \begin{cases} \hat{y}_{o|N_o} \pm \Phi^{-1}(1 - \frac{\alpha}{2}, 0, \hat{\sigma}^2 \Sigma_o) & \text{for SAR} \\ \hat{y}_{o|N_o} \pm t^{-1}(1 - \frac{\alpha}{2}, 0, \hat{\sigma}^2 \Sigma_o, \nu) & \text{for tSAR} \end{cases},$$

where $\Phi^{-1}(1 - \frac{\alpha}{2}, 0, \hat{\sigma}^2 \Sigma_o)$ and $t^{-1}(1 - \frac{\alpha}{2}, 0, \hat{\sigma}^2 \Sigma_o, \nu)$ are the $1 - \frac{\alpha}{2}$ quantiles of the $N(0, \hat{\sigma}^2 \Sigma_o)$ and the $t(0, \hat{\sigma}^2 \Sigma_o, \nu)$ distribution.

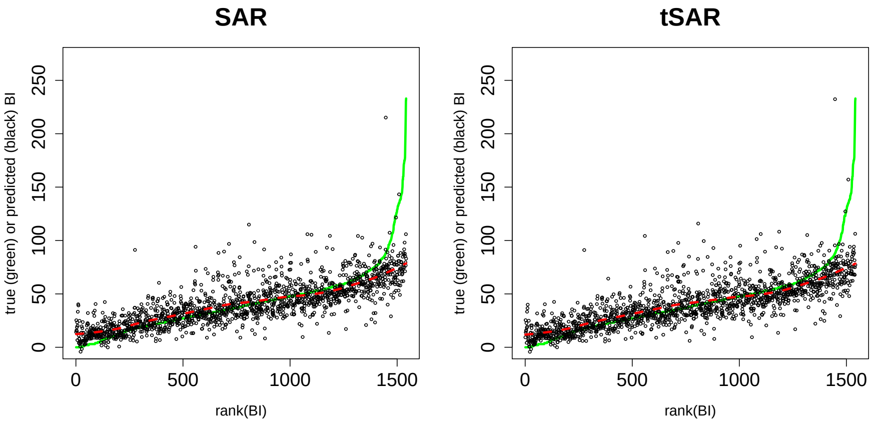

Fig. 5 True vs predicted Burning Index (BI). A smoothed curve for the predicted Burning Index was added in red. For better visualization the Burning Index was ordered.

To perform out of sample prediction, we divide our data set in 10 distinct batches. We use 9 batches for fitting the model and apply the same procedure as before. Our fitted model is the one with the lowest BIC. For the remaining batch data we perform out of sample prediction. Doing this 10 times gives us an out of sample prediction for every location. In every case the fitted model was a tSAR model. For comparison we also take the best SAR model for every case and perform out of sample prediction with this model. The predictions are shown in Figure 5 where we see that there is not a big difference between the SAR and the tSAR model. The prediction is influenced by the estimation of λ and $\boldsymbol{\beta}$ where the SAR and the tSAR model provide similar estimates. The two models differ in the specification of the error distribution which influences confidence intervals. Figure 6 shows the confidence intervals and Table 5 the proportion of data points inside the corresponding confidence interval. We see that, in all three cases of confidence levels, this proportion is closer to the theoretical confidence level for tSAR based confidence intervals. To support this statement we conduct a likelihood ratio test (see Wilks (1938)) for binomial data. We consider a theoretical confidence level of $1-\alpha$. Then we test the null hypothesis that the number of points lying outside the confidence interval is binomial distributed with success probability α against the alternative that it is binomial distributed with a success probability different than α. The results of this test are shown in Table 6. We see that higher p-values are obtained when the tSAR model is used. For the 99% confidence interval the SAR model leads to a very small p-value and the null hypothesis is rejected at the 0.1% level. This can be explained by the fact that the normal distribution is not a good choice to model heavy tailed data.

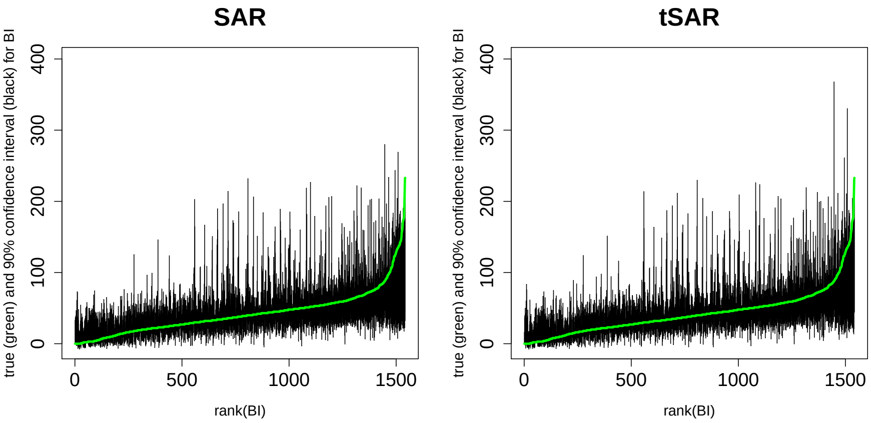

Fig. 6 Burning Index (BI) and its 90% confidence intervals. For better visualization the Burning Index was ordered.

Table 5 Comparison of different confidence intervals. The first column gives the level of the confidence interval. The other two columns show the proportion of data points inside the confidence interval.

	SAR	tSAR
90%	91.05%	90.21%
95%	94.36%	94.55%
99%	97.93%	98.96%

Table 6 Comparison of different confidence intervals. The first column gives the level of the confidence interval. The other two columns show the p-value of the likelihood ratio test.

	SAR	tSAR
90%	0.1622	0.7853
95%	0.2566	0.4265
99%	0.0002	0.8827

6 Outlook

We proposed the tSAR model, an extension of the SAR model for t-distributed errors, which lead to notable improvements in the model fit in our application. The tSAR model showed improvement in the BIC value, its residuals behaved well and it provided more accurate confidence intervals. A natural question which arises is if we can extend the SAR model to other distributions than the t-distribution. Having a closer look at how we approached the tSAR model we can proceed in a similar way for other distributions. We consider the model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \lambda \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X\beta}) + \boldsymbol{\epsilon},$$

where everything except $\boldsymbol{\epsilon}$ is defined as in the SAR model (see Definition 1). We make the more general assumption for the error $\boldsymbol{\epsilon}$ that it has expectation zero, a diagonal variance matrix $\sigma^2\boldsymbol{\Sigma}_\epsilon$ and that $\epsilon_i/(\sigma\sqrt{(\Sigma_\epsilon)_{ii}})$ are identically and independent distributed with density $\phi(\cdot|\boldsymbol{\theta})$, where $\phi(\cdot|\boldsymbol{\theta})$ is the density of a distribution with zero mean, unit variance and parameter vector $\boldsymbol{\theta}$. So one could allow for errors that follow for example a skew-*t* distribution. Note that $\boldsymbol{\theta}$ is empty for location-scale distributions (e.g. the normal distribution). We obtain the density of $\boldsymbol{Y}$ as in Section 3.2 using the density transformation rule as

$$f_Y(\boldsymbol{y}) = |\det(\Sigma_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W}))| \prod_{i=1}^{n} \phi\left(\left(\boldsymbol{\Sigma}_\epsilon^{-\frac{1}{2}}(\mathbf{I}_n - \lambda \boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)_i |0, 1, \boldsymbol{\theta}\right).$$

The regression parameters $\boldsymbol{\beta}$ could be estimated by the generalized least squares estimator and σ^2 as in the SAR model. Then we can form the profile log-likelihood and estimate λ and $\boldsymbol{\theta}$ by numerical optimization. Alternatively one could think about finding estimators of $\boldsymbol{\theta}$ depending on λ such that the dimensionality of the profile log-likelihood can be reduced. It would be interesting to investigate this in more detail for various distributions.

Acknowledgements The first author acknowledges financial support by a research stipend of the Technical University of Munich. The second author is supported by the German Research Foundation through the TUM International Graduate School of Science and Engineering (IGSSE). The third and fourth authors are supported by the German Research Foundation (DFG grants CZ 86/5-1 and CZ 86/4-1). Computations were performed on a Linux cluster supported by DFG grant INST 95/919-1 FUGG.

References

Anselin L, Bera AK (1998) Spatial dependence in linear regression models with an introduction to spatial econometrics. Statistics Textbooks and Monographs 155:237–290

Banerjee S (2005) On geodetic distance computations in spatial modeling. Biometrics 61(2):617–625

Bivand R, Piras G (2015) Comparing implementations of estimation methods for spatial econometrics. Journal of Statistical Software 63(18):1–36, URL http://www.jstatsoft.org/v63/i18/

Box GE, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society Series B (Methodological) pp 211–252

Brent RP (1973) Algorithms for minimization without derivatives. Prentice-Hall, Englewood Cliffs, New Jersey

De Oliveira V, Song JJ (2008) Bayesian analysis of simultaneous autoregressive models. Sankhyā: The Indian Journal of Statistics, Series B (2008-) pp 323–350

Group NWC (2002) Gaining a Basic Understanding of the National Fire Danger Rating System. National Wildfire Coordinating Group

Kariya T, Kurata H (2004) Generalized least squares. John Wiley & Sons

Kotz S, Nadarajah S (2004) Multivariate t-distributions and their applications. Cambridge University Press

Ord K (1975) Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70(349):120–126

Pace RK, Barry R (1997) Sparse spatial autoregressions. Statistics & Probability Letters 33(3):291–297

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data, vol 368. John Wiley & Sons

Whittle P (1954) On stationary processes in the plane. Biometrika pp 434–449

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9(1):60–62