Spatial Statistics 28 (2018) 331–351

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Spatial detrending revisited: Modelling local trend patterns in NO_2-concentration in Belgium and Germany

Svenia Behm [*,1], Harry Haupt [1], Angelika Schmid [1]

University of Passau, Innstraße 27, 94030 Passau, Germany

ARTICLE INFO

Article history:
Received 6 October 2017
Accepted 12 April 2018
Available online 21 April 2018

Keywords:
Stationarity
RIO model
Air pollution
Land use
Nonparametrics

ABSTRACT

Short-term predictions of air pollution require spatial modelling of trends, heterogeneities, and dependencies. Two-step methods allow real-time computations by separating spatial detrending and spatial extrapolation into two steps. Existing methods discuss trend models for specific environments and require specification search. Given more complex environments, specification search gets complicated by potential nonlinearities and heterogeneities. This research embeds a nonparametric trend modelling approach in real-time two-step methods. Form and complexity of trends are allowed to vary across heterogeneous environments. The proposed method avoids ad hoc specifications and potential generated predictor problems in previous contributions. Examining Belgian and German air quality and land use data, local trend patterns are investigated in a data driven way and are compared to results computed with existing methods and variations thereof. An important aspect of our empirical illustration is the heterogeneity and superior performance of local trend patterns for both research regions. The findings suggest that a nonparametric spatial trend modelling approach is a valuable tool for real-time predictions of pollution variables: it avoids specification search, provides useful exploratory insights and reduces computational costs.

* Corresponding author.
E-mail address: svenia.behm@uni-passau.de (S. Behm).

[1] Conflicts of interest: none.

https://doi.org/10.1016/j.spasta.2018.04.004
2211-6753/

1. Introduction

Industrial parks, roads and other sources of fossil fuel combustion processes are responsible for a large share of nitrogen oxides and particulate matters that pollute the air and create severe health risks (Wolf et al., 2017). Information on the location of pollution sources can enhance the identification of local pollution hotspots and trend patterns, even at points where no direct observations are available. Detailed spatial pollution maps have a considerable impact on health policy. An example is the German legislation on banning pollution-intensive cars from cities and its major impact on air pollution (Fensterer et al., 2014).

A well-established source of information for air quality assessment are land use classes. Land use data such as the CORINE land cover inventory encode the usage of a particular territory in land use classes (e.g., Feranec et al., 2016). Frequently, these classes are combined with complementary information on traffic density, demography, topography, and other geographic variables (e.g., Gilliland et al., 2005; Hooyberghs et al., 2006; Sahsuvaroglu et al., 2006; Janssen et al., 2008; Wang et al., 2013; Hennig et al., 2016). A key advantage of land use data is that information on single land use classes can be scaled down when granular data are available, for example on individual exposure to air pollution within a single urban residence (Hennig et al., 2016).

The crucial role of land use information in regression-based models has lead to the notion *Land Use Regression* (LUR). The difference between *using land use indicators in regression* and LUR is that the latter usually relies on the assumption of independence and stationarity of the regression errors (e.g., Gilliland et al., 2005; Ryan and LeMasters, 2007; Hoek et al., 2008). Neglecting such assumptions carries severe potential for ignoring bias and inefficiencies (Montero et al., 2015). Air pollution data are likely to exhibit spatial dependence, because the closer two monitoring sites are located, the more likely they share a common source of pollution or dominant wind direction. There are two main alternatives to combining a regression framework with the modelling of spatial dependencies among individual sites.

(a) In two-step or *residual kriging* methods, a first spatial detrending step allows to filter non-stationarities driven by phenomena such as titration (e.g., Hooyberghs et al., 2006). This is followed by a second (ordinary) kriging step to include the dependence structure in the spatial prediction. Hooyberghs et al. (2006) and Janssen et al. (2008) suggest to use historical data to produce real-time spatial predictions within a two-step *residual interpolation optimised* (RIO) modelling framework. To account for nonstationarities in O_3-concentration across Belgium, Hooyberghs et al. (2006) compute a local spatial trend based on historical measurements using population density as auxiliary data. Janssen et al. (2008) use CORINE land use data instead of population density data in the detrending step and analyse the three pollutants NO_2, O_3, and PM_{10}. The RIO residual kriging procedure has two advantages: First, trend and semivariogram estimation can be done in two separated steps. Second, as long as the crucial assumption of stable spatial trend and semivariogram over time holds, it allows real-time predictions at basically zero computational cost.

(b) Alternatively, *universal kriging* is a one-step method, where the spatial dependence structure and the impacts of the predictors are estimated simultaneously. However, the difference between two-step methods and universal kriging is not always clear-cut (e.g., Mercer et al., 2011), and the latter can also be applied to filtered data. As Montero et al. (2015) point out, splitting up detrending and kriging in two steps is a recommended alternative to avoid ambiguities in universal kriging with regard to the interplay of trend specification and semivariogram estimation. While a correct trend specification is important in both methods to fulfil the requirements for kriging, it remains unclear how to specify the relationship between predictors and pollution with regard to optimising predictive performance.

Two-step methods provide a simple and useful tool for real-time predictions. Their key assumption seems to hold, as average pollution levels are quite stable over time and independent of short term influences, for example over different seasons (e.g., Sahsuvaroglu et al., 2006), or over the span of several years (e.g., Wang et al., 2013). Our work aims at providing further insights into two-step methods such as the RIO residual kriging method, and generalises the method of Janssen et al. (2008) theoretically and empirically. The quality of the trend filter in the first step is crucial for any inferences drawn from the second step. Hence we suggest nonparametric generalisations to adapt the trend

modelling step to general environments, exhibiting different degrees of complexity and heterogeneity in spatial patterns. In particular we suggest to simplify the inclusion of land use classes in the trend estimation step.

In Janssen et al. (2008), every monitoring site is assigned a pollutant-specific land use indicator that describes average pollution based on the relative share of every land use class within the sites' vicinity. This indicator summarises the interplay of constant local characteristics contained in the predictors and is interpreted as a proxy for the long-term total pollution load a single location has to carry. The authors assume that mean and standard deviation of the pollutant can be described by polynomials in the indicator. They do not consider additional predictors controlling for further sources of heterogeneity in spatial trend patterns. To avoid the consequences of misspecifying the trends, we propose to use nonparametric trend regressions. Nonparametrics allow for a data-driven exploration of trend patterns while avoiding specification search based on ad hoc polynomials (and interactions if further predictors are used). We show that multivariate generalisations of the trend functions can be easily accomplished by allowing for different trends for background, industrial and traffic environments.

The simultaneous estimation of a trend function and a pollutant-specific land use indicator (weighting single land use classes) in prediction employed by Janssen et al. (2008) leads to a generated predictor problem. Hence we propose direct inclusion of the information on land use classes as predictors in our trend function. We thoroughly discuss estimation, prediction and comprehensive empirical evidence for Belgian and German air quality and land use data. Our empirical analysis reproduces existing results of Janssen et al. (2008) for Belgium and provides evidence for Belgium and Germany that the suggested modifications perform very well.

The remainder of this article is organised as follows: Section 2 discusses the database used for our empirical investigation. Section 3 explains the statistical theory, including an overview on Janssen et al. (2008) and indicator-based two-step spatial prediction methods. Section 4 provides detailed insights into our results and Section 5 concludes.

2. Data

In the application to German air pollution, we investigate daily maxima of the recorded hourly NO_2-concentration over the time period 1st Jan 2007 to 31st Dec 2012. The data have been obtained from the European Environment Agency (EEA), who maintains AirBase, the European air quality database ([dataset] EEA, European Environment Agency, 2016). The database consists of monitoring data from fixed monitoring sites, measured at regular intervals, as well as meta-information on the monitoring sites involved. One meta-information is the sites' type that can either be "Background", "Industrial", or "Traffic". For a complete description of the meta-data on monitoring site characteristics, we refer to Appendix B.1. Further, we use the CORINE Land Cover 2006 (CLC2006) data layer in a 100 $\times$ 100 m resolution ([dataset] EEA, European Environment Agency, 2010b). For detailed information on CLC data including changes between the four different data layers CLC1990, CLC2000, CLC2006, CLC2012, see Feranec et al. (2016).

In order to make our empirical findings comparable to those of Janssen et al. (2008), we also analyse Belgian hourly NO_2-concentration from AirBase over the time period 1st Jan 2001 to 31st Dec 2006, and the CLC2000 layer, i.e. land use classification in the year 2000 version ([dataset] EEA, European Environment Agency, 2010a). Table 1 shows that German data contain a considerably higher number of monitoring sites and exhibit a quite different distribution over measuring sites' types in comparison to Belgium. While both countries have an equivalent share of background sites, the relative shares of industrial and traffic sites are inverted.

In our analysis we omit daily maximum NO_2 values above 500 $\mu g/m^3$ as well as negative values. Based on the remaining daily maximum values the mean and standard deviation of each monitoring site is calculated, separately for weekdays and weekends. For supplementary information about the data quality of the German and Belgian air pollution data and the data preprocessing we refer to Appendix B.2. Fig. 1 displays the respective boxplots for Belgium and Germany. While the four statistics (mean weekday, mean weekend, st.dev. weekday, st.dev. weekend) for Belgium and Germany differ only slightly with respect to their medians, the interquartile ranges and the ranges

Table 1
Numbers of monitoring sites in Belgium (Germany) that were active within the period 1st Jan 2001 to 31st Dec 2006 (1st Jan 2007 to 31st Dec 2012).

	Background	Industrial	Traffic	Total
Belgium	37 (52.85%)	23 (32.86%)	10 (14.29%)	70
Germany	276 (51.49%)	38 (7.09%)	222 (41.42%)	536

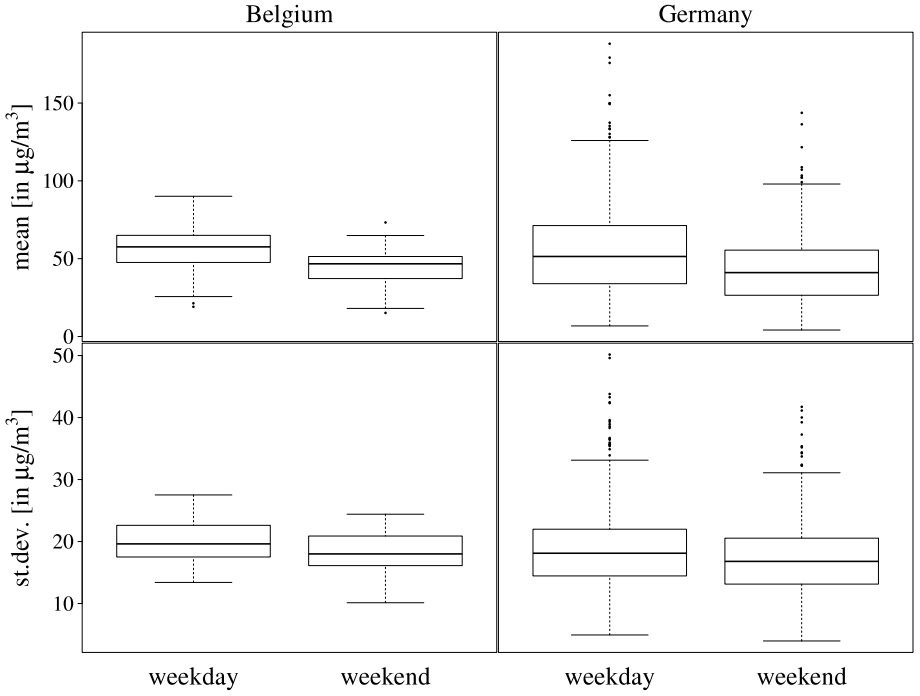

Fig. 1. Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian monitoring site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

between the whiskers are remarkably higher for the German data compared to Belgian data. For both research regions we observe differences between the mean of daily maximum NO_2 concentrations on weekdays and weekends. For the standard deviation of daily maximum NO_2 concentrations only a small difference between weekdays and weekends occurs. In Figs. A.1–A.3 we explore the distribution of the means and standard deviations differentiating by the sites' type. We find that observed differences between Belgium and German data can be traced back to measurements at traffic sites.

Considering the usage of the CLC data in air pollution studies, it is common practice to reclassify the 44 land use classes in the CLC inventory (e.g. Beelen et al., 2009, 2013; Wolf et al., 2017). Following the suggestion of Janssen et al. (2008), we group the 44 classes into eleven more general land use classes. The European Monitoring and Evaluation Programme (EMEP) provides emission data concerning national total, sector and gridded emissions for Europe (see [dataset] EMEP and CEIP, 2014, for detailed information). Those data are classified with regard to their relationship to air pollution, and the classification results in so-called sectors, referred to as SNAP (Selected Nomenclature for reporting of Air Pollutants). Table 2 summarises the resulting classifications and descriptions.

The empirical analysis is conducted with the statistical software R (R Core Team, 2013) using the packages broom (Robinson, 2017), GISTools (Brunsdon and Chen, 2014), gstat (Pebesma, 2004; Gräler

Table 2
Relationship between grouped CLC classes and the equivalent groups in the SNAP sector classification (according to Janssen et al., 2008).

Grouped class	Description	CLC classes	SNAP sectors
Class 1	Continuous urban fabric	1	S2
Class 2	Discontinuous urban fabric, green and sport	2,10,11	S2
Class 3	Industrial or commercial units	3	S3 + S4
Class 4	Road and rail networks and associated land	4	S7
Class 5	Port areas	5	S8
Class 6	Airports	6	S8
Class 7	Mine, dump and construction sites	7–9	S1 + S4 + S5 + S9
Class 8	Arable land	12–14	S10
Class 9	Agricultural areas	15–22	S10
Class 10	Forest and semi natural areas	23–34	S11
Class 11	Wetlands and water bodies	35–44	S11

et al., 2016), np (Hayfield and Racine, 2008), optimx (Nash and Varadhan, 2011; Nash, 2014), raster (Hijmans, 2016), rgdal (Bivand et al., 2017), spatstat (Baddeley et al., 2015), and timeDate (Rmetrics Core Team et al., 2015).

3. Statistical modelling

Assume air pollution at time $t \in D_t$ to be a latent geostatistical random process

$$Y_t(\cdot) = \{Y_t(\boldsymbol{s}) : \boldsymbol{s} \in D_s \subset \mathbb{R}^2\},$$

where D_s refers to the study area. Within the study region D_s define the locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, $n \in \mathbb{N}$. Let $Z_t(\boldsymbol{s})$, where

$$Z_t(\cdot) = \{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in D_s\},$$

denote the data process at time $t \in D_t$. In our computations below let $z_{i,t}$ denote a realisation of $Z_t(\boldsymbol{s}_i)$ at location $\boldsymbol{s}_i$ at time $t \in D_t$. The vector $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,T})$, $i \in 1, \ldots, N$, defines the time series at monitoring site i, the vector $\boldsymbol{z}_t = (z_{1,t}, \ldots, z_{n,t})$, $t \in \{1, \ldots, T\}$, defines measurements for all cross-sectional units or monitoring sites recorded at time t.

Following Cressie (1993) and Diggle and Ribeiro Jr (2007), the relationship between the unobserved geostatistical process and the data process is given by

$$Z_t(\boldsymbol{s}) = Y_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}) \tag{1}$$

with $\epsilon_t(\boldsymbol{s}) \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. If the unobserved geostatistical process $Y_t(\cdot)$ at time $t \in D_t$ is assumed to be a stationary and isotropic Gaussian process, it holds $\forall \boldsymbol{s}, \boldsymbol{s}' \in D_s, \boldsymbol{s} \neq \boldsymbol{s}'$,

$$E[Y_t(\boldsymbol{s})] = \mu, \tag{2a}$$

$$Var[Y_t(\boldsymbol{s})] = \sigma^2, \tag{2b}$$

$$C(h) = Cov[Y_t(\boldsymbol{s}), Y_t(\boldsymbol{s}')] = \sigma^2 \rho(h), \tag{2c}$$

where the autocorrelation function $\rho(h) = Corr[Y_t(\boldsymbol{s}), Y_t(\boldsymbol{s}')]$ depends on the distance $h = \|\boldsymbol{s} - \boldsymbol{s}'\|$, $E[\cdot]$ denotes the expected value, $Var[\cdot]$ the variance, and $C(\cdot)$ the autocovariance function. Under the assumptions stated above, analogous stationarity conditions hold for the data process $Z_t(\cdot)$, and the ordinary kriging predictor $\hat{Y}_t(\boldsymbol{s}_0)$ can be calculated for any $\boldsymbol{s}_0 \in D_s$, $t \in D_t$.

3.1. Spatial trend modelling: parametric polynomials

The RIO technique proposed by Hooyberghs et al. (2006) and Janssen et al. (2008) starts with a detrending step in order to filter the data process $Z_t(\cdot)$ such that stationarity conditions analogous to

(2a)–(2c) hold. The grouped land use classes (see Table 2) enter the equation for the pollutant specific β-index according to

$$\beta(\boldsymbol{s}, r) = log\left[1 + \sum_{k=1}^{11} a_k \cdot sh_k(\boldsymbol{s}, r)\right], \tag{3}$$

where $sh_k(\boldsymbol{s}, r)$ describes the share of the kth class within a circular buffer zone with radius r around location $\boldsymbol{s}$. For the sake of simplicity we omit r and $\boldsymbol{s}$ and write β_i for $\beta(\boldsymbol{s}_i, r)$, β for $\beta(\boldsymbol{s}, r)$ and sh_k for $sh_k(\boldsymbol{s}, r)$. The class weights $a_k,\ k = 1, \ldots, 11$, define the relative impact of the respective class on the concentration of the air pollutant under investigation. Eq. (3) shows how the relative contribution of every land use class is summed up to an overall indicator. This means that a certain share of roads can be equivalent to a certain share of industrialised area, or a larger share of residential area (as the latter are usually relatively small sources of air pollution). Further details on the class weights are given in Table A.1 in the Appendices A and B.

Janssen et al. (2008) assume that spatial trends of mean and standard deviation are functions of the pollutant specific β-index. For the sake of a more general exposition covering the extensions in Section 3.2, we consider trend functions including potential further predictors X,

$$\mu \approx m_\mu(\beta, X), \tag{4a}$$

$$\sigma \approx m_\sigma(\beta, X). \tag{4b}$$

In their application to Belgian data, Janssen et al. (2008) assume that mean and standard deviation in Eqs. (4a) and (4b) can be described by a second and first order polynomial of β, respectively, and do not consider additional predictors X. The functions m_μ and m_σ are estimated in regressions using estimates $\bar{z}$ and s of μ and σ, respectively, based on the time series observed for each measuring site where a distinction is made between weekdays and weekends. For the sake of simplicity we omit further notation.

For both statistics, β is calculated via Eq. (3) and therefore depends on $\boldsymbol{s}$ and $a_1, \ldots, a_{11}$. Under assumption (4a) the coefficients $a_1, \ldots, a_{11}$ in Eq. (3) are optimised through the following numerical optimisation procedure, after defining suitable termination criteria

1. Specify a starting set $a_1^{(1)}, \ldots, a_{11}^{(1)}$ of $a_1, \ldots, a_{11}$ (see Janssen et al., 2008).
2. Regress $\bar{z}_i$ on $m_\mu(\beta_i^{(1)}, X_i)$ where $\beta^{(1)}$ is computed using the set $a_1^{(1)}, \ldots, a_{11}^{(1)}$, and obtain the predictor $\widehat{m_\mu}^{(1)}(\beta_i^{(1)}, X_i)$.
3. Calculate the value of the RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{m_\mu}^{(1)}(\beta_i^{(1)}, X_i) - \bar{z}_i)^2}$.
4. If none of the termination criteria is fulfilled, restart the procedure with a different set $a_1^{(2)}, \ldots, a_{11}^{(2)}$, otherwise the optimal set is found.

Denoting the optimised class weights by $\tilde{a}_1, \ldots, \tilde{a}_{11}$ and the corresponding β-index by $\tilde{\beta}_1, \ldots, \tilde{\beta}_n$, the trend functions for mean and standard deviation can be computed, for every i, as $\hat{\mu}_i = \widehat{m_\mu}(\tilde{\beta}_i, X_i)$ and $\hat{\sigma}_i = \widehat{m_\sigma}(\tilde{\beta}_i, X_i)$, respectively.

According to Janssen et al. (2008), using the fitted values $\hat{\mu}_i$ and $\hat{\sigma}_i$, and given pre-defined reference levels μ^{ref} and σ^{ref}, detrending of the measurement values $z_{i,t}$ can be achieved according to

$$z_{i,t}^{*} = z_{i,t} + (\mu^{ref} - \hat{\mu}_i), \tag{5a}$$

$$z_{i,t}^{**} = (z_{i,t}^{*} - \bar{z}_i^{*})\frac{\sigma^{ref}}{\hat{\sigma}_i} + \bar{z}_i^{*}. \tag{5b}$$

After filtering the monitored data $z_{i,t}$ according to Eqs. (5a) and (5b), we obtain the transformed data $z_{i,t}^{**}$, which we interpret as realisations of $Z_t^{**}(\boldsymbol{s}_i)$, the filtered data process at time $t \in D_t$. Hence, for each $\boldsymbol{s} \in D_s$,

$$E[Z_t^{**}(\boldsymbol{s})] = \mu(\boldsymbol{s}) + (\mu^{ref} - \hat{\mu}(\boldsymbol{s})) \approx \mu^{ref}, \tag{6}$$

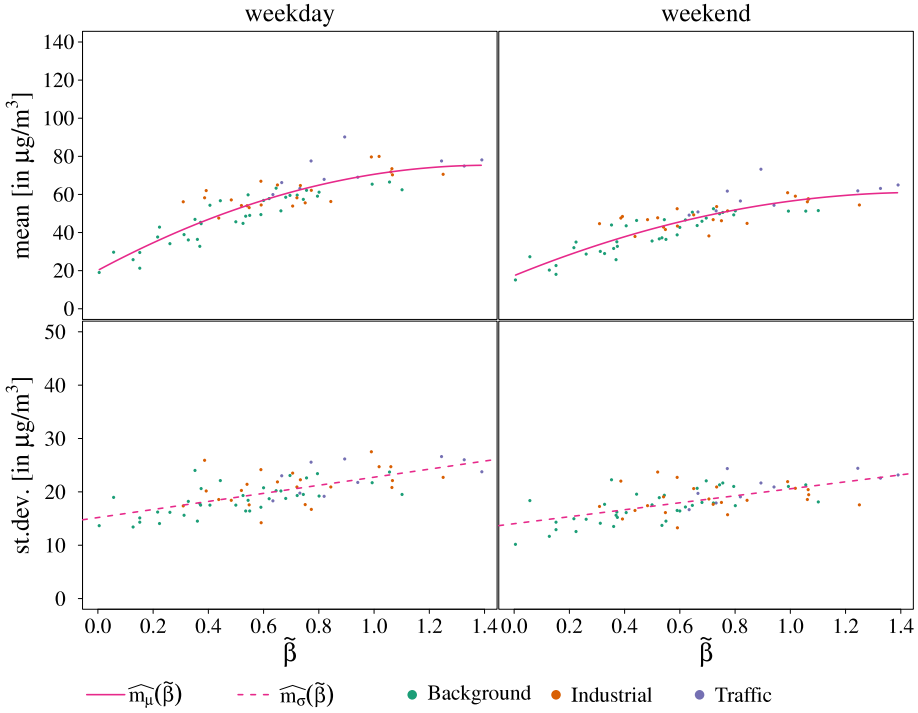

Fig. 2. Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

relying on assumption (4a) in the last transformation, and

$$Var[Z_t^{**}(\boldsymbol{s})] = \left(\sigma^{ref}\right)^2 Var\left[\frac{Z_t^{*}(\boldsymbol{s}) - \bar{Z}^{*}(\boldsymbol{s})}{\hat{\sigma}(\boldsymbol{s})}\right] \approx \left(\sigma^{ref}\right)^2, \tag{7}$$

since the middle term describes the standardisation of $Z_t^*(\boldsymbol{s})$. Eqs. (6) and (7) show that the filtered data process approximately satisfies the (weak) stationarity properties (2a)–(2c) and can be used in the kriging procedure.

Based on all historical detrended measurements $z_{i,t}^{**}$, the semivariogram required for ordinary kriging is estimated. For any $\boldsymbol{s}_0 \in D_s$ at time $t \in D_t$ an interpolated value $\hat{Y}_t^{**}(\boldsymbol{s}_0)$ can be calculated and retrended with regard to the local mean and local standard deviation of the originally monitored process. The retrending formulas can be written as

$$\hat{Y}_t^{*}(\boldsymbol{s}_0) = (\hat{Y}_t^{**}(\boldsymbol{s}_0) - \bar{\hat{Y}}^{**}(\boldsymbol{s}_0))\frac{\hat{\sigma}(\boldsymbol{s}_0)}{\sigma^{ref}} + \bar{\hat{Y}}^{**}(\boldsymbol{s}_0), \tag{8a}$$

$$\hat{Y}_t(\boldsymbol{s}_0) = \hat{Y}_t^{*}(\boldsymbol{s}_0) - (\mu^{ref} - \hat{\mu}(\boldsymbol{s}_0)). \tag{8b}$$

The RIO technique rests on the crucial assumption that both spatial trends and the semivariogram are stable over time, enabling real-time predictions at basically zero computational cost. Real-time predictions are produced in the following way: detrend a new set of observations (at monitoring sites) using the fitted trend functions, interpolate the detrended values using the fitted semivariogram and retrend the interpolated values using the fitted trend functions.

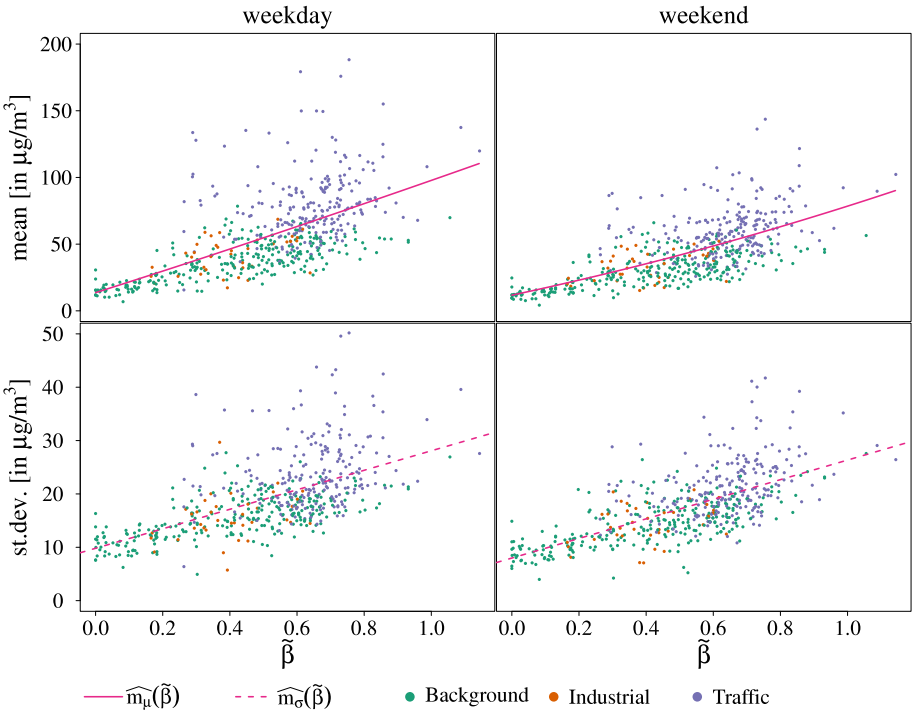

Fig. 3. German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

3.2. Spatial trend modelling: a general nonparametric approach

There are several options to include further predictors in Eqs. (4a) and (4b). In general, the functions m_μ and m_σ can be approximated by higher-order parametric expansions using polynomials of β interacting with (the levels of) X. Such a strategy, however, requires assumptions on the degree of the approximation and a high number of parameters. In order to avoid ad hoc assumptions, potential underspecification, or potentially extensive specification search, a straightforward alternative is to estimate m_μ and m_σ using a nonparametric trend model. Such a model should deliver a more accurate representation of the trend patterns than a specification based on a parametric expansion if the latter is underspecified and the data are sufficiently informative for nonparametric regression (e.g., Haupt et al., 2010). More important and evident from our empirical illustration, nonparametric methods provide explorative insights about the trend patterns driven by β and potential further predictors such as the type of monitoring sites X.

Hence, as nonparametric methods can help to identify the best parametric approximation and to avoid problems of misspecifying the trend functions, we employ a local linear kernel smoothing estimator of $E(\bar{Z}|\beta, X) = m(\beta, X)$ in the trend regression model

$$\bar{Z} = m(\beta, X) + U \qquad \text{with } E(U|\beta, X) = 0, \tag{9}$$

based on Eq. (4a). A generalised least squares estimator is denoted as $\widehat{m}_{LL}$, where $(\widehat{m}_{LL}, \widehat{\gamma})$ minimises

$$\sum_{i=1}^{n} [\bar{Z}_i - m - \gamma(\beta_i - \beta)]^2 K(\mathbf{W}, \mathbf{W}_i, \mathbf{h}),$$

where $\mathbf{W} = (\beta, X)$ denotes the vector of regressors, $K = k_\beta \cdot k_X$ is a product kernel, and $\mathbf{h} = (h_\beta, h_X)'$ is a vector of bandwidths which we estimate using least squares cross validation (see Li and Racine,

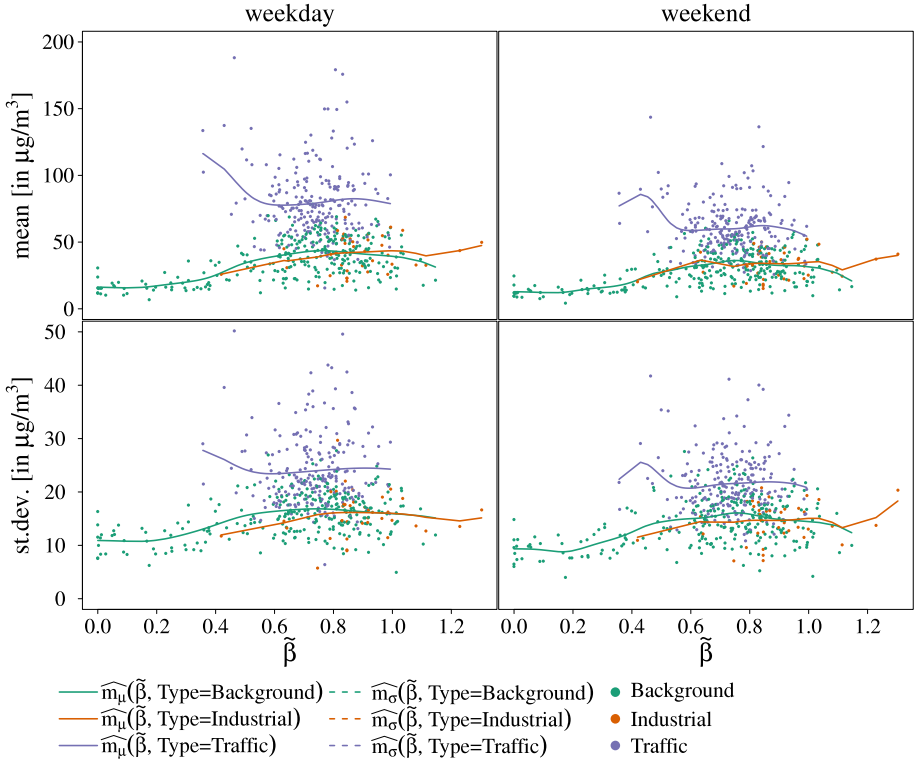

Fig. 4. German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

2004). The use of *mixed* continuous (i.e. pollutant specific β-index) and categorical (i.e. type of monitoring site X) predictors in nonparametric regressions has been discussed extensively in the various works of Li and Racine (e.g., Li and Racine, 2007).

The β-index in Eq. (9) is unknown and has to be computed according to the procedure described in Section 3.1. Hence the estimated β-index $\tilde{\beta}$ is a generated predictor. The potential consequences for estimation and inference in parametric models have been discussed in an abundant literature following the seminal paper of Pagan (1984). In a nonparametric context Sperlich (2009) and Mammen et al. (2012) provide authoritative treatments (see Haupt et al., 2018, for a discussion in the mixed predictor context). Depending on the problem at hand, researchers may prefer to use an aggregated index, but should be aware that generated regressor problems may invalidate the interpretation of the β-index. In the current context the problems can be avoided from the outset if the β-index is not considered. We propose to directly include the information on land use classes and define the categorical predictors

$$X_1 = \underset{k\in\{1,\ldots,11\}}{\operatorname{argmax}} sh_k, \tag{10}$$

$$X_2 = \underset{k\in\{1,\ldots,11\}\setminus X_1}{\operatorname{argmax}} sh_k, \tag{11}$$

determining which classes have the largest and second largest share (within the circular buffer zone around a certain location), respectively. Note that including the third largest class has no remarkable effect. In our application for 534 of 536 German sites and for 69 of 70 Belgian sites, the sum of the shares of the first and second largest class is larger than 50%. The continuous predictor

$$S = sh_{X_1} + sh_{X_2}. \tag{12}$$

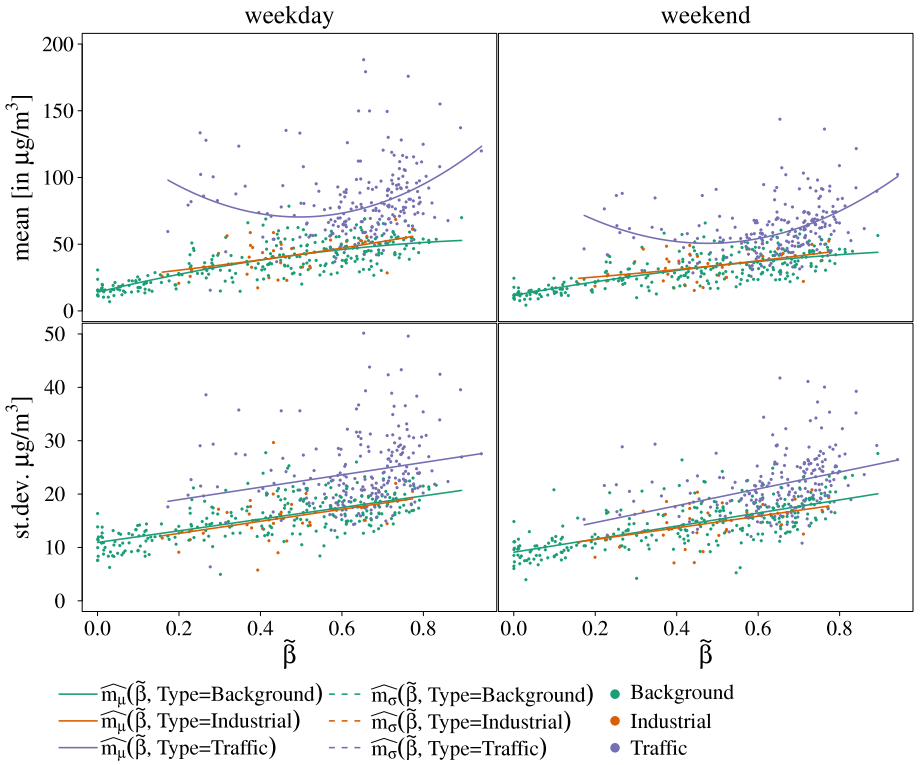

Fig. 5. German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

is defined as the sum of the shares of the first and second largest class. Then, instead of the predictors used in Eq. (9), we consider the categorical predictors X_1, X_2, and the sites' type, and the continuous predictor S.

4. Results

For the sake of exposition, we introduce the following abbreviations: "QL" ("LL") refers to a quadratic (linear) trend for the mean and a linear trend for the standard deviation; "TypeQL" ("TypeLL") allows local trend differing with respect to the type of a monitoring site ("Background", "Industrial", or "Traffic"); "NP" refers to the nonparametric approach with $\beta-$index together with the sites' type; "NPnoBeta" refers to the nonparametric approach without $\beta-$index but with the predictors defined in Eqs. (10)–(12) together with the sites' type.

The estimated QL trend functions for Belgian data are displayed in Fig. 2 and replicate results of Janssen et al. (2008, right plot of Fig. 5 and middle plot of Fig. 8), while Fig. 3 shows the corresponding QL estimates using German data. A global second order polynomial fits the Belgian data quite well, while we observe considerably more heterogeneity in the German data. The curvature is less pronounced in the plots of Fig. 3 and bears no visible difference to the LL trend functions displayed for Germany in Fig. A.8 in Appendix A.

Comparing the trend functions for weekdays and weekends for Belgian as well as for German data, we observe a shift along the y-axis (for both specifications LL and QL). This is in accordance with the boxplots displayed in Fig. 1 and indicates that, on average, the concentration level of NO_2 drops from weekdays to weekends.

The replication of the results of Janssen et al. (2008) in a narrow sense for Belgian data and in a wider sense for German data suggests that the assumption of global trend forms is too restrictive. Determining a global trend form requires an ad hoc specification of polynomial degree and specification search. Previous contributions such as Janssen et al. (2008) do not explicitly discuss this issue. The optimisation of the class weights a_k affects the values of $\tilde{\beta}$, the position of the points along the x-axis and thus the fitted trend function (e.g., compare the range of $\tilde{\beta}$ in Figs. 3 and A.8). To avoid ad hoc specification search and to widen the scope of applicability to heterogeneous environments, we discuss a more general approach to spatial trend fitting and illustrate it with German data. Note that further results for Belgium, completing our empirical analysis, are provided in Figs. A.4–A.7 in Appendix A.

An encompassing approach to trend analysis is the nonparametric regression, following the mixed kernel estimation approach for continuous and categorical predictors of Li and Racine (2004, 2007), compare Eq. (9). Fig. 4 shows estimated NP trend functions for German data based on local linear kernel regressions, where bandwidths are estimated by least squares cross-validation using the default kernel functions proposed by Hayfield and Racine (2008). Trends are calculated by simultaneously smoothing over $\tilde{\beta}$ and the three categories of the sites' type contained in X. We observe substantial differences in local levels and slopes between traffic sites and all other sites indicating that the NO_2 concentration at traffic sites is on average larger than at background or industrial sites. Apart from minor boundary effects visible in the plots for weekend data, the estimates suggest that a piecewise quadratic trend may be sufficiently flexible. The finding of heterogeneity in local trend patterns in Germany based on our visual analysis is confirmed by the quantitative results from the nonparametric approach including the sites' type. The corresponding results on predictive performance are discussed in detail below.

Based on the exploratory insights obtained from the nonparametric regressions, we add dummy variables and interactions as indicators for the monitoring sites' type to the specification QL. The resulting TypeQL trend estimates are shown in Fig. 5. Visual inspection of the results and comparison to Fig. 3 suggest that the specification TypeQL allowing local quadratic trend patterns provides a superior fit to the German data. Again, this finding is supported by an analysis of predictive performance. Equivalent plots for specifications LL and TypeLL for Germany are provided in Figs. A.8 and A.9 in Appendix A.

The trend functions corresponding to the specifications TypeQL and TypeLL reveal substantial differences in local levels and slopes between traffic sites and all other sites in Germany. For Belgian data such clear differences cannot be observed (see Figs. A.6 and A.7 in Appendix A).

For specification NPnoBeta trends are calculated by simultaneously smoothing over S and the categories X_1, X_2, and the sites' type. This specification entails considerably lower computational costs compared to those of NP, as the optimisation of group weights is not required. For German (Belgian) data computation time equals 3.45 h (7 min) to derive the trend functions using NP, compared to 16 s (4.3 s) for NPnoBeta. For NPnoBeta it is not possible to display the estimated trend functions in two-dimensional space, as they depend on one continuous and three unordered categorical predictor variables. In order to evaluate the predictive performance of NPnoBeta compared to the approaches including the β-index, we carry out a leave-one-out cross-validation (LOOCV). In each loop of LOOCV one monitoring site is omitted and the entire RIO technique – consisting of the four steps of optimising group weights, detrending, kriging and retrending (as described in Section 3) – is applied to the remaining sites. For NPnoBeta the optimisation of group weights is no longer necessary and therefore each loop of LOOCV consists of the steps detrending, kriging and retrending. Table 3 summarises the results of LOOCV. As suggested by our visual inspection of the nonparametric trend estimates, allowing the trend functions to differ with the sites' type enhances the predictive performance. Adding an indicator for the sites' type to specifications QL (LL) leads to a performance gain of 13.7% (12.5%) with regard to RMSE for Germany. For Belgium, it lowers the RMSE by 2.0% when the indicator is added to LL, and increases the RMSE by 14.0% when the indicator is added to QL. The latter deterioration of predictive performance in Belgium is due to a single outlier produced in the optimisation process. Avoiding the generated predictor problem by including the information on land use classes directly in NPnoBeta improves (reduces) the predictive performance by 3.4% (1.6%) for German (Belgian) data compared to NP. Table A.2 in Appendix A provides further and more detailed results on our LOOCV

Table 3
Results of LOOCV for different specifications and their predictive performance.

RMSE	QL	LL	TypeQL	TypeLL	NP	NPnoBeta
Germany	20.84	20.82	17.99	18.21	19.07	18.43
Belgium	13.76	13.79	15.69	13.51	13.66	13.88

analysis, revealing that the inclusion of the third largest LUC class has no remarkable effect on the predictive performance with regard to RMSE. Overall we observe that NPnoBeta has a superior (equal) LOOCV performance for Germany (Belgium) while it does not require specification search, avoids generated predictor problems and causes almost zero computational costs.

5. Discussion and conclusions

Approaches for spatial interpolation of air pollutant data require assumptions on stationarity or on trend patterns of the underlying geostatistical random processes. Step-wise procedures based on filtering known or estimated spatial trends bear the advantage of real-time applicability due to their computational and interpretational simplicity. The RIO framework of Hooyberghs et al. (2006) and Janssen et al. (2008) enhances spatial interpolation and predictive performance by exploiting pollution relevant information from local land use patterns. The general applicability of the method hinges on assumptions about ad hoc global trend patterns defined by land use related pollution indicators. Existing methods discuss trend models for specific environments and require specification search. In practice, however, research environments of different size and level of aggregation may exhibit complex nonlinear local trend patterns, driven by spatial heterogeneities and dependencies. Specification search then becomes a troublesome endeavour.

Based on the spatial detrending employed by Janssen et al. (2008), we propose the use of a simple flexible framework for data driven trend modelling and subsequent filtering of the data. A crucial assumption is the selection of further predictors driving the spatial complexity of trend patterns. The various types of monitoring sites are an obvious initial choice for such a predictor. This approach has the advantage of preserving the intuition of larger values of the land use indicator β representing higher local – that is type-specific – levels of pollution, while allowing for type-specific trend levels and slopes.

We propose a nonparametric spatial trend modelling approach using all available predictors. The approach is computationally feasible and does not require ad hoc assumptions on the functional form. It can be used in an exploratory way to identify potential parametric approximations of trend generating mechanisms. In addition, we propose to avoid potential generated predictor problems. This can be done by directly including the information on land use classes, instead of computing a pollution-specific indicator. The performance of the proposed method, existing methods, and variants thereof can be studied by using leave-one-out cross-validation analysis of the predictive performance.

We find that a simple generalisation of the existing methods by using multiple nonparametric regression methods leads to considerable gains in predictive performance while computational costs remain low. Furthermore, the proposed method bears a large potential for exploratory analysis of trending mechanisms while avoiding lengthy trend specification search.

In an empirical study, we first successfully replicate existing results of Janssen et al. (2008) for Belgium using similar but not the same data, and then apply the proposed method to German data. We investigate the assumption of global trend patterns and find strong (weak) evidence against such an assumption for German (Belgian) data. The nonparametric approach can be used to identify local parametric approximations of trend patterns. The overall performance of the proposed method suggests that the nonparametric method is a very good choice for research environments with considerably different complexity. Obvious advantages are that it does not require specification search, avoids generated predictor problems and has almost zero computational costs.

Potential extensions can be considered in several directions. First, it should be kept in mind that the β-values change simultaneously with the functional form, and hence a monotonicity restriction is necessary to preserve the intuition of β as an index representing mean pollution. A non-monotonic

functional form resulting from polynomial or nonparametric trend fits stresses plausibility of this theoretical rationale. The question of imposing monotonicity constraints or not depends on the problem at hand; i.e. whether predictive performance or interpretability is the main objective. Second, statistical tools could be used to provide live monitoring of the crucial assumption of stable trend functions for mean and standard deviation over time. Third, the robustness of the results could be assessed with regard to the choice and aggregation of land use categories as well as the choice of variables determining the trend forms. Fourth, further diagnostics could refer to the uncertainty arising from the stepwise nature of the analysis. There is no clear indication in the original application on how to calculate the uncertainty arising from errors due to trend elimination and kriging, as well as their potential dependence structure.

A flexible two-step procedure reduces the computational demand for spatial now- and forecasts and allows researchers to explore and test suitable trend specifications. The approach is transparent in its single steps and sufficiently general for a wide range of applications.

Acknowledgements

We thank J. Schnurbus, T. Szentimrey, participants of the 4th conference on Spatial Statistics, Lancaster 2017, and an anonymous reviewer for helpful suggestions. All errors are ours.

Appendix A. Tables and figures

Table A.1
Optimised class weights. Following Janssen et al. (2008), class weights a_2, a_{10} and a_{11} are set to 1, 0 and 0, respectively. Therefore the optimisation procedure returns optimal values for the other eight class weights.

Germany	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
QL	2.77	1.00	0.92	0.73	0.71	0.09	0.34	0.10	0.36	0.00	0.00
LL	2.96	1.00	0.92	0.80	0.67	0.08	0.31	0.10	0.35	0.00	0.00
TypeQL	1.76	1.00	1.39	2.09	1.52	1.47	0.91	0.12	0.13	0.00	0.00
TypeLL	3.53	1.00	1.77	3.65	2.21	2.31	1.25	0.33	0.47	0.00	0.00
NP	0.05	1.00	2.07	5.07	1.16	1.17	4.25	2.21	0.81	0.00	0.00
Belgium											
QL	3.49	1.00	1.49	6.00	2.75	1.38	1.73	0.35	0.00	0.00	0.00
LL	1.62	1.00	1.63	3.65	2.10	1.30	1.80	0.40	0.00	0.00	0.00
TypeQL	0.83	1.00	0.96	2.42	1.65	0.95	1.11	0.27	0.00	0.00	0.00
TypeLL	0.89	1.00	1.09	3.16	1.91	0.91	0.13	0.36	0.00	0.00	0.00
NP	0.98	1.00	2.61	6.02	1.10	1.12	3.78	0.75	0.63	0.00	0.00

Table A.2
Results of LOOCV for different specifications and their predictive performance with regard to RMSE.

Germany	QL	LL	TypeQL	TypeLL	NP	NPnoBeta[a]	NPnoBeta[b]
Background	16.70	16.70	12.79	12.93	14.18	13.16	13.18
Industrial	14.52	14.46	13.10	13.25	14.25	15.29	15.45
Traffic	27.06	27.02	25.30	25.63	25.98	25.52	25.16
Overall	20.84	20.82	17.99	18.21	19.07	18.43	18.31
Belgium							
Background	13.16	13.02	12.88	12.80	13.40	13.33	13.57
Industrial	14.63	14.92	14.33	14.09	13.99	14.86	15.10
Traffic	14.02	14.04	29.19	14.81	13.84	13.69	14.35
Overall	13.76	13.79	15.69	13.51	13.66	13.88	14.19

[a] With first and second largest LUC.

[b] With first, second and third largest LUC.

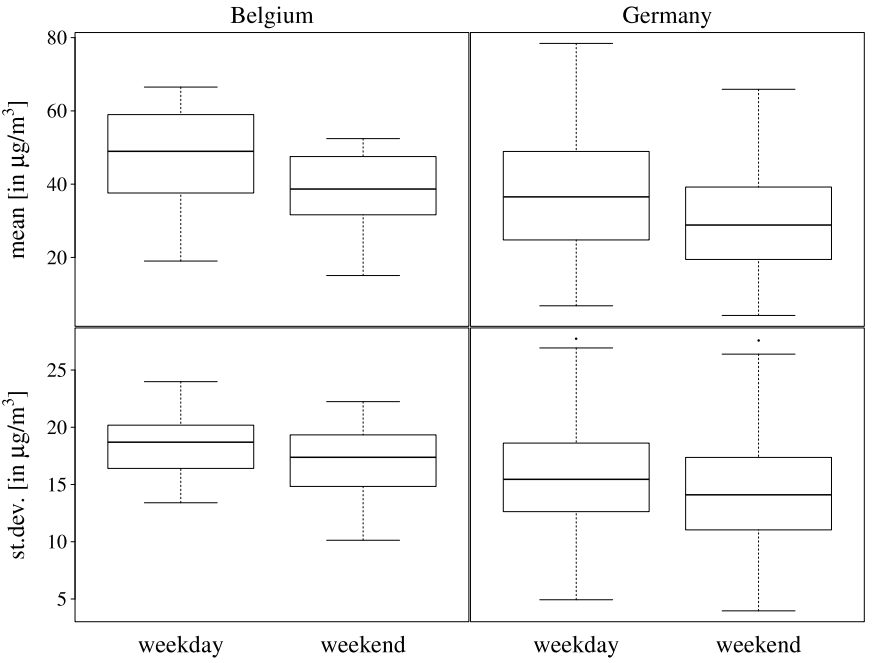

Fig. A.1. Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian background site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

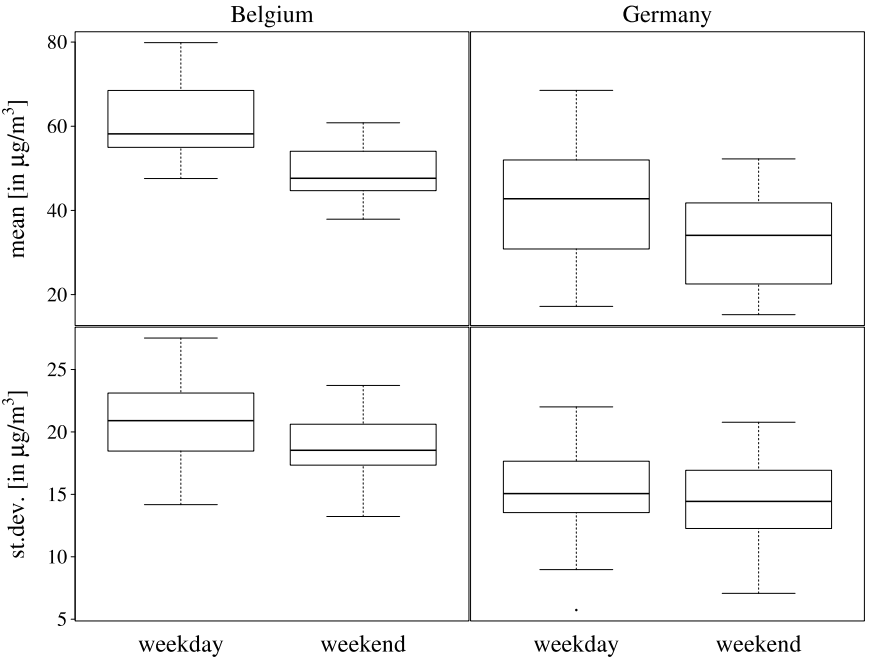

Fig. A.2. Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian industrial site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

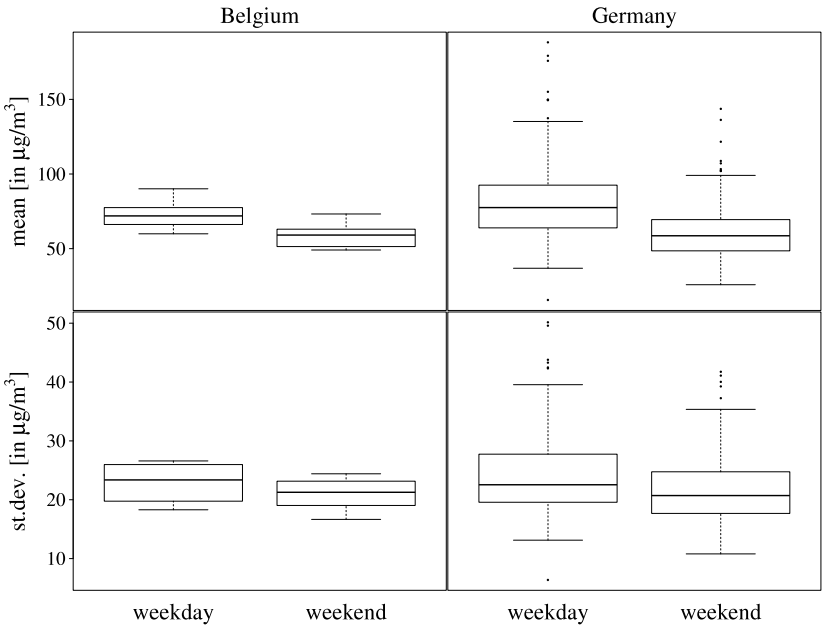

Fig. A.3. Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian traffic site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

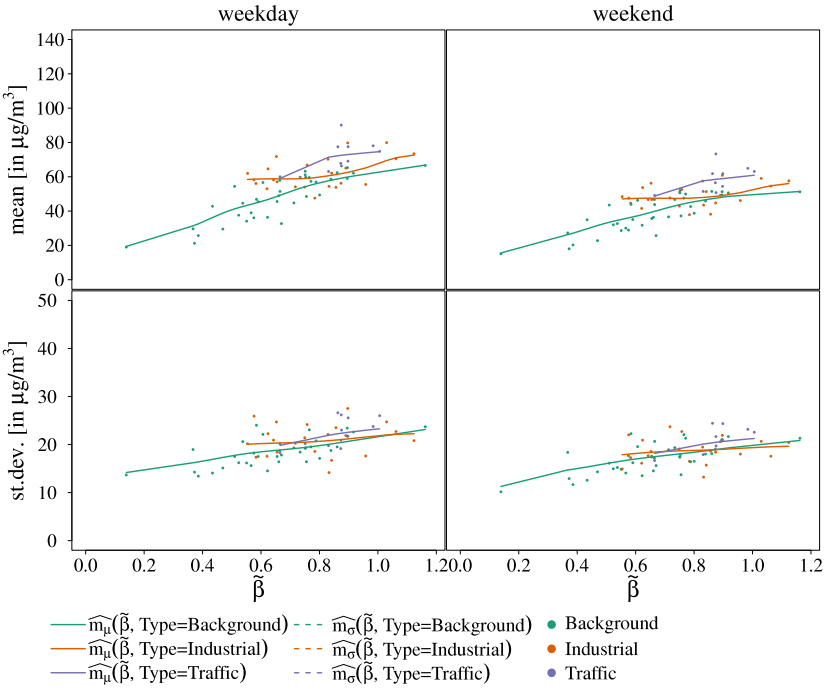

Fig. A.4. Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

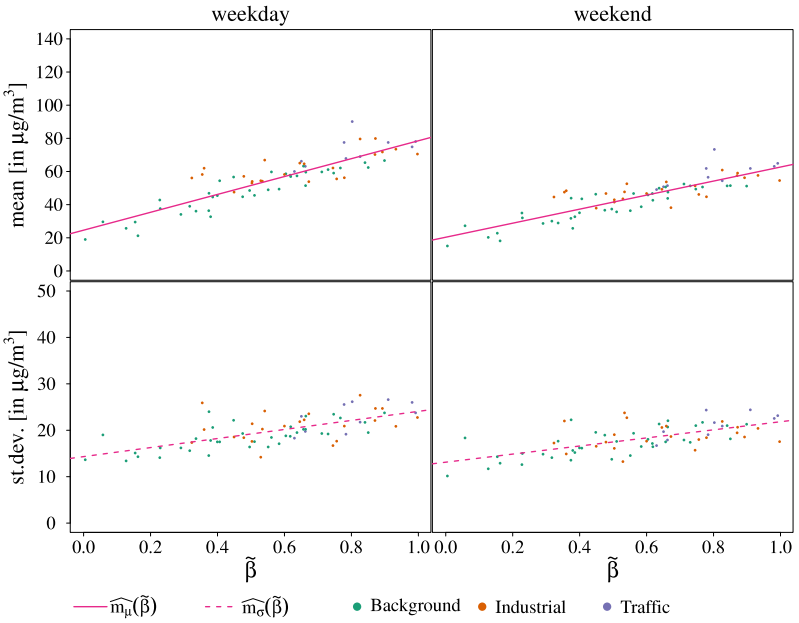

Fig. A.5. Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

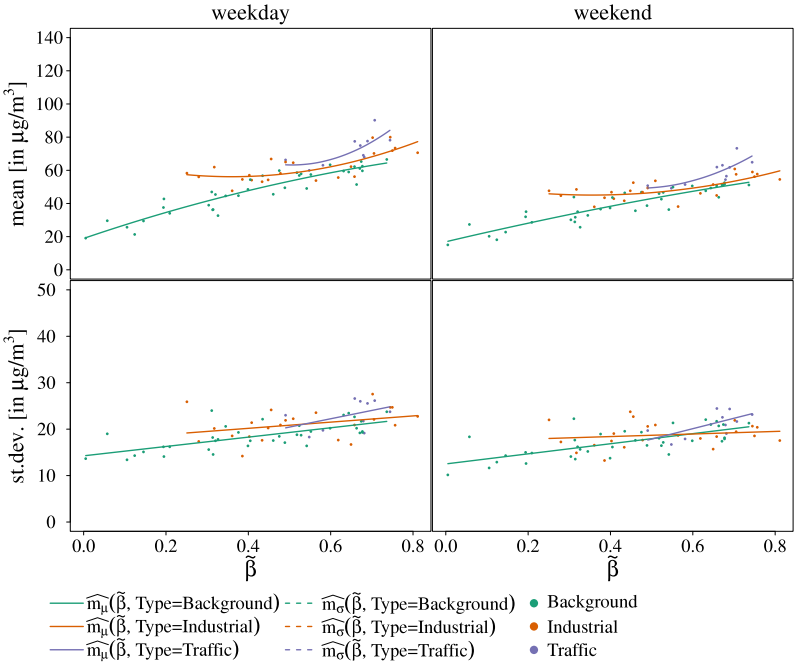

Fig. A.6. Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

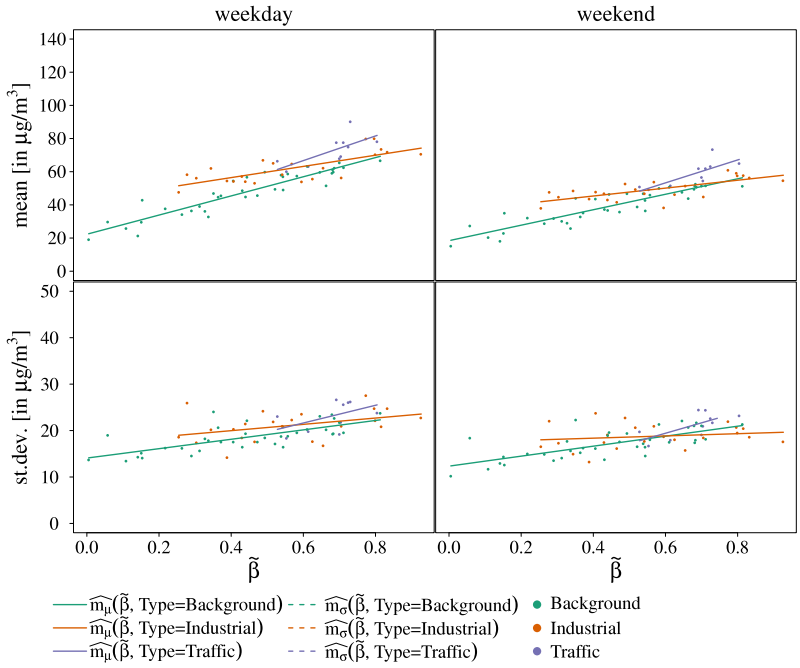

Fig. A.7. Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

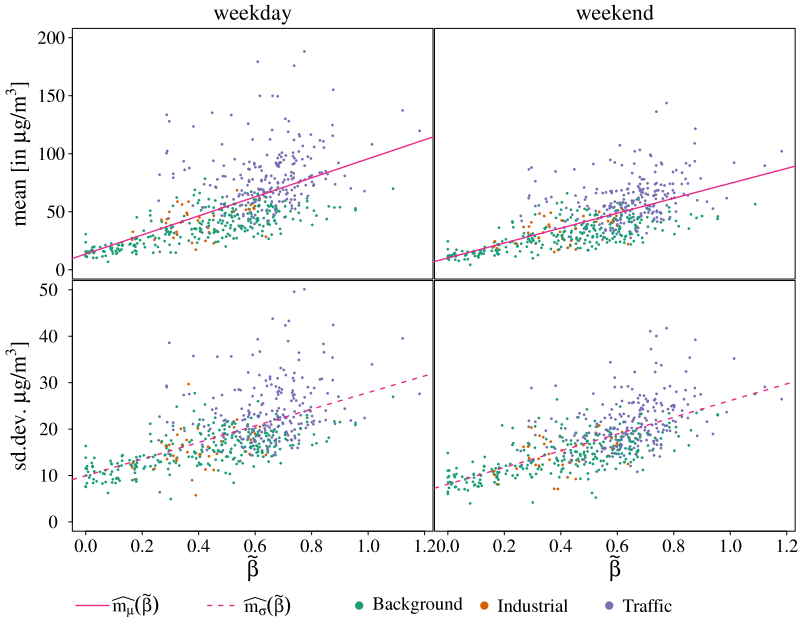

Fig. A.8. German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

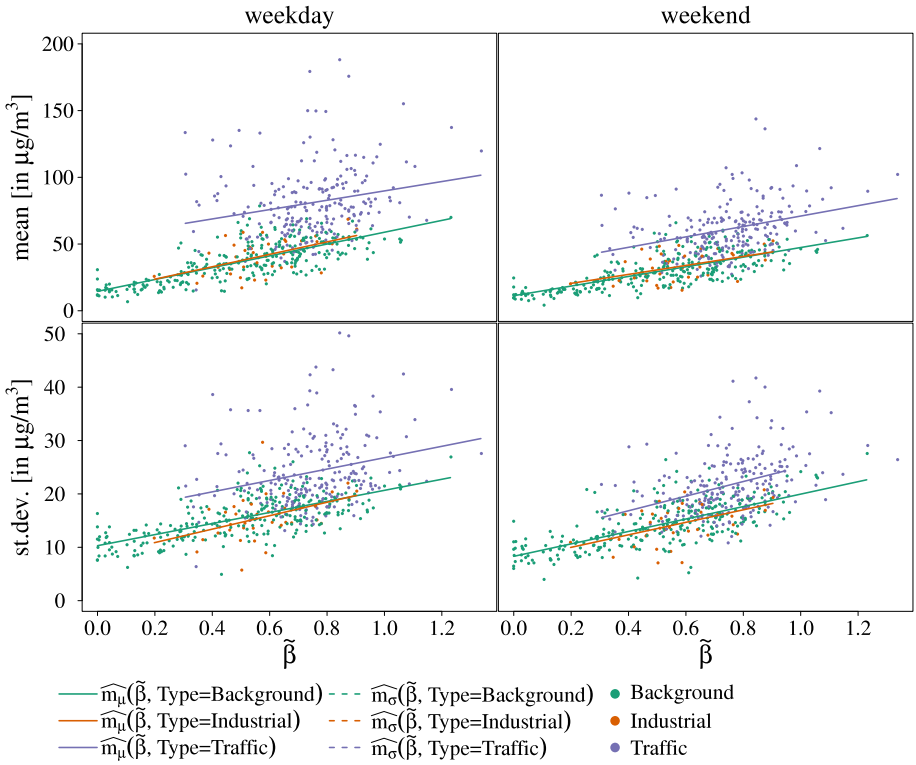

Fig. A.9. German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

Appendix B. Data related descriptions

B.1. Metadata in AirBase

AirBase consists of monitoring data from fixed monitoring sites as well as meta-information on the monitoring sites involved. The following meta-information is provided by AirBase: station European code, station local code, country iso code, country name, station name, station start date, station end date, type of station, station ozone classification, station type of area, station subcat rural back, street type, station longitude deg, station latitude deg, station altitude, station city, lau level1 code, lau level2 code, lau level2 name, EMEP station.

With regard to air pollution analysis the following variables might be of interest:

- type of station — Background, Industrial, Traffic
- station ozone classification — rural, rural background, suburban, urban (the pollutants NO_2 and O_3 are strongly correlated, see Janssen et al., 2008, p. 4889),
- station type of area — rural, suburban, urban
- station subcat rural back — near city, regional, remote
- street type — Canyon street ($L/H < 1.5$), Highway (average speed vehicles > 80 km/h), Unknown, Wide street ($L/H > 1.5$); length (L) of the canyon usually expresses the road distance between two major intersections; height (H) of the canyon

- station longitude deg
- station latitude deg
- station altitude.

In our work we consider station longitude deg, station latitude deg and type of station.

B.2. Data processing and data quality AirBase

In the following we describe how we have processed the hourly recorded NO_2 values and provide information about the data quality. Quality flags in the raw data of the AirBase statistics indicate the quality of each measurement value. A quality flag $>$ 0 indicates valid measurement data. A quality flag $<=$ 0 indicates invalid or missing data ([dataset] EEA, European Environment Agency, 2016).

Belgian AirBase data: The time period 1st Jan 2001 to 31st Dec 2006 has $24 * (365 * 6 + 1) = 52\,584$ h. A full sample with recorded hourly values for each of the 70 monitoring sites would therefore consist of $52\,584 * 70 = 3\,680\,880$ observations. There is no entry in the source data for 815 064 site-date-hour combinations, which corresponds to about 22.14%. This is partly due to the fact that some sites have not recorded the NO_2 concentrations over the whole period, either they have been built up after 1st Jan 2001 or switched off before 31st Dec 2006 or for some time between the 1st Jan 2001 and the 31st Dec 2006. The percentage of either missing or not validated entries in the source data is equal to $371\,497/(3\,680\,880 - 815\,064) \widehat{=} 13.43\%$. We have omitted missing and non validated values from further analysis and have extracted from the daily maximum NO_2 concentration for each site-day combination the remaining data which results in 112 340 maximum values, compared to $70 * (365 * 6 + 1) = 153\,370$ maximum values if data for each site-date combination existed. The Belgian data do not contain any extremely high values (above 500 $\mu g/m^3$) nor any negative daily maximum values.

German AirBase data: The time period 1st Jan 2007 to 31st Dec 2012 has $24*(365*6+2) = 52\,608$ h. A full sample with recorded hourly values for each of the 537 monitoring sites would therefore consist of $52\,608 * 537 = 28\,250\,496$ observations. There is no entry in the source data for 5 391 528 site-date-hour combinations, which corresponds to about 19.08%. This is partly due to the fact that some sites have not recorded the NO_2 concentrations over the complete time, either they have been built up after 1st Jan 2001 or switched off before 31st Dec 2006 or for some time between the 1st Jan 2001 and the 31st Dec 2006. The percentage of either missing or not validated entries in the source data is equal to $1\,547\,472/(28\,250\,496 - 5\,391\,528) \widehat{=} 6.77\%$. We have omitted missing and non validated values from further analysis and have extracted from the daily maximum NO_2 concentration for each site-day combination the remaining data which results in 920 343 maximum values, compared to $537 * (365 * 6 + 2) = 1\,177\,104$ maximum values if data for each site-date combination existed. Omitting missing and non validated values reduces the number of sites from 537 to 536. Further investigation has shown that the source data do not contain any validated data for site DETH082. Three daily maximum values have been removed as they are extremely high (above 500 $\mu g/m^3$) and 58 as they are negative such that finally 920 282 maximum values and 536 sites remain for further analysis.

References

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC Press, London.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D.J., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci. Total Environ. 407 (6), 1852–1867. http://dx.doi.org/10.1016/j.scitotenv.2008.11.048.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K.T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M.J., Dėdelė, A., Gražulevičienė, R., Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., de Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B., de Hoogh, K., 2013. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe –The ESCAPE project. Atmos. Environ. 72, 10–23. http://dx.doi.org/10.1016/j.atmosenv.2013.02.037.

Bivand, R., Keitt, T., Rowlingson, B., 2017. rgdal: bindings for the geospatial data abstraction library. URL: https://CRAN.R-project.org/package=rgdal R package version 1.2-6.

Brunsdon, C., Chen, H., 2014. GISTools: some further GIS capabilities for R. URL: https://CRAN.R-project.org/package=GISTools R package version 0.7-4.

Cressie, N.A.C., 1993. Statistics for spatial data. In: Wiley Series in Probability and Mathematical Statistics, revised ed. Wiley, New York.

[dataset] EEA, European Environment Agency, 2010a. CORINE land cover 2000 raster data, version 13 (05/2010). URL: https://www.eea.europa.eu/ds_resolveuid/b00116e51c79865cf89a84162b8fd21e (Acessed on 29.05.17).

[dataset] EEA, European Environment Agency, 2010b. CORINE land cover 2006 raster data - version 13 (02/2010). URL: https://www.eea.europa.eu/ds_resolveuid/a645109f7a11d43f5d7e275d81f35c61 (Acessed on 29.05.17).

[dataset] EEA, European Environment Agency, 2016. AirBase –European air quality database, version 8. URL: https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8 (Accessed on 20.04.17).

[dataset] EMEP and CEIP, 2014. Present state of emission data. URL: http://www.ceip.at/ms/ceip_home1/ceip_home/webdab_emepdatabase/reported_emissiondata/ (Accessed on 29.05.17).

Diggle, P.J., Ribeiro Jr, P.J., 2007. Model-based geostatistics. Springer, New York, . http://dx.doi.org/10.1007/978-0-387-48536-2.

Fensterer, V., Küchenhoff, H., Maier, V., Wichmann, H.-E., Breitner, S., Peters, A., Gu, J., Cyrys, J., 2014. Evaluation of the impact of low emission zone and heavy traffic ban in Munich (Germany) on the reduction of PM10 in ambient air. Int. J. Environ. Res. Public Health 11 (5), 5094–5112. http://dx.doi.org/10.3390/ijerph110505094.

Feranec, J., Soukup, T., Hazeu, G., Jaffrain, G. (Eds.), 2016. European landscape dynamics: CORINE land cover data. CRC Press, Boca Raton, Florida.

Gilliland, F., Avol, P.K., Jerrett, M., Dvonch, T., Lurmann, F., Buckley, T., Breysse, P., Keeler, G., de Villiers, T., McConnell, R., 2005. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: lessons learned from the Centers for Childrens Environmental Health and Disease Prevention Research. Environ. Health Perspect. 113 (10), 1447–1454. https://doi.org/10.1289/ehp.7673.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R J. 8, 204–218.

Haupt, H., Schnurbus, J., Semmler, W., 2018. Estimation of grouped, time-varying convergence in economic growth. Eco. Sta. forthcoming. http://dx.doi.org/10.1016/j.ecosta.2017.09.001.

Haupt, H., Schnurbus, J., Tschernig, R., 2010. On nonparametric estimation of a hedonic price function. J. Appl. Econometrics 5, 894–901. https://doi.org/10.1002/jae.1186.

Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: The np package. J. Stat. Softw. 27 (5), 1–32. https://doi.org/10.18637/jss.v027.i05.

Hennig, F., Sugiri, D., Tzivian, L., Fuks, K., Moebus, S., Jöckel, K.-H., Vienneau, D., Kuhlbusch, T.A., de Hoogh, K., Memmesheimer, M., et al., 2016. Comparison of land-use regression modeling with dispersion and chemistry transport modeling to assign air pollution concentrations within the Ruhr area. Atmosphere 7 (3), 48. https://doi.org/10.3390/atmos7030048.

Hijmans, R.J., 2016. raster: geographic data analysis and modeling. URL: https://CRAN.R-project.org/package=raster. R package version 2.5-8.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578. http://dx.doi.org/10.1016/j.atmosenv.2008.05.057.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., 2006. Spatial interpolation of ambient ozone concentrations from sparse monitoring points in Belgium. J. Environ. Monit. 8 (11), 1129–1135. http://dx.doi.org/10.1039/b612607n.

Janssen, S., Dumont, G., Fierens, F., Mensink, C., 2008. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmos. Environ. 42 (20), 4884–4903. http://dx.doi.org/10.1016/j.atmosenv.2008.02.043.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statist. Sinica 14, 485–512. URL: http://www.jstor.org/stable/24307205.

Li, Q., Racine, J., 2007. Nonparametric econometrics: theory and practice. Princeton University Press.

Mammen, E., Rothe, C., Schienle, M., 2012. Nonparametric regression with nonparametrically generated covariates. Ann. Statist. 40 (2), 1132–1170. URL: http://www.jstor.org/stable/41713668.

Mercer, L.D., Szpiro, A.A., Sheppard, L., Lindström, J., Adar, S.D., Allen, R.W., Avol, E.L., Oron, A.P., Larson, T., Liu, L.-J.S., Kaufman, J.D., 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NO_x) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Atmos. Environ. 45 (26), 4412–4420. http://dx.doi.org/10.1016/j.atmosenv.2011.05.043.

Montero, J.-M., Fernández-Avilés, G., Mateu, J., 2015. Spatio-temporal prediction and kriging. In: Spatial and spatio-temporal geostatistical modeling and kriging. John Wiley & Sons, Ltd, pp. 266–273. http://dx.doi.org/10.1002/9781118762387.ch8.

Nash, J.C., 2014. On best practice optimization methods in R. J. Stat. Softw. 60 (2), 1–14. https://doi.org/10.18637/jss.v060.i02.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. J. Stat. Softw. 43 (9), 1–14. https://doi.org/10.18637/jss.v043.i09.

Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. Internat. Econom. Rev. 25, 221–247. URL: http://www.jstor.org/stable/2648877.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/.

Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., Byers, J.W., 2015. timeDate: Rmetrics - chronological and calendar objects. URL: https://CRAN.R-project.org/package=timeDate R package version 3012.100.

Robinson, D., 2017. broom: convert statistical analysis objects into tidy data frames. URL: https://CRAN.R-project.org/package=broom R package version 0.4.2.

Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhalation Toxicol. 19 (sup1), 127–133. https://doi.org/10.1080/08958370701495998.

Sahsuvaroglu, T., Arain, A., Kanaroglou, P., Finkelstein, N., Newbold, B., Jerrett, M., Beckerman, B., Brook, J., Finkelstein, M., Gilbert, N.L., 2006. A land use regression model for predicting ambient concentrations of nitrogen dioxide in Hamilton, Ontario, Canada. J. Air Waste Manage. Assoc. 56 (8), 1059–1069. http://dx.doi.org/10.1080/10473289.2006.10464542.

Sperlich, S., 2009. A note on non-parametric estimation with predicted variables. Econom. J. 12, 382–395. https://doi.org/10.1111/j.1368-423X.2009.00291.x.

Wang, R., Henderson, S.B., Sbihi, H., Allen, R.W., Brauer, M., 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmos. Environ. 64, 312–319. http://dx.doi.org/10.1016/j.atmosenv.2012.09.056.

Wolf, K., Cyrys, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., Peters, A., 2017. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. Sci. Total Environ. 579, 1531–1540. http://dx.doi.org/10.1016/j.scitotenv.2016.11.160.