UNIVERSITÄT PASSAU

Wirtschaftswissenschaftliche Fakultät

Four essays on statistical modelling of environmental data

DISSERTATION

zur Erlangung des akademischen Grades

eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)

an der Universität Passau

eingereicht von

Svenia Elena Behm

Passau, Juli 2021

Disputation am:	03.02.2022
Erstgutachter:	Prof. Dr. Harry Haupt Lehrstuhl für Statistik und Data Analytics Universität Passau
Zweitgutachter:	PD Dr. Joachim Schnurbus Lehreinheit für Computergestützte Statistik und Mathematik Universität Passau
Vositz:	Prof. Dr. Markus Diller Lehrstuhl für Betriebswirtschaftslehre mit Schwerpunkt Taxation Universität Passau

Acknowledgements

At this point, I would like to say thank you to everyone who accompanied me during my doctoral studies. In particular, I thank Prof. Dr. Harry Haupt for the supervision and the collaboration during my dissertation. The ongoing constructive discussions with him motivated me to continue researching and seeking clarity. He has always encouraged and supported me to attend summer schools and conferences, which has broadened my professional and social horizons. Further, I thank PD Dr. Joachim Schnurbus for the many helpful suggestions he has made about my work and for being the second examiner of my dissertation as well as Prof. Dr. Markus Diller for chairing the examination board. My thanks also go to my co-authors Prof. Dr. Harry Haupt, Dr. Angelika Schmid, and Dr. Markus Fritsch for the pleasant collaboration and the fruitful debates. A warm thank you also to my colleagues, PD Dr. Joachim Schnurbus, Dr. Markus Fritsch, Ida Bauer, Matthias Wild, Constanze Lehner, Dr. Angelika Schmid, and Dr. Sandra Huber, who enriched my time at the chair. A special thank you to Dr. Angelika Schmid for bringing me to the chair and for being my office mate at the beginning of my dissertation. I am also very grateful to Ida Bauer for sharing the office with me for a couple of years.

My most sincere thanks go to my partner and my family who have always believed in me and have been my greatest mental support. Without my partner, Helmut Schmid, I would probably never have started this thesis and never have finished it. Thank you for your wise and loving attitude, your open ear, your unconditional support, and your patience with me. Finally, I thank my parents, Gabriele and Norbert Behm, for always giving me the opportunity to develop freely in my life.

Contents

List of Figures

List of Tables

1 Introduction

This dissertation deals with geostatistical, time series, and regression analytical approaches for modelling spatio-temporal processes, using air quality data in the applications. The work is structured into four essays the abstracts of which are given in the following.

The first essay is titled "Spatial detrending revisited: Modelling local trend patterns in NO_2-concentration in Beglium and Germany". It is written in co-authorship by Prof. Dr. Harry Haupt and Dr. Angelika Schmid and published in 2018 in Spatial Statistics 28, pp. 331-351 (`https://doi.org/10.1016/j.spasta.2018.04.004`).

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.

The second essay is titled "Predictability of hourly nitrogen dioxide concentration". It is written in co-authorship with Prof. Dr. Harry Haupt and published in 2020 in Ecological Modelling 428, 109076 (`https://doi.org/10.1016/j.ecolmodel.2020.109076`).

Abstract. Temporal aggregation of air quality time series is typically used to investigate stylized facts of the underlying series such as multiple seasonal cycles. While aggregation reduces complexity, commonly used aggregates can suffer from non-representativeness or non-robustness. For example, definitions of specific events such as extremes are subjective and may be prone to data contaminations. The aim of this paper is to assess the predictability of hourly nitrogen dioxide concentrations and to explore how predictability depends on (i) level of temporal aggregation, (ii) hour of day, and (iii) concentration level. Exploratory tools are applied to identify structural patterns, problems related to commonly used aggregate statistics and suitable statistical modeling philosophies, capable of handling multiple seasonalities and non-stationarities. Hourly times series and subseries of daily measurements for each hour of day are used to investigate the predictability of pollutant levels for each hour of day, with prediction horizons ranging from one hour to one week ahead. Predictability is assessed by time series cross validation of a loss function based on out-of-sample prediction errors. Empirical evidence on hourly nitrogen dioxide measurements suggests that predictability strongly depends on conditions (i)-(iii) for all statistical models: for specific hours of day, models based on daily series outperform models based on hourly series, while in general predictability deteriorates with exposure level.

The third essay is titled "Agglomeration and infrastructure effects in land use regression models for air pollution – Specification, estimation, and interpretations". It is written in co-authorship with Dr. Markus Fritsch and published in 2021 in Atmospheric Environment 253, 118337 (`https://doi.org/10.1016/j.atmosenv.2021.118337`).

Abstract. Established land use regression (LUR) techniques such as linear regression utilize extensive selection of predictors and functional form to fit a model for every data set on a given pollutant. In this paper, an alternative to established LUR modeling is employed, which uses additive regression smoothers. Predictors and functional form are selected in a data-driven way and ambiguities resulting from specification search are mitigated. The approach is illustrated with nitrogen dioxide (NO_2) data from German monitoring sites using the spatial predictors longitude, latitude, altitude and structural predictors; the latter include population density, land use classes, and road traffic intensity measures. The statistical performance of LUR modeling via additive regression smoothers is contrasted with LUR modeling based on parametric polynomials. Model evaluation

is based on goodness of fit, predictive performance, and a diagnostic test for remaining spatial autocorrelation in the error terms. Additionally, interpretation and counterfactual analysis for LUR modeling based on additive regression smoothers are discussed.
Our results have three main implications for modeling air pollutant concentration levels: First, modeling via additive regression smoothers is supported by a specification test and exhibits superior in- and out-of-sample performance compared to modeling based on parametric polynomials. Second, different levels of prediction errors indicate that NO_2 concentration levels observed at background and traffic/industrial monitoring sites stem from different processes. Third, accounting for agglomeration and infrastructure effects is important: NO_2 concentration levels tend to increase around major cities, surrounding agglomeration areas, and their connecting road traffic network.

The fourth essay is titled "Outlier detection and visualisation in multi-seasonal time series and its application to hourly nitrogen dioxide concentration". It is written in single authorship and has not been published yet.

Abstract. Outlier detection in data on air pollutant recordings is conducted to uncover data points that refer to either invalid measurements or valid but unusually high concentration levels. As air pollutant data is typically characterised by multiple seasonalities, the task of outlier detection is associated with the question of how to deal with such non-stationarities. The present work proposes a method that combines time series segmentation, seasonal adjustment, and standardisation of random variables. While the former two are employed to obtain subseries of homoskedastic data, the latter ensures comparability across the subseries. Further, the standardised version of the seasonally adjusted subseries represents a scaled measure for the outlyingness of each data point in the original time series from its mean and therefore forms a suitable basis for outlier detection. In an empirical application to data on hourly NO_2 concentration levels recorded at a traffic monitoring site in Cologne, Germany, over the years 2016 to 2019, the common boxplot criterion is used to examine each standardised seasonally adjusted subseries for positive outliers. The results of the analyses are put into their natural temporal order and displayed in a heatmap layout that provides information on when single and sequential outliers occur.

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

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.

Keywords. Stationarity, RIO model, Air pollution, Land use, Nonparametrics.

2.1 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 nonstationarities 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 fulfill 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.

2.2 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 2.7.B.1. Further, we use the CORINE Land Cover 2006 (CLC2006) data layer in a 100×100 meter 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 2.1 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.

Table 2.1: 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

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 2.7.B.2. Fig. 2.1 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 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. 2.A.1-2.A.3 we explore

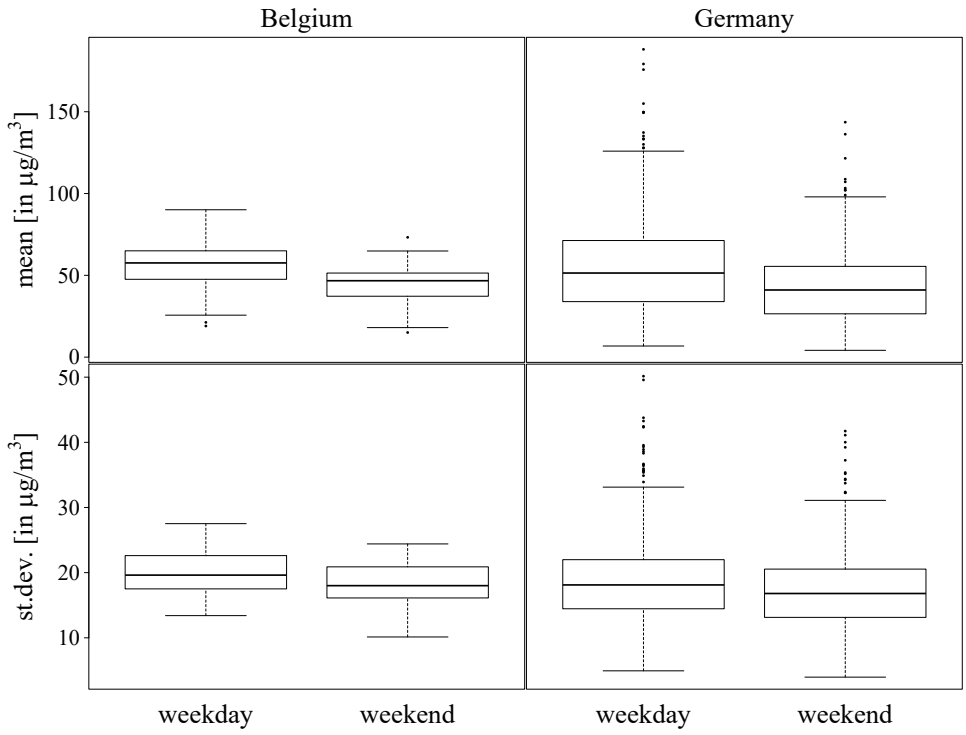


Figure 2.1: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian monitoring site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

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 2.2 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`

Table 2.2: 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

(Pebesma, 2004; Gräler 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).

2.3 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

$$Z_t(\boldsymbol{s}) = Y_t(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}) \tag{2.1}$$

with $\epsilon_t(\boldsymbol{s}) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$. 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 $\forall \boldsymbol{s}, \boldsymbol{s'} \in D_s$, $\boldsymbol{s} \neq \boldsymbol{s'}$,

$$\mathrm{E}[Y_t(\boldsymbol{s})] = \mu, \tag{2.2a}$$

$$\mathrm{Var}[Y_t(\boldsymbol{s})] = \sigma^2, \tag{2.2b}$$

$$\mathrm{C}(h) = \mathrm{Cov}[Y_t(\boldsymbol{s}), Y_t(\boldsymbol{s}')] = \sigma^2 \rho(h), \tag{2.2c}$$

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

2.3.1 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 (2.2a)-(2.2c) hold. The grouped land use classes (see Table 2.2) enter the equation for the pollutant specific β-index according to

$$\beta(\boldsymbol{s}, r) = \log\left[1 + \sum_{k=1}^{11} a_k \cdot sh_k(\boldsymbol{s}, r)\right], \tag{2.3}$$

where $sh_k(\boldsymbol{s}, r)$ describes the share of the k-th 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. (2.3) 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 2.A.1 in the Appendix.

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,

$$\mu \approx m_\mu(\beta, X), \tag{2.4a}$$

$$\sigma \approx m_\sigma(\beta, X). \tag{2.4b}$$

In their application to Belgian data, Janssen et al. (2008) assume that mean and standard deviation in Eqs. (2.4a) and (2.4b) 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. (2.3) and therefore depends on $\boldsymbol{s}$ and $a_1, \ldots, a_{11}$. Under assumption (2.4a) the coefficients $a_1, \ldots, a_{11}$ in Eq. (2.3) 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 $\widehat{\mu}_i = \widehat{m}_\mu(\tilde{\beta}_i, X_i)$ and $\widehat{\sigma}_i = \widehat{m}_\sigma(\tilde{\beta}_i, X_i)$, respectively.

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

according to

$$z^*_{i,t} = z_{i,t} + (\mu^{ref} - \widehat{\mu}_i), \tag{2.5a}$$

$$z^{**}_{i,t} = (z^*_{i,t} - \bar{z}^*_i)\frac{\sigma^{ref}}{\widehat{\sigma}_i} + \bar{z}^*_i. \tag{2.5b}$$

After filtering the monitored data $z_{i,t}$ according to Eqs. (2.5a) and (2.5b), 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 $\boldsymbol{s} \in D_s$,

$$E[Z^{**}_t(\boldsymbol{s})] = \mu(\boldsymbol{s}) + (\mu^{ref} - \widehat{\mu}(\boldsymbol{s})) \approx \mu^{ref}, \tag{2.6}$$

relying on assumption (2.4a) in the last transformation, and

$$Var[Z^{**}_t(\boldsymbol{s})] = \left(\sigma^{ref}\right)^2 Var\left[\frac{Z^*_t(\boldsymbol{s}) - \bar{Z}^*(\boldsymbol{s})}{\widehat{\sigma}(\boldsymbol{s})}\right] \approx \left(\sigma^{ref}\right)^2, \tag{2.7}$$

since the middle term describes the standardisation of $Z^*_t(\boldsymbol{s})$. Eqs. (2.6) and (2.7) show that the filtered data process approximately satisfies the (weak) stationarity properties (2.2a)-(2.2c) 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 $\widehat{Y}^{**}_t(\boldsymbol{s}_0)$ 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

$$\widehat{Y}^*_t(\boldsymbol{s}_0) = (\widehat{Y}^{**}_t(\boldsymbol{s}_0) - \bar{\widehat{Y}}^{**}(\boldsymbol{s}_0))\frac{\widehat{\sigma}(\boldsymbol{s}_0)}{\sigma^{ref}} + \bar{\widehat{Y}}^{**}(\boldsymbol{s}_0), \tag{2.8a}$$

$$\widehat{Y}_t(\boldsymbol{s}_0) = \widehat{Y}^*_t(\boldsymbol{s}_0) - (\mu^{ref} - \widehat{\mu}(\boldsymbol{s}_0)). \tag{2.8b}$$

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.

2.3.2 Spatial trend modelling: A general nonparametric approach

There are several options to include further predictors in Eqs. (2.4a) and (2.4b). 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

$$\bar{Z} = m(\beta, X) + U \qquad \text{with } E(U|\beta, X) = 0, \tag{2.9}$$

based on Eq. (2.4a). A generalised least squares estimator is denoted as $\widehat{m}_{LL}$, where $(\widehat{m}_{LL}, \widehat{\gamma})$ minimises

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

where $\boldsymbol{W} = (\beta, X)$ denotes the vector of regressors, $K = k_\beta \cdot k_X$ is a product kernel, and $\boldsymbol{h} = (h_\beta, h_X)'$ is a vector of bandwidths which we estimate using least squares cross validation (see Li and Racine, 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. (2.9) 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

$$X_1 = \underset{k\in\{1,\ldots,11\}}{\operatorname{argmax}} sh_k, \tag{2.10}$$

$$X_2 = \underset{k\in\{1,\ldots,11\}\setminus X_1}{\operatorname{argmax}} sh_k, \tag{2.11}$$

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

$$S = sh_{X_1} + sh_{X_2}. \tag{2.12}$$

is defined as the sum of the shares of the first and second largest class. Then, instead of the predictors used in Eq. (2.9), we consider the categorical predictors X_1, X_2, and the sites' type, and the continuous predictor S.

2.4 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 nonparamatric approach without $\beta-$index but with the predictors defined in Eqs. (2.10)-(2.12) together with the sites' type.

The estimated QL trend functions for Belgian data are displayed in Fig. 2.2 and replicate results of Janssen et al. (2008, right plot of Fig. 5 and middle plot of Fig. 8), while Fig. 2.3 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. 2.3 and bear no visible difference to the LL trend functions displayed for Germany in Fig. 2.A.8 in 2.7.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. 2.1 and indicates that, on average, the concentration level of NO_2 drops from weekdays to weekends.

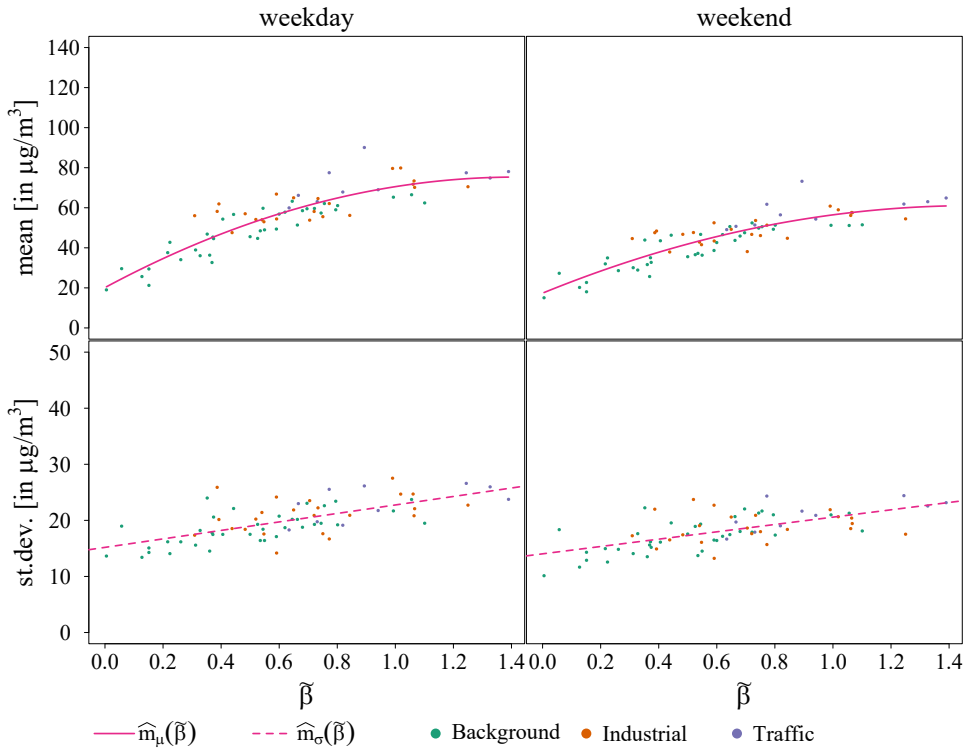


Figure 2.2: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

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 Fig. 2.3 and Fig. 2.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 2.7.A.

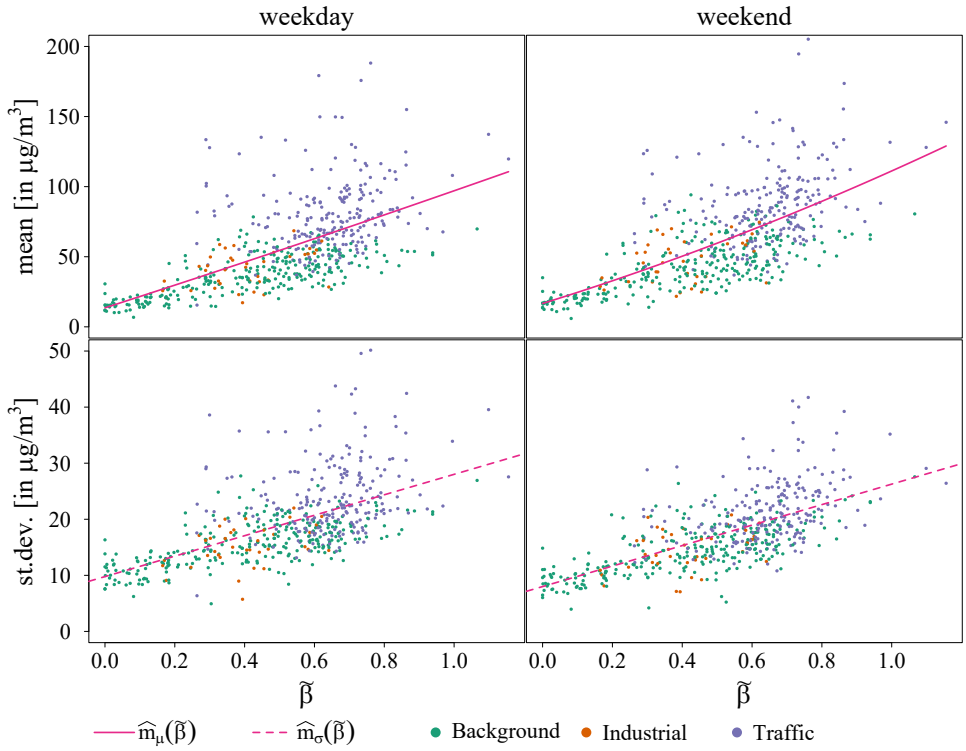


Figure 2.3: German data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation (specification QL).

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. (2.9). Fig. 2.4 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.

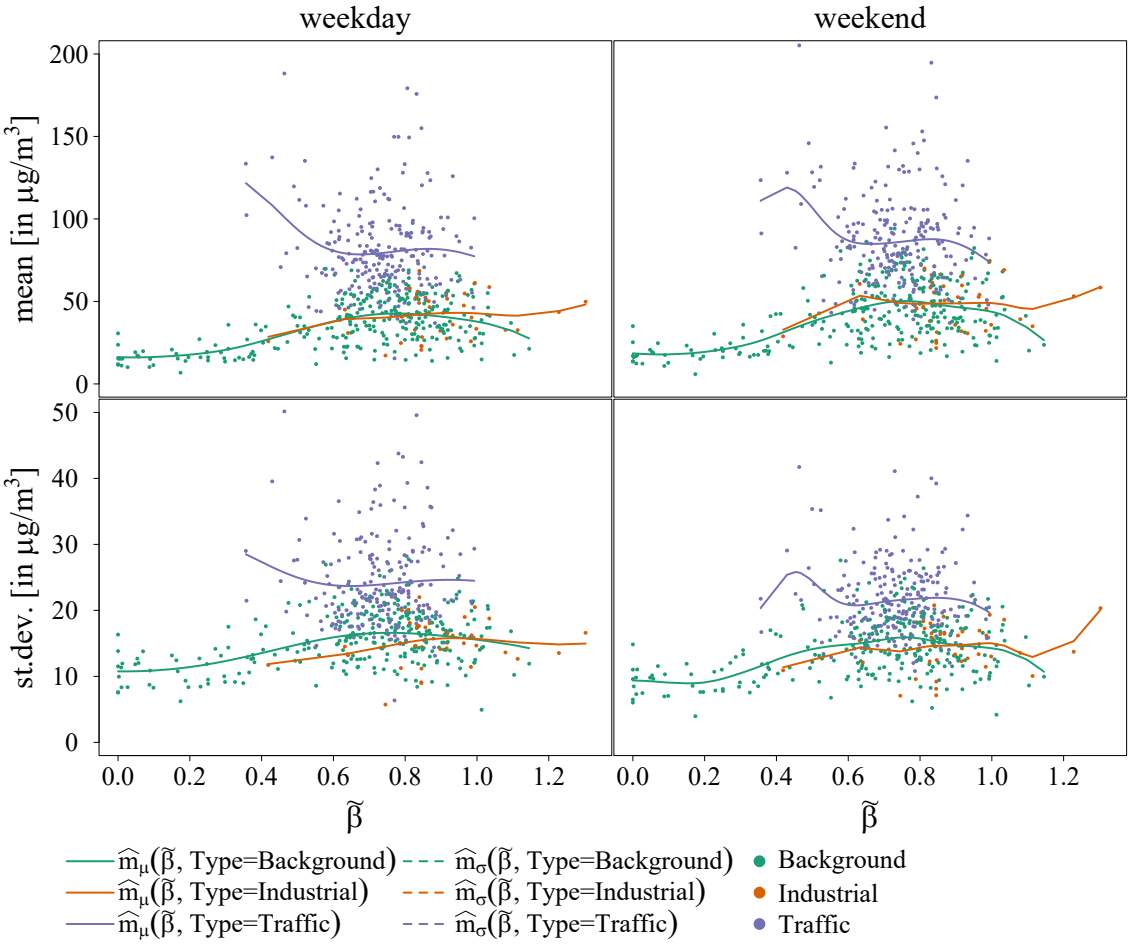


Figure 2.4: German data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

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. 2.5. Visual inspection of the results and comparison to Fig. 2.3 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 speci-

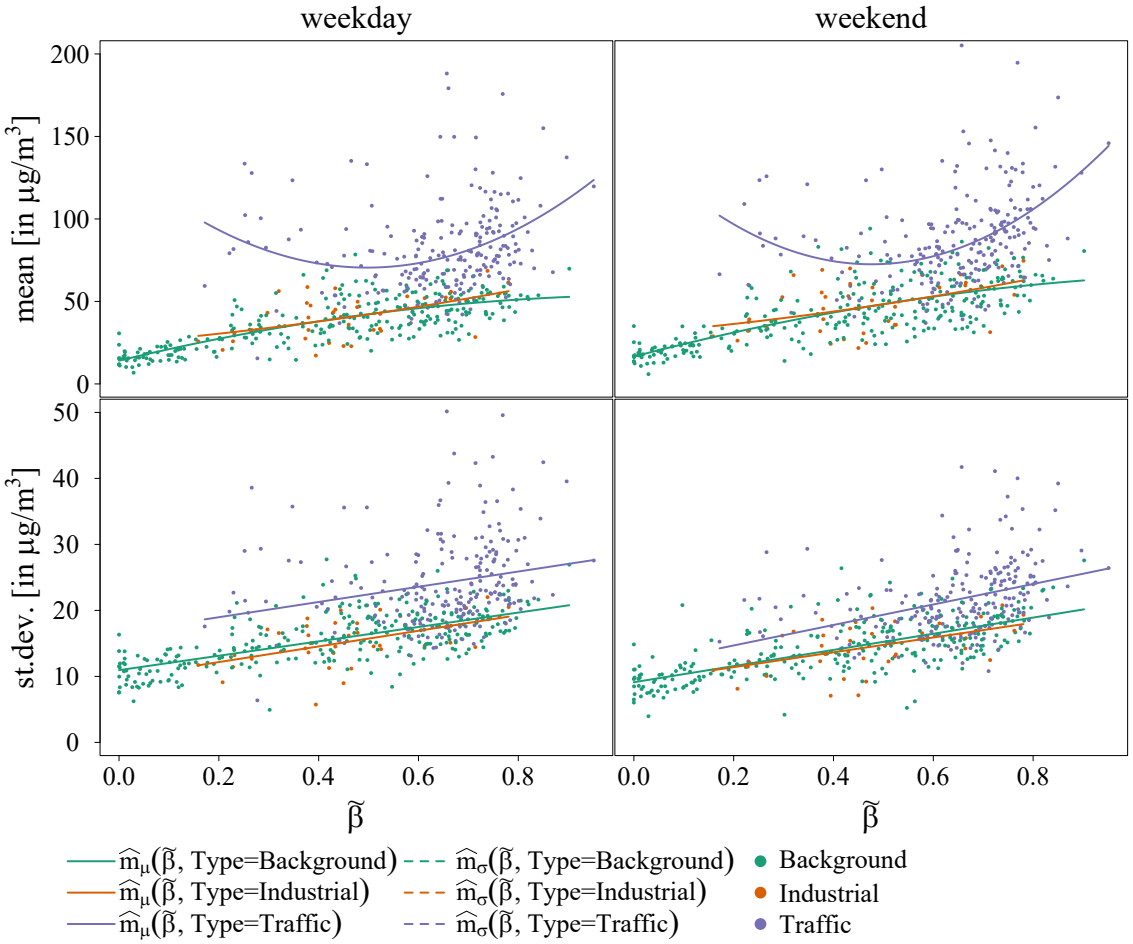


Figure 2.5: German data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

fications LL and TypeLL for Germany are provided in Figs. 2.A.8 and 2.A.9 in 2.7.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. 2.A.6 and 2.A.7 in 2.7.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 hours (7 minutes) to derive the trend functions using NP, compared to 16 seconds (4.3 seconds) 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 above) – 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 2.3 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 2.A.2 in 2.7.A provides further and more detailed results on our LOOCV 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.

Table 2.3: 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

2.5 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.

2.6 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.

2.7 Appendix

2.7.A Tables and figures

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

Germany	QL	LL	TypeQL	TypeLL	NP	NPnoBeta*	NPnoBeta**
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

* with first and second largest LUC ** with first, second and third largest LUC

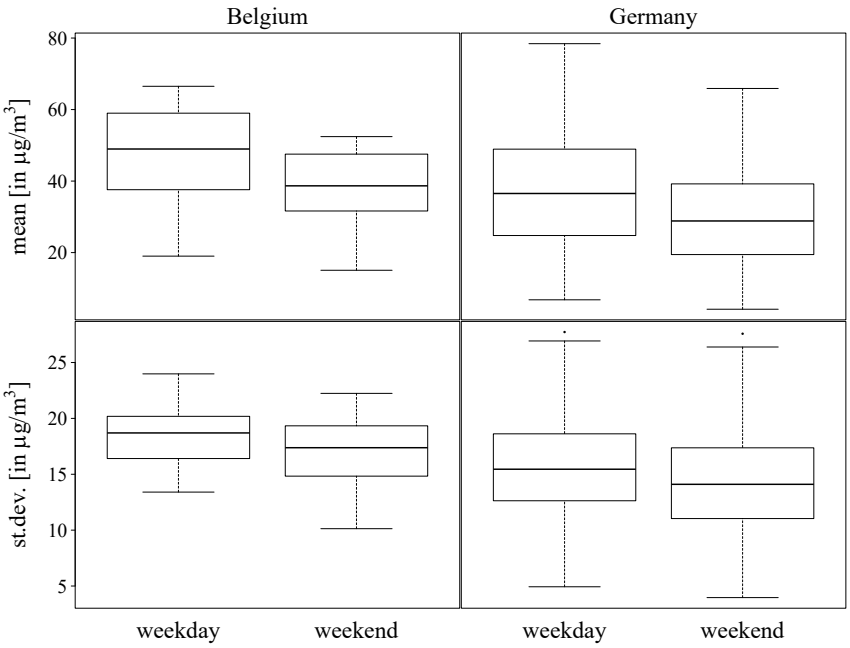


Figure 2.A.1: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian background site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

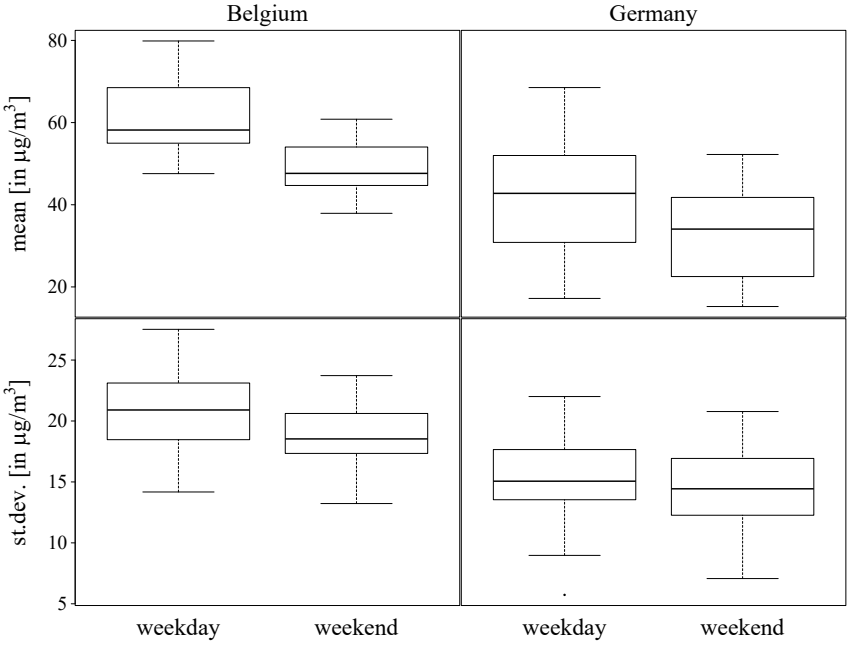


Figure 2.A.2: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian industrial site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

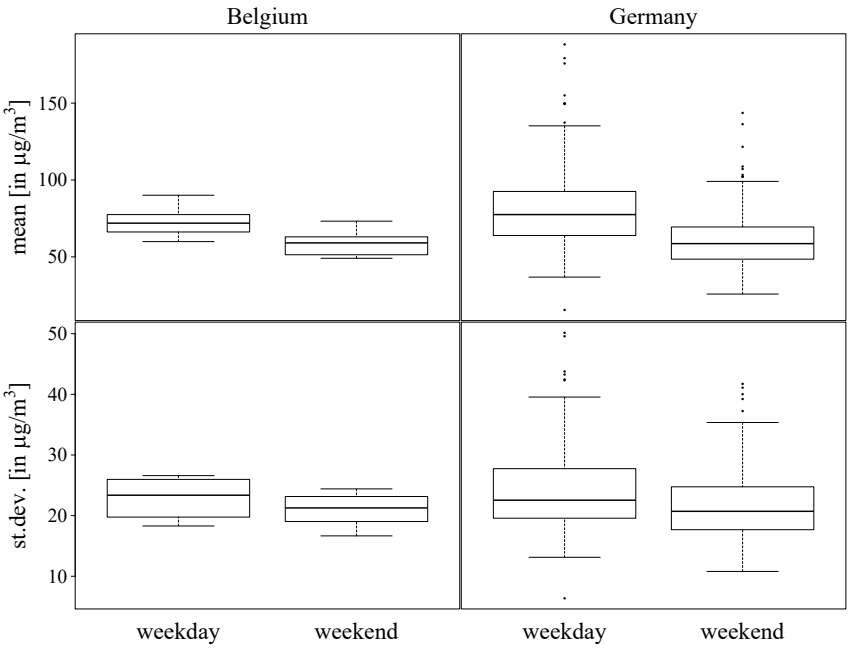


Figure 2.A.3: Top: Boxplots of the mean and standard deviation over the daily maximum NO_2 values of each Belgian traffic site, separately for weekdays and weekends. Bottom: Analogous boxplots for German data.

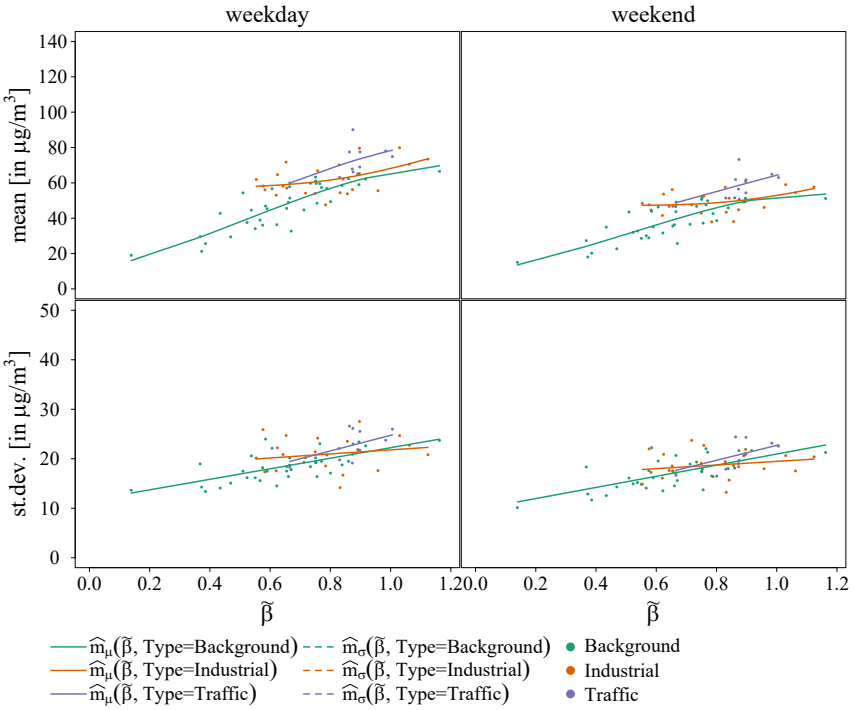


Figure 2.A.4: Belgian data $(\tilde{\beta}_i, \hat{\mu}_i)$ and $(\tilde{\beta}_i, \hat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to the nonparametric approach (specification NP).

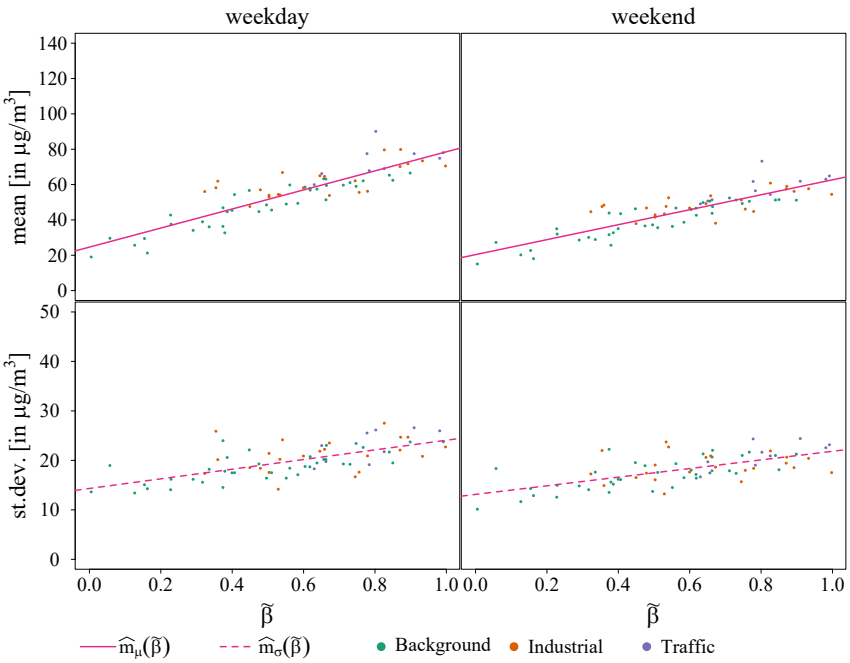


Figure 2.A.5: Belgian data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

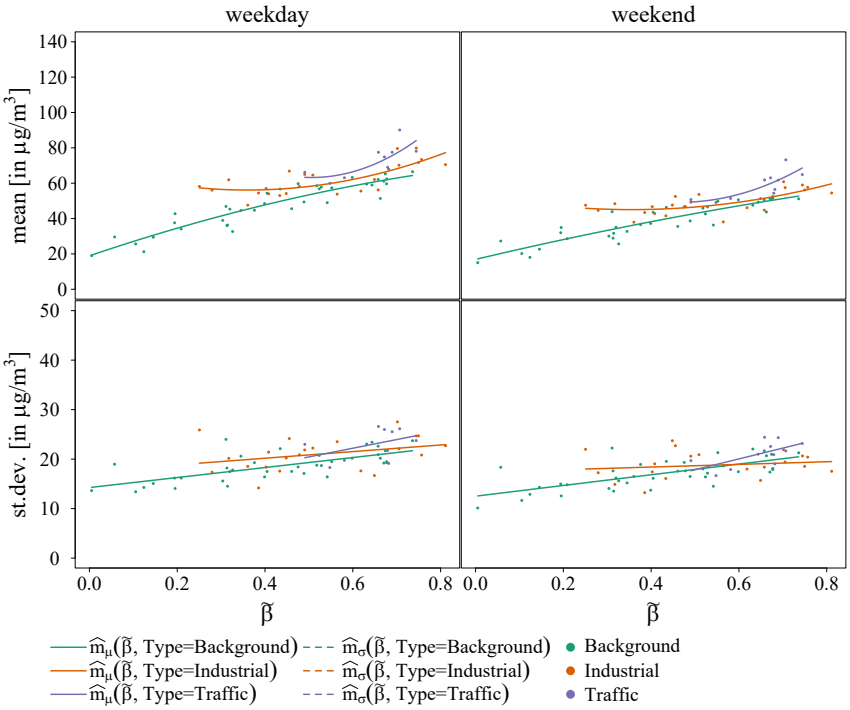


Figure 2.A.6: Belgian data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a quadratic trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeQL).

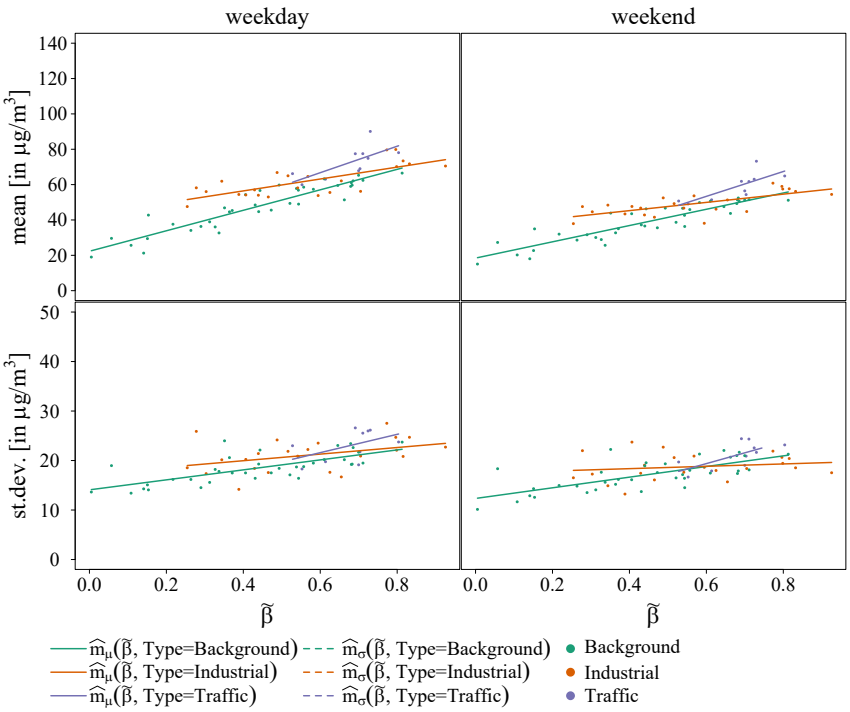


Figure 2.A.7: Belgian data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

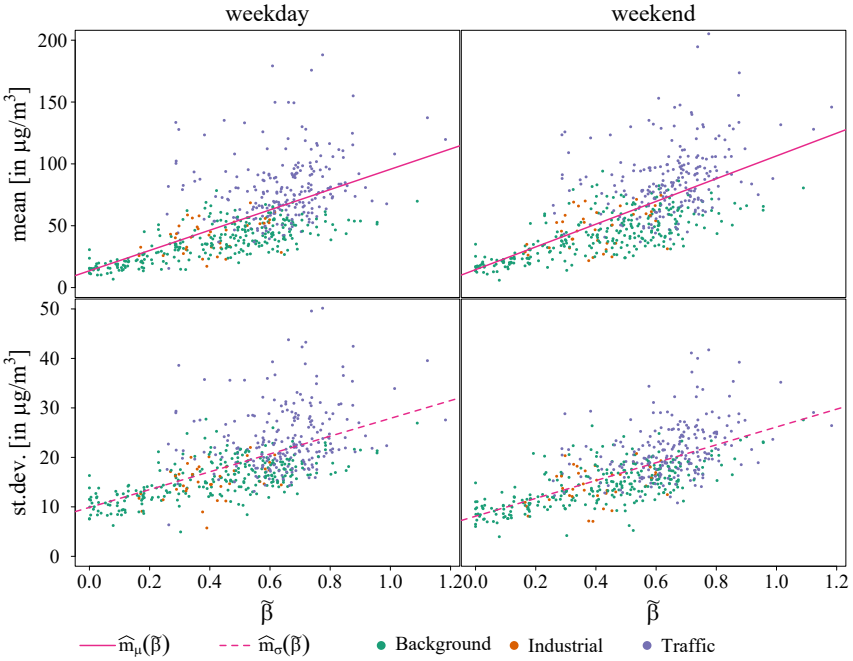


Figure 2.A.8: German data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation (specification LL).

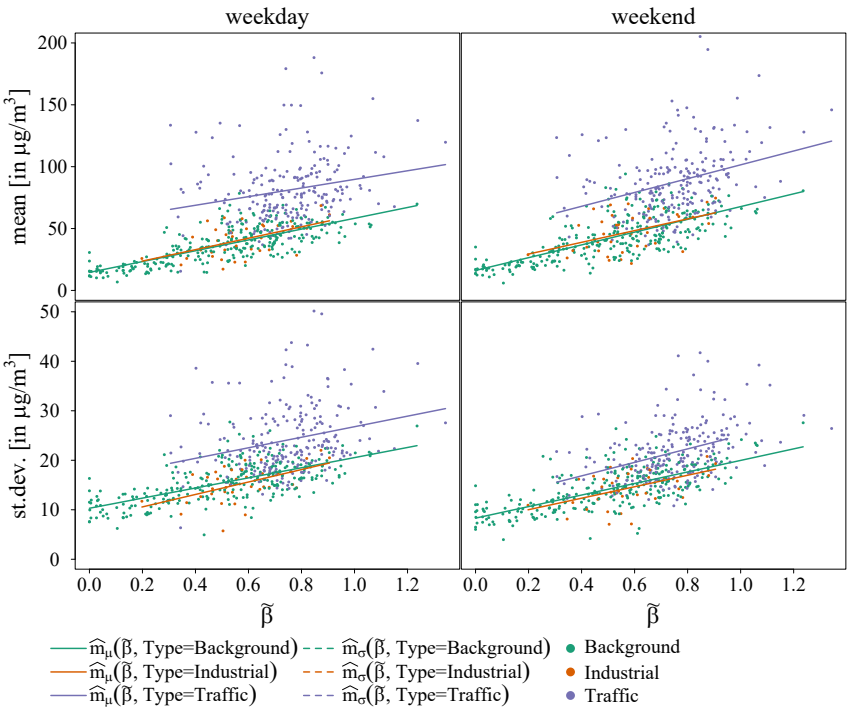


Figure 2.A.9: German data $(\tilde{\beta}_i, \widehat{\mu}_i)$ and $(\tilde{\beta}_i, \widehat{\sigma}_i)$ scatterplots for weekdays and weekends (top left to bottom right); $\tilde{\beta}_i$ and the fitted trend functions correspond to a linear trend for the mean and a linear trend for the standard deviation; both are allowed to differ with an indicator for the sites' type (specification TypeLL).

2.7.B Data related descriptions

2.7.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 > 80km/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.

2.7.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 $24 \cdot (365 \cdot 6 + 1) = 52\,584$ hours. A full sample with recorded hourly values for each of the 70 monitoring sites would therefore consist of $52\,584 \cdot 70 = 3\,680\,880$ 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 $70 \cdot (365 \cdot 6 + 1) = 153\,370$ 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 $24 \cdot (365 \cdot 6 + 2) = 52\,608$ hours. A full sample with recorded hourly values for each of the 537 monitoring sites would therefore consist of $52\,608 \cdot 537 = 28\,250\,496$ 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 $537 \cdot (365 \cdot 6 + 2) = 1\,177\,104$ 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.

2.8 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. Science of the Total Environment 407, 1852–1867. doi: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. Atmospheric Environment 72, 10–23. doi: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. 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 29th May 2017.

[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 29th May 2017.

[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 20th April 2017.

[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 29th May 2017.

Diggle, P.J., Ribeiro Jr, P.J., 2007. Model-based geostatistics. Springer, New York. doi: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 PM_{10} in ambient air. International Journal of Environmental Research and Public Health 11, 5094–5112. doi: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 Children's Environmental Health and Disease Prevention Research. Environmental Health Perspectives 113, 1447–1454.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. The R Journal 8, 204–218. doi:10.32614/RJ-2016-014.

Haupt, H., Schnurbus, J., Semmler, W., 2018. Estimation of grouped, time-varying convergence in economic growth. Econometrics and Statistics forthcoming. doi:10.1016/j.ecosta.2017.09.001.

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

Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27, 1–32. doi: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, 48. doi: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. Atmospheric Environment 42, 7561–7578. doi: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. Journal of Environmental Monitoring 8, 1129–1135. doi:10.1039/b612607n.

Janssen, S., Dumont, G., Fierens, F., Mensink, C., 2008. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment 42, 4884–4903. doi:10.1016/j.atmosenv.2008.02.043.

Li, Q., Racine, J., 2004. Cross-validated local linear nonparametric regression. Statistica

Sinica 14, 485–512. URL: https://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. Annals of Statistics 40, 1132–1170. doi:10.1214/12-AOS995.

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). Atmospheric Environment 45, 4412–4420. doi: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, 266–273. doi:10.1002/9781118762387.ch8.

Nash, J.C., 2014. On best practice optimization methods in R. Journal of Statistical Software 60, 1–14. doi:10.18637/jss.v060.i02.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. Journal of Statistical Software 43, 1–14. doi:10.18637/jss.v043.i09.

Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. International Economic Review 25, 221–247. doi:10.2307/2648877.

Pebesma, E.J., 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30, 683–691. doi: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 char-

acterizing intraurban air pollution exposure. Inhalation Toxicology 19, 127–133. doi: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. Journal of the Air & Waste Management Association 56, 1059–1069. doi:10.1080/10473289.2006.10464542.

Sperlich, S., 2009. A note on non-parametric estimation with predicted variables. Econometrics Journal 12, 382–395. doi: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. Atmospheric Environment 64, 312–319. URL: http://www.sciencedirect.com/science/article/pii/S1352231012009272, doi: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. Science of the Total Environment 579, 1531–1540. doi:10.1016/j.scitotenv.2016.11.160.

3 Predictability of hourly nitrogen dioxide concentration

Abstract. Temporal aggregation of air quality time series is typically used to investigate stylized facts of the underlying series such as multiple seasonal cycles. While aggregation reduces complexity, commonly used aggregates can suffer from non-representativeness or non-robustness. For example, definitions of specific events such as extremes are subjective and may be prone to data contaminations. The aim of this paper is to assess the predictability of hourly nitrogen dioxide concentrations and to explore how predictability depends on (i) level of temporal aggregation, (ii) hour of day, and (iii) concentration level. Exploratory tools are applied to identify structural patterns, problems related to commonly used aggregate statistics and suitable statistical modeling philosophies, capable of handling multiple seasonalities and non-stationarities. Hourly times series and subseries of daily measurements for each hour of day are used to investigate the predictability of pollutant levels for each hour of day, with prediction horizons ranging from one hour to one week ahead. Predictability is assessed by time series cross validation of a loss function based on out-of-sample prediction errors. Empirical evidence on hourly nitrogen dioxide measurements suggests that predictability strongly depends on conditions (i)-(iii) for all statistical models: for specific hours of day, models based on daily series outperform models based on hourly series, while in general predictability deteriorates with exposure level.

Keywords. Air pollution prediction, Nitrogen dioxide, Predictability, Aggregation level, Multiple seasonality.

3.1 Introduction

Many environmental and socio-economic processes are shaped by seasonal variations in annual cycles, while anthropogenic influences typically contribute daily and weekly seasonal patterns. Air pollutants are a particularly complex example of such processes. For example, nitrogen dioxide (NO_2) concentrations have a pronounced daily cycle layered on top of weekly and annual cycles. Assessment and prediction of such processes is of considerable public and political interest. To protect human health, the European Air Quality Directive (2008/50/EC) requires that the member states of the European Union are obliged to monitor air quality and take action if the limit values for air pollutant concentrations are (expected to be) exceeded (Council of the European Union, 2008). The projects "Review of evidence on health aspects of air pollution — REVIHAAP" (WHO, 2013) and "Health risks of air pollution in Europe – HRAPIE" (WHO, 2013) are carried out by the World Health Organization (WHO) to provide a comprehensive scientific evidence-based overview of the links between air pollution and adverse health effects and the associated economic consequences. In the realm of REVIHAAP and HRAPIE so called concentration-response functions are defined to quantify the health impacts due to air pollution and then used as a basis for a cost-benefit analysis of EU air quality policy (Héroux et al., 2015). The reports of REVIHAAP and HRAPIE give evidence that short- and long-term exposure to NO_2 may encourage respiratory and cardiovasular diseases. Recent works reviewing epidemiological studies are, among others, Hoek et al. (2013) and Atkinson et al. (2018). According to the 2018 report on air quality in Europe (EEA, 2018) the ambient air concentrations of particulate matter (PM), ozone (O_3), NO_2 and further pollutants decreased between 2000 and 2016. There are, however, still exceedances of the critical values and related health risks require further research. Hence the development of accurate and reliable methods for assessment and prediction of air quality indicators remains to be an important field of research.

The prediction of local air quality indicators such as NO_2 concentrations is a particularly challenging problem due to the requirement of statistical modeling of periodic components and their potential variation over time. Although air quality data are typically available as hourly time series, analyses often rely on aggregated data, using statistics such as daily maxima or averages. Various modeling frameworks can be used to address the inherent

non-stationarities and to produce predictions for different time horizons (e.g., Cabaneros et al., 2019; Gocheva-Ilieva et al., 2014; Lawson et al., 2011; Moisan et al., 2018; Sharma et al., 2009; Zhao et al., 2018) such as hourly exposure, daily maxima or annual averages. Consequences of aggregation are on the one hand a reduction of the number of seasonal cycles, but on the other hand an inability to predict the precise hour and not only the level of the maximum exposure on the next day. Hence, relevant information can be masked, biased or even deleted in the aggregation process. For times series with multiple seasonal patterns such as local NO_2 concentrations, simple aggregates such as time series of daily maxima still exhibit cyclical patterns of weekly and annual length and hence require complex modeling approaches.

In the literature on multiple seasonal patterns in air quality processes, the focus of removing one or more components of the underlying time series often is on descriptive data analysis: visualizations at different levels of temporal aggregation are used to detect seasonal patterns and discuss their anthropogenic and meteorological drivers. Mayer (1999) conducts an exploratory analysis of hourly data on four air pollutants (nitrogen monoxide (NO), NO_2, ozone (O_3) and the sum of NO_2 and O_3 (O_x)), recorded in Stuttgart, southern Germany, over the period 1981-1993, and identifies annual, weekly and daily cycles. DeGaetano and Doherty (2004) analyze data on meteorology and hourly PM concentration levels at 20 monitoring sites in New York City with respect to their spatial and temporal variation. They consider extreme concentration percentiles and the median concentration level for each day-hour combination. Their findings suggest that the amplitudes of the seasonal patterns vary with percentiles and that daily and weekly patterns are mainly caused by anthropogenic factors whereas annual patterns arise rather due to meteorological variations. Liu et al. (2015) investigate hourly $PM_{2.5}$ and PM_{10} concentration levels in Beijing over the period 2004-2012. They explore the daily and annual seasonal patterns in the data using heatmaps to visualize the distribution of the pollutants in dependence of the month and the time of the day. Further, they provide a thorough discussion about the extent to which the patterns can be traced back to meteorological phenomena using bivariate polar plots and correlations. Moisan et al. (2018) account for multiple seasonalities and predict $PM_{2.5}$ but use multivariate techniques which require auxiliary variables and corresponding predictions, and may thus suffer from additional sources of bias.

One of the few studies dealing with univariate air quality prediction under consideration

of multiple seasonalities is the recent work of Zhao et al. (2018). They use a time series analysis procedure called *Prophet* to investigate weekly and annual seasonality of daily $PM_{2.5}$ concentrations measured at 220 monitoring sites across the United States over the period 2007-2015. Prophet was initially developed for smoothing and predicting daily business data and relies on the generalized additive modeling approach assuming the time series to be additively decomposable into trend, seasonality, holiday, and error component (Taylor and Letham, 2018). Lawson et al. (2011) apply structural time series models to hourly nitrogen oxide data recorded at a monitoring site in Dublin to predict one day (i.e. 24 hourly steps) ahead. Due to the short time period of just 40 days of data, a single daily seasonal pattern is considered and data referring to weekend days are not included in the analysis. In addition, they use separate models for the series of morning and evening peak values for some ad hoc selected hours. As this can be seen as an aggregation of hourly to daily data, the one day ahead predictions correspond to one step ahead predictions. Applications of the Box-Jenkins modeling approach can be found, among others, in Sharma et al. (2009), Kumar and Jain (2010), or Gocheva-Ilieva et al. (2014). They employ seasonal autoregressive integrated moving average (SARIMA) models based on hourly data for six pollutants (over one year) and monthly maxima values of three pollutants (over 15 years), respectively. The prediction horizon is one hour to three days ahead and one month to two years ahead, respectively.

The aim of this work is to assess the predictability of local concentration levels of NO_2 based on hourly measurements generated by local monitoring sites. For the protection of the public against adverse health effects caused by high exposure levels, not only the value of a daily statistic is important for assessing the exposure, but also the hour of day at which a specific exposure (such as the maximum) occurs. Hence our analysis maintains the hour of day reference and considers one hourly time series y_t as well as 24 daily series $y_{t(\iota)}$ (for each hour ι), for every monitoring station in our sample. We identify daily and weekly seasonal patterns and apply suitable state-of-the-art univariate time series modeling techniques capable of handling multiple seasonalities and nonstationarities. Predictability is defined by a loss function based on out-of-sample h-step prediction errors. Using time series cross-validation, we produce out-of-sample predictions for NO_2 concentration levels for each hour of day, with prediction horizons h ranging from one hour to one week. We thoroughly discuss the relation between predictability and hour of day and respective NO_2

concentration level. In detail, concerning a prediction horizon between one and 24 h, the root mean squared error (RMSE) is calculated over the prediction errors that refer to the specific hour of day to be predicted. With regard to a prediction horizon between one and seven days, the hourly and daily models are evaluated under consideration of the hour of day reference. The comparison between daily and hourly models is used to assess whether, and if so, for which hours of the day the aggregation of hourly to daily data improves the predictive performance. To check the robustness of our results, we conduct the analysis for different monitoring sites. In addition to reporting the predictive performance over all h-step prediction errors, we evaluate the models' predictive performance separately for each hour of day to be predicted.

The remainder of the paper is organized as follows: Section 3.2 describes data and exploratory visualization tools to gain insights into the complex seasonal patterns of the data. Section 3.3 introduces modeling framework and methods used to calculate and evaluate predictions. Section 3.4 discusses the estimated models and their corresponding predictive performance, while Section 3.5 concludes.

3.2 Material

In our empirical application, we consider data on the hourly NO_2 concentration level (in $\mu g/m^3$) recorded by the European Environment Agency in four cities in the south eastern part of Bavaria (Germany) over the years 2014 and 2015. The four selected cities differ with respect to population density and monitoring site environment, leading to sufficiently heterogeneous exposure levels for the purposes in this study. In addition, following Cabaneros et al. (2020), sites with a low percentage of missing values between 2004 and 2015 were selected. Tables 3.1 and 3.2 display the characteristics of the selected monitoring sites and descriptive statistics of the raw data, respectively.

Table 3.1: Air quality monitoring sites: population density, coordinates (latitude, longitude, altitude), environment type, and distance from road.

City	Pop. dens.	Lat.	Long.	Alt.	Environ. type	Dist.
Passau	750	48.57363	13.42204	299 m	Urban background	22 m
Regensburg	1890	49.01523	12.10157	337 m	Urban traffic	6.6 m
Landshut	1100	48.53988	12.15700	390 m	Urban traffic	7.4 m
Burghausen	940	48.17718	12.82931	419 m	Suburban background	30 m

Table 3.2: Order statistics (minimum, median, maximum), mean, standard deviation (SD), SD/Mean, and signal-to-noise ratio (SNR) of the hourly data on NO_2 concentrations.

City	Min.	Median	Max.	Mean	SD	SD/Mean	SNR	Missing
Passau	2.04	27.98	121.00	30.94	16.38	0.53	5.53	0.11%
Regensburg	2.47	36.48	173.19	39.27	19.21	0.49	6.21	0.35%
Landshut	2.24	25.58	115.80	27.56	13.60	0.49	6.14	0.66%
Burghausen	0.48	20.21	105.59	22.29	12.83	0.58	4.80	0.30%

Missing values are imputed by applying the `R` function `na.approx()`. Alternative imputations were calculated using the functions `na.seadec()` (from `R` package `imputeTS`) and `na.interp()` (from `R` package `forecast`), both developed for seasonal time series. The use of different imputation techniques did not lead to changes in the interpretation of our empirical results. Predictive gains, especially in situations with higher percentages of missing values, may be obtained by using more complex and computationally demanding imputation techniques specifically designed for certain constellations of missings and data types. Examples are procedures requiring geo-referenced data (e.g. Yi et al., 2016) or methods based on recurrent dynamics, which do not rely on assumptions about the data generating process (e.g. Cao et al., 2018). A novel procedure proposed by Cabaneros et al. (2020) allows spatiotemporal interpolation and hence may also be used for imputation. Following this avenue is, however, beyond the scope of this paper.

All visualizations and the empirical analysis are produced with the statistical software `R` (R Core Team, 2013), in particular the packages `cowplot` (Wilke, 2019), `data.table` (Dowle and Srinivasan, 2019), `forecast` (Hyndman and Khandakar, 2008, Hyndman et al., 2019), `fpp2` (Hyndman, 2018), `ggplot2` (Wickham, 2016), `imputeTS` (Moritz and Bartz-Beielstein, 2017), `lubridate` (Grolemund and Wickham, 2011), `RColorBrewer` (Neuwirth, 2014), `reshape2` (Wickham, 2007), and `xts` (Ryan and Ulrich, 2018).

In our exposition we discuss empirical findings exemplarily for data recorded in Passau, Germany, and provide a summary of results for other monitoring sites used in this study at the end of Section 3.4. Fig. 3.1 exemplarily displays the complete (and imputed) time series of hourly data for Passau.

Due to the complexity of temporal air quality processes, some key properties of the data can hardly be detected by visual inspection of the complete hourly time series. Fig. 3.2 provides a more comprehensive picture of NO_2 variation over daily and weekly cycles. We

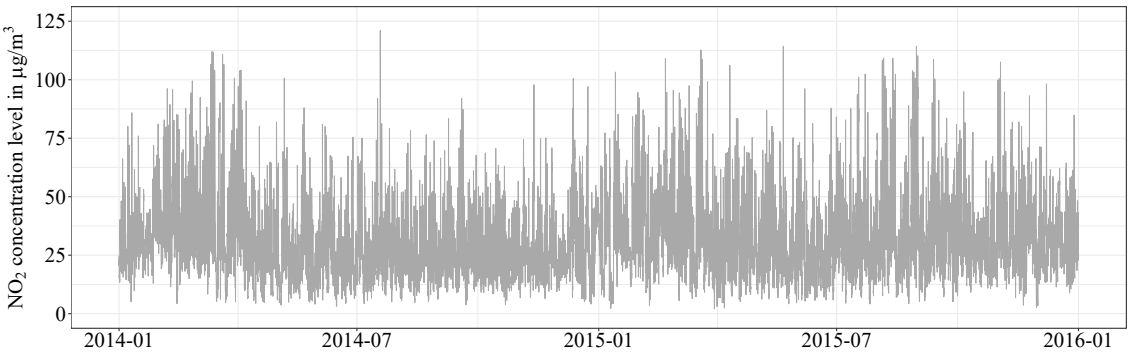


Figure 3.1: Hourly data on NO_2 recorded in 2014 and 2015 at the monitoring site in Passau.

observe differences in the average and median concentration level and the amplitude of the daily series indicating the presence of daily seasonality (e.g. Abdullah et al., 2019; Cabaneros et al., 2020). Between 5pm and 10pm higher mean and median values and higher amplitudes occur in comparison to the other times of day. With all daily series the mean is above the median where the gap between mean and median increases in the evening hours implying an increase in right-skewness. Specifically we highlight the daily time series for 7pm referring to the 60 min between 6pm and 7pm (in the bottom left display), as the typical peak characterizing the bimodal shape of the NO_2 concentration distribution over the hours of a day. Other stylized facts of the daily seasonal pattern, we will use in our discussion, are the 8am series, representing the morning peak, and the 1pm series as the minimum between the peaks. In the bottom right, the boxplots of the daily series for 7pm illustrate the weekly cycle over the days of the week.

The polar plots shown in Fig. 3.3 represent the hourly (left display) and the 7pm daily (right display) time series in a circular layout. Each color refers to a specific week and the corresponding line connects the $24 \cdot 7 = 168$ measurements for the hourly and the seven measurements for the daily time series, respectively. The polar plots reveal the variation of daily and weekly seasonality. We observe higher values in the evening hours, both on weekdays and weekend days. This might be due to an increasing traffic load, people returning from work, or going out on weekends. The NO_2 concentration on Saturday and Sunday mornings usually is remarkably lower as compared to the mornings of the working days.

Fig. 3.4 displays the variation in daily and weekly seasonal patterns over the NO_2 concentration distribution. The plot is constructed by building a weekly series for each day-hour

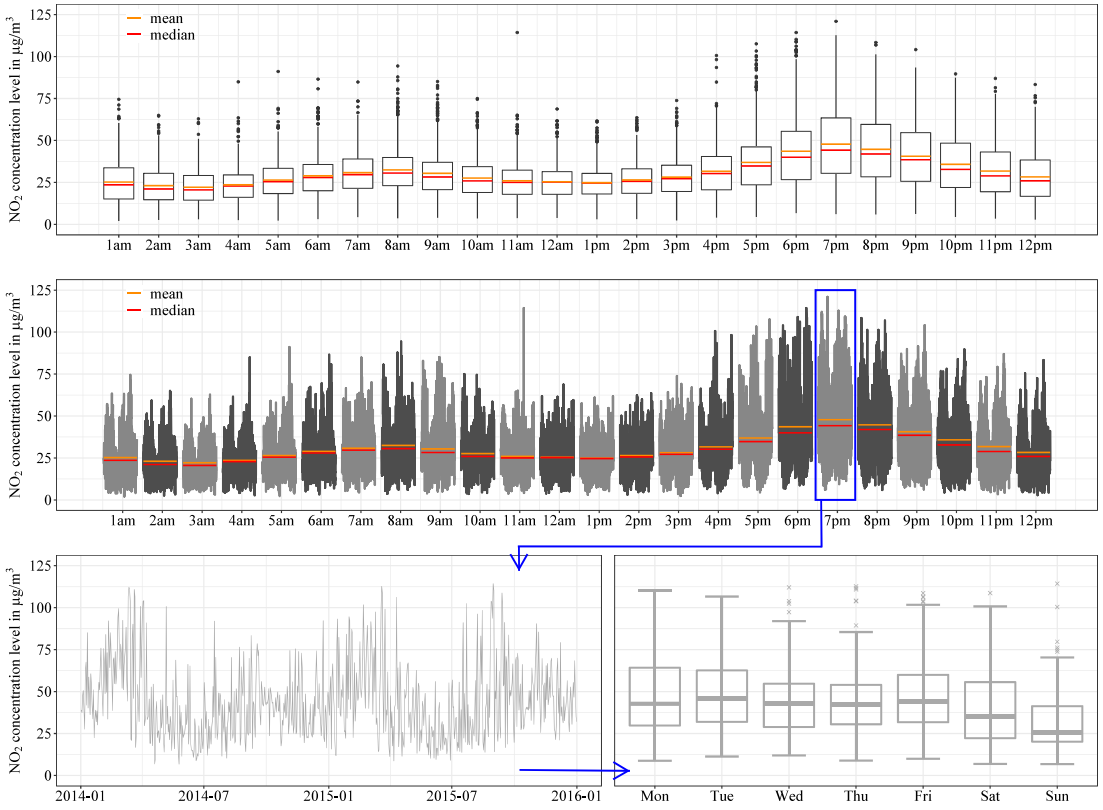


Figure 3.2: Hourly boxplots of daily NO_2 time series for each hour of day over 2014 and 2015 (top); orange and red horizontal lines depict the mean and median values of the daily series, respectively. Daily series for each hour of day over 2014 and 2015 (middle). Daily time series of measurements at 7pm over 2014 and 2015 (bottom left). Daily boxplots of time series of measurements at 7pm over 2014 and 2015 (bottom right).

combination which results in 168 weekly series over 2014 and 2015. For each weekday, the mean and median curves exhibit the typical bimodal shape caused by anthropogenic factors whereas on weekends just one peak occurs in the evening. With regard to the pollution peaks, i.e. values in the top decile (q90_max), we observe more variability as compared to other quantiles (min_q50, q50_q75 and q75_q90) and values that are twice as high as the respective mean and median. In contrast to the smooth curves for mean and central quantiles, the tails of the NO_2 concentration distribution exhibit more wiggliness and remarkable variations in the amplitude. An analysis based on daily mean or median values may draw an overly simplified picture of air pollutant distributions. Vice versa, an analysis based only on daily maxima may be prone to outlier problems and in general may give an unrepresentative picture of air pollutant concentrations since it just incorporates one of the 24 values occurring each day. In any case, the aggregation of the 24 hourly

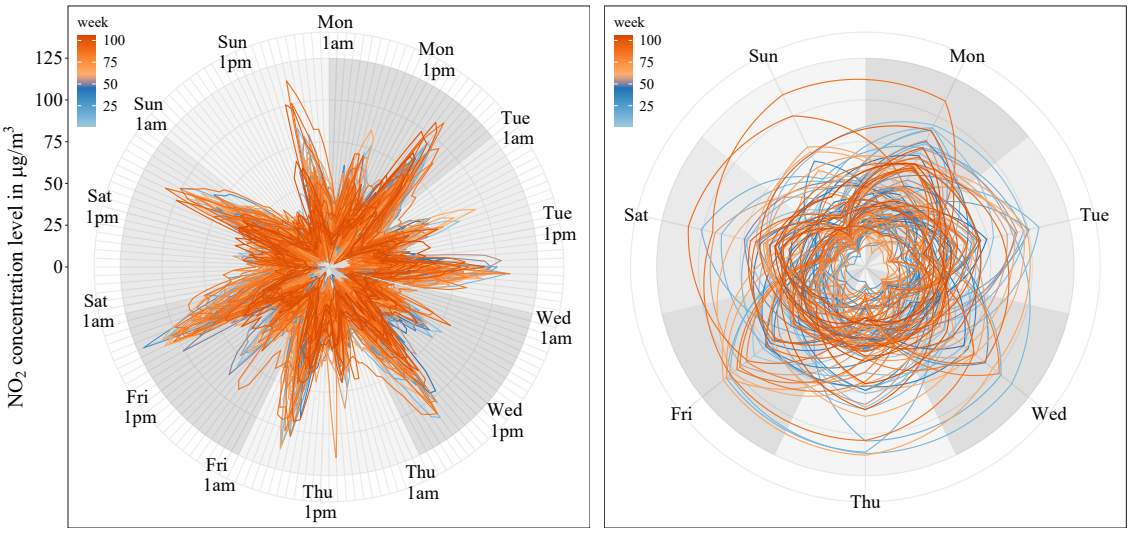


Figure 3.3: Polar plot of hourly NO_2 concentration, where each line refers to a specific week with 168 measurements (left display). Polar plot of daily NO_2 concentration at 7pm, where each line refers to a specific week with seven measurements (right display).

values to a daily statistic such as mean, median, or maximum, causes the hour of day reference to disappear. Based on existing health evidence even one hour peak exposures can have adverse health effects (WHO, 2013). This means that not only the level of daily peaks is relevant to protect the population from exposure to harmful levels, but also the time at which these levels occur.

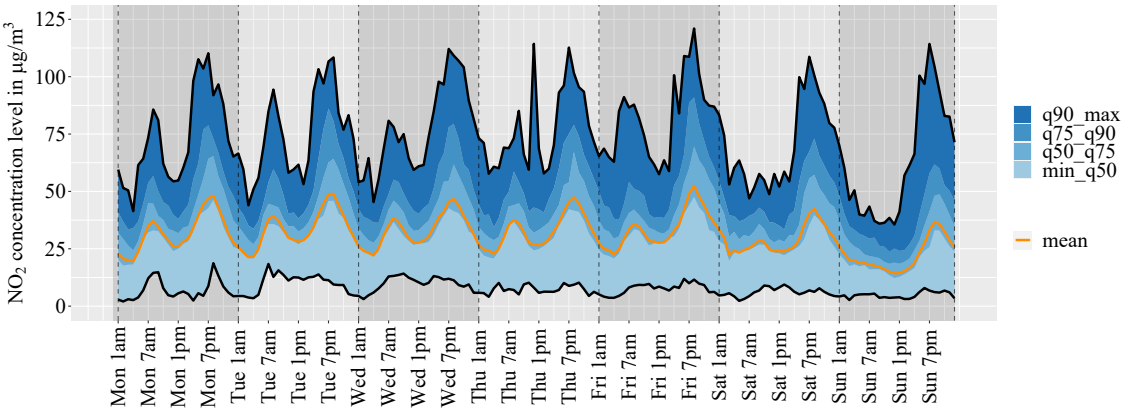


Figure 3.4: Quantiles and mean value of the weekly NO_2 time series for each day-hour combination of the years 2014 and 2015.

If an artificial time series is constructed based on ranks over several series, such as daily

maxima, the resulting series in general does not correspond to an observable time series and has only limited scope for ecological policies. An illustrative example for this claim is given by the two panels in Fig. 3.5: Exemplarily for 8am, 1pm, and 7pm, and two weeks, the NO_2 data is shown in form of boxplots (over 24 NO_2 measurements), one for each day of week. The pink and lightgreen dotted lines connect the median and maximum values, respectively, and the darkgreen, purple and orange solid lines connect the measurements referring to 8am, 1pm, and 7pm. We observe changes in the seasonal figure and the amplitude of the daily series, in particular with respect to the 8am and 7pm series. Regarding the 7pm values, e.g., the sequence of the increases and decreases from Tuesday to Sunday in the left and right panel is exactly reversed. The amplitude of the 8am series is noticeably higher in the left panel than in the right panel. In the left panel, the 7pm values coincide with the maximum values and are remarkably higher as compared to the median values, on five of the seven days they are even marked as outliers. Considering the right panel, the maximum values are outliers on five of the seven days but do not match a specific hour of day. Since the daily maximum values correspond to different hours of the day, it is not possible to provide a hour of day reference using this daily statistic in air quality assessment.

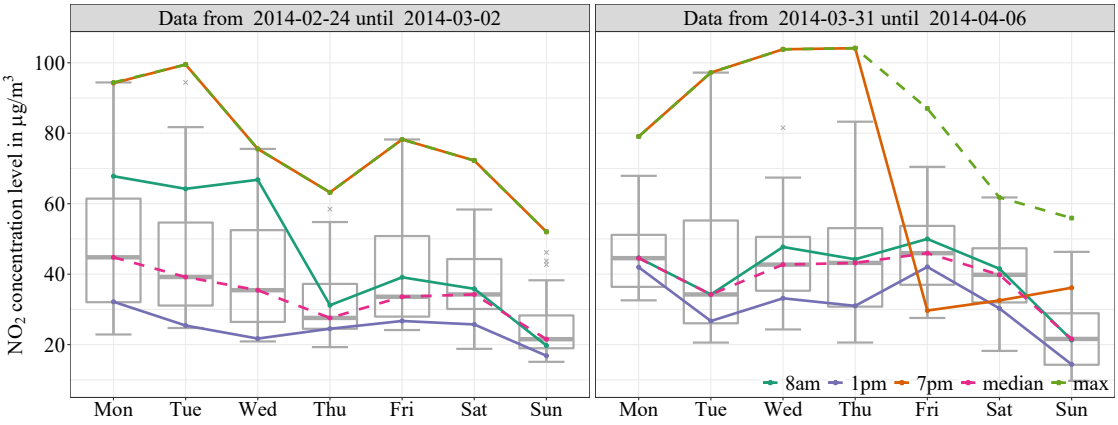


Figure 3.5: Boxplots over NO_2 values for each day of the respective week; dashed and solid lines indicate weekly course of the daily median and maximum values and of the 8am, 1pm, and 7pm values, respectively.

Fig. 3.6 displays the empirical density curves for the 8am, 1pm, and 7pm series indicating remarkable differences in the properties of the three time series. Whereas the values of the 1pm series exhibit a narrow and almost symmetrical distribution, the variation of the 8am series is larger and we observe a tendency to produce more extreme values. This tendency

peaks for the 7pm series, which exhibits by far the largest variation and a pronounced skewness to the right.

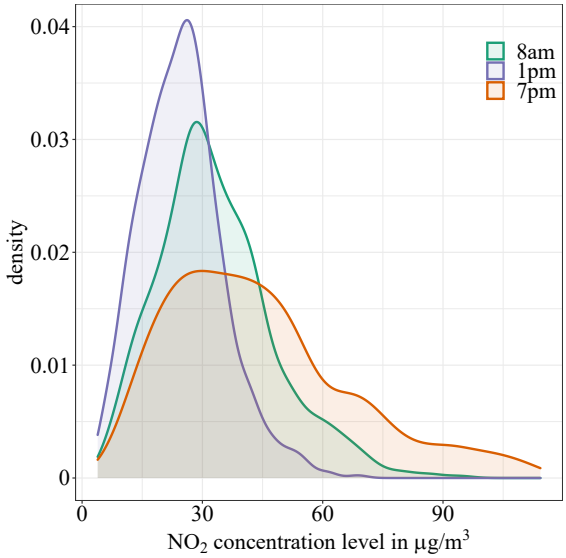


Figure 3.6: Empirical density curves of the daily NO_2 time series referring to 8am, 1pm, and 7pm.

A more detailed picture of the empirical densities is given by Fig. 3.7, where the density curves are estimated for every quarter of each of the two years under consideration. For all three series, the empirical densities vary over the quarters with respect to their mean, variance, and skewness. The 7pm series appears to be less stable over time than the 8am and 1pm series. In particular, the empirical density curve of the 7pm concentration appears to be most volatile over quarters. For our data we do not observe relevant annual changes and hence we will focus on the daily and weekly seasonality.

The exploratory analysis reveals the presence of a pronounced daily and weekly seasonality in the data with slightly changing features over time. This suggests the application of time series modeling techniques capable of dealing with nonstationarities due to complex seasonal patterns. We illustrate that the hour of day reference plays an important role in short-term prediction in order to protect citizens from harmful levels of pollutants. Hence, we will not aggregate from hourly data to daily statistics, e.g. mean, median, or maximum, as the latter statistics do not provide a complete and representative picture of the practically most relevant daily distribution of air pollutants. Therefore, the hour of day reference is maintained when considering hourly models and daily models built upon time series for specific hours, in our exposition exemplarily for 8am, 1pm, and 7pm.

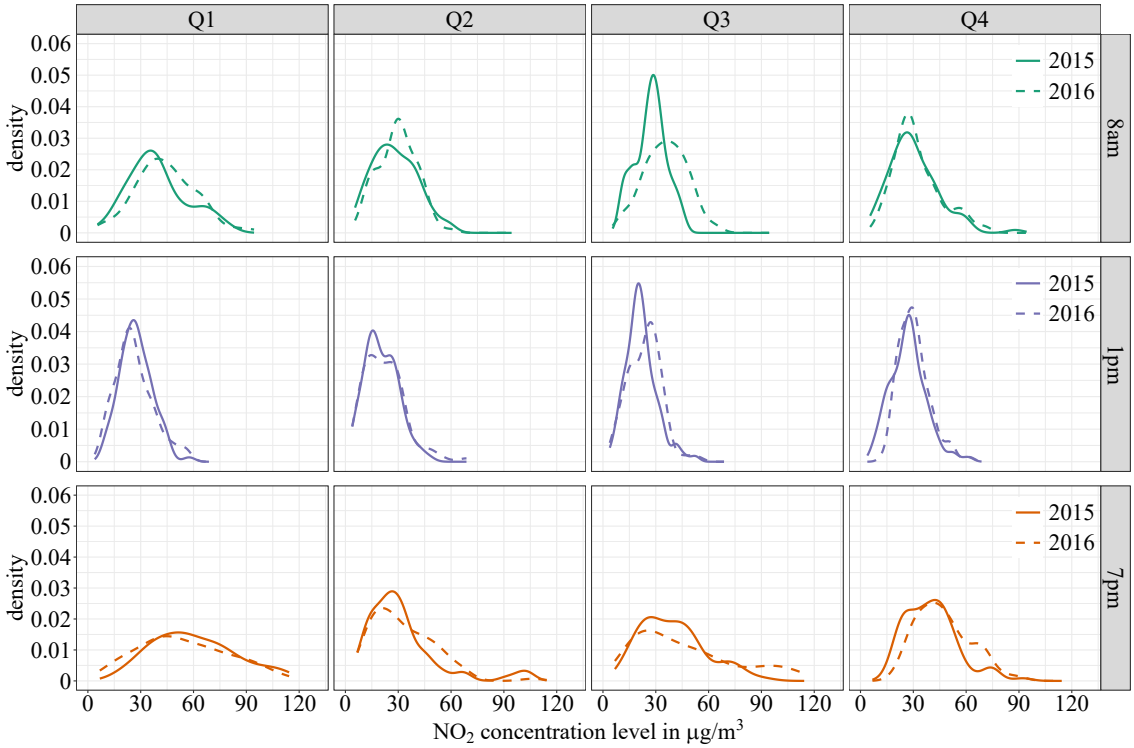


Figure 3.7: Conditional empirical density curves for all eight quarters in 2014 and 2015 of the daily series referring to 8am (top), 1pm (middle), and 7pm (bottom).

According to Figs. 3.5-3.7, the daily series differ with respect to their time series properties, exhibiting non-stationarities of varying degrees and changes in the seasonal figure and the amplitude over time. Due to the observed differences, it seems reasonable to take hourly and daily models into account and to assess whether, and specifically for which hours of the day, the daily models enhance the predictability with respect to predicting the hourly NO_2 concentration level for a specific hour and prediction horizon.

3.3 Methods

The statistical problems arising from the pronounced anthropogenic structures inherent in air pollutant time series bear a resemblance to other areas of research such as the modeling of (local) electricity demand. Examples of resulting modeling tasks are multiple seasonalities, trends, nonlinearities, and associations to related phenomena such as meteorological processes. As a consequence, similar approaches for modeling complex time series can be applied (e.g., Arora and Taylor, 2018; Taylor, 2003, 2010; Taylor and McSharry, 2007). Basically, there are two ways to handle the prediction task in presence of seasonalities (see,

e.g., Bell and Hillmer, 1984, or Harvey, 1989, p. 29). One way is to decompose the time series into seasonal and non-seasonal components which entails three steps with respect to prediction, i.e., de-seasonalizing, prediction, and re-seasonalizing. This approach relies on the crucial assumptions of additivity and orthogonality of the seasonal and non-seasonal components (see Proietti and Riani, 2009). The other way is to incorporate the seasonal patterns in the time series model. In the present work, we consider one modeling approach based on additive decomposition and two approaches that incorporate seasonal structures directly.

We consider observed time series $y_{t(\iota,d,m)}$, $t = 1, \ldots, T$, where ι refers to hour ($\iota = 1, \ldots, 24$, corresponding to 1am,...,12pm), d refers to day, and m refers to month, reflecting the different cycles typically present in NO_2 concentration data. As we will not model annual seasonal effects, all models allow for ARIMA structures to remedy the potential problems of serial correlation in the error process. In our empirical analysis, we study hourly time series y_t, $t = 1, \ldots, T$, and daily time series $y_{t(\iota)}$, $t = 1(\iota), \ldots, T(\iota)$, for a given hour ι. For ease of exposition, we will omit ι whenever the context is clear. We assume a simple but flexible model structure where y_t depends on trend T_t, season S_t, and perturbation R_t. A well known member of this model class is the univariate unobserved components model

$$y_t = h(T_t, S_t, R_t). \tag{3.1}$$

This model is most frequently considered in its additive form $h(T_t, S_t, R_t) = T_t + S_t + R_t$ and R_t modeled as an ARIMA process. The model can be extended to include a cyclical component and exogenous variables (see e.g., Young et al., 1999). There is a huge literature on models of this type, allowing for different mixtures of deterministic and stochastic model components, degrees of flexibility, and computational burdens. The three modeling approaches used in this paper follow the assumption that a linear decomposition of Eq. (3.1) exists, but allow for some nonlinearities by considering the transformed response variable $y_t^{(\lambda)}$, with Box-Cox parameter λ, where

$$y_t^{(\lambda)} = \begin{cases} \frac{y_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(y_t), & \lambda = 0. \end{cases} \tag{3.2}$$

As a consequence, we study regression models where we assume that $y_t^{(\lambda)}$ can be approxi-

mated by functions of trend and seasonal components, such that a dynamically complete specification arises, i.e. a remainder process e_t can be assumed to be Gaussian white noise. Our baseline predictions are generated using seasonal autoregressive integrated moving average (SARIMA) models under the assumption of unchanging seasonal and non-seasonal parameters over time (e.g., Taylor and McSharry, 2007) (see Section 3.1). Second, under the same assumption, we employ harmonic regression models with ARIMA errors (denoted by HarmReg), which rely on decomposing the time series into a seasonal and a non-seasonal component (see Section 3.2). Third, allowing for time varying patterns in all components, we employ the TBATS model of De Livera et al. (2011), which incorporates trigonometric seasonality, Box-Cox transformation, ARIMA errors, trend and seasonal components into a state space framework. Related approaches can be found in Taylor (2003) or Gould et al. (2008) (see Section 3.3). Predictability is analyzed using the cross-validated out-of-sample prediction errors $\hat{e}_{t+h}$ to calculate the RMSE (see Section 3.4) for 75 models (i.e. three hourly and $3 \cdot 24 = 72$ daily models), for each monitoring site in our sample.

3.3.1 Seasonal ARIMA model

The extension of classical ARIMA models to allow for seasonal cycles of various length belongs to the toolbox of many applied researchers and is a frequently employed benchmark in various fields of application, despite the cumbersome structure of seasonal lag polynomials. We employ the formulation of Hyndman and Khandakar (2008) and definition (3.2), and consider a $SARIMA(p,d,q)(P,D,Q)_m$ model given by

$$\Phi(L^m)\phi(L)(1-L^m)^D(1-L)^d y_t^{(\lambda)} = \Theta(L^m)\theta(L)e_t, \tag{3.3}$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ are AR and MA lag polynomials of order p and q, respectively, $\Phi(L^m) = 1 - \Phi_1 L^m - \ldots - \Phi_P L^{P \cdot m}$ and $\Theta(L) = 1 + \Theta_1 L^m + \ldots + \Theta_Q L^{Q \cdot m}$ denote the seasonal AR and MA lag polynomials of order P and Q, respectively, and $e_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ is Gaussian white noise.

The estimation of a $SARIMA(p,d,q)(P,D,Q)_m$ requires to solve several interrelated problems: first, identifying the lag lengths p and q and the relevant mode of filtering for stationarity d, second, estimating the parameters in $\Phi(L^m)$ and $\Theta(L)$, and third, test-

ing model validity. In our application, we use the R-function `Arima()` and select optimal orders according to the AIC (Akaike Information Criterion) due to its asymptotic efficiency and asymptotic equivalence to the final prediction error criterion, which results in testing 324 combinations ($p, q, P, Q \in \{0, 1, 2\}$ and $d, D \in \{0, 1\}$).
The estimated model, exemplarily for a $SARIMA(1,0,1)(0,1,1)_{24}$ model, is given by

$$y_t^{(\widehat{\lambda})} = \widehat{\phi}_1 y_{t-1}^{(\widehat{\lambda})} + y_{t-24}^{(\widehat{\lambda})} - \widehat{\phi}_1 y_{t-25}^{(\widehat{\lambda})} + e_t + \widehat{\theta}_1 e_{t-1} + \widehat{\Theta}_1 e_{t-24} + \widehat{\theta}_1 \widehat{\Theta}_1 e_{t-25}. \tag{3.4}$$

The corresponding equations for a h-step ahead prediction are given by

$$\widehat{y}_{T+h|T}^{(\lambda)} = \begin{cases} \widehat{\phi}_1 y_T^{(\widehat{\lambda})} + y_{T-23}^{(\widehat{\lambda})} - \widehat{\phi}_1 y_{T-24}^{(\widehat{\lambda})} + \widehat{\theta}_1 e_T & \\ \qquad + \widehat{\Theta}_1 e_{T-23} + \widehat{\theta}_1 \widehat{\Theta}_1 e_{T-24}, & h = 1, \\ \widehat{\phi}_1 \widehat{y}_{T+h-1|T}^{(\widehat{\lambda})} + y_{T+h-24}^{(\widehat{\lambda})} - \widehat{\phi}_1 y_{T+h-25}^{(\widehat{\lambda})} & \\ \qquad + \widehat{\Theta}_1 e_{T+h-24} + \widehat{\theta}_1 \widehat{\Theta}_1 e_{T+h-25}, & 1 < h \leq 24, \\ \widehat{\phi}_1 \widehat{y}_{T+24|T}^{(\widehat{\lambda})} + \widehat{y}_{T+1|T}^{(\widehat{\lambda})} - \widehat{\phi}_1 y_T^{(\widehat{\lambda})} + \widehat{\theta}_1 \widehat{\Theta}_1 e_T, & h = 25, \\ \widehat{\phi}_1 \widehat{y}_{T+h-1|T}^{(\widehat{\lambda})} + \widehat{y}_{T+h-24|T}^{(\widehat{\lambda})} - \widehat{\phi}_1 \widehat{y}_{T+h-25|T}^{(\widehat{\lambda})}, & h > 25. \end{cases} \tag{3.5}$$

Note that the prediction equations for $1 < h \leq 24$ can also be written as

$$\begin{aligned} \widehat{y}_{T+h|T}^{(\lambda)} =& \widehat{\phi}_1^h y_T^{(\widehat{\lambda})} + y_{T-(24-h)}^{(\widehat{\lambda})} - \widehat{\phi}_1^h y_{T-24}^{(\widehat{\lambda})} + \widehat{\phi}_1^{h-1} \widehat{\theta}_1 e_T + \widehat{\Theta}_1 e_{T-(24-h)} + \widehat{\phi}_1^{h-1} \widehat{\Theta}_1 \widehat{\theta}_1 e_{T-24} \\ &+ \sum_{j=1}^{h-1} \widehat{\phi}_1^{h-1-j} (\widehat{\phi}_1 + \widehat{\theta}_1) \widehat{\Theta}_1 e_{T-(24-j)}. \end{aligned} \tag{3.6}$$

Applying the inverse Box-Cox transformation of Wooldridge (1992) to the prediction yields

$$\widehat{y}_{T+h|T} = \begin{cases} \left(\widehat{\lambda} y_{T+h|T}^{(\widehat{\lambda})} + 1 \right)^{1/\widehat{\lambda}}, & \widehat{\lambda} \neq 0, \\ \exp\left(y_{T+h|T}^{(\widehat{\lambda})} \right), & \widehat{\lambda} = 0. \end{cases} \tag{3.7}$$

To apply the SARIMA model stated in Eq.(3.3) to $y_{t(\iota)}$, substitute y_t by $y_{t(\iota)}$ and set $m = 7$.

3.3.2 Harmonic regression model with ARIMA errors

For this model class we assume that y_t is additively separable and can be decomposed into a seasonal component S_t, that corresponds to a sum of seasonal terms $s_{j,t}$, $j = 1, \ldots, J$, and a non-seasonal component z_t modeled as an $ARIMA(p, d, q)$ process. Under these assumptions and definition (3.2), a harmonic regression model with ARIMA errors can be stated as

$$y_t^{(\lambda)} = S_t + z_t, \tag{3.8a}$$

$$\phi(L)(1 - L)^d z_t = \theta(L)e_t, \tag{3.8b}$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ are AR and MA lag polynomials of order p and q, respectively, $e_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$ is Gaussian white noise, and the seasonal component in (3.8a) is defined by

$$S_t = \sum_{j=1}^{J} s_{j,t} = \sum_{j=1}^{J} \sum_{k=0}^{k_j} \left(a_{j,k} \cos\left(\frac{2\pi t k}{m_j}\right) + b_{j,k} \sin\left(\frac{2\pi t k}{m_j}\right) \right). \tag{3.9}$$

The seasonal component S_t is assumed to be a periodic process with multiple frequencies and amplitudes. The frequency of $s_{j,t}$ is denoted by m_j and the number of harmonics required for $s_{j,t}$ is denoted by k_j, where $0 < k_j \leq m_j/2$ for even values of m_j, and $0 < k_j \leq (m_j - 1)/2$ for odd values of m_j (e.g., Harvey, 1989, Ch. 3.2, and Shumway and Stoffer, 2017, Ch. 4). The hourly NO_2 time series have two seasonal cycles with frequencies 24 and 168 (i.e., $J = 2$, $m_1 = 24$, and $m_2 = 168$). To apply model (3.8a) and (3.8b) to daily time series $y_{t(\iota)}$, substitute y_t by $y_{t(\iota)}$ and set $J = 1$ and $m_1 = 7$. For estimation, the `R` function `auto.arima()` is applied to hourly data for different combinations of k_1 and k_2 and to daily data for different numbers of k_1 in order to identify the optimal number(s) of harmonics minimizing AIC (see Section 3.1). This includes three steps: first, estimating $\widehat{\lambda}$ for pre-specified k_1 and k_2, second, regressing the transformed response $y_t^{(\widehat{\lambda})}$ on the harmonics $\cos\left(\frac{2\pi t k}{m_j}\right)$ and $\sin\left(\frac{2\pi t k}{m_j}\right)$ $k = 0, \ldots, k_j$, $j = 1, 2$, to obtain the residuals $\widetilde{z}_t$, and third, fitting a family of ARIMA models to the residuals $\widetilde{z}_t$ in order to identify the optimal order in Eq. (3.8b) according to AIC.

The equations for a h-step ahead prediction, exemplarily with $ARIMA(1, 0, 2)$ errors, are

given by

$$\widehat{y}^{(\widehat{\lambda})}_{T+h|T} = \mathrm{E}\left(y^{(\widehat{\lambda})}_{T+h}|\widetilde{z}_T, \widetilde{z}_{T-1}, \ldots, \widetilde{z}_1, \widehat{S}_{T+h}\right) \tag{3.10a}$$

$$= \widehat{S}_{T+h} + \begin{cases} \widehat{\phi}_1 \widetilde{z}_T + \widehat{\theta}_1 e_T + \widehat{\theta}_2 e_{T-1}, & h = 1, \\ \widehat{\phi}_1 \widehat{\widetilde{z}}_{T+1|T} + \widehat{\theta}_2 e_T, & h = 2, \\ \widehat{\phi}_1 \widehat{\widetilde{z}}_{T+h-1|T}, & h > 2. \end{cases} \tag{3.10b}$$

Eq. (3.10a) states that the predicted value corresponds to the expected value conditional on the residuals from the auxiliary regression and the prediction for S_{T+h}, which is, according to Eq. (3.9), obtained by

$$\widehat{S}_{T+h} = \sum_{j=1}^{J} \sum_{k=0}^{k_j} \left(\widehat{a}_{j,k} \cos\left(\frac{2\pi(T+h)k}{m_j}\right) + \widehat{b}_{j,k} \sin\left(\frac{2\pi(T+h)k}{m_j}\right)\right).$$

Note that the procedure described above substantially differs from the dynamic harmonic regression model of Young et al. (1999), which allows for stochastic time-varying parameters in Eq. (3.9) in a state space framework.

3.3.3 TBATS model

De Livera et al. (2011) propose a state space framework to deal with multiple seasonality extending earlier exponential smoothing approaches. Their approach allows for complex (non-integer) seasonality, provides a simple remedy for neglected cycles causing serially correlated errors, and avoids issues with nonlinear exponential smoothing models. The TBATS model is given by

$$y^{(\lambda)}_t = l_{t-1} + \psi b_{t-1} + \sum_{j=1}^{J} s^{(j)}_{t-1} + d_t, \tag{3.11a}$$

$$l_t = l_{t-1} + \alpha d_t, \tag{3.11b}$$

$$b_t = (1-\psi)b + \psi b_{t-1} + \beta d_t, \tag{3.11c}$$

$$s^{(j)}_t = \sum_{k=1}^{k_j} s^{(j)}_{k,t}, \tag{3.11d}$$

$$d_t = \sum_{i=1}^{p} \phi_i d_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t, \tag{3.11e}$$

where the seasonal components in (3.11d) are calculated according to

$$s_{k,t}^{(j)} = s_{k,t-1}^{(j)} \cos\left(\frac{2\pi tk}{m_j}\right) + s_{k,t-1}^{*(j)} \sin\left(\frac{2\pi tk}{m_j}\right) + \gamma_1^{(j)} d_t, \tag{3.12a}$$

$$s_{k,t}^{*(j)} = -s_{k,t-1}^{(j)} \sin\left(\frac{2\pi tk}{m_j}\right) + s_{k,t-1}^{*(j)} \cos\left(\frac{2\pi tk}{m_j}\right) + \gamma_2^{(j)} d_t. \tag{3.12b}$$

Eq. (3.11a) denotes the measurement equation, Eqs. (3.11b) to (3.11d) are the smoothing equations for level, trend, and seasonality, where b denotes the long-term trend, and Eq. (3.11e) incorporates ARIMA structures in the error component, where again $e_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is Gaussian white noise. The smoothness of the seasonal component $s_t^{(j)}$ is controlled by the number of harmonics k_j. For NO_2 data, we observe two seasonalities with frequencies 24 and 168 and have to estimate the Box-Cox parameter λ, the dampening parameter ψ, the smoothing parameters α, β, $\gamma_1^{(j)}$, $\gamma_2^{(j)}$, $j = 1, 2$, the number of harmonics k_1 and k_2, the ARIMA parameters $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$, as well as starting values for the level, for the fourier coefficients $s_{k,t-1}^{j}$ and $s_{k,t-1}^{*(j)}$, $j = 1, 2$, and for the error term. For the application of the TBATS model to $y_{t(\iota)}$, we substitute y_t by $y_{t(\iota)}$ and set $J = 1$ and $m_1 = 7$.

In matrix notation the TBATS model (3.11a) to (3.11e) is given by

$$y_t^{(\lambda)} = \boldsymbol{w}' \boldsymbol{x}_{t-1} + e_t, \tag{3.13a}$$

$$\boldsymbol{x}_t = \boldsymbol{F} \boldsymbol{x}_{t-1} + \boldsymbol{g} e_t. \tag{3.13b}$$

The h-step ahead prediction is given by

$$\widehat{y}_{T+h|T}^{(\lambda)} = \boldsymbol{w}' \boldsymbol{F}^{h-1} \boldsymbol{x}_T, \tag{3.14}$$

and the re-transformation is carried out according to Eq. (3.7).

3.3.4 Procedure for evaluation of predictability

For model evaluation, we conduct a time series out-of-sample cross-validation with extending windows. In each validation loop, the training sample is extended by one time unit; i.e. one hour for hourly (hly) time series and one day for daily (dly) time series, respectively. In order to get reliable initial estimates for further iteration, we split the

data after the first 69 weeks. The resulting training sample then has length 69·168=11592 (69· 7=483) for hourly (daily) models, respectively, which corresponds to approx. 66% (66%) of the data. Given that we consider prediction horizons ranging between 1 and 168 h for hly time series (1 and 7 days for dly time series), we can conduct 5000 hly (240 dly) cross-validation steps, which implies that the testing uses approx. 28.5% (33%) of the data. Note that k-fold cross-validation is not feasible, as all models employed in this study are fully iterative (see Bergmeir et al., 2018). In the following we provide details on the evaluation algorithm.

Hourly time series:

hly 1. Fit the model to an initial training sample $y_{t_0}, \ldots, y_{t_1}$ where $t_0 = 1$ and $t_1 = 69{\cdot}168 = 11592$, i.e. the training sample comprises 69 weeks of data, and denote the estimated model by mod_{t_1}.

hly 2. In the n-th loop, $n = 0, \ldots, N$, $N = 4999$, of the time series cross-validation,

- extend the training sample by n hours;
- update mod_{t_1} for the new training sample and use the updated model to predict one hour to one week ahead, which yields the predicted values $\widehat{y}_{t_1+n+h|t_1+n}$ for each $h = 1, \ldots, 168$;
- calculate the out-of-sample prediction errors $\widehat{e}_{t_1+n+h|t_1+n} = y_{t_1+n+h} - \widehat{y}_{t_1+n+h|t_1+n}$.

hly 3. After 5000 loops, for each prediction horizon h, calculate the out-of-sample RMSE

$$RMSE_h^{hly} = \sqrt{\frac{1}{N+1}\sum_{n=0}^{N} \widehat{e}^2_{t_1+n+h|t_1+n}} \quad . \tag{3.15}$$

hly 4. Filter the estimated prediction errors for hour ι of the respective day and calculate, for each combination of ι and h, the respective RMSE

$$RMSE_{h,\iota}^{hly} = \sqrt{\frac{1}{\lfloor (N+1)/24 \rfloor} \sum_{\substack{n=0 \\ t_1+n+h \text{ refers to } \iota}}^{4992-1} \widehat{e}^2_{t_1+n+h|t_1+n}} \quad , \tag{3.16}$$

where the upper bound is set to 4991 to ensure that the subseries of prediction errors are of equal length for each ι, $\iota = 1, \ldots, 24$.

Daily time series:

dly 1. Fit the model to an initial training sample $y_{t_0(\iota)}, \ldots, y_{t_1(\iota)}$ where $t_0 = 1$ and $t_1 = 69 \cdot 7 = 483$, i.e. the training sample comprises 69 weeks of data, and denote the estimated model by mod_{t_1}.

dly 2. In the n-th loop, $n = 0, \ldots, N$, $N = 239$, of the time series cross-validation,

- extend the training sample by n days;
- update mod_{t_1} for the new training sample and use the updated model to predict one day to one week ahead, which yields $\widehat{y}_{t_1(\iota)+n+h|t_1(\iota)+n}$ for each $h = 1, \ldots, 7$;
- calculate the out-of-sample errors $\widehat{e}_{t_1(\iota)+n+h|t_1(\iota)+n} = y_{t_1(\iota)+n+h} - \widehat{y}_{t_1(\iota)+n+h|t_1(\iota)+n}$.

dly 3. After 240 loops, for each prediction horizon h, calculate the out-of-sample RMSE

$$RMSE_{h,\iota}^{dly} = \sqrt{\frac{1}{N+1}\sum_{n=0}^{N} \widehat{e}^2_{t_1(\iota)+n+h|t_1+n}}. \tag{3.17}$$

The flowchart in Fig. 3.8 graphically summarizes the steps of our empirical analysis.

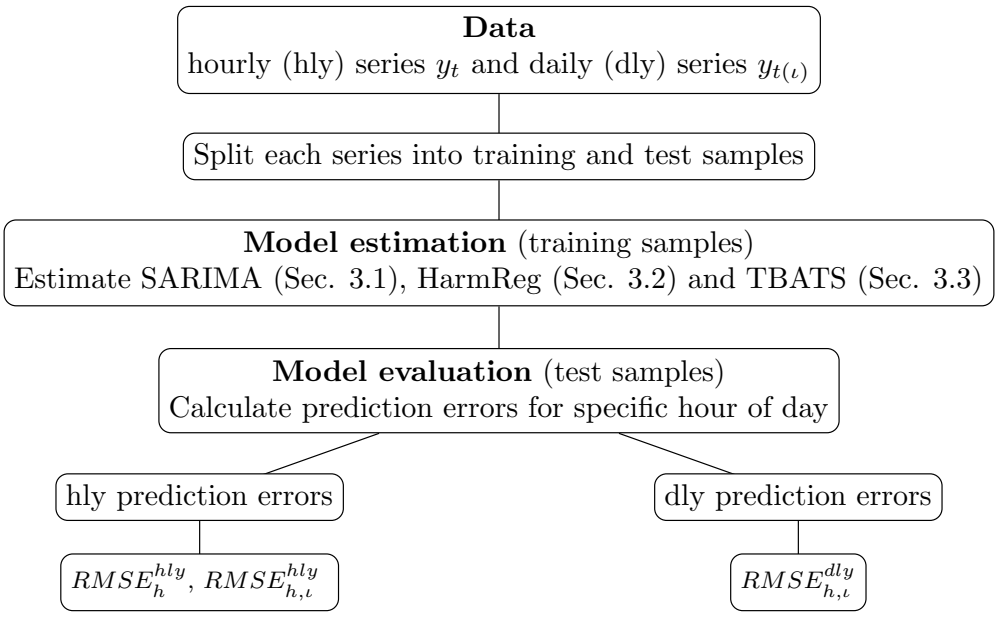


Figure 3.8: Flowchart summarizing the empirical analysis.

3.4 Results and discussion

In this section, we discuss the prediction accuracy of the estimated models using hourly data, y_t, and daily data, $y_{t(\iota)}$, exemplarily for the hours $\iota = 8\text{am}, 1\text{pm}, 7\text{pm}$, as motivated in Section 3.2. Further details are given in the appendix. The evaluation of the models is carried out according to the procedure described in the preceding Section. Note that, due to the estimation procedure of the implemented `R` functions, the fitted values and the residuals refer to one-step ahead predictions and one-step ahead prediction errors, respectively, corresponding to one-hour ahead (using hourly data) and one-day ahead (using daily data) prediction horizons.

As shown in Fig. 3.9, the hourly SARIMA model outperforms the daily SARIMA models for all hours of day concerning the in-sample RMSE values. However, when the focus lies on predicting the air pollutant concentration one-day ahead, it might be reasonable to consider 24-step ahead prediction errors and 1-step ahead prediction errors for the hourly and daily models, respectively, to provide an appropriate comparison between hourly and daily models.

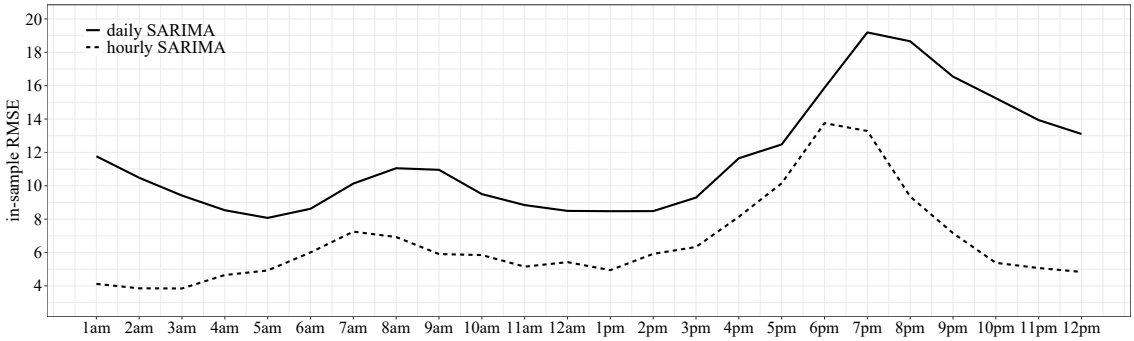


Figure 3.9: In-sample RMSE for each hour of day derived from hourly and daily SARIMA models.

The upper panel of Fig. 3.10 displays, for the last week of the initial training sample and the first week of the initial test sample, the fitted and predicted values derived from the hourly models. Concerning the training period, all models fit the data remarkably well. As SARIMA just incorporates the daily seasonality the pattern in the predicted values repeats every 24 h. Differences in the daily course of the NO_2 concentration levels from weekdays to weekend days can clearly be seen in the predicted values referring to HarmReg and TBATS. The morning and evening modes are less pronounced between $T + 72$ and

$T + 120$ which correspond to Saturday and Sunday. The lower panel of Fig. 3.10 zooms into the upper panel and displays the last day of the initial training sample and the first day of the initial test sample.

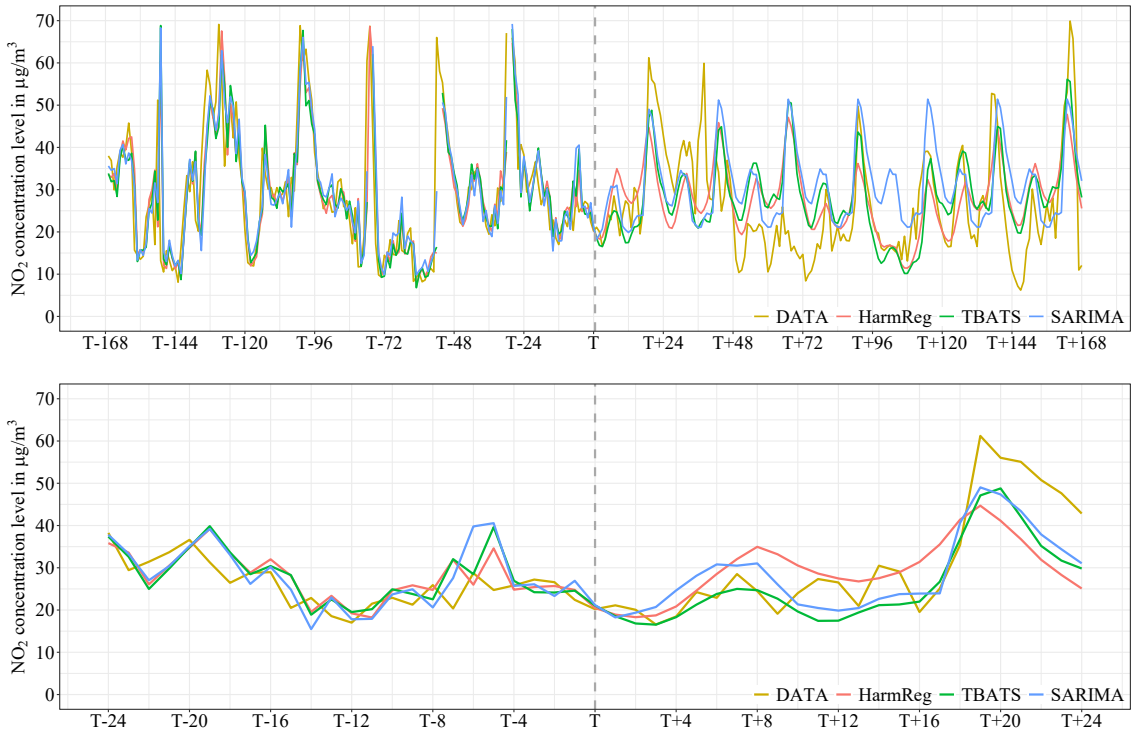


Figure 3.10: Fitted and predicted values derived from hourly models where the upper (lower) panel refers to last week (day) of training sample and first week (day) of test sample; the respective cutoff is marked by the vertical dashed line and the course of observed data is drawn in gold colour.

Fig. 3.11 displays boxplots of the fitted SARIMA model values for each hour of day derived from the hourly (upper panel) and the 24 daily (lower panel) models. Both panels in Fig. 3.11 reveal the bimodal daily pattern. The (interquartile) range, the gap between median and mean value and the number of outliers are higher for the hourly model compared to the daily models, in particular concerning the evening hours.

Analogous plots for the residuals in Fig. 3.12 reveal a tendency that the use of hourly data provides a more accurate fit to the training data in comparison to the use of daily data. Figures such as Fig. 3.11 and Fig. 3.12 for the fitted values and the residuals derived from the other two modeling approaches, HarmReg and TBATS, provide an analogous conclusion.

Fig. 3.13 shows the out-of-sample RMSE values for a prediction horizon from 1 to 24 h

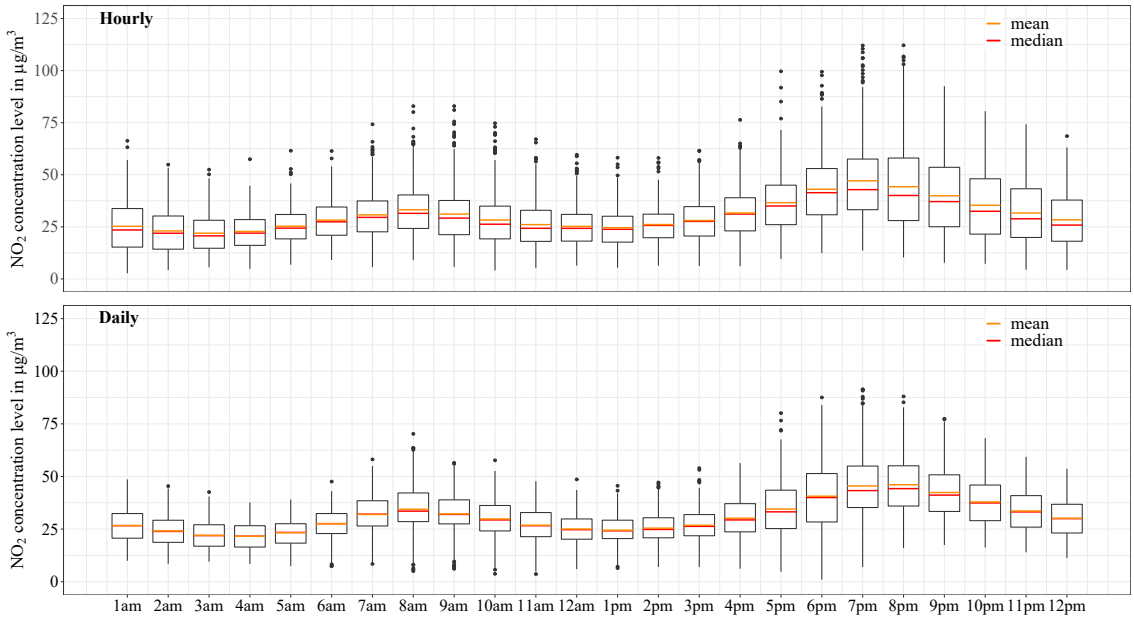


Figure 3.11: Boxplots of fitted values for each hour of day derived from the hourly (upper panel) and daily (lower panel) SARIMA models; the orange and red horizontal lines indicate the mean and median of the fitted values, respectively.

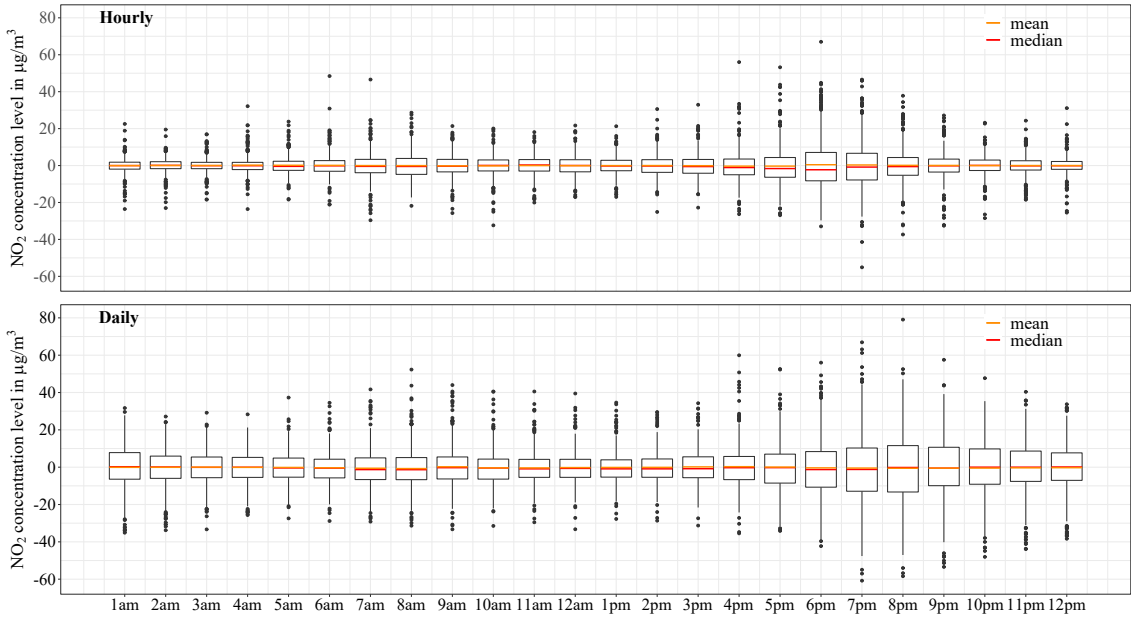


Figure 3.12: Boxplots of residuals for each hour of day derived from the hourly (upper panel) and daily (lower panel) SARIMA models; the orange and red horizontal lines indicate the mean and median of the residuals, respectively.

calculated according to Eq. (3.15) and reveals that SARIMA outperforms the other two models with respect to the RMSE values built over all prediction errors.

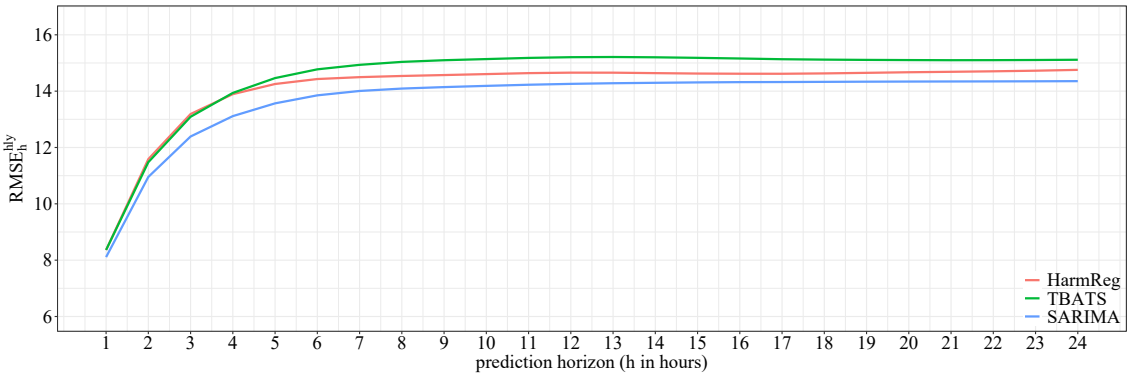


Figure 3.13: RMSE of hourly models in dependence of prediction horizon (compare Eq. (3.15)).

According to Fig. 3.14, there is no remarkable change in the out-of-sample RMSE values concerning a prediction horizon from 1 to 7 days. SARIMA still outperforms the other two models with respect to the RMSE values built over all prediction errors.

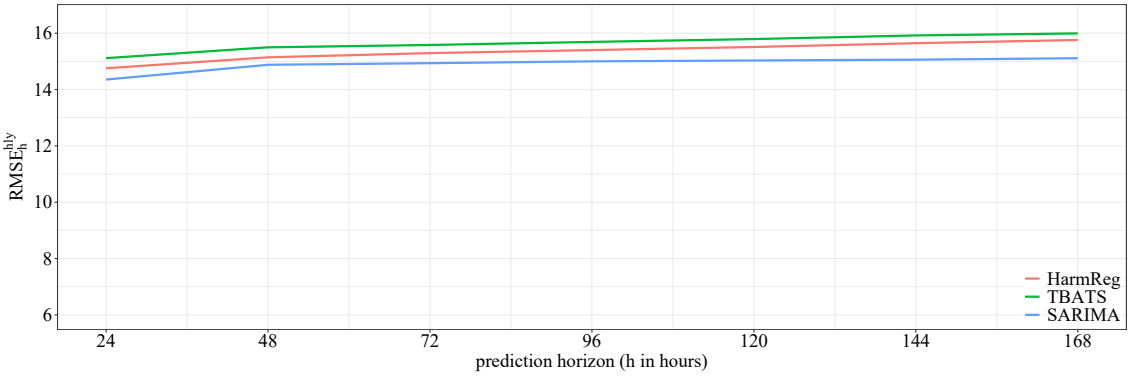


Figure 3.14: RMSE of hourly models in dependence of prediction horizon (compare Eq. (3.15)).

Regarding the panels in Fig. 3.15, a different conclusion is drawn. The prediction performance of each model depends on the hour of day to be predicted. In terms of RMSE for 8am (upper panel of Fig. 3.15), HarmReg outperforms TBATS and SARIMA for $h > 4$. Concerning 1pm (middle panel of Fig. 3.15), HarmReg slightly outperforms SARIMA and TBATS for $h > 8$. According to the lower panel of Fig. 3.15, the 7pm values are best predicted by SARIMA. There are remarkable differences in the RMSE values for different hours of day, which might be due the various degrees of variability and complexity in the daily series $y_{t(\iota)}$ (compare Fig. 3.6 and Fig. 3.7).

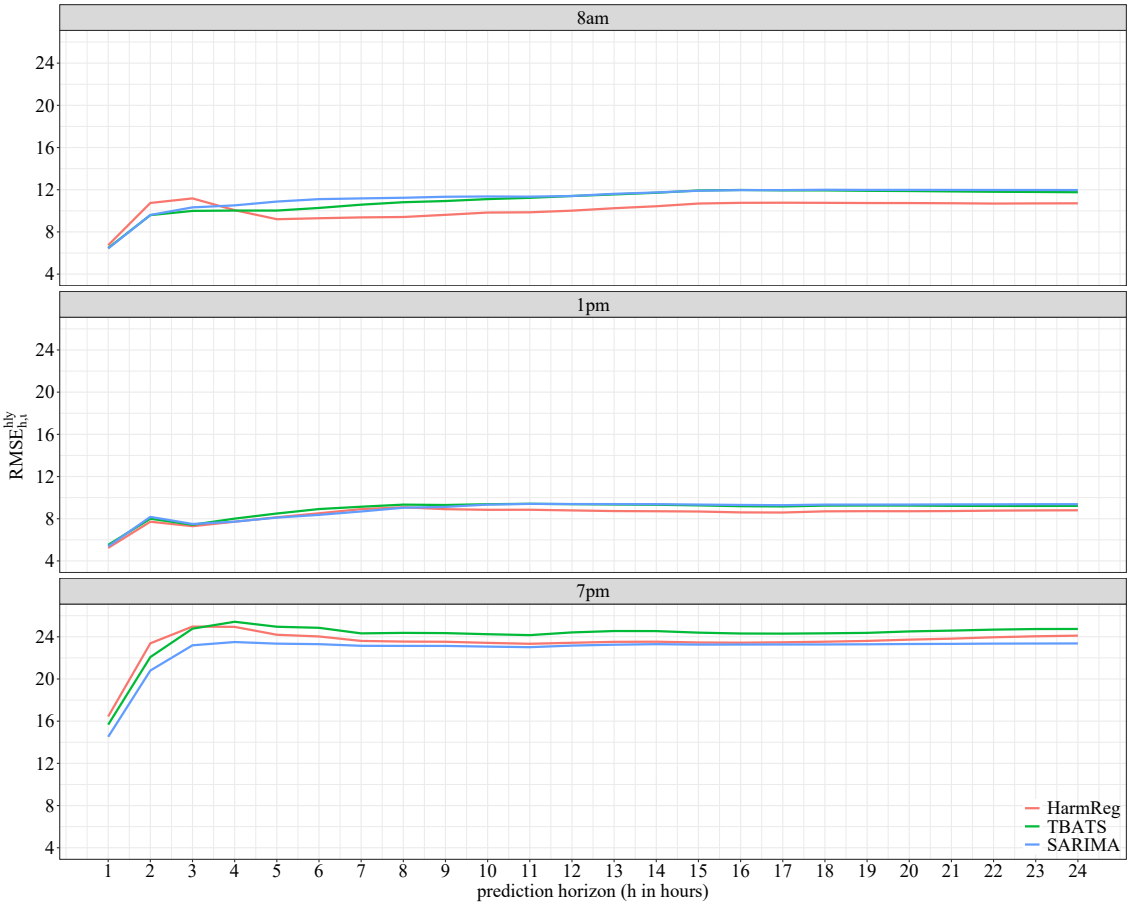


Figure 3.15: RMSE of hourly models in dependence of prediction horizon and hour of day to be predicted (compare Eq. (3.16)).

A comparison of the model performance in terms of the aggregation level of the data is shown in Fig. 3.16, where the RMSE for prediction horizons $h = 1, \ldots, 7$ is depicted for the daily and hourly models. The formulas to obtain the RMSE values correspond to those stated in Eqs. (3.16) and (3.17). Note that the scale of the ordinate in the panels of Fig. 3.16 is fixed in order to ease model comparison with respect to the predictability of the hours of day. The left and right panel suggest to build models based on daily series, whereas the middle panel shows the lowest RMSE values for the hourly HarmReg model. In the right panel, in particular, RMSE values are remarkably lower for models using daily series.

We conclude our discussion with Tables 3.3 and 3.4, summarizing the previous results and contrasting them for all four monitoring sites in this study.

For a prediction horizon of 1, 24 and 168 h(s), the overall RMSE for each hourly model is reported in Table 3.3. There are no remarkable differences in the RMSE values between

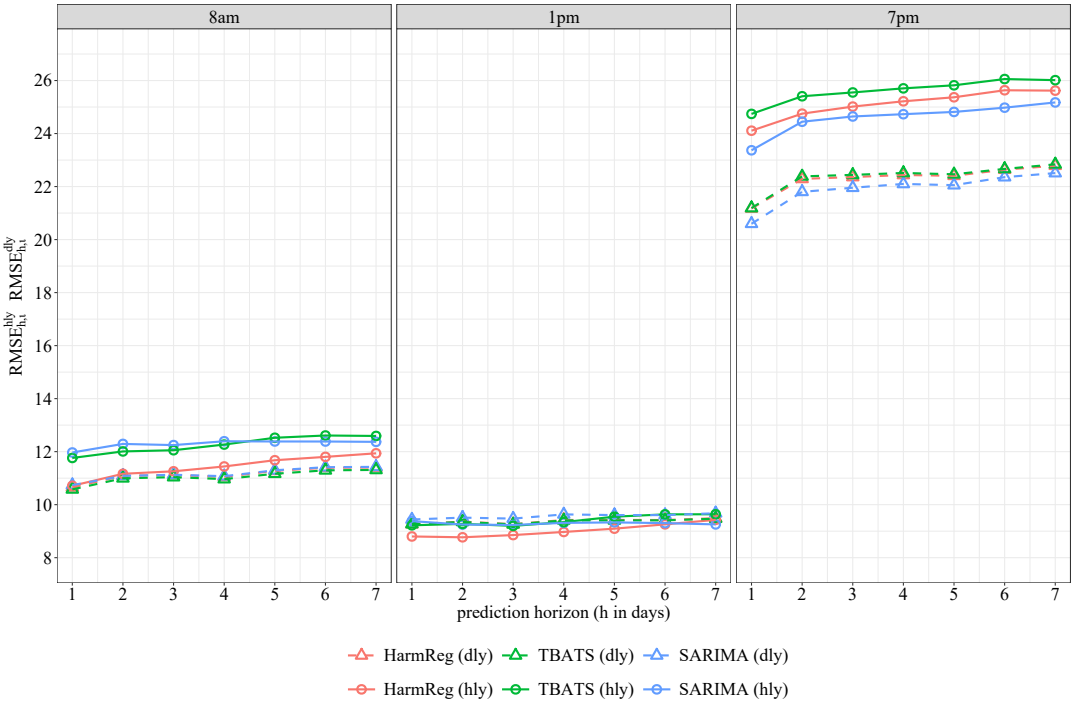


Figure 3.16: RMSE of daily and hourly models in dependence of prediction horizon and hour of day to be predicted (compare Eqs. (3.16) and (3.17)).

the three modeling techniques. The RMSE increases with the prediction horizon up to 24 h ahead and changes only slightly afterwards. This can be explained as follows: The underlying time series is stationary at every position (hour) s of the seasonal cycle. Hence, with increasing prediction horizon h beyond $h = 24$, for a prediction

$$\widehat{y}_{T+h|T}, \quad h = 25, 26, \ldots$$

only the predictive power of the long-term mean of the time series (at every position s) remains and the prediction error converges to the long-run variance of the underlying process. Note that this prediction only has information up to time period T but not about the latest seasonal cycle (i.e., $y_{T+1}, \ldots, y_{T+24}$). Theoretically, and given a consistent estimate of the long-term mean, the predictive power of the long-term mean should not deteriorate (substantially) as long as the properties of the underlying process do not change (substantially) (see Brockwell and Davis, 1991, Sec. 5.2).

Exemplarily for SARIMA and a prediction horizon of 24 h (1 day) and 168 h (7 days), the RMSE in dependence of the hour of day to be predicted is shown in Table 3.4. For SARIMA (as well as HarmReg and TBATS) and several hours we observe a tendency that

		Prediction horizon		
City	Method	1 h	24 h	168 h
Passau	SARIMA	8.11	14.36	15.11
	HarmReg	8.36	14.76	15.76
	TBATS	8.36	15.11	15.99
Regensburg	SARIMA	9.78	19.01	19.59
	HarmReg	9.85	18.95	20.77
	TBATS	10.07	20.90	22.18
Landshut	SARIMA	6.62	11.50	11.78
	HarmReg	6.85	11.90	12.57
	TBATS	6.60	11.38	11.63
Burghausen	SARIMA	6.91	10.68	11.31
	HarmReg	6.96	10.61	11.73
	TBATS	6.81	10.56	11.72

Table 3.3: RMSE of hourly models for all monitoring sites and different prediction horizons (compare Eq. (3.15)).

		Prediction horizon: hly		Prediction horizon: dly	
City	Hour of day	24 h	168 h	1 day	7 days
Passau	8am	11.97	12.37	10.74	11.43
	1pm	9.38	9.26	9.44	9.68
	7pm	23.37	25.17	20.60	22.51
Regensburg	8am	17.19	17.64	16.11	16.36
	1pm	13.39	13.71	12.35	12.83
	7pm	30.64	32.21	26.28	27.35
Landshut	8am	12.24	12.17	10.79	10.58
	1pm	9.82	10.40	9.54	10.05
	7pm	13.64	13.97	14.92	15.02
Burghausen	8am	8.89	9.33	9.74	10.10
	1pm	7.00	7.36	6.68	7.13
	7pm	15.86	17.01	12.23	13.05

Table 3.4: RMSE of daily and hourly models for all monitoring sites for SARIMA model and different prediction horizons and hours of day to be predicted (compare Eqs. (3.16) and (3.17)).

the use of daily data reduces RMSE in comparison to the use of hourly data. Further results are available from the authors upon request.

3.5 Conclusions

Predictions of air pollutant concentrations play an important role in protecting the population from potential adverse health effects. The development of appropriate prediction models is challenging as temporal air quality processes exhibit multiple seasonal patterns and non-stationarities. There are many studies on predicting air quality taking into account one seasonality but few on multi-seasonal prediction models. Furthermore, model evaluation is generally based on all prediction errors without reference to the hour to be predicted, e.g. time of day. We propose a framework to assess the predictability of local concentration levels of NO_2 (or other air pollutants) based on hourly measurements generated by local monitoring sites. Our analysis maintains the hour of day reference and considers hourly time series and daily series (for each hour) and every monitoring site. Predictability is assessed via loss functions based on out-of-sample h-step prediction errors. Using time series cross-validation, we produce out-of-sample predictions for NO_2 concentration levels for each hour of day, with prediction horizons h ranging from one hour to one week. In an empirical analysis, we apply state-of-the art multi-seasonal univariate prediction models to hourly NO_2 data from four monitoring sites in Germany. We thoroughly discuss the relation between predictability and hour of day and respective NO_2 concentration level separately for each hour of day to obtain a comprehensive picture of the model performance. We find that prediction accuracy strongly depends on the hour of day to be predicted. For hours that tend to exhibit relatively high pollutant levels the predictability of both, hourly and daily models, deteriorates. For some hours, a prediction gain can be achieved by building the models on daily instead of hourly data. Similar findings for the other monitoring sites suggest at least some regional robustness of our results. In future extensions of this work, further performance criteria can be investigated, such as a hit or success rate for correct forecasts of the hour of maximal exposure. The use of univariate methods has the advantage of being independent of additional and potentially costly predictors. On the other hand, the inclusion of covariates, e.g. information on weather or traffic, may improve the predictability in general and of hours of day with high exposure in particular.

3.6 Acknowledgements

We thank two anonymous referees, an associate editor, Marek Brabec, Phillip Sibbertsen, James Taylor, Matthias Wild, and the participants of The International Society for Ecological Modelling Global Conference 2019 in Salzbug/Austria and the participants of the Workshop on Statistical Modelling of Environmental Data 2019 in Passau/Germany for helpful comments and suggestions. All remaining errors are ours.

3.7 Appendix A

This appendix provides details on estimation setups and results for models based on hourly and daily data recorded in 2014 and 2015 at the monitoring site in Passau. Table 3.A.1 gives an overview over the meta parameters of the estimated hourly and daily models. We focus on details for HarmReg and TBATS, further results are available upon request.

Table 3.A.1: Overview over the number of harmonics and the order of ARIMA, SARIMA and ARMA processes of the estimated hourly and daily models.

Data	Method	Number of harmonics and/or order of ARIMA, SARIMA or ARMA process
y_t	HarmReg	$k_1 = 11$, $k_2 = 13$, $p = 5$, $d = 1$, $q = 1$
	SARIMA	$p = 1$, $d = 0$, $q = 1$, $P = 1$, $D = 1$, $Q = 1$
	TBATS	$k_1 = 5$, $k_2 = 5$, $p = 0$, $q = 0$
$y_{t(8am)}$	HarmReg	$k_1 = 3$, $p = 1$, $d = 1$, $q = 2$
	SARIMA	$p = 1$, $d = 1$, $q = 1$, $P = 1$, $D = 1$, $Q = 1$
	TBATS	$k_1 = 3$, $p = 1$, $q = 0$
$y_{t(1pm)}$	HarmReg	$k_1 = 3$, $p = 1$, $d = 1$, $q = 1$
	SARIMA	$p = 0$, $d = 1$, $q = 1$, $P = 0$, $D = 1$, $Q = 1$
	TBATS	$k_1 = 3$, $p = 0$, $q = 0$
$y_{t(7pm)}$	HarmReg	$k_1 = 2$, $p = 0$, $d = 1$, $q = 2$
	SARIMA	$p = 0$, $d = 1$, $q = 2$, $P = 0$, $D = 1$, $Q = 1$
	TBATS	$k_1 = 3$, $p = 0$, $q = 1$

Details HarmReg

We have applied the function `auto.arima()` to the training sample y_t for 192 different combinations of k_1 and k_2 with $k_1 = 1, \ldots, 12$ and $k_2 = 1, \ldots, 16$ to identify the optimal number of harmonics in HarmReg. Fig. 3.A.1 displays how the deciles of the resulting AIC values are distributed over the combinations of k_1 and k_2 where the combination

$k_1 = 11$ and $k_2 \in \{13, 14\}$ returns the minimum AIC value. The residuals derived from the harmonic regression with $k_1 = 11$ and $k_2 = 13$ harmonics, representing daily and weekly seasonal patterns, respectively, follow an $ARIMA(5, 1, 1)$-process.

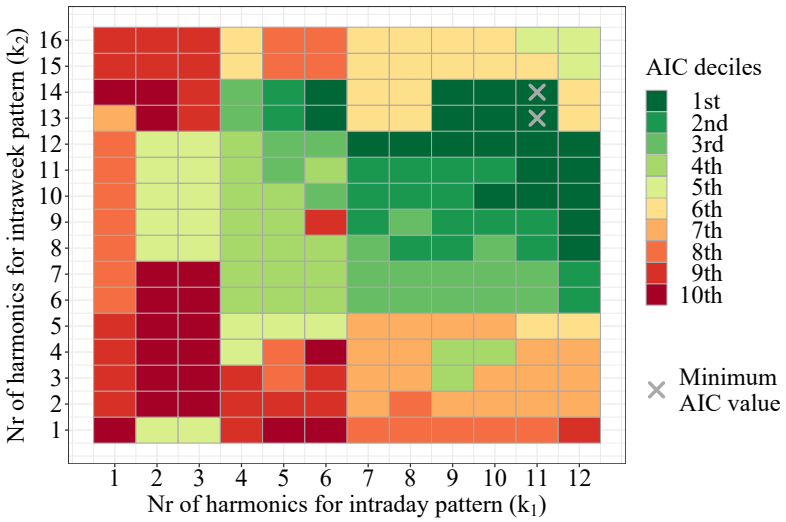


Figure 3.A.1: Distribution of the AIC deciles over different combinations of k_1 and k_2 in HarmReg obtained by applying `auto.arima()` with fourier terms as regressors to the training sample.

For the daily models the identification of the optimal number of harmonics is less cumbersome since the weekly seasonality implies $m_1 = 7$ and thus $0 < k_1 \leq 3$ (see subsection 3.3.2).

Details TBATS

An extract from the output of the estimated TBATS model for the hourly data is given by

```
TBATS(0.207, {0,0}, 0.801, {<24,5>, <168,5>})
  Lambda: 0.206589
  Alpha: 1.204514
  Beta: -0.2897133
  Damping Parameter: 0.80129
  Gamma-1 Values: 0.004004553 -0.001977883
  Gamma-2 Values: 0.001286855 0.0001531337
```

The estimated number of harmonics is five for both, the daily and weekly seasonal component, the Box-Cox and damping parameter equals 0.207 and 0.801, respectively, and the estimated error follows a white noise process. The smoothing parameters of the seasonal components are quite small, the smoothing parameter for the level and trend equation

is equal to 1.205 and -0.290, respectively. Fig. 3.A.2 shows for the first five weeks of the initial training sample the components of the hourly TBATS model. The component *observed* refers to the Box-Cox transformed data that are decomposed into *level*, *slope*, *season1* (daily seasonality), *season2* (weekly seasonality), and *residuals*; *season* refers to the sum of *season1* and *season2*. Note that the first day of the data under consideration, 1st January 2014, is a Wednesday and the drops in *season2* refer to the weekends. The component *level* changes much faster than the seasonal components which is in accordance to the values of the corresponding smoothing parameters. Further, we observe that the value range of *season* is about a fifth of the value range of *level*. We may understand *level* as the baseline NO_2 exposure and *season* as the variation due to the day of the week and the hour of day.

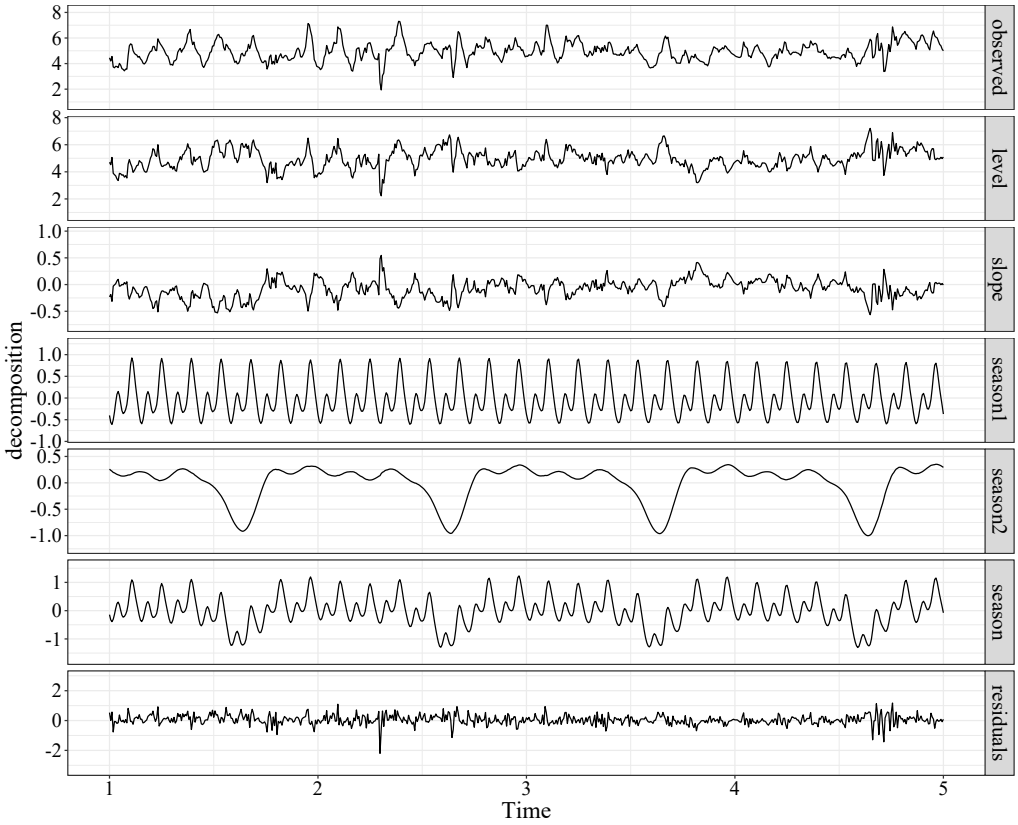


Figure 3.A.2: Components of the hourly TBATS model for the first five weeks of the initial training sample.

The estimated daily TBATS models for 8am, 1pm, and 7pm are `TBATS(0.108, 1,0, -, <7,3>)`, `TBATS(0.512, 0,0, -, <7,3>)` and `TBATS(0.234, 0,1, -, <7,3>)`, respectively.

3.8 References

Abdullah, S., Nasir, N.H.A., Ismail, M., Ahmed, A.N., Jarkoni, M.N.K., 2019. Development of ozone prediction model in urban area. Internation Journal of Innovative Technology and Exploring Engineering 8, 2263–2267. doi:10.35940/ijitee.J1127.0881019.

Arora, S., Taylor, J.W., 2018. Rule-based autoregressive moving average models for forecasting load on special days: A case study for France. European Journal of Operational Research 266, 259–268. doi:10.1016/j.ejor.2017.08.056.

Atkinson, R.W., Butland, B.K., Anderson, H.R., Maynard, R.L., 2018. Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies. Epidemiology 29, 460–472. doi:10.1097/EDE.0000000000000847.

Bell, W.R., Hillmer, S.C., 1984. Issues involved with the seasonal adjustment of economic time series. Journal of Business & Economic Statistics 2, 291–320. doi:10.2307/1391266.

Bergmeir, C., Hyndman, R.J., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis 120, 70–83. doi:10.1016/j.csda.2017.11.003.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. Springer. URL: https://link.springer.com/book/10.1007%2F978-1-4419-0320-4.

Cabaneros, S.M., Calautit, J.K., Hughes, B., 2020. Spatial estimation of outdoor NO_2 levels in Central London using deep neural networks and a wavelet decomposition technique. Ecological Modelling 424, 109017. doi:10.1016/j.ecolmodel.2020.109017.

Cabaneros, S.M., Calautit, J.K., Hughes, B.R., 2019. A review of artificial neural network models for ambient air pollution prediction. Environmental Modelling & Software 119, 285–304. doi:10.1016/j.envsoft.2019.06.014.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., Li, Y., 2018. Brits: Bidirectional recurrent imputation for time series. Advances in Neural Information Processing Systems , 6775–6785.

Council of the European Union, 2008. Directive 2008/50/EC on ambient air quality and cleaner air for Europe. Official Journal of the European Communities .

De Livera, A.M., Hyndman, R.J., Snyder, R.D., 2011. Forecasting time series with complex

seasonal patterns using exponential smoothing. Journal of the American Statistical Association 106, 1513–1527. doi:10.1198/jasa.2011.tm09771.

DeGaetano, A.T., Doherty, O.M., 2004. Temporal, spatial and meteorological variations in hourly $PM_{2.5}$ concentration extremes in New York City. Atmospheric Environment 38, 1547–1558. doi:10.1016/j.atmosenv.2003.12.020.

Dowle, M., Srinivasan, A., 2019. data.table: Extension of 'data.frame'. URL: https://CRAN.R-project.org/package=data.table. r package version 1.12.2.

EEA, 2018. Air quality in Europe – 2018 report. doi:10.2800/777411.

Gocheva-Ilieva, S.G., Ivanov, A.V., Voynikova, D.S., Boyadzhiev, D.T., 2014. Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach. Stochastic Environmental Research and Risk Assessment 28, 1045–1060. doi:10.1007/s00477-013-0800-4.

Gould, P.G., Koehler, A.B., Ord, J.K., Snyder, R.D., Hyndman, R.J., Vahid-Araghi, F., 2008. Forecasting time series with multiple seasonal patterns. European Journal of Operational Research 191, 207–222. doi:10.1016/j.ejor.2007.08.024.

Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. Journal of Statistical Software 40, 1–25. doi:10.18637/jss.v040.i03.

Harvey, A.C., 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press. doi:10.1017/CBO9781107049994.

Héroux, M.E., Anderson, H.R., Atkinson, R., Brunekreef, B., Cohen, A., Forastiere, F., Hurley, F., Katsouyanni, K., Krewski, D., Krzyzanowski, M., Künzli, N., Mills, I., Querol, X., Ostro, B., Walton, H., 2015. Quantifying the health impacts of ambient air pollutants: Recommendations of a WHO/Europe project. International Journal of Public Health 60, 619–627. doi:10.1007/s00038-015-0690-y.

Hoek, G., Krishnan, R.M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J.D., 2013. Long-term air pollution exposure and cardio-respiratory mortality: A review. Environmental Health 12, 43. doi:10.1186/1476-069X-12-43.

Hyndman, R., 2018. fpp2: Data for 'Forecasting: Principles and Practice' (2nd Edition). URL: https://CRAN.R-project.org/package=fpp2. R package version 2.3.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2019. forecast: Forecast-

ing functions for time series and linear models. URL: http://pkg.robjhyndman.com/forecast. R package version 8.7.

Hyndman, R., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 26, 1–22. doi:10.18637/jss.v027.i03.

Kumar, U., Jain, V.K., 2010. ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). Stochastic Environmental Research and Risk Assessment 24, 751–760. doi:10.1007/s00477-009-0361-8.

Lawson, A.R., Ghosh, B., Broderick, B., 2011. Prediction of traffic-related nitrogen oxides concentrations using structural time-series models. Atmospheric Environment 45, 4719–4727. doi:10.1016/j.atmosenv.2011.04.053.

Liu, Z., Hu, B., Wang, L., Wu, F., Gao, W., Wang, Y., 2015. Seasonal and diurnal variation in particulate matter PM_{10} and $PM_{2.5}$ at an urban site of Beijing: Analyses from a 9-year study. Environmental Science and Pollution Research 22, 627–642. doi:10.1007/s11356-014-3347-0.

Mayer, H., 1999. Air pollution in cities. Atmospheric Environment 33, 4029–4037. doi:10.1016/S1352-2310(99)00144-2.

Moisan, S., Herrera, R., Clements, A., 2018. A dynamic multiple equation approach for forecasting $PM_{2.5}$ pollution in Santiago, Chile. International Journal of Forecasting 34, 566–581. doi:10.1016/j.ijforecast.2018.03.007.

Moritz, S., Bartz-Beielstein, T., 2017. imputeTS: Time series missing value imputation in R. The R Journal 9, 207–218. URL: 10.32614/RJ-2017-009, doi:10.32614/RJ-2017-009.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. URL: https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-2.

Proietti, T., Riani, M., 2009. Transformations and seasonal adjustment. Journal of Time Series Analysis 30, 47–69. doi:10.1111/j.1467-9892.2008.00600.x.

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/.

Ryan, J.A., Ulrich, J.M., 2018. xts: eXtensible Time Series. URL: https://CRAN.R-project.org/package=xts. r package version 0.11-2.

Sharma, P., Chandra, A., Kaushik, S.C., 2009. Forecasts using Box–Jenkins models for the ambient air quality data of Delhi City. Environmental Monitoring and Assessment 157, 105–112. doi:10.1007/s10661-008-0520-2.

Shumway, R.H., Stoffer, D.S., 2017. Time series analysis and its applications: With R examples. Springer. doi:10.1007/978-3-319-52452-8.

Taylor, J.W., 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. Journal of the Operational Research Society 54, 799–805. doi:10.1057/palgrave.jors.2601589.

Taylor, J.W., 2010. Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research 204, 139–152. doi:10.1016/j.ejor.2009.10.003.

Taylor, J.W., McSharry, P.E., 2007. Short-term load forecasting methods: An evaluation based on european data. IEEE Transactions on Power Systems 22, 2213–2219. doi:10.1109/TPWRS.2007.907583.

Taylor, S.J., Letham, B., 2018. Forecasting at scale. The American Statistician 72, 37–45. doi:10.1080/00031305.2017.1380080.

WHO, 2013. Health risks of air pollution in Europe – HRAPIE project recommendatiaons for concentration-response functions for cost-benefit analysis of particulate matter, ozone and nitrogen dioxide. URL: http://www.euro.who.int/__data/assets/pdf_file/0006/238956/Health_risks_air_pollution_HRAPIE_project.pdf. Accessed on November 20, 2019.

WHO, 2013. Review of evidence on health aspects of air pollution – REVIHAAP project: Technical report. URL: http://www.euro.who.int/__data/assets/pdf_file/0004/193108/REVIHAAP-Final-technical-report.pdf. Accessed on September 13, 2019.

Wickham, H., 2007. Reshaping data with the 'reshape' package. Journal of Statistical Software 21, 1–20. doi:10.18637/jss.v021.i12.

Wickham, H., 2016. ggplot2: Elegant graphics for data nalysis. Springer-Verlag New York. doi:10.1007/978-0-387-98141-3.

Wilke, C.O., 2019. cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. URL: https://CRAN.R-project.org/package=cowplot. r package version 1.0.0.

Wooldridge, J.M., 1992. Some alternatives to the Box-Cox regression model. International

Economic Review 33, 935–955. doi:10.2307/2527151.

Yi, X., Zheng, Y., Zhang, J., Li, T., 2016. ST-MVL: Filling missing values in geo-sensory time series data, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016. 9–15. URL: https://www.microsoft.com/en-us/research/publication/st-mvl-filling-missing-values-in-geo-sensory-time-series-data/.

Young, P.C., Pedregal, D.J., Tych, W., 1999. Dynamic harmonic regression. Journal of Forecasting 18, 369–394. doi:10.1002/(SICI)1099-131X(199911)18:6<369::AID-FOR748>3.0.CO;2-K.

Zhao, N., Liu, Y., Vanos, J.K., Cao, G., 2018. Day-of-week and seasonal patterns of $PM_{2.5}$ concentrations over the United States: Time-series analyses using the Prophet procedure. Atmospheric Environment 192, 116–127. doi:10.1016/j.atmosenv.2018.08.050.

4 Agglomeration and infrastructure effects in land use regression models for air pollution – Specification, estimation, and interpretations

Abstract. Established land use regression (LUR) techniques such as linear regression utilize extensive selection of predictors and functional form to fit a model for every data set on a given pollutant. In this paper, an alternative to established LUR modeling is employed, which uses additive regression smoothers. Predictors and functional form are selected in a data-driven way and ambiguities resulting from specification search are mitigated. The approach is illustrated with nitrogen dioxide (NO_2) data from German monitoring sites using the spatial predictors longitude, latitude, altitude and structural predictors; the latter include population density, land use classes, and road traffic intensity measures. The statistical performance of LUR modeling via additive regression smoothers is contrasted with LUR modeling based on parametric polynomials. Model evaluation is based on goodness of fit, predictive performance, and a diagnostic test for remaining spatial autocorrelation in the error terms. Additionally, interpretation and counterfactual analysis for LUR modeling based on additive regression smoothers are discussed.

Our results have three main implications for modeling air pollutant concentration levels: First, modeling via additive regression smoothers is supported by a specification test and exhibits superior in- and out-of-sample performance compared to modeling based on parametric polynomials. Second, different levels of prediction errors indicate that NO_2 concentration levels observed at background and traffic/industrial monitoring sites stem from different processes. Third, accounting for agglomeration and infrastructure effects is important: NO_2 concentration levels tend to increase around major cities, surrounding agglomeration areas, and their connecting road traffic network.

Keywords. Land use regression, Additive regression smoothers, Spatial cross-validation, Counterfactual analysis, Nitrogen dioxide, Exposure to air pollution.

4.1 Introduction

Results from epidemiology suggest that exposure to air pollution is a risk factor for developing malign tumours, respiratory and cardiovascular diseases (see, e.g., Tang et al., 2017; Amini et al., 2020 and the review articles Hoek et al., 2013; Atkinson et al., 2018) that increase mortality and may even offset positive effects attributed to outdoor physical exercise (Sinharay et al., 2018). Health effects of long-term exposure to air pollution are typically assessed by tracking the health of a study cohort while monitoring its exposure to pollutants by: Direct measurement via portable devices (Sinharay et al., 2018), measurements of nearby air quality monitoring sites (Ostro et al., 2010 use the closest site; Pope III et al., 2002 average over all sites in a city), satellite remote sensing (Wang et al., 2017), indicators of exposure (Beelen et al., 2014), or approximation based on the characteristics of the surroundings of a participant by, e.g., regression-based techniques such as land use regression (LUR; Johnson et al., 2010); for an extensive overview, see Hoek (2017). A number of recent applications of LUR modeling of air pollutant concentration levels employed standardized model selection techniques to choose predictors and functional form based on empirical data (Beelen et al., 2009, 2013; Wu et al., 2015; Eeftens et al., 2016; Rahman et al., 2017; Wolf et al., 2017; Lu et al., 2020b). The techniques in the studies relied on parametric polynomials of degree one to approximate air pollutant concentration levels and provided a reasonable fit to the data. The approaches were frequently verified by comparing expected and estimated effect signs. Drawbacks result from extensive specification search: Implementing the specification search process leads to ambiguities; additionally, substantial effort needs to be spent when developing large-scale models to match the study area of national cohorts, as the models are fitted separately for each pollutant and area (Hoek, 2017).

Other applications of LUR modeling used techniques from statistical learning: Approaches based on regularization and shrinkage like lasso and ridge regression reduce the predictor space by employing loss functions that rely on a fit term and a penalty term; nonparametric techniques such as random forests, (tree-based) boosting, artificial neural networks, support vector regression; and ensembles (models consisting of multiple individual models), which produce predictions by computing the (weighted) average of predictions obtained from individual models (see, e.g., Russo et al., 2013; Singh et al., 2013; Brokamp et al.,

2017; Alimissis et al., 2018; Vizcaino and Lavalle, 2018; Chen et al., 2019; Berrocal et al., 2020; Lu et al., 2020a). The techniques often exhibit black-box character and so-called meta-parameters need to be tuned. This typically requires profound knowledge of the algorithms in context with, e.g., comparable data structures; checking for confounders; or appropriate setup for meta-parameter tuning and evaluation of predictive performance (Riley, 2019).

Further applications of LUR modeling utilized regression smoothers in generalized additive models (GAM; see Hastie and Tibshirani, 1990; Wood, 2017). Within the framework, complex multivariate functionals are approximated by additive decompositions of univariate and bivariate smooths (see, e.g., Hart et al., 2009; Yanosky et al., 2009, 2014; Zhang et al., 2018; Chen et al., 2019). Modeling approaches based on additive regression smoothers are able to account for typical characteristics of continuous environmental processes such as air pollution: Local heterogeneity, potential nonlinearities, and complex dependence structures which vary over geographic space. We also refer to the latter as complex spatial association structures – structures that are captured by spatial predictors, while controlling for structural predictors.

Popular alternatives to LUR modeling not considered in this paper are dispersion modeling (DM) (see, e.g., de Hoogh et al., 2014; Fallah-Shorshani et al., 2017) and the geostatistical approach kriging (see, e.g, Mercer et al., 2011; Behm et al., 2018); for comparison studies of DM and LUR, see Gulliver et al. (2011); de Hoogh et al. (2014); for kriging and LUR, see Beelen et al. (2009); Mercer et al. (2011).

In this paper, we propose to decompose the effects of the spatial and structural predictors into additively linked univariate and bivariate smooths. LUR models based on parametric polynomials and additive regression smoothers are estimated. We include structural and spatial characteristics into the models and account for different monitoring site types. The statistical performance of the models is evaluated by their fit, a specification test, and prediction error measures. We characterize the performance of the models and provide empirical evidence that mean annual nitrogen dioxide (NO_2) concentration levels observed at different monitoring site types arise from different processes. We illustrate the processes based on three exemplary locations, elaborate on model interpretations, and compare our results with previous LUR studies.

The remainder of this paper is organized as follows: Section 4.2 reviews LUR modeling

based on parametric polynomials and highlights structural differences of LUR modeling based on additive regression smoothers; we further describe how response and predictors were obtained and detail model evaluation and software. Section 4.3 characterizes the empirical distribution of the response and correlations among the predictors; the section then illustrates the statistical performance of the two LUR approaches: In-sample and out-of-sample metrics are obtained from different validation schemes and monitoring site types. Section 4.4 focuses on LUR modeling based on additive regression smoothers. The section discusses interpolation maps and counterfactual analysis, compares our modeling results with previous LUR studies, and details the strengths and limitations of the approach. Section 4.5 concludes.

4.2 Material and methods

This section discusses the structural modeling assumptions involved in LUR modeling based on additive regression smoothers and parametric polynomials. Additionally, we introduce the data and detail how we evaluate the out-of-sample performance of the models. The functions and data used to generate the results in this paper are provided in an `R` package (Fritsch and Behm, 2021b; `https://github.com/markusfritsch/smoothLUR`).

4.2.1 Modeling based on additive regression smoothers

We estimated LUR models to predict conditional mean annual NO_2 concentration levels (in $\mu g/m^3$) using spatial and structural predictors. The former were longitude, latitude, and altitude; the latter were population density, land use classes, and road traffic intensity measures. We modeled mean annual NO_2 concentration levels Y via

$$Y = \mu(\boldsymbol{Z}, \boldsymbol{X}) + \varepsilon, \tag{4.1}$$

where $\mathbf{Z}$ and $\mathbf{X}$ denote structural and spatial characteristics and ε is a remainder term. The Kolmogorov-Arnold representation theorem (Kolmogorov, 1956; Arnol'd, 1957) suggests that functions of multiple predictors can be approximated by additively linked functions of lower dimension (i.e., a smaller number of predictors). We decomposed function $\mu(\cdot)$ from Equation (4.1) into a structural component $g(\cdot)$ and a spatial component $\eta(\cdot)$:

$$Y = g(\mathbf{Z}) + \eta(\mathbf{X}) + \varepsilon. \tag{4.2}$$

By analogous reasoning, we decomposed the functions $g(\cdot)$ and $\eta(\cdot)$ into additively linked smooth functions of low-dimension. Note that we assume additive separability of the two modeling components $\mathbf{Z}$ and $\mathbf{X}$. In principle, any desired interactions of the components could be included – yet, this requires additional degrees of freedom. We modeled the response Y via univariate or bivariate smooth functions of $\mathbf{Z}$ and $\mathbf{X}$ and obtained the LUR model based on additive regression smoothers

$$Y = \beta_0 + \sum_p s_{u,p}(Z_p) + s_b(X_{\text{Lon}}, X_{\text{Lat}}) + s_{u,A}(X_{\text{Alt}}) + \varepsilon. \tag{4.3}$$

This yields a generalized additive model (GAM) structure (Hastie and Tibshirani, 1990; Wood, 2017). The model components $s(\cdot)$ represent smooth functions; $s_{u,p}$ and $s_{u,A}$ are univariate smooths and s_b is a bivariate smooth. We used splines to model the smooth functions in Equation (4.3) to account for the typical characteristics of continuous environmental processes such as air pollution: Local heterogeneity, potential nonlinearities, and complex dependence structures which vary over geographic space. The degree of smoothness of the splines is chosen data-driven by trading-off goodness of fit to the data with a so-called "roughness penalty", which penalizes variation of the slope of the function. We used univariate and bivariate regression splines to model univariate and bivariate smooths (for details see, e.g., Ruppert et al., 2003; Wood, 2003; Hastie et al., 2009).

4.2.2 Modeling based on parametric polynomials

A special case of Equation (4.3) are LUR models based on parametric polynomials, for example,

$$Y = \beta_0 + \sum_i \beta_i Z_i + \sum_j \delta_j X_j + \varepsilon. \tag{4.4}$$

In Equation (4.4), β_i and δ_j denote coefficients of parametric polynomials corresponding to structural predictors Z_i and spatial predictors X_j. Models of the form of Equation (4.4)

were frequently employed in LUR studies (see, e.g., Hoek et al., 2008; Beelen et al., 2009, 2013; Eeftens et al., 2016; Wolf et al., 2017). In these studies, choosing the functional form – the predictors to include in the model via a parametric polynomial from a large set of potential predictors and the corresponding degree of the polynomial – was conducted by a forward stepwise procedure.

An established procedure utilizing linear regression modeling is the standardized ESCAPE procedure (see, e.g., Beelen et al., 2013; Eeftens et al., 2016; Wolf et al., 2017). This procedure consists of the following steps: First, specify a univariate model for the conditional distribution of the response Y including only the predictor that maximizes adjusted R-squared ($\overline{\mathrm{R}}^2$). At subsequent steps, add the predictor to the model that yields the maximum increase in $\overline{\mathrm{R}}^2$ and fulfills three criteria simultaneously: (i) $\overline{\mathrm{R}}^2$ increases by at least one percent, (ii) estimated coefficient conforms with pre-specified direction of effect, and (iii) directions of effects of predictors already included in the model do not change. When the procedure is stopped, all predictors included in the model with p-value larger than 0.1 are removed sequentially. Note that many alternative criteria could be used to guide specification search in forward stepwise procedures. Examples are other in- and out-of-sample metrics and corresponding thresholds for including predictors, further procedures for excluding predictors such as residual-based checks and formal tests, and different thresholds (for an overview of various metrics and criteria, see, e.g., Bennett et al., 2013).

4.2.3 Response and predictors

We obtained data on mean annual NO_2 concentration levels (response) from the German air quality monitoring network provided by the European Environment Agency (EEA, 2017). We constructed predictor data by dividing Germany into a 1 x 1 km grid and used values computed for grid cell centers to represent the predictor values for the whole grid cell. The values were computed based on considering relative fractions of areas, numeric values, or total lengths within a circle of radius (also referred to as buffer) 1 km around each grid cell center. In our LUR models, we used CORINE land cover data, shapefiles on administrative regions, a digital terrain model grid, population density at municipality key level, and road traffic network data. Table 4.1 shows the potential predictors, their units, and the data sources: Land cover classes which indicate the percentage of area in a buffer that is covered by residential areas of high population density such as residential

blocks (HighDens), low population density such as detached and semi-detached houses (LowDens), industrial areas (Ind), traffic infrastructure besides seaports and airports (Transp), seaports (Seap), airports (Airp), areas currently under construction (Constr), urban green spaces (UrbGreen), agriculture (Agri), forest (Forest); population density measured via inhabitants per km^2 (PopDens); a proxy for road traffic intensity obtained by adding up lengths of federal autobahn (FedAuto), primary roads (PriRoad), secondary roads (SecRoad), and local routes (LocRoute) within each buffer; longitude (Lon), and latitude (Lat) expressed via geocoordinates, and altitude in meters (Alt). Similar to studies by Beelen et al. (2009, 2013); Wolf et al. (2017); Vizcaino and Lavalle (2018), the rightmost column of Table 4.1 contains expected directions of effect (on average and ceteris paribus) of predictors on the response. With an increase in population density, e.g., mean annual NO_2 concentration levels are expected to increase on average – holding all other observable and unobservable influences constant. A thorough description of the data is given in Fritsch and Behm (2021a).

4.2.3.1 Air pollution measurement data

Mean annual NO_2 concentration levels were collected at 246 (19) background and 157 (22) traffic and industrial monitoring sites across Germany (Rhine-Ruhr metropolitan area). There are particular positioning requirements for the two types of monitoring sites: Background monitoring sites are representative of a wider area; traffic and industrial sites are located in close proximity to local pollution sources such as major roads or industrial areas (EEA, 2017). The left (right) plot in Fig. 4.1 displays spatial locations of monitoring sites across Germany (Rhine-Ruhr metropolitan area), where the border of the Rhine-Ruhr metropolitan area is depicted by the darkbronze line. Dots indicate background (bronze) and traffic/industrial monitoring sites (gold).

Germany covers an area of 357,386 km^2 (as of December 31, 2015; see Statistisches Bundesamt, 2020), with the Rhine-Ruhr metropolitan area contributing an area of roughly 7,000 km^2 (computed in `R`). As depicted by Fig. 4.1, monitoring sites are unevenly distributed across Germany – being more frequently located in large cities and relatively sparse in rural areas.

Table 4.1: Overview of potential predictors and data sources; CORINE classes 1-25 grouped into 10 classes according to Beelen et al. (2009); buffer radius 1 km; area measured as relative area within buffer (%), length as absolute length within buffer (meters). Rightmost column indicates expected directions of effect (on average and ceteris paribus) of predictors on mean annual NO_2 concentration levels.

GIS dataset	**Predictor**	**Name**	**Unit**	**Effect**
Land cover (EEA, 2016)	High density residential	HighDens	area (%)	+
	Low density residential	LowDens	area (%)	+
	Industry	Ind	area (%)	+
	Transport	Transp	area (%)	+
	Seaports	Seap	area (%)	+
	Airports	Airp	area (%)	+
	Construction	Constr	area (%)	+
	Urban Greenery	UrbGreen	area (%)	−
	Agriculture	Agri	area (%)	±
	Forest	Forest	area (%)	−
German administrative regions (BKG, 2015b)	Population density	PopDens	inhabitants per km^2	+
Road traffic network (EuroGeographics, 2018)	Federal autobahn	FedAuto	length (m)	+
	Primary roads	PriRoad	length (m)	+
	Secondary roads	SecRoad	length (m)	+
	Local routes	LocRoute	length (m)	+
Geocoordinates (EEA, 2017)	Longitude, Latitude	Lon, Lat	WGS84	±
Topography (EEA, 2017; BKG, 2015a)	Altitude	Alt	m	−

4.2.3.2 Land use proxies

We employed CORINE land cover data 2012, Version 18 in raster format (resolution 100 x 100 m) as proxy for land use. The raw data are available from the European Environment Agency (EEA, 2016) and contain information on land cover over Europe: One of 44 CORINE land cover classes is attributed to each raster cell. We grouped CORINE classes 1-25 into ten classes according to Beelen et al. (2009) and computed the proportion of surface area of each grouped class in a buffer of radius 1 km around each grid cell center. We excluded Seap, Airp, and Constr in model estimation, as mostly zeros were observed for these predictors.

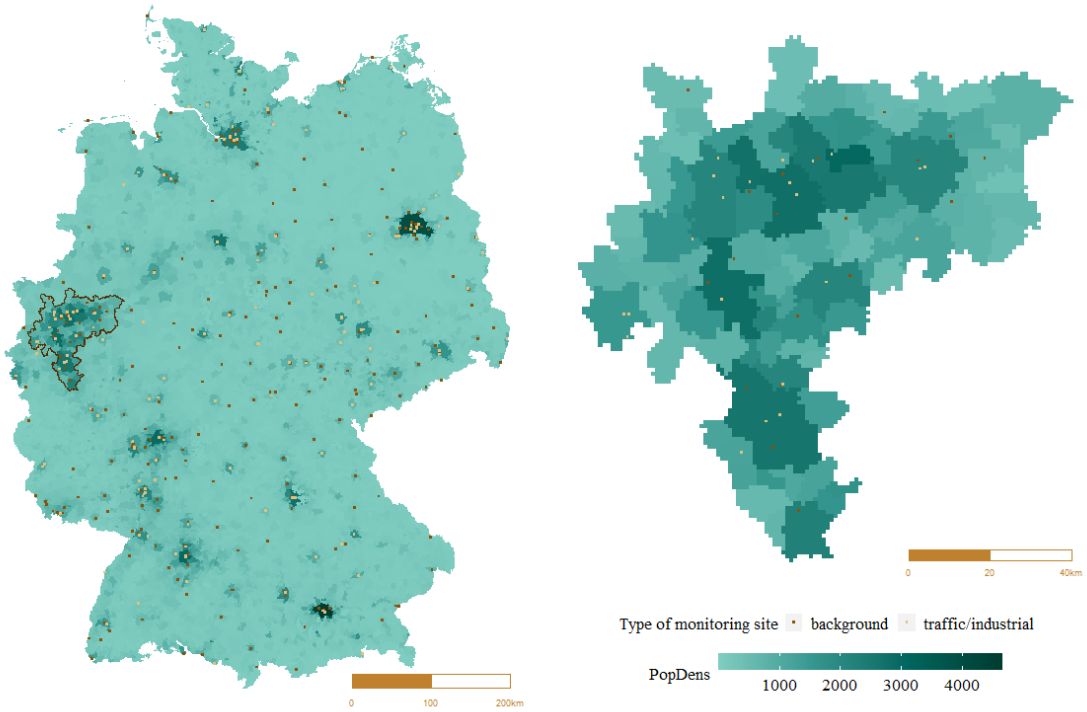


Figure 4.1: Map of Germany (left) and Rhine-Ruhr metropolitan area (right) representing population density at municipality key level; darker shades of green indicate higher population density; maps also show locations of monitoring sites for which mean annual NO_2 concentration levels were available for 2015; background monitoring sites given as bronze, traffic/industrial as gold dots; border of Rhine-Ruhr metropolitan area is depicted by darkbronze line (left).

4.2.3.3 Population density and road traffic network

The maps shown in Fig. 4.1 are colored according to population density at municipality key level, where population density increases with darker shades of green. As indicated by the right plot of Fig. 4.1, the Rhine-Ruhr metropolitan area is constituted of areas with high population density. We derived the number of inhabitants per km^2 from data provided by the Federal Government for Geo-Information and Geodesy (BKG, 2015b). For traffic intensity, we used a proxy computed by total length of four different types of roads in buffers of radius 1 km around each grid cell center: Federal autobahn, primary roads, secondary roads, and local routes. Data on the German road traffic network were obtained from EuroGeographics (2018).

4.2.3.4 Topography and geocoordinates

The meta information on monitoring sites contained geocoordinates and altitudes. Longitude and latitude of each grid cell of the 1 x 1 km grid across Germany were represented by the respective grid cell centers based on the World Geodetic System (WGS84). We derived values for altitude for each grid cell center from the digital terrain model grid of width 200 m obtained from the Federal Government for Geo-Information and Geodesy (BKG, 2015a).

4.2.4 Model evaluation and software

We evaluated model performance by computing the in-sample metric $\overline{\mathrm{R}}^2$ and derived the out-of-sample metrics root mean square error (RMSE) and mean absolute error (MAE) from different validation schemes: Leave-one-out cross-validation (LOOCV), repeated spatially stratified k-fold cross-validation (KFCV), and repeated spatially stratified hold-out validation (HOV). We conducted KFCV by adjusting the following algorithm for spatial stratification:

Algorithm 1: Algorithm for repeated k-fold cross-validation

Data: $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x})$ of sample size n, with $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_P)$ and $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$

Result: Out-of-sample metrics $RMSE$ and MAE

Initialization: Set the number of seeds R and the number of folds K;

for $r = 1, \ldots, R$ **do**

  draw random vector ι of length n containing numbers $1, \ldots, K$ with identical frequency;

  add ι as a column to data;

  for $k = 1, \ldots, K$ **do**

    assign all rows of the data for which $\iota \neq k$ to training set: $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x})_{r,k}$; sample size n_k;

    assign all rows of the data for which $\iota = k$ to test set: $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x})_{r,-k}$; sample size n_{-k};

    use training set to fit $\mathrm{par}_{\mathrm{r,k}}$ and $\mathrm{smooth}_{\mathrm{r,k}}$;

    compute $RMSE_{r,k} = \sqrt{\frac{1}{n_{-k}} \sum_{i=1}^{n_{-k}} (\widehat{y}_{i,r,k} - y_{i,r,k})^2}$;

    compute $MAE_{r,k} = \frac{1}{n_{-k}} \sum_{i=1}^{n_{-k}} |\widehat{y}_{i,r,k} - y_{i,r,k}|$;

  average over k to obtain $RMSE_r$ and MAE_r;

average over r to obtain $RMSE$ and MAE;

return *RMSE and MAE*

In Algorithm 1, vector ι is used for constructing K folds from the data. To adjust the algorithm for spatial stratification, draw the random vector ι such that it balances across the desired feature(s) or geographic regions. We balanced over the 16 German federal states and carried out the spatially stratified sampling such that each observation is held out exactly once; we used the function create_folds() from the `R`-package `splitTools` (Mayer, 2020) to obtain spatially stratified samples. Additionally, we set the number of folds $K = 10$, the number of seeds $R = 100$, and chose the numbers from 1 to 100 as seeds. The LOOCV and repeated HOV procedures deviate just slightly from Algorithm 1: For repeated HOV, averaging over K folds is not necessary; for LOOCV, set K equal to the number of observations n: Sampling, setting a seed, and computing the corresponding averages is redundant.[1]

We also computed Moran's I statistic with function `Moran.I()` from `R`-package `ape` (Paradis and Schliep, 2019). The null hypothesis considered is that there is no remaining spatial autocorrelation in the error terms. A rejection of the null is frequently interpreted as an indication that the model specification under consideration is spatially incomplete (or misspecified).

All computations in this paper were carried out with the statistical software `R`, version 4.0.2 (R Core Team, 2013) using packages `ape` (Paradis and Schliep, 2019), `cowplot` (Wilke, 2019), `data.table` (Dowle and Srinivasan, 2020), `ggplot2` (Wickham, 2016), `ggpubr` (Kassambara, 2020), `mgcv` (Wood, 2003, 2017), `raster` (Hijmans, 2020), `RColorBrewer` (Neuwirth, 2014), `rgdal` (Bivand et al., 2020), `rgeos` (Bivand and Rundel, 2020), `RgoogleMaps` (Löcher, 2020), `sp` (Pebesma and Bivand, 2005; Bivand et al., 2013), and `splitTools` (Mayer, 2020).

4.3 Results

The flow chart shown in Fig. 4.2 gives an overview of the input data and the modeling stages of the empirical analysis. The following section depicts the results of the individual modeling stages.

We illustrate descriptives on mean annual NO_2 concentration levels, depending on the type of monitoring site. We consider all monitoring sites, background monitoring sites,

[1] For further reading on validation schemes in the context of modeling air pollutant concentration levels, we refer the interested reader to Johnson et al. (2010), Huang et al. (2018), Zhang et al. (2018), Chen et al. (2019), Berrocal et al. (2020), and Ren et al. (2020).

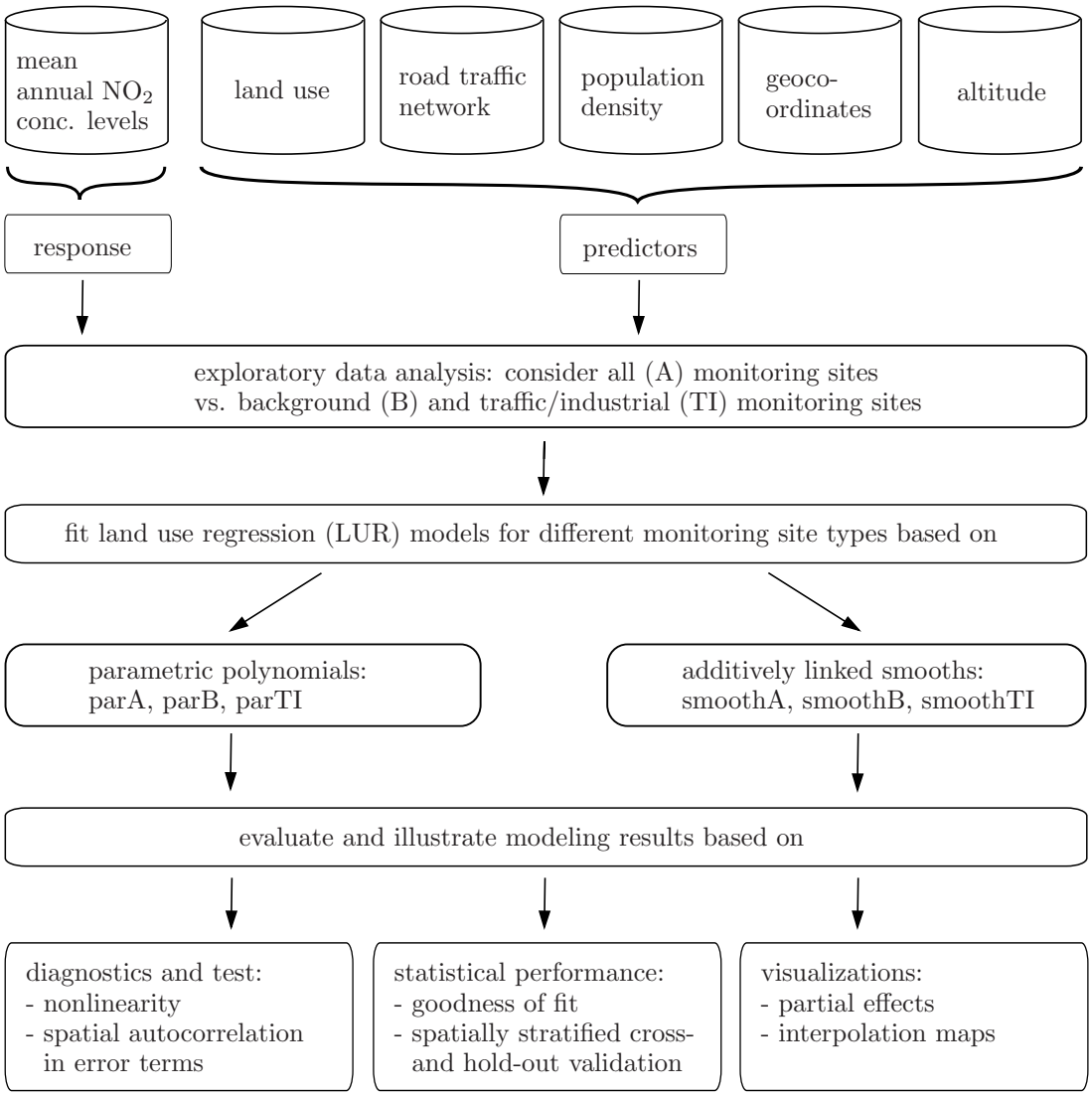


Figure 4.2: Flow chart of input data and modeling stages of empirical analysis.

and traffic/industrial monitoring sites. Pairwise correlations among the predictors are investigated. We contrast LUR modeling based on parametric polynomials and additive regression smoothers in terms of model fit and predictive performance for different validation schemes and types of monitoring sites; additionally, results from a specification test are given.

4.3.1 Model input data

Table 4.2 provides descriptives on mean annual NO_2 concentration levels for all 403, 246 background, and 157 traffic/industrial monitoring sites.

Table 4.2: Descriptives characterizing empirical distribution of mean annual NO_2 concentration levels depending on type of monitoring site; measures include mean, standard deviation, five number summary, and total number of observations n; lines give figures for all, background, and traffic/industrial monitoring sites.

Type of monitoring site	n	Mean	SD	Min	q25	Median	q75	Max
All	403	25.39	14.79	2.53	14.73	22.41	33.24	87.23
Background	246	17.39	7.74	2.53	11.34	17.62	22.90	39.00
Traffic/Industrial	157	37.92	14.51	10.69	27.49	37.17	47.20	87.23

Table 4.2 shows higher mean annual NO_2 concentration levels, variations, and (interquartile) ranges for traffic/industrial sites as compared to background monitoring sites. Fig. 4.3 provides histograms and corresponding empirical density curves for mean annual NO_2 concentration levels recorded at background (bronze) and traffic/industrial (gold) monitoring sites.

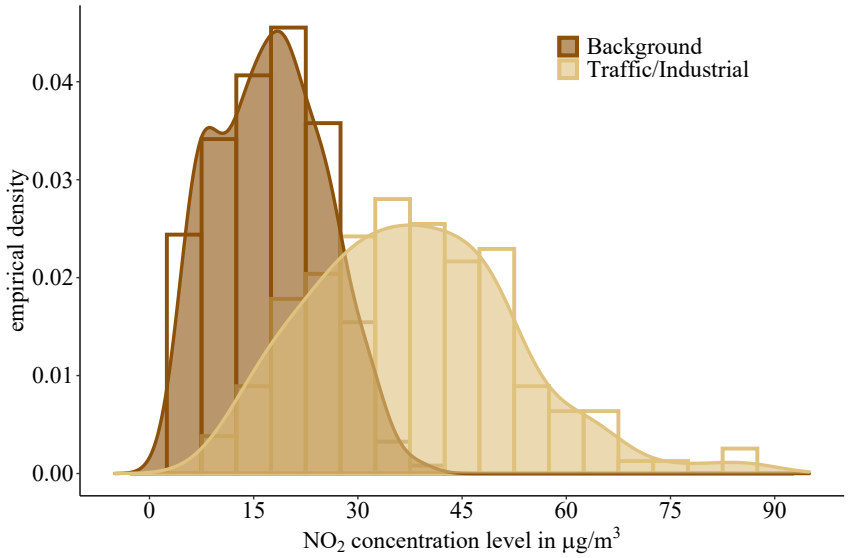


Figure 4.3: Histogram and empirical density curve for mean annual NO_2 concentration levels at background (bronze) and traffic/industrial (gold) monitoring sites.

Since high correlations (collinearities) among predictors lead to imprecisely estimated coefficient estimates, we considered pairwise Bravais-Pearson correlations for all predictors employed in our LUR models. Fig. 4.4 displays a correlation plot, where negative correlations are indicated in bronze and positive correlations in gold. The color shade represents the strength of correlation, with light colors reflecting correlations that are close to zero and dark colors indicating correlations close to one in absolute value. We excluded the predictors Airp, Seap, and Constr since we observed mostly zeros. There were no indi-

cations of collinearity problems. Pairwise correlations across predictors observed at all monitoring sites ranged from −0.62 to 0.53.

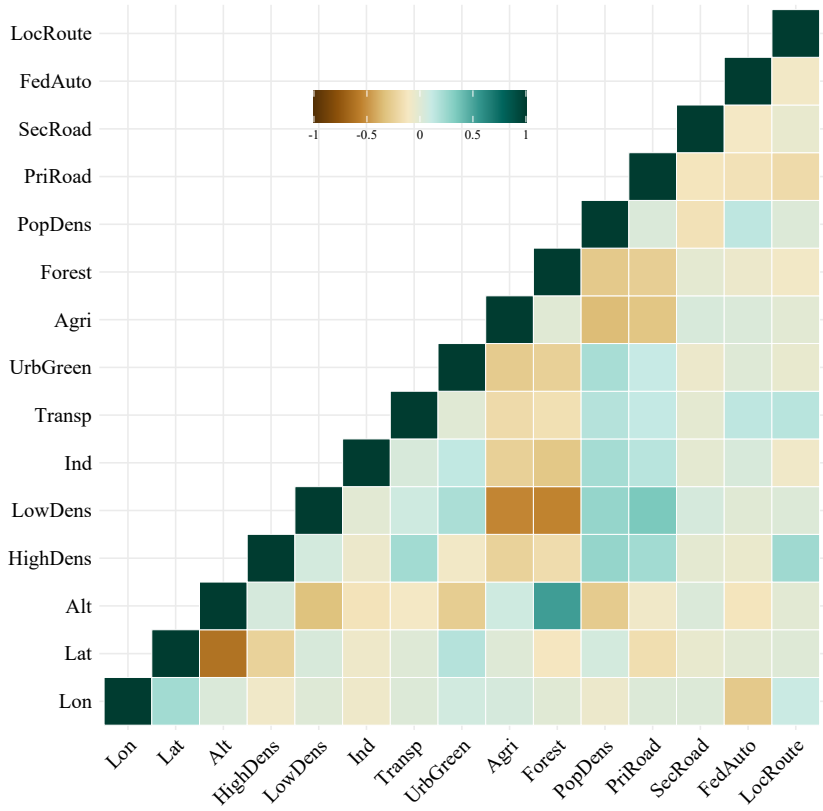


Figure 4.4: Pairwise Bravais-Pearson correlations of predictors observed at all monitoring sites; Airp, Seap, and Constr, where mostly zeros were observed, are excluded; color scale ranges from darkbronze to darkgreen corresponding to correlations from −1 to 1; light colors represent correlations close to zero, dark colors correlations close to one in absolute value; minimum (maximum) observed correlation −0.62 (0.53).

4.3.2 Modeling results

We modeled mean annual NO_2 concentration levels with two different LUR approaches: LUR modeling based on parametric polynomials of degree one and LUR modeling based on additive regression smoothers. First, we fitted models on data from all monitoring sites and evaluated the statistical performance. We then used only background or traffic/industrial monitoring site data and illustrated the differences in estimated concentration levels across monitoring site types.

4.3.2.1 Modeling based on all monitoring sites

Table 4.3 shows results for fitting a LUR model based on parametric polynomials by the ESCAPE procedure. Similar to recent LUR studies of Beelen et al. (2013); Eeftens et al. (2016); Wolf et al. (2017), we only considered parametric polynomials of degree one and used data observed at all monitoring sites for model fitting. We referred to the LUR model as parA; the capital letter indicates the data used for model fitting.

Table 4.3: LUR model based on parametric polynomials parA fitted by ESCAPE procedure; only polynomials of degree one of structural and spatial predictors are considered; data from all monitoring sites used for model fitting; $\overline{R}^2$ and Moran's I statistic to test for spatial autocorrelation in error terms given for parA.

Pred.	Est.	SE
(Int.)	16.795	2.425
PopDens	0.006	0.001
PriRoad	0.002	< 0.001
HighDens	41.810	5.382
LowDens	13.160	1.997
Lon	-0.988	0.216
SecRoad	0.002	0.001
$\overline{R}^2$	0.57	
Moran's I (p-value)	-0.18 (< 0.001)	

Six predictors reflecting agglomeration and infrastructure effects were selected when fitting parA. The model yields an $\overline{R}^2$ of 0.57; Moran's I statistic rejected the null of no spatial autocorrelation in the error terms and indicated model misspecification. When an interaction of Lon and Lat is included in the LUR model, the effect is selected instead of Lon. All other modeling results remain qualitatively identical.

We also fitted a LUR model based on additive regression smoothers smoothA to the data. The model accounts for nonlinearities and complex spatial association structures: We included Lon and Lat via a bivariate smooth; all other predictors are included via univariate smooths. Fig. 4.5 illustrates the partial effects of smoothA and corresponding 95% confidence intervals. According to the partial effects plots, there are no clear indications for nonlinearities. Five predictors were smoothed out, as the estimated degrees of freedom (edf) reduced to zero; similar to parA, the predictors included in smoothA reflect agglomeration and infrastructure effects.

The interpolation maps displayed in Fig. 4.6 were derived from smoothA. Both maps

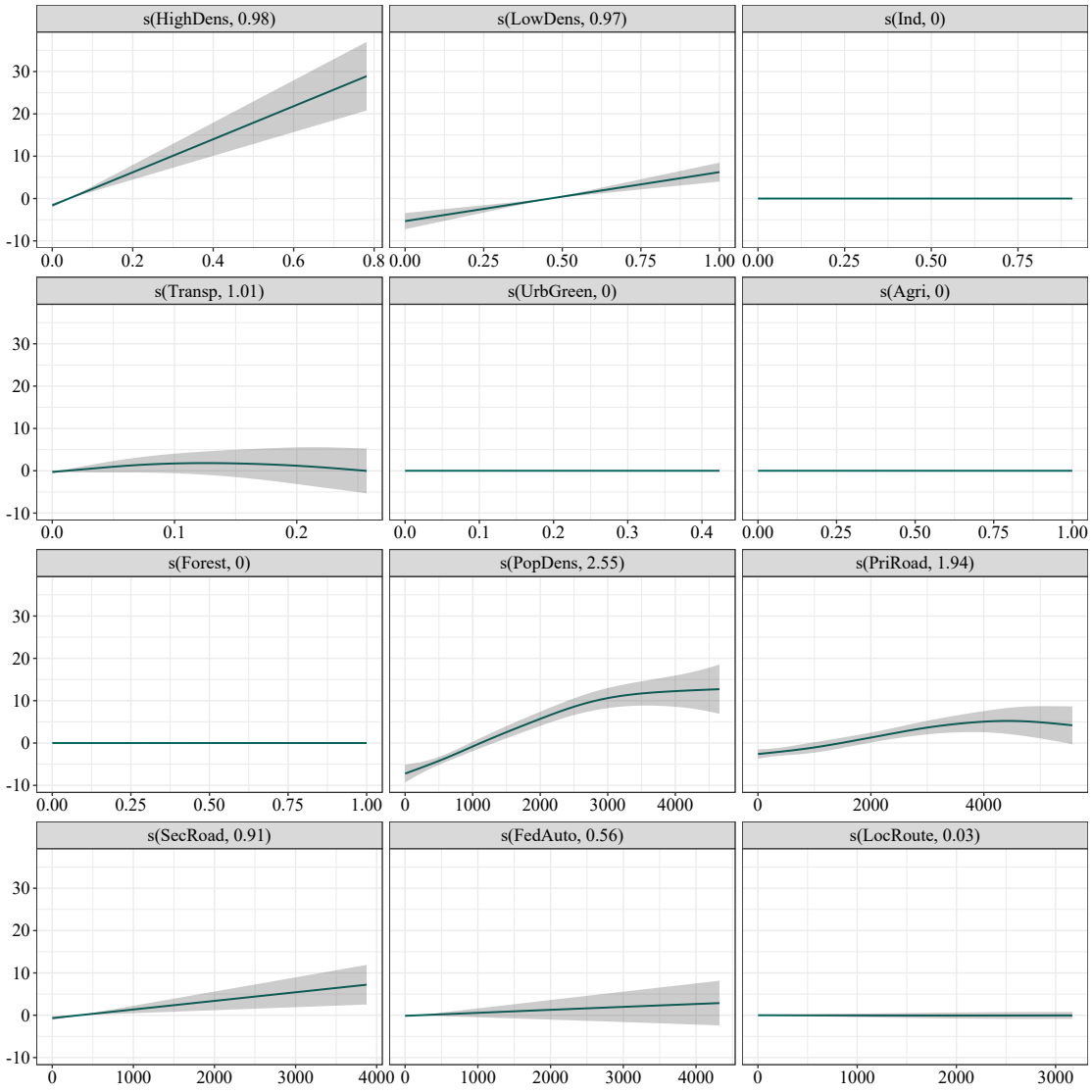


Figure 4.5: Partial effects of LUR model based on additive regression smoothers smoothA; darkgreen line depicts univariate regression spline estimate of $s_{u,p}$ for respective z_p, grey area marks corresponding 95 % pointwise confidence bands. Numbers in parentheses refer to estimated degrees of freedom (edf) for spline estimates and indicate curvature of effect of predictors (details, see Wood, 2017); roughness of spline effect increases with edf, where 1 corresponds to linear parametric effect; 0 implies effect is smoothed out.

were created based on gridded data. The right map visualizes conditional mean annual NO_2 concentration level across all grid cells. Values were obtained by adding all structural effects for a grid cell to the corresponding spatial effect and ranged from roughly 2.5 to 65.1. We replaced 951 out-of-range predictions (0.3% of the 356, 791 predictions) by minimum or maximum observed concentration level. Values tended to be higher in two particular areas: (1) In cities and surrounding agglomeration areas compared to rural areas; (2) In

grid cells which include the road traffic network linking the agglomeration areas. The latter results from the positive partial effect of the road traffic intensity predictors (in particular PriRoad and SecRoad) on mean annual NO_2 concentration levels (see Fig. 4.5). In the right map of Fig. 4.6, the effect is visible through thin orange and red lines. The left plot of Fig. 4.6 displays the part of conditional mean annual NO_2 concentration levels attributable to the spatial effect. Values for the grid cells were obtained by adding up the bivariate smooth of Lon, Lat and the univariate smooth of Alt and ranged from roughly -12.9 to 3.9. The spatial effect tended to decrease when moving from southwest to northeast. This highlights the presence of complex spatial association structures which were captured by the bivariate and univariate smooth of the spatial predictors.

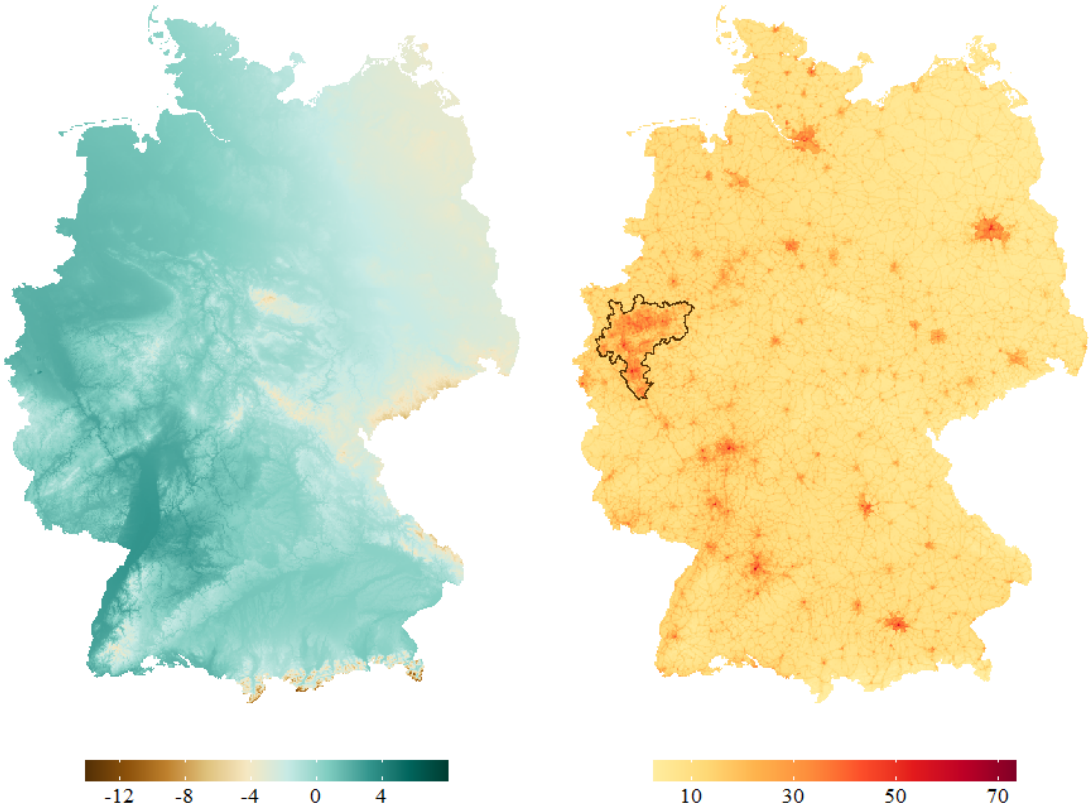


Figure 4.6: Interpolation maps derived from LUR model based on additive regression smoothers smoothA in 1 x 1 km resolution; all monitoring site types used for model fitting; maps visualize conditional mean annual NO_2 concentration levels (right) and part which is attributable to spatial effect (left); the latter consists of sum of bivariate smooth of Lon and Lat $s_b(X_{\text{Lon}}, X_{\text{Lat}})$ and univariate smooth of Alt $s_{u,A}(X_{\text{Alt}})$; out-of-range predictions replaced by minimum (maximum) observed mean annual NO_2 concentration level; border of Rhine-Ruhr metropolitan area depicted by darkbronze line.

We evaluated models parA and smoothA based on LOOCV, repeated spatially stratified K-fold cross-validation (KFCV, with $K = 10$), and repeated spatially stratified HOV; out-of-sample metrics RMSE and MAE were computed for all validation schemes; we used the 16 federal states of Germany as geographic regions for spatial stratification and balanced the observations in the training and test sample accordingly. Table 4.4 illustrates the results together with $\overline{R}^2$ and the Moran's I statistic.

Table 4.4: In-sample metric $\overline{R}^2$, Moran's I statistic, and out-of-sample metrics RMSE and MAE for LUR model based on parametric polynomials parA and LUR model based on additive regression smoothers smoothA; all monitoring site types used for model fitting; validation schemes: Leave-one-out cross-validation (LOOCV), K-fold cross-validation (KFCV, with $K = 10$), and hold-out validation (HOV).

			LOOCV		KFCV		HOV	
	$\overline{R}^2$	Moran's I (p-value)	RMSE	MAE	RMSE	MAE	RMSE	MAE
parA	0.57	-0.18 (< 0.001)	9.87	7.30	9.92	7.37	9.68	7.20
smoothA	0.59	-0.23 (< 0.001)	9.98	7.27	9.96	7.25	9.72	7.12

The metrics displayed in Table 4.4 are similar for both LUR models; Moran's I statistic indicates that the models are misspecified. As monitoring sites are divided into two categories, we considered the predictive performance for both types of monitoring sites separately.

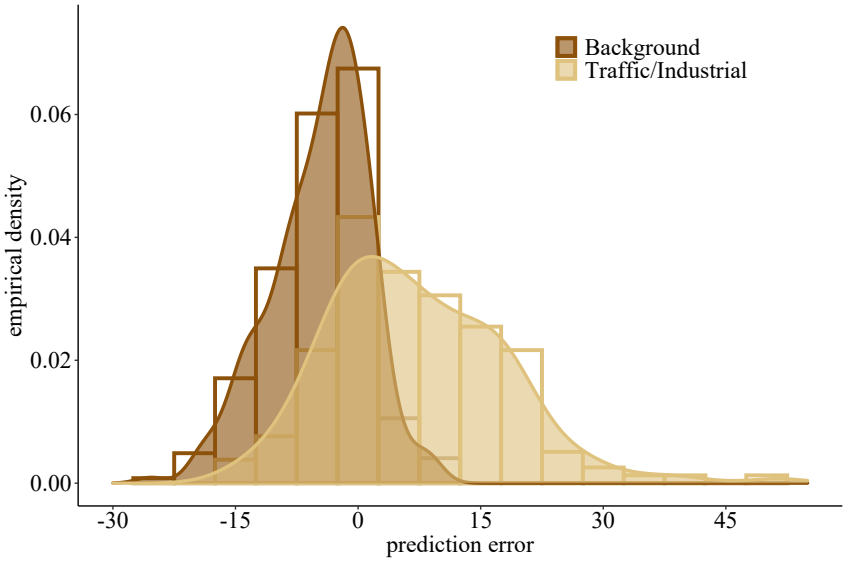


Figure 4.7: Histogram and empirical density curve for LOOCV prediction errors at background (bronze) and traffic/industrial (gold) monitoring sites for LUR model based on additive regression smoothers smoothA.

Fig. 4.7 shows histograms and empirical density curves for LOOCV prediction errors based

on smoothA for background (bronze) and traffic/industrial monitoring sites (gold). The plot illustrates that prediction errors and their dispersion tend to be higher for the latter sites. The results are robust across validation schemes and qualitatively identical for parA. This indicates that concentration levels observed at background and traffic/industrial monitoring sites may stem from two different processes.

4.3.2.2 Modeling depending on the type of monitoring site

We fitted LUR models for concentration levels observed at background and traffic/industrial monitoring sites. We estimated a LUR model based on parametric polynomials and additive regression smoothers for both types of sites. We refer to the former models with parB and parTI and to the latter models with smoothB and smoothTI; the capital letters indicate the data used for model fitting. Table 4.5 shows coefficient estimates, $\overline{\mathrm{R}}^2$ and the Moran's I statistic for parB and parTI.

Table 4.5: LUR models based on parametric polynomials parB and parTI fitted by ESCAPE procedure using parametric polynomials of structural and spatial predictors; only polynomials of degree one are considered; parB uses data observed at background monitoring sites; parTI uses traffic/industrial monitoring site data; $\overline{\mathrm{R}}^2$ and Moran's I statistic to test for spatial autocorrelation in error terms given for both models.

parB			parTI		
Pred.	Est.	SE	Pred.	Est.	SE
(Int.)	129.151	9.894	(Int.)	97.005	26.505
PopDens	0.002	< 0.001	PopDens	0.008	0.001
Forest	-9.760	1.061	PriRoad	0.002	0.001
FedAuto	0.003	0.001	HighDens	20.503	7.056
Agri	-7.013	0.968	LowDens	17.446	3.864
Lat	-2.143	0.190	Lat	-1.651	0.512
Alt	-0.010	0.002			
$\overline{\mathrm{R}}^2$	0.76			0.54	
Moran's I (p-value)	0.19 (< 0.001)			0.12 (0.06)	

Both models contain structural and spatial predictors. Six predictors were selected in model parB, five in parTI. While $\overline{\mathrm{R}}^2$ was higher for model parB, Moran's I statistic indicated model misspecification for both models ($\alpha = 0.1$).

We only provide a brief summary of the modeling results of smoothB and smoothTI in the following, as the key characteristics of the models were similar to smoothA (see Section 4.3.2.1; more detailed results are included in the Appendix). Five predictors were

smoothed out in both models. Nonlinearities were present for predictor PopDens for smoothB, while there were no clear deviations from nonlinearity for smoothTI. The two models captured agglomeration and infrastructure effects and complex spatial association structures. Predicted conditional mean annual NO_2 concentration levels based on smoothTI exceeded those based on smoothB.

Table 4.6: In-sample metric $\overline{R}^2$, Moran's I statistic, and out-of-sample metrics RMSE and MAE for LUR models based on parametric polynomials parB, parTI and LUR models based on additive regression smoothers smoothB, smoothTI; capital letters indicate monitoring site types used for model fitting: Background (B), traffic/industrial (TI); validation schemes include leave-one-out cross-validation (LOOCV), K-fold cross-validation (KFCV, with $K = 10$), and hold-out validation (HOV).

			LOOCV		KFCV		HOV	
	$\overline{R}^2$	Moran's I (p-value)	RMSE	MAE	RMSE	MAE	RMSE	MAE
parB	0.76	0.19 (< 0.001)	3.88	2.99	3.94	3.02	3.88	2.99
smoothB	0.85	-0.03 (0.44)	3.34	2.55	3.33	2.54	3.35	2.58
parTI	0.54	0.12 (0.06)	10.04	8.15	10.35	8.37	10.01	8.12
smoothTI	0.62	0.05 (0.37)	10.23	8.22	10.31	8.31	9.90	8.04

Table 4.6 shows $\overline{R}^2$, Moran's I statistic, and out-of-sample metrics RMSE and MAE for the LUR models based on parametric polynomials of degree one and the LUR models based on additive regression smoothers. The main implications of the table are: (i) $\overline{R}^2$ is higher for LUR models based on additive regression smoothers; (ii) for background monitoring sites, RMSE and MAE are lower for smoothB compared to parB – while there are only minor differences between smoothTI and parTI; (iii) there are no substantial differences across validation schemes or out-of-sample metrics; (iv) Moran's I statistic does not indicate model misspecification for smoothB and smoothTI. Overall, the results support the more flexible specifications: Additively linked univariate and bivariate smooths are suitable to account for spatial heterogeneity in mean annual NO_2 concentration levels. Additionally, fitting separate LUR models for background and traffic/industrial monitoring site data improves the statistical performance of the models.

We used the two LUR models based on additive regression smoothers smoothB and smoothTI to create interpolation maps for the Rhine-Ruhr metropolitan area in 1 x 1 km resolution. Fig. 4.8 shows the maps. The left plot was created by predicting mean annual NO_2 concentration levels for all grid cells with smoothB; for the right plot, smoothTI was used. Values range from 10.2 to 37.3 (left map) and from 11.6 to 57.1 (right map); simi-

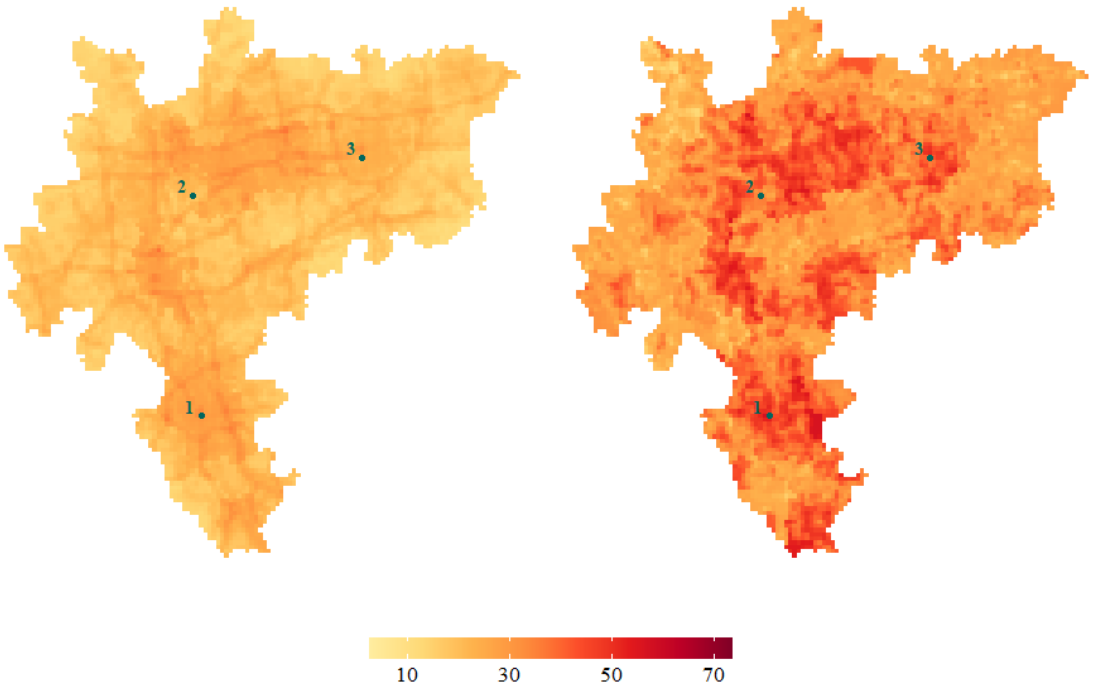


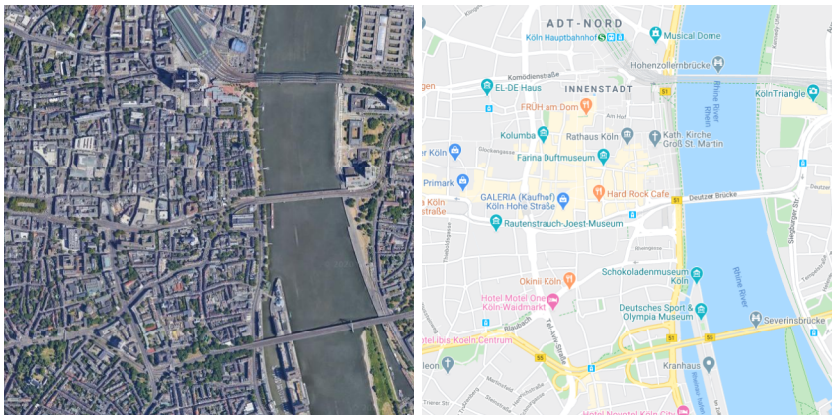
Figure 4.8: Interpolation maps for background (left) and traffic/industrial (right) conditional mean annual NO_2 concentration levels across Rhine-Ruhr metropolitan area in 1 x 1 km resolution; predictions derived from LUR models based on additive regression smoothers smoothB and smoothTI; out-of-range predictions replaced by minimum (maximum) observed concentration level; points 1, 2, and 3 mark grid cell centers located in Cologne city center, southern countryside of Mühlheim an der Ruhr, and suburb of Dortmund.

lar to Fig. 4.6, the interpolation maps highlight agglomeration and infrastructure effects: Mean annual NO_2 concentration levels are higher in large cities, the surrounding areas, and at locations of the connecting road traffic network. Predictions from smoothTI clearly exceeded those from smoothB.

4.3.2.3 Modeling city, suburban, and rural regions

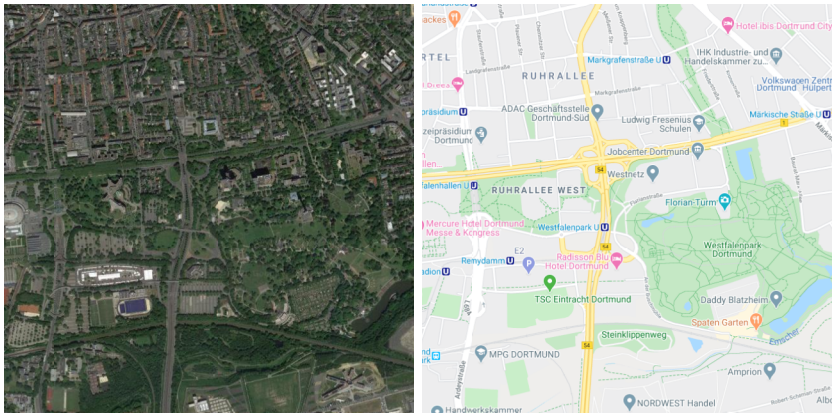
We illustrate our modeling results based on three exemplary locations marked in Fig. 4.8: Cologne city center (point 1), southern countryside of Mühlheim an der Ruhr (point 2), and suburb of Dortmund (point 3). Fig. 4.9 gives a visual impression of the grid cells corresponding to the three points.

For point 1, the area is characterized by inhabited areas and primary roads. The area around point 2 is covered mostly by agricultural areas and green spaces; for point 3 it



(a) City center of Cologne (point 1)

(b) Southern countryside of Mühlheim an der Ruhr (point 2)



(c) Suburb of Dortmund (point 3)

Figure 4.9: Satellite (left) and map (right) images extracted from Google Maps for grid cells centered around points 1, 2, and 3.

is mostly inhabited areas, green spaces, industrial areas, and infrastructure. Table 4.7 summarizes structural predictor values for the grid cells corresponding to the three points.

Table 4.7: Structural predictors of three locations marked in Fig. 4.8 which indicate grid cell centers located in Cologne city center (point 1), southern countryside of Mühlheim an der Ruhr (point 2), and suburb of Dortmund (point 3). Predictors HighDens, LowDens, Ind, Transp, UrbGreen, Agri, and Forest in %; PopDens in inhabitants per km^2; PriRoad in meters road length; predictors with only zero values for all three points omitted.

Point	HighDens	LowDens	Ind	Transp	UrbGreen	Agri	Forest	PopDens	PriRoad
1	0.36	0.38	0.03	0.01	0.00	0.00	0.00	2,607	3,702
2	0.00	0.17	0.00	0.00	0.13	0.67	0.00	1,855	771
3	0.00	0.37	0.15	0.14	0.27	0.00	0.08	2,095	3,849

Some predictors showed considerable variation across grid cells. For predictors measured on relative scale (area in %), predictor variations ranged from 0.08 (Forest) to 0.67 (Agri). For PopDens, values ranged from $1,855$ to $2,607$ (inhabitants per km^2) and for PriRoads, values ranged from 771 to $3,849$ (meters road length); we omitted predictors which had values of zero for all three points.

We used the LUR models based on additive regression smoothers smoothB and smoothTI to predict conditional mean annual NO_2 concentration levels for the three points. Table 4.8 shows the values and illustrates that the predictions varied substantially across models: Predictions from smoothTI clearly exceed those from smoothB, with differences decreasing from point 1 (urban area) to point 3 (suburban area) to point 2 (rural location). Values from smoothA are given for comparison.

Table 4.8: Predictions from LUR models based on additive regression smoothers smoothB, smoothTI, and smoothA for three locations marked in Fig. 4.8; locations indicate grid cell centers located in Cologne city center (1), southern countryside of Mühlheim an der Ruhr (2), and suburb of Dortmund (3).

Point	smoothB	smoothTI	smoothA
1 (Cologne, city center)	32.58	51.32	52.07
2 (Mühlheim an der Ruhr, countryside)	19.52	31.10	24.69
3 (Dortmund, suburb)	23.50	42.51	36.49

4.4 Discussion

The focus of the following discussion is on LUR models based on additive regression smoothers. We discuss interpretation of the complex effects arising from the interplay of

multiple predictors: We illustrate how interpolation maps can be used for counterfactual analysis of individual exposure to air pollution. Additionally, we compare our modeling results to previous LUR studies and point out the strengths and limitations of the approach employed in this paper.

4.4.1 Interpretation and counterfactual analysis for LUR models based on additive regression smoothers

Interpolation maps can be generated by evaluating LUR models based on additive regression smoothers on a given grid. The maps are a concise visual summary of the fitted models. As such, they illustrate the complex effects arising from the interplay of multiple predictors. When suitable models are employed, interpolation maps provide a comprehensive picture of exposure to air pollutant concentration levels and their spatial variation: Fig. 4.6 highlights the effect of agglomeration areas and the connecting road traffic network on mean annual NO_2 concentration levels. Potential applications for interpolation maps reach beyond visual representations of modeling results; the maps can be employed for counterfactual analysis. We use points 1, 2, 3 and LUR models smoothB and smoothTI to provide an example motivated by recent results from epidemiology: Assume an individual takes a walk through an urban green space or roams a street with high traffic intensity. The effect on exposure to air pollution can be approximated by contrasting the two maps displayed in Fig. 4.8. Recent results by Sinharay et al. (2018) suggest that taking these differences into account is important. They linked walking in highly polluted areas with adverse health effects, which may even offset positive effects attributed with outdoor physical exercise.

Besides counterfactual analysis, interpolation maps may also be useful in a number of further settings. First, trends in air pollutant concentrations levels over time can be investigated for past data by comparing maps across years (see, e.g., Hart et al., 2009); second, potential future outcomes resulting from different scenarios for particular predictors can be considered based on predicted air pollutant concentration levels (see, e.g., Vizcaino and Lavalle, 2018); third, interpolation maps of background air pollutant concentration levels could be used to reflect baseline exposure of individuals in epidemiologic studies which is enriched by more detailed information – for example, models for work-place exposure, more fine-grained effects such as regional anthropogenic factors (harbors, airports, and

fossil power plants), or regional natural factors (such as proximity to the ocean).

4.4.2 Comparison to previous LUR studies

Recent studies which modeled mean annual NO_2 concentration levels were conducted by, e.g., Beelen et al. (2013); Eeftens et al. (2016); Wolf et al. (2017); Vizcaino and Lavalle (2018); Chen et al. (2019). These studies used data from monitoring sites located in Germany, Switzerland, or the European Union. The studies by Beelen et al. (2013); Eeftens et al. (2016); Wolf et al. (2017) employed parametric polynomials of degree one together with predictor selection (ESCAPE procedure). Beelen et al. (2013) summarized the application of the ESCAPE procedure to 36 major cities across the EU-25 based on data measured from October 2008 until April 2011. Different models resulted for different cities, where R^2 ranged from 0.55 to 0.92 ($\overline{R}^2$ not reported) and RMSE (obtained from LOOCV) ranged from 2.1 to 12.0 ($\mu g/m^3$). Eeftens et al. (2016) employed the procedure to data from Switzerland collected from January 2011 until December 2012. They considered fitting separate models to alpine and non-alpine regions and fitting separate models to different regions. According to Eeftens et al. (2016), the latter yielded superior predictive performance; $\overline{R}^2$ ranged from 0.46 to 0.89 and RMSE (obtained from LOOCV) ranged from 3.0 to 8.9 ($\mu g/m^3$). Wolf et al. (2017) used data from Augsburg (Germany) measured in between March 2014 and April 2015 and obtained an $\overline{R}^2$ of 0.94 (RMSE not reported). Vizcaino and Lavalle (2018) employed data from 2010 and fitted LUR models based on random forests across the EU-28. They reported R^2 ranging from 0.4 to 0.64 ($\overline{R}^2$ and RMSE not reported). Chen et al. (2019) also used data from 2010 and fitted various statistical learning models such as artificial neural networks, boosted regression trees, and support vector regression besides approaches that are based on GAM. Reported R^2 ranged from 0.59 to 0.95 ($\overline{R}^2$ not reported); for the out-of sample metrics, RMSE ranged from 9.0 to 9.6 ($\mu g/m^3$) for 5-fold cross-validation and from 11.5 to 14.6 ($\mu g/m^3$) for hold-out validation.

We fitted model smoothA based on data from all monitoring sites. The model yielded an $\overline{R}^2$ of 0.59, RMSE was in between 9.72 and 9.98 ($\mu g/m^3$), and MAE ranged from 7.20 to 7.27 ($\mu g/m^3$). When we considered models fitted for different types of monitoring sites, model smoothB exhibited the best statistical performance for the considered metrics. The model yielded an $\overline{R}^2$ of 0.85, RMSE was in between 3.33 and 3.35 ($\mu g/m^3$), and MAE

ranged from 2.54 to 2.58 ($\mu g/m^3$). Overall, our results were similar to the LUR studies mentioned above. Note, that comparability of modeling results across all mentioned studies is limited for the following reasons: (i) Data from different sources and years were employed; (ii) data from different monitoring site types were used; (iii) studies focused on different regions; (iv) different model performance metrics were reported (not all studies reported out-of-sample metrics) or different techniques were used to obtain the metrics (LOOCV, KFCV, and HOV based on different seeds, choices for the number of folds, and their stratification).

4.4.3 Strengths and limitations

There are a number of strengths and limitations of LUR modeling based on additive regression smoothers employed in this paper. First, as we used a 1 x 1 km grid in our analysis, our models miss out on fine-scale variation of the process inducing mean annual NO_2 concentration levels. The resolution is likely too coarse to account for the leveling off of the effects of local pollution sources at traffic/industrial monitoring sites. This may partly explain why there are no substantial differences in Table 4.6 between the LUR models based on parametric polynomials and additive regression smoothers regarding predictive performance. To overcome this drawback, the background concentration levels captured by LUR models based on additive regression smoothers could be complemented by models which capture regional particularities and pollutant-specific behavior.

Second, we only considered a single buffer of radius 1 km in our analysis. Investigating further buffers may be valuable when considering mean annual NO_2 concentration levels observed at traffic/industrial monitoring sites in greater detail. We plan to return to this aspect in subsequent work and incorporate more detailed data on traffic intensity. Note that including further buffer sizes into the analysis may lead to increased correlations among the predictors. We recommend a careful correlation analysis similar to the one carried out in our empirical application, as the predictor chosen by the algorithm may be arbitrary in the presence of high correlations.

Third, LUR modeling based on additive regression smoothers determines the appropriate number of degrees of freedom for the different effects based on the data. Basically, the functional form and its flexibility for approximating mean annual NO_2 concentration levels are determined based on the data. This is a key difference to LUR modeling

based on parametric polynomials, where specifying the functional form and/or predictor selection are delicate tasks which need to be performed by the analyst. Compared to the ambiguities resulting from specification search in LUR modeling based on parametric polynomials, LUR modeling based on additive regression smoothers "let's the data speak for themselves". This carries the potential of uncovering interesting effects, e.g., nonlinear and non-monotonic relationships between response and predictors, and can also illustrate limitations of the data (when yielding implausible effect shapes).

Fourth, we impose additive separability of the different model components in the empirical analysis. In principle, arbitrary interactions between predictors can be included in LUR models based on additive regression smoothers by using additional degrees of freedom. Due to the infinitely large number of possible combinations of interaction depth and involved predictors, this is generally not advisable. Instead, including interaction effects should be guided by subject-matter knowledge on phenomena which impact dispersion and decay of different pollutants.

Clear advantages of data-driven modeling in LUR based on additive regression smoothers are: (i) the approach provides a data-driven way to validate conventional LUR models based on parametric polynomials or improve the models based on the modeling insights; (ii) no knowledge about physical and chemical processes is required; (iii) the approach is straightforward to apply and provides a comparable and reliable methodology for modeling background air pollution concentration levels – an aspect which was outlined by administrators of major cities across the European Union as an important obstacle when estimating background air pollution concentration levels (Viana et al., 2020); (iv) in contrast to techniques from statistical learning, which often yield black-box models, LUR modeling based on additive regression smoothers yields interpretable effects that can be visualized and investigated regarding their plausibility. Since we provide all codes, functions, and datasets employed in the empirical analysis in a freely accessible online repository, our work is fully reproducible. We provide a comprehensive illustration of the employed modeling approach, which forms the basis for further analysis and can be extended in various directions such as different regions and pollutants.

4.5 Conclusions

LUR modeling based on additive regression smoothers allows to account for typical characteristics of air pollution processes, i.e., local heterogeneity, potential nonlinearities, and spatial heterogeneity and dependence, in a flexible, data-driven way. We illustrated the approach based on mean annual NO_2 concentration levels, visualized the estimated effects, and evaluated the statistical performance of the models. Overall, our results indicate that mean annual NO_2 concentration levels observed at background and traffic/industrial monitoring sites stem from two different processes. Comparisons with LUR models based on parametric polynomials and a test for spatial misspecification support LUR modeling based on additive regression smoothers. Finally, we generated interpolation maps of resolution 1 x 1 km based on LUR models employing additive regression smoothers. The interpolation map of the total effects shows higher mean annual NO_2 concentration levels in major cities and surrounding agglomeration areas compared to rural locations. A similar effect is visible for the road traffic network connecting the different agglomeration areas.

4.6 Acknowledgements

We thank two anonymous referees, the editor, Harry Haupt, Joachim Schnurbus, the participants of the Urban Transitions 2018 in Sitges, Barcelona, Spain, and the participants of CFE-CMStatistics 2018 in Pisa, Italy, for many helpful comments, discussions, and remarks. All errors are ours.

4.7 Appendix

4.7.A Employed `R`-packages

Table 4.A.1 shows all employed `R`-packages, the corresponding package versions, release dates and the corresponding references. All packages given in the table below are hosted on "The Comprehensive R Archive Network" (CRAN; accessible via `https://cran.r-project.org/`). Using the package versions given in Table 4.A.1 together with the datasets and codes provided in the online repository `https://github.com/markusfritsch/smoothLUR` ensures full reproducibility of the reported results.

Table 4.A.1: Overview of employed R-packages, corresponding package versions, release dates, and references.

R-package	Version	Date	Reference
ape	5.4	2020-06-03	Paradis and Schliep (2019)
cowplot	1.1.0	2020-09-08	Wilke (2019)
data.table	1.13.0	2020-07-24	Dowle and Srinivasan (2020)
ggplot2	3.3.2	2020-06-19	Wickham (2016)
ggpubr	0.4.0	2020-06-27	Kassambara (2020)
mgcv	1.8-31	2019-11-09	Wood (2003, 2017)
raster	3.3-13	2020-07-17	Hijmans (2020)
RColorBrewer	1.1-2	2014-12-07	Neuwirth (2014)
rgdal	1.5-15	2020-08-04	Bivand et al. (2020)
rgeos	0.5-3	2020-05-08	Bivand and Rundel (2020)
RgoogleMaps	1.4.5.3	2020-02-12	Löcher (2020)
sp	1.4-2	2020-05-20	Pebesma and Bivand (2005); Bivand et al. (2013)
splitTools	0.2.1	2020-04-18	Mayer (2020)

4.7.B Supplementary material

4.7.B.1 Modeling based on parametric polynomials

Table 4.B.1 shows LUR models based on parametric polynomials for modeling mean annual NO_2 concentration levels and combines Tables 4.3 and 4.5. The models were fitted according to the ESCAPE procedure. The set of potential predictors employed structural and spatial predictors. Similar to recent LUR studies of Beelen et al. (2013); Eeftens et al. (2016); Wolf et al. (2017), we only considered parametric polynomials of degree one. The table shows models fitted based on all monitoring sites (parA), background monitoring sites (parB), and traffic/industrial monitoring sites (parTI). All models displayed in Table 4.B.1 include PopDens, at least one predictor reflecting road traffic intensity, and at least one spatial predictor. Model parB provided the best fit to the data and $\overline{R}^2$ was similar to values reported in literature (see, e.g., Beelen et al., 2013). According to Moran's I statistic, there were indications for model misspecification for parA and parB (significance level $\alpha = 0.1$). We also fitted LUR models based on parametric polynomials without spatial predictors. Compared to the results shown in Table 4.B.1, $\overline{R}^2$ was lower and there were similar indications for model misspecification.

Model misspecification may result from neglected nonlinearities in one or multiple predictors (i.e., inappropriate chosen functional form). Fig. 4.B.1 displays a diagnostic tool for

investigating parB for neglected nonlinearities: Partial residual plots. The plots are based on modeling results from parB and constructed by plotting the sum of the residuals of parB and partial fits of predictors ($\widehat{\beta}_i \cdot z_i$ for structural predictors and $\widehat{\beta}_j \cdot x_j$ for spatial predictors) against the corresponding predictor values (z_i or x_j). The dashed bronze line results from fitting a simple linear regression model to the data, while the solid darkgreen line displays the fit of a univariate smooth. The partial residual plots in Fig. 4.B.1 suggest that there are nonlinearities in the data that parB cannot account for and that the functional form of the model may be misspecified. For models parA and parTI, there were no clear indications for nonlinearities.

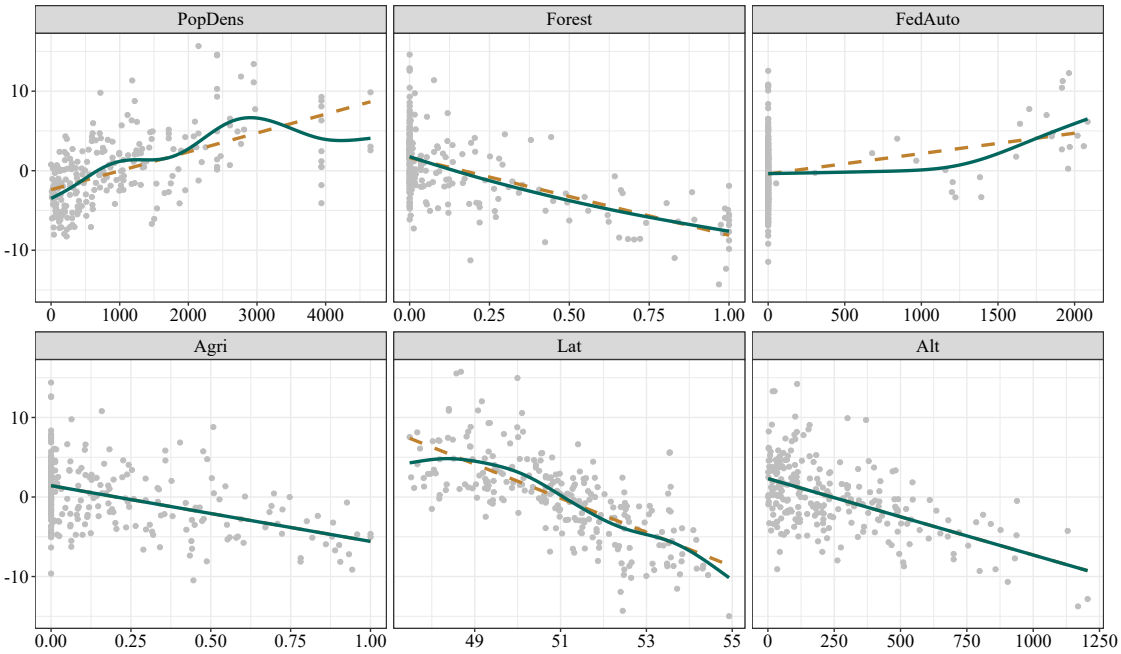


Figure 4.B.1: Partial residual plots derived from LUR model based on parametric polynomials parB, which includes structural and spatial predictors; for respective predictor, ordinate refers to sum of residuals of parB and values of $\widehat{\beta}_i \cdot z_i$ (for structural predictors) or $\widehat{\beta}_j \cdot x_j$ (for spatial predictors); abscissa refers to values of z_i or x_j; dashed bronze line indicates linear fit, smooth darkgreen curve depicts univariate smooth fit for respective scatterplot.

Table 4.B.1: LUR models based on parametric polynomials parA, parB, and parTI fitted by ESCAPE procedure using parametric polynomials of degree one; set of potential predictors includes structural and spatial predictors; parA uses all, parB background, and parTI traffic/industrial monitoring sites for model fitting; $\overline{R}^2$ and statistic to test for spatial autocorrelation in error terms (Moran's I) given for both models.

parA			parB			parTI		
Pred.	Est.	SE	Pred.	Est.	SE	Pred.	Est.	SE
(Int.)	16.795	2.425	(Int.)	129.151	9.894	(Int.)	97.005	26.505
PopDens	0.006	0.001	PopDens	0.002	< 0.001	PopDens	0.008	0.001
PriRoad	0.002	< 0.001	Forest	-9.760	1.061	PriRoad	0.002	0.001
SecRoad	0.002	0.001	FedAuto	0.003	0.001	HighDens	20.503	7.056
HighDens	41.810	5.382	Agri	-7.013	0.968	LowDens	17.446	3.864
LowDens	13.160	1.997	Lat	-2.143	0.190	Lat	-1.651	0.512
Lon	-0.988	0.216	Alt	-0.010	0.002			
$\overline{R}^2$	0.57			0.76			0.54	
Moran's I (p-value)	-0.18 (< 0.001)			0.19 (< 0.001)			0.12 (0.06)	

4.7.B.2 Modeling based on additive regression smoothers

We estimated three different LUR models based on additive regression smoothers using data observed at different types of monitoring sites: smoothA (all), smoothB (background), and smoothTI (traffic/industrial monitoring sites). We obtained the LUR models, by modeling all structural and spatial effects in Equation (4.3) via regression splines. For all structural predictors and the spatial predictor Alt, we used univariate smooths, while we modeled Lon and Lat via a bivariate smooth. Note that the employed smoothers choose the curvature of individual components data-driven; this may lead to smooths being effectively reduced to parametric polynomials of degree one.

Partial effects of individual univariate smooths for specification smoothB and corresponding 95% confidence intervals are shown in Fig. 4.B.2. For univariate smooths, we found: (i) nonlinearities in PopDens; (ii) effect of predictors HighDens, Ind, Forest, PriRoad, and FedAuto effectively reduced to parametric polynomials of degree one (edf roughly 1); (iii) predictors Transp, UrbGreen, Agri, SecRoad, and LocRoute were smoothed out (edf roughly 0). There were no indications for nonlinear effects for smoothTI.

The left plot of Fig. 4.B.3 displays the part of estimated background mean annual NO_2 concentration levels attributed to the spatial component only, when simultaneously accounting for all structural predictors. The map visualizes the sum of the spatial effects of smoothB consisting of a bivariate smooth of Lon, Lat and a univariate smooth of Alt. To obtain the maps, two steps were required: First, a LUR model based on additive regression smoothers was fitted to mean annual NO_2 concentration levels. Second, predicted values for conditional mean annual NO_2 concentration levels were computed based on observed predictors for each $1\,\mathrm{x}\,1\,\mathrm{km}$ grid cell in which no monitoring site was positioned. The interpolation maps visualize two types of values: Fitted values for grid cells that contain monitoring sites – when the location of the grid cell center and the monitoring site coincides – and predicted values for all other grid cells. In Fig. 4.B.3, values range from -11.0 to 5.3 and tend to decrease when moving from southwest to northeast. This highlights the presence of complex spatial association structures which are captured by the bivariate and univariate smooths of the spatial predictors. Conditional mean annual NO_2 concentration level estimates for a grid cell by smoothB resulted from adding all structural effects for a respective grid cell to the corresponding spatial effect and are visualized in the right plot of Fig. 4.B.3. Values ranged from 2.5 to 39.0. We replaced 385 out-of-range predictions

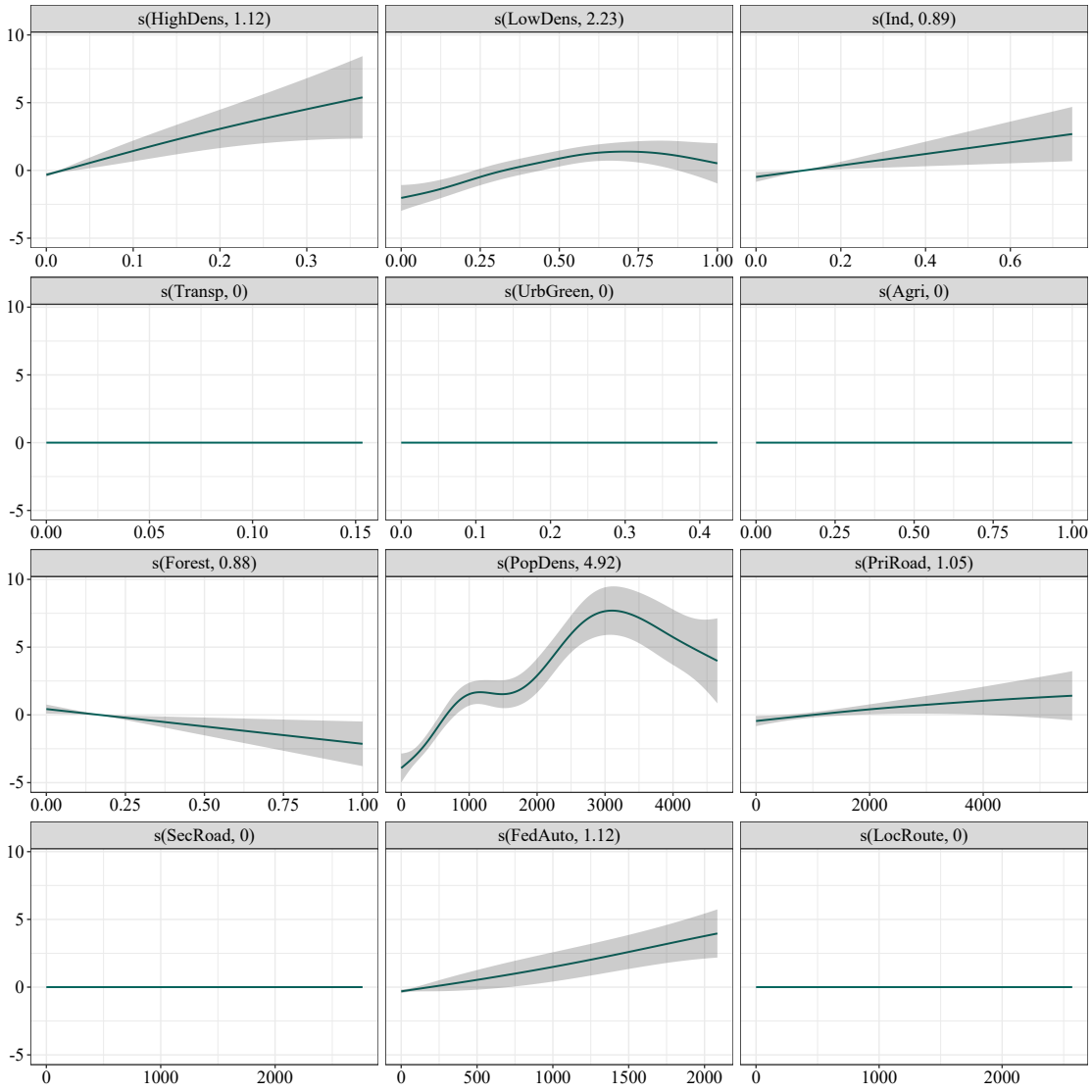


Figure 4.B.2: Partial effects derived from LUR model based on additive regression smoothers smoothB; darkgreen line depicts univariate thin plate regression spline estimate of $s_{u,p}$ for respective z_p, grey area marks corresponding 95 % pointwise confidence bands. Numbers in parentheses refer to estimated degrees of freedom (edf) for spline estimates and indicate curvature of effect of predictors (details, see Wood, 2017); roughness of spline effect increases with edf, where 1 corresponds to linear parametric effect; 0 implies predictor is smoothed out.

(0.1% of the 356, 791 predictions) by minimum or maximum observed mean annual NO_2 concentration level. Fig. 4.B.3 highlights that estimated mean annual NO_2 concentration levels are higher in cities and surrounding agglomeration areas compared to rural areas. Further, the road traffic network (in particular federal autobahn) linking agglomeration areas is visible across Germany through darker thin lines. This observation is in accor-

dance to Fig. 4.B.2, which indicates a positive partial effect of FedAuto on mean annual NO_2 concentration levels. Maps based on model smoothTI showed higher concentration levels and also highlighted agglomeration and infrastructure effects.

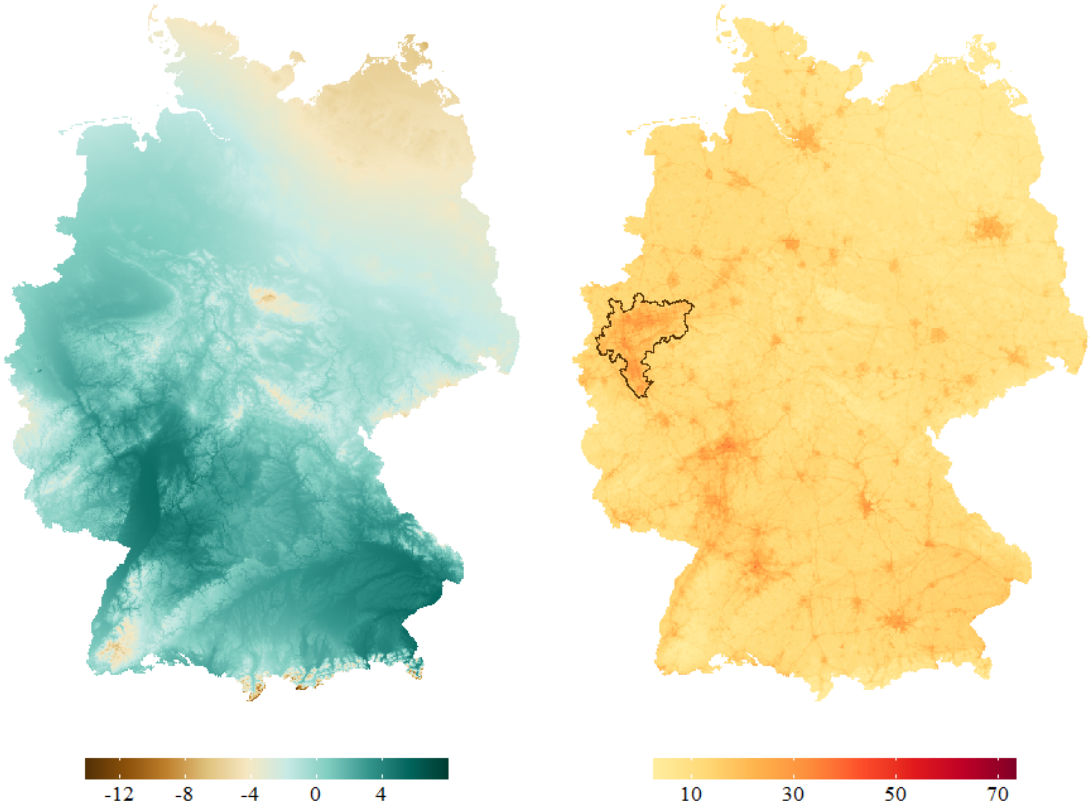


Figure 4.B.3: Interpolation maps derived from LUR model based on additive regression smoothers smoothB in 1 x 1 km resolution; all monitoring site types used for model fitting; maps visualize conditional mean annual NO_2 concentration levels (right) and part which is attributable to spatial effect (left); the latter consists of sum of bivariate smooth of Lon and Lat $s_b(X_{\text{Lon}}, X_{\text{Lat}})$ and univariate smooth of Alt $s_{u,A}(X_{\text{Alt}})$; out-of-range predictions replaced by minimum (maximum) observed mean annual NO_2 concentration level; border of Rhine-Ruhr metropolitan area depicted by darkbronze line.

The interpolation map displayed on the right-hand side in Fig. 4.6 is based on smoothA and highlights the effect of major cities, the surrounding agglomeration areas, and the connecting road traffic network for conditional mean annual NO_2 concentration levels. In contrast to Fig. 4.B.3, smoothA accounts for the road traffic network by including the predictors PriRoad and SecRoad into the model. For smoothTI, the interpolation maps also highlight major cities, surrounding agglomeration areas, and infrastructure effects and are not displayed here.

4.7.B.3 Interpretation of univariate and bivariate smooths

Consider the LUR model based on parametric polynomials given in Equation (4.4). The effect of a one unit increase of Z_i on Y (on average, c.p.) may be interpreted globally as

$$\frac{\partial \, \mathrm{E}(Y|\mathbf{Z}, \mathbf{X})}{\delta Z_i} = \beta_i. \tag{4.B.1}$$

For Equation (4.3), we suggest to interpret the effects of predictors based on differences in effects relative to corresponding changes in predictors. The effect of a change in Z_p on Y (on average, c.p.) in Equation (4.3) is

$$\frac{\partial \, \mathrm{E}(Y|\mathbf{Z}, \mathbf{X})}{z_{p_2} - z_{p_1}} = \frac{s_{u,p}(z_{p_2}) - s_{u,p}(z_{p_1})}{z_{p_2} - z_{p_1}}, \tag{4.B.2}$$

where z_{p_1}, z_{p_2} are located in the domain of Z_p; $s_{u,p}(Z_p)$ denote univariate smooth functions. The interpretation can be transferred to the bivariate smooth effect of $(X_{\mathrm{Lon}}, X_{\mathrm{Lat}})$ on Y in Equation (4.3) accordingly: The numerator contains the difference of two bivariate effects evaluated at the respective $(X_{\mathrm{Lon}}, X_{\mathrm{Lat}})$ and the numerator is a suitable distance metric (e.g., Euclidean or Manhattan distance of the two locations).

Fig. 4.B.4 exemplarily illustrates the interpretation of spatial effect $s_b(X_{\mathrm{Lon}}, X_{\mathrm{Lat}}) + s_{u,A}(X_{\mathrm{Alt}})$ (left plot) and structural effect $s_{u,p}(\mathrm{PopDens})$ (right plot) for two locations based on the proposition in Equation (4.B.2). At location 1 (Cologne city center), population density is higher than at location 2 (southern countryside of Mühlheim an der Ruhr). In order to derive the effect of a change in location, we divided the difference of the estimated spatial effects at the two locations by a suitable distance metric. For the univariate smooth of PopDens, we computed differences of the partial effect of PopDens evaluated at both locations; then, we divided this by the difference of PopDens between both locations. When differences of predictor values in the denominator of Equation (4.B.2) tend to zero, the effect of the change in location corresponds to the local slope of the approximating smooth and, therefore, has a similar interpretation as Equation (4.B.1) in a sufficiently small region.

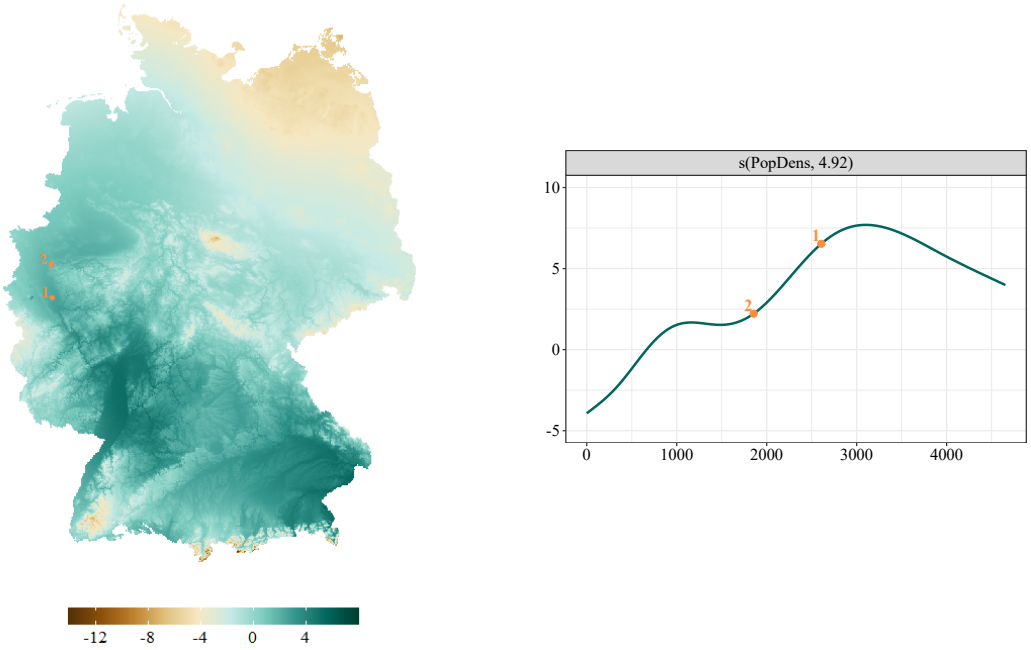


Figure 4.B.4: Estimated spatial effect consisting of $s_b(X_{\text{Lon}}, X_{\text{Lat}}) + s_{u,A}(X_{\text{Alt}})$ over Germany in 1 x 1 km resolution (left) and univariate smooth $s_{u,p}(\text{PopDens})$ – both derived from LUR models based on additive regression smoothers smoothB; orange dots indicate grid cell centers located in Cologne city center (point 1), southern countryside of Mühlheim an der Ruhr (point 2) (left) and corresponding population densities (right).

4.8 References

Alimissis, A., Philippopoulos, K., Tzanis, C.G., Deligiorgi, D., 2018. Spatial estimation of urban air pollution with the use of artificial neural network models. Atmospheric Environment 191, 205–213. doi:10.1016/j.atmosenv.2018.07.058.

Amini, H., Dehlendorff, C., Lim, Y.H., Mehta, A., Jørgensen, J.T., Mortensen, L.H., Westendorp, R., Hoffmann, B., Loft, S., Cole-Hunter, T., Bräuner, E.V., Ketzel, M., Hertel, O., Brandt, J., Jensen, S.S., Christensen, J.H., Geels, C., Frohn, L.M., Backalarz, C., Simonsen, M.K., Andersen, Z.J., 2020. Long-term exposure to air pollution and stroke incidence: A Danish nurse cohort study. Environment International 142, 105891. doi:10.1016/j.envint.2020.105891.

Arnol'd, V.I., 1957. On functions of three variables, in: Doklady Akademii Nauk SSSR, Russian Academy of Sciences. 679–681.

Atkinson, R.W., Butland, B.K., Anderson, H.R., Maynard, R.L., 2018. Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies. Epi-

demiology 29, 460–472. doi:10.1097/EDE.0000000000000847.

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. Science of the Total Environment 407, 1852–1867. doi: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. Atmospheric Environment 72, 10–23. doi:10.1016/j.atmosenv.2013.02.037.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W.W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A.S., Lanki, T., Turunen, A., Oftedal, B., Nystad, W., Nafstad, P., De Faire, U., Pedersen, N.L., Östenson, C.G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K.T., Overvad, K., Ellermann, T., Eeftens, M., Peeters, P.H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Hampel, R., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Künzli, N., Schindler, C., Schikowski, T., Adam, M., Phuleria, H., Vilier, A., Clavel-Chapelon, F., Declercq, C., Grioni, S., Krogh, V., Tsai, M.Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Brunekreef, B., Hoek, G., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project. The Lancet 383, 785–795.

Behm, S., Haupt, H., Schmid, A., 2018. Spatial detrending revisited: Modelling local

trend patterns in NO_2-concentration in Belgium and Germany. Spatial Statistics 2, 331–351. doi:10.1016/j.spasta.2018.04.004.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1–20. doi:10.1016/j.envsoft.2012.09.011.

Berrocal, V.J., Guan, Y., Muyskens, A., Wang, H., Reich, B.J., Mulholland, J.A., Chang, H.H., 2020. A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. Atmospheric Environment 222, 117130. doi:10.1016/j.atmosenv.2019.117130.

Bivand, R., Keitt, T., Rowlingson, B., 2020. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. URL: https://CRAN.R-project.org/package=rgdal. R package version 1.5-16.

Bivand, R., Pebesma, E., Gomez-Rubio, V., 2013. Applied spatial data analysis with R. Second ed., Springer, New York. doi:10.1007/978-1-4614-7618-4.

Bivand, R., Rundel, C., 2020. rgeos: Interface to Geometry Engine - Open Source ('GEOS'). URL: https://CRAN.R-project.org/package=rgeos. R package version 0.5-3.

BKG, 2015a. Federal Government for Geo-Information and Geodesy, DGM200 GK3 GRID-ASCII, GeoBasis-DE / BKG 2015. URL: https://gdz.bkg.bund.de/index.php/default/digitale-geodaten.html. Accessed on 27th October 2020.

BKG, 2015b. Federal Government for Geo-Information and Geodesy, VG250-EW Ebenen GK3 Shape, GeoBasis-DE / BKG 2015. URL: https://gdz.bkg.bund.de/index.php/default/digitale-geodaten.html. Accessed on 27th October 2020.

Brokamp, C., Jandarov, R., Rao, M.B., LeMasters, G., Ryan, P., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment 151, 1–11. doi:10.1016/j.atmosenv.2016.11.066.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H., Martin, R.V.,

Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environment International 130, 104934. doi:10.1016/j.envint.2019.104934.

Dowle, M., Srinivasan, A., 2020. data.table: Extension of 'data.frame'. URL: https://CRAN.R-project.org/package=data.table. R package version 1.13.0.

EEA, 2016. European Environment Agency CORINE Land Cover (CLC) 2012 raster data – Version 18.5.1 (09/2016). URL: http://land.copernicus.eu/pan-european/corine-land-cover/clc-2012/view. Acessed on 27th October 2020.

EEA, 2017. European Environment Agency, Air Quality e-Reporting. URL: b21a537e763e4ad9ac8ccffe987d6f77. Accessed on 27th October 2020.

Eeftens, M., Meier, R., Schindler, C., Aguilera, I., Phuleria, H., Ineichen, A., Davey, M., Ducret-Stich, R., Keidel, D., Probst-Hensch, N., Künzli, N., Tsai, M., 2016. Development of land use regression models for nitrogen dioxide, ultrafine particles, lung deposited surface area, and four other markers of particulate matter pollution in the Swiss SAPALDIA regions. Environmental Health 15, 53. doi:10.1186/s12940-016-0137-9.

EuroGeographics, 2018. Euroglobalmap (egm), v9.0. URL: https://eurogeographics.org/products-and-services/open-data/. Accessed on 27th April 2020.

Fallah-Shorshani, M., Shekarrizfard, M., Hatzopoulou, M., 2017. Evaluation of regional and local atmospheric dispersion models for the analysis of traffic-related air pollution in urban areas. Atmospheric Environment 167, 270–282. doi:10.1016/j.atmosenv.2017.08.025.

Fritsch, M., Behm, S., 2021a. Data for modeling nitrogen dioxide concentration levels across Germany. Submitted for publication.

Fritsch, M., Behm, S., 2021b. smoothLUR: Functions and data for smooth land use regression modeling. URL: https://github.com/markusfritsch/smoothLUR. R package version 0.1.1.

Gulliver, J., de Hoogh, K., Fecht, D., Vienneau, D., Briggs, D., 2011. Comparative assessment of GIS-based methods and metrics for estimating long-term exposures to air pollution. Atmospheric Environment 45, 7072–7080. doi:10.1016/j.atmosenv.2011.

09.042.

Hart, J.E., Yanosky, J.D., Puett, R.C., Ryan, L., Dockery, D.W., Smith, T.J., Garshick, E., Laden, F., 2009. Spatial modeling of PM_{10} and NO_2 in the continental United States, 1985–2000. Environmental Health Perspectives 117, 1690–1696. doi:10.1289/ehp.0900840.

Hastie, T., Tibshirani, R., 1990. Generalized additive models. Chapman and Hall, London. doi:10.1201/9780203753781.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction. Second ed., Springer, New York. doi:10.1007/978-0-387-84858-7.

Hijmans, R.J., 2020. raster: Geographic data analysis and modeling. URL: https://CRAN.R-project.org/package=raster. R package version 3.3-13.

Hoek, G., 2017. Methods for assessing long-term exposures to outdoor air pollutants. Current Environmental Health Reports 4, 450–462. doi:10.1007/s40572-017-0169-5.

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. Atmospheric Environment 42, 7561–7578. doi:10.1016/j.atmosenv.2008.05.057.

Hoek, G., Krishnan, R.M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J.D., 2013. Long-term air pollution exposure and cardio-respiratory mortality: A review. Environmental Health 12, 43. doi:10.1186/1476-069X-12-43.

de Hoogh, K., Korek, M., Vienneau, D., Keuken, M., Kukkonen, J., Nieuwenhuijsen, M.J., Badaloni, C., Beelen, R., Bolignano, A., Cesaroni, G., Pradas, M.C., Cyrys, J., Douros, J., Eeftens, M., Forastiere, F., Forsberg, B., Fuks, K., Gehring, U., Gryparis, A., Gulliver, J., Hansell, A.L., Hoffmann, B., Johansson, C., Jonkers, S., Kangas, L., Katsouyanni, K., Künzli, N., Lanki, T., Memmesheimer, M., Moussiopoulos, N., Modig, L., Pershagen, G., Probst-Hensch, N., Schindler, C., Schikowski, T., Sugiri, D., Teixidò, O., Tsai, M.Y., Yli-Tuomi, T., Brunekreef, B., Hoek, G., Bellander, T., 2014. Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. Environment International 73, 382–392. doi:10.1016/j.envint.2014.08.011.

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., Liu, Y., 2018. Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. Environmental Pollution 242, 675–683. doi:10.1016/j.envpol.2018.07.016.

Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S., Özkaynak, H., 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmospheric Environment 44, 3660–3668. doi:10.1016/j.atmosenv.2010.06.041.

Kassambara, A., 2020. ggpubr: 'ggplot2' based publication ready plots. URL: https://CRAN.R-project.org/package=ggpubr. R package version 0.4.0.

Kolmogorov, A.N., 1956. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. Doklady Akademii Nauk SSSR 108, 179–182.

Löcher, M., 2020. RgoogleMaps: Overlays on static maps. URL: https://CRAN.R-project.org/package=RgoogleMaps. R package version 1.4.5.3.

Lu, M., Schmitz, O., de Hoogh, K., Kai, Q., Karssenberg, D., 2020a. Evaluation of different methods and data sources to optimise modelling of NO_2 at a global scale. Environment International 142, 105856. doi:10.1016/j.envint.2020.105856.

Lu, M., Soenario, I., Helbich, M., Schmitz, O., Hoek, G., van der Molen, M., Karssenberg, D., 2020b. Land use regression models revealing spatiotemporal co-variation in NO_2, NO, and O_3 in the Netherlands. Atmospheric Environment 223, 117238. doi:10.1016/j.atmosenv.2019.117238.

Mayer, M., 2020. splitTools: Tools for data splitting. URL: https://CRAN.R-project.org/package=splitTools. R package version 0.2.1.

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). Atmospheric Environment 45, 4412–4420. doi:10.1016/j.atmosenv.2011.05.043.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. URL: https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-2.

Ostro, B., Lipsett, M., Reynolds, P., Goldberg, D., Hertz, A., Garcia, C., Henderson,

K.D., Bernstein, L., 2010. Long-term exposure to constituents of fine particulate air pollution and mortality: Results from the California Teachers Study. Environmental Health Perspectives 118, 363–369. doi:10.1289/ehp.0901181.

Paradis, E., Schliep, K., 2019. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528. doi:10.1093/bioinformatics/bty633.

Pebesma, E.J., Bivand, R., 2005. Classes and methods for spatial data in R. R News 5, 9–13. URL: https://CRAN.R-project.org/doc/Rnews/.

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 287, 1132–1141. doi:10.1001/jama.287.9.1132.

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/.

Rahman, M.M., Yeganeh, B., Clifford, S., Knibbs, L.D., Morawska, L., 2017. Development of a land use regression model for daily NO_2 and NO_x concentrations in the Brisbane metropolitan area, Australia. Environmental Modelling & Software 95, 168–179. doi:10.1016/j.envsoft.2017.06.029.

Ren, X., Mi, Z., Georgopoulos, P.G., 2020. Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environment International 142, 105827. doi:10.1016/j.envint.2020.105827.

Riley, P., 2019. Three pitfalls to avoid in machine learning. Nature 572, 27–29. doi:10.1038/d41586-019-02307-y.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511755453.

Russo, A., Raischel, F., Lind, P.G., 2013. Air quality prediction using optimal neural networks with stochastic variables. Atmospheric Environment 79, 822–830. doi:10.1016/j.atmosenv.2013.07.072.

Singh, K.P., Gupta, S., Rai, P., 2013. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment 80, 426–437.

doi:10.1016/j.atmosenv.2013.08.023.

Sinharay, R., Gong, J., Barratt, B., Ohman-Strickland, P., Ernst, S., Kelly, F.J., Zhang, J.J., Collins, P., Cullinan, P., Chung, K.F., 2018. Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls: A randomised, crossover study. The Lancet 391, 339–349. doi:10.1016/S0140-6736(17)32643-0.

Statistisches Bundesamt, 2020. Size of territory: Länder, reference date. URL: https://www-genesis.destatis.de/genesis/online. Accessed on 27th October 2020.

Tang, G., Zhao, P., Wang, Y., Gao, W., Cheng, M., Xin, J., Li, X., Wang, Y., 2017. Mortality and air pollution in Beijing: The long-term relationship. Atmospheric Environment 150, 238–243. doi:10.1016/j.atmosenv.2016.11.045.

Viana, M., de Leeuw, F., Bartonova, A., Castell, N., Ozturk, E., Ortiz, A.G., 2020. Air quality mitigation in European cities: Status and challenges ahead. Environment International 143, 105907. doi:10.1016/j.envint.2020.105907.

Vizcaino, P., Lavalle, C., 2018. Development of European NO_2 land use regression model for present and future exposure assessment: Implications for policy analysis. Environmental Pollution 240, 140–154. doi:10.1016/j.envpol.2018.03.075.

Wang, Y., Shi, L., Lee, M., Liu, P., Di, Q., Zanobetti, A., Schwartz, J.D., 2017. Long-term exposure to $PM_{2.5}$ and mortality among older adults in the southeastern US. Epidemiology 28, 207–214. doi:10.1097/EDE.0000000000000614.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Second ed., Springer, New York. doi:10.1007/978-3-319-24277-4.

Wilke, C.O., 2019. cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. URL: https://CRAN.R-project.org/package=cowplot. R package version 1.1.0.

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. Science of the Total Environment 579, 1531–1540. doi:10.1016/j.scitotenv.2016.11.160.

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65, 95–114. doi:10.1111/1467-9868.00374.

Wood, S.N., 2017. Generalized additive models: An introduction with R. Chapman & Hall/CRC Texts in Statistical Science. second ed., CRC press, Boca Raton. doi:10.1201/9781315370279.

Wu, J., Li, J., Peng, J., Li, W., Xu, G., Dong, C., 2015. Applying land use regression model to estimate spatial variation of $PM_{2.5}$ in Beijing, China. Environmental Science and Pollution Research 22, 7045–7061. doi:10.1007/s11356-014-3893-5.

Yanosky, J.D., Paciorek, C.J., Laden, F., Hart, J.E., Puett, R.C., Liao, D., Suh, H.H., 2014. Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. Environmental Health 13, 63. doi:10.1186/1476-069X-13-63.

Yanosky, J.D., Paciorek, C.J., Suh, H.H., 2009. Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the northeastern and midwestern United States. Environmental Health Perspectives 117, 522–529. doi:10.1289/ehp.11692.

Zhang, Z., Wang, J., Hart, J.E., Laden, F., Zhao, C., Li, T., Zheng, P., Li, D., Ye, Z., Chen, K., 2018. National scale spatiotemporal land-use regression model for $PM_{2.5}$, PM_{10} and NO_2 concentration in China. Atmospheric Environment 192, 48–54. doi:10.1016/j.atmosenv.2018.08.046.

5 Outlier detection and visualisation in multi-seasonal time series and its application to hourly nitrogen dioxide concentration

Abstract. Outlier detection in data on air pollutant recordings is conducted to uncover data points that refer to either invalid measurements or valid but unusually high concentration levels. As air pollutant data is typically characterised by multiple seasonalities, the task of outlier detection is associated with the question of how to deal with such non-stationarities. The present work proposes a method that combines time series segmentation, seasonal adjustment, and standardisation of random variables. While the former two are employed to obtain subseries of homoskedastic data, the latter ensures comparability across the subseries. Further, the standardised version of the seasonally adjusted subseries represents a scaled measure for the outlyingness of each data point in the original time series from its mean and therefore forms a suitable basis for outlier detection. In an empirical application to data on hourly NO_2 concentration levels recorded at a traffic monitoring site in Cologne, Germany, over the years 2016 to 2019, the common boxplot criterion is used to examine each standardised seasonally adjusted subseries for positive outliers. The results of the analyses are put into their natural temporal order and displayed in a heatmap layout that provides information on when single and sequential outliers occur.

Keywords. Outlier detection; Multiple seasonality; Time series segmentation; Boxplot criterion; Nitrogen dioxide

5.1 Introduction

Today, air pollution is one of the major concerns in regard to its adverse effects on ecology and human health with poor air quality particularly affecting urban dwellers. In terms of human health, nitrogen dioxide (NO_2) is the air pollutant of most relevance as it is a "key precursor of a range of secondary pollutants" (WHO, 2006, Ch. 12). There is empirical evidence that increased short-term exposure to NO_2 is associated with adverse health outcomes, e.g. admission to hospital for stroke or mortality from stroke (Shah et al., 2015) and a lower lung function (Rice et al., 2013; Panis et al., 2017; Dauchet et al., 2018; Strassmann et al., 2021). The latter studies refer to cohorts of healthy adults and areas with relatively low concentration levels that are mostly below the official WHO limit of $200\,\mu g/m^3$ for hourly and $40\,\mu g/m^3$ for annual NO_2 concentration levels (WHO, 2006, Ch. 12). To reduce air pollution burdens, monitoring and assessment of air quality is an ongoing task (Council of the European Union, 2008). Thereby, the detection of outliers, i.e. unusually high concentration levels, is of particular interest. A detected outlier indicates either unwanted data, e.g. measurement errors, or a valid but unusually high concentration level (Van Zoest et al., 2018). In the former case, outlier detection may act as an auxiliary tool for data validation (Čampulová et al., 2018). Outliers of the second category may help decision-makers in formulating effective mitigation strategies, insofar as the causes of outliers can also be investigated and efforts can be made to prevent their occurrence (Martínez Torres et al., 2020). Concerning outlier detection in NO_2 processes, the question arises on how to deal with the presence of non-stationarities, in particular multiple seasonalities, i.e. intra-day and intra-week cycles that are mainly driven by anthropogenic factors and an intra-year cycle that is rather caused by natural conditions (see, e.g., Behm and Haupt, 2020, and references therein).

In the literature, there are few studies on outlier detection in univariate hourly air pollutant processes that account for the cyclical behaviour in various ways.Van Zoest et al. (2018) examine outliers in hourly data on NO_2 over one year. They group the measurements into temporal categories to account for anthropogenically driven seasonalities. For every category, they construct confidence intervals where points falling outside the confidence interval are labelled outliers. Čampulová et al. (2017) and Čampulová et al. (2018) employ nonparametric regression to hourly data on particulate matter over one month. While the

former use methods from statistical process control to investigate the regression residuals for outliers, the latter divide the regression residuals into homogeneous segments via change point analysis and investigate each segment for outliers. Functional data analysis provides an alternative approach to outlier detection in hourly data on air pollutants (Febrero et al., 2008; Sguera et al., 2016; Martínez Torres et al., 2020). Thereby, the hourly data are summarised to, for example, daily curves that are called functional trajectories. Ahead of outlier detection, Febrero et al. (2008) and Sguera et al. (2016) separate the trajectories into two groups referring to weekdays and weekend days, while Martínez Torres et al. (2020) consider one group of trajectories. Outlier detection is carried out by using depth measures for functional data that quantify the centrality of a given curve within each group of trajectories.

The aim of the present work is to conduct outlier detection in multi-seasonal data based on a scaled measure for the deviation of each data point from its mean. For this objective, a generic method that relies on time series segmentation, seasonal adjustment, and standardisation of random variables is proposed and illustrated in an application to hourly data on NO_2 concentration levels recorded at a traffic monitoring site in Cologne, Germany, over the years 2016 to 2019. In a first step, the hourly time series is divided into $24 \cdot 7 = 168$ weekly subseries whereby each subseries refers to a specific combination of hour of day and day of week and exhibits only an intra-year cycle. The idea of dividing a multi-seasonal time series according to its cyclical patterns goes back to Gladyshev (1961) and is also discussed in, e.g., Jones and Brelsford (1967), Pagano (1978), and Franses (1994). In a second step, each subseries is seasonally adjusted using the framework of an additive components model, i.e. by decomposing the time series into a seasonal and a non-seasonal component, defining an appropriate estimate for the seasonal component, and subtracting this estimate from the time series (Bell and Hillmer, 1984). In the present study, the seasonal component is estimated by regressing the subseries on fourier terms. In a third step, the regression residuals are standardised which yields a scaled measure for the deviation of each data point in the respective weekly time series from its mean, i.e. a suitable basis for outlier detection. In the present study, the common boxplot criterion is chosen to examine each set of standardised regression residuals for positive outliers. As the data under consideration are already validated, the outliers detected with the proposed method correspond to hourly concentration levels that are unusually high for the combi-

nation of hour of day and day of week the subseries refers to. The results from the 168 separate analyses are put into their natural temporal order and presented in a heatmap layout which provides insights into when single and sequential outliers occur. Thereby, the length of an outlier sequence can be understood as an indicator for the persistence of unusually high concentration levels and long sequences may be of particular interest with respect to developing mitigation strategies, insofar as the causes thereof can be investigated. The statistical software `R`, version 4.0.5 (R Core Team, 2013), is used to conduct the empirical analyses and produce the visualisations. Details on the employed packages are given in the Appendix.

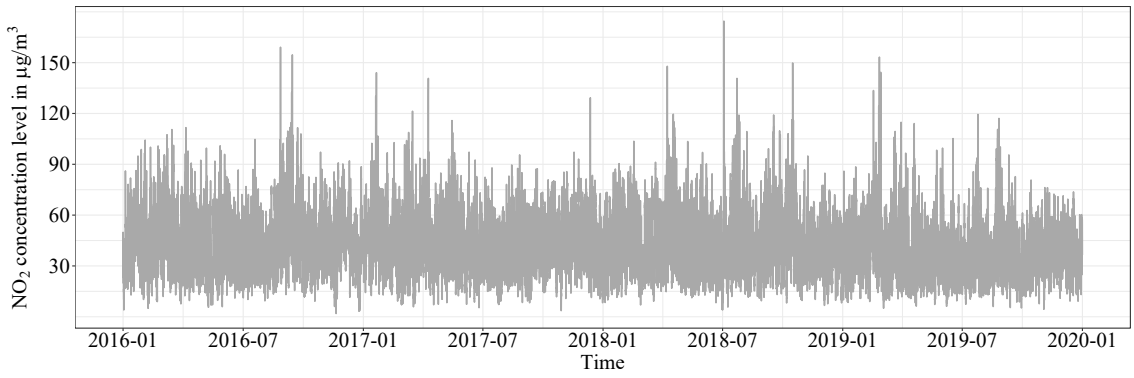
The remainder of this article is organised as follows: Section 5.2 introduces the heatmap layout and outlines the modelling framework. In Section 5.3, the empirical data are described and results of the empirical application are presented and discussed. Section 5.4 sums up, gives an outlook on potential extensions, and concludes.

5.2 Methods

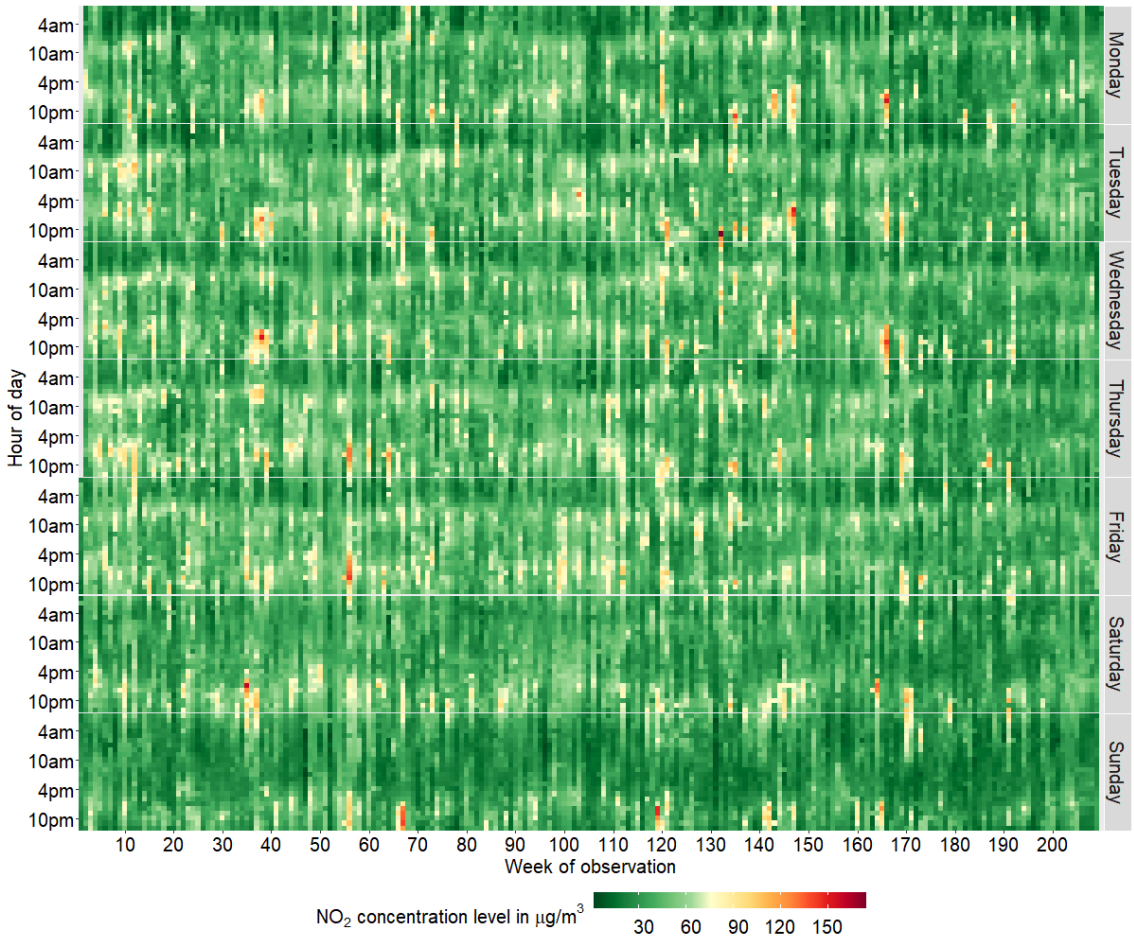
Let y_t, $t = 1, \ldots, T$, denote the observed hourly time series that is divided into subseries $y_{t(\iota,\nu)}$, $t(\iota,\nu) = 1(\iota,\nu), \ldots, T(\iota,\nu)$, where ι refers to hour of day ($\iota = 1, \ldots, 24$, corresponding to 1am, ...,12pm) and ν refers to day of week ($\nu = 1, \ldots, 7$ corresponding to Monday, ..., Sunday). Each weekly subseries is analysed separately, but the $24 \cdot 7 = 168$ subseries are put into their natural temporal order and presented in form of a heatmap to visualise the results of each analysis step.

5.2.1 Heatmap layout

Fig. 5.1 introduces the heatmap layout that is used throughout this work and refers to the time series of hourly data on NO_2 concentration levels recorded over years 2016 to 2019 at monitoring site Turiner Straße in Cologne. Fig. 5.1(a) shows a classical time series plot that provides little information about the time series apart from the presence of some obvious peaks. In particular, it remains unclear how the hourly data are distributed over the seasonal cycles. Fig. 5.1(b) provides a more comprehensive picture. It displays the hourly data in form of a heatmap and illustrates the distribution of the hourly data across hours of day, days of week, and weeks of observation. Each row refers to a specific combination of hour of day and day of week, i.e. each row represents one of the 168



(a) Time series plot.



(b) Heatmap where each row refers to specific combination of hour of day and day of week and each column refers to specific week of observation; cells are coloured according to recorded NO_2 concentration levels.

Figure 5.1: Time series plot (a) and heatmap (b) referring to hourly data on NO_2 concentration levels recorded over years 2016 to 2019 at monitoring site Turiner Straße in Cologne.

subseries, and each column refers to a specific week of observation. Each cell corresponds to one specific combination of date and hour and is coloured according to the recorded NO_2 concentration level. To ease the readability of the matrix, observations recorded on Mondays between midnight and 1am are arranged in the first row. As 1st January 2016 was a Friday the most left column is empty for Monday to Thursday, and, as 31st December 2019 was a Tuesday, the most right column is empty for Wednesday to Sunday. With regard to the present example, most cells in Fig. 5.1(b) are coloured green indicating relatively low or moderate NO_2 concentration levels. With decreasing proportions, cells are coloured yellow, orange, and red indicating that the number of observations decreases towards the upper tail of the distribution of the observed NO_2 concentration levels. The darkred cells correspond to the peaks that are visible in the time series plot with the maximum value referring to Tuesday, 3rd July 2018, 11pm, in observation week 132.

5.2.2 Modelling framework

Henceforth, each weekly time series $y_{t(\iota,\nu)}$, i.e. each row in Fig. 5.1(b), is examined for outliers separately. Therefore, remaining seasonalities are removed from each weekly time series and the seasonally adjusted data are standardised to obtain a scaled measure for the deviation of the weekly time series from its mean. Based on the standardised data, the cutoff, i.e. the upper fence of the common boxplot, is computed and data points exceeding this cutoff are labelled outliers. To simplify the formal exposition, ι and ν are omitted whenever the context is clear. For given ι and ν, the following assumptions are made.

Assumption 5.1 *The weekly time series y_t is additively separable and can be decomposed into a systematic term μ_t (seasonal component) and an error term e_t (non-seasonal component)*

$$y_t = \mu_t + e_t. \tag{5.2.1}$$

Assumption 5.2 *The systematic component μ_t in Eq. (5.2.1) corresponds to the sum of fourier terms*

$$\mu_t = \sum_{k=0}^{K} \left(a_k \cos(\omega_k t) + b_k \sin(\omega_k t)\right) = a_0 + \sum_{k=1}^{K} \left(a_k \cos(\omega_k t) + b_k \sin(\omega_k t)\right), \tag{5.2.2}$$

where $\omega_k = (2\pi k)/m$, $k = 1, \ldots, K$, and $m = 365.25/7$.

The parameter K in Eq. (5.2.2) determines the number of fourier terms that are superimposed to approximate the seasonal component μ_t and the coefficients a_k and b_k indicate the amplitude of the k-th cosine and sine curve, respectively. The idea to approximate the seasonal component of a time series by a sum of fourier terms goes back to Slutzky (1937) who showed that the summation of sine curves is a suitable way to express the regular behaviour of cyclic processes. It follows from Eq. (5.2.2) that the coefficients a_k and b_k are assumed to be time-invariant which is why the model stated in Eq. (5.2.4) is a fixed frequency regression model. Note that this fixed modelling design substantially differs from the harmonic regression model of Young et al. (1999) which allows for stochastic time-varying coefficients in Eq. (5.2.2).

Assumption 5.3 *The error term e_t in Eq. (5.2.1) follows a stationary $ARMA(p,q)$ process*

$$\phi(L)e_t = \theta(L)\epsilon_t, \tag{5.2.3}$$

where $\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$ and $\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$ are AR and MA lag polynomials of order p and q, respectively, and $\epsilon_t \overset{iid}{\sim} WN(0, \sigma_\epsilon^2)$ is white noise with $\mathrm{E}(\epsilon_t^4) < \infty, \forall t$.

In matrix notation, Eq. (5.2.1) can be stated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{5.2.4}$$

with $\boldsymbol{y}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ being a Tx1, a $(2K+1)$x1, and a Tx1 vector, respectively,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{2K-1} \\ \beta_{2K} \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \\ b_1 \\ \vdots \\ a_K \\ b_K \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_T \end{pmatrix},$$

and $\boldsymbol{X}$ the Tx$(2K+1)$ design matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,2K-1} & x_{1,2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{T,1} & x_{T,2} & \cdots & x_{T,2K-1} & x_{T,2K} \end{pmatrix} = \begin{pmatrix} 1 & \cos(\omega_1) & \sin(\omega_1) & \cdots & \cos(\omega_K) & \sin(\omega_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos(T\omega_1) & \sin(T\omega_1) & \cdots & \cos(T\omega_K) & \sin(T\omega_K) \end{pmatrix}.$$

The ordinary least squares (OLS) estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5.2.5}$$

is used to obtain the seasonally adjusted weekly subseries, i.e. the vector of residuals

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}. \tag{5.2.6}$$

Theorem 5.1 *The OLS estimator defined in Eq. (5.2.5) is an unbiased estimator of* $\boldsymbol{\beta}$.

Proof 5.1 *It follows from Assumption 5.2 that the design matrix* $\boldsymbol{X}$ *has deterministic entries and is of full rank. Further, it follows from Assumption 5.3 that* $\boldsymbol{e}$ *has zero mean. Thus, it is*

$$\mathrm{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\right] = \boldsymbol{\beta}.$$

Theorem 5.1 confirms that the OLS estimator is an appropriate choice to approximate the systematic component in Eq.(5.2.1) and thus to seasonally adjust the weekly subseries. While the original weekly time series is non-stationary and heteroskedastic, the set of regression residuals is homoskedastic and reflects the deviation of the data from its mean. To ensure comparability across the 168 series of regression residuals, each series of OLS residuals is scaled according to

$$\widehat{\boldsymbol{e}}^s = \frac{\widehat{\boldsymbol{e}}}{\widehat{\sigma}_e}, \tag{5.2.7}$$

with

$$\widehat{\sigma}_e = \left(\frac{1}{T}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}\right)^{1/2}. \tag{5.2.8}$$

As the OLS residuals have zero mean, the scaled residuals in Eq. (5.2.8) correspond to the standardised residuals. Basically, any distance-based outlier criterion is applicable to the standardised residuals $\widehat{\boldsymbol{e}}^s$. Using the most common boxplot criterion, a standardised

residual is labelled a positive outlier when it exceeds the cutoff c^U defined by

$$c^U = \mathrm{q}_{0.75}(\widehat{\boldsymbol{\epsilon}}^s) + 1.5 \cdot \mathrm{IQR}(\widehat{\boldsymbol{\epsilon}}^s), \tag{5.2.9}$$

with $\mathrm{q}_{0.75}(\cdot)$ and $\mathrm{IQR}(\cdot)$ denoting the upper quartile and the interquartile range, respectively. Outliers that are detected based on the cutoff defined in Eq. (5.2.9) correspond to outliers in the observed hourly time series.

Theorem 5.2 *Given Assumptions A.1-A.3, the estimator $\widehat{\sigma}_e$ defined in Eq. (5.2.8) is a consistent estimator of σ_e.*

Proof 5.2 *The proof relies on Wold's representation theorem and the law of large numbers for L^1-mixingales. Details are given in the Appendix.*

Theorem 5.2 provides a rationale for the standardisation of the residuals according to Eq. (5.2.7). Although, in the present work, the boxplot criterion is also applicable to non-standardised regression residuals and the main reason for standardising the residuals is to ensure comparability across the subseries, it should be mentioned that standardised variables are the general starting point for a wide range of approaches to outlier detection. The standardised version of the regression residuals represents a scaled measure for the deviation of each data point from its mean and therefore a scaled measure for the outlyingness of each data point which is why the standardised version of the residuals forms a suitable basis for outlier detection. An early work on outlier criteria based on standardised samples is published by Thompson (1935). The findings of Thompson (1935) and their extension by Pearson and Sekar (1936) are later used to develop two well known outlier tests, namely the extreme studentised deviate test, also known as Grubb's test (Grubbs, 1950), and the generalised extreme studendised deviate test, also known as Rosner's test (Rosner, 1983). For a comprehensive overview over outlier tests and issues involved with outlier detection, the interested reader is referred to Hawkins (1980) and Barnett and Lewis (1994).

Today, apart from outlier tests, the boxplot criterion that became popular with the work of Tukey (1977) is widely used for outlier detection. Its application is simple and does not require a priori information on the number of outliers to be detected. It is to be noted that the definition stated by Eq. (5.2.9) is one of many options to specify the boxplot-based cutoff point. When the interest lies in more extreme values, one could increase

the multiplier constant of the interquartile range or substitute the interquartile range by, e.g., the interdecile range. Using very high percentiles is not advisable as they might be distorted by outlying data points, a problem in outlier detection that is called masking (Hoaglin et al., 1986). Generally, the boxplot criterion is resistant against the problem of masking since it relies on the first, second, and third quartile. Though, the problem of masking might arise when the sample size is small or the data is highly skewed. Alternative boxplot-based definitions for such data settings are discussed in, e.g., Carling (2000); Hubert and Vandervieren (2008); Walker et al. (2018).

5.3 Empirical application

The empirical data refer to validated hourly NO_2 concentration levels (in $\mu g/m^3$) recorded at a traffic monitoring site in Cologne, Germany, over the years 2016 to 2019, and are retrieved from the air quality database provided by the European Environment Agency (EEA, 2021). The monitoring site is located in the northern part of the city center in the direction of the main train station on the sidewalk of a six-lane main street running from north to south. Its surrounding is continuously built up with multistory residential and commercial buildings.[1] The percentage of missing entries in the source data is equal to 5.03% and the longest sequence of missing values in the 168 weekly subseries is equal to three. Therefore, it is reasonable to impute missing values by applying the function `na.approx()` to each weekly subseries where leading/trailing missing values are substituted by the nearest non-missing value.

Fig. 5.2 illustrates the methodology described in Section 5.2 exemplarily for the subseries referring to Monday 8pm ($\iota = 20$ and $\nu = 1$). The function `tslm()` from the package `forecast` (Hyndman and Khandakar, 2008; Hyndman et al., 2021) is used to derive the OLS estimator $\widehat{\boldsymbol{\beta}}$ defined in Eq. (5.2.5). The number of harmonics K in Eq. (5.2.2) is chosen between one and four such that the AIC (Akaike Information Criterion) is minimised. In the top left display of Fig. 5.2, the time series of hourly NO_2 measurements is shown where the green curve depicts the harmonic fit $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The top right display refers to the series of regression residuals $\widehat{\boldsymbol{e}}$, while the two bottom displays refer to the standardised regression residuals $\widehat{\boldsymbol{e}}^s$. The cutoff based on the boxplot criterion c^U is depicted by the dashed

[1] The description of the monitoring sites' location is taken from `https://www.lanuv.nrw.de/umwelt/luft/immissionen/messorte-und-werte` (accessed 28th June 2021).

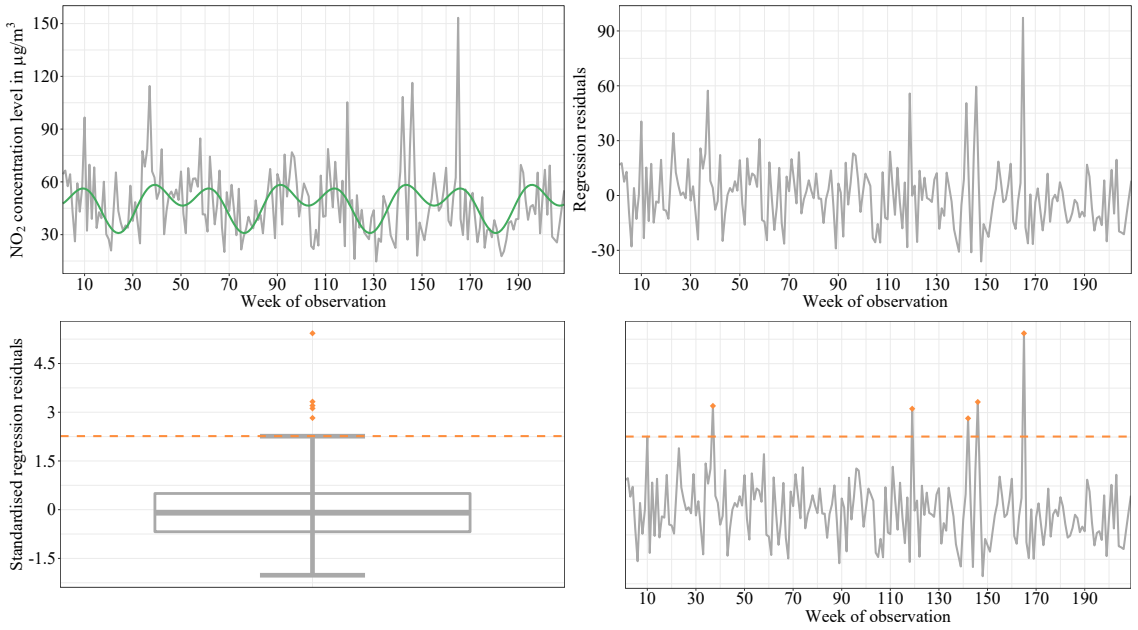


Figure 5.2: Illustration of the methodology for, exemplarily, the subseries referring to Monday 8pm; top left: time series of hourly NO_2 measurements and harmonic fit depicted by green curve; top right: series of regression residuals; bottom left: boxplot of standardised regression residuals; bottom right: series of standardised regression residuals; cutoff based on boxplot criterion depicted by dashed horizontal orange line and outliers marked by orange diamonds.

horizontal orange line. For the considered subseries, five single outliers are detected and marked by the orange diamonds.

Fig. 5.3 displays, in an analogous heatmap layout to Fig. 5.1(b), the harmonic fit $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ for $\iota = 1, \ldots, 24$ and $\nu = 1, \ldots, 7$. Placing the 168 subseries of estimated mean hourly NO_2 concentration levels back in their natural temporal order reveals some interesting patterns. While there is an overall tendency to lower values in summer[2], the yearly repetitive patterns strongly depend on day of week and hour of day. They are more prominent for weekdays as compared to weekend days and for morning and evening hours as compared to the remaining hours. These findings suggest that the intra-day and intra-week patterns of NO_2 concentration levels are strongly driven by anthropogenic factors. It is worth noting that although a separate model is estimated for each of the 168 subseries, the distribution of fitted values shown in Fig. 5.3 is smooth across the models with no visible discontinuity.

[2]The weeks of observation 23-36, 75-88, 127-140, and 179-192 refer to the German summer period June to August.

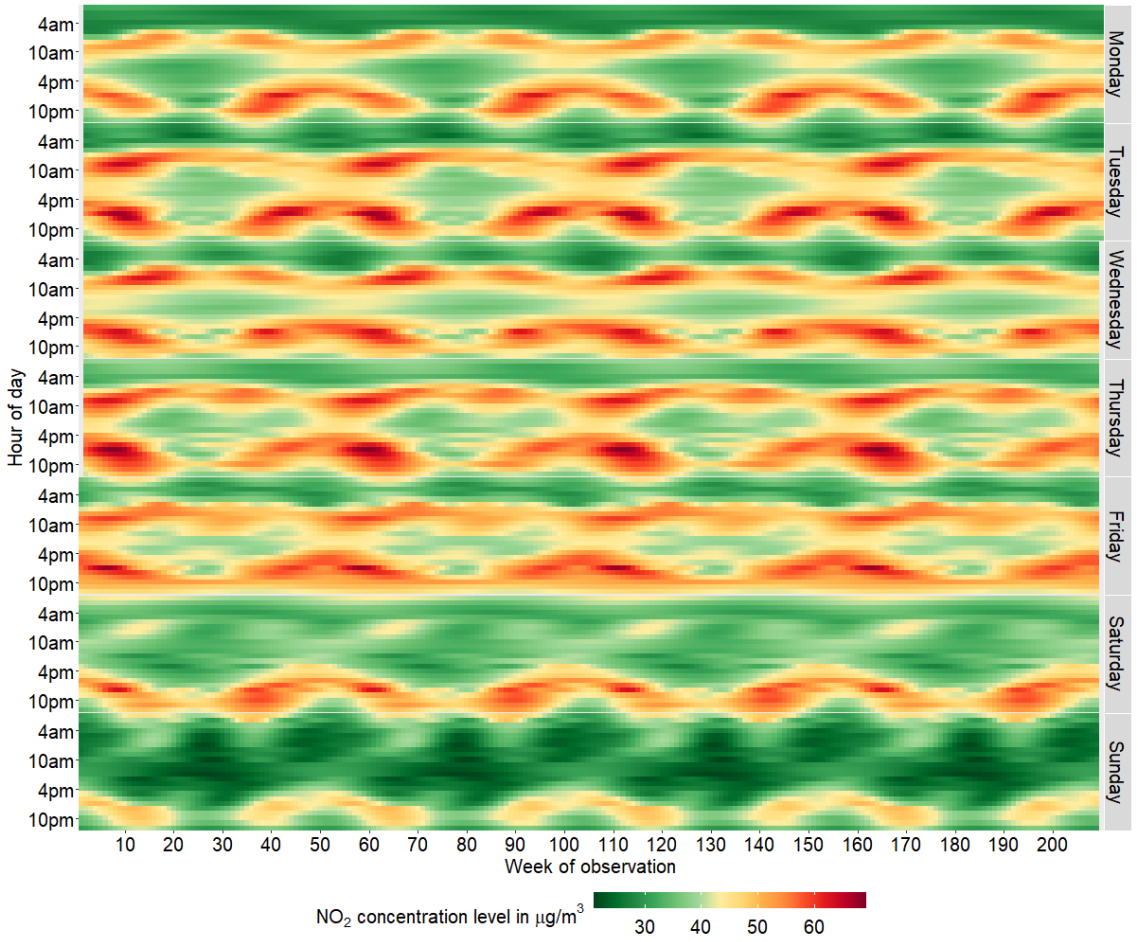


Figure 5.3: Heatmap showing distribution of estimated mean hourly NO_2 concentration levels over years 2016 to 2019 at monitoring site Turiner Straße in Cologne; each row refers to specific combination of hour of day and day of week; each column refers to specific week of observation; cells are coloured according to fitted values obtained from 168 harmonic regression models where a separate model is developed for each row, i.e. weekly subseries of recorded NO_2 concentration levels.

Fig. 5.4 displays a heatmap that illustrates the distribution of the standardised residuals according to Eq. (5.2.7), i.e. $\widehat{\boldsymbol{e}}^s = (\widehat{e}_1^s, \ldots, \widehat{e}_T^s)'$, and the occurrence of the outliers based on the boxplot criterion defined by the cutoff c^U in Eq. (5.2.9). Cells coloured green correspond to negative standardised residuals indicating an overestimation of the observed NO_2 concentration level for the specific combination of date and hour. From yellow to orange to red, the values of standardised residuals are positive and increase where outliers, i.e. standardised residuals for which it holds $\widehat{e}_t^s > c^U$, are marked by cells bordered black. Overall, it appears that there is a slight tendency to overestimation in the last year of the observation period (2019, weeks of observation 158-210). Referring to Fig. 5.1(b), this

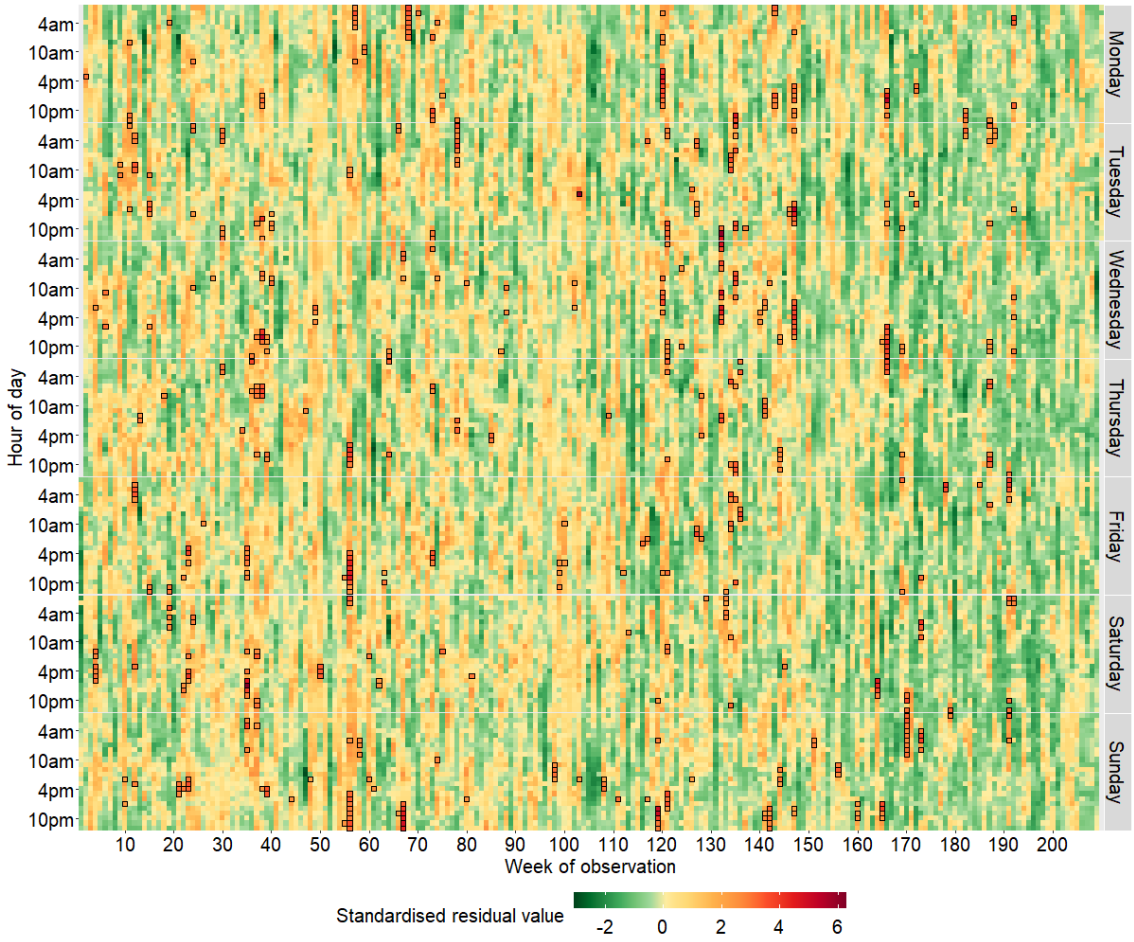


Figure 5.4: Heatmap showing distribution of standardised residual values over years 2016 to 2019 at monitoring site Turiner Straße in Cologne; each row refers to specific combination of hour of day and day of week; each column refers to specific week of observation; cells are coloured according to standardised residual values obtained from 168 harmonic regression models where a separate model is developed for each row, i.e. weekly subseries of recorded NO_2 concentration level; cells bordered black refer to outliers based on boxplot criterion.

may be due to the fact that very low NO_2 concentration values (dark green coloured cells) tend to be observed in 2019. Regarding the outliers, Fig. 5.4 provides for each outlier information on when it occurs and whether it is a single outlier, i.e. the preceding and succeeding value is a non-outlier, or an element of an outlier sequence, i.e. at least the preceding or succeeding value is also an outlier. Note that in order to provide this information not only the analysis results of the subseries from which the outlier originates are needed but also the analysis results of the subseries that refer to the preceeding and succeeding hours.

Taking together the results of the 168 separate analyses, a total of 555 outliers are detected, which corresponds to 1.58% of the data. For nine of the 168 subseries, no outlier is detected. Table 5.1 summarises how often outlier sequences of a certain length are detected. An outlier sequence of length equal to one is synonymous with a single outlier. According

Table 5.1: Overview over occurrence of detected outlier sequences depending on sequence length.

Length of outlier sequence	1	2	3	4	5	7	8	10	11	13
Number of occurrences	156	68	30	14	6	2	2	2	1	2

to the figures in Table 5.1, less than one third of the outliers (156) are single outliers while the majority of outliers (399) occurs in sequences of length equal or larger than two. In the observation period under consideration, there are two outlier sequences of length equal to 13 which means that 13 sequential hourly recordings of NO_2 concentration levels are outliers based on the boxplot criterion.

To improve the visibility of the outliers and ease the investigation of their occurrence, an alternative to Fig. 5.4 is given by Fig. 5.5. Cells are coloured according to the length of the outlier sequence they refer to. An outlier sequence of length equal to zero (grey) and one (lightgreen) is synonymous with a non-outlier and a single outlier, respectively. From yellow to orange to red, the length of the outlier sequence increases. The first (second) outlier sequence of length equal to 13 starts on Sunday, 22nd January 2017, at 4pm (Sunday, 9th April 2017, at 6pm) in week of observation 56 (67) and ends on Monday, 23rd January 2017, at 5am (Monday, 10th April 2018, at 7pm) in week of observation 57 (68). An accumulation of outliers can be observed on the weekdays Monday to Wednesday of observation weeks 120 to 150. The proportion of outliers in this period is 4.43 % which is remarkably higher than the average proportion of 1.58 %.

5.4 Discussion and conclusions

The introduced method for outlier detection in multi-seasonal time series combines time series segmentation, seasonal adjustment, and standardisation of random variables. Hourly data is divided into weekly subseries and each subseries is seasonally adjusted via harmonic regression. Standardisation of the regression residuals ensures comparability across the subseries and yields a scaled measure for the deviation of each data point from its mean, i.e. each standardised residual represents a scaled measure for the outlyingness

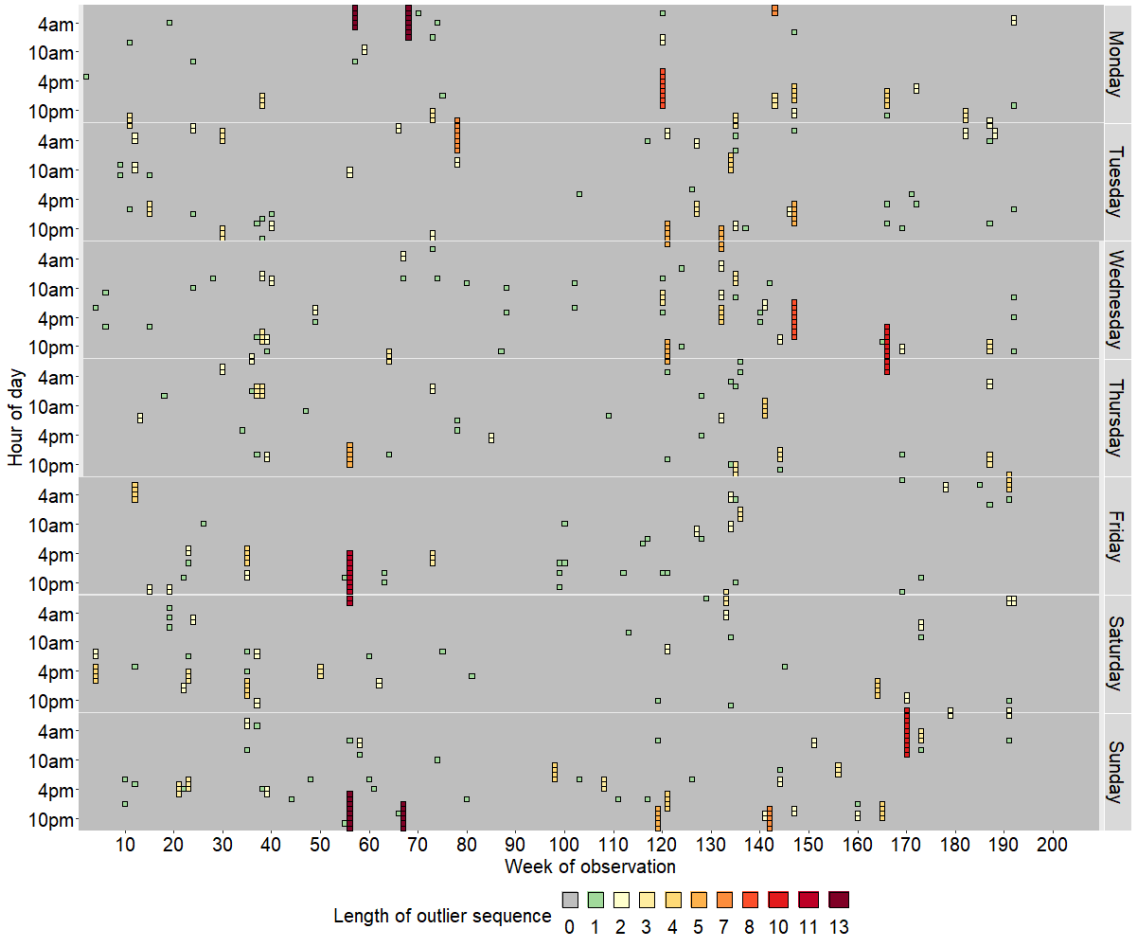


Figure 5.5: Heatmap showing distribution of outlier sequences over years 2016 to 2019 at monitoring site Turiner Straße in Cologne; each row refers to specific combination of hour of day and day of week; each column refers to specific week of observation; cells are coloured according to length of outlier sequence they refer to; length equal to zero (grey cells) and one (lightgreen cells) indicates non-outlier and single outlier, respectively.

of the corresponding data point in the weekly subseries. In an empirical application to hourly data on NO_2 concentration levels recorded at a traffic monitoring site in Cologne, Germany, over the years 2016 to 2019, the common boxplot criterion is used to examine the standardised regression residuals for positive outliers that correspond to unusually high concentration levels for given combination of hour of day and day of week. At every modelling step, a heatmap is presented where the 168 subseries are put into their natural temporal order. The structure of the heatmaps illustrates the temporal distribution of the observed data, the modelling fit, and the occurrence of outliers across observation weeks, days of week, and hours of day. In particular, the heatmaps provide information on when an outlier occurs and how long it persists, i.e. they visualise outlier sequences

of various lengths. In the data under consideration, a total of 555 positive outliers are detected, which corresponds to 1.58% of the data, while, for nine of the 168 subseries, no outliers are detected. Less than one third of the detected outliers are singles outliers and the longest sequence of detected outliers has a length equal to 13 which occurs twice in the period under consideration.

Dividing the hourly data on NO_2 concentration levels into 168 weekly subseries has some advantages. First, since the hourly series is subdivided to the extent that the subseries are free of intra-day and intra-week cycles one circumvents the ambiguity of choosing appropriate time windows in order to account for anthropogenically induced seasonalities. Second, the subseries only reveal one seasonality which facilitates the step of seasonal adjustment. For a time horizon of four years, each subseries still consists of over 200 observations such that estimating up to nine regression coefficients in a harmonic regression model with up to four fourier terms is reasonable and does not bear the risk of overfitting. Third, sequential observations in the weekly subseries are separated by a time span of a week which allows, from a climatological point of view, to assume potential autocorrelation to be negligibly small.

The introduced methodology relies on the assumption that a harmonic regression model with fourier terms and time-invariant regression coefficients is an appropriate choice to approximate the systematic component. From visual inspection of the heatmap showing the distribution of the estimated mean hourly NO_2 concentration levels (Fig. 5.3), the harmonic regression approach seems to work well. The heatmap reveals the typical seasonal patterns of NO_2 concentration levels and is smooth across the 168 models, with no visible discontinuities. Nevertheless, it is clear that any misspecification errors made in the regression step are passed on to the residuals which may cause data points to be erroneously identified as outliers. Evaluating the robustness of the introduced methodology against misspecification is beyond the scope of this paper and remains the subject of further research.

The proposed method can be extended into the spatial dimension by applying it to several spatial locations in a study region, e.g. to all monitoring sites in Cologne. This way of examining spatio-temporal data for outliers can provide insights into whether detected outliers and, in particular, long outlier sequences occur contemporaneously at several locations or whether they are spatially local events. Although the focus of the present

work lies on detecting positive outliers in hourly NO_2 dioxide concentration levels, the proposed method is also suitable for detecting negative outliers and applicable to any multi-seasonal time series data. Examples are, among others, data on other air pollutants, weather phenomena such as cloud cover, wind speed, and temperature, electricity demand and consumption, traffic volume, taxi rides, or call volumes to hospitals. In particular, the application of the introduced method to data that is known to be correlated with the NO_2 concentration level, e.g. metereological data, may uncover the causes for the outliers detected in the present work.

Overall, this paper presents a generic procedure for outlier detection in univariate time series with multiple seasonalities that forms a suitable basis for a variety of applications and extensions.

5.5 Acknowledgements

I thank Harry Haupt and Joachim Schnurbus for helpful comments and suggestions. All errors are my own.

5.6 Appendix A

R-packages

Detailed information on the employed R-packages is given in Table 5.A.1.

Table 5.A.1: Overview of employed R-packages, corresponding package versions, release dates, and references.

R-package	Version	Date	Reference
forecast	8.14	2021-03-11	Hyndman and Khandakar (2008); Hyndman et al. (2021)
ggplot2	3.3.2	2020-06-19	Wickham (2016)
gridExtra	2.3	2017-09-09	Auguie (2017)
lubridate	1.7.10	2021-02-26	Grolemund and Wickham (2011)
RColorBrewer	1.1-2	2014-12-07	Neuwirth (2014)
scales	1.1.1	2020-05-11	Wickham and Seidel (2020)
zoo	1.8-8	2020-05-02	Zeileis and Grothendieck (2005)

Proof of Theorem 5.2

The proof relies strongly on Hamilton (2020, Ch.7). Under Assumption 5.3 and according to the Wold representation theorem (Wold, 1938), the error term can also be stated as

$$e_t = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j}, \tag{5.A.1}$$

where $\epsilon_t \overset{\text{iid}}{\sim} WN(0, \sigma_\epsilon^2)$, $\Psi_0 = 1$, and $\sum_{j=0}^{\infty} \Psi_j^2 < \infty$. Using Eq. (5.A.1), the population variance can be written as

$$\begin{aligned}
\mathrm{E}(e_t^2) &= \mathrm{E}\left(\sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j} \sum_{j'=0}^{\infty} \Psi_{j'} \epsilon_{t-j'} \right) && \text{(5.A.2a)} \\
&= \mathrm{E}\left(\sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \Psi_j \Psi_{j'} \epsilon_{t-j} \epsilon_{t-j'} \right) && \text{(5.A.2b)} \\
&= \sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \Psi_j \Psi_{j'} \, \mathrm{E}\left(\epsilon_{t-j} \epsilon_{t-j'} \right). && \text{(5.A.2c)}
\end{aligned}$$

The last step (interchange of limits and expectations) is allowed because $\mathrm{E}(|\epsilon_{t-j}\epsilon_{t-j'}|) < \infty$ and $\sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} |\Psi_j \Psi_{j'}| = \sum_{j=0}^{\infty} |\Psi_j| \sum_{j'=0}^{\infty} |\Psi_{j'}| < \infty$. Define

$$\begin{aligned}
\eta_t &= e_t^2 - \mathrm{E}(e_t^2) && \text{(5.A.3a)} \\
&= \left(\sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \Psi_j \Psi_{j'} \epsilon_{t-j} \epsilon_{t-j'} \right) - \left(\sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \Psi_j \Psi_{j'} \, \mathrm{E}\left(\epsilon_{t-j} \epsilon_{t-j'} \right) \right) && \text{(5.A.3b)} \\
&= \sum_{j=0}^{\infty} \sum_{j'=0}^{\infty} \Psi_j \Psi_{j'} \left[\epsilon_{t-j} \epsilon_{t-j'} - \mathrm{E}\left(\epsilon_{t-j} \epsilon_{t-j'} \right) \right]. && \text{(5.A.3c)}
\end{aligned}$$

The expectation of η_t conditional on $\Omega_{t-m} = \{\epsilon_{t-m}, \epsilon_{t-m-1}, \ldots\}$ for $m > 1$ is given by

$$\mathrm{E}(\eta_t | \Omega_{t-m}) = \sum_{j=m}^{\infty} \sum_{j'=m}^{\infty} \Psi_j \Psi_{j'} \left[\epsilon_{t-j} \epsilon_{t-j'} - \mathrm{E}(\epsilon_{t-j} \epsilon_{t-j'}) \right]. \tag{5.A.4}$$

The expected absolute value of the expression in Eq. (5.A.4) is bounded by

$$\mathrm{E}|\mathrm{E}(\eta_t|\Omega_{t-m})| = \mathrm{E}\left|\sum_{j=m}^{\infty}\sum_{j'=m}^{\infty}\Psi_j\Psi_{j'}\left[\epsilon_{t-j}\epsilon_{t-j'} - \mathrm{E}(\epsilon_{t-j}\epsilon_{t-j'})\right]\right| \tag{5.A.5a}$$

$$\leq \mathrm{E}\left(\sum_{j=m}^{\infty}\sum_{j'=m}^{\infty}|\Psi_j\Psi_{j'}|\cdot|\epsilon_{t-j}\epsilon_{t-j'} - \mathrm{E}(\epsilon_{t-j}\epsilon_{t-j'})|\right) \tag{5.A.5b}$$

$$\leq \sum_{j=m}^{\infty}\sum_{j'=m}^{\infty}|\Psi_j\Psi_{j'}|\cdot M, \tag{5.A.5c}$$

for some $M < \infty$. With

$$\lim_{m\to\infty}\sum_{j=m}^{\infty}\sum_{j'=m}^{\infty}|\Psi_j\Psi_{j'}| = \lim_{m\to\infty}\sum_{j=m}^{\infty}|\Psi_j|\sum_{j'=m}^{\infty}|\Psi_{j'}| = 0,$$

it follows that η_t is an L^1-mixingale with respect to Ω_t. Further, η_t is uniformly integrable as

$$\mathrm{E}(\eta_t^2) = \mathrm{E}\left\{e_t^4 - 2e_t^2\,\mathrm{E}(e_t^2) + \left[E(e_t^2)\right]^2\right\} \tag{5.A.6a}$$

$$= \mathrm{E}(e_t^4) - \left[\mathrm{E}(e_t^2)\right]^2 \tag{5.A.6b}$$

$$= \sum_{j=0}^{\infty}\sum_{j'=0}^{\infty}\sum_{l=0}^{\infty}\sum_{l'=0}^{\infty}\Psi_j\Psi_{j'}\Psi_l\Psi_{l'}\,\mathrm{E}\left(\epsilon_{t-j}\epsilon_{t-j'}\epsilon_{t-l}\epsilon_{t-l'}\right) - \left[\sum_{j=0}^{\infty}\sum_{j'=0}^{\infty}\Psi_j\Psi_{j'}\,\mathrm{E}\left(\epsilon_{t-j}\epsilon_{t-j'}\right)\right]^2 \tag{5.A.6c}$$

$$< \infty. \tag{5.A.6d}$$

By the law of large numbers for L^1-mixingales it follows

$$\frac{1}{T}\sum_{t=1}^{T}\eta_t = \frac{1}{T}\sum_{t=1}^{T}\left[e_t^2 - \mathrm{E}(e_t^2)\right] \xrightarrow{\mathrm{p}} 0, \tag{5.A.7}$$

from which

$$\frac{1}{T}\sum_{t=1}^{T}e_t^2 \xrightarrow{\mathrm{p}} E(e_t^2). \tag{5.A.8}$$

Substituting e_t in Eq. (5.A.8) by the regression residuals yields the estimator given in Eq. (5.2.8). ■

5.7 References

Auguie, B., 2017. gridExtra: Miscellaneous functions for 'grid' graphics. URL: https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Barnett, V., Lewis, T., 1994. Outliers in statistical data. John Wiley and Sons, Chichester.

Behm, S., Haupt, H., 2020. Predictability of hourly nitrogen dioxide concentration. Ecological Modelling 428. doi:10.1016/j.ecolmodel.2020.109076. 109076.

Bell, W.R., Hillmer, S.C., 1984. Issues involved with the seasonal adjustment of economic time series. Journal of Business & Economic Statistics 2, 291–320. doi:10.2307/1391266.

Čampulová, M., Michálek, J., Mikuška, P., Bokal, D., 2018. Nonparametric algorithm for identification of outliers in environmental data. Journal of Chemometrics 32, e2997. doi:10.1002/cem.2997.

Čampulová, M., Veselík, P., Michálek, J., 2017. Control chart and six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}. Atmospheric Pollution Research 8, 700–708. doi:10.1016/j.apr.2017.01.004.

Carling, K., 2000. Resistant outlier rules and the non-Gaussian case. Computational Statistics and Data Analysis 33, 249—258. doi:10.1016/S0167-9473(99)00057-2.

Council of the European Union, 2008. Directive 2008/50/EC on ambient air quality and cleaner air for Europe. Official Journal of the European Communities .

Dauchet, L., Hulo, S., Cherot-Kornobis, N., Matran, R., Amouyel, P., Edmé, J.L., Giovannelli, J., 2018. Short-term exposure to air pollution: Associations with lung function and inflammatory markers in non-smoking, healthy adults. Environment International 121, 610–619. doi:10.1016/j.envint.2018.09.036.

EEA, 2021. European Environment Agency, Air Quality e-Reporting. URL: b21a537e763e4ad9ac8ccffe987d6f77. Accessed on 12th May 2021.

Febrero, M., Galeano, P., González-Manteiga, W., 2008. Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. Environmetrics 19, 331–345. doi:10.1002/env.878.

Franses, P.H., 1994. A multivariate approach to modeling univariate seasonal time series.

Journal of Econometrics 63, 133–151. doi:10.1016/0304-4076(93)01563-2.

Gladyshev, E.G., 1961. Periodically correlated random sequence. Soviet Mathematics 2, 385–388.

Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. Journal of Statistical Software 40, 1–25. doi:10.18637/jss.v040.i03.

Grubbs, F.E., 1950. Sample criteria for testing outlying observations. Annals Of Mathematical Statistics 21, 27–58. URL: https://www.jstor.org/stable/2236553.

Hamilton, J.D., 2020. Time series analysis. Princeton University Press. doi:10.1515/9780691218632.

Hawkins, D.M., 1980. Identification of outliers. Chapman & Hall, New York.

Hoaglin, D.C., I., B., T., J.W., 1986. Performance of some resistant rules for outlier labeling. Journal of the American Statistical Association 81, 991–999. URL: http://www.jstor.org/stable/2289073.

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis 52, 5186–5201. doi:10.1016/j.csda.2007.11.008.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2021. forecast: Forecasting functions for time series and linear models. URL: http://pkg.robjhyndman.com/forecast. R package version 8.14.

Hyndman, R., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 26, 1–22. doi:10.18637/jss.v027.i03.

Jones, R.H., Brelsford, W.M., 1967. Time series with periodic structure. Biometrika 54, 403–408. doi:10.2307/2335032.

Martínez Torres, J., Pastor Pérez, J., Sancho Val, J., McNabola, A., Martínez Comesana, M., Gallagher, J., 2020. A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. Mathematics 8, 225.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. URL: https://CRAN.R-project.org/package=RColorBrewer. R package version 1.1-2.

Pagano, M., 1978. On periodic and multiple autoregressions. The Annals of Statistics 6, 1310–1317. URL: http://www.jstor.org/stable/2958718.

Panis, L.I., Provost, E.B., Cox, B., Louwies, T., Laeremans, M., Standaert, A., Dons, E., Holmstock, L., Nawrot, T., De Boever, P., 2017. Short-term air pollution exposure decreases lung function: A repeated measures study in healthy adults. Environmental Health 16. doi:10.1186/s12940-017-0271-z.

Pearson, E.S., Sekar, C.C., 1936. The efficiency of statistical tools and a criterion for the rejection of outlying observations. Biometrika 28, 308–320. doi:10.2307/2333954.

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/.

Rice, M.B., Ljungman, P.L., Wilker, E.H., Gold, D.R., Schwartz, J.D., Koutrakis, P., Washko, G.R., O'Connor, G.T., Mittleman, M.A., 2013. Short-term exposure to air pollution and lung function in the Framingham Heart Study. American Journal of Respiratory and Critical Care Medicine 188, 1351–1357. doi:10.1164/rccm.201308-1414OC.

Rosner, B., 1983. Percentage points for a generalized ESD many-outlier procedure. Technometrics 25, 165–172. URL: http://www.jstor.org/stable/1268549.

Sguera, C., Galeano, P., Lillo, R.E., 2016. Functional outlier detection by a local depth with application to NO_x levels. Stochastic Environmental Research and Risk Assessment 30, 1115–1130. doi:10.1007/s00477-015-1096-3.

Shah, A.S.V., Lee, K.K., McAllister, D.A., Hunter, A., Nair, H., Whiteley, W., Langrish, J.P., Newby, D.E., Mills, N.L., 2015. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. BMJ 350:h1295. doi:10.1136/bmj.h1295.

Slutzky, E., 1937. The summation of random causes as the source of cyclic processes. Econometrica 5, 105–146. URL: http://www.jstor.org/stable/1907241.

Strassmann, A., de Hoogh, K., Röösli, M., Haile, S.R., Turk, A., Bopp, M., Puhan, M.A., Group, S.N.C.S., 2021. NO_2 and $PM_{2.5}$ exposures and lung lunction in Swiss adults: Estimated effects of short-term exposures and long-term exposures with and without adjustment for short-term deviations. Environmental Health Perspectives 129, 017009. doi:10.1289/EHP7529.

Thompson, W.R., 1935. On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard distribution. Annals of Mathematical Statistics , 214–219URL: https://www.jstor.org/stable/2957692.

Tukey, J.W., 1977. Exploratory data analysis. Volume 2. Reading, MA: Addison-Wesley.

Van Zoest, V.M., Stein, A., Hoek, G., 2018. Outlier detection in urban air quality sensor networks. Water, Air, & Soil Pollution 229, 111. doi:10.1007/s11270-018-3756-7.

Walker, M.L., Dovoedo, Y.H., Chakraborti, S., Hilton, C.W., 2018. An improved boxplot for univariate data. The American Statistician 72, 348–353. doi:10.1080/00031305.2018.1448891.

WHO, 2006. Air quality guidelines: Global update 2005. Copenhagen: WHO Regional Office for Europe.

Wickham, H., 2016. ggplot2: Elegant graphics for data nalysis. Springer-Verlag New York. doi:10.1007/978-0-387-98141-3.

Wickham, H., Seidel, D., 2020. scales: Scale functions for visualization. URL: https://CRAN.R-project.org/package=scales. r package version 1.1.1.

Wold, H., 1938. A study in the analysis of stationary time series. Ph.D. thesis. Almqvist & Wiksell. Stockholm.

Young, P.C., Pedregal, D.J., Tych, W., 1999. Dynamic harmonic regression. Journal of Forecasting 18, 369–394. doi:10.1002/(SICI)1099-131X(199911)18:6<369::AID-FOR748>3.0.CO;2-K.

Zeileis, A., Grothendieck, G., 2005. zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software 14, 1–27. doi:10.18637/jss.v014.i06.