River-Mediated Dynamic Environmental Factors and Perinatal Data Analysis

Jonathan Rathjens

2021

Dissertation

Department of Statistics
TU Dortmund University

Supervisor	Prof. Dr. Katja Ickstadt

Examination commission:

Chair	PD Dr. Jürgen Hölzer (Ruhr-University Bochum)
Referee	Prof. Dr. Katja Ickstadt
Referee	Prof. Dr. Andreas Groll
Assistant	Prof. Dr. Jörg Rahnenführer

Veröffentlicht als Dissertation in der Fakultät Statistik der Technischen Universität Dortmund,
Dortmund, Deutschland.

Published as a dissertation submitted to the Department of Statistics of TU Dortmund University,
Dortmund, Germany.

Acknowledgements

I would like to thank Professor Katja Ickstadt for her continual support as a supervisor and long-time mentor, the way she introduced me to the world of science, our many interesting joint research and teaching activities, and for always encouraging me to work on this project.

This work represents parts of the research carried out by our statistical-medical project group, in which we have enjoyed working together productively for many years. I give many thanks to all those involved and wish them all the best in continuing the work. In particular, I would like to thank Dr. Jürgen Hölzer from the Department of Medicine in Bochum for his manifold help, his advice and dedication in leading the project.

All our work would not have been possible without Arthur Kolbe, my medical counterpart in our group. We have shared all our thoughts and ideas for many years, which has been a tremendous support and guidance for me. I would also like to thank all the others, namely Eva Becker and Constanze Lüttke, who had diligently worked in our project in the past.

Professor Nadja Klein from the Humboldt University of Berlin supported me greatly in developing one of the two main parts of this work. I would like to thank her for her profound commitment in this cooperation and for always giving me great advice.

I would like to sincerely thank Professors Andreas Groll and Jörg Rahnenführer for their effort as reviewers and members of the examination commission, respectively. I wish them and the entire Department all the best for the future, and additionally, many thanks to my former colleagues for our time and ardent work together.

In behalf of our project group, I would like to thank Katharina Olthoff (d. 2019) and Dr. Sabine Bergmann from the NRW State Agency for Nature, Environment and Consumer Protection (LANUV) for their supportive and helpful cooperation, consultation and provision of data.

Many thanks to the Quality Assurance Office (qs-nrw) at the Medical Association Westphalia-Lippe, in particular Hans-Joachim Bücker-Nott and Heike Jaegers, for providing access to the perinatal registry data, their kind support and our valuable discussions.

Our project group and I kindly thank Stiftung Mercator for funding parts of our work.

Finally, I graciously thank my wife Larisa Rathjens and my dear friends and former colleagues Marie Beisemann, Esther Denecke and Leo Geppert for their meticulous proof-reading and many very helpful suggestions.

Jonathan Rathjens

Abstract

Perfluorooctanoic acid (PFOA) and related per- and polyfluoroalkyl substances, a group of man-made persistent organic chemicals employed for many products, are widely distributed in the environment. Adverse health effects may occur even at low exposure levels. A large-scale PFOA contamination of drinking water resources, especially of the river Ruhr, was detected in North Rhine-Westphalia, Germany, in summer 2006. Subsequent measurements are available from the water supply stations along the river and elsewhere. The first state-wide environmental-epidemiological study on the general population analyses these secondary data together with routinely collected perinatal registry data, to estimate possible developmental-toxic effects of PFOA exposure, especially regarding birth weight (BW).

Drinking water data are temporally and spatially modelled to assign estimated exposure values to the residents. A generalised linear model with an inverse link deals with the steeply decreasing temporal data pattern at mainly affected stations. Confirmed by a river-wide joint model, the river's segments between the main junctions are the most important factor to explain the spatial structure, besides local effects. Deductions from stations to areal units are made possible via estimated supply proportions.

Regression of perinatal data with BW as response usually includes the gestational age (GA) as an important covariate in polynomial form. However, bivariate modelling of BW and GA is recommended to distinguish effects on each, on both, and between them. Bayesian distributional copula regression is applied, where the marginals for BW and GA as well as the copula representing their dependence structure are fitted independently and all parameters are estimated conditional on covariates. While a Gaussian is suitable for BW, the skewed GA data are better modelled by the three-parametric Dagum distribution. The Clayton copula performs better than the Gumbel and the symmetric Gaussian copula, although the lower tail dependence is weak. A non-linear trend of BW on GA is detected by the standard polynomial model. Linear effects of biometric and obstetric covariates and also of maternal smoking on BW mean are similar in both models, while the distributional copula regression also reveals effects on all other parameters.

The local PFOA exposure is spatio-temporally assigned to the perinatal data of the most affected town of Arnsberg and so included in the regression models. No significant effect results and a relatively high amount of noise remains. Perspectively and for larger regions, this can be dealt with by exposure modelling on area level using dependence information, by allowing further asymmetry in the bivariate distribution of BW and GA, and by respecting geographical structures in birth data.

Keywords: distributional copula regression, drinking water contamination, perfluorooctanoic acid, perinatal registry data, river modelling, two spatial level data.

Contents

List of Tables

List of Figures

Preliminary remark

The text below, along with the major part of the tables and figures, is partly based on the articles by Rathjens *et al.* (2021b) and Rathjens *et al.* (2021a), which give some main statistical results of the reported research project, and of which the author of this dissertation is essentially responsible with regard to execution of research and writing.

Specifically, the following sections represent edited, partly extended and rearranged versions of the said publications:

- Sections 1.1, 1.2, 2.1, 2.2, 2.3.3, 2.4, 2.5 and 2.7, along with
  - Tables 2.2, 2.3, 2.4 and 2.6,
  - Figures 2.1, 2.2, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, A.1, A.2, A.3 and A.4,

  are partly based on Rathjens *et al.* (2021b).
- All sections from Chapter 3 as well as Sections 4.2.3, 4.3 and 4.4, along with
  - Tables 3.1, 3.3 and 3.4,
  - all figures from Chapter 3,

  are partly based on Rathjens *et al.* (2021a).

The corresponding individual source references are omitted in the following text.

Chapter 1

Introduction

This dissertation contributes to applied statistics in the fields of environmental medicine and epidemiology. A long lasting cooperation project has been established to investigate the possible toxic effect of per- and polyfluoroalkyl substances (PFASs), especially of perfluorooctanoic acid (PFOA), on foetal development. A secondary data analysis is conducted regarding the general population in the state of North Rhine-Westphalia (NRW), Germany, which has been affected by environmental contamination.

The results obtained so far, from the statistical perspective, are reported here. Although designed for a very specific data analysis, the approaches detailed below are general enough to be transferred to similar situations in other fields of application.

The research project has brought about a complex study to find specific solutions for exploration, modelling and inference in unusual, partly unique data situations. Major statistical issues include the spatial distribution and temporal progression of the environmental exposure, in conjunction with bivariate regression analysis of birth data. Both established and innovative statistical methods are applied.

Specifically, the exposure to PFOA is presumed to be mainly caused by drinking water, especially from the river Ruhr in NRW. The temporal progression of this dynamic contamination has to be modelled using routine measurement data, to obtain reliable, stable information, continuously over a long period of time. Various generalised linear and mixed models depending on time are applied to this. Furthermore, these values have to be spatially assigned in accordance with the water supply structure, respecting possible dependencies along a river, and to be realigned on the spatial level of perinatal registry data.

The birth data from the NRW state perinatal registry are analysed to estimate effects of the environmental factor and other explanatory variables on perinatal parameters. The data have been collected for quality assurance in obstetric health care. Complete records from many years are available. To investigate effects of PFASs on foetal development, the birth weight is of primary interest. However, to gain deeper insights into the relationships between the variables involved, the birth weight is bivariately modelled together with the gestational age, all conditional on biometric and medical covariates as well as on the spatio-temporally assigned environmental exposure. A recent Bayesian distributional regression approach by Klein and Kneib (2016) using copulas is applied, for the first time to birth data with two continuous response variables. While no significant PFOA effect is observed in this way, the bivariate analysis enables some insights from the perspective of obstetric research.

Below, a review on the relevance of PFASs in the environment and their physiological effects is given, including potential developmental toxicity, which motivates the further research. Section 1.2 summarises the major PFASs contamination event in NRW prior to 2006. The subsequent environmental medical research project is detailed in Section 1.3, including an account of a human biomonitoring cohort study in the most affected town of Arnsberg, an overview of the available state-wide secondary data sets, the major research questions and aims, and an overview of other sub-projects, most of them further detailed below. An overview of the whole dissertation is given in Section 1.4.

1.1 Background on perfluorooctanoic acid (PFOA)

Per- and polyfluoroalkyl substances (PFASs) name a group of man-made persistent organic chemicals produced since the 1940's which are employed in a wide variety of products used by consumers and industry (OECD, 2004; ATSDR, 2018). Their biochemical persistence and stability have led to pollution of the environment worldwide and an internal exposure of the general human population (cf., e.g.,Buck *et al.*, 2011). The lead substances perfluorooctanoic acid (PFOA) and perfluorooctane sulphonic acid (PFOS) are readily absorbed after ingestion or inhalation, not metabolised and very slowly excreted via urine and faeces. Several studies report different estimates of the biological half-life of PFOA in the human organism, but usually of about two to three years (Table 4.1 on page 63).

The most important PFASs source for the human population is food (Fromme *et al.*, 2009). In regions with contaminated drinking water, according to Vestergren and Cousins (2009), tap water contributes to more than 75% of the estimated intake in comparison to other exposure pathways (like diet, air, consumer articles). In the highly affected region in NRW, consumption of contaminated drinking water is associated with increased PFOA concentrations in human plasma (Hölzer *et al.*, 2008).

The toxicity of PFASs is currently being discussed among toxicologists, epidemiologists and regulatory agencies. Developmental and liver toxicity (Lau *et al.*, 2004; Lau, 2012), elevated blood lipids (Frisbee *et al.*, 2010; Steenland *et al.*, 2009) and immunotoxicity (Grandjean *et al.*, 2012; Looker *et al.*, 2014; NTP, 2016) are assessed as major health effects associated with the exposure to PFOA and PFOS. Other relevant impairments to health, like the incidence of type 2 diabetes in humans (Sun *et al.*, 2018), have been reported.

In particular, analyses of animal experimental data (see Lau *et al.*, 2007 for a meta-analysis) and human epidemiological data concluded that developmental exposure to PFOA adversely affects human health; they are based on sufficient evidence of decreased foetal growth in both human and non-human mammalian species. In animal experiments, impairments of mouse mammary development at relatively low PFOA concentrations are known (Macon *et al.*, 2011). Human epidemiological studies indicate reduced birth weight with increasing internal PFASs exposure (Maisonet *et al.*, 2012), even if this is still at background level (Apelberg *et al.*, 2007; Fei *et al.*, 2007). Nolan *et al.* (2009) report associations of PFOA exposure with birth weight and gestational age for a case of contaminated drinking water. For a cohort of mothers and children selected from the general population, Whitworth *et al.* (2012) report slightly lower standardised birth weight values for the quartile of mothers with the highest PFASs concentrations in blood plasma. In a large epidemiological study, a delay in pubertal development was observed for children exposed to elevated levels of PFOA (Lopez-Espinosa *et al.*, 2011). Meta-analyses of animal and human epidemiological studies on the association of PFOA exposure with foetal growth can be found in Johnson *et al.* (2014), Koustas *et al.* (2014) and Lam *et al.* (2014). Bach *et al.* (2016) report reduced birth weight in a birth cohort with increased PFOS exposure. However, the mechanisms of toxicity have not yet been sufficiently clarified. Detailed toxicological assessments have been published lately (EPA, 2016; ATSDR, 2018; EFSA CONTAM, 2018; UBA HBM-Kommission, 2018; NJDEP, 2019).

The European Food Safety Authority (EFSA) lowered the tolerable weekly intake from 10,500 (1050) (EFSA, 2008) to 6 (13) ng per kg body weight for PFOA (PFOS) and concluded that a considerable proportion of the population currently is exceeding these levels (EFSA CONTAM, 2018). Only recently, EFSA also evaluated the exposure to the sum of four different PFASs and recommended an even lower tolerable weekly intake of 4.4 ng per kg body weight for the sum of PFOA, PFOS, PFHxS and PFNA (EFSA CONTAM, 2020). The German drinking water commission lowered their precautionary action value for pregnant women, breastfeeding mothers and infants under 24 months of age to 50 ng PFOA per litre of drinking water (Trinkwasserkommission, 2020); the lifelong tolerable guide value is currently set to 100 ng/l (Trinkwasserkommission, 2016). Among the effects evaluated by the EFSA and the drinking water commission are lipid metabolism, immunotoxicity and developmental effects.

The findings on possible developmental toxicity of PFASs motivate research concerning associations between maternal PFASs exposure before or during pregnancy and the newborn children's weight. As the drinking water exposure pathway predominates in regions, where it is considerably above background level, water contamination can be used as a surrogate marker of the human internal exposure (see Section 4.2.4 for a model verification).

1.2 Contamination in North Rhine-Westphalia (NRW)

A large-scale contamination of drinking water with PFOA and other PFASs has been discovered in the summer of 2006 in North Rhine-Westphalia (NRW), Germany (Skutlarek *et al.*, 2006). The principal cause has been the use of PFASs-polluted soil conditioner on more than 1300 farmlands with subsequent contamination of surface waters. Further PFOA-contaminations of both surface and groundwater by sewage plants and other sources like fire extinguishing foam have occurred in different parts of NRW (LANUV, 2011), though with minor effects on drinking water.

A particularly severe contamination has affected the river Ruhr and its tributary Möhne (Skutlarek *et al.*, 2006; Wilhelm *et al.*, 2008). Large quantities of PFASs from polluted soil conditioner have been washed out from a field near the town of Brilon and conducted to the Möhne. Both rivers are important resources of the regional drinking water: More than 20 water supply stations are installed along them (Figure 1.1), supplying a large region of NRW with about four to five million inhabitants, especially the Ruhr metropolitan area.

PFASs are generally not removed during riverbank filtration and conventional water processing; some water supply stations have subsequently installed activated charcoal filters in order to reduce the PFOA-concentrations in drinking water below the guidance value of the German drinking water commission (Trinkwasserkommission, 2016).

Since 2006, the NRW state environmental agency has implemented an extensive monitoring programme of PFASs in relevant environmental media, including soil, water and drinking water (LANUV, 2011). The latter constitutes the main part of the data set on PFASs concentrations in drinking water (Table 1.1) used in the subsequent environmental medical research project.

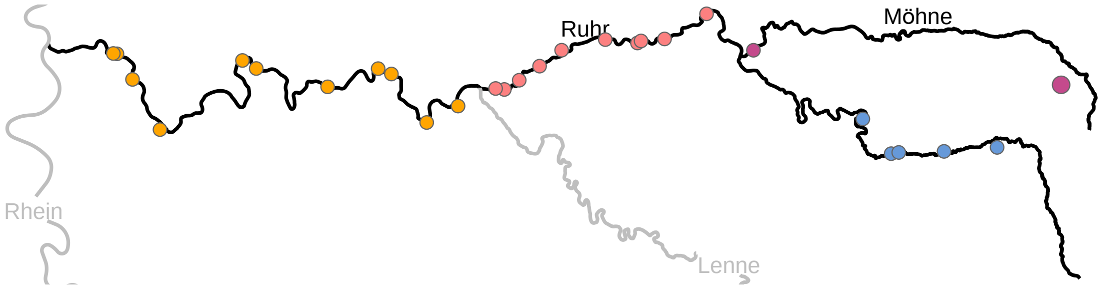

Figure 1.1: The river Ruhr and its tributaries Möhne and Lenne with water supply stations depending on these surface waters. An important contamination site has been discovered near the town of Brilon and the Möhne spring.

1.3 The 'PerSpat' project

After the discovery of the NRW contamination, the NRW State Agency for Nature, Environment and Consumer Protection (*Landesamt für Natur, Umwelt und Verbraucherschutz*; LANUV) has assessed the pollution by soil conditioner on farmlands near the town of Brilon (Wilhelm *et al.*, 2008; LANUV, 2011). The pollution has turned out to be extremely high and has mainly affected the rivers Möhne and Ruhr. The drinking water in parts of the town of Arnsberg is from the Möhne and had been severely contaminated for at least two years.

Therefore, a human biomonitoring cohort study (Hölzer *et al.*, 2008) on individual internal exposure (PFASs concentrations in blood serum) and consumption behaviour has been started in 2006 by the Department of Hygiene, Social and Environmental Medicine at the Ruhr-University Bochum (RUB), obtaining unique data on the human internal exposure in this region. These data include 90 children from Arnsberg who went to school that year, together with their mothers. The study has been extended to male adults from Arnsberg and control groups from comparable towns in the region not affected by PFASs contamination. Data from anglers consuming fish from lake Möhne, through which the contaminated river flows, have also been analysed (Hölzer *et al.*, 2011). Follow-up examinations of the biomonitoring cohort have been conducted several times in the following years (Hölzer *et al.*, 2009; Brede *et al.*, 2010) until 2017 (Bacher, 2020).

As the rivers Möhne and Ruhr are important for the drinking water supply of several million inhabitants in large parts of NRW, the project 'PerSpat' ('Perfluoroalkyl Spatial') has been started in 2014 to investigate the dynamics and the spatial distribution of PFASs in rivers and drinking water resources as well as potential health risks. It is a collaboration between the Department of Hygiene, Social and Environmental Medicine (PD Dr. Jürgen Hölzer) at the RUB and the Chair of Mathematical Statistics with Applications in Biometrics (Prof. Dr. Katja Ickstadt) at TU Dortmund University.

The LANUV has provided PFASs measurement data from the drinking water supply besides further information (overview in Table 1.1); the drinking water data set has been complemented with data collected from water suppliers by the Department of Hygiene, Social and Environmental Medicine.

The question of potential developmental toxicity is addressed by using data from the perinatal registry (overview in Table 1.2) provided by another cooperation partner, the quality assurance office (qs-nrw) located at the medical association Westphalia-Lippe.

Table 1.1: Data set overview for PFASs concentration measurements in the NRW drinking water.

Collection	measurements	various water suppliers
	management	LANUV, RUB
	purpose	protection of drinking water after contamination
	size	about 4500 measurements
	variables	PFOA, PFOS, further PFASs
Time	accuracy	day of sampling (partly with hour and minute)
	start	summer 2006 (for regions at risk)
	end	ongoing along river Ruhr
	further information	partly time of interventions, esp. filter installations
Space	water supply stations	about 700 in NRW, data from about 250
	water supply areas	about 450 in NRW, data from about 200
	further information	assignment of stations and areas; geodata of areas; importance of stations: e.g., location, water amount

Table 1.2: NRW perinatal data set overview.

Collection	measurements	hospitals and other childbirth institutions
	management	competent institute for quality assurance and transparency in health care in accordance with German § 137a SGB V
	purpose	quality assurance in health care
	size	about 150 000 observations per year, about 250 biometric, medical and other variables
	time	here: 2003–2014 (ongoing routine survey)
Variables	regarding newborn	e.g., weight, height, sex, Apgar score
	regarding mother	e.g., age, previous pregnancies, height, smoking; pregnancy risks, diseases: e.g., hypertension, (pregnancy) diabetes
	regarding pregnancy	e.g., expected date of birth, gestational age
	regarding delivery	e.g., mode, complications (soft-tissue injury, fever, . . .)
	Temporal structure	day of birth
	Spatial structure	mother's place of residence: about 850 postal code areas

The ‘PerSpat’ project is a large-scale study using these secondary data on the general population in NRW, to evaluate associations of Perfluorooctanoic Acid (PFOA) and other PFASs with perinatal parameters, especially the newborn children’s weight.

Major statistical issues include the modelling of concentration data from drinking water, spatio-temporal assignment of the resulting predictions to the perinatal registry data and advanced regression models in birth data analysis. Another question is whether PFOA contamination of the local drinking water is an appropriate surrogate marker (proxy) for the internal exposure of the residents, and how to model the pathway between them.

The main focus is on regions affected by the Möhne-Ruhr contamination incident, although other parts of NRW are also of interest, where some lower but detectable PFOA concentrations have been found besides many regions with no detection. Due to higher data variability, measurements considerably above critical values and its long half-life, PFOA is of primary interest among all available PFASs.

After preliminary works mainly on Markov random field modelling and spatial alignment issues (e.g., Goeken *et al.*, 2013), several further statistical theses have been written within the project and co-supervised by the author of this dissertation. Becker (2016) deals with drinking water data below the limit of quantification (details in Section 2.3.1) and preliminary modelling approaches including interpolation and regression with change points (Section 2.3.2). Becker (2017) continues the latter with a focus on Kriging (Section 2.3.4) and Gaussian processes in hydrology (Section 2.3.5) and gives reasons to presume homogeneity of contamination data within water supply areas (Section 2.3.3). Ganme (2017) investigates geographical structures in spatially aggregated birth weight data as well as their associations with air pollution and urbanity (Section 4.1.1). Kohlenbach (2019) explores the applicability of various spatial realignment approaches to the two main data sets (Section 4.2.1). Müller (2020) applies longitudinal mixed models to analyse the Arnsberg cohort data for associations of internal PFASs exposure and lipide concentrations in blood, separately for children and adults, with further covariates like age, height, weight, alcohol and tobacco consumption; the only significant effect is found for PFOA in children, but many values are missing and the need for imputation turns out to be an important issue.

Further work on the Arnsberg cohort has been performed, including Cox regression applied to puberty development and further longitudinal modelling of the internal exposure conditional on drinking water contamination and consumption behaviour (see Section 4.1.2 for the latter).

Main parts of the project have been presented at scientific conferences or published in scientific journals. Apart from the results reported below (Rathjens *et al.*, 2021b,a), this includes regression analyses of the complete perinatal data set (Kolbe *et al.*, 2016), pharmacokinetic modelling of the internal exposure to PFOA (Kolbe *et al.*, 2019b) as well as weighted geographical data realignment using population density data (Kolbe *et al.*, 2019a). Bacher (2020) reports a follow-up examination of the Arnsberg cohort, in which associations between PFASs exposure and the effectiveness of an influenza vaccination are studied, too.

1.4 Outline

The remainder of this dissertation is thematically structured, according to the two principal data sets and the respective analyses. The drinking water analyses are reported first, as they include modelling of the environmental exposure, such that their results are necessary for the further epidemiological research.

Within the chapters, the structure is partly chronological to illustrate the research process. Regarding both the drinking water and the perinatal data set, the respective chapters open with a data description (Sections 2.1 and 3.1), followed by a short review on potentially relevant methodological literature related to statistical challenges with the given data (Sections 2.2 and 3.2). Afterwards, the conducted analyses are detailed, followed by some conclusions, discussions and perspectives (Sections 2.7, 2.8, 3.6, 3.7 and 3.8).

Specifically, the analyses of drinking water contamination data in Chapter 2 start with the report of various preliminary studies (Section 2.3), which are necessary to prepare the main study, or to rule out the applicability of certain temporal and spatial analysis approaches. As first part of the main study, temporal generalised regressions models are fitted and evaluated in Section 2.4, to explore the dynamic PFOA measurements of the stations along the river Ruhr. In order to conclude from station-wise contamination values to those of the so-called water supply areas, the supply relationships are quantitatively estimated in terms of a matrix of shares (Section 2.5). A river-wide generalised mixed model for all Ruhr data is developed in Section 2.6.

The perinatal data regression analyses in Chapter 3 are restricted to the subset from the town of Arnsberg, for reasons of computability and the town's relevance. They begin with the standard univariate polynomial model (Section 3.3), followed by a comprehensive report of the innovative application of a Bayesian distributional copula model (Section 3.4), including a brief theoretical introduction, processing, model selection and evaluation. Both models' performances and results are compared in Section 3.5.

Chapter 4 approaches the original research question of a potential effect of PFOA on foetal development, in a spatial context. After reports of related sub-projects (Section 4.1), the spatio-temporal assignment of local PFOA concentrations in drinking water to the perinatal observations is detailed in Section 4.2. PFOA effect estimations using the Arnsberg perinatal subset are reported in Section 4.3, followed by some final remarks (Section 4.4).

More detailed outlines can be found at the beginning of the respective chapters. The final Chapter 5 closes with some more general considerations on the subject, complementary to the more specified issues dealt with in the respective closing sections of the chapters before.

Chapter 2

Drinking water contamination analyses

In this chapter, the PFOA measurement data from the drinking water supply of North Rhine-Westphalia (NRW), Germany, are described, explored, and finally modelled to obtain reliable and stable estimations of the average contaminations and their variability for any location and point of time. The resulting predicted values constitute the exposure variable used to estimate a possible PFOA effect in perinatal analyses, as reported in Chapter 4. But the temporal and spatial structures of persistent chemical contaminations in drinking water along a river are also investigated in their own right, to get insights in the relevant and universal characteristics of such data and to find model classes generally applicable in situations like this.

The main focus is on stations along the river Ruhr, where considerably more comprehensive data are available, a consequence of being affected by the Möhne-Ruhr contamination incident (Section 1.2). Among the available PFASs data, the focus is on PFOA: This is due to its higher data variability, measurements considerably above critical values, and its long half-life.

The data are described in more detail in Section 2.1, with an overview on their features and the numbers of observations as well as an outline of the river network in NRW and its usage for drinking water (Section 2.1.1); the typical temporal structures to be found in the various locations are differentiated in Section 2.1.2. Section 2.2 gives a brief overview of literature on modelling irregular measurement series, also with a spatial component, and such with a focus on river networks.

With this starting point, several preliminary studies and sub-projects within the project 'PerSpat' are reported in Section 2.3. They are necessary to prepare the main study, or to rule out the applicability of certain temporal and spatial analysis approaches. The first issue is the handling of measurements below the limit of quantification (Section 2.3.1). Preliminary approaches to model the temporal data from individual water supply stations, especially using kernel smoothing, are briefly explained in Section 2.3.2. To motivate joint spatial analyses within certain regional clusters of water supply stations, Section 2.3.3 points out the homogeneous behaviour of many neighbouring stations, especially when supplying the same area; here, the available data from the drinking water network are also used for comparison and evaluation. Several spatio-temporal approaches, which have not been considered further, are briefly named in Sections 2.3.4 (Kriging) and 2.3.5.

As first part of the main study, temporal generalised regression models (Section 2.4) are fitted to the individual stations' data (Section 2.4.1) and the spatial structure of model results along the river Ruhr is initially explored (Section 2.4.2). Section 2.5 focuses on the pathways in the water supply of the state: The shares of water are estimated to allow conclusions to be drawn from contamination values at stations to those of the so-called water supply areas. A river-wide generalised mixed model for all Ruhr data is developed in Section 2.6; based on the previous findings (Section 2.6.1), random effects are included in Section 2.6.2 and the river's structure is respected in an enhanced approach using Bayesian modelling (Section 2.6.3).

Conclusions from this chapter are summarised and discussed in Section 2.7, followed by an outlook on possible alternative modelling approaches (Section 2.8).

All data preprocessing steps, visualisations and analyses have been conducted using the `R` environment (R Core Team, 2020) along with its basic packages. Processing of geographical data and drawing of maps have been conducted using the `R` packages `rgdal` (Bivand *et al.*, 2019) and `tmap` (Tennekes, 2018). See Appendix B.2.1 for documentation.

2.1 Description of PFOA concentration data from NRW

The PFOA measurements (overview in Table 1.1 on page 5) have been conducted on behalf of the local drinking water suppliers; the data have been collected by the NRW State Agency for Nature, Environment and Consumer Protection (*Landesamt für Natur, Umwelt und Verbraucherschutz*; LANUV) and by the Department of Hygiene, Social and Environmental Medicine of the Ruhr-University Bochum.

2.1.1 Data amount and features

The data consist of measurements from completely purified, ready-to-drink water in both water supply stations (3349 observations) and networks (536) since summer 2006. Observations are available until 2016 and partially ongoing. Furthermore, there are some samples drawn from raw water (418) and during water treatment (87) at the stations. Many samples, especially the latter, are non-detects. See Figure A.2 on page 81 for a cross-sectional overview of the spatial distribution of PFOA concentrations. Table 2.1 shows the structure of the measurement data set.

Table 2.1: Structure (variables) of PFOA measurement data set.

Station:	name of water supply station
Area:	name of water supply area
Time:	day the sample has been taken
LoQ:	whether the concentration is under the limit of quantification
Value:	PFOA concentration or limit of quantification in ng/l
Type:	whether the sample is drawn from an area (network), from a station, from its raw water or during treatment

NRW is geographically divided in some 450 water supply areas (cf., Figure 2.1), with any data available from about 200 areas. For the stations, many of them being very small, any data are available from about 250 of 700 stations. There is a complex, neither injective nor surjective assignment between areas and stations; see Section 2.5 for the water supply proportions; minor changes in areal division, station running and assignment have occurred in the course of the years. The number of non-equidistant measurements per time period varies between locations and periods according to the contamination risk.

More detailed information on the amount of data per station and their ranges and temporal patterns can be found in Table 2.2. According to this, the data availability is satisfying for all affected stations,

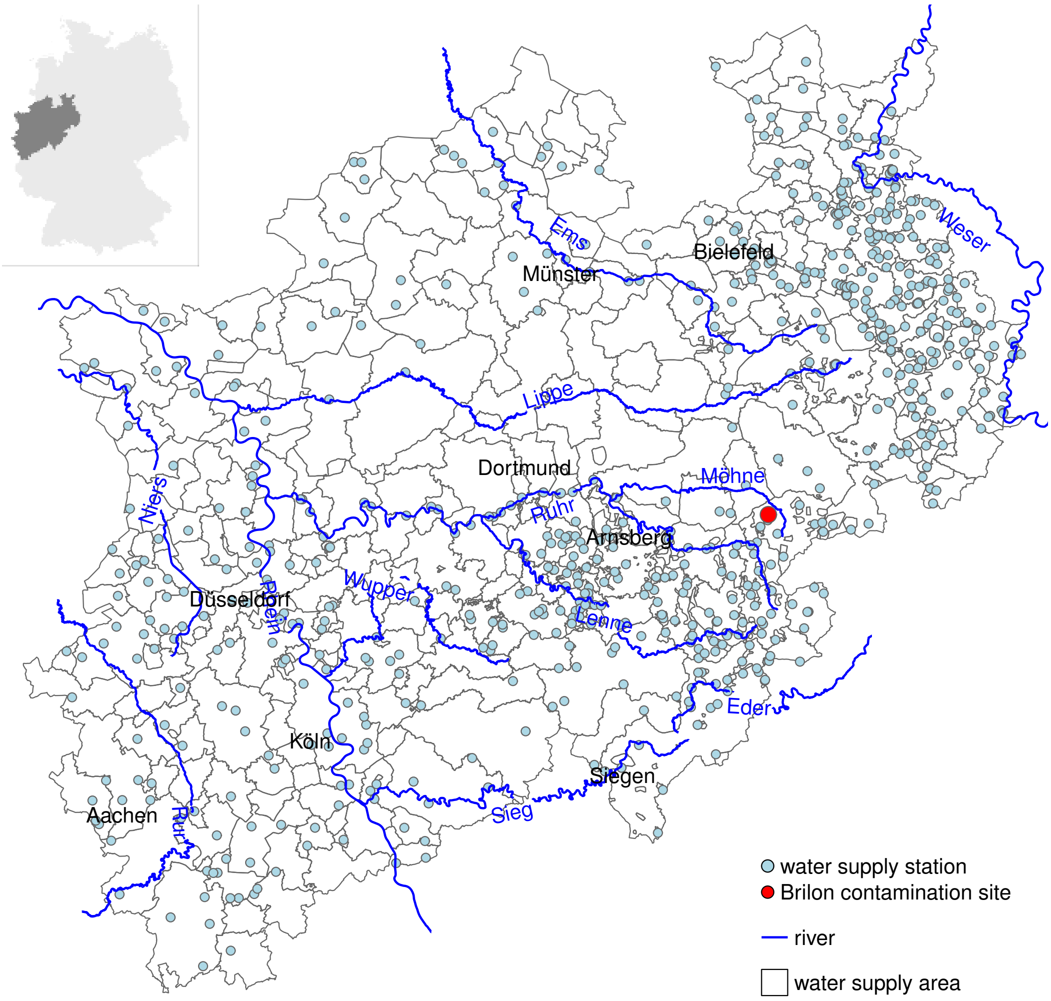

Figure 2.1: NRW divided into water supply areas, with water supply stations and the major rivers which are used as drinking water resources. An important contamination discovered in 2006 is located near the Möhne spring. (Above: NRW within Germany.)

i.e., along the river Ruhr (a total of 2484 observations at 27 stations), and all of them feature at least rather high values; stations not depending on the river Ruhr are much less observed, typically with low values, although there are exceptions. Table 2.3 details the data availability with respect to water supply areas: Where some of the stations supplying a particular area are Ruhr dependent, there are typically enough data to assess the area's contamination risk by deduction from the stations.

An important issue in preparation of modelling are exceptionally high values at the beginning of the observed period of time: The most affected supply station of Möhnebogen at the Ruhr's tributary Möhne near Arnsberg features a PFOA maximum of 640 ng/l. On the other hand, values below the limit of quantification (LoQ, mostly 10 ng/l) are frequent in less affected regions and later periods of time. A kind of discretisation is caused by a data accuracy of 10 ng/l frequently given.

Apart from the measurement data, there is further information from water supply operators on the most likely non-existence of PFOA contamination at certain stations or areas over longer periods of

Table 2.2: Number of water supply stations by data availability and temporal data pattern of the PFOA concentrations. (In this rough classification of typical observed situations, 'very high' means PFOA concentrations about and above the guide value of 100 ng/l, 'rather high' values are about and above 50 ng/l. LoQ: limit of quantification.)

Data availability	Temporal data pattern	Number of stations	
(or other information)		Ruhr	others
many (more than 25 data)	decrease, with very high values	17	1
	decrease, with rather high values	6	0
	diffuse, with rather high values	4	1
	all values below LoQ	0	1
few	single very high value	0	2
	decrease, with rather high values	0	3
	diffuse, with rather high values	0	1
	low values	0	16
	all values below LoQ	0	142
raw water / treatment	low values	0	4
	all values below LoQ	0	53
expert statement	(no contamination risk)	0	56
none	—	0	≈ 400
	total	27	≈ 675

time. For some of these places, there are a few values, or even just one, and these measurements usually confirm the water supplier's assumption of no relevant PFASs contamination.

With data being on two spatial levels, stations and areas, there is also information on their assignment and the supplied and demanded water amounts (used in Section 2.5).

The main rivers of NRW, with water supply stations along them, are shown in Figure 2.1. A graph of their directions of flow and their junctions (a simplified spatial structure, comprising of river segments) is shown in Figure A.1 on page 80. As explored in Sections 2.3.3 and 2.4, a certain homogeneity may be presumed for river-depending stations within the same river segment. Apart from the main rivers, there are many stations taking water from smaller tributaries, but with almost no measurement data available.

For the river Ruhr, there are five water supply stations upstream beyond the Möhne mouth, the station of Möhnebogen on the river Möhne, and twenty-one stations downstream from both rivers' junction (Figure 1.1 on page 4).

In summary, the PFOA concentrations of all water supply stations along the river Ruhr are well observed, beside water network data from many water supply areas depending on them. For other regions, there are fewer measurements per station, usually below the limit of quantification, or none.

2.1.2 Temporal data patterns for individual water supply stations

Water supply stations along the river Ruhr (Figure 1.1 on page 4) are focussed on, with numerous measurements covering a longer period of time. For data availability and patterns, cf., Table 2.2. According to this, there are two main types of temporal data patterns, a rather diffuse one (Figure 2.2) and a striking decrease (Figure 2.3).

The first typical temporal structure is found at stations which are 1) less affected by the contamination

Table 2.3: Number of water supply areas by contamination risk and data availability per station: Considering the water supply stations with a relevant supply for a particular area (estimated share of more than 5%, Section 2.5), it is distinguished whether all or some of these stations are potentially affected by the Möhne-Ruhr contamination incident, and whether any data are available from all or some of these stations (right the respective numbers of areas among them with only one supplying station).

Ruhr contamination – Data availability	Number of areas	Single-station areas among them
all stations affected	23	14
– all with data	23	14
– some with data	0	—
– none with data	0	0
some stations affected	9	—
– all with data	8	
– some with data	1	
– none with data	0	
no station affected	423	271
– all with data	181	110
– some with data	27	—
– none with data	215	161

incident, being upstream beyond the Möhne mouth (like Mengesohl, Figure 2.2 left), 2) partly supplied by groundwater (like Volmarstein, centre) or 3) far downstream at the river Ruhr (like Styrum West, right). Measurements observed here feature no or hardly detectable trends and comparatively small variability. A certain decline in background pollution may be reflected, too.

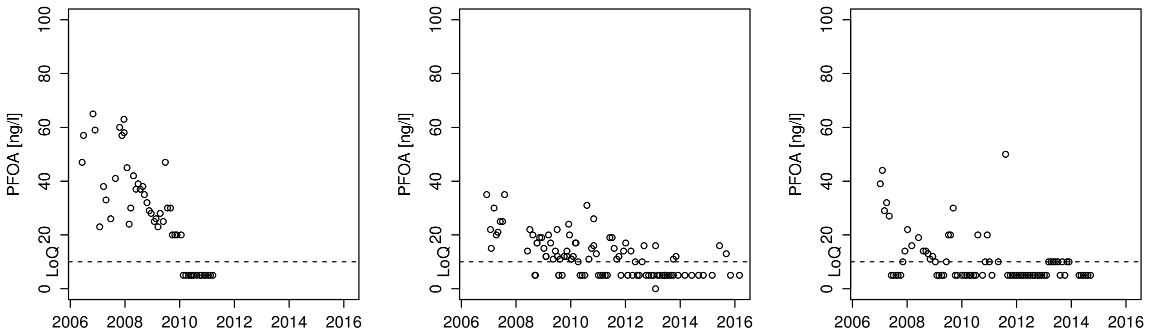

Figure 2.2: Measurement series of PFOA from exemplary water supply stations with comparatively low PFOA values and no or a weak trend. Left: Mengesohl; centre: Volmarstein; right: Styrum West.

A second type is a more or less smooth decline over the years, though some short-term trends may be distinguished. The decrease is considerably stronger for the first couple of years (e.g., Essen-Überruhr, Figure 2.3).

This decrease at the beginning is the strongest at the outstanding station of Möhnebogen, where some water filtering measures have taken place after discovery of the contamination. Filters eliminate all PFASs after installation, become gradually less efficient and have to be reactivated after a while (Figure 2.4). Thus, some outliers, change points, and even seasonality may occur for this station (and

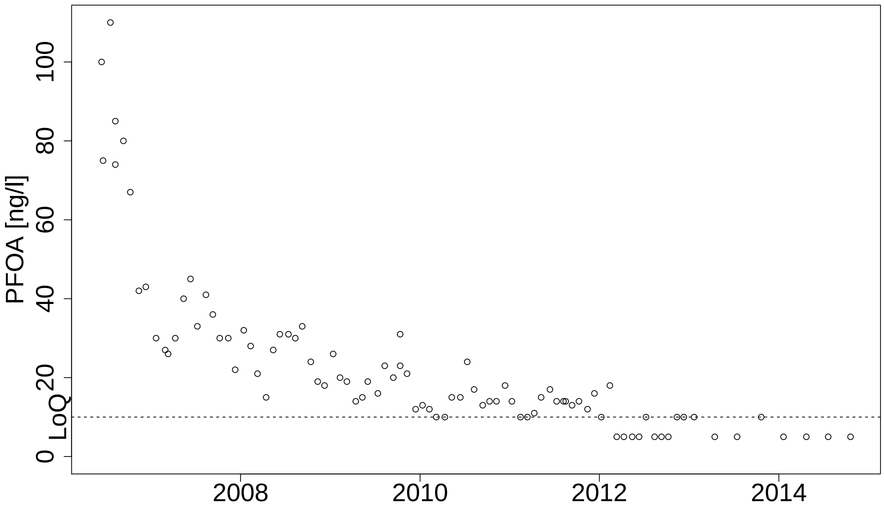

Figure 2.3: Measurement series of PFOA from Essen-Überruhr station as an example for a smooth decrease.

also for some others downstream, which have installed filters too, if filtering data were available).

Apart from these differing observed types of temporal structures, there are periods of missing data in some locations: The density of measurements in the course of time and space varies widely. Thus, extrapolation should be performed with caution, especially when using rather simple regression or interpolation methods.

In summary, there are two typical trends in the PFOA values of water supply stations at the river Ruhr – beside the special case of Möhnebogen station. In particular, a characteristic form of 'decreasing decrease' can be distinguished. Apart from the Ruhr-depending stations, there are but a few with sufficient data to identify a trend (one example are the stationary measurements around the limit of quantification in Haltern, see Figure 2.6 on page 19).

2.2 Methodological background

This empirical data analysis has been started rather open-minded with regard to model classes. Ultimately, a model in continuous time and discrete space is preferred due to irregular measurement series and the predominant role of water supply stations along a river.

However, an autoregressive model for temporal smoothing of irregularly observed time series based on wavelets has been developed by Salcedo *et al.* (2012), which also allows for non-stationarity. A Bayesian model for predicting values at regular points in time by Nieto-Barajas and Sinha (2015) and another for periodic signals by Beelaerts *et al.* (2012) each assume a stationary process. The trend is taken into account by Quick *et al.* (2015) through a hierarchical Bayesian model for a spatio-temporal process, where gradients are estimated to detect sudden changes. Further proposals taking into account censored values are given in Section 2.3.1.

As for approaches in spatial scenarios similar to that of the 'PerSpat' project, and for profound model developments, see, e.g., the smoothing models by White *et al.* (2017) for areas being but partially observed: Data are primarily given on area level with a Markov property assumed, whereas the focus

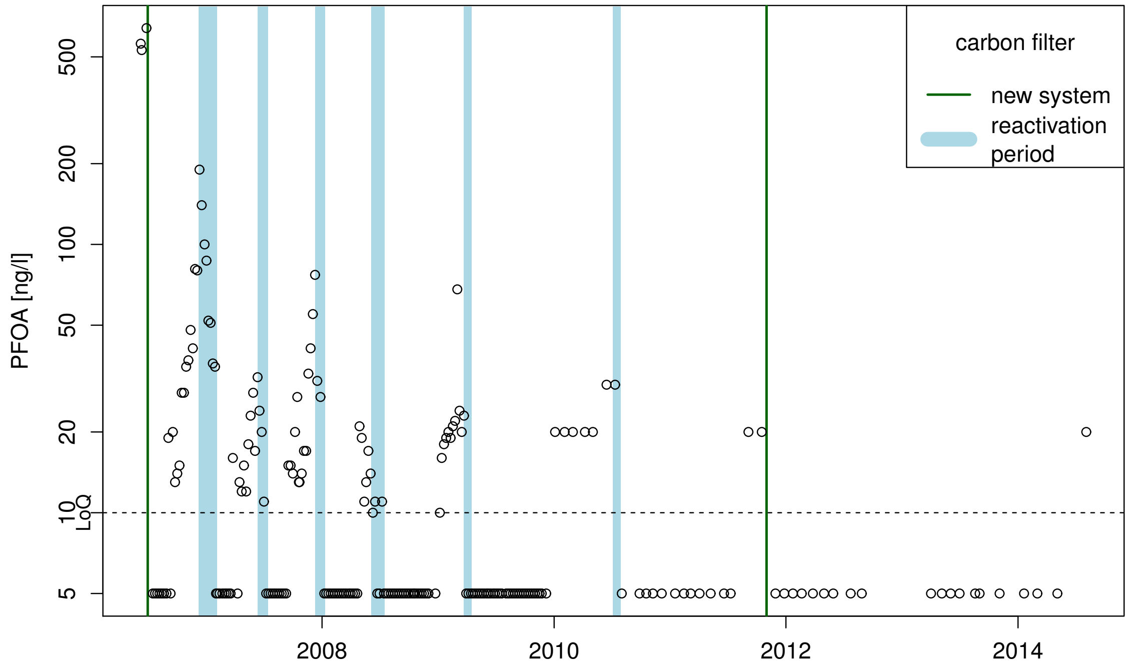

Figure 2.4: Measurement series of PFOA from Möhnebogen station including times of carbon filtering interventions.

of the ‘PerSpat’ project is on the temporal modelling first, and also on encountering complicated spatial relationships. Bayesian hierarchical discrete-space approaches, along with choice considerations for the intrinsic CAR prior, can be found in Keefe *et al.* (2019); again, the neighbourhood structure is more specific than found in the given situation. Grollemund *et al.* (2019) present a regression model using functional predictors as representations of the temporal structures in spatio-temporal analysis; however, after exploring the given data patterns, the functional predictors are in some way replaced by a simpler generalised linear model, whose link function may be regarded as an expression of a simple functional relationship.

As a frequentistic alternative, Tang *et al.* (2019) develop a semi-parametric copula model to cope with the spatio-temporal dependence and to predict values for new points of time and sites. This is rather restrictive, being Markovian in time, whereas continuous, irregularly measured temporal data are modelled within ‘PerSpat’. Wang and Sun (2019) propose a spatial regression model with local polynomials modelling the coefficients as a function of space, but without an explicit temporal component. An autoregressive conditional heteroscedasticity model with respect to spatial neighbourhood is developed by Otto *et al.* (2018) and augmented to the spatio-temporal case, but where data are observed at or arranged to discrete points in time.

As water supply stations are partly supplied from rivers, a specific structure can be assumed for the spatial dependence, determined by distances along the river, directions, junctions and water amounts. An overview of this field can be found in Cressie *et al.* (2006), who develop a covariance model for Kriging. O’Donnell *et al.* (2014) propose a flexible regression model depending on a spatially smoothed expected value function. Here, temporal, especially seasonal influences can also be integrated. Model comparisons for hydrological variables on large river networks by Kriging and regression can be found in Laaha *et al.* (2014). All such models are typically designed for far more complex river networks than considered in the 'PerSpat' project.

The set-up of consistent spatial correlation matrices along more complex river networks is described as a particular challenge (cf., e.g., Peterson *et al.*, 2007). With more data from groundwater-stations not directly adjacent to a river than are available in the 'PerSpat' project, the river could be regarded as an inhomogeneous line source of contamination (cf., e.g., Ayub *et al.*, 2019). Another option would be the incorporation of more detailed river network features, such as water amounts, stream velocity and smaller tributaries (cf., Ver Hoef *et al.*, 2006), if corresponding data were available.

Further methods and application scenarios for risk assessment, when measurement data are available at more steps along water supply systems than in the 'PerSpat' project, can be found in Roozbahani *et al.* (2013). Causal remote effect estimation, when both influence and outcome quantities are measured along a river and over time, unlike in 'PerSpat', is performed by Saul *et al.* (2019) via the parametric g-formula.

After all, the unique, unusual and complex spatial and temporal structure of the given drinking water data requires a specifically designed modelling that deviates from the methods mentioned in the literature. Shin *et al.* (2011a) give another such practical example of estimating PFOA contamination in the region around a chemical plant, respecting many local environmental circumstances, which determine the transport of pollutants in air, soil, ground-, surface and drinking water.

2.3 Preparation and preliminary approaches

2.3.1 Measurements below the limit of quantification

Many data of PFOA concentrations in drinking water are reported as below the limit of quantification (LoQ). In these cases, the measurement technology is not accurate enough to quantify a concentration as being above zero, and the measurement error is too large compared to the magnitude measured. Such data should not be set to zero (Helsel, 2006), whether or not a statistical model allows this.

A simplifying approach is to group drinking water pollution into categories (cf., Nolan *et al.*, 2009 and Maisonet *et al.*, 2012). Here, data below LoQ can be interpreted as the smallest value of an ordinal variable. Lee and Helsel (2005) use a regression of the values above LoQ on their quantiles. For temporal interpolation, spatial regression and prediction may be more targeted, replacing the spatial dimension by time. Here, Rathbun (2006) proposes an imputation of values below LoQ using the covariance structure to improve Kriging estimation. This is compared with simpler substitutions by Schelin and Sjöstedt-de Luna (2014) and turns out to be particularly suitable in the case of a normal distribution assumption. A Bayesian alternative assuming a Gaussian random field is found in de Oliveira (2005), where non-stationarity of the process is allowed by introducing space- or time-dependent regressors. Kriging methods for spatial interpolation with values below LoQ are further presented and compared by Saito and Goovaerts (2000) and Knotters *et al.* (1995). If these measurements are used as regressor, Nie *et al.* (2010) propose approaches to include the below-LoQ-statement in the regression model.

Within the 'PerSpat' project, Becker (2016) has worked on this issue, referring to definitions and approaches from Currie (2004), Armbruster and Pry (2008) and Helsel (2006), among others:

Randomisation is often recommended, be it from a uniform distribution on $[0, \mathrm{LoQ}]$ or from a Gaussian distribution around $\frac{\mathrm{LoQ}}{2}$. However, the variability of randomised values can be higher as for actual measurements, and the resulting ‘artificial’ data can feature an ‘accuracy’ which is not found in real data: In all relevant cases, the LoQ is reported as 10 ng/l; but many water supply stations report data which are also accurate to 10 ng/l (i.e.: below LoQ, 10, 20, 30, ...). Therefore, it has not seemed appropriate to introduce values with a semblance of higher accuracy. Furthermore, models for stations with rather low measurements have turned out to be quite sensitive to this randomisation. (On a side note, measurement accuracy is not necessarily the same as reporting accuracy. And higher values tend to have higher absolute inaccuracy due to serial dilution in chemical analysis.)

So, a simpler recommended approach is applied for dealing with data below LoQ, namely to fix them to a certain value, most commonly $\frac{\text{LoQ}}{2}$, which in this case is 5 ng/l.

2.3.2 Temporal modelling

Besides the LoQ issue and several data cleansing steps, Becker (2016) applies preliminary approaches to model the temporal structure of data from water supply stations within the ‘PerSpat’ project. Constant, linear or spline interpolation tend to overfit the given data and does not always follow the natural decreasing long-term trend, where longer periods without data occur. Smoothing, by running-mean or Gaussian kernel, suffers from similar problems, extrapolation is often not reliable. Regression on time with a quadratic term has a sufficient prediction performance in many cases. The optimal model is different from station to station. A universal approach, like regression, that is less sensitive to single data points, should be preferred. Where change points (installation and replacement of carbon filters) are known, the benefit of piecewise modelling becomes apparent.

As a simple but rather straightforward Bayesian smoothing approach, a conjugate Gamma-Gamma-model is also applied. Let a random variable X_t represent the concentration in drinking water for an arbitrary point of time t and λ_t be a rate parameter at t such that

$$X_t|\lambda_t \sim \Gamma(\zeta, \lambda_t).$$

The time-independence of the shape parameter ζ is motivated by an assumed constant relative measurement error in terms of the coefficient of variation:

$$c_v := \frac{\sqrt{\text{Var}(X_t)}}{\mathbb{E}(X_t)} = \frac{\sqrt{\zeta/\lambda_t^2}}{\zeta/\lambda_t} = \frac{1}{\sqrt{\zeta}}.$$

Measurement errors are an important source of variability besides natural fluctuations. The absolute error depends on the measured magnitude due to serial dilution in chemical analysis. According to LANUV, 2011, $c_v \approx 0.2$ is assumed in case of the NRW PFOA measurements, resulting in $\zeta \approx 25$. Furthermore, the Gamma distribution is a natural approach for positive data, without need of transformation (as with the log-normal distribution), which would be very sensitive with small values (cf., the LoQ issue in Section 2.3.1). When $\zeta = 25$ is fixed as a first approach, a conjugate prior Gamma distribution

$$\lambda_t \sim \Gamma(\xi_t, \psi_t)$$

emerges. To incorporate smoothing in the calculation of the posterior distribution of any λ_{t_0}, weighted data are used, so that the parameters are updated by

$$\begin{aligned} \xi_t &\dashrightarrow \xi_t + \zeta \cdot K \cdot \sum_{k=1}^{K} z_k, \\ \psi_t &\dashrightarrow \psi_t + K \cdot \sum_{k=1}^{K} x_t^{(k)} z_k. \end{aligned}$$

The data $x_t^{(k)}$ of observations $k = 1, \ldots, K$, are weighted according to their temporal distances

$$d_{\text{temp}}(t, t_0) := |t - t_0|$$

to the respective t_0. Weights are then obtained from, e.g., a Epanechnikov kernel:

$$z_k := z(t(k), t_0) := \max\left\{0, \quad \frac{3}{4} \cdot \left(1 - \frac{d^2_{\text{temp}}(t, t_0)}{s^2_{\text{temp}}}\right)\right\}$$

with smoothing parameter s_{temp} controlling the kernel's bandwidth. Various empirical, vague and informative priors for the λ_t's are considered as well as adjustments of d_{temp} and s_{temp} and also modelling ζ. However, depending on the availability of data in the considered period of time, results seem either overfitted or the posterior depends too much on the prior or its variance is too large.

2.3.3 Data homogeneity of water supply areas

In many cases, PFOA values for water supply areas have to be deduced from the relevant water supply stations' data in order to assess the residents' exposure. It is considered whether the involved stations' temporal trends are similar enough ('homogeneous') for joint modelling and whether these data and the respective network data, if any, are sufficiently homogeneous in the course of time. Underlying data analyses within the 'PerSpat' project are reported by Becker (2017).

Homogeneity or inhomogeneity are presumed on the basis of exploratory results, particularly on the two distinguishable main types of temporal data patterns at the stations (Section 2.1.2), the stations' locations with respect to the affected river Ruhr (cf., Section 2.1.1) and the structures found in predictions for Ruhr-dependent stations (Section 2.4.2).

As illustrated in Table 2.2 on page 11, the occurrence of the typical data patterns is strongly correlated with the stations' dependence from the river Ruhr, where high values with a steep decrease in the course of time can be found. Moreover, these patterns are most similar for stations within the same segment of the river, regarding both data and predictions, where prediction intervals often overlap or are near to each other, along the course of time, as shown in Figures 2.8 on page 26 and A.3 on page 82. In other locations, there are typically low values and few measurements.

Therefore, an area is presumed as homogeneous, if it is essentially supplied by stations, which are in the same situation with regard to their location and their temporal data pattern. Examples for the homogeneity of data, especially of such with a decreasing pattern, are shown in Figure 2.5.

On the other hand, there are a few areas with some supplying stations depending on the contaminated river Ruhr and some being non-affected (cf., the middle part of Table 2.3 on page 12). An example of an inhomogeneously supplied area is shown in Figure 2.6: Three stations supply the COE_3 area, two of them are dependent on the contaminated river Ruhr, the third (with a water share of about 40%) is

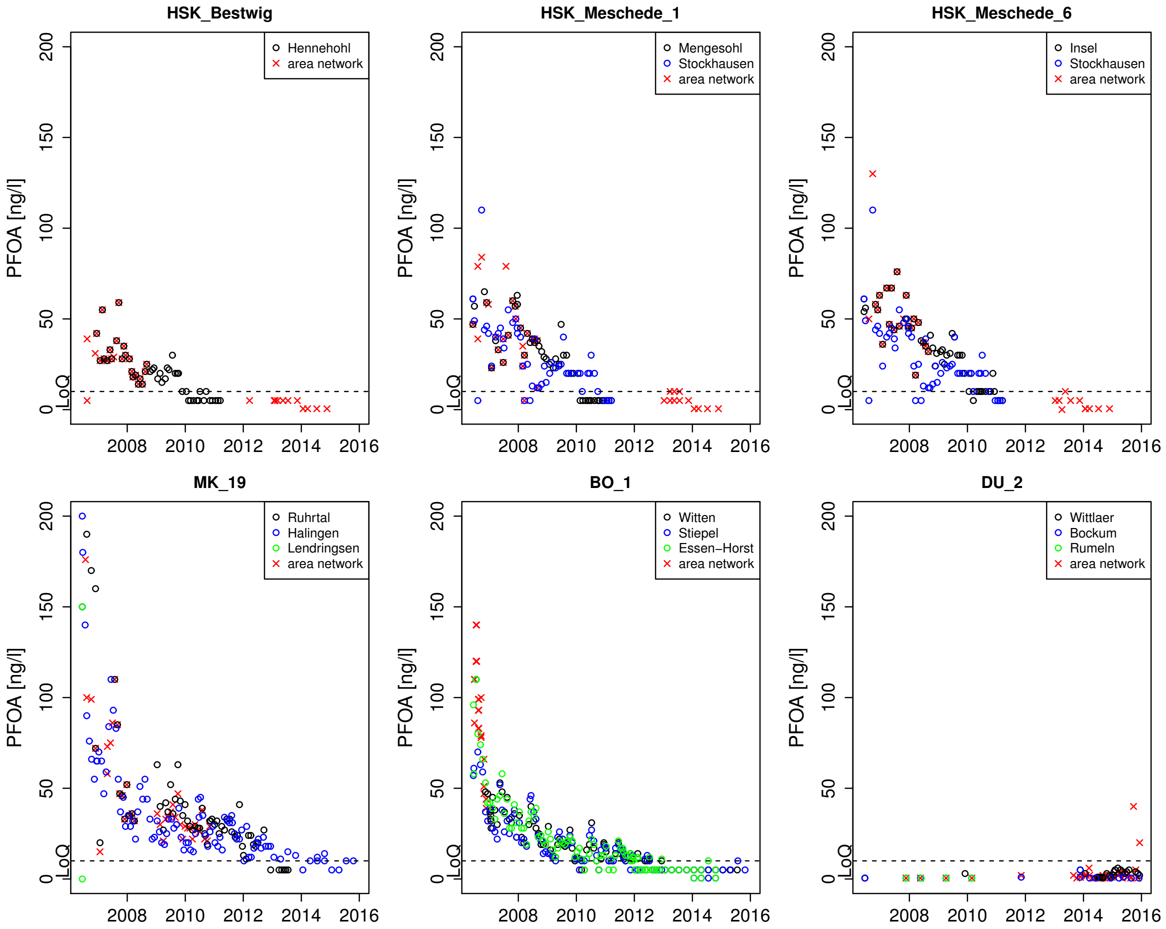

Figure 2.5: Comparison of PFOA measurements from the network of some water supply areas (red crosses) and from their relevant supplying stations.

non-affected.

A reliable double-checking of the homogeneity is limited to cases with sufficient amounts of data from the respective stations or from the area's water network, representing the same periods of time. Larger amounts of station data are mostly given along the river Ruhr (cf., Tables 2.2 and 2.3 from page 11). Network data are rare, especially for the most interesting period of time. (E.g., there are none for the interesting case of the inhomogeneous COE_3 area.) Where available, the range of the network data seems not to substantially differ from the respective concurrent station samples (cf., Figure 2.5).

For Ruhr-dependent areas, data from the respective stations are combined, to search for clusters in the space of time and values (using the R package EMCluster, Chen and Maitra, 2015). The same is done with network data, where available. It is not possible to distinguish the data and (re-)discover the stations in this way.

Another approach to check whether concurrent station data are distinguishable is a regression model similar to those in Section 2.4 for the combined measurement data of an area's supplying stations, with additional dummy variables representing the stations. For areas completely dependent on (a segment of) the river Ruhr, the resulting station effects are not significant or at most very weak. For inhomogeneous areas, such effects are found.

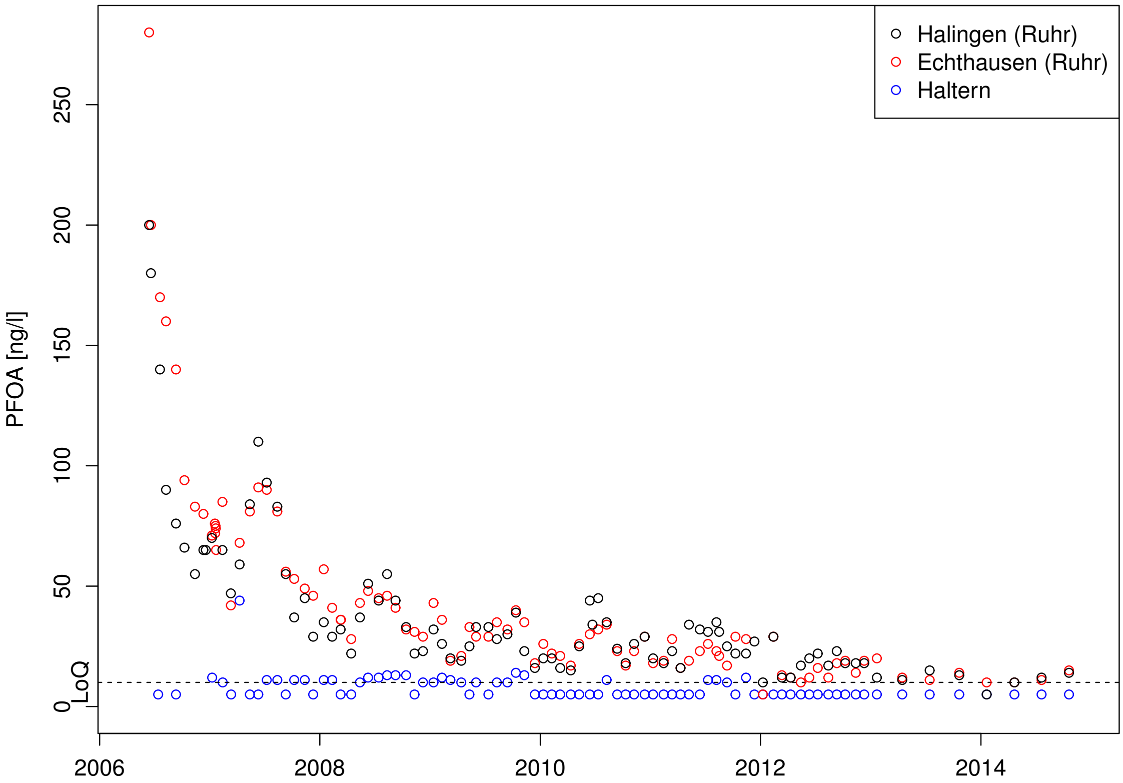

Figure 2.6: Measurement series of PFOA from the three water supply stations supplying the COE_3 water supply area.

Both approaches allow the conclusion for the vast majority of water supply areas (cf., the upper and lower part of Table 2.3 on page 12) that there is no evidence against a hypothesis of homogeneity, where adequate data are available (i.e., some of the about 30 stations with many data according to Table 2.2 on page 11 are involved). This avoids uncertainty by averaging PFOA concentrations. Perspectively, joint models can be fitted to the respective water supply stations, to obtain estimates for a whole area's exposure. Where waters from the river Ruhr and other sources are involved (about 9 areas according to Table 2.3), this situation is more difficult.

2.3.4 Kriging along rivers

Within the 'PerSpat' project, Becker (2017) has worked on Kriging approaches in hydrology, referring to Cressie *et al.* (2006), and Skøien *et al.* (2003, 2006), among others:

Kriging approaches along the river turn out to be inappropriate, since they do not respect the river segments, which dominate the spatial data structure (Section 2.4). Furthermore, even for the well observed river Ruhr, the data are too sparse to analyse proper variograms, and sometimes the assumption of isotropy seems violated. Ordinary and universal Kriging lead to sufficient local estimates when cross-validated along the river Ruhr, but perform worse near the Möhnebogen water supply station. It is questionable whether the continuous Kriging solution is an efficient way to merely estimate concentrations at discrete, definite stations, as in the given situation.

Topological (Top-) Kriging (Skøien *et al.*, 2006, using the R package `rtop`, Skøien *et al.*, 2014; cf., e.g., Laaha *et al.*, 2014) is designed for prediction on the level of the rivers' catchment areas. This would be useful to model measurements along the river network in NRW, thereby summarising several stations each. Nonetheless, with the stations' values connected to water supply areal units, it does not seem productive to introduce a third spatial level. Furthermore, no geographical data on the actual catchment areas are available.

In general and state-wide, the problems of a varying data density, up to entirely non-observed regions, and of a regionally different strength of dependence remain hindering for all Kriging approaches. For the river Ruhr, the rather complicated models, especially Top-Kriging, seem exaggerated for an 'linear' spatial river structure. Therefore, these approaches are dropped because of the non-suitable spatial data structure. For problems of the performance of spatial methods, especially (Top-)Kriging, compared to regression approaches for river-related catchment area data, see, e.g., Brunner *et al.* (2018).

2.3.5 Further spatial and spatio-temporal modelling along rivers

A lack of data in vast regions also affects the applicability of comprehensive state-wide smoothing models. Moreover, water supply areas which are isolated within the supply relationship structure and additional spatial dependencies would have to be respected.

As a preliminary spatio-temporal approach, a Bayesian Gamma-Gamma model (a generalisation of that from Section 2.3.2) has been applied. It focusses on the mean surface to predict PFOA values for any water supply station, water supply area and point of time. It suffers from the same limitations as the temporal-only model, such that the reliability and uncertainty in prediction vary widely according to the presence of data.

Within the 'PerSpat' project, Becker (2017) has worked on the application of Gaussian processes to model the PFOA concentration, mainly along the river Ruhr, referring to Rigby and Stasinopoulos (2005, 2014), among others, using the `R` packages `CompRandFld` (Padoan and Bevilacqua, 2015) and `geoR` (Ribeiro Jr *et al.*, 2020). The form and degree of spatial dependence has been explored via the empirical semivariogram, but it turned out to be non-monotonic. So, it has been difficult to identify a clear spatial structure, partly due to the small number of points (i.e., stations) in space. It has been found useful to exclude the stations upstream of the Möhne mouth before modelling. Instead of geographical coordinates, kilometres along the river have been used for spatial information, so space being actually one-dimensional. As a preliminary spatio-temporal approach, the time has been included as a second dimension, but without useful semivariogram results. In conclusion, the applied methods seem to be restricted to actual two-dimensional geographical coordinates.

In this context, Becker (2017) has also considered Gaussian Markov random fields to model PFOA concentrations at stations along the river Ruhr, conditional on the neighbouring stations. Again, the clarity of the spatial structure, and thus the applicability of the model with a reasonable smoothing parameter, is affected by sudden changes (especially river junctions) and the possibility of prediction proves to be different depending on the location. Models have been fitted separately for several points of time. Based on this, the time has been experimentally included by establishing a spatio-temporal neighbourhood structure. However, the MCMC simulation has turned out to be extremely difficult due to the large number of surface parameters, or, reducing this number, too less points in time could be considered.

Preliminary spatial and spatio-temporal approaches (like smoothing, Markov random fields) mostly pay special attention to the modelling on two spatial levels: stations and water supply areas. They partly use a proportion- or weight-matrix as developed in Section 2.5. Other terms are included to the area-level equations to allow for additional uncertainty there. However, as data on this level (from the network) are not numerous and meaningful enough, more complicated approaches are refrained from and the focus is set on modelling at station level, where data are extensive. Conclusions from stations to areas regarding PFOA contamination by means of the weight-matrix remain, therefore, deterministic, but without further speculative model assumptions and excessive computation.

2.4 Temporal GLMs for individual water supply stations

It is investigated, which classes are appropriate to model the dynamics (temporal progression) of PFOA concentration data at a water supply station. Models are fitted individually to a station's data, but it is of main interest to obtain a consistent approach along the river Ruhr.

Let X_{it} denote the PFOA concentration for time t, now distinguished by water supply station i. Reasonable units of time are months or days.

2.4.1 Model choice

The data are continuous with respect to time with an irregular measurement process: There are periods of time with a high 'data density' (up to two measurements within a few days) and others without any data for several month. Therefore, regression models are applied with the time as a covariate instead of autoregressive models.

An appropriate model has to cope with the varying shapes of decrease (cf., Figures 2.2 and 2.3 from page 12), which may also imply some heteroscedasticity. Several generalised linear models (GLMs, Table 2.4) with various distribution families for X_{it} and link functions

$$g\left(\mathbb{E}X_{it}\right) = g(\mu_i(t)) = \eta_i(t) = \alpha_i + \beta_i \cdot t$$

as well as data transformations are considered.

For the latter (see Item 1 in Appendix B.2.1 for documentation), a pre-transformation like $\ln(X_{it})$, $\sqrt{X_{it}}$ or $\frac{1}{\sqrt{X_{it}}}$ and usage of a simple linear regression leads to very large residuals for periods of high measurements. For the pre-transformation $\frac{1}{X_{it}}$, the instability of the estimation is even worse. Additionally, results are difficult to interpret, when re-transformed after the model fit.

Instead of that, while sticking to the simple linear regression, the time covariate is experimentally transformed (see Item 1 in Appendix B.2.1 for documentation): Among the usual fractional polynomials, the model

$$\mathbb{E}X_{it} = \alpha_i + \beta_i \frac{1}{\sqrt{t}} \tag{2.1}$$

results in a particularly good fit. Results are included to Table 2.4. However, it is not translation invariant, and a starting point t_0 has to be chosen, preferably close to the beginning of the measurements (in this situation the beginning of the year 2006) and consistently for all stations. From an environmental scientific point of view, the results are difficult to interpret in terms of the change of concentration during a given span of time.

Another preliminarily considered approach is a Poisson regression, where data are discretisised motivated by their frequently given accuracy of 10 ng/l (i.e., 0 for values below the limit of quantification, $1 \mathrel{\hat{=}} 10$ ng/l, $2 \mathrel{\hat{=}} 20$ ng/l, …). However, as there are many locations and periods of time, where all measurements are far from zero, this model seems not appropriate, and its fit has not always converged. For the same reason, models with zero-inflation are not applied.

For the remaining GLMs without pre-transformation of data (see Item 2 in Appendix B.2.1 for documentation), at first all combinations of continuous distribution families and link functions, such that the model fit converges for all stations, are identified (shown in Table 2.4). Among these, a Gaussian distribution with an inverse link

$$(\mathbb{E}X_{it})^{-1} = \alpha_i + \beta_i \cdot t$$

turns out to yield the best fit in terms of the residual sum of squares (RSS) for most of the stations, especially those with the typical ‘decreasing decrease’. But also the Gamma distribution with its natural inverse link is often close to it in that sense. Table 2.4 shows the goodness of fit of the various models; the RSS is considered, as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) are not directly comparable between GLM families. Figure 2.7 illustrates the suitability of the models with regard to the estimated mean function as well as their capability to reflect the data’s variance.

While the variance $\mathrm{Var}(\hat{\mu}_i)$ of the mean response estimator is always a function of time and is further affected in case of non-identity link functions, a proper choice of the data variance function turns out to be more relevant in terms of magnitude: $\mathrm{Var}(X_{it})$ is constant for all μ_i in case of the Gaussian family, leading to a certain probability of negative predictions, but also to a closer prediction interval in the period of high values. For the Gamma family, it holds

$$\mathrm{Var}(X_{it}) = \phi_i^2 \mu_i^2(t)$$

with a dispersion factor ϕ_i, so it is possible to interpret and estimate the measurements’ standard error as a constant multiple of the mean. This is in accordance with some assumptions on measurement technique and natural fluctuation (LANUV, 2011). On the other hand, the prediction interval is very large in the period of high values (perhaps corresponding with an inferior measurement technique in the early period) and rather small later on (notwithstanding that an increased uncertainty is assumed for measurements close to the limit of quantification). However, all models seem to sufficiently cover the data variance.

The inverse link is very suitable for modelling the ‘decreasing decrease’ type of temporal data patterns; for less definite trends, especially along the upper Ruhr, others would do (cf. Table 2.4). In some cases, a slight heteroscedasticity remains. Outlying data at the measurements’ beginning have a strong leverage that may lead to a certain instability, especially when ‘predicting the past’. With the inverse link being superior compared to the log link, the decrease turns out to be steeper than exponential for highly contaminated places.

Any such models are applicable for situations without change points or outliers as described in Section 2.1.2. Where carbon filtering interventions are known, a piece-wise regression is applied.

With regard to easy interpretability of the parameters, the goodness of the predictions, and the perspective model extension (Section 2.6), the Gaussian GLM with an inverse link is focussed on.

A preliminary polynomial regression approach (Section 2.3.2) has also been dropped in favour of

Table 2.4: Goodness of fit of various GLMs (distribution families with link functions) in terms of the residual sum of squares by water supply station along the river Ruhr (best fitting model per station in bold). The last column represents a linear model (Gaussian family and identity link), where the time transform from Equation (2.1) is applied.

			Gaussian		Gamma	(time
	Station	identity	log	inverse	inverse	transform)
Upper Ruhr	Hennenohl	**2509**	2594	2969	3492	2797
	Insel	**3960**	5205	6991	10436	8761
	Mengesohl	**4822**	6403	8012	12069	9204
	Stockhausen	9309	**8349**	8448	8655	8913
	Langel	**14313**	15219	18680	26437	20501
Middle Ruhr	Echthausen	88504	45406	**14310**	21009	27617
	Warmen	113172	35174	**27297**	92980	33468
	Fröndenberg	20550	**14131**	16808	23934	14807
	Ruhrtal	42814	30823	**23334**	23354	23793
	Halingen	52522	35329	**21342**	24756	22985
	Hengsen	45081	26048	**20154**	20515	20240
	Villigst	117266	54163	**32504**	33629	39224
	Ergste	52451	34579	**22754**	25543	25000
	Westhofen 1	83985	39238	**21610**	22606	29586
	Westhofen 2	52549	30258	**21761**	21861	22315
Lower Ruhr	Hengstey	131908	**79197**	84974	130458	81677
	Volmarstein	3501	3207	**3110**	**3110**	3122
	Ruhrstrasse	6356	5454	**5061**	5132	5096
	Witten	3926	**3007**	3342	3740	3336
	Stiepel	12004	8328	7003	7146	**6886**
	Essen-Horst	17017	7938	5645	6071	**5572**
	Essen-Überruhr	16607	8839	**4616**	4710	5053
	Essen-Kettwig	11580	6569	**4859**	5016	5358
	Dohne	13437	11544	**10594**	10679	10912
	Styrum Ost	**7256**	7386	7499	7682	7823
	Styrum West	5738	5499	5103	5272	**5087**

this simpler, more interpretable model with few parameters. Table 2.5 evaluates polynomial (GLM) regressions on time with a linear, an additional quadratic, and an additional cubic term; for the simple linear model, even a higher degree is not of much use; for a Gaussian distribution with inverse link, there are but minor differences between the degrees.

2.4.2 Results

Figure 2.8 shows a cross-section $\hat{\mu}_i(t^*)$ of the mean response estimation results along the river Ruhr. Further points of time can be found in Figure A.3 on page 82, a longitudinal overview in Figure A.4 on page 85. See Item 2 in Appendix B.2.1 for documentation.

The results show a clear distinction between the river's segments: The abrupt increase of PFOA concentrations below the junction of Ruhr and Möhne is reflected as well as a decrease below the mouth of the Ruhr's tributary Lenne. The estimations for most of the stations along the lower Ruhr tend to feature smaller uncertainty intervals.

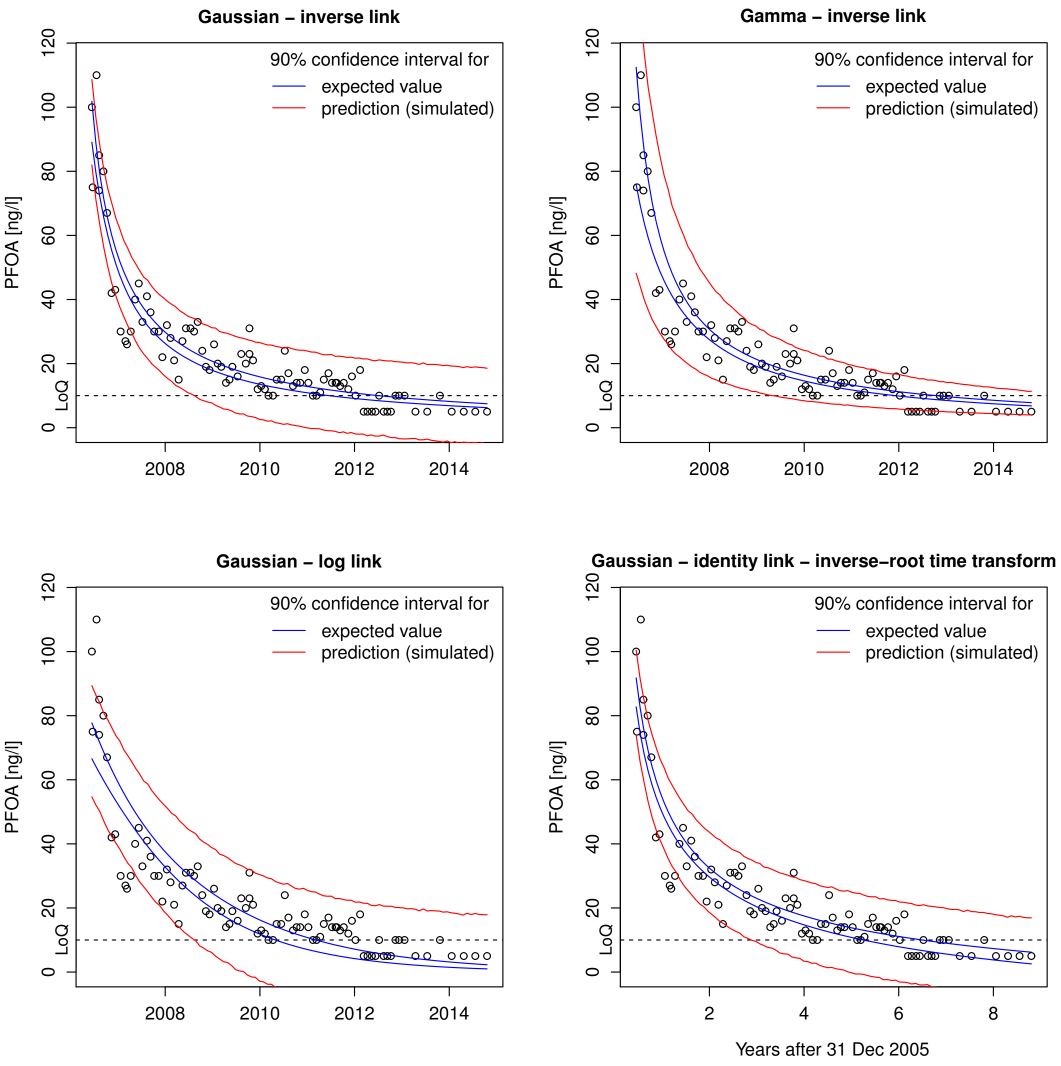

Figure 2.7: (GLM) Regression for PFOA measurements depending on time: point-wise 90% confidence intervals for the expected value (blue) and simulated 90% prediction intervals (red) for various models exemplarily fitted to the data of Essen-Überruhr.

Within the segments, no definite trend is recognizable; local effects of the individual stations seem to overlay a trend along the river.

To conclude, the GLM with Gaussian distribution and an inverse link function is a convenient way of functional representation of temporal trends in PFOA data of the water supply stations along the river Ruhr. This consistent approach opens the possibility to jointly model all data from this river, and so to distinguish station-related random effects from the general temporal progression. Such a model should respect the river segments and is developed in Section 2.6.

2.5 Station-to-area supply proportions

With data (and also possible models) on two spatial levels: water supply stations and water supply areas, the proportions are of strong interest, how much of an area's water stems from the respective stations, in

Table 2.5: Exemplary GLM polynomial fitting results for Essen-Überruhr: Regression on time with a linear, an additional quadratic, and an additional cubic term, respectively. Especially for the best fitting inverse link, there is not much gain from a higher degree.

Link	Degree	RSS	AIC	BIC
identity	1	16607	738	746
identity	2	10710	700	710
identity	3	7608	671	684
inverse	1	4616	622	629
inverse	2	4491	621	631
inverse	3	4108	615	627

order to mediate between the two levels. However, there is no direct information on this question (apart from rough estimations by a few local drinking water suppliers).

Instead of this, these supply proportions are estimated using data on the stations' water supply amounts and the areas' water demands together with the qualitative information which stations supply which areas. See Item 3 in Appendix B.2.1 for documentation.

Let $\tilde{\mathbf{a}} = (\tilde{a}_1, \ldots, \tilde{a}_m)^T$ be the supply of stations $i = 1, \ldots, m$ and $\tilde{\mathbf{b}} = \left(\tilde{b}_1, \ldots, \tilde{b}_n\right)^T$ the demands of areas $j = 1, \ldots, n$. The data (say, cubic metres per year) are somewhat vague and, therefore, standardised to $a_i := \frac{\tilde{a}_i}{\tilde{\mathbf{a}}^T \mathbf{1}_m}$ and $b_j := \frac{\tilde{b}_j}{\tilde{\mathbf{b}}^T \mathbf{1}_n}$, respectively, in order to obtain matching totals.

Formally, a weight matrix

$$\mathbf{W} = \begin{pmatrix} w_{11} & \ldots & w_{1m} \\ \vdots & & \vdots \\ w_{n1} & \ldots & w_{nm} \end{pmatrix} \in [0,1]^{n \times m}$$

is searched for, such that area j gets the w_{ji}'s part of its water from station i, which means that the water amounts $w_{ji} b_j$ of the areas sum up to the stations total supply:

$$a_i = \sum_{j=1}^{n} w_{ji} b_j, \quad i = 1, \ldots, m, \tag{2.2}$$

with area-wise restrictions

$$\sum_{i=1}^{m} w_{ji} = 1, \quad j = 1, \ldots, n. \tag{2.3}$$

Such a deconvolution problem (cf., e.g., Cutler, 1978) is generally not uniquely solvable.

In the given situation, the problem is simplified by several facts:

1. Only a small number of stations and areas are connected at all. And these relations are known at least qualitatively, albeit not in numbers. Therefore, it is known for far most of the cases that a station i^* does not supply an area j^* and so there are many further restrictions $w_{j^*i^*} = 0$ when calculating $\mathbf{W}$.

2. If an area is supplied by a single station, the respective weight $w_{j^*i^*} = 1$ is known prior to solving the system of Equations (2.2) and (2.3), even if this relation is not bijective.

3. Due to the many zeros in $\mathbf{W}$, the equation system of No. (2.2) and (2.3) is actually split in, often

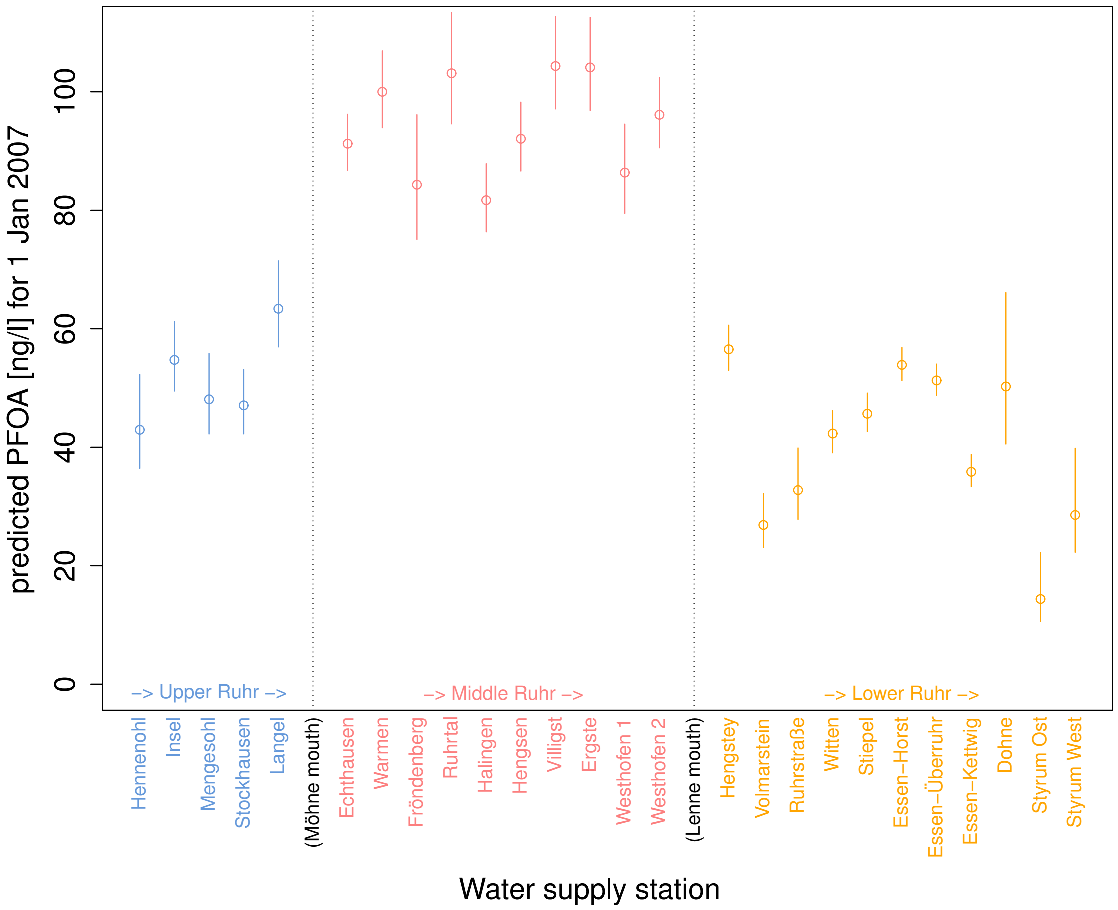

Figure 2.8: Predictions (estimates and approximate 95% confidence intervals of the expected value) of the PFOA values on 1 January 2007, from GLMs with Gaussian distribution and inverse link, individually for each water supply station along the river Ruhr.

very small, partial problems, now in matrix notation:

$$\mathbf{a}_{(h)} = \mathbf{W}_{(h)}^T \mathbf{b}_{(h)}, \quad \mathbf{W}_{(h)} \mathbf{1}_{m_h} = \mathbf{1}_{n_h}, \quad h = 1, \ldots, H. \tag{2.4}$$

Each partial problem is to solve a non-zero block $\mathbf{W}_{(h)} \in [0,1]^{n_h \times m_h}$ of $\mathbf{W}$, concerning m_h stations and n_h areas. Moreover, a large number of supply relations are unique, i.e., $m_h = n_h = 1$ (cf., Table 2.6).

Using only the qualitative information from Item 1, i.e., whether it holds $w_{ji} = 0$ or $w_{ji} > 0$ (for all areas j and all stations i), an ‘adjacency matrix’ $\mathbf{G} \in \{0,1\}^{n \times n}$ of the areas is set up, indicating whether two areas are (partly) supplied by the same station (or several of them). $\mathbf{G}$ is transformed to a diagonal structure of non-zero blocks with the help of the `R` packages `igraph` (Csardi and Nepusz, 2006) and `Matrix` (Bates and Mächler, 2019). These blocks of areas from $\mathbf{G}$ with their respective affiliated stations are, therefore, separated from each other in terms of the supply structure. (A later, nuanced version of this is shown in Figure 2.9.) Thus, the partial problems referred to in Item 3, regarding the non-zero blocks of $\mathbf{W}$, are set up from these blocks of $\mathbf{G}$.

With that, $H = 322$ linear equation systems like No. (2.4) are obtained. Of these, 27 are non-trivial, i.e., $m_h, n_h \geq 2$. Each comprises of $m_h + n_h$ independent equations like No. (2.2) and (2.3) to solve a certain number of unknown quantities $w_{ji} \in (0,1)$.

In 6 of these 27 cases, the system is determined. In 20 systems, there are more equations than unknowns and a numerical solution for the least squares problem

$$\left\| \mathbf{a}_{(h)} - \mathbf{W}_{(h)}^T \mathbf{b}_{(h)} \right\|_2^2 \longrightarrow \min_{\mathbf{W}_{(h)}}!, \quad \mathbf{W}_{(h)} \mathbf{1}_{m_h} = \mathbf{1}_{n_h}$$

is found with the help of the R package `NlcOptim` (Chen and Yin, 2019) performing optimisation with constraints; the obtained estimated proportions are exact enough given the vague data; in particular, the by far largest system ($m_h = 72$, $n_h = 82$), containing many of the Ruhr dependent stations, is solvable. Only one system is under-determined: it concerns a small region ($m_h = 3$, $n_h = 6$) not related to the Ruhr and is not considered further.

Table 2.6: Overview on water supply relations: number of connections between stations and areas with relevant shares of water ($> 5\%$ of an area's demand).

Number of areas	**285**	81	46	19	8	16
Supplying stations per area	1	2	3	4	5	≥ 6
Number of stations	22	**578**	73	16	5	8
Supplied areas per station	none	1	2	3	4	≥ 5

The obtained proportions of water supply from stations to areas are important to estimate the stations' relevance for the NRW water supply in general. They are also used to assess the areas' PFOA burden and thereby the residents' risk by weighted averaging. This is particularly relevant if the PFOA concentration is inhomogeneous within an area, especially one supplied from both the Ruhr and other rivers or groundwater (Section 2.3.3).

The supply matrix **W** also illustrates to what extent the spatial units (water supply areas) are contiguous in terms of shared water: Blocks of 'adjacent' areas are determined as above; the resulting Figure 2.9 shows one large region of such potentially correlated areas (with the river Ruhr in their south) and several smaller clusters. On the other hand, there is a large number of single isolated areas.

Therefore, the obtained supply matrix is a key to understand the unusual discrete spatial structure of NRWs drinking water supply. It is the basis for models on supply area level, including analyses of network data.

2.6 Joint modelling along the river Ruhr

As pointed out in Section 2.1, the availability of PFOA contamination data strongly varies within the state of NRW. The by far largest part of data is from the water supply stations along the river Ruhr. These measurements have also been taken more regularly in the course of the years. Therefore, it is obvious to further focus on this river, as also done before, especially in Section 2.4.

The tributary Möhne is a special case when considering spatial relationships along the river Ruhr. It is the principal source of contamination, what has become evident in Section 2.4. But there is only one relevant water supply station directly depending on it: Möhnebogen with its outstandingly high

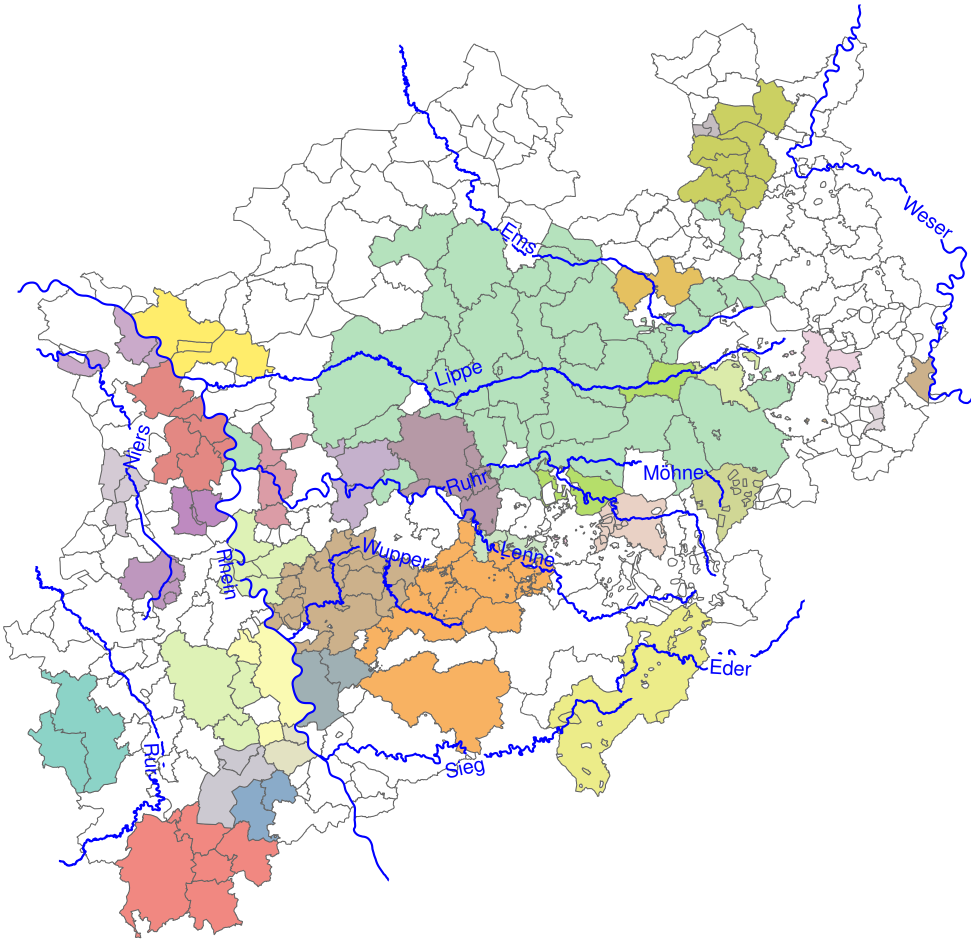

Figure 2.9: NRW water supply areas combined according to their potential correlation (i.e., two areas have relevant ($> 5\%$) shares of water from the same station): A coloured area is correlated with at least one other area of the same colour and not correlated with any area of other colours. White coloured areas are not correlated with any other area.

measurement values and clearly visible fluctuations caused by filtering measures (Section 2.1.2 with Figure 2.4). As these points of intervention are known, a separate modelling treatment is possible and indicated, realised by piece-wise regression. When modelling the Ruhr stations jointly, it is therefore reasonable to exclude the Möhne. In this way, the space becomes actually one-dimensional.

2.6.1 Initial situation

Section 2.4 concludes that regression on time is a reasonable approach to model measurement series at individual water supply stations, as a functional representation of a temporal trend. It is also concluded that certain generalised linear models (GLMs) are consistently applicable to all stations; in particular, a Gaussian distribution with an inverse link is considered; other families and links can be justified, too. To introduce this notation again, let a random variable X_{it} represent the PFOA concentration in drinking water at a station i and time t. GLMs

$$\begin{aligned} X_{it} &\sim N\left(\frac{1}{\eta_i(t)}, \sigma_e^2\right) \\ \eta_i(t) &= \alpha_i + \beta_i \cdot t. \end{aligned} \tag{2.5}$$

with fixed intercepts α_i and slopes β_i are individually fitted for all water supply stations $i = 1, \ldots, m$ along the river Ruhr.

When applying spatio-temporal statistics to all the Ruhr data, a model that is continuous in time (irregular measurement series) and discrete in space (stations) is naturally of interest. Considering contamination as continuously distributed along the river is also reasonable, and various approaches have been applied (Section 2.3), though without striking success. Furthermore, measurements are not directly from the river, but from within the stations after some stages of treatment. For these and other reasons, individual effects of the stations are likely and they should be prominently respected in joint modelling.

Another conclusion from Section 2.4 is the importance of the river segments, distinguished as: upstream beyond the Möhne mouth (here called upper Ruhr), between the Möhne and the largest tributary Lenne (middle Ruhr), and below the Lenne mouth (lower Ruhr); see Figure 1.1 on page 4. The predictive means in individual station models and their uncertainties can be clearly distinguished between these segments (Figure 2.8 on page 26). Within the segments, however, a spatial structure is not apparent and local effects may prevail (Section 2.4.2).

Therefore, some regression models based on those from Section 2.4 are developed and applied jointly to all the Ruhr data, to explore and distinguish effects of stations and river segments.

2.6.2 River-wide mixed GLM

The model in Equation system (2.5) is augmented by fixed effects for the river segments $s = 1, \ldots, S$, besides the overall intercept and slope, to identify significant differences in water contamination caused by the rivers' junctions (see Item 4 in Appendix B.2.1 for documentation). To avoid over-parametrisation, the upper Ruhr segment is used as baseline, resulting in just $s = 2, 3$. Modelling the segments by random effects does not seem reasonable, as the contamination process of the river as a whole is rather deterministic, with local and spontaneous random fluctuations found at the levels of stations and measurements; furthermore, the number of segments is small and such effects could hardly be presumed to have a common variance, considering data and predictions from previous sections.

Additionally, random effects p_i, q_i of the stations are introduced, again with respect to both intercept and slope. Here, the usage of fixed effects would mean an overemphasis of local, individual phenomena, leading to a kind of overfitting, besides a cumbersome number of model parameters. Instead, it is intended to capture possible random fluctuations on station level and to explore, whether there is a spatial structure in the resulting effects.

With these specifications and unlike in (2.5), the river-wide model

$$\begin{array}{rcl} X_{it} & \sim & N\left(\frac{1}{\eta_i(t)}, \sigma_e^2\right) \\ \eta_i(t) & = & \alpha + \beta \cdot t + \alpha_{(s)} + \beta_{(s)} \cdot t + p_i + q_i \cdot t \\ p_i & \sim & N(0, \sigma_p^2) \\ q_i & \sim & N(0, \sigma_q^2) \\ (i & = & 1, \ldots, m) \\ (s & = & 2, \ldots, S) \\ & & (i \text{ within } s) \end{array} \tag{2.6}$$

is fitted to all joined data of the Ruhr stations.

Several variations (inclusion and exclusion of terms, subgroups of data, etc.) of this model are tested and compared by the BIC, using the `R` package `lme4` (Bates *et al.*, 2015):

The introduction of station random effects p_i, q_i, with overall variances σ_p^2, σ_q^2 (independent of the segments) increases the goodness of fit only slightly or not at all. This result is even sensitive to randomisation of values below the limit of quantification (Section 2.3.1). The residuals (with variance σ_e^2) are predominant over these effects.

On the other hand, when fitting individual models to the segments, i.e.,

$$\begin{array}{rcl} X_{it} & \sim & N\left(\frac{1}{\eta_i(t)}, \sigma_e^2\right) \\ \eta_i(t) & = & \alpha + \beta \cdot t + p_i + q_i \cdot t \\ p_i & \sim & N(0, \sigma_p^2) \\ q_i & \sim & N(0, \sigma_q^2) \\ (i & = & 1, \ldots, m_s) \end{array}$$

for $s = 1, 2, 3$, the station random effects are in some cases stronger than before. This raises the possibility that the random effect variance differs between segments, and motivates a further model improvement (Section 2.6.3).

Experimentally, a quantitative spatial distance, in terms of kilometres along the river, is added to the linear predictor $\eta_i(t)$, but turns out to be irrelevant in terms of changing the BIC, if river segments are present in the model. This is also another indication to stay with a discrete spatial structure.

In contrast to the approach in Equation system (2.6), a hierarchical random effect modelling of stations within segments does not prove useful, as the goodness of fit is weaker than in other models and no significant overall intercept and slope emerge.

2.6.3 Enhanced Bayesian mixed GLM

As the variability is likely to differ between the river segments, the model in Equation system (2.6) is further augmented by distinguishing the station random effect population per segment:

$$
\begin{array}{rcl}
X_{it} & \sim & N\left(\dfrac{1}{\eta_i(t)}, \sigma_e^2\right) \\
\eta_i(t) & = & \alpha + \beta \cdot t + \alpha_{(s)} + \beta_{(s)} \cdot t + p_{i(s)} + q_{i(s)} \cdot t \\
p_{i(s)} & \sim & N(0, \sigma_{p,s}^2) \\
q_{i(s)} & \sim & N(0, \sigma_{q,s}^2) \\
(i & = & 1, \ldots, m_s) \\
(s & = & (1,)2, \ldots, S).
\end{array}
\qquad (2.7)
$$

As this goes beyond the usual estimation routines, and to have a closer look on the random effects' behaviour, a Bayesian model is implemented in the `Stan` software (Stan Development Team, 2020b), using the `R` interface package `rstan` (Stan Development Team, 2020a). The default non-informative priors are employed for $\alpha, \beta, \alpha_{(s)}, \beta_{(s)}, \sigma_e^2, \sigma_{p,s}^2$ and $\sigma_{q,s}^2$, $s = (1,)2, \ldots, S$. See Item 2 in Appendix B.2.1 for documentation. Table 2.7 shows the posteriors of variance and fixed effect parameters.

Table 2.7: Quantiles of the sampled posterior distributions for the standard deviations (non-squared) of residuals and random effects as well as for the fixed effects from the model in Equation system (2.7). The residual variance σ_e^2 is not directly comparable to the random effect variances due to the non-linear model specification. A significance mark $*$ indicates that the 95% credible interval does not include zero. (The specific numbers are not directly informative regarding substantive interpretation in this complicated model with a non-linear link function – apart from σ_e.)

	Quantiles					
	2.5%	10%	50%	90%	97.5%	
σ_e	13.2157	13.3665	13.6280	13.8989	14.0478	$*$
$\sigma_{p,1}$	0.0003	0.0007	0.0036	0.0111	0.0193	$*$
$\sigma_{q,1}$	0.0002	0.0004	0.0009	0.0022	0.0042	$*$
$\sigma_{p,2}$	0.0051	0.0065	0.0098	0.0153	0.0206	$*$
$\sigma_{q,2}$	0.0007	0.0010	0.0015	0.0022	0.0028	$*$
$\sigma_{p,3}$	0.0002	0.0005	0.0050	0.0122	0.0190	$*$
$\sigma_{q,3}$	0.0002	0.0006	0.0012	0.0021	0.0027	$*$
α	-0.0503	-0.0477	-0.0412	-0.0352	-0.0320	$*$
β	0.0073	0.0079	0.0089	0.0099	0.0109	$*$
$\alpha_{(2)}$	-0.0296	-0.0251	-0.0177	-0.0095	-0.0060	$*$
$\beta_{(2)}$	-0.0011	-0.0002	0.0010	0.0022	0.0029	
$\alpha_{(3)}$	-0.0606	-0.0559	-0.0467	-0.0369	-0.0305	$*$
$\beta_{(3)}$	0.0050	0.0060	0.0075	0.0089	0.0097	$*$

Figure 2.10 represents the posteriors of the stations' random intercepts $p_{i(s)}$ and confirms that it is reasonable to model them with segment-wise distributions. While there are hardly no random effects of stations along the upper Ruhr, their variance $\sigma_{p,s}$ is larger for the middle Ruhr (Table 2.7). However, even here, many random effects are still not significant. A certain similarity to the station-individual predictions in Figure 2.8 on page 26 (apart from the sign) holds especially for the lower Ruhr. A spatial structure in terms of a smooth decrease is indicated, albeit very weak.

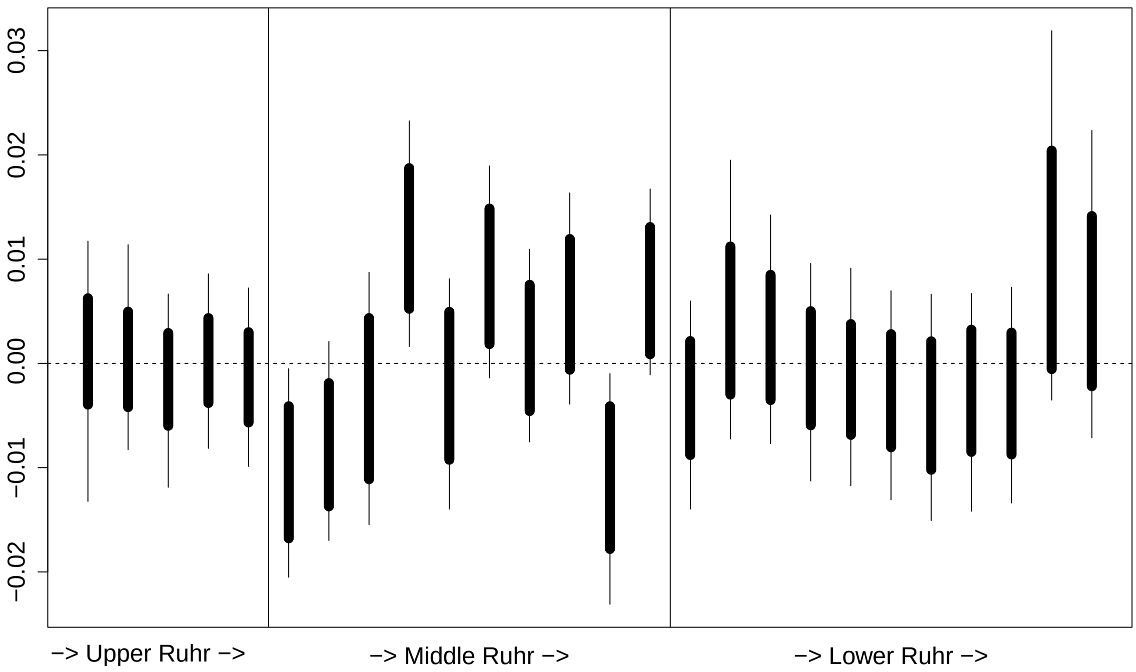

Figure 2.10: Credible intervals (80% and 95%) for the random intercepts $p_{i(s)}$ $(i = 1,\dots,m_s; s = 1,2,3)$ by water supply station along the river Ruhr, obtained from the sampled posterior distributions from the model in Equation system (2.7). (The specific numbers on the axis are not very informative regarding substantive interpretation in this complicated model with a non-linear link function.)

Results for the random slopes (time dependence) are similar. On a side note, the stations' intercepts and slopes are generally somewhat correlated in all models, not only the joint models. Nonetheless, it is canonical to always include the intercept where a non-linear link function is present. Moreover, it is not clear from the outset whether individual effects can be found rather regarding the influence of time (random slope) or absolutely (random intercept).

After all, some of the stations along the middle Ruhr feature random effects in the model from Equation system (2.7), which significantly differ from zero. But as before, the residuals (with variance σ_e^2) are still predominant over them.

For the fixed effects with regard to both intercept and slope (time dependence), the distinctiveness of all river segments is mostly significant.

2.7 Conclusions and discussion

Given the potential health risks of PFASs, especially PFOA, and the contamination incident affecting the rivers Möhne and Ruhr, measurement data from water supply stations and within the drinking water network are explored.

The data comprise of quite complete measurement series along the river Ruhr and only few measurements each at other water supply stations. Along the river, a steeply decreasing temporal data pattern is observed at mainly affected stations, as opposed to a more diffuse structure upstream; in later years after the contamination, the decrease is weaker. In other regions, no PFOA is usually detected.

A consistent temporal modelling approach is of interest for comparison of results along the river

Ruhr and for building joint models. A Gaussian generalised linear model with an inverse link function turns out to have the best fit to the data among regression approaches, hinting to a more than exponential contamination decrease after the incident. If there are crucial events such as change points or extreme outliers, a piece-wise regression is applied.

When exploring the spatial structure of the predictions resulting from this model, the important role of the river segments (between the main junctions) is found: The PFOA concentration at stations considerably varies between the Ruhr segments, but less and without clear trends within. Joint modelling of all data from Ruhr-dependent stations confirms the significant effects of the segments. Random effects of stations occur, especially along the middle course of the river Ruhr below the Möhne mouth, but are comparatively weak. Such a model is smoother and generalises better than the station-wise approach; it is also applicable to predict PFOA values with a more regional perspective, especially in water supply areas completely dependent on a single Ruhr segment.

The importance of random effects and the absence of a clear spatial trend of declining concentrations downstream hint at possible station-related effects. The treatment of the water on its way from the river to a station's outlet varies, particularly with regard to filtration, both technically and naturally through the river bank. Some stations operate with a certain share of non-contaminated groundwater, but these proportions are not constant and no reliable data are available. Another aspect are additional minor PFOA discharges, especially from sewage plants affecting the river Ruhr (cf. LANUV, 2011). In fact, there seems to be a slightly increasing trend along parts of the lower Ruhr.

The spatial structure of the water supply consists of two levels, stations and water supply areas, and assessment of the residents' PFOA risk by area is ultimately of interest. Therefore, inferences from the station-level data are necessary for prediction at the area level. An approximate supply proportion matrix is obtained from rough information on the water amounts. Furthermore, from the exploratory results, homogeneity of data from stations along the river Ruhr, which supply the same areas, can be presumed. Therefore, the uncertainty when aggregating data in such areas is low. For those areas supplied from both the river Ruhr and unaffected stations, the uncertainty is higher, but aggregation is possible using weights from the proportion matrix. In summary, reasonable conclusions from stations to areas are generally possible.

From the perspective of environmental medicine, the steep decrease of PFOA concentrations in water, after the contamination is discovered and handled, is relevant for the assessment of the internal exposure. Studies on the Arnsberg cohort within the 'PerSpat' project (cf., Section 4.2.4), including pharmacokinetic modelling, emphasise the predominant importance of time periods with very high burden; later minor PFOA intakes are then less relevant for the temporal progression of the internal exposure (see also Bartell *et al.*, 2010; Seals *et al.*, 2011; Russell *et al.*, 2015; Li *et al.*, 2018). Modelling this early stage with adequate precision is therefore important. For later years, the decrease in water is weaker and a certain level of detectable contamination remains over a long time; this is also relevant, as even a low level PFOA burden is suspected of adverse health effects.

Data and predictions not only show a severe contamination of the river Ruhr downstream from the Möhne mouth, but also further upstream from this alleged principal PFOA source, albeit significantly weaker. The widespread distribution of this substance should always be kept in mind.

Predicted PFOA concentrations in drinking water are used to estimate concentrations in blood serum of NRW residents, depending on time and place of residence, and also used in the analysis of health-related data from the state-wide birth registry (Chapter 4). The developed modelling approaches are transferable to other or more complex river systems and to other substances solute in water or to similar scenarios.

2.8 Perspectives

As the environmental epidemiological interest lies primarily in the prediction of PFOA concentrations per water supply area, it is straightforward to change the perspective in modelling accordingly. E.g., in general spatio-temporal notation with discrete points of time, let

$$\nu := \left(\nu^{(1)}, \dots, \nu^{(T)}\right) := \begin{pmatrix} \nu_{11} & \dots & \nu_{1T} \\ \vdots & & \vdots \\ \nu_{n1} & \dots & \nu_{nT} \end{pmatrix}$$

represent the expected values of PFOA concentration at time $t = 1, \dots, T$ and area $j = 1, \dots, n$. To respect both spatial levels of data, the same is denoted for the water supply stations $i = 1, \dots, m$ by

$$\mu := \left(\mu^{(1)}, \dots, \mu^{(T)}\right) := \begin{pmatrix} \mu_{11} & \dots & \mu_{1T} \\ \vdots & & \vdots \\ \mu_{m1} & \dots & \mu_{mT} \end{pmatrix}.$$

As before, the spatial levels are connected by a weight matrix $\mathbf{W}$ as developed in Section 2.5, in the simplest case as $\nu = \mathbf{W}\mu$. More general, an areal random effect like $\omega \sim \mathrm{N}\left(\mathbf{0}_n, \tau^2 \mathbf{I}_n\right)$ may be introduced to avoid a deterministic conclusion from stations to their supplied areas, when further uncertainty is likely:

$$\nu = \left(\mathbf{W}\mu^{(1)} + \omega, \dots, \mathbf{W}\mu^{(T)} + \omega\right) = \mathbf{W}\mu + \mathbf{1}_T' \otimes \omega.$$

This is contrary to the previous goal of a model that is continuous in time. Furthermore, sufficient data would be required on area level, i.e., from the network.

With the given amount of data, all models are reliable only for the river Ruhr, but generally applicable to all main rivers of the state. Useful predictions are not possible for largely unobserved regions or periods of time.

Extrapolations are challenging with all methods, especially for the early period prior to the first measurements in 2006. Yet, the duration of the Brilon contamination has been estimated as two to four years from the fertilizer's quantity, the drainage, and the residents' internal exposure (Hölzer *et al.*, 2009, Skutlarek *et al.*, 2006). This information should be respected when 'predicting the past' using the above models.

Comprehensive spatial models should respect the complex potential state-wide spatial correlation structure. It is explored using the supply proportions. There is one large region of potentially correlated areas near the river Ruhr, but also many other supply areas which are isolated or grouped to small clusters. This information is important to set up a covariance matrix for usage in spatial models.

Given the results from the Ruhr data, it also seems reasonable to focus on the river segments when building up a neighbourhood structure based on water supply stations and their locations along the river network in NRW. In the future, river-induced correlations may be estimated from the Ruhr data. Spatial correlations caused by groundwater may be relevant, too, and, therefore, are possible extensions of the neighbourhood structure. The existence of several contamination sources should be taken into account, too.

Chapter 3

Perinatal data analyses

This chapter is concerned with regression analyses of the perinatal registry data from North Rhine-Westphalia (NRW), Germany. The main interest is on the estimation of influences on the birth weight (BW). There are reasons for bivariate modelling of BW jointly with the gestational age (GA, the duration of pregnancy). This is taken up by a copula model within a framework of Bayesian distributional regression, as the first application of this innovative method to birth data with two continuous response variables.

The local PFOA concentration in drinking water is later added as another environmental explanatory variable (Chapter 4).

The region along the upper course of the river Ruhr in NRW, precisely the town of Arnsberg, is focussed on: It is of particular interest due to variable and in parts very high PFOA concentrations in drinking water during several years. Furthermore, the external exposure to PFOA is well assessed and confirmed by human biomonitoring data on the internal exposure of a cohort of residents (Hölzer *et al.*, 2008). A constrained data analysis also eases computability in Bayesian distributional regression.

It is investigated, which family of one-parametric copulas, which families of marginal distributions and which linear predictors are most suitable for the given perinatal registry data. Using the identified copula model, the results are compared to a standard univariate regression model for BW and the effects of biometric, perinatal, environmental and socio-economic covariates on BW, GA and their dependence are estimated.

The chapter opens with a presentation of the data in more detail (Section 3.1). This is followed by a short literature review on bivariate modelling in biometrics, especially for birth data respecting GA, on copula applications, especially in regression, and on options to model several parameters conditional on covariates (Section 3.2).

The analyses begin with the development of the standard univariate model (Section 3.3), where GA is included in functional form as a polynomial. Explanation and application of the Bayesian distributional copula model follow in Section 3.4, beginning with a brief theoretical introduction and the existent implementation (Section 3.4.1). After some preliminary considerations and pre-processing steps (Section 3.4.2), the optimal marginal families for BW and GA are selected in Section 3.4.3, the copula in Section 3.4.4, all including variable selection. Substantive results are given in Section 3.4.5. The performances and results of both the polynomial and the copula model are compared in Section 3.5.

Various aspects of data modelling are discussed in Section 3.6, including the usage of secondary data and their quality (Section 3.6.1), the dependence structure of BW and GA, which may be more difficult to deal with after all (Section 3.6.2), and some medical interpretations and the benefit of bivariate modelling thereby (Section 3.6.3). Conclusions of this chapter follow in Section 3.7, some perspective aspects in Section 3.8.

Besides the application of a developer version of the `BayesX` software (Belitz *et al.*, 2020) for the distributional regression models, routine calculations have been performed using the `R` environment (R Core Team, 2020), see Appendix B.2.2 for documentation.

3.1 Perinatal registry data from NRW

The perinatal registry data (overview in Table 1.2 on page 5) are collected by all hospitals and are combined and processed by the quality assurance office located at the state medical association, for the purpose of quality assurance in obstetrical health care. An overview of collection and administration of the German perinatal registry data can be found in Reeske *et al.* (2011), another use of them as secondary data in a scientific study. Within the 'PerSpat' project, these secondary data collected from 2003 until 2014 are used. They comprise of about 1.7 million records and more than 200 biometric, medical and social variables on mother and child, pregnancy, birth and treatment.

They are anonymised by removing all personal information apart from the postal code of the mother's residence. Further data cleansing steps are performed, in particular regarding the plausibility of GA. Analyses are restricted to singleton births, to children born alive without malformations and to postal codes within NRW.

When restricted to the town of Arnsberg, 6442 birth cases are observed within the period from 2003 until 2014. Those where values of relevant variables are not given are removed, leaving a total of 4451 observations.

The response variables are BW (measured in g with varying accuracy, mean: 3390, standard deviation: 517) and GA (clinically estimated, in days, mean: 277, standard deviation: 12), the former being of primary interest (Figure 3.1).

Individual relevant explanatory variables are pre-selected from the perinatal registry data. This is done in accordance with the literature (e.g., Gardosi *et al.*, 1995; Schwartz *et al.*, 2010; Frederick *et al.*, 2008; Thompson *et al.*, 2001) and with previous findings within the 'PerSpat' project (Kolbe *et al.*, 2016). The specific variables are: the child's sex, the number of previous pregnancies of the mother, whether the child has been delivered by Caesarean section (sectio), whether the birth has been induced, the mother's age, the mother's height, the mother's body mass index (BMI) at the beginning of pregnancy, the gain of weight of the mother during pregnancy, the number of cigarettes the mother reports to smoke per day, whether the mother is single and whether the mother is employed. Some descriptive characteristics can be found in Table 3.1.

Temporal and spatial information is given by the date of birth and the postal code of the mother's residence. These are later used to include the average PFOA concentration in drinking water for a given place and time as a spatio-temporally assigned covariate (see Section 4.2).

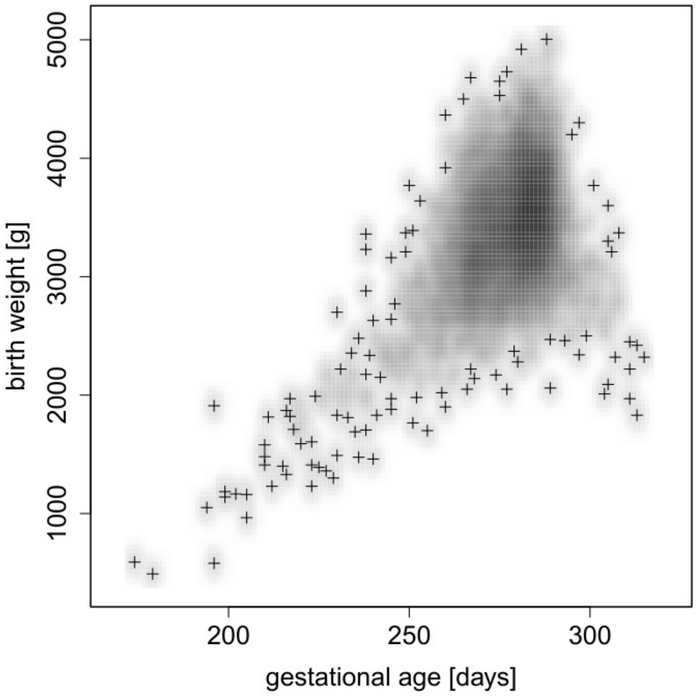

Figure 3.1: Observations of birth weight and gestational age (density of the point cloud represented by shading: darker shade is for more points, + stands for a single isolated point).

Table 3.1: Descriptive characteristics for covariates from the perinatal data.

l	Covariate	Unit	Description
1	sex		female: 47%
2	previous pregnancies		0: 38%, 1: 32%, 2: 16%
3	sectio		24%
4	induction		26%
5	maternal age	years	mean: 29.4, s.d.: 5.5
6	maternal height	cm	mean: 167.0, s.d.: 6.7
7	maternal BMI	$\mathrm{kg\,m^{-2}}$	mean: 25.2, s.d.: 5.3
8	maternal gain of weight	kg	mean: 10.4, s.d.: 5.7
9	maternal smoking	cigs./day	no: 87%, $\leq$10 cigs.: 8%
10	mother is single		7%
$L = 11$	mother is employed		42%

3.2 Methodological background

When analysing perinatal (newborn infants') data with BW as the response variable of primary interest, it is essential to adjust for GA, which is often reported as the quantitatively most important covariate (e.g., Skjærven *et al.*, 2000; Fang *et al.*, 2007; Weiss *et al.*, 2014; Frederick *et al.*, 2008). Augmenting linear models, it may be included as a polynomial or in other parametric functional forms (e.g., Gardosi *et al.*, 1995; Salomon *et al.*, 2007). Another widespread alternative is a binary response with a class such as 'small for GA' (e.g., Gage, 2003; Polakowski *et al.*, 2009; Thompson *et al.*, 2001).

In contrast to univariate approaches like these, consideration of bivariate (or multivariate) outcomes is frequent in biometric research, such as meta-analysis (Berkey *et al.*, 1998), clinical trials (Braun, 2002), dose-response-modelling in developmental toxicology (Regan and Catalano, 1999), measuring heavy metal concentrations in the environment (Pozza *et al.*, 2019), or simultaneous estimation of human exposure to jointly occurring chemicals (Sy *et al.*, 2020). In spatial statistics, a Bayesian hierarchical model for multiple responses can be found in Gelfand and Vounatsou (2003), who generalise the conditional autoregressive (CAR) model accordingly with proper priors; the application to children's biometric data is also proposed there.

In gynaecological and obstetric research, modelling of a bivariate response comprising of both BW and GA is recommended (e.g., Schwartz *et al.*, 2010; Gage, 2003; Fang *et al.*, 2007), but not common. Schwartz *et al.* (2010) use a finite mixture model in a Bayesian analysis of the discretisised two-dimensional response, because of suspected subgroups in the population and to increase flexibility. GA, which is often only known in full weeks, is taken into account as a censored variable. Neelon *et al.* (2014) generalise this approach from categorical to normally distributed covariates; in addition, spatial effects and area regressors are included. A two-dimensional mixture model of normal distributions for BW and GA is also used by Gage (2003) and estimated by maximum likelihood estimation (MLE) or EM methodology. For binary response variables representing BW by GA in relation with mortality, Gage *et al.* (2008) propose a mixture model of logistic regressions that depends on the density of covariates, to account for subgroups and unobserved heterogeneity. A mixture model for several response variables is also used by Zhu *et al.* (2012), to determine the membership to subgroups by logistic regression.

A flexible way in following the recommendation of bivariate analysis is to use copulas (Nelsen, 2006) to simultaneously and independently estimate the univariate marginal distributions and the dependence structure via the copula. More flexible non-Gaussian copula families assume less symmetries for the data; instead, upper or lower tail dependence can be modelled. There is a vast literature on copula modelling, also in the regression context. Effects of covariates on the parameter of various copula families are modelled by Acar *et al.* (2013). Krämer *et al.* (2013) work on MLE in regression with several outcomes joint by copulas. For Bayesian copula application in human biology, see, e.g., Dalla Valle *et al.* (2018). Another approach for regression problems is to represent the multivariate density by a (D-)vine copula (Kraus and Czado, 2017; Cooke *et al.*, 2020).

An important issue in regression analysis of data with non-Gaussian shapes is to model several parameters of a (possibly multivariate) distribution, and especially other parameters than the location, conditional on covariates. Vector generalised linear (VGLMs) and additive models (VGAMs), first proposed by Yee and Wild (1996), further developed and comprehensively presented by Yee (2015), enable this for a wide range of distribution families, also beyond the exponential family class. For many univariate cases, GAMs for location, scale and shape parameters are already introduced by Rigby and Stasinopoulos (2005) and much work has been done on them since, see in particular Wood (2017) and Rigby *et al.* (2019). While all this does not include copulas, Song *et al.* (2009), on the other hand, develop simultaneous MLE for VGLMs using Gaussian copulas, but the estimation of correlation matrices does not depend on covariates. Finally, Vatter and Chavez-Demoulin (2015) develop a GAM, where a conditional equivalent to Sklar's theorem (Sklar, 1959) is considered and the dependence structure of two response variables is represented by a copula, and a single dependence parameter is a function of covariates; under certain regularity conditions and limited to continuous responses with the whole real axis as support, they present penalised MLE.

Within the 'PerSpat' project, the VGLM approach of Yee (2015) has had an essential part in preliminary considerations for modelling the perinatal response variables, in particular using a Gamma distribution for GA.

However, as the Bayesian interpretation of parameters, data and their relationships fits well to the goal of modelling arbitrary parameters conditional on covariates, the Bayesian distributional copula regression models developed by Klein and Kneib (2016) are applied: Besides modelling the two response variables depending on several covariates, these models comprise of another predictor to independently model the parameter of one-parametric copulas between the two of them in a flexible manner. The procedure is outlined in Figure 3.2 on page 45 and generally applicable to data from various areas of applications, including the biometric studies named above. It is a natural approach to analyse a bivariate response with an asymmetric joint distribution and skewed marginals as found in the data (Figure 3.1 on page 38). Klein *et al.* (2019) also consider bivariate copula regression with GA and low BW measured as a continuous and a binary variable, respectively, conditional on various biometric and clinical variables in a spatial context.

3.3 Standard polynomial regression

Instead of bivariate regression for BW and GA, univariate analyses for one of them are common in gynaecological and obstetric research (e.g., Skjærven *et al.*, 2000; Weiss *et al.*, 2014; Frederick *et al.*, 2008), perhaps adjusted for the other, or with a dichotomous response like 'small for GA' (e.g., Polakowski *et al.*, 2009; Thompson *et al.*, 2001).

For the remainder of this chapter, let K be the number of observations, y_{k1} the observed BW and y_{k2} the observed GA, $k = 1,\dots,K$, from continuous response variables Y_1 and Y_2. Observations x_{kl} of L covariates x_l, $l = 1,\dots,L$, are considered (not to be confused with the PFOA concentration variable denoted by X in Chapter 2).

In preliminary studies (see Item 1 in Appendix B.2.2 for documentation), a regression model is confirmed as the most suitable among univariate BW models, where GA is included as a covariate in the form of a polynomial P_γ of degree three:

Polynomial regression models

$$y_{k1} = \beta_0 + P_\gamma(y_{k2}) + \sum_l \beta_l x_{kl} + \varepsilon_k,$$

$k = 1,\dots,K$, with independent $\varepsilon_k \sim N(0,\sigma^2)$ are applied for BW response, with observed GA y_{k2} as explanatory variable and some of the further covariates numbered $l = 1,\dots,L$ (see Table 3.1 on page 38).

Among the usual fractional polynomials (Royston and Sauerbrei, 2008) of degree one or two, the resulting mean prediction errors are very close to each other. With regard to the residual sum of squares, the Akaike information criterion, the Bayesian information criterion and the maximum prediction error (i.e., for outlying data), the polynomial

$$P_\gamma(y_{k2}) = \gamma_1 y_{k2}^2 + \gamma_2 y_{k2}^3$$

performs best. However, a model with 'full' polynomial

$$P_\gamma(y_{k2}) = \gamma_1 y_{k2} + \gamma_2 y_{k2}^2 + \gamma_3 y_{k2}^3$$

is even better in this respect; it is in accordance with gynaecological and obstetric literature (e.g., Gardosi *et al.*, 1995; Salomon *et al.*, 2007) and therefore preferred. Standard diagnostic plots show no hints on violations of model assumptions; in particular, Gaussian assumption is thoroughly fulfilled; moreover, the residuals display no anomalies when plotted against gestational age.

Therefore, this standard model is applied in a Bayesian interpretation using the software `BayesX` (Belitz *et al.*, 2020, see Section 3.4.1) for further evaluation and comparison to the Bayesian distributional copula regression. Estimates of regression coefficients are summarised in Table 3.4 on page 51.

3.4 Bayesian copula distributional regression

Below, the employed class of Bayesian conditional copula models is outlined within a distributional regression framework, developed by Klein and Kneib (2016). The focus is on such model components that are relevant for the analyses in this chapter. Especially, the predictors of the structured additive regression models (Fahrmeir *et al.*, 2004) are restricted to linear components, due to computational reasons and because most of the explanatory variables are binary. Furthermore, there is no hierarchical modelling, in particular not in relation to space, as a small region is considered, divided in only few spatial units, which are equivalent to the environmental exposure to be included later. The more general perspective can be found in Klein and Kneib (2016) and the references therein.

From now on, bivariate responses with observations (y_{k1}, y_{k2}), $k = 1, \dots, K$, are considered. Let probability density functions be denoted by f_1 and f_2 and cumulative distribution functions (CDFs) by F_1 and F_2, respectively, for the continuous response variables Y_1 and Y_2 as before.

3.4.1 Bivariate copula models and implementation

A bivariate copula is defined by a CDF $C_\rho : [0,1] \times [0,1] \rightarrow [0,1]$ such that the joint CDF of Y_1 and Y_2 can be written as

$$F(y_1, y_2) = C_\rho(F_1(y_1), F_2(y_2)) =: C_\rho(u, v).$$

Sklar's theorem (Sklar, 1959) ensures that C_ρ always exists and is unique for continuous Y_1 and Y_2 while F_1 and F_2 are uniformly distributed. With a copula density $c_\rho(\cdot,\cdot)$, the joint density of Y_1 and Y_2 can be written as

$$f(y_1, y_2) = c_\rho(F_1(y_1), F_2(y_2)) \cdot f_1(y_1) \cdot f_2(y_2)$$

and a conditional density as

$$f_{1|2}(y_1|y_2) = c_\rho(F_1(y_1), F_2(y_2)) \cdot f_1(y_1). \tag{3.1}$$

While this representation is unconditional, the results can be extended to the regression context (Patton, 2006), see also Section 3.2.

There are various families of copulas, characterised by a parameter ρ representing the degree and form of dependence between Y_1 and Y_2. Implemented for application in these analyses are the Gaussian copula family with density

$$\begin{aligned} c_\rho(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}\cdot\frac{\rho}{1-\rho^2}\cdot\left\{\rho(\Phi^{-1}(u))^2 - 2\cdot\Phi^{-1}(u)\cdot\Phi^{-1}(v) + \rho(\Phi^{-1}(v))^2\right\}\right], \\ \rho \in (-1,1) \end{aligned}$$

(where Φ^{-1} is the inverted CDF (quantile function) of a univariate standard Gaussian distribution), the Clayton copula family with density

$$\begin{aligned} c_\rho(u,v) = (1+\rho)(uv)^{-1-\rho}\left(u^{-\rho}+v^{-\rho}-1\right)^{-2-1/\rho}, \\ \rho \in (0,+\infty), \end{aligned}$$

and the Gumbel copula family with density

$$\begin{aligned} c_\rho(u,v) = \frac{1}{uv}(-\ln u)^{\rho-1}(-\ln v)^{\rho-1}\exp\left(-z^{1/\rho}\right)\left(z^{2/\rho-2}-(1-\rho)z^{1/\rho-2}\right), \\ \text{where } z := (-\ln u)^\rho + (-\ln v)^\rho, \\ \rho \in (1,+\infty). \end{aligned}$$

The Gaussian copula allows for linear dependence and symmetry only; its response distribution is equivalent to that of a standard multivariate regression model. The others are more flexible and assume less symmetry. The Clayton copula allows for a stronger non-linear dependence between the two variables within the region of their extremely low values (tail dependence), whereas the Gumbel copula allows for upper tail dependence (Nelsen, 2006).

The two marginal distribution families can be chosen independently from each other and from the copula. Besides the Gaussian distribution $N(\mu,\sigma^2)$ with expected value μ and variance σ^2, a Dagum distribution with density

$$f_{p,a,b}(y) = \frac{ap}{y}\cdot\frac{(y/b)^{ap}}{((y/b)^a+1)^{p+1}},$$

shape parameters $p > 0$ and $a > 0$ and dispersion parameter $b > 0$ is implemented.

All this is considered conditional on covariates. For any parameter θ of the joint distribution of Y_1 and Y_2, being either one of the assumed marginal distributions or of the copula (i.e., $\theta \in \{\mu,\sigma^2,p,a,b,\rho\}$ in this case), a linear predictor

$$\eta^{(\theta)} = \beta_0^{(\theta)} + \beta_1^{(\theta)}x_1 + \ldots + \beta_L^{(\theta)}x_L$$

is formed from the covariates, possibly just from a part of them or even reduced to the intercept. Link functions h_θ with $\theta = h_\theta^{-1}(\eta^{(\theta)})$ are specified appropriately for the respective parameter spaces:

$$\begin{aligned}
\mu &= \eta^{(\mu)}, & \\
\theta &= \exp(\eta^{(\theta)}) & \text{for } \theta \in \{\sigma^2, p, a, b\}, \\
\rho &= \frac{\eta^{(\rho)}}{\sqrt{1+(\eta^{(\rho)})^2}} & \text{for the Gaussian copula,} \\
\rho &= \exp(\eta^{(\rho)}) & \text{for the Clayton copula,} \\
\rho &= \exp(\eta^{(\rho)})+1 & \text{for the Gumbel copula.}
\end{aligned}$$

The covariates to be included to the linear predictor $\eta^{(\theta)}$ can be separately selected for all parameters $\theta \in \{\mu, \sigma^2, p, a, b, \rho\}$.

All models are estimated using a developer version of the BayesX software (Belitz *et al.*, 2020), which implements fully Bayesian inference based on Markov chain Monte Carlo simulation techniques, to simultaneously and independently assess the covariates' influences on location, dispersion or shape parameters of the marginals and on the copula parameter. BayesX also includes routines for spatial data, hierarchical models and structured additive models (Fahrmeir *et al.*, 2004) in general, see Klein and Kneib (2016) for details. BayesX offers rotation of copula data $(u,v) \in (0,1)^2$ by 90° or −90°, to cover all four possible directions of tail dependence.

3.4.2 Preliminaries, preparation and outline of procedure

Bayesian distributional copula regression models (Section 3.4.1) are applied to the perinatal registry data. The models are evaluated using the anonymised excerpt of $K = 4451$ observations representing the five postal code areas of the town of Arnsberg in the upper Ruhr region, North Rhine-Westphalia, Germany, for the years of 2003 until 2014 (Section 3.1).

Besides the BayesX software, calculations have been performed using the R environment (R Core Team, 2020), with the Dagum distribution from the VGAM package (Yee, 2020) and copula distributions from copula (Hofert *et al.*, 2020). See Appendix B.2.2 for documentation.

In preliminary studies, univariate frequentistic linear regression models for both marginal response variables: birth weight (BW) y_1 and gestational age (GA) y_2, ignoring the respective other, have also been considered. In both cases, no hints on heteroscedasticity or other striking patterns in the residuals have been found. However, the Gaussian assumption has seemed somewhat violated in the lower tail of the data, especially for the GA model, resulting in a slightly higher variance of lower fitted values. The latter model has also revealed a separation of the predicted values in two groups, presumably corresponding to whether children have been delivered by Caesarean section.

After preparation steps of data import and cleansing, the response data of observations $k = 1, \ldots, K$ are standardised for numerical reasons only, without affecting the results, since the original responses can easily be recovered by linear back-transformation. Specifically, to employ a Gaussian for the marginals in BayesX, data-independent values for mean and standard deviation in a reasonable scale are used to yield

$$\tilde{y}_{k1} := \frac{y_{k1} - 3500}{500} \text{ for BW and } \tilde{y}_{k2} := \frac{y_{k2} - 280}{14} \text{ for GA.}$$

To apply the Dagum marginal (with positive support), BW is normalised to

$$\tilde{y}_{k1} := \frac{y_{k1}}{500},$$

while GA is also inverted to a more appropriate shape by

$$\tilde{y}_{k2} := \frac{322 - y_{k2}}{14}, \tag{3.2}$$

to have the main part of the data closer to zero and the tail on the right. 322 days = 46 weeks exceed the maximum observable GA.

The choice of the optimal copula regression model is a stepwise procedure, outlined in Figure 3.2. Marginal distribution families are chosen by applying Gaussian and Dagum models to both univariate responses; in all these four models and for all parameters therein ($\theta \in \{\mu, \sigma^2\}$ or $\theta \in \{p, a, b\}$, respectively), covariates are selected if the 95% credible intervals of their respective $\beta_l^{(\theta)}$'s do not include zero; the optimal models per family are compared by probability integral transform values, quantile residuals and log-scores (Section 3.4.3). The optimal marginals of BW and GA are combined with all possible copula families (including rotations); covariates for the copula parameter are selected and these models compared by information criteria (Section 3.4.4). For the linear predictors of all parameters, all covariates listed in Table 3.1 on page 38 are considered, without interactions. The final model is evaluated in terms of prediction performance and substantive results (Sections 3.4.5 and 3.5).

3.4.3 Marginal model selections

After non-significant covariates are excluded, different univariate distributional models (Klein *et al.*, 2015) are fitted (see Item 2 in Appendix B.2.2 for documentation) and compared using logarithmic scores (log-scores, Gneiting and Raftery, 2007). To compute the log-scores, a four-fold cross-validation is implemented, for which the observations are randomly assigned to subsamples of equal size. Using the estimated model based on three subsamples, individual log-scores for the respective left-out subsample are computed using the `R` package `scoringRules` (Jordan *et al.*, 2019). See Item 3 in Appendix B.2.2 for documentation. The average log-scores for Dagum and Gaussian distribution are very close to each other in the case of BW; however, for GA, the Dagum distribution has a notably better fit (Table 3.2).

Table 3.2: Mean log-scores to evaluate prediction performances of the two models for the marginal responses. A lower score indicates a better fit.

	Marginal	
Model	Birth weight	Gestational age
Gaussian	7.66	4.07
Dagum	7.63	3.74

These findings are confirmed by graphical evaluation (see Item 3 in Appendix B.2.2 for documentation) of the probability integral transform values $F(y; \hat{\beta})$ and the corresponding normalised quantile residuals $\Phi^{-1}(F(y; \hat{\beta}))$ (cf., Dunn and Smyth, 1996), where posterior means of the respective $\beta_l^{(\theta)}$'s are employed, and y (BW or GA) is on the standardised scale. The theoretically expected uniform distribution of the $F(y; \hat{\beta})$'s is well recognisable for the Dagum fits, while it is strongly violated for the Gaussian

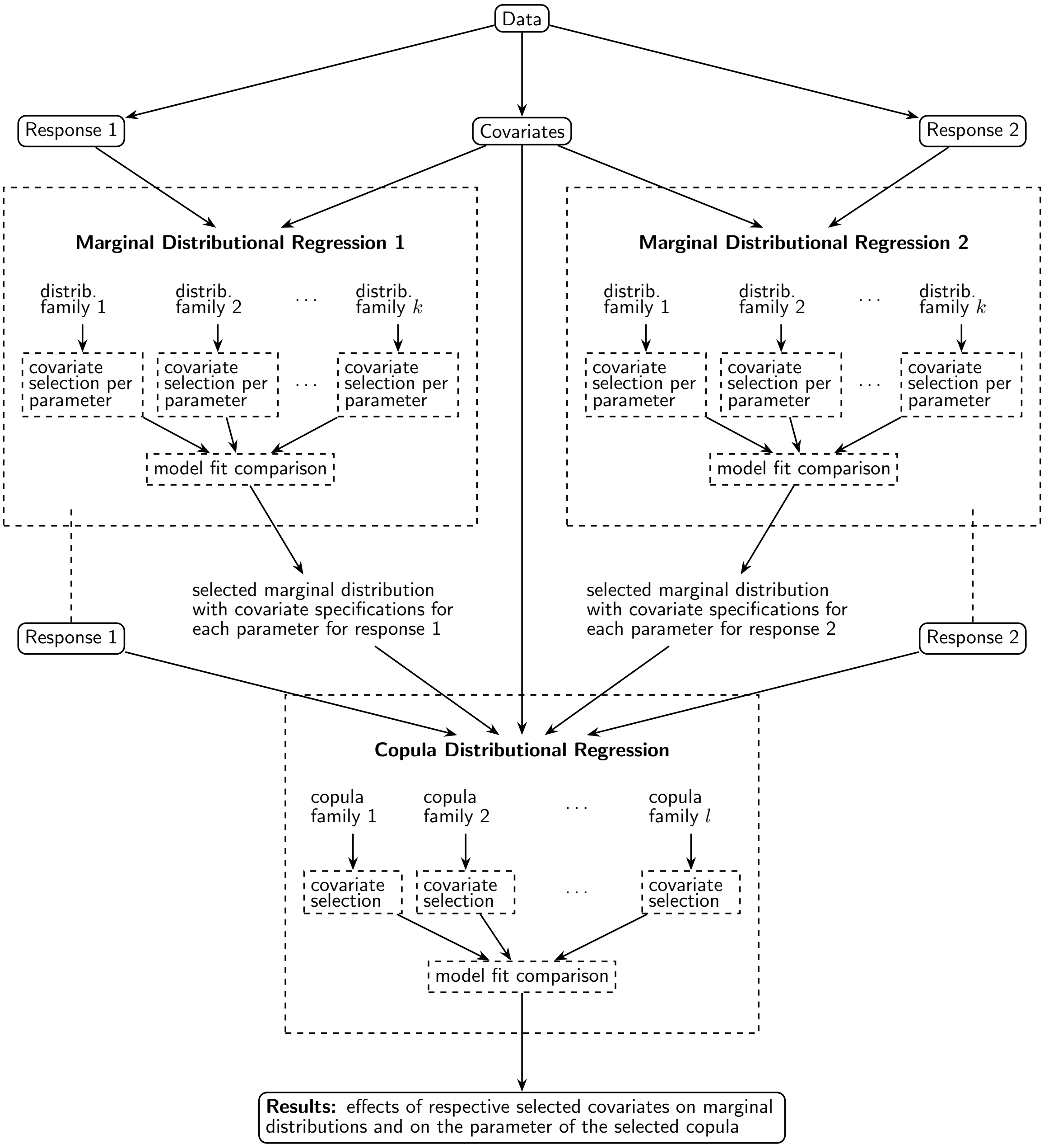
Data
Response 1
Covariates
Response 2
Marginal Distributional Regression 1
distrib. family 1
distrib. family 2
. . .
distrib. family k
covariate selection per parameter
covariate selection per parameter
. . .
covariate selection per parameter
model fit comparison
Marginal Distributional Regression 2
distrib. family 1
distrib. family 2
. . .
distrib. family k
covariate selection per parameter
covariate selection per parameter
. . .
covariate selection per parameter
model fit comparison
selected marginal distribution with covariate specifications for each parameter for response 1
selected marginal distribution with covariate specifications for each parameter for response 2
Response 1
Response 2
Copula Distributional Regression
copula family 1
copula family 2
. . .
copula family l
covariate selection
covariate selection
. . .
covariate selection
model fit comparison
Results: effects of respective selected covariates on marginal distributions and on the parameter of the selected copula

Figure 3.2: Outline of the proposed procedure to choose optimal marginal and copula families in bivariate Bayesian distributional regression.

fit of GA; a similar structure as for the latter remains also for the Gaussian fit of BW, but considerably weaker (Figure 3.3). A quantile-quantile-plot of the $\Phi^{-1}(F(y;\hat{\beta}))$'s (Figure 3.4) shows a slightly better fit of the Dagum model in either case, especially in the lower range, but more striking for GA, where a distinguishable structure remains with the Gaussian. These results are convincing to use the Dagum distribution for the GA marginal in the further analyses.

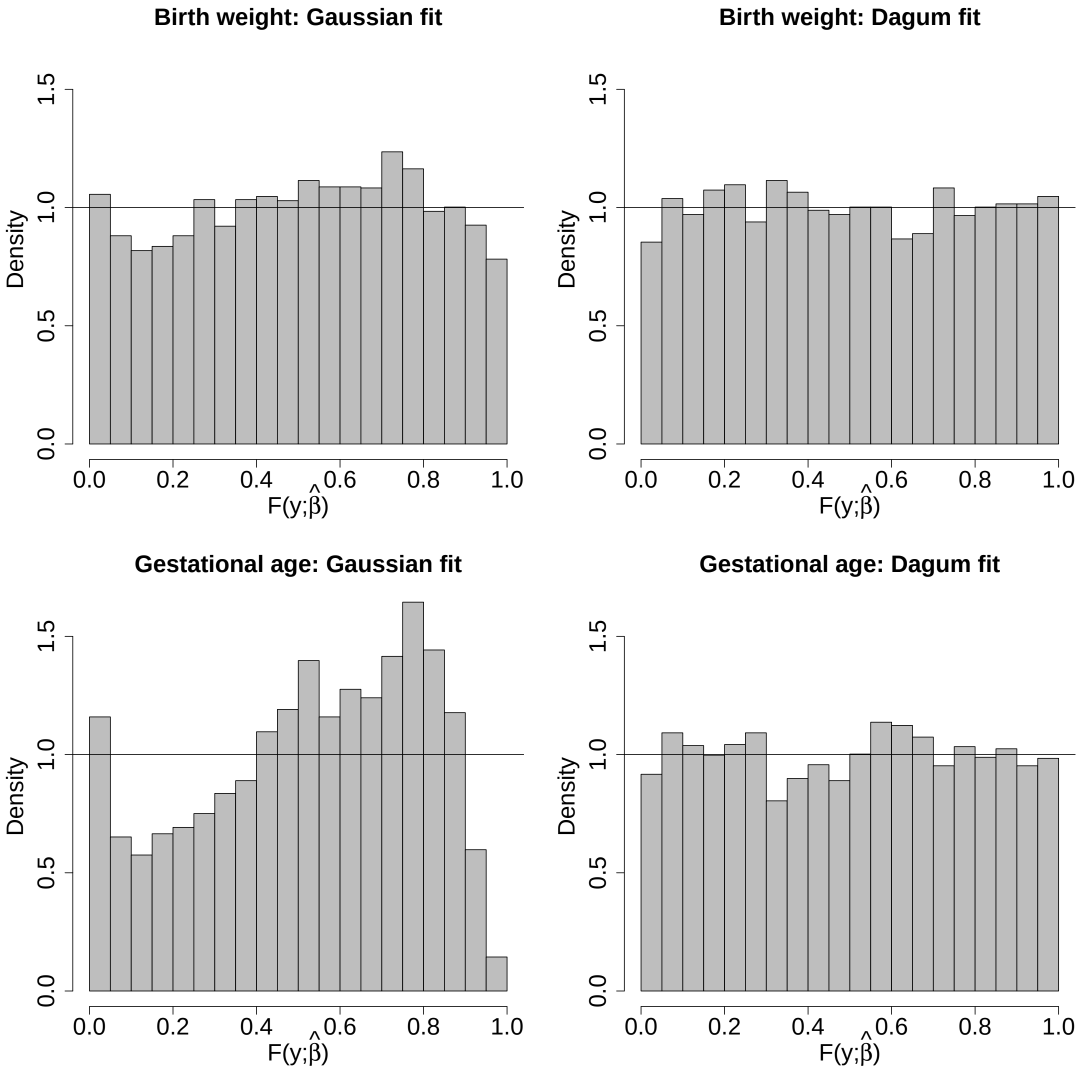

Figure 3.3: Histograms of probability integral transform values $F(y;\hat{\beta})$ (with posterior means of the β's plugged in) for the two models and the two marginal response variables.

For BW, however, the results are less clear and there are reasons to remain with the Gaussian distribution in case of doubt: In this application, influences of covariates on the mean (and variability) of BW are of primary interest – characteristics directly represented by the $N(\mu,\sigma^2)$ parametrisation; and there is at least some evidence above that this distribution family does not fit essentially worse than the other. In this parametrisation, effects of covariates are easily and directly interpretable. By contrast, the interpretation of effects on the three Dagum parameters is quite complicated and sometimes ambiguous in terms of substantive results (cf., Section 3.4.5). Results from the Gaussian fit are furthermore directly comparable to other studies, especially with standard regression models (cf., Section 3.5). So, the Gaussian is used

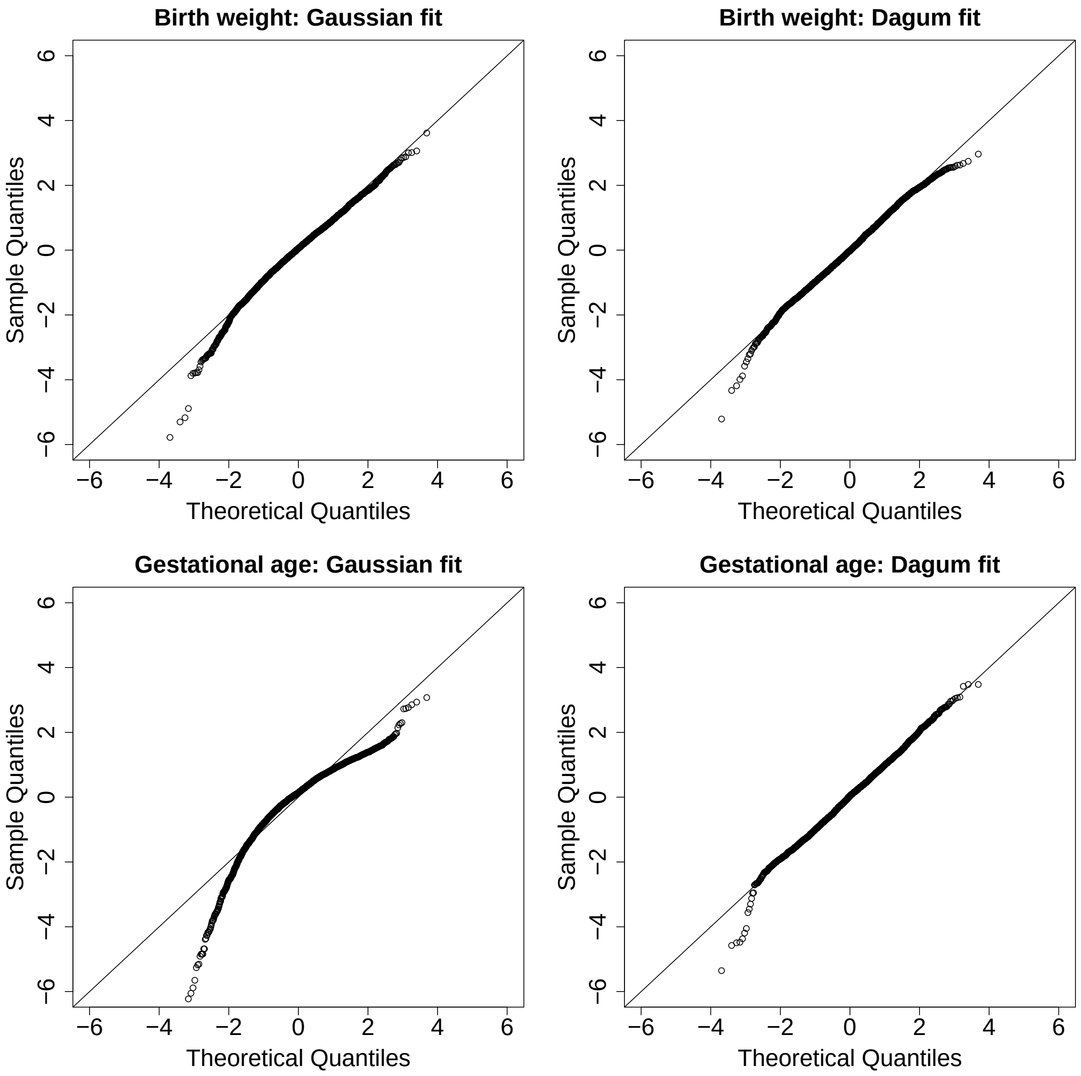

Figure 3.4: Quantile-quantile-plots of randomised quantile residuals $\Phi^{-1}(F(y;\hat{\beta}))$ (with posterior means of the β's plugged in) for the two models and the two marginal response variables.

as BW marginal in the further analyses.

3.4.4 Copula model selection

As a motivation, the usefulness of flexible copula regression and non-linear dependence structures is illustrated by comparing confidence intervals of rank correlation coefficients (R package spearmanCI, de Carvalho, 2018) of BW and GA from subsets of data selected by certain covariate choices: In case of a discrete covariate, the respective levels are considered; in case of a continuous covariate, observations are ordered by the covariate values and grouped together into four subsets of equal size. See Item 4 in Appendix B.2.2 for documentation. Differences between these levels or subsets are especially pronounced for the sectio and weight-gain covariates (Figure 3.5). Therefore, it seems reasonable to apply distributional regression methods also with respect to the copula parameter.

The Gaussian, Clayton, and Gumbel copula are compared, the latter two also with copula data (u,v)

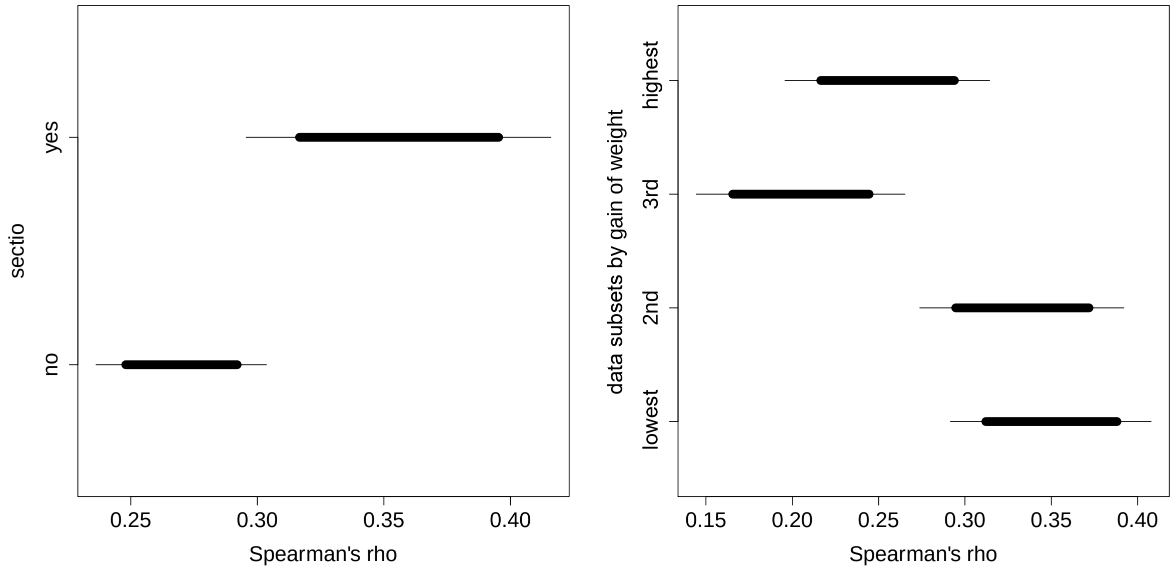

Figure 3.5: Rank correlations of birth weight and gestational age (80% and 95% confidence intervals) from subsets of data selected by certain covariate choices: left: Caesarean section, two groups; right: maternal gain of weight, ordered observations grouped into four equal subsets; other covariates do not feature such visible differences between subsets.

rotation by 90° and $-90°$, using the deviance information criterion (DIC, Spiegelhalter *et al.*, 2002) and the widely application information criterion (WAIC, Watanabe, 2010). See Item 5 in Appendix B.2.2 for documentation. Too many covariates with respect to the ρ parameter of the Clayton and Gumbel copulas lead to technical problems with the given amount of data, where the MCMC sampling is unable to initiate. Therefore, they are pre-selected based on the variability of correlation coefficients reported above and by tentatively adding covariates one by one. With respect to the three copula families, the Clayton copula model without rotation yields the best DIC and WAIC values.

In conclusion from the model fitting, a Clayton copula with Dagum marginal for GA and Gaussian marginal for BW emerges as the best. The predictor specifications for each of the six model parameters ρ, p, a, b, μ and σ^2 are given in Table 3.3, the respective link functions specified in Section 3.4.1 are employed.

3.4.5 Evaluation of effects

Based on the results from Sections 3.4.3 and 3.4.4, the Bayesian bivariate distributional copula regression model (see specification in Table 3.3) is applied to the perinatal registry data.

Significant influences, i.e., $0 \notin \mathrm{CI}_{95\%}\left(\beta_l^{(\theta)}\right)$, of covariates on BW's mean are quantified in Table 3.4. Apart from this, covariates also significantly influence other model parameters (see overview in Table 3.3). BW's scale (σ^2) is higher for male children, in case of sectio, for higher maternal BMI and if the mother smokes.

For the Dagum distribution of gestational age (GA), the shape parameter p is higher in case of induction and lower in case of sectio. The shape parameter a increases with the maternal gain of weight and is lower in case of sectio, induction and if the mother smokes. The scale parameter b increases

Table 3.3: Final bivariate copula model specification and result overview: composition of the linear predictor $\eta^{(\theta)}$ from the covariates, per parameter $\theta \in \{\mu, \sigma^2, p, a, b, \rho\}$ of the chosen marginal and copula families, together with the employed link functions. Included covariates are marked by $+$, $-$ or $\circ$, where $+$ and $-$ denote significant positive and negative effects, respectively, and $\circ$ is for no significant effect in the final evaluation.

	Response:	Birth weight		Gestational age			Copula:
	Family:	Gaussian		Dagum			Clayton
	Parameter:	μ	σ^2	p	a	b	ρ
Components of $\eta^{(\theta)}$	sex (female)	$-$	$-$				
	prev. pregn.	$+$				$+$	
	sectio	$-$	$+$	$-$	$-$	$+$	$+$
	induction	$+$		$+$	$-$	$-$	
	mat. age						
	mat. height	$+$					
	mat. BMI	$+$	$+$				
	weight gain	$+$			$+$		
	smoking	$-$	$+$		$-$		
	single	$\circ$					
	employed			$\circ$			
Link: $\eta^{(\theta)} = \dots$		μ	$\ln \sigma^2$	$\ln p$	$\ln a$	$\ln b$	$\ln \rho$

with the number of previous pregnancies and in case of sectio, and is lower in case of induction. If the distribution's

$$\text{median } = b \cdot \left(-1 + 2^{1/p}\right)^{-1/a},$$

the monotonically increasing link functions and the inverting transformation of the data from Equation (3.2) on page 44 are considered, these results can be qualitatively interpreted such that GA is higher for decreasing p or b or increasing a, e.g., with increasing maternal gain of weight. But it becomes also obvious that this interpretation is generally rather difficult. It leads to no consistent results in the case of sectio or induction.

For the copula parameter, only the information, whether the child has been delivered by Caesarean section, emerges as a stably estimated significant influence. Taking the intercept into account, the dependence between BW and GA measured in this way turns out to be surprisingly weak, in fact not far from independence ($\rho \approx 0.40$ for children delivered by Caesarean section, $\rho \approx 0.14$ for the others), although significantly positive.

Examples of bivariate predictions depending on covariate values (levels) are shown in Figure 3.6. See Item 6 in Appendix B.2.2 for documentation.

3.5 Comparison of standard and copula approach

A four-fold cross-validation is performed and the obtained prediction samples of BW from copula and standard model are compared to the observed values by log-scores as in Section 3.4.3. See Item 7 in Appendix B.2.2 for documentation.

To make the copula model's results comparable, after estimation of the bivariate joint distribution,

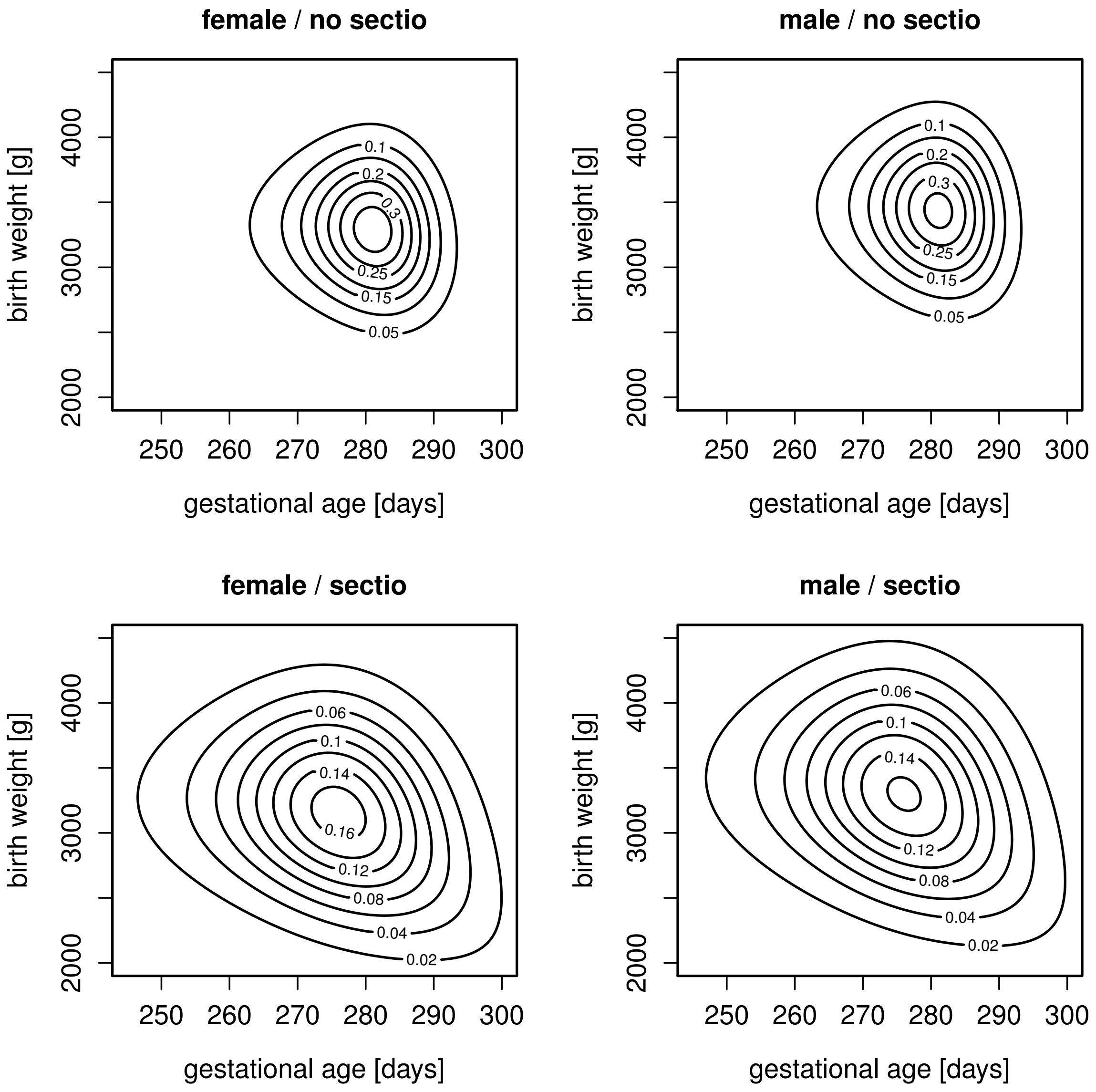

Figure 3.6: Bivariate density of birth weight and gestational age, as predicted from the copula model with posterior mean parameters plugged in, conditional on certain selected exemplary covariate values (the others are fixed to: maternal height: 170 cm, maternal BMI: 20 kg m^{-2}, maternal gain of weight: 10 kg; and all others set to 'no' or 0, respectively).

Table 3.4: Regression coefficients (posterior mean and standard deviation) regarding the parameter μ of (standardised) BW, estimated in the polynomial and the copula distributional regression model.

Covariate	Coefficients in regression models			
	Polynomial		Copula	
sex (female)	-0.2908	$(\pm\, 0.0237)$	-0.2896	$(\pm\, 0.0273)$
previous pregnancies	0.0583	$(\pm\, 0.0085)$	0.0484	$(\pm\, 0.0065)$
sectio	not signif.		-0.2907	$(\pm\, 0.0535)$
induction	not signif.		0.0870	$(\pm\, 0.0280)$
maternal age	not signif.		not signif.	
maternal height	0.0279	$(\pm\, 0.0018)$	0.0290	$(\pm\, 0.0015)$
maternal BMI	0.0302	$(\pm\, 0.0023)$	0.0368	$(\pm\, 0.0029)$
maternal gain of weight	0.0144	$(\pm\, 0.0022)$	0.0249	$(\pm\, 0.0024)$
maternal smoking	-0.0308	$(\pm\, 0.0029)$	-0.0416	$(\pm\, 0.0031)$
mother is single	-0.1271	$(\pm\, 0.0480)$	not signif.	
mother is employed	not signif.		not signif.	
gestational age	-2.6145	$(\pm\, 0.1717)$		
squared gestational age	0.0112	$(\pm\, 0.0007)$		
cubic gestational age	$-15\text{E}{-}6$	$(\pm\, 0.9\text{E}{-}6)$		

predictions of BW conditional on GA are of interest. To evaluate the conditional distribution with the density from Equation (3.1) on page 42, random numbers are drawn via rejection sampling with a uniform envelope extended to a large enough range. Thereby, the observed GA values y_{k2}, parameter estimates $\hat{\theta}$ obtained from samples of the posterior $\hat{\beta}_l^{(\theta)}$'s, and the covariate values of the respective observations are used.

For the standard model, there results an average log-score of 7.41; for the copula model, with respect to BW response conditional on GA, it is 7.67. Thus, the standard model designed for this specific situation performs better than the more general copula model, when reduced to the prediction of BW.

For the vast majority of BW predictions, the distributional copula regression model is close to univariate polynomial regression. The residual and comparison plots in Figure 3.7 (see Item 7 in Appendix B.2.2 for documentation) show how the models agree, especially in mean (bottom left). However, extremely low BWs are correctly predicted by the polynomial alone (top left), while their observations diverge from the copula model predictions (top right: fitted values are in the range of the main part of data, but residuals are too far to the negative). A closer range of predictions from the copula model is also visible (top right). The residual plot for GA from copula regression (Figure 3.7, bottom right) also reveals rather poor prediction of extremely low values, which often coincide with very low BW; besides, there emerge two distinguishable groups of GA predictions, presumably in connection with the highly influential sectio and induction covariates.

Due to independent and simultaneous estimation of marginals and copula, estimates of the regression coefficients with regard to BW's mean are very similar in both models, but their significances differ (Table 3.4 on page 51).

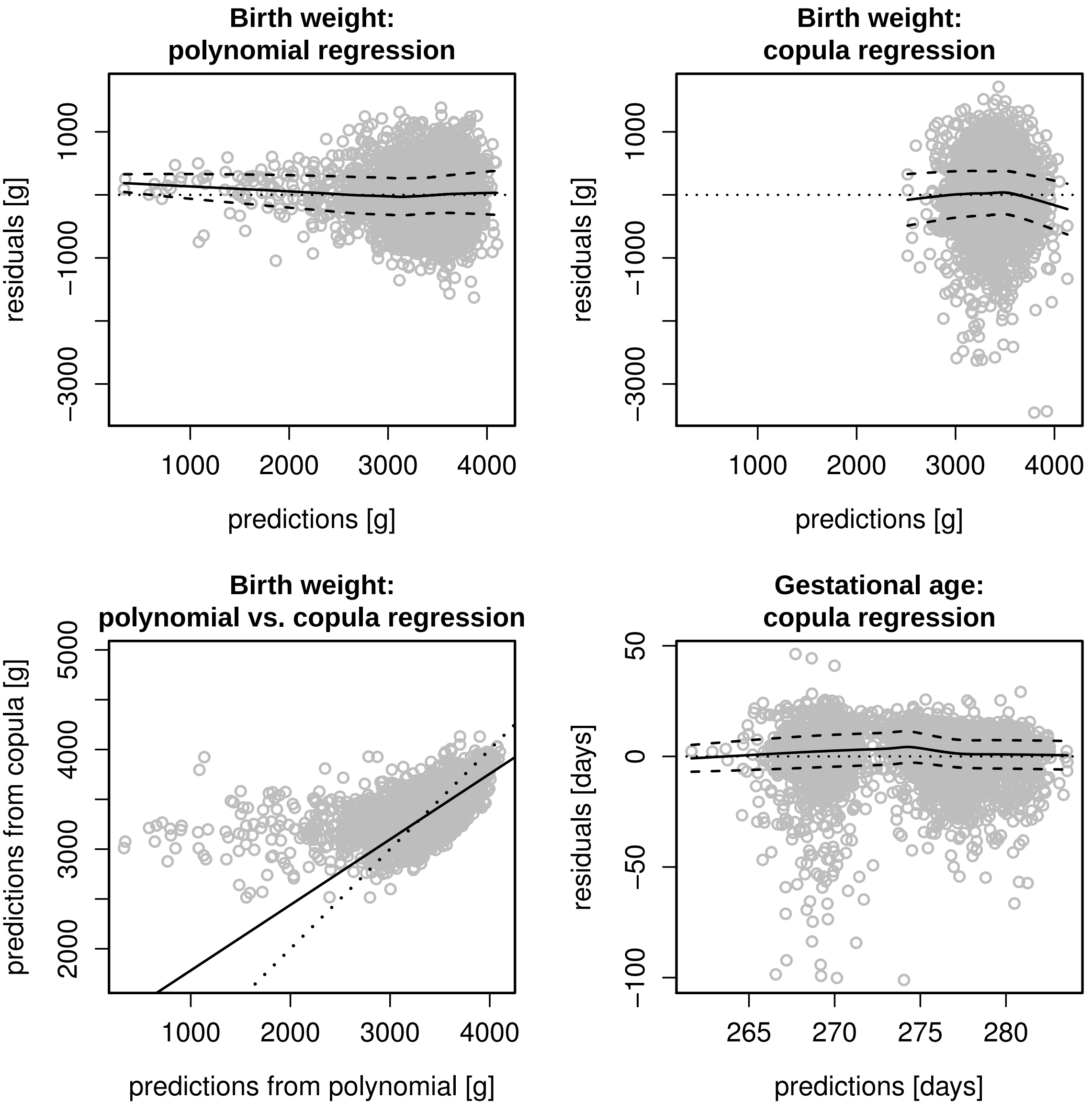

Figure 3.7: Top: residual plots for birth weight from polynomial (left) and copula distributional (right) regression model, each with smoothed mean and standard deviation lines; bottom right: the same for gestational age; bottom left: predictions of birth weight from polynomial and copula distributional regression model, plotted against each other per observation, with bisecting line (dotted) and robustly estimated principal axis of the plotted data (‘direction of main point cloud’, solid); predictions for all figures obtained from a cross-validation study.

3.6 Discussion

3.6.1 Data quality

The secondary data from the perinatal registry have not originally been collected to be scientific material, but for quality assurance. As such, they are nonetheless very informative with regard to procedures in obstetric health care, like the birth mode (Caesarean section, induction), which turns out as an important covariate. Perinatal registry data from Germany have already been used in scientific studies, e.g., by Reeske *et al.* (2011).

On the other hand, measurement accuracy varies (e.g., one hospital reports birth weight (BW) accurate to 1 g, another to 10 g). This holds also for gestational age (GA), where the unit of days found in the data, instead of weeks, is not common and varying accuracy of partially rounded data is possible.

In general, GA data are subject to uncertainty of reporting, measurement, clinical estimation and documentation (e.g., Polakowski *et al.*, 2009), although they have been carefully checked for plausibility in the 'PerSpat' project. However, the socio-economic status of a family can presumably be represented worse than in other studies, which are designed for this purpose. Information on nationality is given, but cannot be used like 'ethnicity' as found in, e.g., US studies.

Maternal smoking is self-reported and perhaps biased towards a socially desirable answer; nonetheless, these data are accurate enough such that an effect of smoking in line with other studies from the literature (see Section 3.6.3) is detected despite the remaining noise.

Another issue is related to data accuracy: Although both response variables are justifiably considered as continuous, their measurement is actually not. As pointed out, BW values are but accurate up to certain levels, and GA is given in discrete units of time. This affects the (u, v)-data after non-linear transformation by CDFs in the copula framework, which are then strictly speaking not uniformly distributed. However, modelling them conditional on covariates, the diagnostic plots above hint at adequate uniform distributions.

3.6.2 Gestational age and dependence structure

There are strong effects of all three polynomial terms of GA in the univariate model and the increasing trend of the mean BW along GA decreases again towards the end. This phenomenon is also reported in other studies (e.g., Skjærven *et al.*, 2000; Voigt *et al.*, 1996) and could be an effect of medical decisions to deliver foetuses with high weights rather early by induction or sectio and to avoid such treatments for a longer time when foetal weight is low.

The dependence structure between BW and GA responses is difficult to estimate and perhaps ambiguous. Information criteria of the non-rotated Clayton copula's fit and of that rotated by $-90°$ are quite close to each other. The dependence is estimated as rather weak in terms of the copula parameter; especially, there is no visible tail in predictions from this model (Figure 3.6 on page 50). Although these results seem to be in contradiction to the expectation, and so one of the intended uses of the copula model is doubtful, they can be assumed to be effects of the said trend reversal of BW in the region of very high GA.

Prior to analysis, the Clayton copula has been assumed to be a little more appropriate for the given data situation, as it is able to model lower tail dependence. A particularly visible dependence of GA and BW can be found for extremely low values of these variables (Figure 3.1 on page 38). However, when the transformation of the GA data from Equation (3.2) on page 44 is taken into account, the lower tail is actually located in the region of high GA, where the mean BW tends to become lower again and, thus, the dependence may be different than in the main part of the data. Now, as the Clayton copula without rotation is indeed fitted best to the given transformed data, the region of high GA is confirmed to represent the actual tail and, thus, the few extremely outlying data for low GA are not reflected by this model.

All this and Figure 3.1 hint at two tails in the data (in the regions of both high and low GA), which cannot both be modelled by the considered copula families. In this case, the less prominent tail towards high gestational ages (the lower tail after pre-inversion of data) would be more important due to more observations. However, the main part of data in their ‘centre’ seems to predominate any estimation.

In any case, this estimation of tail dependence is very sensitive to GA observation. Any data inaccuracies, which are generally possible for GA (e.g., Polakowski *et al.*, 2009), have an impact on regression models.

3.6.3 Model comparison and substantive evaluation

An important benefit of the distributional copula regression model are visible differences between groups, with respect to both scale and dependence: Figure 3.6 on page 50 shows examples of predictions, distinguished by sex and sectio. It becomes apparent, that the variability and structure of the response data is deeper explained, when influences of covariates on more parameters than only the means are allowed – unlike in a standard regression model (Section 3.5).

The relative closeness of both models’ birth weight predictions (Figure 3.7 on page 52, bottom left) for most of the data is remarkable, as the purpose of the copula model is more general and only a partial, one-dimensional result is considered, whilst the polynomial is more specified for the data situation. However, when only birth weight is of interest, it remains much more efficient to stick with a univariate regression approach.

Considering both models together, conclusions are obtained that go beyond effects of covariates on BW. A striking example are the possible relationships between BW, GA and the Caesarean section covariate (cf., Tables 3.3 and 3.4 from page 49): The latter has a significant influence on BW according to the copula model, where GA is separately estimated, while this does not hold for the standard model, where GA is present as a significant covariate. The sectio covariate also significantly influences the parameters of the Dagum distribution of GA in the copula model as well as the copula parameter (Section 3.4.5). According to these results, the influence of the sectio covariate is in fact manifold (cf., e.g., Stotland *et al.*, 2004), but this can only be discovered using the bivariate model, which provides more extensive conclusions in this respect. In the standard model, the importance of the sectio covariate disappears; it is presumably predominated and in parts mediated by GA. Similar considerations hold for the induction covariate.

As a different example, both models agree with respect to the significant effect of smoking on BW (Table 3.4). There is also an effect on GA found in the bivariate model (Section 3.4.5, Table 3.3), but only with respect to one Dagum parameter and, thus, presumably less important. So, there seems to be no mediation by GA in the standard model. The influence of smoking on both BW, GA and on the risk of pre-term birth or ‘small for GA’ has also been found in many studies with univariate responses (e.g., Ko *et al.*, 2014; Polakowski *et al.*, 2009; Kyrklund-Blomberg and Cnattingius, 1998; Li *et al.*, 1993).

3.7 Conclusions

For regression analyses regarding birth weight (BW), the bivariate modelling jointly with the gestational age (GA) emerges as very productive. The results allow insights into the relationships between these two variables and others, e.g., Caesarean section, avoiding mediation.

Distributional regression, where any parameter of the bivariate distribution is estimated conditional on covariates, is an appropriate instrument to explain the variability and structure of the perinatal registry data in more depth. While a Gaussian distribution is well fitted to the marginal BW data, the heavily skewed GA data are better modelled by the more flexible Dagum distribution. Effects of many explanatory variables on both BW and GA can be distinguished. A copula model is useful to simultaneously and independently estimate the dependence structure and the marginals. The perinatal data are fitted better by the lower-tail Clayton copula than by the Gaussian and the Gumbel. However, the estimated dependence is weak and the result not entirely convincing with regard to possible tail dependence.

As the trend of BW along GA tends to reverse and BW becomes lower again for observations with very high GA, this bivariate structure may be more asymmetric than what the Clayton copula can cope with. The effect of GA on BW is better explained by a univariate polynomial regression, especially regarding data outside the main cloud of observations.

The models allow the inclusion of the local PFOA concentration in drinking water as an additional environmental explanatory variable, spatio-temporally assigned using the mother's postal code and the date of birth (Chapter 4).

3.8 Perspectives

The employed Dagum distribution fits well to the strongly asymmetric GA data, but other families might also be suitable. They should be just as flexible and, therefore, have several parameters including shape, even when the structure of results is unfavourable for substantial interpretation. By contrast, a Gaussian distribution assumption is not implausible for BW conditional on covariates, and it should be preferred for the sake of interpretability and comparability of results, despite slight model fit discrepancies.

Since the available methods cannot reflect further asymmetry but the latter is likely and visible in bivariate plots of data (Figure 3.1 on page 38) and residuals (Figure 3.7 on page 52), other copula families as well as specific data transformations might be useful for complicated bivariate response data shapes as here. E.g., the skewed t-copula allows for strong asymmetry and non-linearity (Sun *et al.*, 2008), but estimation and interpretation of the multiple parameters are inconvenient compared to the applied one-parametric representation of dependence structure.

When observations with low BW and low GA are of particular interest, it is recommendable to stick with the conventional approach to model the binary response 'small for GA' instead of continuous BW data. In the same distributional regression context with copulas as here, 'small for GA' is modelled jointly with GA by Klein *et al.* (2019). Another approach is an extreme value copula (Gudendorf and Segers, 2010), as there are few data in this region of interest; while it has not been possible to properly model the 'angle' of the tail in preliminary frequentistic studies using the Galambos copula, this issue has to be further addressed.

Chapter 4

Data integration and evaluation

It is possible to assess effects of PFOA contamination in drinking water on perinatal parameters, in particular on the newborn children's weight, by integrating the exposure data modelled along space and time in Chapter 2 into appropriate models for birth data as developed in Chapter 3. Using the copula model, this analysis is restricted to data from the town of Arnsberg, which have mainly been worked with in Chapter 3, mostly for computational reasons and the town's distinctive role as the most affected place in NRW.

In this chapter, the result of this combined model is reported along with a number of preliminary considerations and preparation steps. The latter include the necessary spatial realignment of data, as the spatial level of the drinking water supply areas differs from that of the perinatal data (postal codes). Furthermore, some aspects on the assignment of individual exposure information as well as possible other spatial effects have to be considered.

Much of this work is related to the Arnsberg cohort introduced in Section 1.3. The participants are not included in the perinatal registry data set, as the children are born around the year 2000, and would not be identifiable anyway. But data from this biomonitoring study group serve to explore the relationships between external and internal exposure, and so to reason that usage of drinking water data is meaningful.

This chapter begins with reports on related sub-projects (Section 4.1). One of these is to explore the relevance of spatial relationships within the perinatal data set, not considering the PFOA exposure; spatial autocorrelation of the mean birth weight as well as possible influences of urbanity and air pollution on it are estimated (Section 4.1.1). Furthermore, data from the Arnsberg cohort and its control groups are used to explore the temporal progression of PFOA concentrations in blood serum (i.e., the internal exposure) after the end of the major contamination; influences of consumption behaviour, estimated external exposure and biometric covariates on the internal exposure are also estimated (Section 4.1.2).

Section 4.2 explains several steps of the assignment of individually modelled PFOA exposure values to residents. Starting with the realignment issue, common solutions from the literature are briefly reviewed and their applicability to the given situation is evaluated in Section 4.2.1. The ultimately preferred simpler approach using population density data is outlined in Section 4.2.2. Section 4.2.3 focuses on the town of Arnsberg as an example to explain the spatial structure of the drinking water supply and the deduction of local contamination values on postal code level. The internal exposure is addressed in Section 4.2.4, where the spatio-temporal predictions from Chapter 2 are used in a pharmacokinetic model. The perinatal data from Arnsberg, combined with modelled drinking water data, are evaluated in Section 4.3, followed by some concluding remarks in Section 4.4.

4.1 Preliminary and related studies

4.1.1 Urbanity, air pollution and birth weight in NRW

Within the 'PerSpat' project, Ganme (2017) has worked on geographical structures in the NRW registry data regarding birth weight and on associations between birth weight, air pollution and urbanity, referring to Romão *et al.* (2013) and Basu *et al.* (2014), among others:

The birth weight data have been spatially aggregated by averaging, first on their natural level of postal codes, later on the level of cities and municipalities to match the air pollution data. The population density in the areal units has been used as the measure of urbanity. A spatial autocorrelation of mean birth weight (i.e., correlation between neighbouring postal codes or municipalities) has been confirmed by spatial autoregression models. Significant effects of the spatially varying factors of population density and various air pollutants on mean birth weight have been found, but also collinearity between the factors.

To further explore the spatial structure of the perinatal data, Kohlenbach (2019) has applied a regression model on individual level and afterwards aggregated the residuals on postal code level. There is a correlation between these mean residuals and the area sizes. As smaller postal code areas are usually found in urban regions, this is another hint that urbanity correlates with birth weight.

Perspectively, when spatially analysing perinatal data, a spatial structure of some individual covariates should be taken into account, usually as random effect in a mixed model. E.g., 'ethnicity' (unlike nationality) as found in US studies would be subject to spatial analyses, since the 'ethnic' composition of the population may vary within a state, especially between urban and rural regions.

4.1.2 Studies on the Arnsberg cohort

Beside the major works by Hölzer *et al.* (2008); Hölzer *et al.* (2009); Hölzer *et al.* (2011); Bacher (2020), additional smaller studies on the human biomonitoring cohort from Arnsberg, NRW, and its control groups (cf., Section 1.3) have been conducted within the 'PerSpat' project. Graphical and regression analyses on the temporal progression of PFOA concentrations in blood (internal exposure) include data on individual tap water consumption behaviour, among others, and estimations of the local water contamination based on models from Chapter 2.

Overall, a decrease of the internal exposure is clearly visible for both adults and children from Arnsberg. When logarithmic PFOA concentrations are considered, this trend is linear for several years from 2006 on. This corresponds to an exponential decay, such that the decrease is proportional to the current value, which allows a reliable estimation of the biological half-life. Measurements from successive years are strongly correlated, but this correlation becomes weaker with an increasing time lag. Including data from the 2017 follow-up, the decrease becomes slightly weaker in later years, especially for the now adolescent girls. For the latter, the trend is also observed along the individual age instead of time.

A visible, but very weak linear decrease of logarithmic PFOA concentrations is also found for the control groups, which have not been affected by specific environmental contamination events with PFASs, but are exposed at background level. Anglers fishing from the contaminated lake Möhne are grouped by their internal exposure; those who report to eat fish several times a week are over-represented in the high level groups. For mothers from Arnsberg who report to have given birth to another child, the internal exposure visibly decreases more in the respective year, albeit very weakly.

In cross-sectional regression models for adults from Arnsberg for several years, significant influences of age, sex and body mass index (BMI) on the internal exposure to PFOA are found. Older and female participants are thus more exposed, while a higher BMI is associated with lower exposure. Sex and age also significantly influence the differences of PFOA concentrations between the first examination in 2006 and those 4 years later, where older and female participants show a slower decay in their internal exposure.

When the estimated external exposure via drinking water is taken into account, there is also a significantly positive effect of the presumed PFOA intake about 2006 on the blood concentrations in the course of time. In contrast, the later minor intakes have no effect.

The longitudinal analyses suffer from a rather large number of missing values, as many participants have not taken part in all follow-up examinations.

4.2 Assignment of individual exposure to PFOA

Daily estimations of the PFOA concentrations in drinking water are modelled on the spatial level of water supply stations and, afterwards, of so-called water supply areas (Chapter 2). However, these are still spatially misaligned to the perinatal registry data, which comprise of individual birth records but are spatially allocated on postal code level. Below, this issue is first addressed in general (Sections 4.2.1 and 4.2.2), and afterwards illustrated for the particular relevant town of Arnsberg (Section 4.2.3).

There also remains the question of appropriate temporal assignment, i.e., which point or period of time within the water contamination process can represent the individual exposure that is relevant for influencing birth weight. The assumed long biological half-life of PFOA is important here. Addressing this issue is one of the purposes of modelling the temporal progression of PFOA concentration in the human organism, as presented in Section 4.2.4.

4.2.1 Spatial realignment approaches

In the given situation, there are two spatial data sets of disparate resolution (misalignment, a block-to-block change-of-support-problem). The second division of space is not a refinement of the first, but can deviate from it in every respect (non-nested, cf. Mugglin *et al.*, 2000). Overviews of the situation and solution approaches can be found in, e.g., Gelfand *et al.* (2001) and Gotway and Young (2002).

A joint analysis of misaligned spatial structures is also possible by transition to continuous models by infinitesimal refinement of the resolution. A Bayesian Poisson-Gamma point process model is developed by Wolpert and Ickstadt (1998). Best *et al.* (2000) base an epidemiological Poisson regression with individual and spatial covariates on this. Model comparisons with Poisson-Gamma and semi-parametric approaches can be found in Best *et al.* (2005) and Sturtz and Ickstadt (2014). However, these and other epidemiological methods are usually restricted to count data. While these have been extensively researched on, especially concerning disease frequencies, Gaussian and other continuous responses, as needed in the 'PerSpat' project, are comparatively rare in spatial epidemiology. Within the 'PerSpat' project, Goeken *et al.* (2013) have fitted Markov random field models under certain restrictions, which do not allow for individual covariates, unlike point process models.

Various approaches to handle spatially misaligned data sets are considered. Kohlenbach (2019) has explored their applicability to the situation within the 'PerSpat' project:

The general change-of-support-problem is outlined by Gelfand *et al.* (2001) and there addressed by Bayesian Kriging. The approach is extended to hierarchical regression by Zhu *et al.* (2003). Because the data cannot be considered as continuous in space (at least for drinking water), as required, this approach turns out not to be directly applicable.

Estimation of spatial processes to solve the problem by means of geostatistics is described by Gotway and Young (2007), specialisations by Krivoruchko *et al.* (2011). This approach is generally applicable to model one of the data sets and so to obtain predictions at the spatial level of the other. However, the estimation of the required correlation matrix has been practically difficult and resulted in extremely varying and even negative predictions of PFOA concentrations. For the moment, a fixed correlation matrix of the water supply areas based on river dependency and the estimated station-to-area-supply proportions has been used to apply the algorithm.

Mugglin *et al.* (2000) consider all intersections ('atoms') of the areal units of both data sets and estimate the variables' distribution parameters on atom-level in Bayesian hierarchical models. The approach is adapted to a spatial grid structure by Agarwal *et al.* (2002). As it is only elaborated for count data, Kohlenbach (2019) considers the number of births with low birth weight per areal unit, together with the local PFOA concentration and covariates from the perinatal registry data, aggregated in the areal units. However, the MCMC evaluation did not produce stable, plausible estimates, perhaps due to a violated Poisson assumption.

4.2.2 Realignment using population density

An individual, simpler realignment approach is ultimately preferred over those from Section 4.2.1. As reported in Kolbe *et al.* (2019a), a postal area's PFOA concentration is averaged from the respective overlapping water supply areas, weighted by population density data from the German national census of 2011 (Statistische Ämter des Bundes und der Länder, 2015):

The population density is given on a grid of 100 m × 100 m cells by the number of inhabitants per cell. This resolution is fine enough compared to both postal code and drinking water supply areas and also compared to their intersections ('atoms' as considered by Mugglin *et al.*, 2000).

The PFOA concentration is known for any atom, and so is the number of inhabitants per atom. Using these numbers, a postal area can be interpreted as a weighted sum of its atoms, and its PFOS concentration calculated as the respective weighted mean.

In this way, the estimation of the external exposure to PFOA for the population of a postal code area is not biased, when a sparsely populated part is supplied with water with a different PFOA concentration than in the main part. Since no more precise location information than the postal code is known for the data of the perinatal registry, this approach is the best possible and correct on average.

4.2.3 Arnsberg contamination assessment

The town of Arnsberg is of particular interest due to comparably high PFOA contamination of drinking water resources, namely the rivers Möhne and Ruhr, from about 2003 until about 2007 (Skutlarek *et al.*, 2006; Hölzer *et al.*, 2008). As consumption of contaminated drinking water is supposed to account for the major part of the human external exposure (Section 1.1 and references therein), it is reasonable to use PFOA concentration in drinking water as a surrogate marker, at least as a local average. However, the individual *in*ternal exposure (PFOA concentration in blood serum) may not be well represented by these data without information on individual water consumption, time of residence and additional biometric data; these aspects are elaborated in another sub-project within 'PerSpat' briefly reported in Section 4.2.4.

Estimated PFOA concentrations in the Arnsberg drinking water are included as an additional explanatory variable to the linear predictors of the regression models selected in Chapter 3. They are spatio-temporally assigned to the perinatal data by postal code (Section 4.2.2) and day of birth. Regarding the latter, a time-lag of one year is employed (Section 4.2.4), such that, e.g., the local drinking water contamination of the future mother's postal code area on 15 April 2007, is used for a perinatal data observation from 15 April 2008.

Every water supply area within Arnsberg is supplied by a single station. Four of the five postal code areas are uniquely located within a water supply area. The westernmost postal code area comprises several districts and is divided between four water supply areas, so their values are averaged. With that, the lowest PFOA concentration is obtained in the eastern part of the town (supplied by a Ruhr-dependent station upstream of the Möhne mouth). In the western part (supplied from Möhne and Ruhr, but also from stations using non-contaminated groundwater), the average concentration in 2006 is about four times higher. Finally, in the central part (solely supplied from the Möhne), it is almost twice as high again. An example of this spatial exposure situation for Arnsberg can be found in Figure 4.1.

The earliest measurements at water supply stations are from the middle of 2006, after the contamination has been discovered and assumed to have started two to four years earlier (Skutlarek *et al.*, 2006; Hölzer *et al.*, 2008). Thus, the exact temporal progression of the PFOA concentrations in drinking water is not known, nor to be reliably modelled as in Section 2.4, for the interesting early period of time. For the moment, a linear increase is presumed, starting at the the limit of quantification of 10 ng/l in the middle of 2003 and ending at station-specific values obtained from the model results for the middle of 2006. This is based on the provisional presumption that the polluted soil conditioner had been steadily brought out in the course of the years. In fact, there are neither records on the discharge nor on the PFOA concentrations in the soil conditioner, and the process of the solution of PFOA in water and of the washout is more complex. For the later period of time from the middle of 2006 on, daily predictions from the station-specific models in Section 2.4 are used, which are significantly declining in the course of time.

4.2.4 Internal exposure modelling

The question whether and how to conclude from the PFOA contamination in the local drinking water (external exposure), via a certain amount of water intake by the residents, to the concentration of PFOA in blood serum (internal exposure) is an important step within the 'PerSpat' project: It has been addressed by Kolbe *et al.* (2019b) and continues to be worked on. Shin *et al.* (2011b) is the principal reference and an important example for modelling and verification of the internal exposure to PFOA following an environmental contamination incident.

The concentration in drinking water decreases within rather short periods of time, after the contamination process in the environment is interrupted (cf., data and model results in Chapter 2), and even more rapid after interventions like filtering measures (cf., Figure 2.4 on page 14). In contrast, the internal exposure is very stable, due to the long biological half-life, so it needs a long time for a decrease in the environment to take effect here.

The respective physiological process is represented by a simplifying one-compartment pharmacoki-

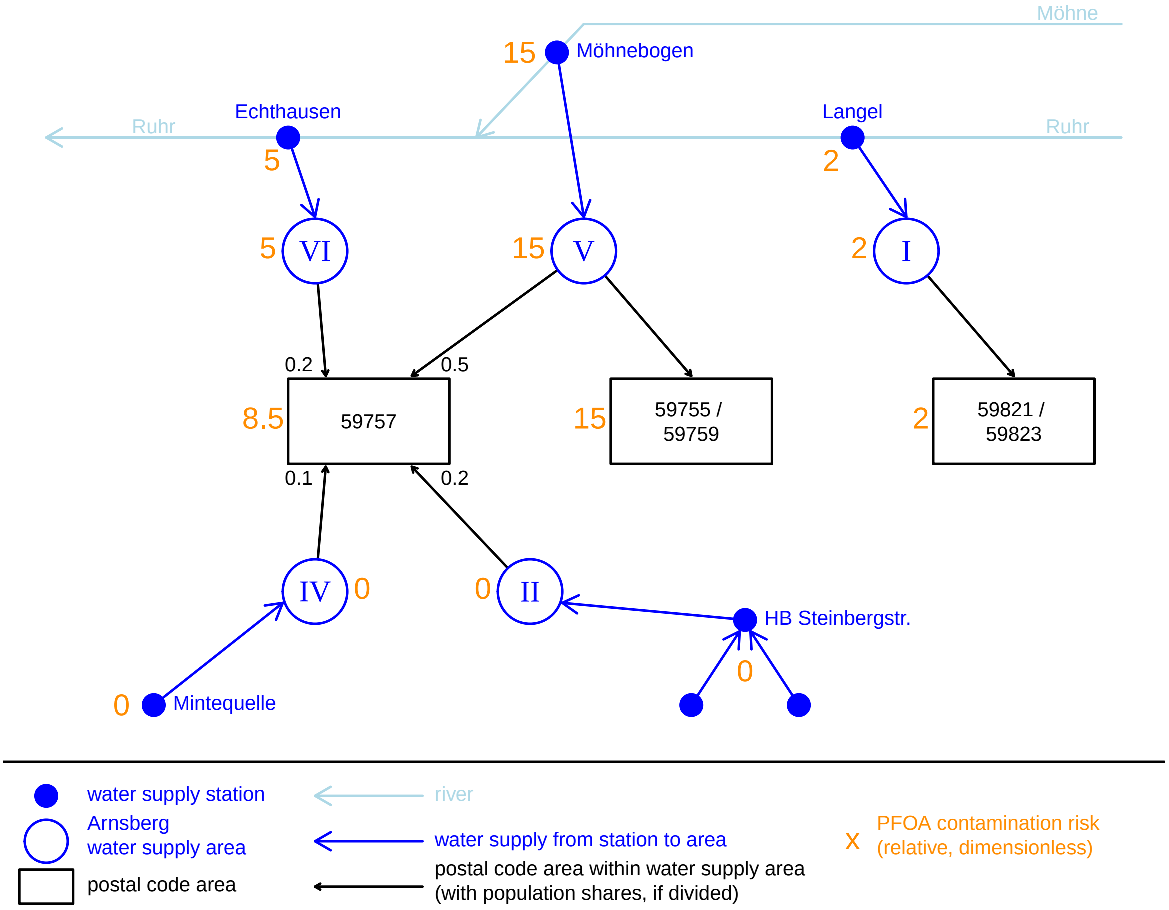

Figure 4.1: Schema of the water supply of the town of Arnsberg in the upper Ruhr region, NRW, Germany. The town is partly supplied from formerly contaminated rivers, its districts being affected to varying degrees. The contamination risk is given in relative, dimensionless units, roughly estimated from modelled data (Chapter 2) for the middle of the year 2006.

netic ADME model (absorption – distribution – metabolism – excretion). It depends on the individual body weight and age, as associations with the PFOA concentration in blood serum are found for both of them in the Arnsberg cohort study (Hölzer *et al.*, 2008; Hölzer *et al.*, 2009; see also Section 4.1.2). The absolute amount of PFOA in the body (here denoted by M_t for a point of time t) is the product of blood concentration, body weight and the volume of distribution, a physiological constant. Changes within the period from time t_0 to t_1 are modelled to be

$$M_{t_1} = M_{t_0} \cdot e^{-\kappa} + D_{t_1 - t_0} \left(\frac{1 - e^{-\kappa}}{\kappa} \right),$$

where κ is an elimination constant determined by the biological half-life. The PFOA intake $D_{t_1-t_0}$ during this period of time is the product of the concentration in water and the individual tap water consumption, to which a certain background exposure is added and all is downweighted by a bioavailability constant. Consumption is age-specifically presumed to be in accordance with the German national nutrition survey (Heuer *et al.*, 2015). The individual age is also needed for modelling the total intake caused by background exposure. In this context, a starting value M_0 has to be determined; it is presumed that there was no additional environmental contamination in NRW (Section 1.2) until 2002 (Skutlarek *et al.*, 2006; Hölzer *et al.*, 2008) and, thus, only background exposure has to be respected for this period. A special model is set up for children, where the body weight is presumed to increase over time.

The local PFOA concentration in water is estimated by spatial realignment as described in Section 4.2.2. Thereby, the water supply areas' concentrations have been deduced from those of the respective water supply stations using the weight matrix from Section 2.5. The stations' values are taken from the models in Chapter 2. All this is modelled per day.

The model is verified using blood sample data from mothers in the Arnsberg cohort and the control group, with 150 participants each and several observations in the course of the years. The predictions from the model, using individual body weight and age, generally agree well with the observations on average, although predictions are slightly too high for some age groups. In the model result, the affected group from Arnsberg and the control group are well distinguishable. However, the observations feature a much wider variance than the modelled values, presumably due to unknown or unconsidered individual effects. Individual tap water consumption is not respected, as the corresponding survey data are not thoroughly complete, accurate and comparable.

To conclude, for persons with a high external exposure at the beginning, later minor concentration peaks in water as well as later periods of lower contamination are hardly relevant for the progression of the internal exposure. Furthermore, a time-lag between the peaks of external and internal exposure of a few months up to a year is recognisable. Therefore, the PFOA concentration in the local drinking water one year before a birth event is considered relevant for possible influences on perinatal parameters. Perspectively, the total external exposure over a longer period of time is equally reasonable.

The model verification is limited to data from younger female participants, but these are of principal interest for the 'PerSpat' project. The assessment of water contamination prior to 2006 is still problematic; for the moment, a linear increase is presumed (cf., Section 4.2.3), as the polluted soil conditioner had been steadily brought out in the course of the years. Physiological constants have to be fixed as model parameters; in particular, various values of the volume of distribution (Thompson *et al.*, 2010; Shin *et al.*, 2011b) and of the biological half-life (Table 4.1) are plausible and have been reported; the rather short half-life of 2.3 years deduced by Bartell *et al.* (2010) leads to predictions which fit best to the cohort data.

4.3 Effect evaluation for Arnsberg perinatal data

With regard to a possible effect of PFOA, a first and preliminary substantive evaluation of the models developed in Chapter 3 is conducted.

The 4451 complete observations from the perinatal registry data are considered, where the postal code of the mother's residence is from Arnsberg (Section 3.1). As no individual information on tap water consumption or a more precise place of residence is given, the spatio-temporal assignment of average PFOA contamination outlined in Section 4.2.3 is employed and the respective values are included per postal code and day of birth as an additional explanatory variable to the perinatal data.

Applying both the standard polynomial and the distributional copula regression, no significant linear effect of the modelled PFOA concentration is found with either.

A simplified representation of this temporally structured covariate is also experimentally included

Table 4.1: Biological half-life estimates for PFOA in human organism.

Source	Study group	Sample size	Median (years)	95% Confidence interval
Olsen *et al.* (2007)	retired production workers (2 female)	26	3.4	[3.0, 4.1]
Brede *et al.* (2010)	Arnsberg cohort	65	3.3	
Bartell *et al.* (2010)	exposed adults (C8 health project)	200	2.3	[2.1, 2.4]
Seals *et al.* (2011)	highly exposed adults (C8 health project)	602	2.9	[2.3, 3.8]
Seals *et al.* (2011)	less exposed adults (C8 health project)	971	8.5	[7.1, 10.1]
Zhang *et al.* (2013)	young females	20	1.8	stand. err.: 0.3
Zhang *et al.* (2013)	males and older females	66	1.7	stand. err.: 0.4
Russell *et al.* (2015)	general population (modelled)		2.4	
Li *et al.* (2018)	formerly exposed children and adults	106	2.7	[2.5, 2.9]

to the models, as a first exploration of possible temporal effects: A dummy variable states whether an observation is from the years of 2004 until 2007 (with regard to the time-lagged PFOA variable) and, thus, from within the period of presumed very high PFOA concentrations in drinking water, but with no significant effect resulting.

If, on the other hand, the linear time itself over the entire observation period is included as a covariate, it shows a significant, but extremely weak effect on the mean parameter of birth weight, which decreases in the course of the years.

These and other aspects are currently used for further developments within the ‘PerSpat’ project to better represent the internal exposure and thereby gain more reliable results for a PFOA effect (cf., Sections 4.2.4 and 4.4).

4.4 Discussion and perspectives

A possible effect of PFOA on perinatal variables is supposed to be weak and difficult to measure, compared to influences of other environmental exposures such as maternal smoking (Section 3.6.3) or air pollution (Section 4.1.1). Therefore, and as there remains much noise in the perinatal data after either model fit, it is not surprising not to find an effect of PFOA.

Furthermore, there are complex relationships and uncertainties on the path between drinking water contamination and internal exposure. Shin *et al.* (2011b) give an important example of modelling and verification of the internal exposure to PFOA, using their assessment of the local drinking water contamination (Shin *et al.*, 2011a). For the ‘PerSpat’ project, similar considerations, pharmacokinetic modelling and comparisons using the Arnsberg cohort data are briefly reported in Section 4.2.4. However, important information on an individual level needed for that are not given in the perinatal data.

The other way around, data from the children in the Arnsberg cohort can be analysed with models from Chapter 3, as far as possible with the smaller number of observations. Here, the individual internal

PFOA burden is known, but perinatal data, especially birth weight and gestational age, would have to be collected retrospectively. However, these children are born around the year 2000 and so are presumably not affected by a major contamination event before or during pregnancy. The focus of analyses in this cohort is mainly on possible adverse effects on the later development (Section 1.3).

PFOA's long biological half-life of several years is important for the reason of the study, but estimations vary (Table 4.1), such that an important model assumption is uncertain. This makes it difficult to decide on a point or period of time, where concentrations in water correspond to the individual burden during pregnancy (cf., Section 4.2.4), even when equal and constant tap water consumption is assumed. In addition, for the whole of NRW, the mothers' consumptions have not been recorded in the perinatal data set.

To deeper explain perinatal data, more complex generalised additive models, especially using splines, could be considered, where non-linear relationships are possible (Fahrmeir *et al.*, 2004). This holds for the temporal modelling of PFOA concentrations, but also for the spatial dimension in future studies when larger regions are considered. There, further information such as neighbourhood of postal codes or water supply relationships could be used. Since lower birth weights are observed in some urban regions, an according spatial dependence structure can also be included (Section 4.1.1). This and other model enhancements may make it possible to more easily detect even very weak effects.

Chapter 5

Reflections

This study consists of two main parts of an epidemiological secondary data analysis, presented in Chapters 2 and 3. The first part deals with modelling the environmental factor of perfluorooctanoic acid (PFOA) concentration in drinking water for assessment of the external exposure. The second part consists of a regression analysis of perinatal data, where the birth weight as the outcome of interest and a number of important other explanatory variables are included. Both parts are of interest as statistical analyses of complex data structures, both enable important insights to substantive matters and allow for future applications in other fields with similar data situations.

The original question and ultimate goal of the 'PerSpat' research project is the potential effect of PFOA and other per- and polyfluoroalkyl substances on foetal development. However, this fades into the background for the time being, as the two main issues of this study have to be addressed first. Later, the original question becomes an application of the developed models (Section 4.3). Moreover, the reported negative result is restricted to one location, lacks some deeper biological considerations on PFOA pathways, and is, therefore, somewhat preliminary within 'PerSpat'. To complete the approach and to strengthen these aspects, related research is currently conducted in another sub-project with a more medical point of view. Starting with Kolbe *et al.* (2016) and Kolbe *et al.* (2019b), computationally easier regression analyses of the whole perinatal registry data are performed using the PFOA exposure predictions obtained above.

Given the challenging data situation and the fact that models are considered or developed in parts of the analyses, which are rather complicated in terms of structure or computational effort, the results are surprisingly simple. It becomes apparent once again that highly refined models need to be matched by equally sophisticated data quality with complete background information.

For instance, when modelling PFOA concentrations jointly from all water supply stations along the river Ruhr, the relevance of the river's segments is obvious, but more local trends cannot be distinguished. Individual effects of stations may depend on properties such as partial groundwater usage, of which no precise information is available. The variability of individual measurements outweighs such random effects. As a result from such and other imponderables like estimated area supply proportions, risk assessment for PFOA in drinking water may be reliable on average but is still subject to uncertainty. However, the developed approaches are transferable to nearly any persistent substance solute in surface waters.

Remaining noise is also found when modelling the perinatal registry data, mainly caused by natural fluctuations, since a highly variable and individual process is observed. Despite of this, data are accurate enough to gain relevant insights, which are of interest for gynaecological and obstetric research in general. Joint modelling of birth weight and gestational age as a bivariate outcome is confirmed to be important to distinguish effects on both variables, on each of them separately, and between them, in particular with regard to explanatory variables like Caesarean section or maternal smoking. In this situation, Bayesian distributional regression is very useful to reveal structures in the data beyond effects on the means.

However, the applied copula model in distributional regression is computationally demanding, i.e., long-running and in some cases unstable. The estimability of the conditional dependence between the two outcomes is technically limited. Analyses have to be restricted to a small subset of the perinatal registry data; this is contrary to the purpose of a large-scale secondary data analysis, which results in a dilemma between data usage and applicability of models. On the other hand, too small an amount of data may also lead to a lack of computability. For the time being, while the complex model proves somewhat unsuitable for practical daily application, it is more efficient to focus on a simpler approach, like the applied polynomial regression. An easier representation of dependencies is another option. For future applications to the large data set, the technical realisation of the copula distributional regression will be adapted accordingly. In any case, allowance for non-linearity and asymmetry should be preserved.

All this renders it very difficult to determine a possible effect of PFOA on perinatal parameters from the given material in a serious manner, the more as the effect is presumed to be comparably weak. However, the considerations and developments on the path to this goal prove useful in their own right, for other applications and for gaining substantive knowledge.

References

Acar, E. F., Craiu, R. V., and Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, **7**, 2822–2850.

Agarwal, D. K., Gelfand, A. E., and Silander Jr., J. A. (2002). Investigating tropical deforestation using two-stage spatially misaligned regression models. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**(3), 420–439.

Apelberg, B. J., Witter, F. R., Herbstman, J. B., Calafat, A. M., Halden, R. U., Needham, L. L., and Goldman, L. R. (2007). Cord serum concentrations of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in relation to weight and size at birth. *Environmental Health Perspectives*, **115**(11), 1670–1676.

Armbruster, D. A. and Pry, T. (2008). Limit of blank, limit of detection and limit of quantitation. *The Clinical Biochemist Reviews*, **29**(Suppl. 1), S49–S52.

ATSDR (2018). Toxicological profile for perfluoroalkyls – draft for public comment. Technical report, Agency for Toxic Substances and Disease Registry.

Ayub, R., Messier, K. P., Serre, M. L., and Mahinthakumar, K. (2019). Non-point source evaluation of groundwater nitrate contamination from agriculture under geologic uncertainty. *Stochastic Environmental Research and Risk Assessment*, **33**(4), 939–956.

Bach, C. C., Bech, B. H., Nohr, E. A., Olsen, J., Matthiesen, N. B., Bonefeld-Jørgensen, E. C., Bossi, R., and Henriksen, T. B. (2016). Perfluoroalkyl acids in maternal serum and indices of fetal growth: the Aarhus birth cohort. *Environmental Health Perspectives*, **124**(6), 848–854.

Bacher, S. (2020). *Follow-Up-Untersuchung 2017 der inneren Belastung von Mutter-Kind-Paaren und Männern in Gebieten erhöhter Trinkwasserbelastung mit perfluorierten Verbindungen*. Dissertation, Ruhr-University Bochum.

Bartell, S. M., Calafat, A. M., Lyu, C., Kato, K., Ryan, P. B., and Steenland, K. (2010). Rate of decline in serum PFOA concentrations after granular activated carbon filtration at two public water systems in ohio and west virginia. *Environmental Health Perspectives*, **118**(2), 222–228.

Basu, R., Harris, M., Sie, L., Malig, B., Broadwin, R., and Green, R. (2014). Effects of fine particulate matter and its constituents on low birth weight among full-term infants in California. *Environmental Research*, **128**, 42–51.

REFERENCES

Bates, D. and Mächler, M. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-18.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48.

Becker, E. (2016). *Modellierung der PFT-Trinkwasserdaten der Wasserwerke an der Ruhr und Möhne*. Internship report, TU Dortmund University.

Becker, E. (2017). *Räumlich-zeitliche Modellierung der PFAS-Exposition über das Trinkwasser*. Master's thesis, TU Dortmund University.

Beelaerts, V., Bauwens, M., and Pintelon, R. (2012). Time series reconstruction from unequally spaced natural archive data. *Mathematical Geosciences*, **44**(3), 283–307.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2020). *BayesX - software for Bayesian inference in structured additive regression models*. `http://www.bayesx.org`.

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., and Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, **17**(22), 2537–2550.

Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, **95**(452), 1076–1088.

Best, N. G., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**(1), 35–59.

Bivand, R., Keitt, T., and Rowlingson, B. (2019). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-8.

Braun, T. M. (2002). The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials*, **23**(3), 240–256.

Brede, E., Wilhelm, M., Göen, T., Müller, J., Rauchfuss, K., Kraft, M., and Hölzer, J. (2010). Two-year follow-up biomonitoring pilot study of residents' and controls' PFC plasma levels after PFOA reduction in public water system in Arnsberg, Germany. *International Journal of Hygiene and Environmental Health*, **213**(3), 217–223.

Brunner, M. I., Furrer, R., Sikorska, A. E., Viviroli, D., Seibert, J., and Favre, A.-C. (2018). Synthetic design hydrographs for ungauged catchments: a comparison of regionalization methods. *Stochastic Environmental Research and Risk Assessment*, **32**(7), 1993–2023.

Buck, R. C., Franklin, J., Berger, U., Conder, J. M., Cousins, I. T., de Voogt, P., Jensen, A. A., Kannan, K., Mabury, S. A., and van Leeuwen, S. P. J. (2011). Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integrated Environmental Assessment and Management*, **7**(4), 513–541.

Chen, W.-C. and Maitra, R. (2015). *EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution*. R package version 0.2-5.

Chen, X. and Yin, X. (2019). *NlcOptim: Solve Nonlinear Optimization with Nonlinear Constraints*. R package version 0.6.

Cooke, R. M., Joe, H., and Chang, B. (2020). Vine copula regression for observational studies. *AStA – Advances in Statistical Analysis*, **104**, 141–167.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**(2), 127–150.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems, 1695.

Currie, L. A. (2004). Detection and quantification limits: basic concepts, international harmonization, and outstanding ("low-level") issues. *Applied Radiation and Isotopes*, **61**, 145–149.

Cutler, D. J. (1978). Numerical deconvolution by least squares: Use of prescribed input functions. *Journal of Pharmacokinetics and Biopharmaceutics*, **6**(3), 227–241.

Dalla Valle, L., Leisen, F., and Rossini, L. (2018). Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society – Series C: Applied Statistics*, **67**(3), 523–548.

de Carvalho, M. (2018). *spearmanCI: Jackknife Euclidean / empirical likelihood inference for Spearman's rho*. R package version 1.0.

de Oliveira, V. (2005). Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics*, **14**(1), 95–115.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**(3), 236–244.

EFSA (2008). Perfluorooctane sulfonate (PFOS), perfluorooctanoic acid (PFOA) and their salts scientific opinion of the panel on contaminants in the food chain. *EFSA Journal*, **653**, 1–131.

EFSA CONTAM (2018). Risk to human health related to the presence of perfluorooctane sulfonic acid and perfluorooctanoic acid in food. *EFSA Journal*, **16**(12), 1–295.

EFSA CONTAM (2020). Risk to human health related to the presence of perfluoroalkyl substances in food. *EFSA Journal*, **18**(9), 1–391.

EPA (2016). Health effects support document for perfluorooctanoic acid (PFOA). Technical Report 822-R-16-003, U.S. Environmental Protection Agency.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, **14**(3), 731–761.

Fang, F., Stratton, H., and Gage, T. B. (2007). Multiple mortality optima due to heterogeneity in the birth cohort: A continuous model of birth weight by gestational age-specific infant mortality. *American Journal of Human Biology*, **19**(4), 475–486.

Fei, C., McLaughlin, J. K., Tarone, R. E., and Olsen, J. (2007). Perfluorinated chemicals and fetal growth: A study within the danish national birth cohort. *Environmental Health Perspectives*, **115**(11), 1677–1682.

Frederick, I. O., Williams, M. A., Sales, A. E., Martin, D. P., and Killien, M. (2008). Pre-pregnancy body mass index, gestational weight gain, and other maternal characteristics in relation to infant birth weight. *Maternal and Child Health Journal*, **12**, 557–567.

Frisbee, S. J., Shankar, A., Knox, S. S., Steenland, K., Savitz, D. A., Fletcher, T., and Ducatman, A. M. (2010). Perfluorooctanoic acid, perfluorooctanesulfonate, and serum lipids in children and adolescents: results from the C8 health project. *Archives of Pediatrics & Adolescent Medicine*, **164**(9), 860–869.

Fromme, H., Tittlemier, S. A., Völkel, W., Wilhelm, M., and Twardella, D. (2009). Perfluorinated compounds – exposure assessment for the general population in western countries. *International Journal of Hygiene and Environmental Health*, **212**(3), 239–270.

Gage, T. B. (2003). Classification of births by birth weight and gestational age: an application of multivariate mixture models. *Annals of Human Biology*, **30**(5), 589–604.

Gage, T. B., Fang, F., and Stratton, H. (2008). Modeling the pediatric paradox: Birth weight by gestational age. *Biodemography and Social Biology*, **54**(1), 95–112.

Ganme, S. T. (2017). *Geographical Structures in Birth Data*. Master's thesis, TU Dortmund University.

Gardosi, J., Mongelli, M., Wilcox, M., and Chang, A. (1995). An adjustable fetal weight standard. *Ultrasound in Obstetrics & Gynecology*, **6**, 168–174.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**(1), 11–25.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, **2**(1), 31–45.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

Goeken, N., Ickstadt, K., Schäfer, M., Bücker-Nott, H.-J., Wilhelm, M., and Hölzer, J. (2013). Statistical approaches to evaluating the association between mother's exposure to perfluoroalkyl substances (PFASs) and birth outcome in North Rhine-Westphalia. In ISES and ISIAQ, editors, *Abstracts of the 2013 Conference of the International Society of Environmental Epidemiology (ISEE)*, pages 5030, P–2–14–01.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**(458), 632–648.

Gotway, C. A. and Young, L. J. (2007). A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, **16**(1), 115–135.

Grandjean, P., Andersen, E. W., Budtz-Jørgensen, E., Nielsen, F., Mølbak, K., Weihe, P., and Heilmann, C. (2012). Serum vaccine antibody concentrations in children exposed to perfluorinated compounds. *Journal of the American Medical Association*, **307**(4), 391–397.

Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2019). Bayesian functional linear regression with sparse step functions. *Bayesian Analysis*, **14**(1), 111–135.

Gudendorf, G. and Segers, J. (2010). *Copula Theory and Its Applications*, chapter 'Extreme-Value Copulas', pages 127–145. Springer, Berlin.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, **65**, 2434–2439.

Heuer, T., Krems, C., Moon, K., Brombach, C., and Hoffmann, I. (2015). Food consumption of adults in Germany: results of the German National Nutrition Survey II based on diet history interviews. *British Journal of Nutrition*, **113**(10), 1603–1614.

Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2020). *copula: Multivariate dependence with copulas*. R package version 1.0-0.

Hölzer, J., Midasch, O., Rauchfuss, K., Kraft, M., Reupert, R., Angerer, J., Kleeschulte, P., Marschall, N., and Wilhelm, M. (2008). Biomonitoring of perfluorinated compounds in children and adults exposed to perfluorooctanoate-contaminated drinking water. *Environmental Health Perspectives*, **116**(5), 651–657.

Hölzer, J., Göen, T., Rauchfuss, K., Kraft, M., Angerer, J., Kleeschulte, P., and Wilhelm, M. (2009). One-year follow-up of perfluorinated compounds in plasma of German residents from Arnsberg formerly exposed to PFOA-contaminated drinking water. *International Journal of Hygiene and Environmental Health*, **212**(5), 499–504.

Hölzer, J., Göen, T., Just, P., Reupert, R., Rauchfuss, K., Kraft, M., Müller, J., and Wilhelm, M. (2011). Perfluorinated compounds in fish and blood of anglers at Lake Mohne, Sauerland area, Germany. *Environmental Science & Technology*, **45**(19), 8046–8052.

Johnson, P. I., Sutton, P., Atchley, D. S., Koustas, E., Lam, J., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide – evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, **122**(10), 1028–1039.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, **90**(12), 1–37.

Keefe, M. J., Ferreira, M. A. R., and Franck, C. T. (2019). Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors. *Bayesian Analysis*, **14**(1), 181–209.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. *Statistics and Computing*, **26**(4), 841–860.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Annals of Applied Statistics*, **9**(2), 1024–1052.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S., and McGovern, M. E. (2019). Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, **38**(3), 413–436.

Knotters, M., Brus, D. J., and Oude Voshaar, J. H. (1995). A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, **67**, 227–246.

Ko, T.-J., Tsai, L.-Y., Chu, L.-C., Yeh, S.-J., Leung, C., Chen, C.-Y., Chou, H.-C., Tsao, P.-N., Chen, P.-C., and Hsieh, W.-S. (2014). Parental smoking during pregnancy and its association with low birth weight, small for gestational age, and preterm birth offspring: A birth cohort study. *Pediatrics & Neonatology*, **55**(1), 20–27.

Kohlenbach, J. (2019). *Methodenvergleich zum Realignment räumlicher Daten am Beispiel von Perinataldaten und Trinkwasserbelastung*. Master's thesis, TU Dortmund University.

Kolbe, A., Rathjens, J., Becker, E., Bücker-Nott, H.-J., Olthoff, K., Wilhelm, M., Ickstadt, K., and Hölzer, J. (2016). Exposure to PFOA and birth outcome in North Rhine-Westphalia, Germany. *Environmental Health Perspectives*, **ISEE**, P2–193.

Kolbe, A., Rathjens, J., Becker, E., Olthoff, K., Bergmann, S., Bücker-Nott, H.-J., Ickstadt, K., and Hölzer, J. (2019a). The PerSpat-project: Integration of national census data for spatial alignment of birth registry and drinking water data from the Ruhr region. *Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*. DocAbstr. 226.

Kolbe, A., Rathjens, J., Becker, E., Bücker-Nott, H.-J., Ickstadt, K., and Hölzer, J. (2019b). PFOA exposure assessment in North Rhine-Westphalia, Germany, linking birth registry data with tap water concentrations. *Environmental Epidemiology*, **3**(ISEE), 212.

Koustas, E., Lam, J., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide – evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, **122**(10), 1015–1027.

Krämer, N., Brechmann, E. C., Silvestrini, D., and Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, **53**(3), 829–839.

Kraus, D. and Czado, C. (2017). D-Vine copula based quantile regression. *Computational Statistics & Data Analysis*, **110**, 1–18.

Krivoruchko, K., Gribov, A., and Krause, E. (2011). Multivariate areal interpolation for continuous and count data. *Procedia Environmental Sciences*, **3**, 14–19.

Kyrklund-Blomberg, N. B. and Cnattingius, S. (1998). Preterm birth and maternal smoking: Risks related to gestational age and onset of delivery. *American Journal of Obstetrics and Gynecology*, **179**(4), 1051–1055.

Laaha, G., Skøien, J. O., and Blöschl, G. (2014). Spatial prediction on river networks: comparison of top-kriging with regional regression. *Hydrological Processes*, **28**, 315–324.

Lam, J., Koustas, E., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide – evidence-based medicine meets environmental health: Integration of animal and human evidence for PFOA effects on fetal growth. *Environmental Health Perspectives*, **122**(10), 1040–1051.

LANUV (2011). Verbreitung von PFT in der Umwelt: Ursachen – Untersuchungsstrategie – Ergebnisse – Maßnahmen. LANUV-Fachbericht 34, Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen.

Lau, C. (2012). Perfluorinated compounds. In A. Luch, editor, *Molecular, Clinical and Environmental Toxicology*, volume 101 of *Experientia Supplementum*, pages 47–86. Springer, Basel.

Lau, C., Butenhoff, J. L., and Rogers, J. M. (2004). The developmental toxicity of perfluoroalkyl acids and their derivatives. *Toxicology and Applied Pharmacology*, **198**(2), 231–241.

Lau, C., Anitole, K., Hodes, C., Lai, D., Pfahles-Hutchens, A., and Seed, J. (2007). Perfluoroalkyl acids: A review of monitoring and toxicological findings. *Toxicological Sciences*, **99**(2), 366–394.

Lee, L. and Helsel, D. R. (2005). Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. *Computers & Geosciences*, **31**, 1241–1248.

Li, C. Q., Windsor, R. A., Perkins, L., Goldenberg, R. L., and Lowe, J. B. (1993). The impact on infant birth weight and gestational age of cotinine-validated smoking reduction during pregnancy. *Journal of the American Medical Association*, **269**(12), 1519–1524.

Li, Y., Fletcher, T., Mucs, D., Scott, K., Lindh, C. H., Tallving, P., and Jakobsson, K. (2018). Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occupational and Environmental Medicine*, **75**(1), 46–51.

Looker, C., Luster, M. I., Calafat, A. M., Johnson, V. J., Burleson, G. R., Burleson, F. G., and Fletcher, T. (2014). Influenza vaccine response in adults exposed to perfluorooctanoate and perfluorooctanesulfonate. *Toxicological Sciences*, **138**(1), 76–88.

Lopez-Espinosa, M.-J., Fletcher, T., Armstrong, B., Genser, B., Dhatariya, K., Mondal, D., Ducatman, A., and Leonardi, G. (2011). Association of perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) with age of puberty among children living near a chemical plant. *Environmental Science & Technology*, **45**(19), 8160–8166.

Macon, M. B., Villanueva, L. R., Tatum-Gibbs, K., Zehr, R. D., Strynar, M. J., Stanko, J. P., White, S. S., Helfant, L., and Fenton, S. E. (2011). Prenatal perfluorooctanoic acid exposure in CD-1 mice: low-dose developmental effects and internal dosimetry. *Toxicological Sciences*, **122**(1), 134–145.

Maisonet, M., Terrell, M. L., McGeehin, M. A., Christensen, K. Y., Holmes, A., Calafat, A. M., and Marcus, M. (2012). Maternal concentrations of polyfluoroalkyl compounds during pregnancy and fetal and postnatal growth in British girls. *Environmental Health Perspectives*, **120**(10), 1432–1437.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, **95**(451), 877–887.

Müller, K. (2020). *Einflüsse perfluorierter Tenside auf Lipidblutkonzentrationen in einer Kohortenstudie*. Bachelor's thesis, TU Dortmund University.

Neelon, B., Gelfand, A. E., and Miranda, M. L. (2014). A multivariate spatial mixture model for areal data: Examining regional differences in standardized test scores. *Journal of the Royal Statistical Society – Series C: Applied Statistics*, **63**(5), 737–761.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, New York, 2nd edition.

Nie, L., Chu, H., Liu, C., Cole, S. R., Vexler, A., and Schisterman, E. F. (2010). Linear regression with an independent variable subject to a detection limit. *Epidemiology*, **21**(Suppl. 4), S17–S24.

Nieto-Barajas, L. E. and Sinha, T. (2015). Bayesian interpolation of unequally spaced time series. *Stochastic Environmental Research and Risk Assessment*, **29**(2), 577–587.

NJDEP (2019). Interim specific ground water criterion for perfluorooctanoic acid (PFOA, C8). Technical support document, New Jersey Department of Environmental Protection – Division of Science and Research.

Nolan, L. A., Nolan, J. M., Shofer, F. S., Rodway, N. V., and Emmett, E. A. (2009). The relationship between birth weight, gestational age and perfluorooctanoic acid (PFOA)-contaminated public drinking water. *Reproductive Toxicology*, **27**(3), 231–238.

NTP (2016). *Immunotoxicity Associated with Exposure to Perfluorooctanoic Acid or Perfluorooctane Sulfonate*. NTP Monograph. Office of Health Assessment and Translation – Division of the National Toxicology Program.

O'Donnell, D., Rushworth, A., Bowman, A. W., and Scott, E. M. (2014). Flexible regression models over river networks. *Journal of the Royal Statistical Society – Series C: Applied Statistics*, **63**(1), 47–63.

OECD, editor (2004). *Results of Survey on Production and Use of PFOS, PFAS and PFOA, Related Substances and Products/Mixtures Containing these Substances*, volume 19 of *OECD Environment, Health and Safety Publications – Series on Risk Management*. Inter-Organization Programme for the Sound Management of Chemicals, Environment Directorate – Organisation for Economic Co-Operation and Development.

Olsen, G. W., Burris, J. M., Ehresman, D. J., Froehlich, J. W., Seacat, A. M., Butenhoff, J. L., and Zobel, L. R. (2007). Half-life of serum elimination of perfluorooctanesulfonate, perfluorohexanesulfonate, and perfluorooctanoate in retired fluorochemical production workers. *Environmental Health Perspectives*, **115**(9), 1298–1305.

Otto, P., Schmid, W., and Garthoff, R. (2018). Generalised spatial and spatiotemporal autoregressive conditional heteroscedasticity. *Spatial Statistics*, **26**, 125–145.

Padoan, S. A. and Bevilacqua, M. (2015). Analysis of random fields using CompRandFld. *Journal of Statistical Software*, **63**(9), 1–27.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review*, **47**(2), 527–556.

Peterson, E. E., Theobald, D. M., and Ver Hoef, J. M. (2007). Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology*, **52**(2), 267–279.

Polakowski, L. L., Akinbami, L. J., and Mendola, P. (2009). Prenatal smoking cessation and the risk of delivering preterm and small-for-gestational-age newborns. *Obstetrics & Gynecology*, **114**(2), 318–325.

Pozza, L. E., Bishop, T. F. A., and Birch, G. F. (2019). Using bivariate linear mixed models to monitor the change in spatial distribution of heavy metals at the site of a historic landfill. *Environmental Monitoring and Assessment*, **191**, 472.

Quick, H., Banerjee, S., and Carlin, B. P. (2015). Bayesian modeling and analysis for gradients in spatiotemporal processes. *Biometrics*, **71**(3), 575–584.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rathbun, S. L. (2006). Spatial prediction with left-censored observations. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**(3), 317–336.

Rathjens, J., Kolbe, A., Hölzer, J., Ickstadt, K., and Klein, N. (2021a). Bivariate analysis of birth weight and gestational age depending on environmental exposures: Bayesian distributional regression with copulas. Pre-print, `https://arxiv.org/abs/2104.14243`.

Rathjens, J., Becker, E., Kolbe, A., Ickstadt, K., and Hölzer, J. (2021b). Spatial and temporal analyses of perfluorooctanoic acid in drinking water for external exposure assessment in the Ruhr metropolitan area, Germany. *Stochastic Environmental Research and Risk Assessment*, **35**(6), 1127–1143.

Reeske, A., Kutschmann, M., Razum, O., and Spallek, J. (2011). Stillbirth differences according to regions of origin: an analysis of the German perinatal database, 2004-2007. *BMC Pregnancy and Childbirth*, **11**(1), 63.

Regan, M. M. and Catalano, P. J. (1999). Bivariate dose-response modeling and risk estimation in developmental toxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**(3), 217–237.

Ribeiro Jr, P. J., Diggle, P. J., Schlather, M., Bivand, R., and Ripley, B. (2020). *geoR: Analysis of Geostatistical Data*. R package version 1.8-1.

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

Rigby, R. A. and Stasinopoulos, D. M. (2014). Automatic smoothing parameter selection in gamlss with an application to centile estimation. *Statistical Methods in Medical Research*, **23**(4), 318–332.

Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape – Using GAMLSS in R*. Chapman and Hall/CRC, Boca Raton.

Romão, R., Pereira, L. A. A., Saldiva, P. H. N., Pinheiro, P. M., Braga, A. L. F., and Martins, L. C. (2013). The relationship between low birth weight and exposure to inhalable particulate matter. *Cadernos de saude publica*, **29**, 1101–1108.

Roozbahani, A., Zahraie, B., and Tabesh, M. (2013). Integrated risk assessment of urban water supply systems from source to tap. *Stochastic Environmental Research and Risk Assessment*, **27**(4), 923–944.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-Building – A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester.

Russell, M. H., Waterland, R. L., and Wong, F. (2015). Calculation of chemical elimination half-life from blood with an ongoing exposure source: The example of perfluorooctanoic acid (PFOA). *Chemosphere*, **129**, 210–216.

Saito, H. and Goovaerts, P. (2000). Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site. *Environmental Science & Technology*, **34**(19), 4228–4235.

Salcedo, G. E., Porto, R. F., Roa, S. Y., and Momo, F. R. (2012). A wavelet-based time-varying autoregressive model for non-stationary and irregular time series. *Journal of Applied Statistics*, **39**(11), 2313–2325.

Salomon, L.-J., Bernard, J.-P., de Stavola, B., Kenward, M., and Ville, Y. (2007). Poids et taille de naissance: courbes et équations. *Journal de gynécologie obstétrique et biologie de la reproduction*, **36**(1), 50–56.

Saul, B. C., Hudgens, M. G., and Mallin, M. A. (2019). Downstream effects of upstream causes. *Journal of the American Statistical Association*, **114**(528), 1493–1504.

Schelin, L. and Sjöstedt-de Luna, S. (2014). Spatial prediction in the presence of left-censoring. *Computational Statistics and Data Analysis*, **74**, 125–141.

Schwartz, S. L., Gelfand, A. E., and Miranda, M. L. (2010). Joint Bayesian analysis of birthweight and censored gestational age using finite mixture models. *Statistics in Medicine*, **29**(16), 1710–1723.

Seals, R., Bartell, S. M., and Steenland, K. (2011). Accumulation and clearance of perfluorooctanoic acid (PFOA) in current and former residents of an exposed community. *Environmental Health Perspectives*, **119**(1), 119–124.

Shin, H.-M., Vieira, V. M., Ryan, P. B., Detwiler, R., Sanders, B., Steenland, K., and Bartell, S. M. (2011a). Environmental fate and transport modeling for perfluorooctanoic acid emitted from the Washington Works Facility in West Virginia. *Environmental Science & Technology*, **45**(4), 1435–1442.

Shin, H.-M., Vieira, V. M., Ryan, P. B., Steenland, K., and Bartell, S. M. (2011b). Retrospective exposure estimation and predicted versus observed serum perfluorooctanoic acid concentrations for participants in the C8 Health Project. *Environmental Health Perspectives*, **119**(12), 1760–1765.

Skjærven, R., Gjessing, H. K., and Bakketeig, L. S. (2000). Birthweight by gestational age in Norway. *Acta Obstetricia et Gynecologica Scandinavica*, **79**(6), 440–449.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

Skøien, J. O., Blöschl, G., and Western, A. W. (2003). Characteristic space scales and timescales in hydrology. *Water Resources Research*, **39**(10).

Skøien, J. O., Merz, R., and Blöschl, G. (2006). Top-kriging-geostatistics on stream networks. *Hydrology and Earth System Sciences*, **10**(2), 277–287.

Skøien, J. O., Blöschl, G., Laaha, G., Pebesma, E., Parajka, J., and Viglione, A. (2014). rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks. *Computers & Geosciences*, **67**, 180–190.

Skutlarek, D., Exner, M., and Farber, H. (2006). Perfluorinated surfactants in surface and drinking waters. *Environmental Science and Pollution Research International*, **13**(5), 299–307.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, **65**(1), 60–68.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society – Series B: Statistical Methodology*, **64**(4), 583–639.

Stan Development Team (2020a). *RStan: the R interface to Stan*. R package version 2.21.2.

Stan Development Team (2020b). *Stan Modeling Language Users Guide and Reference Manual*. Version 2.21.

Statistische Ämter des Bundes und der Länder (2015). *Zensus 2011 – Methoden und Verfahren*. Statistisches Bundesamt, Wiesbaden.

Steenland, K., Tinker, S., Frisbee, S., Ducatman, A., and Vaccarino, V. (2009). Association of perfluorooctanoic acid and perfluorooctane sulfonate with serum lipids among adults living near a chemical plant. *American Journal of Epidemiology*, **170**(10), 1268–1278.

Stotland, N. E., Hopkins, L. M., and Caughey, A. B. (2004). Gestational weight gain, macrosomia, and risk of cesarean birth in nondiabetic nulliparas. *Obstetrics & Gynecology*, **104**(4), 671–677.

Sturtz, S. and Ickstadt, K. (2014). Comparison of Bayesian methods for flexible modeling of spatial risk surfaces in disease mapping. *Biometrical Journal*, **56**(1), 5–22.

Sun, Q., Zong, G., Valvi, D., Nielsen, F., Coull, B., and Grandjean, P. (2018). Plasma concentrations of perfluoroalkyl substances and risk of type 2 diabetes: A prospective investigation among U.S. women. *Environmental Health Perspectives*, **126**(3), 037001.

Sun, W., Rachev, S., Stoyanov, S. V., and Fabozzi, F. J. (2008). Multivariate skewed Student's t copula in the analysis of nonlinear and asymmetric dependence in the German equity market. *Studies in Nonlinear Dynamics & Econometrics*, **12**(2).

Sy, M. M., Garcia-Hidalgo, E., Jung, C., Lindtner, O., von Goetz, N., and Greiner, M. (2020). Analysis of consumer behavior for the estimation of the exposure to chemicals in personal care products. *Food and Chemical Toxicology*, **140**(111320).

Tang, Y., Wang, H. J., Sun, Y., and Hering, A. S. (2019). Copula-based semiparametric models for spatiotemporal data. *Biometrics*, **75**(4), 1156–1167.

Tennekes, M. (2018). tmap: Thematic maps in R. *Journal of Statistical Software*, **84**(6), 1–39.

Thompson, J., Lorber, M., Toms, L.-M. L., Kato, K., Calafat, A. M., and Mueller, J. F. (2010). Use of simple pharmacokinetic modeling to characterize exposure of Australians to perfluorooctanoic acid and perfluorooctane sulfonic acid. *Environment International*, **36**(4), 390–397.

Thompson, J. M. D., Clark, P. M., Robinson, E., Becroft, D. M. O., Pattison, N. S., Glavish, N., Pryor, J. E., Rees, K., and Mitchell, E. A. (2001). Risk factors for small-for-gestational-age babies: The Auckland birthweight collaborative study. *Journal of Paediatrics and Child Health*, **37**(4), 369–375.

Trinkwasserkommission (2016). Fortschreibung der vorläufigen Bewertung von Per- und polyfluorierten Chemikalien (PFC) im Trinkwasser – Begründungen der vorgeschlagenen Werte im Einzelnen. Technical report, Trinkwasserkommission des Bundesministeriums für Gesundheit.

Trinkwasserkommission (2020). Empfehlung des Umweltbundesamtes: Umgang mit per- und polyfluorierten Alkylsubstanzen (PFAS) im Trinkwasser. Technical report, Trinkwasserkommission des Bundesministeriums für Gesundheit.

UBA HBM-Kommission (2018). Ableitung von HBM-I-Werten für Perfluoroktansäure (PFOA) und Perfluoroktansulfonsäure (PFOS) – Stellungnahme der Kommission "Humanbiomonitoring" des Umweltbundesamts. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, **61**(4), 878–885.

Vatter, T. and Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, **141**, 147–167.

Ver Hoef, J. M., Peterson, E., and Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics*, **13**(4), 449–464.

Vestergren, R. and Cousins, I. T. (2009). Tracking the pathways of human exposure to perfluorocarboxylates. *Environmental Science & Technology*, **43**(15), 5565–5575.

Voigt, M., Schneider, K. T. M., and Jährig, K. (1996). Analyse des Geburtengutes des Jahrgangs 1992 der Bundesrepublik Deutschland. *Geburtshilfe und Frauenheilkunde*, **56**, 550–558.

Wang, W. and Sun, Y. (2019). Penalized local polynomial regression for spatial data. *Biometrics*, **75**(4), 1179–1190.

Watanabe, S. (2010). Asymptotic equivalence of Bayesian cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Weiss, E., Krombholz, K., and Eichner, M. (2014). Fetal mortality at and beyond term in singleton pregnancies in Baden-Wuerttemberg/Germany 2004–2009. *Archives of Gynecology and Obstetrics*, **289**(1), 79–84.

White, P., Gelfand, A., and Utlaut, T. (2017). Prediction and model comparison for areal unit data. *Spatial Statistics*, **22**, 89–106.

Whitworth, K. W., Haug, L. S., Baird, D. D., Becher, G., Hoppin, J. A., Skjærven, R., Thomsen, C., Eggesbo, M., Travlos, G., Wilson, R., Cupul-Uicab, L. A., Brantsæter, A. L., and Longnecker, M. P. (2012). Perfluorinated compounds in relation to birth weight in the norwegian mother and child cohort study. *American Journal of Epidemiology*, **175**(12), 1209–1216.

Wilhelm, M., Kraft, M., Rauchfuss, K., and Hölzer, J. (2008). Assessment and management of the first German case of a contamination with perfluorinated compounds (PFC) in the region Sauerland, North Rhine-Westphalia. *Journal of Toxicology and Environmental Health – Part A*, **71**(11–12), 725–733.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**(2), 251–267.

Wood, S. N. (2017). *Generalized Additive Models – An Introduction with R*. Chapman and Hall/CRC, Boca Raton, 2nd edition.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models – With an Implementation in R*. Springer, New York.

Yee, T. W. (2020). *VGAM: Vector generalized linear and additive models*. R package version 1.1-3.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society – Series B: Statistical Methodology*, **58**(3), 481–493.

Zhang, Y., Beesoon, S., Zhu, L., and Martin, J. W. (2013). Biomonitoring of perfluoroalkyl acids in human urine and estimates of biological half-life. *Environmental Science & Technology*, **47**(18), 10619–10627.

Zhu, B., Dunson, D. B., and Ashley-Koch, A. E. (2012). Adverse subpopulation regression for multivariate outcomes with high-dimensional predictors. *Statistics in Medicine*, **31**(29), 4102–4113.

Zhu, L., Carlin, B. P., and Gelfand, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, **14**, 537–557.

Appendix A

Additional figures on drinking water contamination

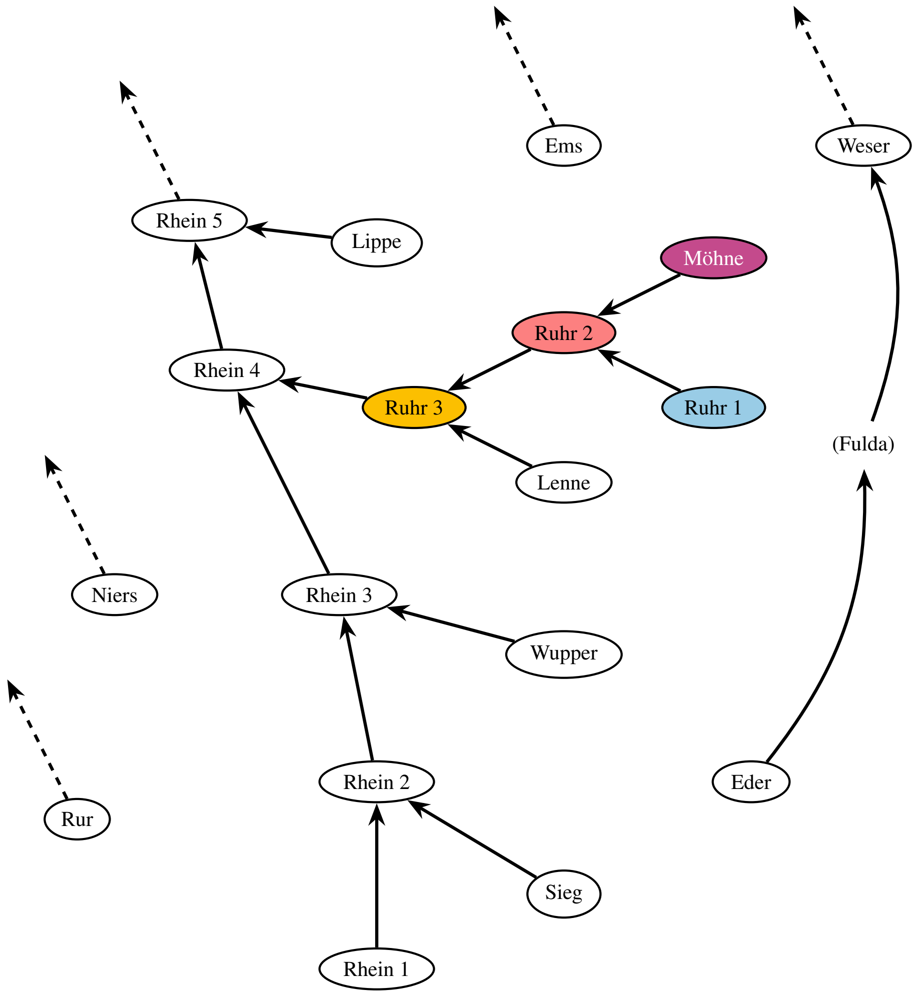

Figure A.1: River network in NRW relevant for drinking water supply: Every vertex represents a river segment between two important junctions, with the most affected Ruhr and Möhne region coloured as in Figures 2.8, A.3 and A.4 The paths represent the connection along the course of the water.

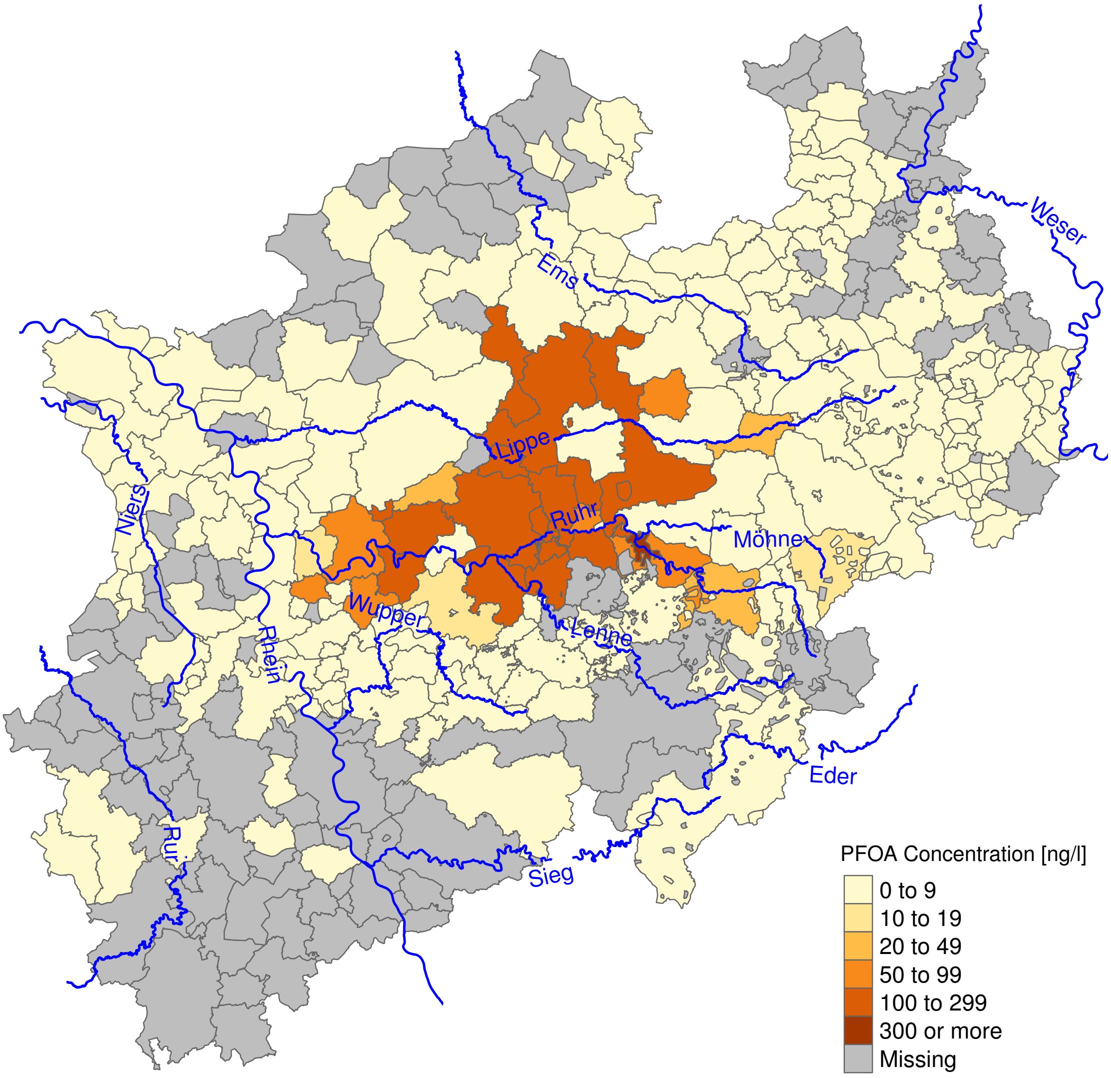

Figure A.2: PFOA concentration per water supply area measured in the respective network. The data closest to 1 July 2006 are used. Where no network samples are available, the respective supplying stations' measurements are averaged with weights according to Section 2.5. (With the limit of quantification at 10 ng/l, the respective areas in light yellow are below it.)

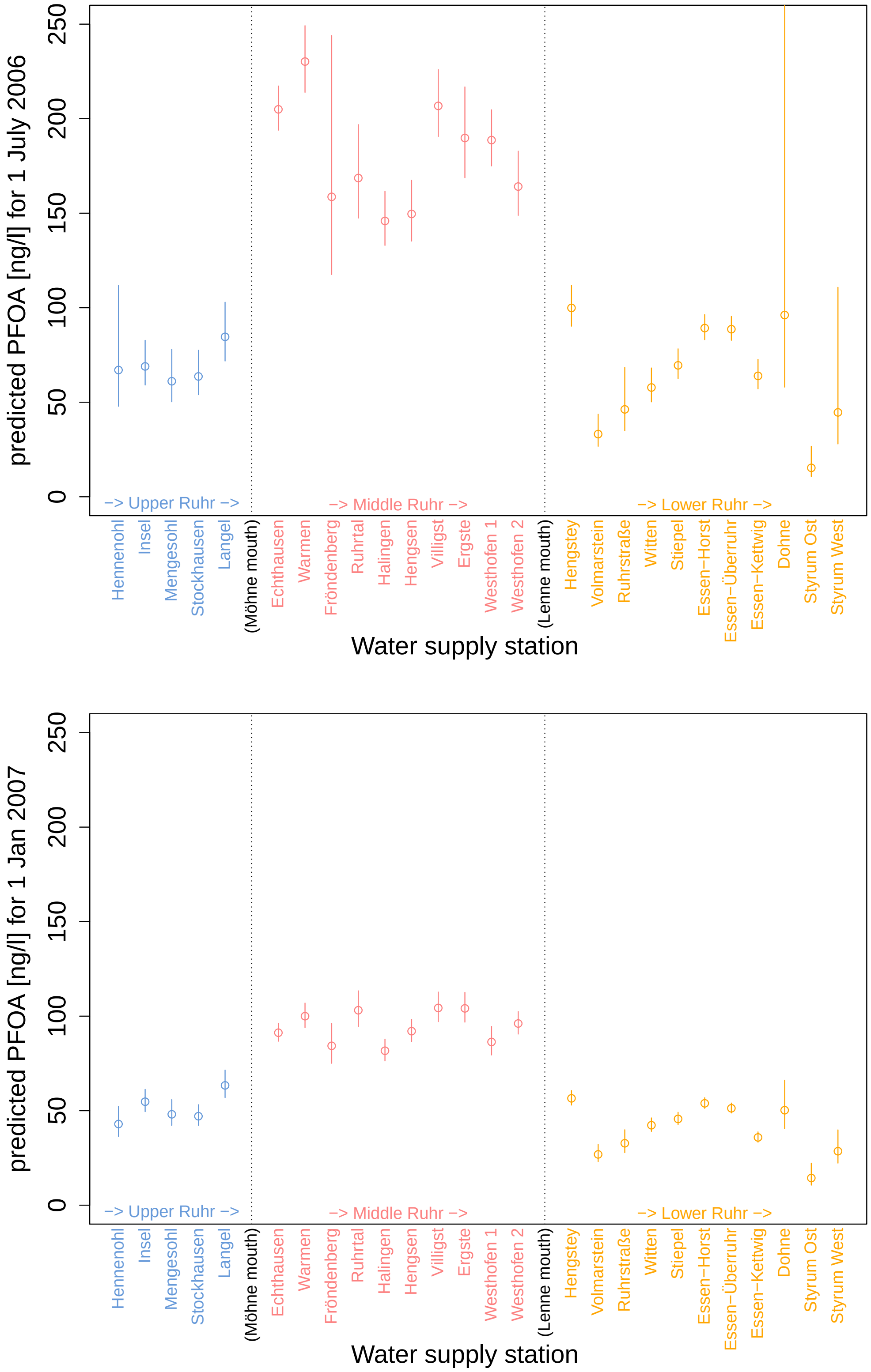

Figure A.3: Predictions (estimates and approximate 95% confidence intervals of the expected value) of the PFOA values from 2006 until 2009, from GLMs with Gaussian distribution and inverse link, individually for each station along the river Ruhr (page 1/3).

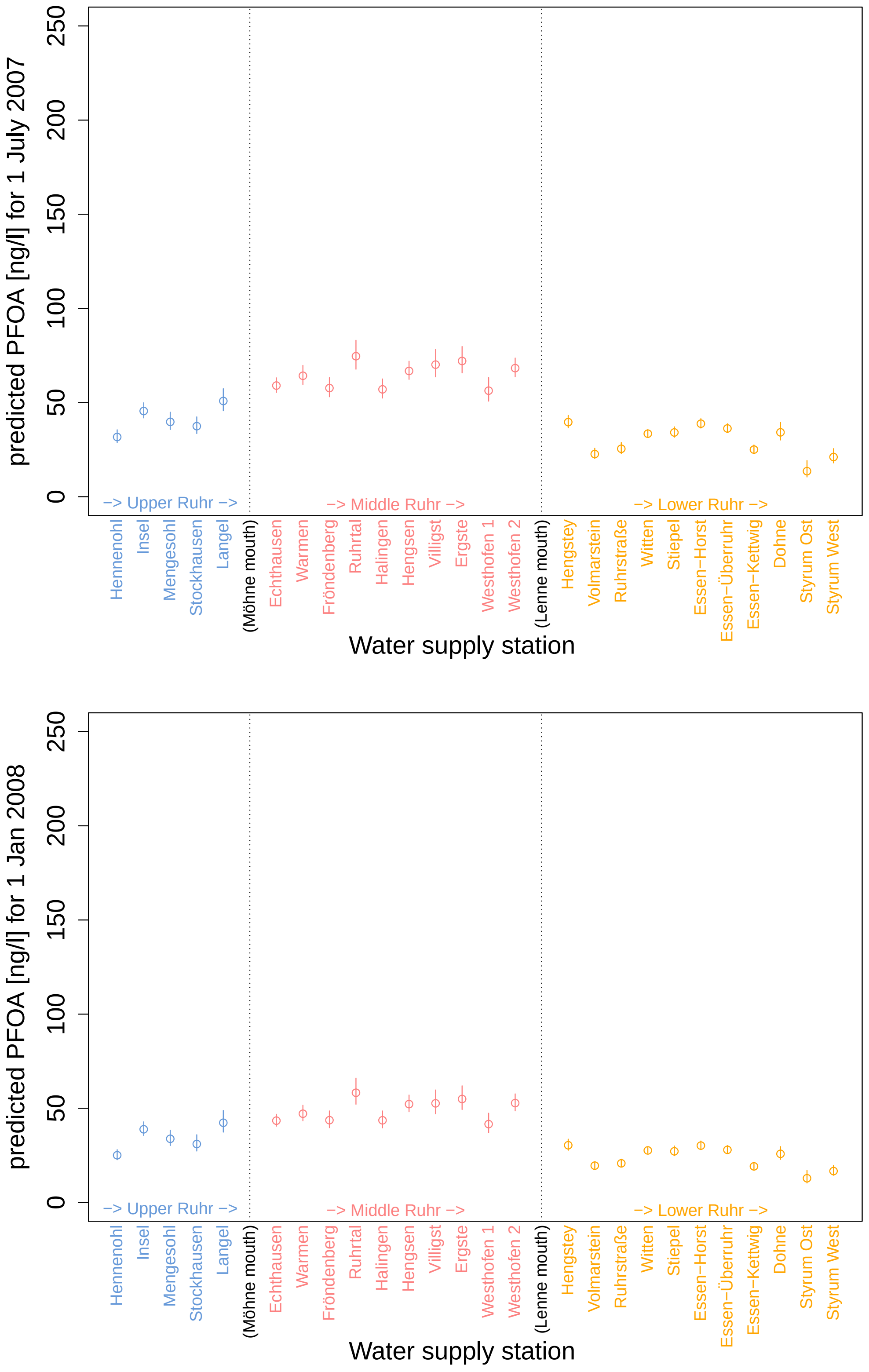

Figure A.3: PFOA predictions for six points of time (page 2/3).

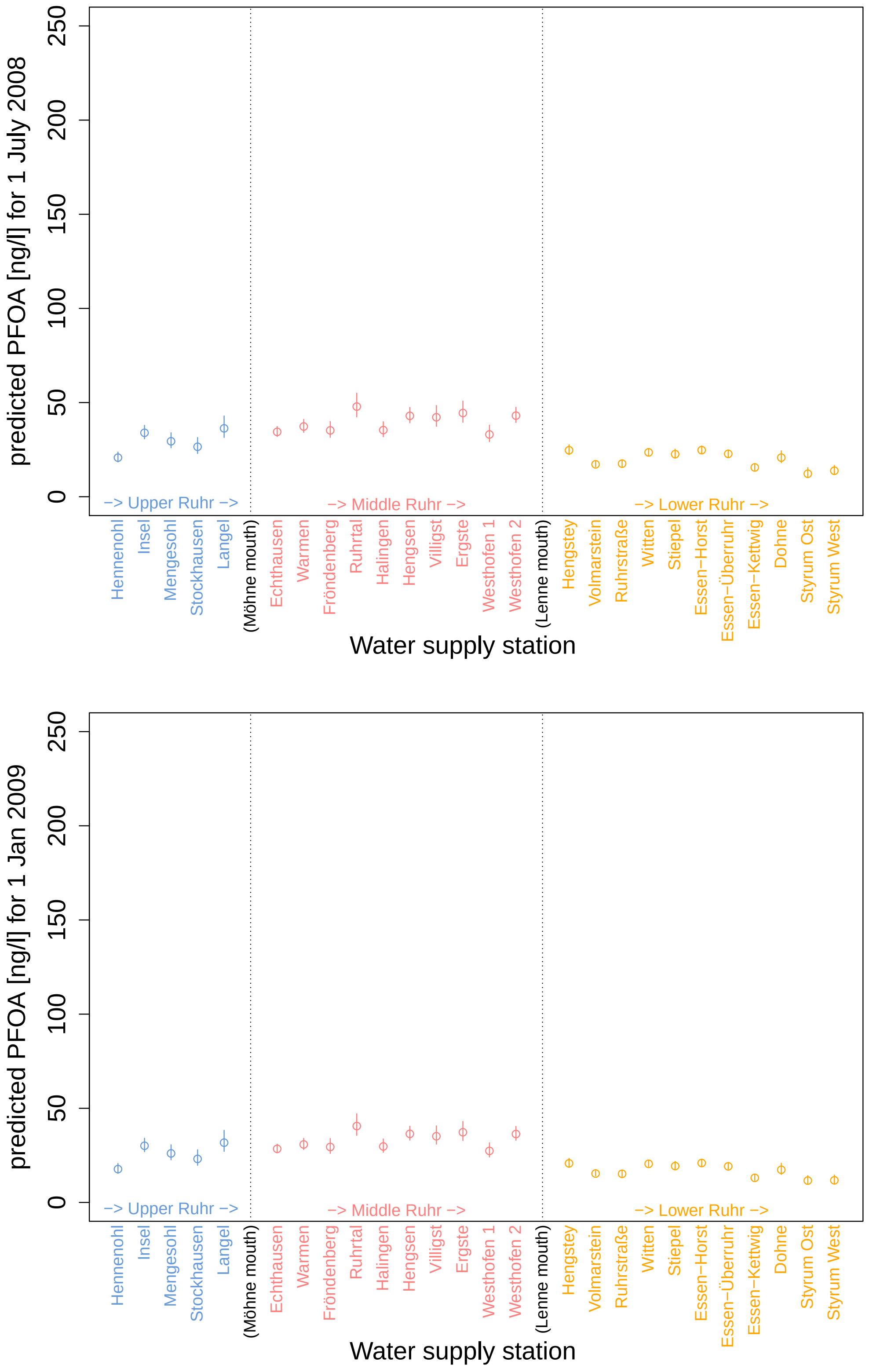

Figure A.3: PFOA predictions for six points of time (page 3/3).

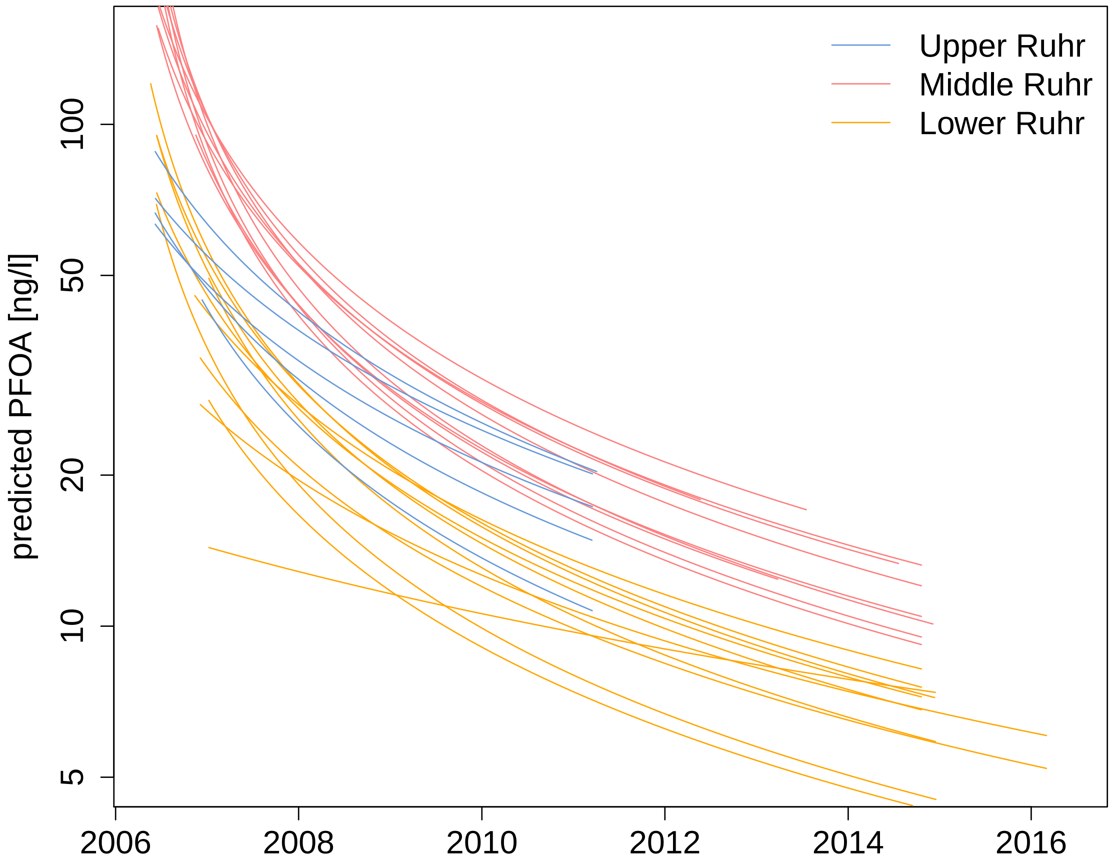

Figure A.4: Predictions of the PFOA values for all stations along the river Ruhr, from GLMs with Gaussian distribution and inverse link, for the respective periods of time in which data are available for the station.

Appendix B

Documentation

B.1 Data sources and availability

Due to the partly restricted data availability, and as the 'PerSpat' project group does not hold the copyright for any of the data sets, it is not possible to deposit them in a public repository.

- Measurement data from the relevant drinking water supply stations along the river Ruhr are publicly available from the LANUV at `www.elwasweb.nrw.de/elwas-hygrisc/twbericht/pft_tw.php?exhibit-use-local-resources` (12 June 2021).
- Other measurement data from the drinking water supply as well as further information (see Table 1.1) have been confidentially provided by the LANUV and some water supply station operators to the Department of Hygiene, Social and Environmental Medicine of the RUB.
- Some of the maps have been produced using geographical data from the German Federal Agency for Cartography and Geodesy (© GeoBasis-DE / BKG, 2019; license: `www.govdata.de/dl-de/by-2-0`; data: `wms_vg250-ew`, `wfs_dlm1000`).

- The perinatal registry data are available from the quality assurance office (*Geschäftsstelle Qualitätssicherung*, qs-nrw) located at the medical association Westphalia-Lippe. Restrictions apply to the availability of these data, which are used under license for the 'PerSpat' project, and so are not publicly available. These data can only be accessed on the premises of qs-nrw.

- Data from the Arnsberg biomonitoring cohort and its control groups are administered by the Department of Hygiene, Social and Environmental Medicine of the RUB and are confidential.
- Population density data from the German national census are publicly available at `www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html`
(12 June 2021).

B.2 Overview on computational steps

Commented programme code files in `R`, `Stan` or `BayesX`, respectively, are available from the public repository *OSF* at `https://osf.io/rfn8q/?view_only=644fff3f6c5d428398404faafeb648af`. They consist of exemplary code snippets and functions, to make the procedures transparent, which are reported in the essential parts of Chapters 2 and 3, although no data can be published alongside. The calculations are not documented in full; rather, repetitions of similar steps (e.g., with other covariates), data loading and writing, or standard applications are omitted or noted in comments.

Specifically, the essential analysis steps for the drinking water contamination data are listed in Section B.2.1 and for the perinatal registry data in Section B.2.2. Documentations regarding other parts of this dissertation are refrained from, as far as they are from other sub-projects within ‘PerSpat’, not meaningful without the actual data, preliminary, or inessential.

B.2.1 Drinking water contamination analyses

The subdirectory `2_water` contains the following files, representing essential parts of the drinking water data analyses in Chapter 2, Sections 2.4, 2.5 and 2.6:

1. The file

   - `1_prelim_transf.R`,

   written in `R`, contains preliminary studies in search for regression models for measurements depending on time, with pre-transformation of data or time using fractional polynomials (Section 2.4.1).

2. The file

   - `2_station_wise_glm.R`,

   written in `R`, contains GLM applications for PFOA measurements depending on time, individually for all water supply stations, in search for the optimal combination of distribution family and link function (Section 2.4.1) with results for the best fitting GLM (Section 2.4.2).

3. The file

   - `3_supply_proport.R`,

   written in `R`, contains the estimation of water supply proportions (shares, how much of a supply area’s water stems from a water supply station) from Section 2.5.

4. The files

   - `4a_ruhr_mixed.R` and
   - `4b_ruhr_mixed_bayes.R`,

   written in `R` with `Stan` code included, contain mixed model applications for PFOA measurements, jointly for all water supply stations along the river Ruhr, with several preliminary variations (Section 2.6.2) and with random effect distributions per river segment (Section 2.6.3), respectively.

B.2.2 Perinatal data analyses

The subdirectory `3_perinatal` contains the following files, representing essential parts of the perinatal data analyses in Chapter 3, Sections 3.3, 3.4 and 3.5:

1. The file

   - `1_fractional_polynomials.R`,

written in R, contains a preliminary study of fractional polynomials to find a parametric function of the gestational age in a regression model for the birth weight (Section 3.3).

2. The files
   - 2a_gaussian_marginal.txt and
   - 2b_dagum_marginal.txt,

   written in BayesX, contain the marginal distributional regression models: a Gaussian for the birth weight and a Dagum distribution for the gestational age (Section 3.4.3).

3. The files
   - 3a_marginal_choice_logscores.R and
   - 3b_marginal_choice_pit_qr.R,

   written in R, contain the procedures to compare the fit of marginal Gaussian and Dagum distributions for birth weight and gestational age, by log-scores and by diagnostic plots, respectively (Section 3.4.3).

4. The file
   - 4_conditional_correlation.R,

   written in R, is to calculate plots of correlations of birth weight and gestational age conditional on covariates (Section 3.4.4).

5. The files
   - 5a_gaussian_copula.txt,
   - 5b_clayton_copula.txt and
   - 5c_gumbel_copula.txt,

   written in BayesX, contain the three possible copula distributional regression models for the joint distribution, with a Gaussian, a Clayton and a Gumbel copula, respectively (Section 3.4.4).

6. The file
   - 6_copula_prediction_plots.R,

   written in R, is to calculate contour plots of two-dimensional predictions from the Clayton copula model for birth weight and gestational age (Section 3.4.5).

7. The files
   - 7a_copula_vs_polynomial_logscores.R and
   - 7b_copula_vs_polynomial_plots.R,

   written in R, contain the cross validation study to compare the fit of the bivariate Clayton copula model (afterwards reduced to the distribution of birth weight conditional on gestational age) to the univariate polynomial model for birth weight, by log-scores and graphically by predictions, respectively (Section 3.5).