technische universität dortmund

RUHR UNIVERSITÄT BOCHUM RUB

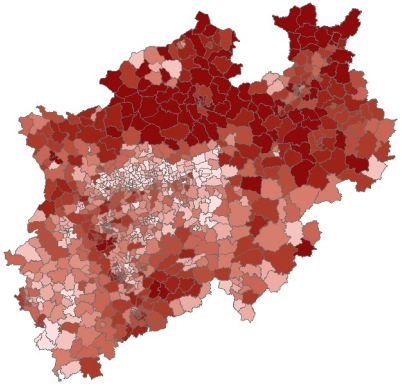
# Bayesian Bivariate Distributional Copula Regression of Birth Weight and Gestational Age

**Jonathan Rathjens**[1] Arthur Kolbe[2] Hans-Joachim Bücker-Nott[3] Jürgen Hölzer[2]
Nadja Klein[4] Katja Ickstadt[1]

**4 December 2020**

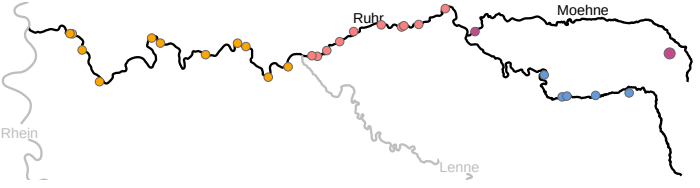
[1]TU Dortmund University [2]Ruhr-University Bochum [4]Humboldt University of Berlin
[3]Medical Association Westphalia-Lippe, Quality Assurance Office (qs-nrw)

## Project “PerSpat”



mean birth weights in NRW in 2010 by postal code

- environmental epidemiological study on general population
- associations of Perfluorooctanoic Acid (PFOA) with perinatal parameter, esp. birth weight**?**
- data on PFOA in drinking water used for local predictions and individual risk assessment
- state-wide perinatal registry data for North Rhine-Westphalia (NRW), Germany




contaminated river Ruhr

## Today: Birth Data Analysis

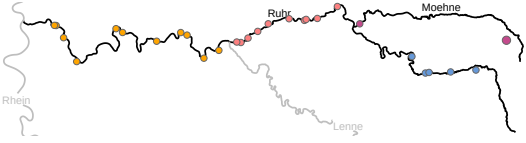
- consider bivariate response (birth weight and gestational age – cf., e.g., Gage, 2003; Ananth and Platt, 2004; Schwartz et al., 2010)
- modelled using copulas in distributional regression (Klein and Kneib, 2016)
- model fit of marginal and copula families
- covariate selection
- comparison to usual univariate regression for birth weight (cf., e.g., Gardosi et al., 1995; Salomon et al., 2007)
- introduction of PFOA exposure risk to analysis

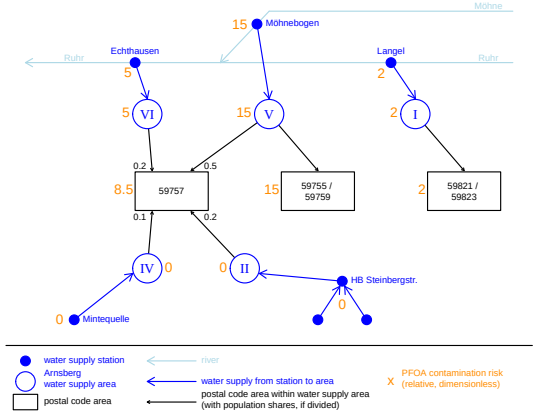
## Data Source and Overview

- collected by hospitals
- combined and processed at state medical association for quality assurance in obstetrical healthcare
- from 2003 until 2014
- about 1.7 million records
- more than 200 biometric, medical and social variables on mother and child, pregnancy, birth and treatment
- restricted to singletons born alive within NRW, and anonymised

## Data Subset

- PFOA contamination in drinking water of Arnsberg, NRW, about 2004
- restriction to data from this town
- 4451 complete records
  (of 6442 with missings, mainly of social variables and smoking)

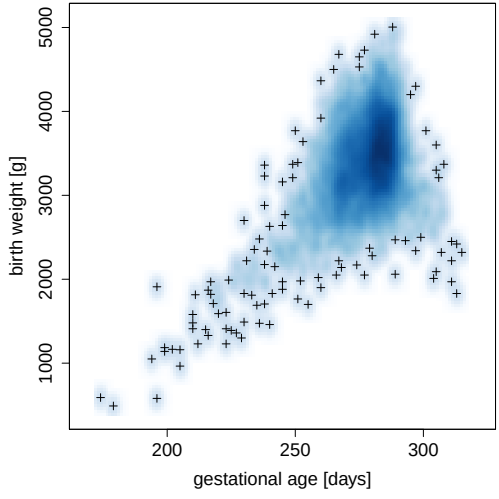



Arnsberg contamination risk by postal code in 2006

## Response Variables

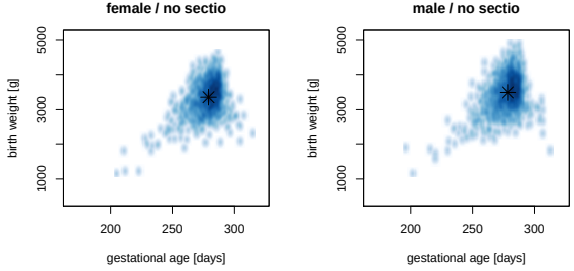
- birth weight ($y_1$) measured in g, of primary interest
- gestational age ($y_2$) in days, usually as most important influence (e.g., Gardosi et al., 1995)



## Covariates of Interest

- the child's sex,
- the number of previous pregnancies of the mother,
- whether the child has been delivered by Cesarean section,
- whether the birth has been induced,
- the mother's age in years,
- the mother's height in **cm**,
- the mother's body mass index measured in $\mathbf{kg/m^2}$ at the beginning of pregnancy,
- the gain of weight of the mother during pregnancy measured in **kg**,
- the number of cigarettes the mother reports to smoke per day,
- whether the mother is single and
- whether the mother is employed.

## Conditional on Covariates




Data Situation

## Distributional Regression

- joint response density $f$ with copula density $c_\rho$:

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

- marginal parameters (location, scale, shape, ...) according to distribution families
- copula parameter $\rho$
- GLM with linear predictors

$$\eta^{(\theta)} = \beta_0^{(\theta)} + \beta_1^{(\theta)} x_1 + \ldots + \beta_m^{(\theta)} x_m$$

  and response functions $g_\theta(\eta^{(\theta)}) = \theta$, individually for all parameters $\theta$
- several combinations of copula and marginal families implemented in `BayesX` (Belitz et al., 2015)
- MCMC estimation with flat normal priors

## Copula families

Gauss:

$$c_\rho(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{\rho}{1-\rho^2}\left(\rho(\Phi^{-1}(u))^2 - 2\Phi^{-1}(u)\Phi^{-1}(v) + \rho(\Phi^{-1}(v))^2\right)\right),$$

Clayton:

$$c_\rho(u,v) = (1+\rho)(uv)^{-1-\rho}\left(u^{-\rho} + v^{-\rho} - 1\right)^{-2-\frac{1}{\rho}},$$

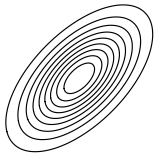
Gumbel:

$$c_\rho(u,v) = \frac{1}{uv}(-\ln u)^{\rho-1}(-\ln v)^{\rho-1}\exp\left(-h^{\frac{1}{\rho}}\right)\left(h^{\frac{2}{\rho}-2} - (1-\rho)h^{\frac{1}{\rho}-2}\right),$$

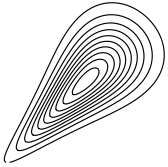
where

$$h = (-\ln u)^\rho + (-\ln v)^\rho$$

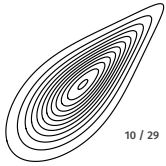
Gauss with $\rho = 0.6$



Clayton with $\rho = 1.5$



Gumbel with $\rho = 2$

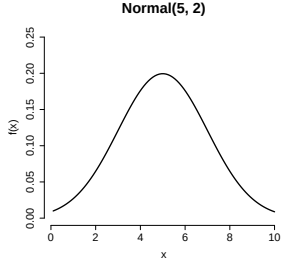


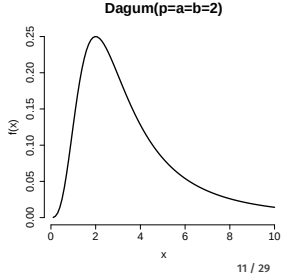
## Marginal Families

- Normal distribution $N(\mu, \sigma^2)$
- Dagum distribution:

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

- independently chosen for both response variables



## Univariate Polynomial Model

- standard approach (cf., e.g., Gardosi et al., 1995; Salomon et al., 2007)
- application for comparison to bivariate model regarding performance of birth weight prediction

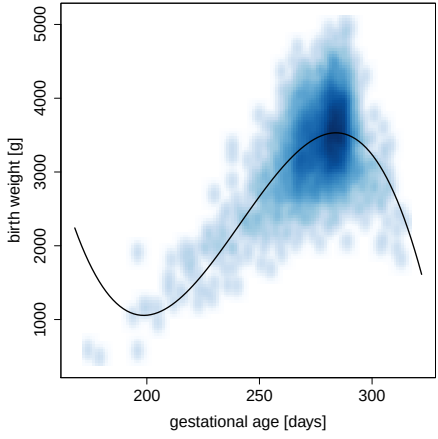
- standard polynomial regression

$$y_1 = \beta_0 + p_\lambda(y_2) + \beta_1 x_1 + \ldots + \beta_m x_m + \epsilon$$

- 'full' polynomial

$$p_\lambda(y_2) = \lambda_1 y_2 + \lambda_2 y_2^2 + \lambda_3 y_2^3$$

has fitted best in preliminary study




(without further covariates)

## Work Flow

1. data preparation
   - import, cleansing, NA-omitting
   - response standardisation
2. choice of marginal families
   - Normal and Dagum models for each univariate response
   - select covariates with sign. influence ($0 \notin CI_{95\%}(\beta)$)
   - compare families by quantile plots and log-scores
3. choice of copula family
   - Gauss, (rotated) Clayton and (rotated) Gumbel copula, combined with optimal marginals
   - select covariates with influence on copula (dependence) parameter
   - compare families by DIC and WAIC
4. evaluation of final model
   - estimated regression coefficients
   - compare to standard polynomial regression approach
   - add PFOA exposure risk, and time

## Data Preparation

- response data transformations for numerical reasons:
- data-independent pre-standardisation for Normal distribution:

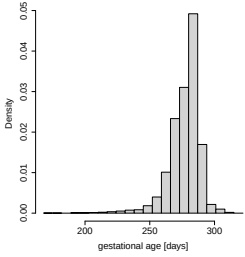
$$\tilde{y}_1 = \frac{y_1 - 3500}{500}, \qquad \tilde{y}_2 = \frac{y_2 - 280}{14}$$

- pre-normalisation for Dagum distribution:

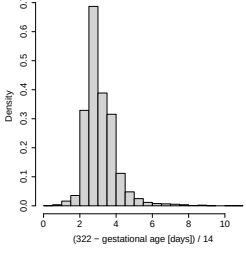
$$\tilde{y}_1 = \frac{y_1}{500}$$

with additional shift for gestational age:

$$\tilde{y}_2 = \frac{322 - y_2}{14}$$

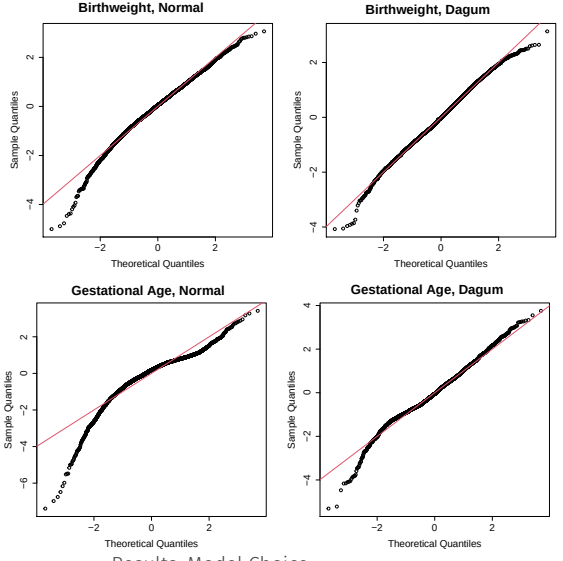



shifted to



## Marginals

quantile plots of residuals ($\Phi^{-1}(F(y;\hat{\beta}))$ with posterior means of the $\beta$'s):



## Marginals

### log-scores

- sample $y$ from a sample of the respective $\beta$'s posteriors and
- compare to observations by log-scores obtained with `scoringRules::logs_sample`
- calculate mean scores:

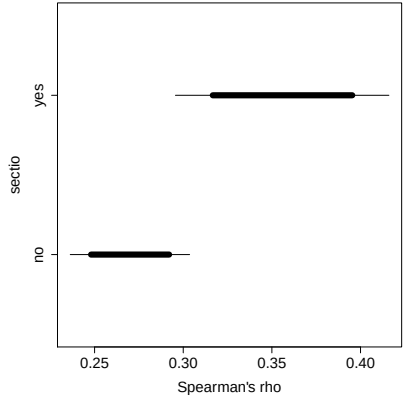
| | birth weight | gest. age |
|---|---|---|
| Normal | 7.58 | 3.97 |
| Dagum | 7.56 | 3.73 |

### family selection

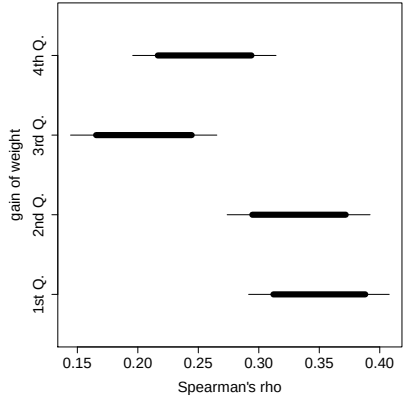
- **Dagum model for gestational age** (better score, convincing quantile plot)
- in doubt, we use **Normal model for birth weight** due to
  - interpretability (influences on mean birth weight of primary interest)
  - comparability to other approaches
  - ease of computation

## Exploration: Conditional Correlation

rank correlation of birth weight and gestational age (80% and 95% confidence intervals) from data subsets according to covariate values:




by sectio




by quartiles of maternal gain of weight

## Copula families

### information criteria

| | DIC | WAIC |
|---|---|---|
| Gauss | Inf | Inf |
| Gumbel | Inf | Inf |
| – 90° | Inf | Inf |
| **Clayton** | **20 981** | **21 245** |
| – 90° | 21 293 | 21 302 |

### selection

- possible **90°** rotation of Gumbel and Clayton copula, to cover all possible directions of tail dependence
- technical MCMC problems when using to many covariates in $\eta^{(\rho)}$
- therefore pre-selection based on Gauss copula result and conditional correlation estimation

- (non-rotated) Clayton copula fits best
- in accordance with lower tail visible in data**?**
- but with low estimates of $\rho$, not far from independence
- copula parameter not influenced by most of the covariates

## Significant Covariate Influences in Final Model

### on birth weight's mean

- − sex (female)
- + previous pregnancies
- − sectio
- + induction
- + maternal height
- + maternal BMI
- + maternal gain of weight
- − maternal smoking

### on birth weight's scale

- − sex
- + sectio
- + maternal BMI
- + maternal smoking

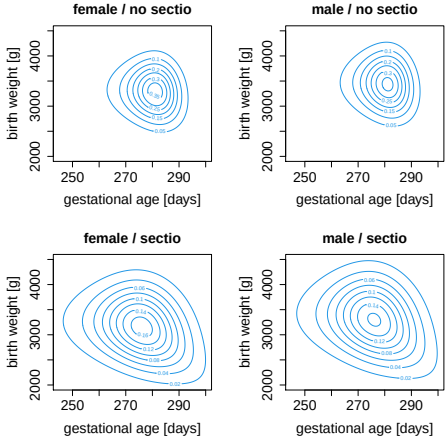
### on Dagum parameters for gestational age

- ○ previous pregnancies
- ○ sectio
- ○ induction
- ○ maternal gain of weight
- ○ maternal smoking
- ○ mother is employed

### on copula parameter (dependence)

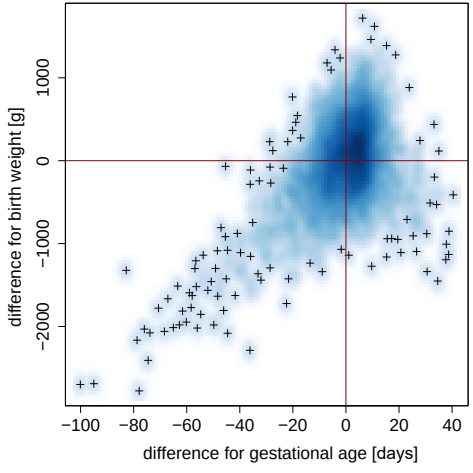
- + sectio

## Predictions



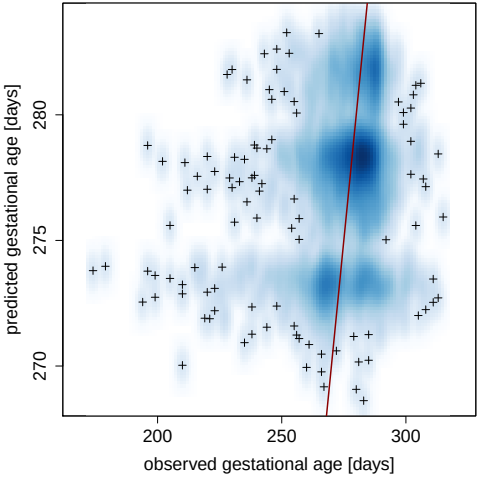

density of $f(y_1, y_2)$ from copula model with posterior mean parameters for certain covariate values (others: maternal height: 170 **cm**, maternal BMI: 20 $\mathbf{kg/m^2}$, maternal gain of weight: 10 **kg**, all others set to 'no'/0)

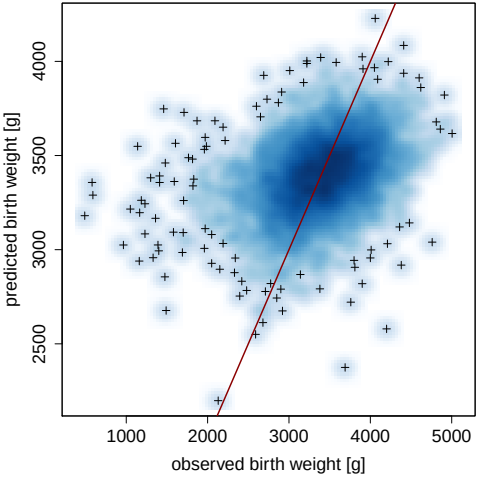
## ‘Residuals’ of Copula Model (Two-dimensional)



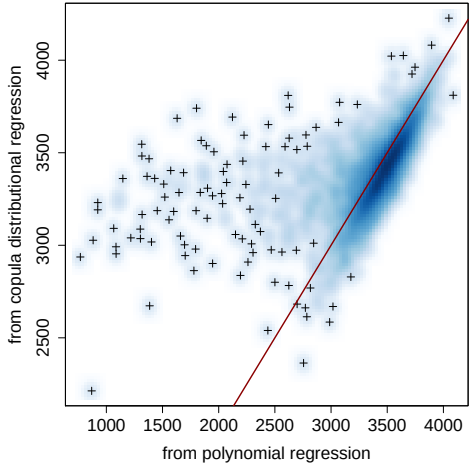

observed – predicted values

## ‘Residuals’ of Copula Model (Dimension-wise)



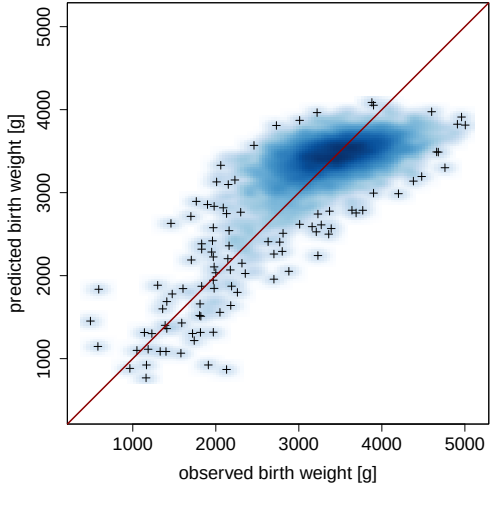
## Comparison to Univariate Polynomial Regression




predicted birth weight

Results: Evaluation

## Univariate Polynomial Regression

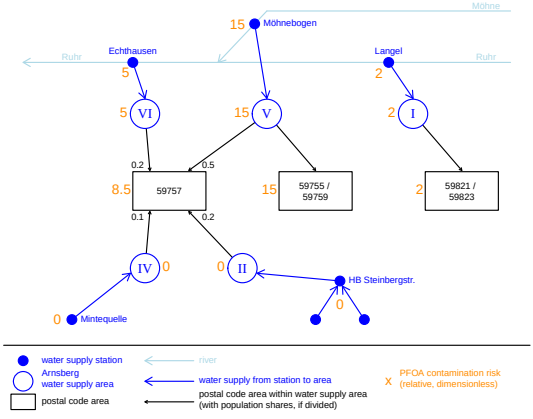


## Regression Results for Mean Birth Weight

Posterior mean (and sd.) of regression coefficients for parameter $\mu$ of (standardised) birth weight:

| Covariate | Polynomial | | $\mu$ from copula model | |
|---|---|---|---|---|
| sex (female) | -0.2908 | (± 0.0237) | -0.2896 | (± 0.0273) |
| previous pregnancies | 0.0583 | (± 0.0085) | 0.0484 | (± 0.0065) |
| sectio | not signif. | | -0.2907 | (± 0.0535) |
| induction | not signif. | | 0.0870 | (± 0.0280) |
| maternal height | 0.0279 | (± 0.0018) | 0.0290 | (± 0.0015) |
| maternal BMI | 0.0302 | (± 0.0023) | 0.0368 | (± 0.0029) |
| maternal gain of weight | 0.0144 | (± 0.0022) | 0.0249 | (± 0.0024) |
| maternal smoking | -0.0308 | (± 0.0029) | -0.0416 | (± 0.0031) |
| mother is single | -0.1271 | (± 0.0480) | not signif. | |
| gestational age | -2.6145 | (± 0.1717) | | |
| squared gestational age | 0.0112 | (± 0.0007) | | |
| cubic gestational age | -0.0000 | (± 0.0000) | | |

# Introduction of PFOA exposure risk




Arnsberg contamination risk by postal code in 2006

- estimated PFOA concentrations in drinking water (Rathjens et al., 2020)
- by postal code and time
- add to polynomial regression and to birth weight mean regression in copula model
- also time as linear covariate

⇒

- no significant effect of PFOA risk in either model
- significant effect of time in copula model (birth weight becoming less)

## Discussion

- birth weight sufficiently modelled with normal distribution
- gestational age better with Dagum distribution
- distributional regression useful to allow for influences on scale and dependence
- with regard to birth weight, the result of the more general copula model is close to the more specified univariate polynomial model
- rather weak dependence between response variables in copula model
- much variance remains unexplained
- lower tail in data not reflected in copula predictions (main data part in center predominating)

## Perspectives

- cross-validated prediction results
- temporal modelling:
- more complex GAM, e.g., with splines
- individual PFOA data from Arnsberg cohort study
- spatial modelling (with larger region)

## References

Ananth, C. V. and Platt, R. W. (2004). Reexamining the Effects of Gestational Age, Fetal Growth, and Maternal Smoking on Neonatal Mortality. BMC Pregnancy Childbirth, **4**(1), 1–9.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). BayesX – Bayesian Inference in Structured Additive Regression Models.

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

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

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying bayesian approach. Stat. Comput., **26**(4), 841–860.

Rathjens, J., Becker, E., Kolbe, A., Ickstadt, K., and Hölzer, J. (2020). Spatial and Temporal Analyses of Perfluorooctanic Acid in Drinking Water for External Exposure Assessment in the Ruhr Metropolitan Area, Germany. Stoch. Environ. Res. Risk Assess., (in production).

Salomon, L.-J., Bernard, J.-P., de Stavola, B., Kenward, M., and Ville, Y. (2007). Poids et taille de naissance: Courbes et Équations. J. Gynecol. Obstet. Biol. Reprod., **36**(1), 50–56.

Schwartz, S. L., Gelfand, A. E., and Miranda, M. L. (2010). Joint Bayesian Analysis of Birthweight and Censored Gestational Age Using Finite Mixture Models. Stat. Med., **29**(16), 1710–1723.