

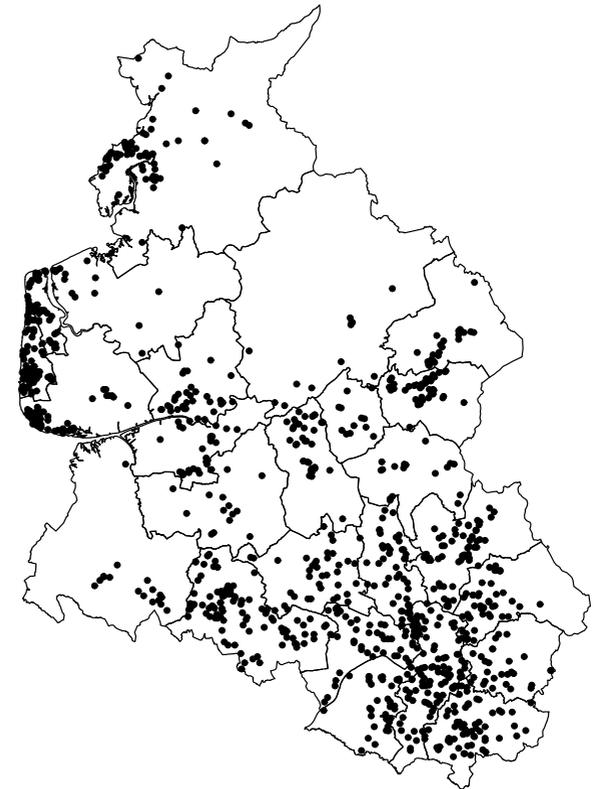
# A mixed model approach for geoadditive hazard regression

Thomas Kneib & Ludwig Fahrmeir  
Department of Statistics, Ludwig-Maximilians-University Munich

1. Leukemia survival data
2. Geoadditive hazard regression
3. Mixed model based inference
4. Results
5. Software
6. Discussion

## Leukemia survival data

- Survival time of adults after diagnosis of acute myeloid leukemia.
- 1,043 cases diagnosed between 1982 and 1998 in Northwest England.
- 16 % (right) censored.
- **Continuous** and **categorical** covariates:
  - age* age at diagnosis,
  - wbc* white blood cell count at diagnosis,
  - sex* sex of the patient,
  - tpi* Townsend deprivation index.
- **Spatial information** in different resolution.



- Classical Cox **proportional hazards model**:

$$\lambda(t; x) = \lambda_0(t) \exp(x' \gamma).$$

- **Baseline-hazard**  $\lambda_0(t)$  is a nuisance parameter and **remains unspecified**.
- Estimate  $\gamma$  based on the partial likelihood.
- Questions / Limitations:
  - Estimate the baseline **simultaneously** with covariate effects.
  - **Flexible** modelling of covariate effects (e.g. nonlinear effects, interactions).
  - **Spatially correlated** survival times.
  - **Non-proportional hazards** models / **time-varying effects**.

⇒ Geoadditive hazard regression models.

## Geoadditive hazard regression

- Replace usual parametric predictor with a **flexible semiparametric** predictor

$$\lambda(t; \cdot) = \lambda_0(t) \exp[f_1(\text{age}) + f_2(\text{wbc}) + f_3(\text{tpi}) + f_{\text{spat}}(s_i) + \gamma_1 \text{sex}]$$

and **absorb the baseline**

$$\lambda(t; \cdot) = \exp[f_0(t) + f_1(\text{age}) + f_2(\text{wbc}) + f_3(\text{tpi}) + f_{\text{spat}}(s_i) + \gamma_1 \text{sex}]$$

where

- $f_0(t) = \log(\lambda_0(t))$  is the **log-baseline-hazard**,
- $f_1, f_2, f_3$  are **nonparametric** functions of age, white blood cell count and deprivation, and
- $f_{\text{spat}}$  is a **spatial** function.

- $f_0(t), f_1(\text{age}), f_2(\text{wbc}), f_3(\text{tpi})$ : **P-splines**
  - Approximate  $f_j$  by a B-spline of a certain degree (basis function approach).
  - Penalize differences between parameters of adjacent basis functions to ensure smoothness.
  - Alternatives: **Random walks**, more general **autoregressive priors**.
- $f_{\text{spat}}(s)$ : District-level analysis
  - **Markov random field approach**.
  - Generalization of a first order random walk to two dimensions.
  - Consider two districts as **neighbors** if they share a common boundary.
  - Assume that the expected value of  $f_{\text{spat}}(s)$  is the **average** of the function evaluations of adjacent sites.

- $f_{spat}(s)$ : Individual-level analysis
  - **Stationary Gaussian random field** (kriging).
  - Spatial effect follows a zero mean stationary Gaussian stochastic process.
  - Correlation of two arbitrary sites is defined by an intrinsic **correlation function**.
  - Low-rank approximations to Gaussian random fields.
- Extensions
  - Cluster-specific **frailties**.
  - **Surface smoothers** based on two-dimensional P-splines.
  - **Varying coefficient terms** with continuous or spatial effect modifiers.
  - **Time-varying effects** based on varying coefficient terms with survival time as effect modifier.
- Geoadditive hazard regression handles all model terms in a **unified way**.

- Express  $f_j$  as the product of a **design matrix**  $Z_j$  and **regression coefficients**  $\beta_j$ .
- Rewrite the model in matrix notation as

$$\log(\lambda(t; \cdot)) = Z_0(t)\beta_0 + Z_1\beta_1 + Z_2\beta_2 + Z_3\beta_3 + Z_{spat}\beta_{spat} + U\gamma.$$

- Bayesian approach: Assign an appropriate **prior** to  $\beta_j$ .
- Frequentist approach: Assume (correlated) **random effects distribution** for  $\beta_j$ .
- All priors can be cast into the **general form**

$$p(\beta_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j\right)$$

where  $K_j$  is a **penalty matrix** and  $\tau_j^2$  is a **smoothing parameter**.

- Type of the covariate and prior beliefs about the smoothness of  $f_j$  determine special  $Z_j$  and  $K_j$ .

## Mixed model based inference

- Each parameter vector  $\beta_j$  can be partitioned into an **unpenalized part** (with flat prior) and a **penalized part** (with i.i.d. Gaussian prior), i.e.

$$\beta_j = Z_j^{unp} \beta_j^{unp} + Z_j^{pen} \beta_j^{pen}$$

- This yields a **variance components model**

$$\eta = X^{unp} \beta^{unp} + X^{pen} \beta^{pen}$$

with

$$p(\beta^{unp}) \propto \text{const} \quad \beta^{pen} \sim N(0, \Lambda)$$

and

$$\Lambda = \text{blockdiag}(\tau_0^2 I, \dots, \tau_{spat}^2 I).$$

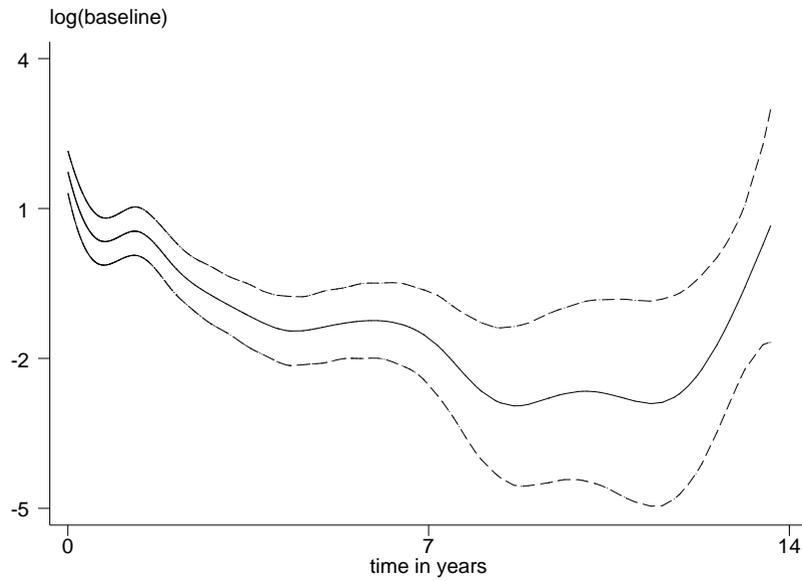
- Regression coefficients are estimated via a **Newton-Raphson-algorithm**.
- Numerical integration has to be used to evaluate the log-likelihood and its derivatives.

- The variance components representation with proper priors allows for **restricted maximum likelihood / marginal likelihood** estimation of the variance components:

$$L(\Lambda) = \int L(\beta^{unp}, \beta^{pen}, \Lambda) p(\beta^{pen}) d\beta^{pen} d\beta^{unp} \rightarrow \max_{\Lambda}.$$

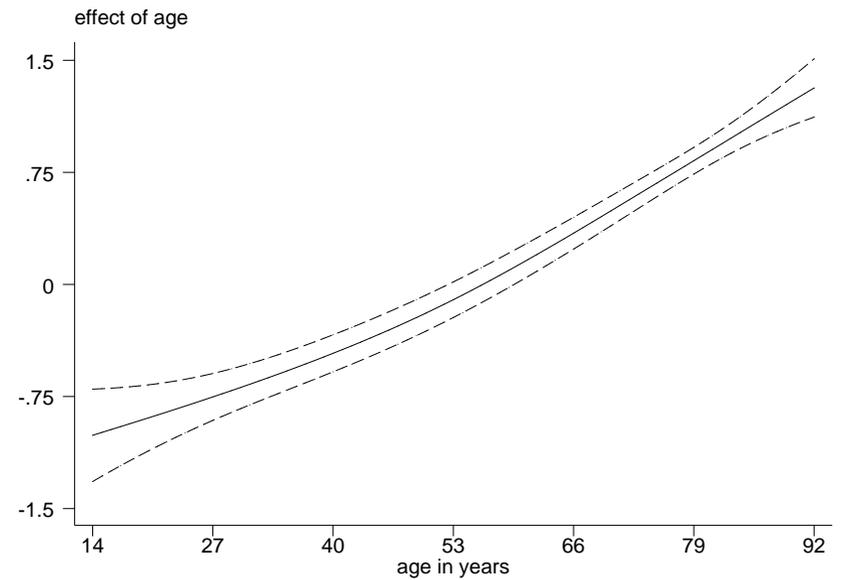
- The marginal likelihood can not be derived analytically.
- Some approximations lead to a simple Fisher-scoring-algorithm.
- Proved to work well in simulations and applications.
- We obtain **empirical Bayes / posterior mode** estimates.

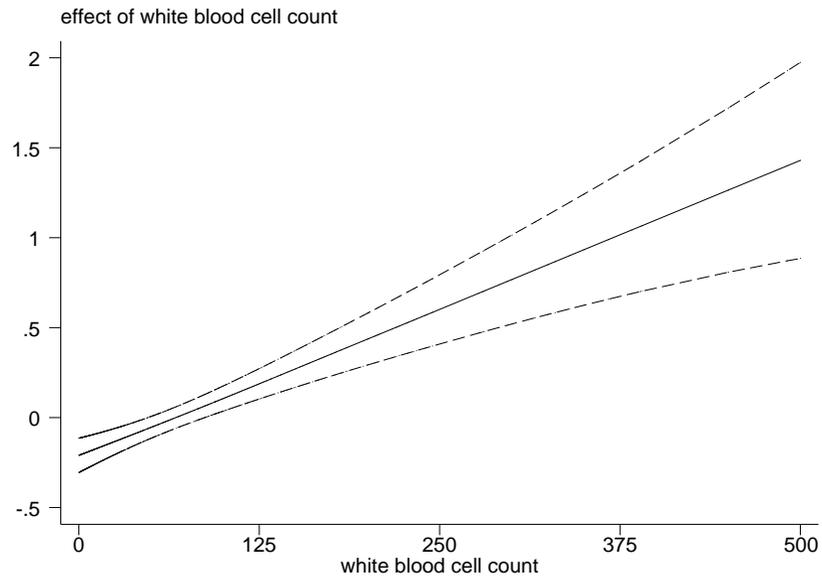
# Results



Log-baseline hazard.

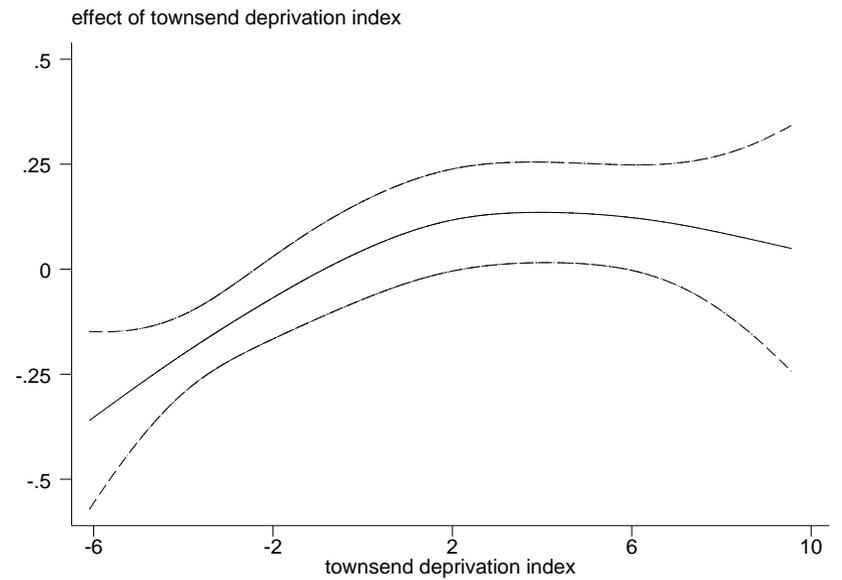
Effect of age at diagnosis.

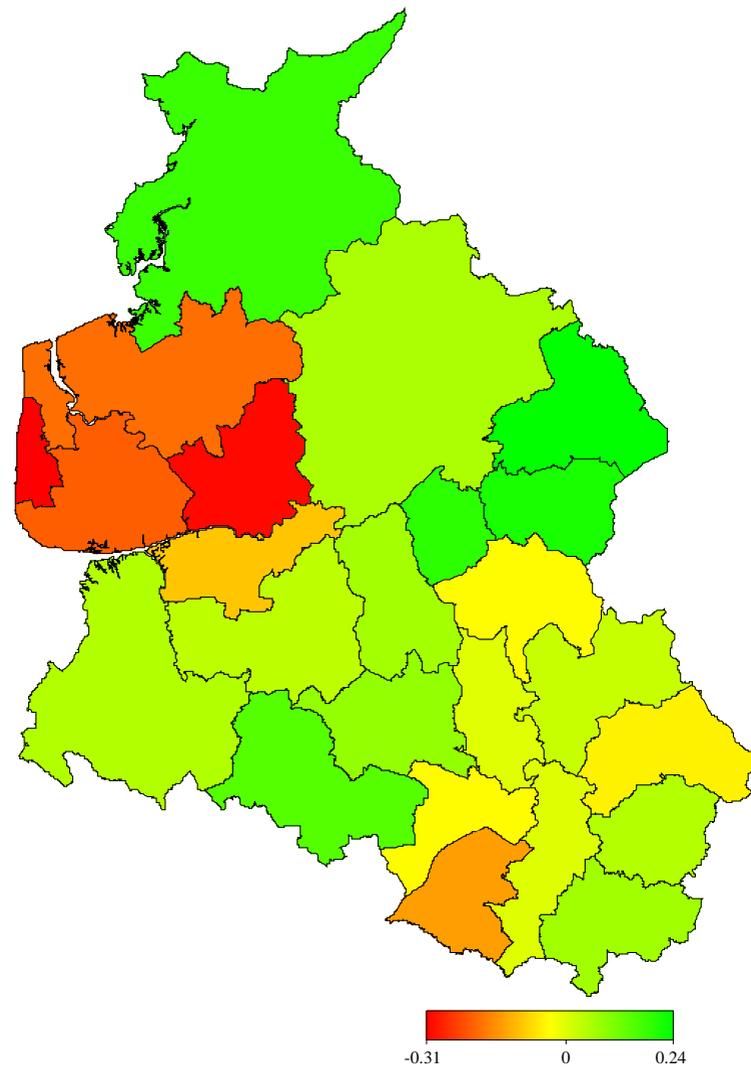




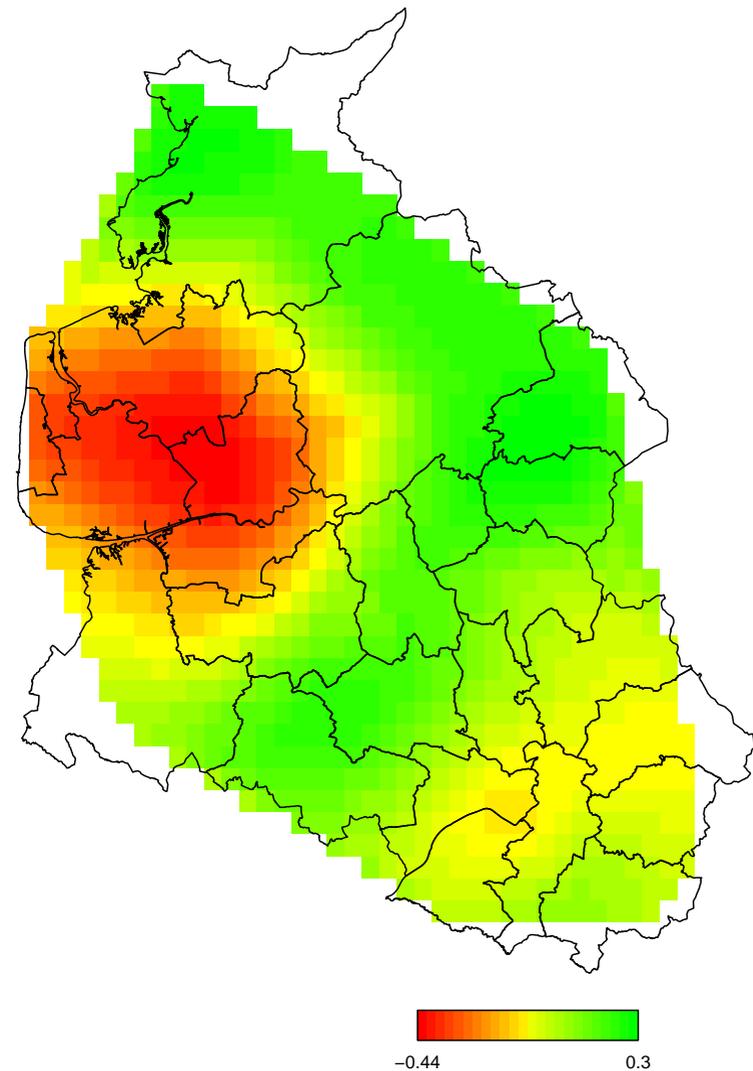
Effect of white blood cell count.

Effect of deprivation.





District-level analysis



Individual-level analysis

## Software

- Estimation was carried out using BayesX, a public domain software package for Bayesian inference.



- Available from

<http://www.stat.uni-muenchen.de/~lang/bayesx>

- Features (within a mixed model setting):
  - Responses: Gaussian, Gamma, Poisson, Binomial, ordered and unordered multinomial, **Cox models**.
  - Nonparametric estimation of the **log-baseline** and **time-varying effects** based on P-splines.
  - Continuous covariates and time scales: Random Walks, P-splines, autoregressive priors for seasonal components.
  - **Spatial Covariates**: Markov random fields, stationary Gaussian random fields, two-dimensional P-Splines.
  - Interactions: Two-dimensional P-splines, varying coefficient models with continuous and spatial effect modifiers.
  - Random intercepts and random slopes.

## Discussion

- Comparison with fully Bayesian approach based on MCMC (Hennerfeind et al., 2003):

### Cons:

- Credible intervals rely on asymptotic normality.
- Only plug-in estimates for functionals.
- Approximations for marginal likelihood estimation.

### Pros:

- No questions concerning mixing and convergence.
- No sensitivity with respect to prior assumptions on variance parameters.
- Somewhat better point estimates (in simulations).
- Numerical integration is required less often.

- **Future work:**
  - More general censoring / truncation schemes.
  - Event history / competing risks models

## References

- Kneib, T. and Fahrmeir, L. (2004): A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400, University of Munich.
- Fahrmeir, L., Kneib, T. and Lang, S. (2004): Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14, 715-745.
- Available from

`http://www.stat.uni-muenchen.de/~kneib`