Gemischte Modelle zur Schätzung geoadditiver Regressionsmodelle

Thomas Kneib & Ludwig Fahrmeir
Institut für Statistik, Ludwig-Maximilians-Universität München

1. Regressionsmodelle für geoadditive Daten
2. Beispiel I: Waldschäden im Forstgebiet Rothenbuch
3. Strukturiert additive Regressionsmodelle
4. Schätzverfahren basierend auf gemischten Modellen
5. Software
6. Beispiel II: Akute myeloische Leukämie

Deutsche Statistische Gesellschaft

20.05.2005

LMU

Regressionsmodelle für geoadditive Daten

- Regressionsmodelle: Generalisierte lineare Modelle, Modelle für kategoriale Daten, Modelle der Überlebenszeitanalyse.
- Übliche Struktur: Zielvariable wird in Abhängigkeit von stetigen und kategorialen Einflussgrößen modelliert.
- Geoadditive Daten: Zusätzlich ist räumliche Information gegeben.
- Probleme rein parametrischer Modellierungen:
  - Räumliche und zeitliche Korrelationen,
  - zeitlich oder räumlich variierende Effekte,
  - nichtlineare Effekte,
  - komplexe Interaktionen,
  - unbeobachtete Heterogenität.

Beispiel I: Waldschäden im Forstgebiet Rothenbuch

- Jährliche Erhebung von Waldschäden von 1983 bis 2001.
- 83 Buchen in einem 15 km $\times$ 10 km großen Gebiet.
- Abhängige Variable: Entlaubungsgrad von Buche i in Jahr t, gemessen in drei geordneten Kategorien:

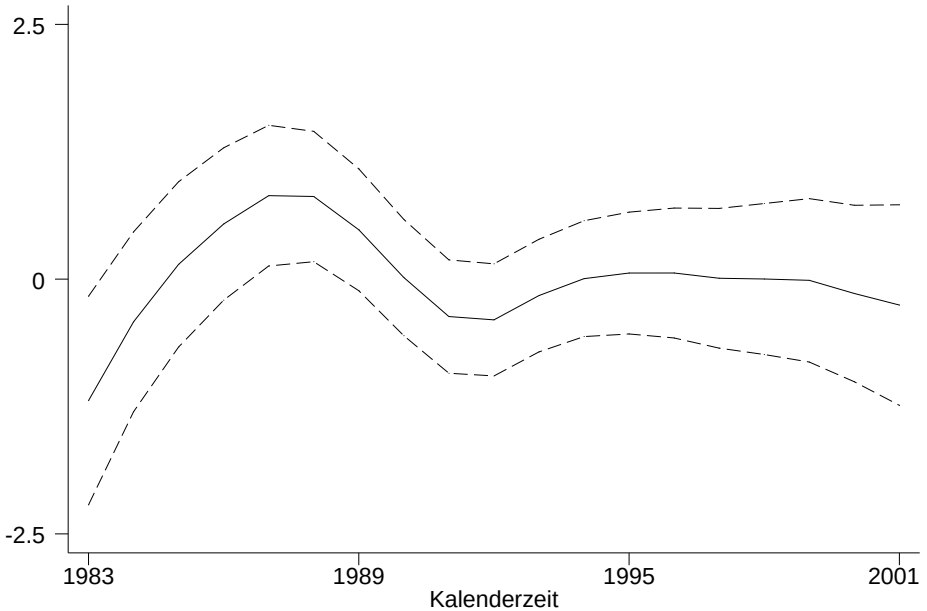
  $y_{it} = 1$ keine Entlaubung der Baumkrone,
  $y_{it} = 2$ weniger als 25% Entlaubung,
  $y_{it} = 3$ mehr als 25% Entlaubung.

- Kumulatives Probitmodell:

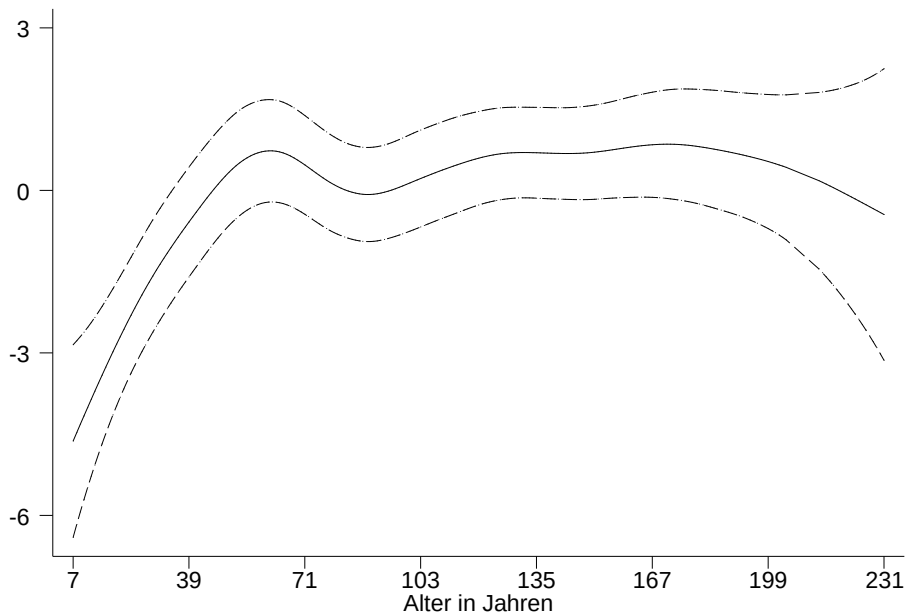
$$P(y_{it} \leq r) = \Phi\left[\theta_r - f_1(t) - f_2(alter_{it}) - f_3(t, alter_{it}) - f_{spat}(s_i) - u_{it}'\gamma\right]$$

  mit der Verteilungsfunktion der Standardnormalverteilung Φ und Schwellenwerten $-\infty = \theta_0 < \theta_1 < \theta_2 < \theta_3 = \infty$

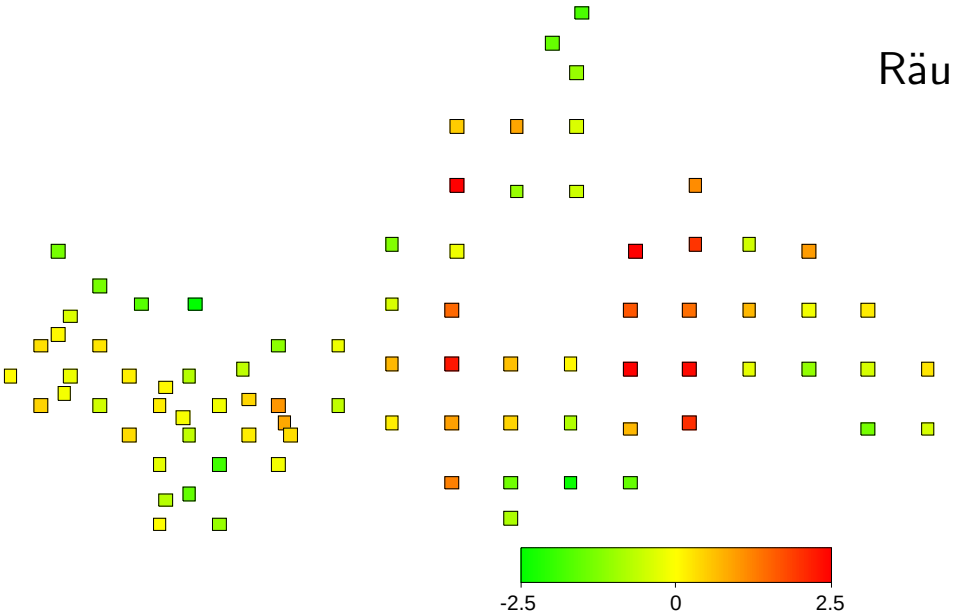
Geschätzter Trend



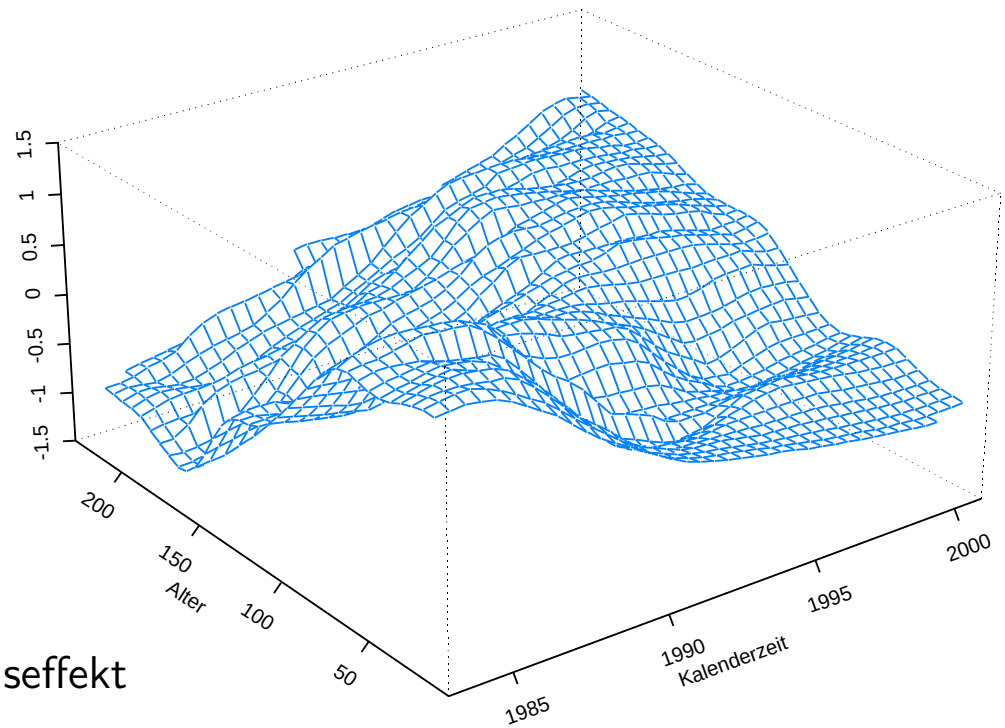
Effekt des Alters



Räumlicher Effekt



Interaktionseffekt



Strukturiert additive Regression

- Idee: Ersetze den üblichen parametrischen Prädiktor durch einen flexiblen semiparametrischen Prädiktor mit
  - nonparametrischen Effekten von Zeitskalen und stetigen Kovariablen (P-Splines),
  - räumlichen Effekten (Markov-Zufallsfelder, Kriging),
  - Interaktionsoberflächen (zweidimensionale P-Splines),
  - variierenden Koeffizienten (stetige und räumliche Effektmodifizierer),
  - zufälligen Effekten (Random Intercepts und Random Slopes).

- P-Splines:
  - Approximiere $f_j(x_j)$ durch große Zahl von B-Spline-Basisfunktionen.
  - Bestrafe Differenzen benachbarter Regressionskoeffizienten.
  - Alternativen: Random Walks oder allgemeinere autoregressive Modellierungen.
- Zweidimensionale P-Splines:
  - Definiere zweidimensionale B-Spline-Basisfunktionen (Tensorprodukte univariater B-Splines).
  - Bestrafung über Prioris aus der räumlichen Statistik.

- Markov Zufallsfelder:
  - Geeignet für Regionendaten.
  - Definiere Nachbarschaften für die Regionen.
  - Erwartungswert von $f_{spat}(s)$ ist der Mittelwert der Funktionsauswertungen benachbarter Regionen.
- Stationäre Gaussfelder (Kriging):
  - Geeignet für exakte Lokationen.
  - Der räumliche Effekt folgt einem stationären stochastischen Prozess.
  - Die Korrelation zweier Punkte wird durch eine intrinsische Korrelationsfunktion beschrieben.
- Eine bayesianische Formulierung des Modells erlaubt es, alle Effekte in einem einheitlichen Rahmen zu betrachten.

Schätzverfahren basierend auf gemischten Modellen

- Alle Effekte f_j können als Produkt einer Designmatrix Z_j und eines Vektors von Regressionskoeffizienten β_j geschrieben werden.

- Bayesianischer Ansatz: Prioris für die Regressionskoeffizienten.

- Generelle Form:

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

  K_j ist eine Strafmatrix und τ_j^2 ein Glättungsparameter.

- Im Allgemeinen besitzt K_j nicht vollen Rang.

$\Rightarrow$ Reparametrisierung des Modells in ein Varianzkomponentenmodell erlaubt die Anwendung von Methodik für gemischte Modelle.

- Zerlege

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

$$p(\beta_j^{unp}) \propto const \qquad \beta_j^{pen} \sim N(0, \tau_j^2 I)$$

$\Rightarrow$ Varianzkomponentenmodell:

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

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

$$\Lambda = \text{blockdiag}(\tau_j^2 I).$$

- Empirische Bayes-Schätzer können iterativ bestimmt werden:
  - Posteriori Modus-Schätzer über penalisierte Maximum Likelihood-Schätzung für die Regressionskoeffizienten.
  - Marginale Likelihood / Restricted Maximum Likelihood für die Varianzparameter:

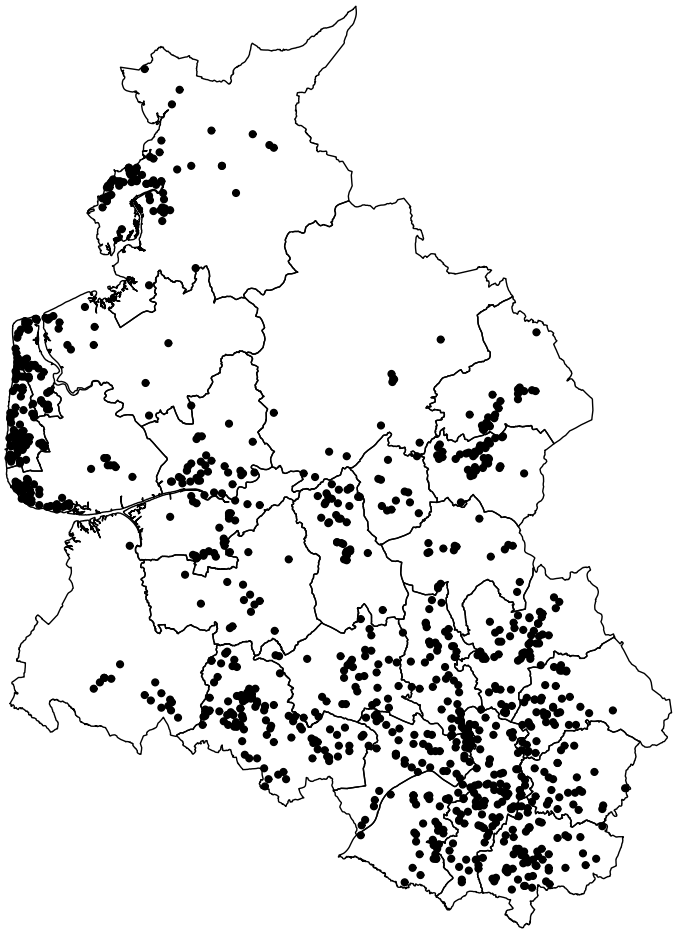
$$L(\Lambda) = \int L(\beta^{unp}, \beta^{pen}, \Lambda) p(\beta^{pen}) d\beta^{unp} d\beta^{pen} \longrightarrow \max_{\Lambda}.$$

Software

- Implementiert im Software-Paket BayesX.
- Möglich ist die Schätzung geoadditiver Regressionsmodelle für
  - univariate Exponentialfamilien (GLMs),
  - kategoriale abhängige Variablen mit geordneten und ungeordneten Kategorien,
  - stetige Überlebenszeiten.
- Erhältlich unter

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

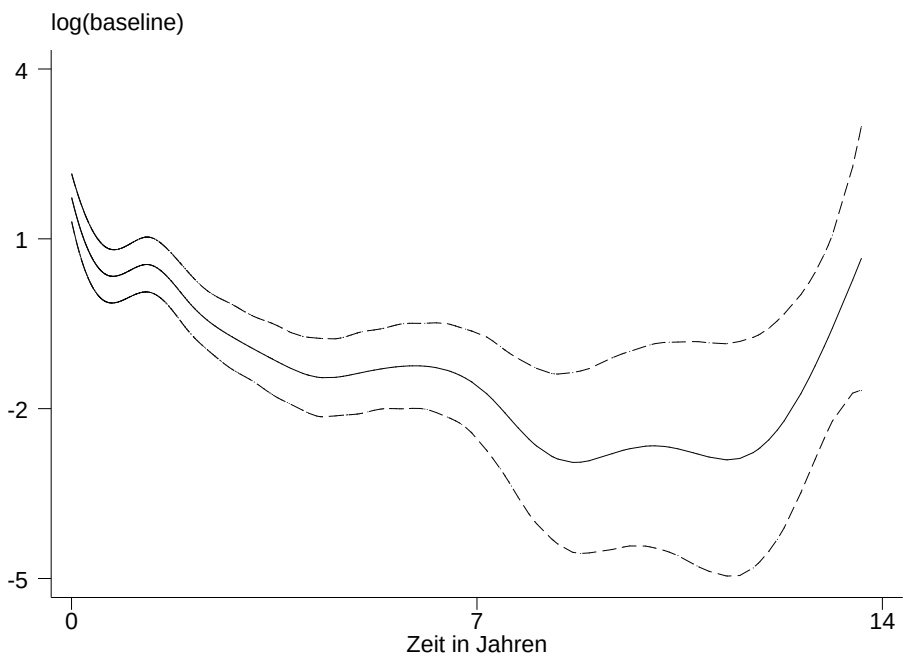
Beispiel II: Akute myeloische Leukämie



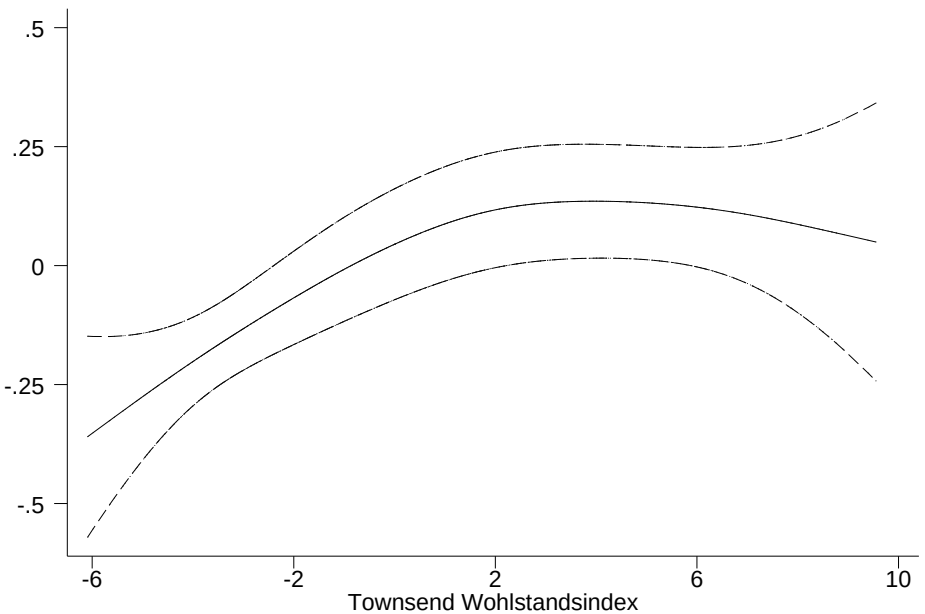
- Überlebenszeit von Erwachsenen nach der Diagnose akuter myeloischer Leukämie.
- 1.043 Fälle, diagnostiziert von 1982 bis 1998 in Nordwest England.
- Circa 13 % Rechtszensierung.
- Räumliche Information in verschiedenen Auflösungen.
- Modell für die Hazardrate:

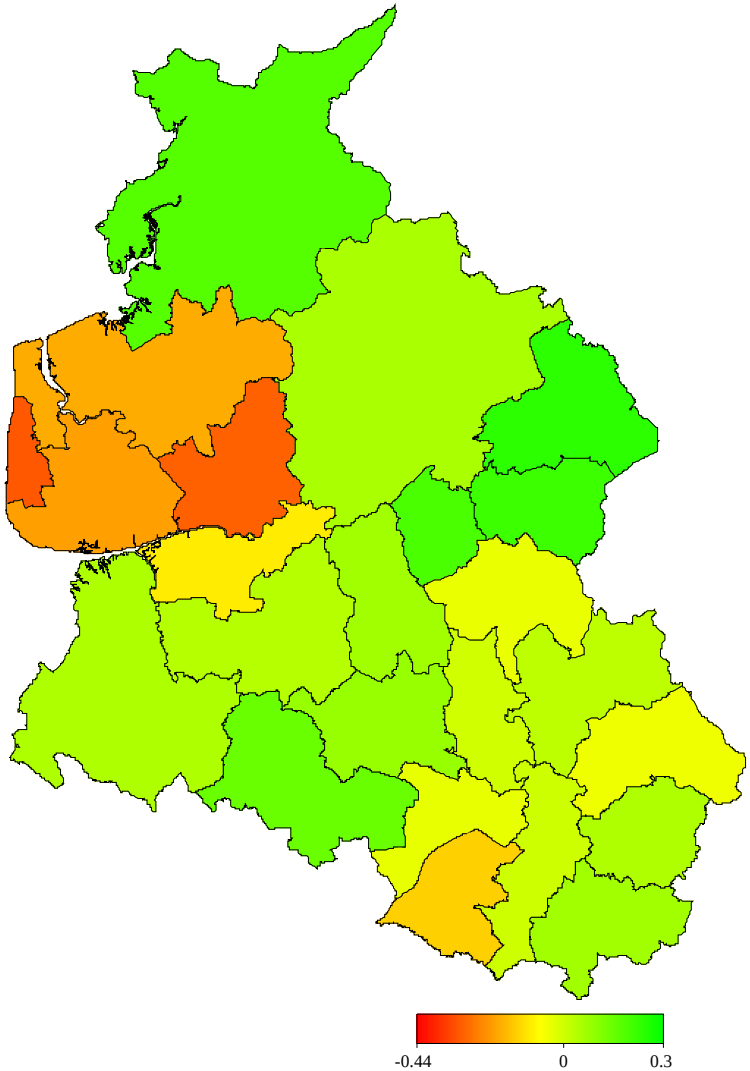
$$\lambda(t;\cdot) = \lambda_0(t)\exp[f_1(alter) + f_2(wb) + f_3(twi) + f_{spat}(s) + \gamma sex]$$

Logarithmierte Baseline-Hazardrate

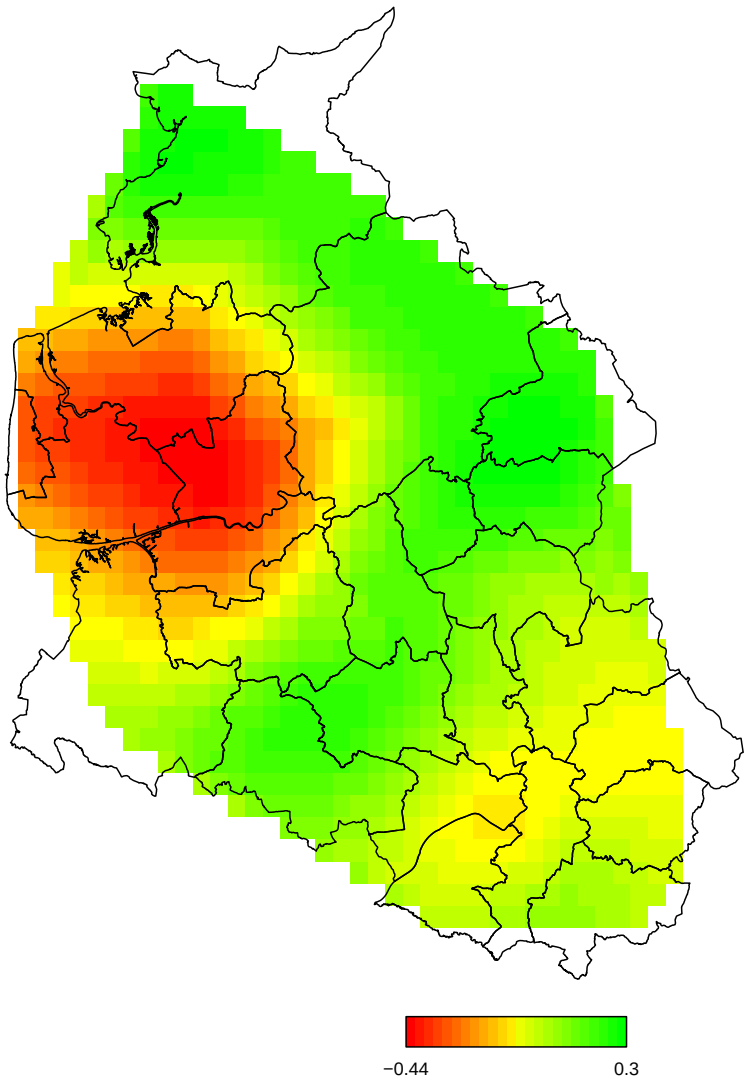


Effekt des Wohlstandsindex





Analyse basierend auf Distrikten



Analyse basierend auf Koordinaten

Diskussion

- Bayesianisches Modell zur Analyse komplexer, geoadditiver Regressionsmodelle ohne die Verwendung computerintensiver MCMC-Simulationstechniken.
- Eng verwandt mit penalisierter Likelihood-Schätzung in einem frequentistischen Kontext.
- Erweiterungsmöglichkeiten:
  - Allgemeinere Modelle für kategorialen Response, z.B. über korrelierte latente Variablen.
  - Allgemeinere Zensierungsmechanismen für Überlebenszeiten, z.B. Intervallzensierung oder Linkstrunkierung.
  - Anisotrope räumliche Effekte.
  - Dreidimensionale Erweiterungen von P-Splines.

Literatur

- Fahrmeir, L., Kneib, T. & Lang, S. (2004): Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14, 715-745.

- Kneib, T. & Fahrmeir, L. (2005): Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.

- Kneib, T. & Fahrmeir, L. (2004): A mixed model approach for structured hazard regression. SFB 386 Discussion Paper 400, University of Munich.

- Erhältlich unter

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