

BayesX: Bayesianische Inferenz in strukturiert additiven Regressionsmodellen

Thomas Kneib

Institut für Statistik
Ludwig-Maximilians-Universität München



1.12.2006



Outline

- Disease Mapping (Analyse räumlich korrelierter Zähldaten).
- BayesX (Vorstellung anhand exemplarischer Anwendungen).

Disease Mapping: Datenstruktur

- Analyse der **geografischen Variation** des Erkrankungs- oder Mortalitätsrisikos bezüglich ein Krankheit.
- Ziel: **Identifikation unbekannter Risikofaktoren** / Beurteilung von **Kovariableneffekten**.
- In der Regel ist die betrachtete Krankheit selten und nicht ansteckend.
- Interessierende Zielvariable: Anzahl der Todesfälle y_s aufgrund einer Krankheit in Regionen $s = 1, \dots, S$ (z.B. Bundesländern, Landkreisen, Gemeinden, etc.).
- Verteilung der Zielvariablen:

$$y_s \sim B(n_s, \pi_s), \quad s = 1, \dots, S.$$

n_s = Bevölkerungsgröße in Region s , π_s = Wahrscheinlichkeit an der Krankheit zu sterben. Die Wahrscheinlichkeit π_s darf über die Regionen hinweg variieren.

- Für **große Bevölkerungen** n_s und **kleine Wahrscheinlichkeiten** π_s kann die Binomialverteilung durch die Poisson-Verteilung approximiert werden:

$$y_s \sim P(\lambda_s)$$

mit $\lambda_s = n_s \pi_s$.

- Um die Raten λ_s vergleichbar zu machen, wird das **erwartete Risiko** e_s eingeführt.
- Beispiel: Sei π die bekannte, globale Mortalitätsrate bezüglich der interessierenden Krankheit. Dann ist das erwartete Risiko in Region s gegeben durch

$$e_s = n_s \pi.$$

- Daraus ergibt sich die Verteilung

$$y_s \sim P(e_s \lambda_s).$$

- e_s wird als **Offset** bezeichnet und kann auch weitere Information (z.B. über die Altersverteilung in den Regionen) enthalten.

Disease Mapping: Explorative Analysen

- Beispiel: Mortalität bezüglich Mundhöhlenkrebs in Deutschland zwischen 1985 und 1990.
- Beobachtete und erwartete Fallzahlen liegen auf Kreisebene vor.
- Zur deskriptiven Analyse nimmt man zunächst an, dass die Fallzahlen y_s unabhängig sind, also keine räumlichen Korrelationen vorliegen.

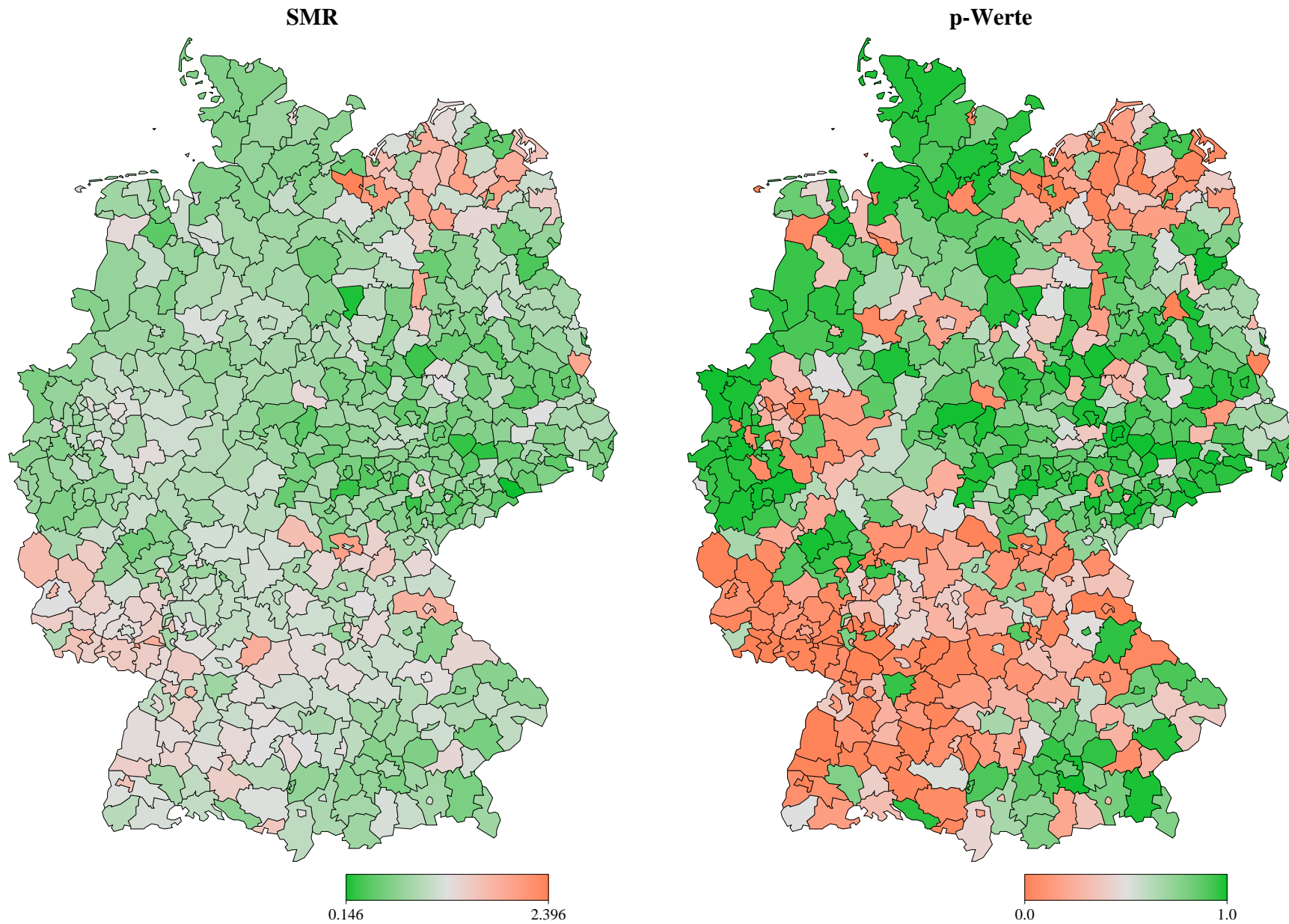
- Dann erhält man als Maximum-Likelihood-Schätzer für λ_s die **Standard-Mortalitätsraten**

$$\hat{\lambda}_{ML} = \frac{y_s}{e_s}.$$

- Alternativ können **p-Werte**

$$p_s = P(Y_s \geq y_s)$$

unter der Annahme $Y_s \sim P(e_s)$ bestimmt werden.



- Probleme der deskriptiven Ansätze:

- **Räumliche Korrelationen** werden nicht berücksichtigt. Die Standardfehler der geschätzten Regressionskoeffizienten stimmen nicht (in der Regel unterschätzt).
- Die Reliabilität der Aussagen hängt wesentlich von der Zahl erwarteter Todesfälle ab. Beispiel: Die Standardabweichung der Standardmortalitätsraten ist gegeben durch

$$\text{sd}(\hat{\lambda}_{ML}) = \frac{\sqrt{y_s}}{e_s}.$$

- Keine Berücksichtigung von **Kovariablen**.

⇒ **Regressionsmodelle mit räumlichen Effekten.**

Disease Mapping: Räumliche Regressionsmodelle

- Log-lineares Poisson-Modell:

$$E(y_s) = \exp(\eta_s),$$
$$\eta_s = x'_s \gamma = \gamma_0 + x_{s1} \gamma_1 + \dots + x_{sp} \gamma_p.$$

- Erweitere das Modell um den Offset, einen **räumlich korrelierten** Effekt β_s und einen **räumlich unkorrelierten** Effekt b_s :

$$\eta_s = \log(e_s) + x'_s \gamma + \beta_s + b_s.$$

- Idee: Unterscheide zwischen räumlich strukturierten und räumlich unstrukturierten Einflussgrößen.
- In realen Anwendungen teilweise schwierig zu trennen.

- Unstrukturierter räumlicher Effekt:

$$b_s \text{ i.i.d. } N(0, \tau_b^2)$$

(i.i.d. zufällige Effekte, Frailties).

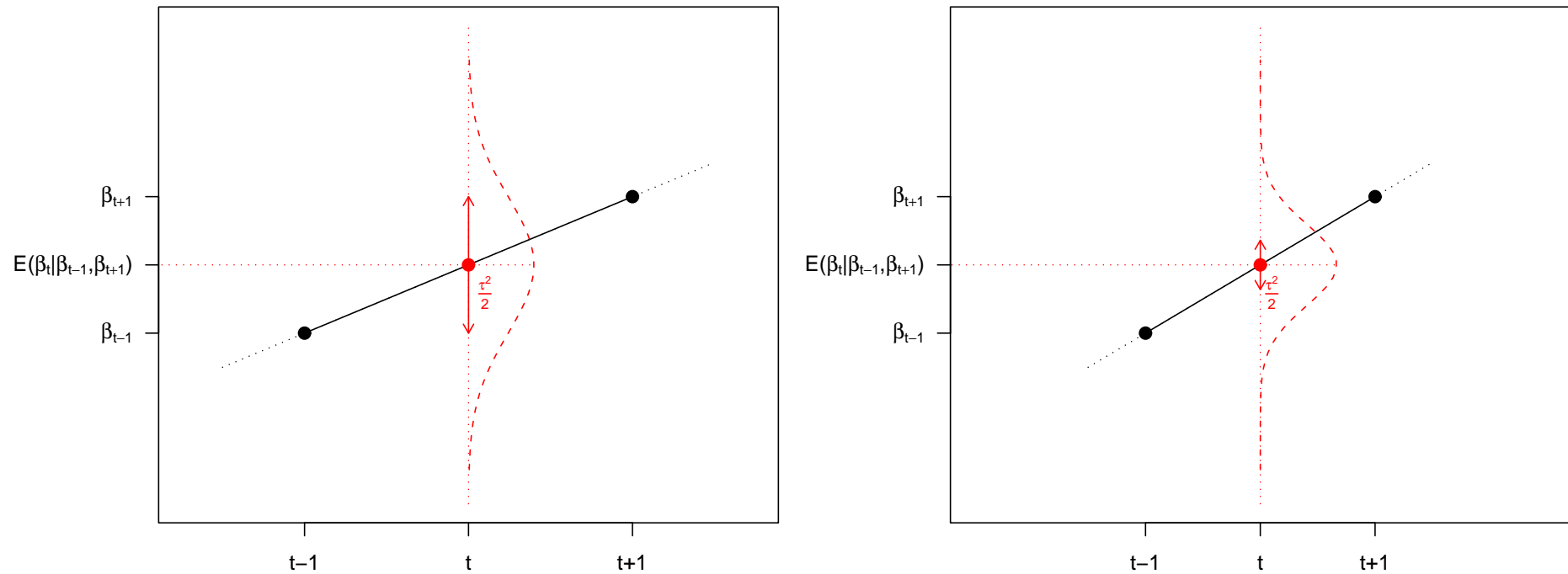
- Strukturierter räumlicher Effekt:

$$[\beta_s \mid \beta_1, \dots, \beta_{s-1}, \beta_{s+1}, \dots, \beta_S] \sim N \left(\frac{1}{N_s} \sum_{r \in \delta_s} \beta_r, \frac{\tau_\beta^2}{N_s} \right)$$

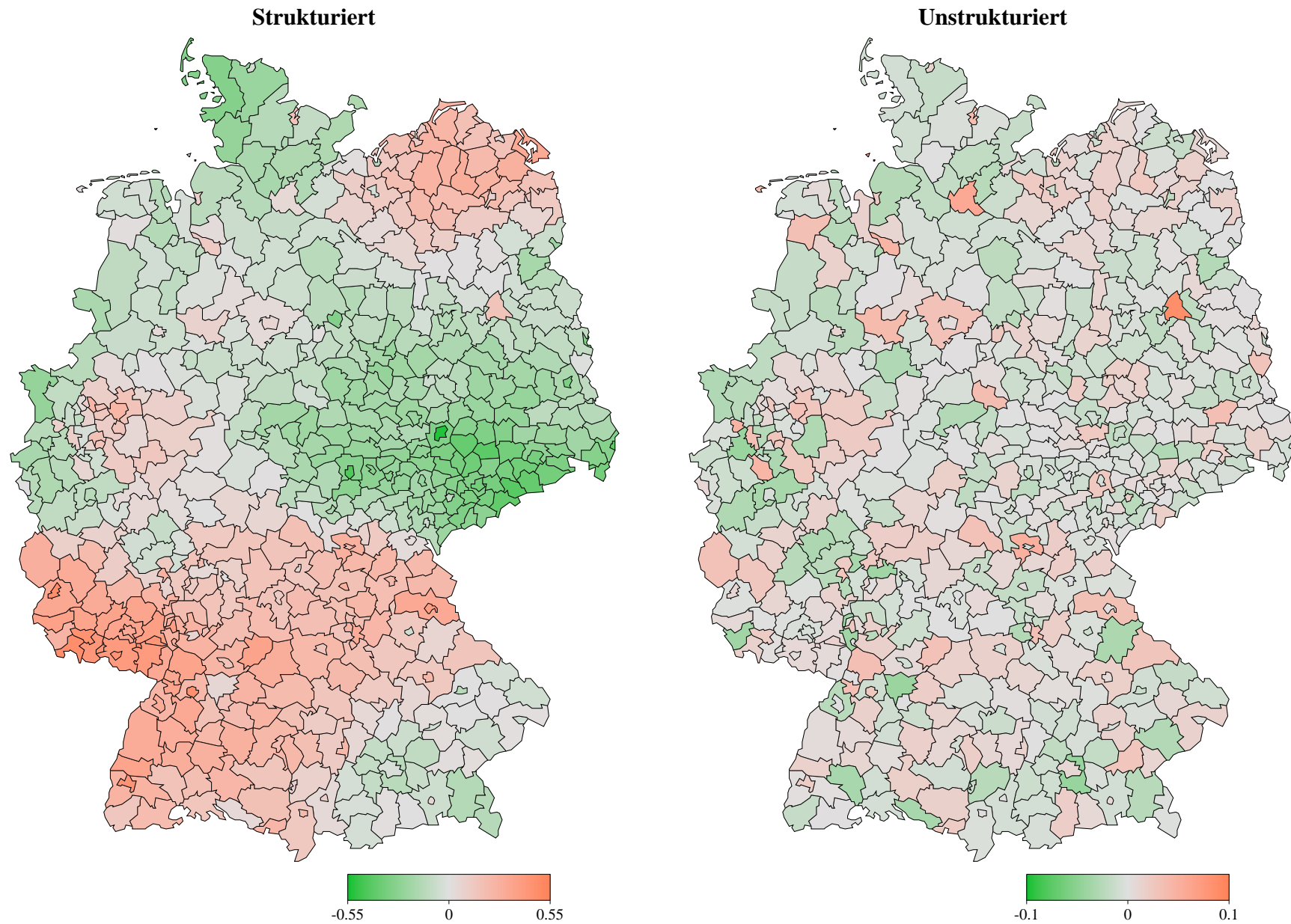
(Intrinsisches Gauß Markov Zufallsfeld).

- $\delta_s =$ Menge der Nachbarn von Region s , $N_s = |\delta_s| =$ Anzahl der Nachbarn.
- Der erwartete Effekt in Region s ist der Mittelwert aus den Effekten der benachbarten Regionen.
- Die Varianz des Effekts ist umgekehrt proportional zur Anzahl der Nachbarn.

- Verallgemeinert den zeitlichen Random Walk zu einem räumlichen Modell.



- Die Varianzen τ_b^2 und τ_β^2 können als **Glättungsparameter** interpretiert werden.



Strukturiert additive Regression

- Gemeinsame Struktur vieler Regressionsmodelle: Charakteristika der Verteilung der Zielvariablen werden in Abhängigkeit eines **linearen Prädiktors** $u'\gamma$ geschrieben, z.B.

$$\mathbb{E}(y|u) = h(u'\gamma) \quad (\text{Generalisierte lineare Modelle})$$

oder

$$\lambda(t|u) = \lambda_0(t) \exp(u'\gamma) \quad (\text{Cox Modell})$$

- **Strukturiert additive Prädiktoren** ergeben sich aus beliebigen Kombinationen von
 - Nonparametrischen Effekten metrischer Kovariablen und Zeitskalen,
 - Räumliche Effekte,
 - Interaktionsoberflächen,
 - Zufälliger Effekte (Random Intercepts und Random Slopes),
 - und einer Reihe möglicher Erweiterungen.

- Beispiel:

$$\mathbb{E}(y_{it}|\cdot) = h \left[u'_{it}\gamma + f_{trend}(t) + f_1(x_{it1}) + f_2(x_{it2}) + f_{1|2}(x_{it1}, x_{it2}) + f_{spat}(s_{it}) + b_i \right]$$

- Generische Repräsentation:

$$\eta = u'\gamma + f_1(z_1) + \dots + f_p(z_p)$$

mit verschiedenen Typen von Funktionen f und generischen Kovariablen z .

- Diese Repräsentation erleichtert die Beschreibung der Inferenzverfahren.
- Der Vektor der Funktionsauswertungen f_j kann immer geschrieben werden als

$$f_j = Z_j \xi_j$$

mit Designmatrix Z_j und Regressionskoeffizienten ξ_j .

- **Generische Form** der Priori-Verteilung für ξ_j :

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

- $K_j \geq 0$ ist eine **Strafmatrix** mit $\text{rank}(K_j) = k_j \leq d_j = \dim(\xi_j)$.
- $\tau_j^2 \geq 0$ lässt sich als **Varianz-** oder (inverser) **Glättungsparameter** interpretieren.
- Enge Verbindung zu **penalisierten Likelihood-Ansätzen**. Strafterme ergeben sich aus der Log-Priori:

$$P_{\lambda_j}(\xi_j) = \log[p(\xi_j|\tau_j^2)] = -\frac{1}{2}\lambda_j\xi_j'K_j\xi_j, \quad \lambda_j = \frac{1}{\tau_j^2}.$$

- **Volle Bayes-Inferenz:**
 - Alle Parameter (insbesondere auch die Varianzen τ^2) sind zufällig und werden mit Prioris versehen.
 - Schätzung üblicherweise über **MCMC-Verfahren**.
 - Übliche Schätzer: **Posteriori-Erwartungswert**, Posteriori-Median.
- **Empirische Bayes-Schätzung:**
 - Unterscheide zwischen **primär interessierenden Parametern** (den Regressionskoeffizienten) und **Hyperparametern** (den Varianzen).
 - Nur die primär interessierenden Parameter werden mit Prioris versehen.
 - Schätze die Hyperparameter durch Maximierung der **marginalen Posteriori-Verteilung**.
 - Einsetzen dieser Schätzer in die gemeinsame Posteriori und Maximierung bezüglich der primär interessierenden Parameter ergibt **Posteriori-Modus-Schätzer**.

Software

- BayesX - Software für empirische und volle Bayes-Inferenz in strukturiert additiven Regressionsmodellen.



- Erhältlich unter

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

- Entwickelt gemeinsam mit Andreas Brezger (HVB, München) und Stefan Lang (Leopold-Franzens-Universität, Innsbruck).
- Beiträge von sieben weiteren Helfern.
- Computerintensive Teile in C++ implementiert.
- Grafische Benutzeroberfläche in Java.
- Anbindung an R ist in Arbeit.

KFZ-Versicherung in Belgien

- Versicherungsdaten aus zwei belgischen KFZ-Versicherungen.
- Stichprobe bestehend aus ca. 160.000 Versicherungsnehmern.
- Ziel: Separate Modelle für die Schadenshöhen und die Schadenshäufigkeiten um Prämien basierend auf Kovariablen des Versicherungsnehmers kalkulieren zu können.
- Primär interessierende Größen: Schadenshöhen y_i und Schadenshäufigkeiten h_i eines Versicherungsnehmers.
- **Kovariablen:**
 - vage* Alter des Fahrzeugs
 - page* Alter des Versicherungsnehmers
 - hp* Pferdestärke des Fahrzeugs
 - bm* Bonus-Malus Score
 - s* Distrikt in Belgien
 - v* Vektor weiterer (kategorialer) Kovariablen.

- **Geoadditive Modelle:**

- Normalverteilungsmodell für die logarithmierten Schadenshöhen $\log(y)$:

$$\log(y) \sim N(\eta, \sigma^2)$$

mit

$$\eta = f_1(vage) + f_2(page) + f_3(bm) + f_4(hp) + f_{spat}(s) + v'\zeta.$$

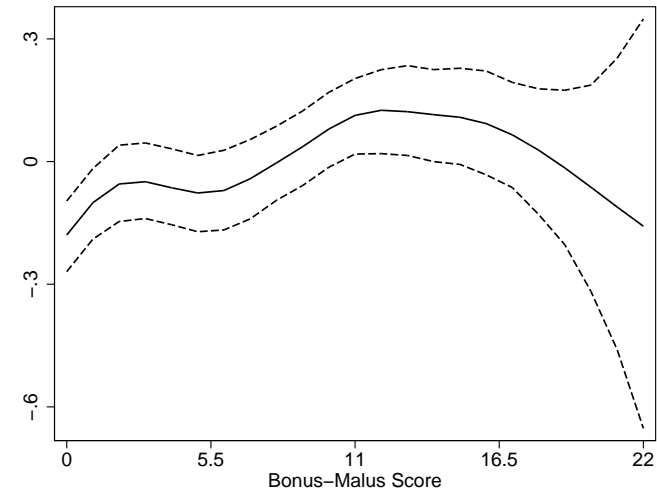
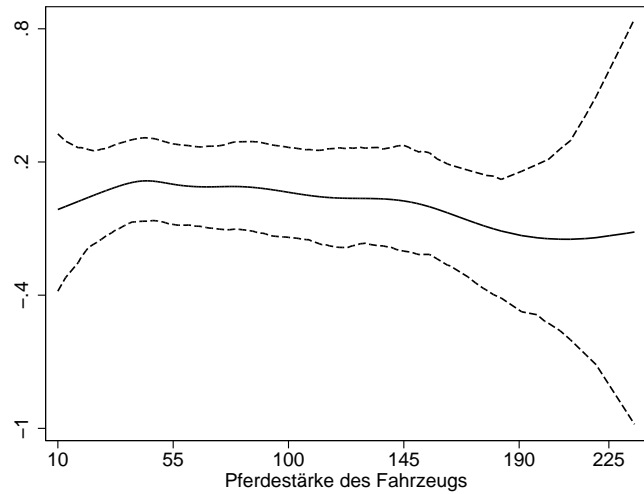
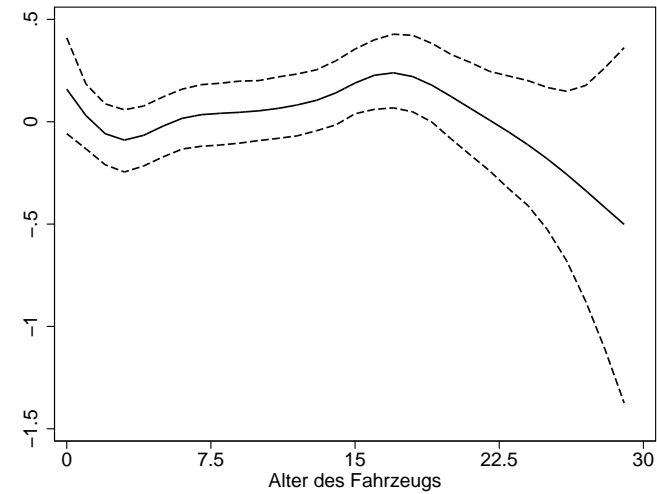
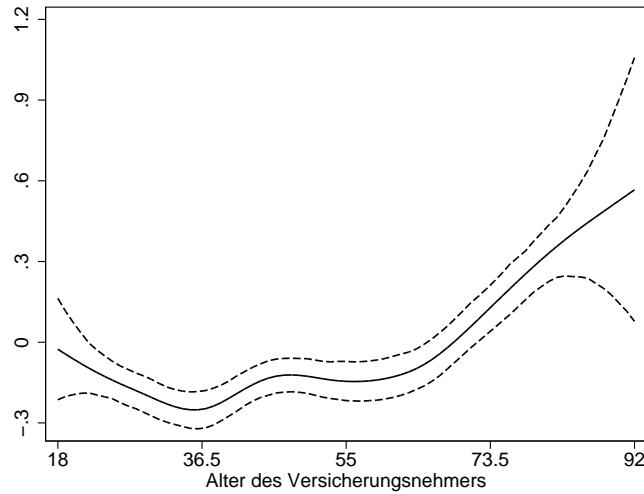
- Poisson Modell für die Schadenshäufigkeiten h_i :

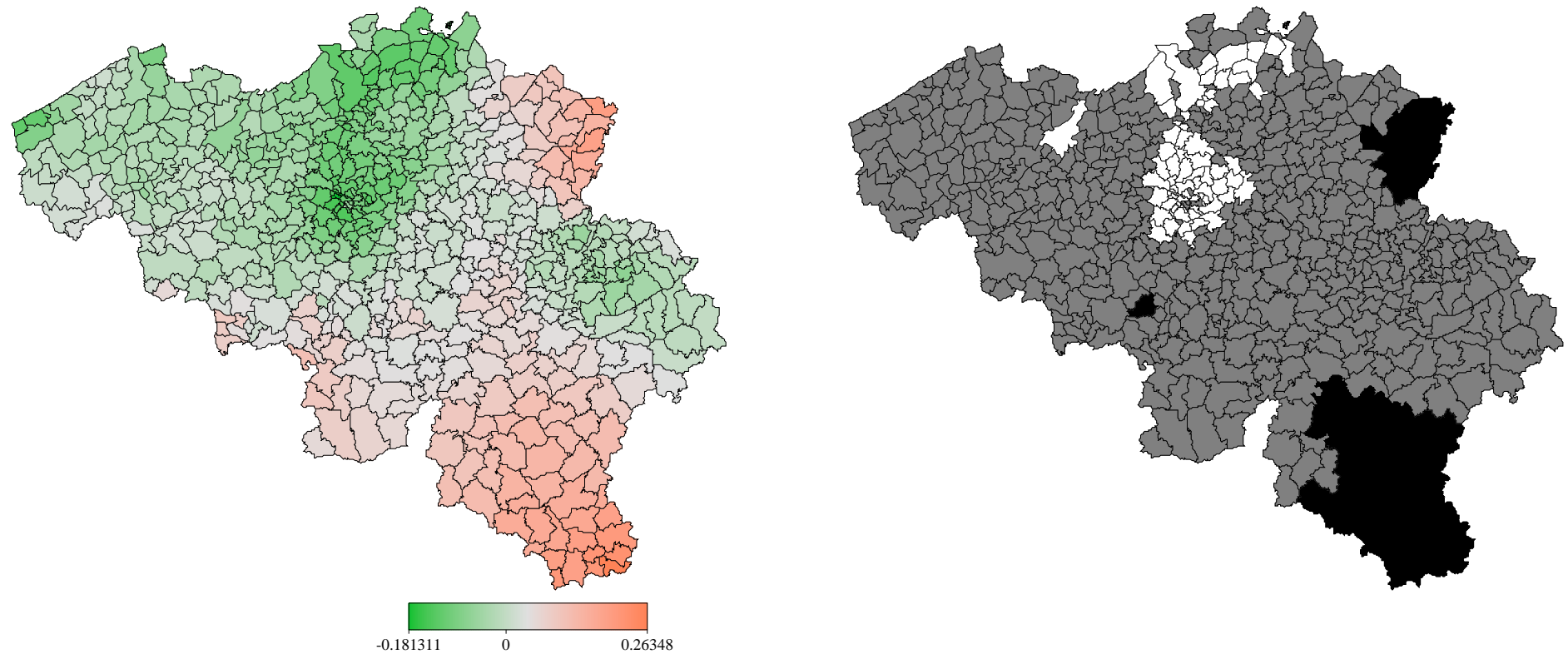
$$h \sim Po(\exp(\eta))$$

mit

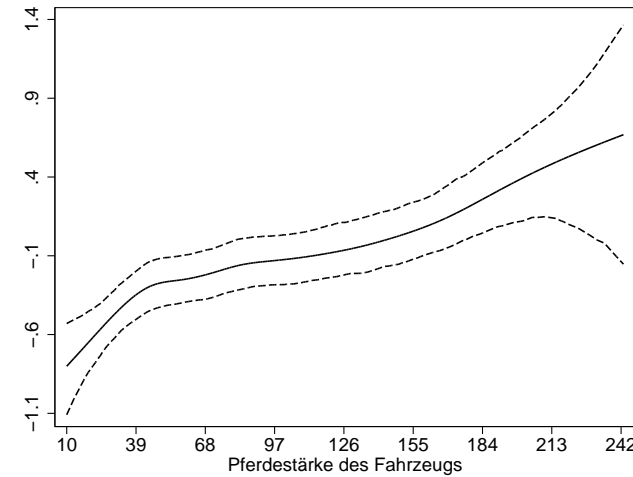
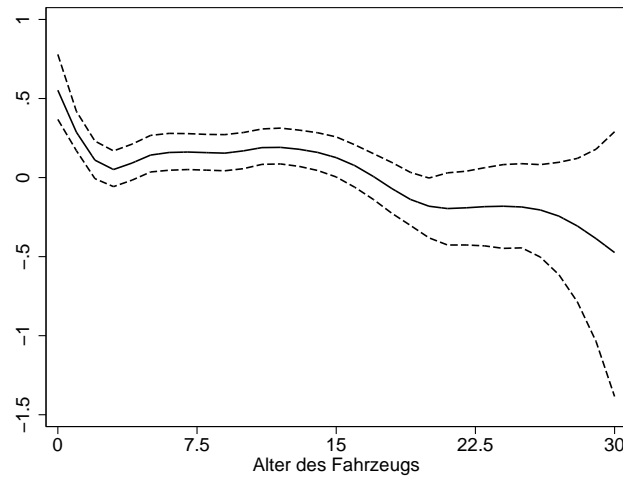
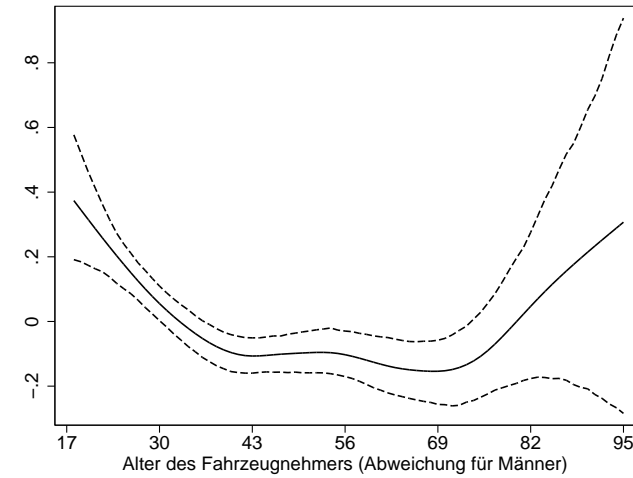
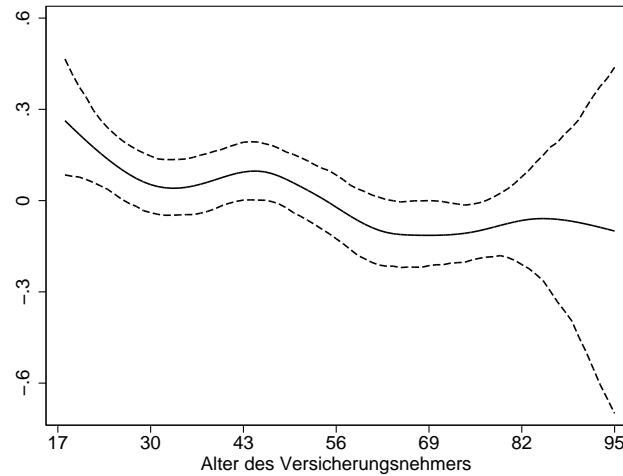
$$\eta = f_1(vage) + f_2(page) + f_3(page)sex + f_3(bm) + f_4(hp) + f_{spat}(s) + v'\zeta.$$

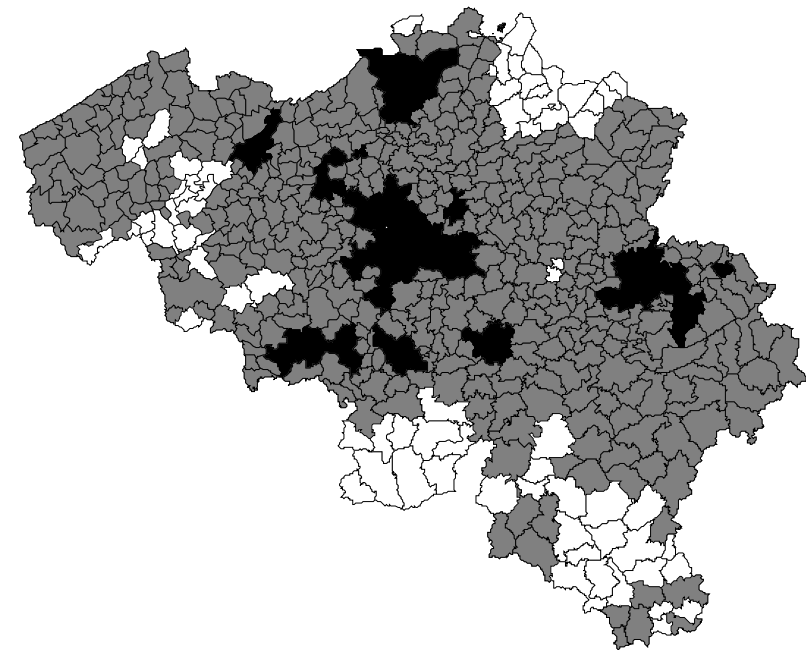
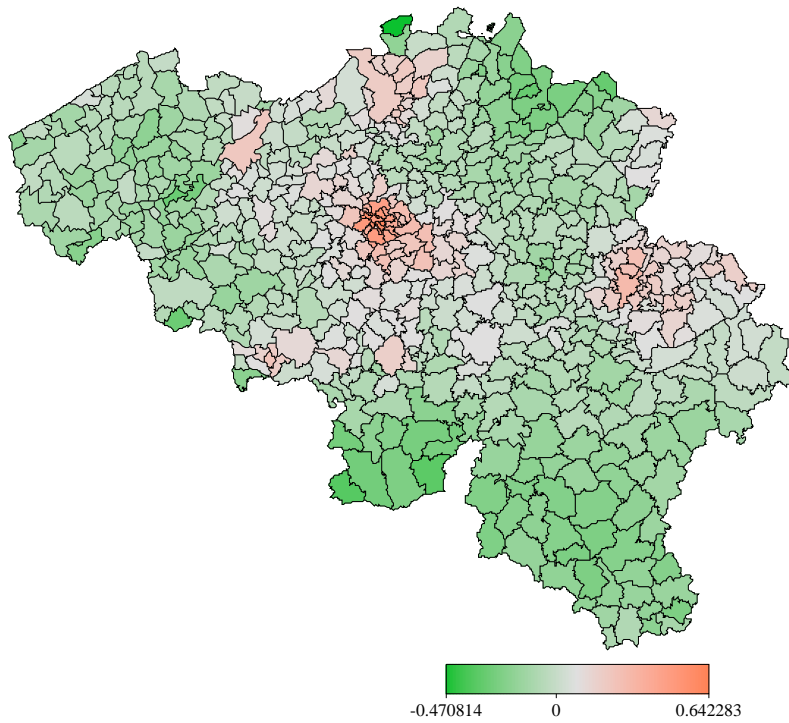
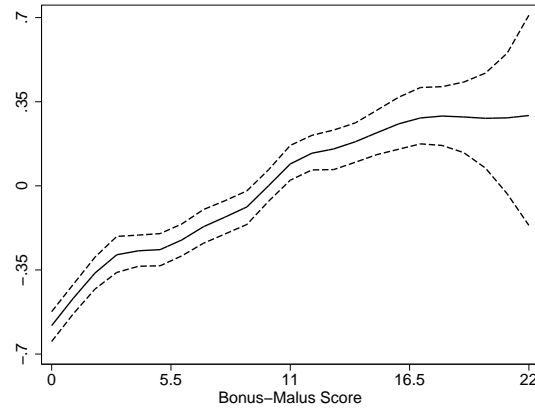
● Ergebnisse für die Schadenshöhen:





- Ergebnisse für die Schadenshäufigkeiten:





- **Univariate Exponentialfamilien** die in BayesX für die Zielvariable zugelassen sind:
 - Normalverteilung,
 - Bernoulli- und Binomialverteilung,
 - Poissonverteilung and Zero-Inflated Poissonverteilung,
 - Gammaverteilung,
 - Negativ-Binomialverteilung.
- **Kategoriale Zielvariablen:**
 - Kumulative und sequentielle Modelle für geordnete Kategorien.
 - Multinomiales Logit- und Probitmodell für ungeordnete Kategorien.
 - Kategorienspezifische Effekte / Kovariablen.

- Analyse **stetiger Überlebenszeiten**:
 - Modelle vom Cox-Typ zur Modellierung des Hazardrate.
 - Simultane Schätzung der Baseline-Hazardrate und der Kovariableneffekte.
 - Zeitvariierende Effekte und Kovariablen.
 - Beliebige Kombinationen von Rechts-, Links- und Intervallzensierung sowie Linkstrunkierung.
- Neueste Erweiterung: **Multi-State-Modelle**.

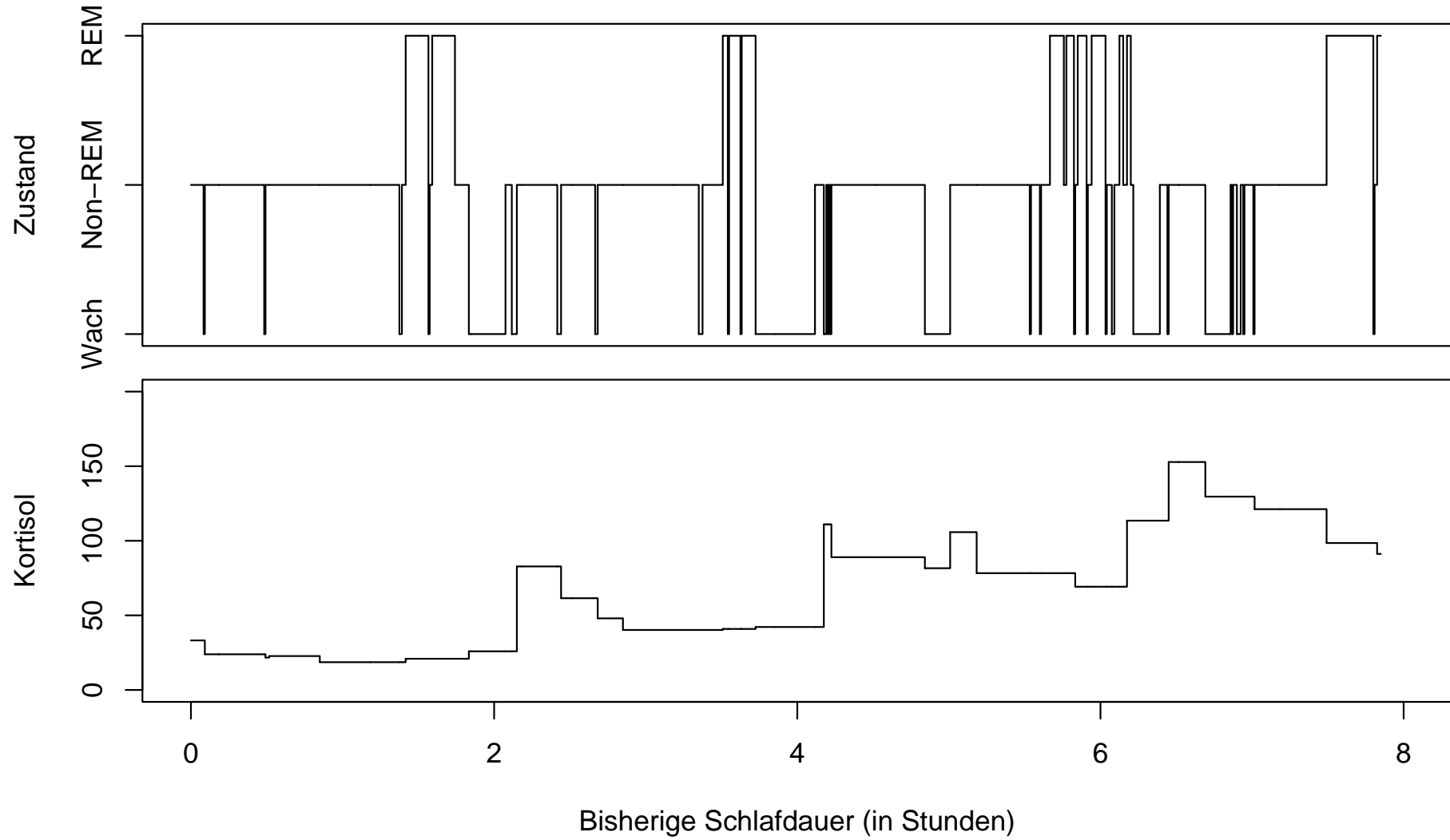
Analyse des menschlichen Schlafs: Multi-State-Modelle

- Multi-State-Modelle definieren eine allgemeine Klasse statistischer Modelle zur Beschreibung der **Entwicklung diskreter Phänomene in stetiger Zeit**.

- Beispiel: Der menschliche Schlafverlauf mit den Zuständen

Wach	Wachphasen
REM	Rapid Eye Movement Phasen (Traumphasen)
Non-REM	Non-REM Phasen (kann genauer unterteilt werden)

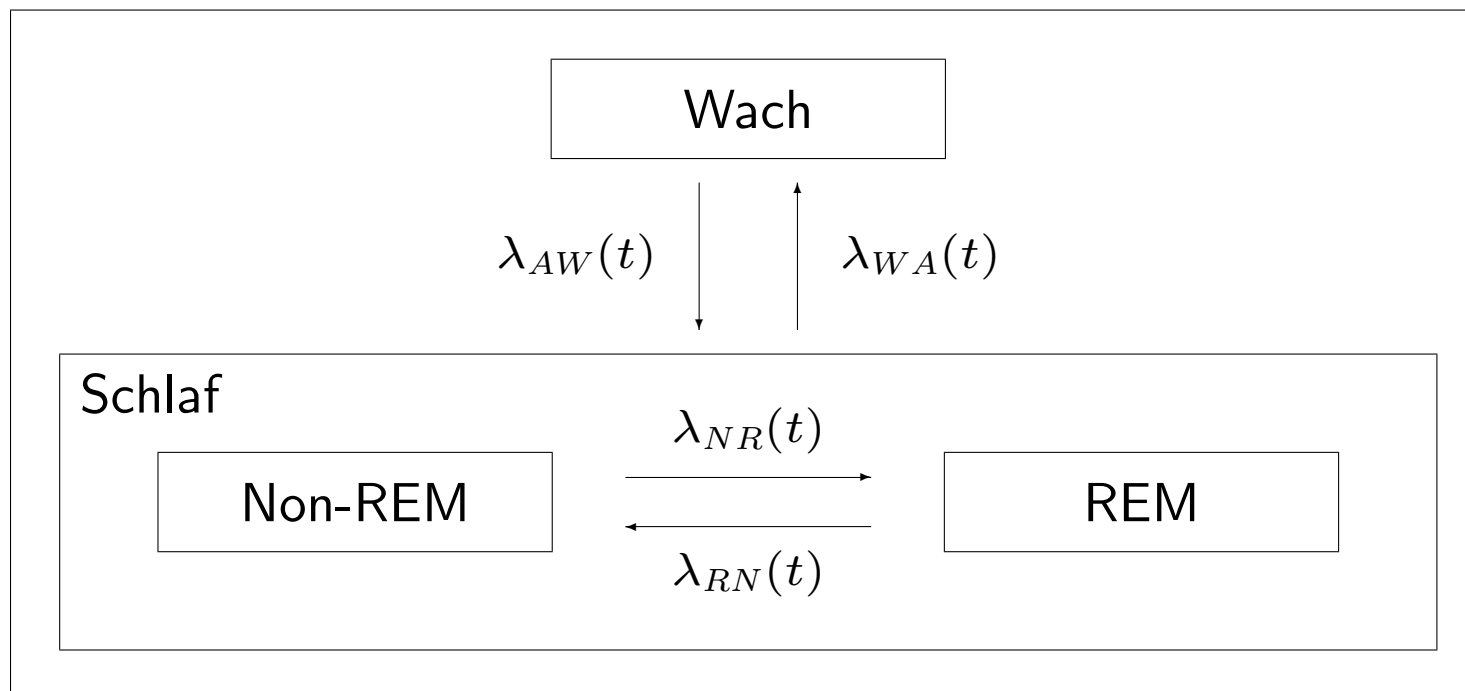
- Mögliche Ziele der Analyse:
 - Beschreibung der Dynamik des menschlichen Schlafs.
 - Analyse des Zusammenhangs zwischen nächtlicher Hormonausschüttung und Schlafverlauf.
 - Vergleich des Schlafverlaufs kranker und gesunder Personen.



- **Datenerhebung:**
 - Rohdaten: Elektroenzephalographische (EEG) Messungen im Abstand von 30 Sekunden (anschließend klassifiziert in die drei Schlaftypen).
 - Messung der Hormonausschüttung durch Blutproben im Abstand von ca. 10 Minuten.
 - Eine Probenacht, um die Teilnehmer der Studie mit den Schlafbedingungen vertraut zu machen.
- Die Datenstruktur von Multi-State Modellen ähnelt der von **Markov-Prozessen**.
- Solche einfachen, parametrischen Modelle sind hier nicht ausreichend, da
 - sich die Übergangintensitäten zwischen Schlafzuständen über die Nacht verändern.
 - Individuelles Schlafverhalten in Form von Kovariablen berücksichtigt werden muss.
 - zusätzlich unbeobachtete Heterogenität modelliert werden muss (nur wenige Kovariablen erhoben).

⇒ **Semiparametrisches Modell für die Übergangsintensitäten.**

- Mathematisch elegant lassen sich semiparametrische Multi-State Modelle im Rahmen von Zählprozessen behandeln.
- Vereinfachtes Modell für die Übergänge:



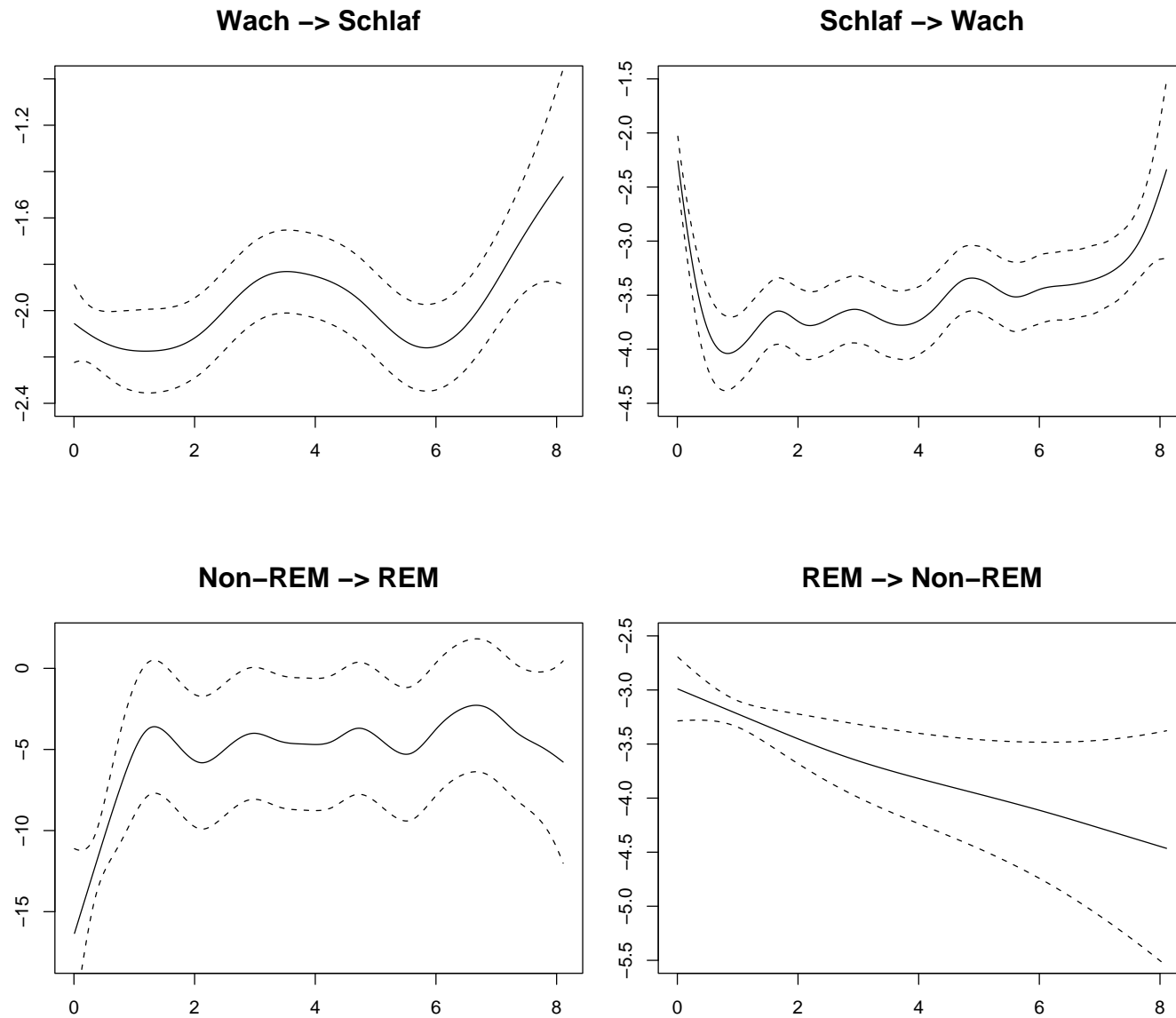
- Spezifikation der Übergangintensitäten:

$$\begin{aligned}\lambda_{AW,i}(t) &= \exp \left[\gamma_0^{(AW)}(t) + s_i \beta^{(AW)} + b_i^{(AW)} \right] \\ \lambda_{WA,i}(t) &= \exp \left[\gamma_0^{(WA)}(t) + s_i \beta^{(WA)} + b_i^{(WA)} \right] \\ \lambda_{NR,i}(t) &= \exp \left[\gamma_0^{(NR)}(t) + c_i(t) \gamma_1^{(NR)}(t) + s_i \beta^{(NR)} + b_i^{(NR)} \right] \\ \lambda_{RN,i}(t) &= \exp \left[\gamma_0^{(RN)}(t) + c_i(t) \gamma_1^{(RN)}(t) + s_i \beta^{(RN)} + b_i^{(RN)} \right]\end{aligned}$$

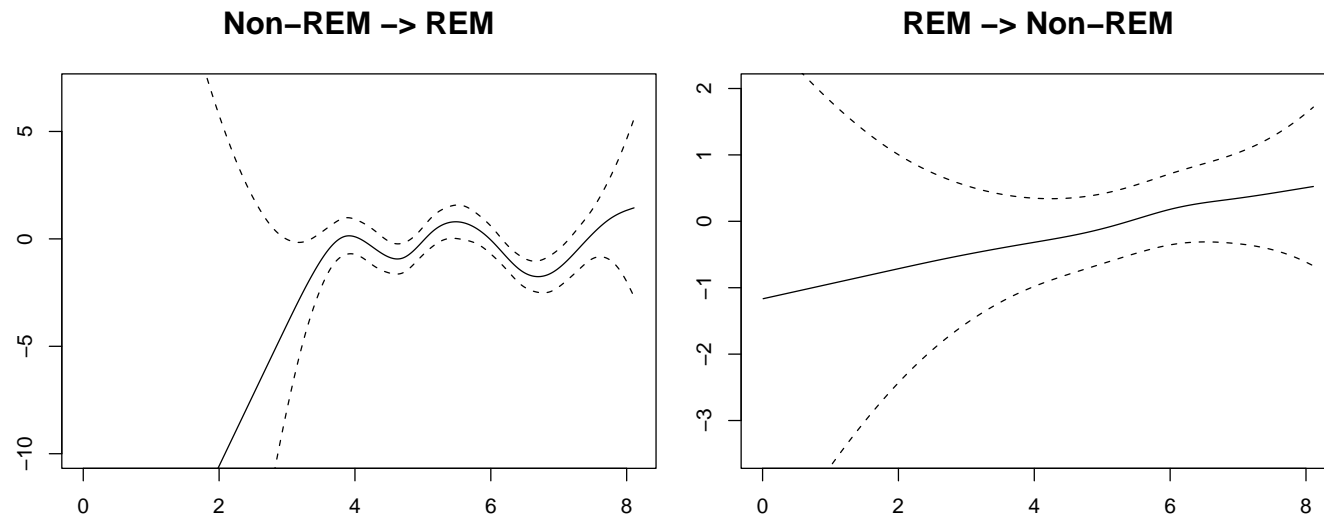
wobei

$$\begin{aligned}c_i(t) &= \begin{cases} 1 & \text{Kortisol} > 60 \text{ n mol/l zum Zeitpunkt } t \\ 0 & \text{Kortisol} \leq 60 \text{ n mol/l zum Zeitpunkt } t, \end{cases} \\ s_i &= \begin{cases} 1 & \text{männlich} \\ 0 & \text{weiblich,} \end{cases} \\ b_i^{(j)} &= \text{übergangs- und individuen-spezifische zufällige Effekte.}\end{aligned}$$

● Baseline Effekte:



- Zeitvariierende Effekte für hohen Kortisolspiegel:



Diskussion

- Strukturiert additive Regression:
 - Flexible Erweiterung bekannter parametrischer Modellklassen.
 - Nonparametrische Funktionsschätzung.
 - Räumliche Effekte.
 - Zufällige Effekte.
 - Interaktionsoberflächen.
- Univariate und kategoriale Zielvariablen, Überlebenszeiten, Multi-State-Modelle.
- Benutzerfreundliche Schätzung in BayesX.

- Geplante Erweiterungen:
 - Intervallzensurierung bei Multi-State-Modellen.
 - Bayesianische Regularisierungs-Prioris (ähnlich wie LASSO).
 - Bayesianische Messfehler-Behandlung in strukturiert additiven Regressionsmodellen.

Referenzen

- BREZGER, KNEIB & LANG (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).
- BREZGER, A. & LANG, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* **50**, 967–991.
- HENNERFEIND, A., BREZGER, A. & FAHRMEIR, L. (2006) Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065-1075.
- FAHRMEIR, L., KNEIB, T. & LANG, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.
- FAHRMEIR, L. & LANG, S. (2001a). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society C* **50**, 201–220.

- FAHRMEIR, L. & LANG, S. (2001b) Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics* **53**, 10–30.
- KNEIB, T. (2006). Geoadditive hazard regression for interval censored survival times. *Computational Statistics and Data Analysis*, **51**, 777–792.
- KNEIB, T. & FAHRMEIR, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics* **62**, 109–118.
- KNEIB, T. & FAHRMEIR, L. (2006). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, to appear.
- KNEIB, T. & HENNERFEIND, A. (2006) Bayesian semiparametric multi-state models. SFB 386 Discussion Paper 502.
- LANG, S. & BREZGER, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.