Hornung, Roman:

Analyse von Wildunfalldaten mit Hilfe räumlicher Poissonprozesse

Masterarbeit, Wintersemester 2011

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.38468

Analyse von Wildunfalldaten mit Hilfe räumlicher Poissonprozesse

Masterarbeit von **Roman Hornung**

(korrigierte Version)

Betreuer:

Prof. Dr. Torsten Hothorn

Dipl. Stat. Monia Mahling

Institut für Statistik

Ludwig–Maximilians–Universität München

12. Dezember 2011

Inhaltsverzeichnis

1. Einleitung **1**

2. Theorie **4**

2.1. Allgemeines zu räumlichen Punktprozessen 4

2.2. Räumliche Poissonprozesse 5

2.3. Schätzung mit Berman-Turner-Device 6

2.4. Generalisierte lineare Modelle 9

2.5. Generalisierte additive Modelle mit Regressionssplines 13

2.6. Maßzahlen 22

2.6.1. K-Funktion 23

2.6.2. L-Funktion 26

3. Datenmaterial **27**

3.1. Repräsentation räumlicher Daten und Koordinatensysteme 27

3.1.1. Raster- und Vektor-Daten-Modell 27

3.1.2. Kartenabbildungen und Koordinatensysteme 28

3.2. Wildunfalldaten 30

3.3. Daten zum Wildverbiss 30

3.4. CORINE Land Cover-Daten 33

3.5. Straßen-Daten 38

3.6. Kurvigkeit 41

3.7. Straßenlänge 43

3.8. Nachtlichter 45

4. Analyse **48**

4.1. Modellierung als räumliche Poissonprozesse 48

4.1.1. Dummypunkte auf regulärem Gitter 51

4.1.2. Dummypunkte auf den Straßen 79

4.2. Modellierung durch logistisches Regressionsmodell 96

4.3. Modellierung mit negativer Binomialverteilung 108

4.4. Analyse der Zeitpunkte 125

5. Zusammenfassung, Ausblick und Vergleich mit anderen Ergebnissen **130**

Literaturverzeichnis **137**

Abbildungsverzeichnis **141**

A. Nachweis approximativer Äquivalenz von räumlichem Poissonprozess mit Dummypunkten auf den Straßen und binärem (logistischem) Modell **145**

1. Einleitung

Jedes Jahr geschehen in Deutschland ca. 200.000 Autounfälle durch Kollisionen mit Rehen. Dies entspricht fast 20% der Anzahl der in der Jagd erlegten Rehen und führt zu ungefähr 3000 Verletzten, darunter 50 Todesfällen, bei Kosten von ca. 490 Millionen Euro. Mit wachsendem Verkehrsaufkommen werden diese Zahlen in den nächsten Jahren voraussichtlich weiter steigen. Bis auf Abzäunungen zeigten die meisten Maßnahmen zur Verringerung der Wildunfallzahl wenig Effekt (Hothorn et al., in Vorb.).

In dieser Arbeit soll nun die Intensität von Unfällen mit Rehen in Bayern mit ausgewählten räumlichen Einflußgrößen in Verbindung gesetzt werden. Im Weiteren wird häufig anstatt von Rehunfällen auch verallgemeinernd von Wildunfällen gesprochen. Während die akkurate Prädiktion der Intensität der Unfälle eher sekundär ist, besteht das Hauptziel in dieser Analyse darin, die Form des Einflusses der Kovariablen zu ergründen. Die betrachteten Kovariablen sind: Wildverbissintensität, Landnutzung, Straßentyp, Kurvigkeit der Straßen, Straßenlänge, Nachtlichter und die geografische Lokation. Unter diesen nimmt Wildverbiss eine besondere Stellung ein. Die Verbissintensität steht in engem Zusammenhang mit der Wilddichte. Auf Letztere kann durch Änderungen der Jagdquoten aktiv Einfluss genommen werden. Demnach bietet Wildverbiss den einzigen Angriffspunkt, um die Anzahl an Wildunfällen zu reduzieren (Hothorn et al., in Vorb.).

Die Analyse geschieht zu einem großen Teil mit Hilfe von räumlichen Poisson-Punktprozessen unter Einschluss von Kovariablen. Dabei zeigten sich teilweise große Diskrepanzen, je nachdem, ob die Schätzung auf Bereiche um die Straßen konzentriert

wurde, oder ob unrealistischerweise angenommen wurde, dass sich die Unfälle überall ereignen können. Erstere Variante ist approximativ gleich einem logistischen Regressionsmodell, wie in der Arbeit nachgewiesen und anhand der Wildunfalldaten aufgezeigt wird. Der Vorteil bei dieser Herangehensweise ist, dass sich die Schätzung besonders stark auf die Straßen konzentriert. Zum Vergleich wird auch ein gewöhnliches generalisiertes additives Modell (GAM) verwendet, wobei die Verteilungsannahme für die Zielgröße dabei - die Anzahl an Wildunfällen auf Quadraten der Seitenlänge 1225 m - die negative Binomialverteilung ist und als Offset die logarithmierte Straßenlänge aufgenommen wird. Bei letzterer Herangehensweise findet zwar auch eine Aggregierung statt, jedoch sind die Flächen so klein, dass man ohne nennenswerten Informationsverlust annehmen kann, dass die Intensität in diesem Bereich konstant ist. Neben der direkten Modellierung der Anzahlen an Wildunfällen pro Straßenmeter unterscheidet sich dieses Modell auch darin, dass durch die Aggregierung kategoriale Kovariablen wie die Landnutzung auch als Anteile berücksichtigt werden können, sodass eine andere Art Information übergeben wird.

Ein weiterer Teil beschäftigt sich mit der Schätzung der zeitlichen Intensität von Rehunfällen unter Annahme eines zeitlichen bzw. eindimensionalen Poisson-Punktprozesses.

Zur Analyse der räumlichen Wildunfallintensität konnten 45790 Fälle berücksichtigt werden, davon 18834 aus dem Jahr 2006 und 26956 aus dem Jahr 2009 und für die Untersuchung der zeitlichen Intensität 74735, davon 34177 aus 2006 und 40558 aus 2009.

Der Aufbau der Arbeit gliedert sich wie folgt: Im zweiten Kapitel wird zunächst die notwendige Theorie erläutert. Dazu gehören räumliche Poisson-Punktprozesse, sowie generalisierte lineare und additive Modelle. Weitere Abschnitte behandeln die K-Funktion und die eng verwandte L-Funktion, welche gängige Maße sind, um Aufschluß über abstoßendes bzw. anziehendes Verhalten eines Punktmusters zu erhalten. Der erste Teil des dritten Kapitels behandelt zunächst die Formate, in denen die Kovariablen zur Verfügung standen und beschreibt nach einer Erklärung von Kartenprojektionen im Allgemeinen den Aufbau des in Deutschland gebräuchlichen Gauß-Krüger-Koordinatensystems, in

dem die Koordinaten der Wildunfälle angegeben sind. Im zweiten Teil werden die einzelnen Einflußgrößen und deren Aufbereitung bzw. Gewinnung beschrieben. Das vierte Kapitel beschreibt das genaue Vorgehen bei der Analyse und geht im Einzelnen auf die Ergebnisse ein. Im letzten Kapitel werden die Ergebnisse noch einmal gebündelt und in diskutierender Weise zusammen gefasst und insbesondere, sofern möglich, mit denen von Hothorn et al. (in Vorb.) und Kaldhusdal (2011) verglichen. Der Hauptunterschied zu deren Herangehensweisen besteht darin, dass die Analyse nicht aggregiert auf Gemeindeebenen erfolgt, sondern dass zugelassen wird, dass sich die Intensität von Lokation zu Lokation ändern kann.

2. Theorie

2.1. Allgemeines zu räumlichen Punktprozessen

In der räumlichen Statistik sind die Beobachtungen generell Realisationen eines räumlichen Prozesses $Y = \{Y_w, w \in W\}$, wobei w eine Lokationsvariable ist und Y_w Werte in einem Zustandsraum E annehmen kann. Im Spezialfall eines (*unmarkierten*) räumlichen Punktprozesses ist $E = W$ und Y abzählbar. Die Positionen von Punkten im Zustandsraum entsprechen bereits den Realisationen des Prozesses. W wird hier als *Beobachtungsfenster* bezeichnet. Im Falle der Wildunfälle wäre das die Karte von Bayern in Gauß-Krüger-Koordinaten. Wenn zusätzlich zu den Lokationen x_i in $\mathbb{R}^2$ jedem Objekt eine oder mehrere zufällige Eigenschaften $m(x_i) \in M$, sogenannte Marken zugeordnet werden, spricht man von einem *markierten* räumlichen Punktprozess. M heißt dabei Markenraum. Mit $N(B)$ wird im Weiteren die zufällige Anzahl an Punkten in einem Bereich $B \subset W$ bezeichnet. $X_i \in W$ sei die Zufallsvariable, die die Lokation x_i des i-ten Objekts beschreibt. Dann ergibt sich für einen unmarkierten räumlichen Punktprozess $Y = \{X_i; i = 1, \ldots, N(W)\}$ und einen markierten räumlichen Punktprozess $Y = \{(X_i, m(X_i)); i = 1, \ldots, N(W)\}$ (Gaetan und Guyon, 2010; Møller und Waagepetersen, 2004; Schmid, 2010).

Im Weiteren werden nur noch unmarkierte räumliche Punktprozesse betrachtet. Die erwartete Anzahl von Punkten in B wird als *Intensitätsmaß* $\Lambda(B)$ bezeichnet. Die Dichte

des Intensitätsmaßes heißt *Intensitätsfunktion* $\lambda(\cdot)$. Es gilt also:

$$\Lambda(B) = \int_B \lambda(\xi) d\xi, \quad B \subseteq W. \tag{2.1}$$

Ein zentrales Konzept in der Punktprozesstheorie ist das der Stationarität. Ein Punktprozess Y heißt stationär, wenn Y und der um $x \in \mathbb{R}^d$ verschobene Prozess $Y_x := \{Y + x\}$ dieselbe Verteilung haben. In diesem Fall kann man die Intensitätsfunktion als erwartete Anzahl von Punkten pro Einheitsquadrat oder -würfel interpretieren, da diese bei Stationarität nicht von der Lokation abhängt und bezeichnet sie auch kurz als *Intensität* λ (Illian et al., 2008).

2.2. Räumliche Poissonprozesse

Räumliche Poissonprozesse modellieren unabhängige Punkte. Dies ist in der Praxis zumeist eine unrealistische Annahme. Dennoch spielen räumliche Poissonprozesse eine zentrale Rolle, da sie zum einen der Grundbaustein sind, auf den komplexere Modelle aufbauen und zum anderen, in den in Abschnitt 2.6 betrachteten Maßzahlen als Referenzmodelle dienen (Møller und Waagepetersen, 2007). Im Falle der Wildunfälle wurde ein räumlicher Poissonprozess angewendet, da davon ausgegangen wurde, dass nach Adjustierung auf die ausgewählten Einflussgrößen eine mögliche Abhängigkeit der Punkte (beinahe vollständig) aufgeklärt werden kann und daher kein komplexeres Modell notwendig ist.

Ein Punktprozess Y auf W ist ein räumlicher *Poisson-Punktprozess*, wenn folgende Eigenschaften erfüllt sind (Schmid, 2010):

1. Für jedes $B \subseteq W$ mit $\Lambda(B) < \infty$ gilt $N(B) \sim Po(\Lambda(B))$, d.h. $N(B)$ ist Poisson-verteilt mit Erwartungswert $\Lambda(B)$.

2. Sind $B_1, \ldots, B_k$ disjunkte Borelmengen, so sind $N(B_1), \ldots, N(B_k)$ unabhängig.

Der Knappheit halber werden räumliche Poisson-Punktprozess im Weiteren meist nur Poissonprozesse genannt. Ein wichtiger Spezialfall sind stationäre Poissonprozesse, welche als *homogen* bezeichnet werden. Bei diesen ist die erwartete Anzahl an Punkten also nur abhängig von der Größe des betrachteten Bereichs und nicht von der Lokation in W. Nicht-stationäre Poissonprozesse heißen entsprechend *inhomogenene* Poissonprozesse (Møller und Waagepetersen, 2004). Interpretieren könnte man $\lambda(x)$ etwas informell als die erwartete Anzahl von Punkten im Einheitsquadrat um den Punkt x, wobei allerdings nicht berücksichtigt wird, dass der Wert von $\lambda(\cdot)$ in diesem Bereich variert. Im Fall der Wildunfälle kann man das aber vernachlässigen, da als Einheit Meter bzw. Quadratmeter gewählt wurde und die Intensitätsfunktion in einem Quadratmeter als konstant angesehen werden kann. In Abschnitt 4.1 wird mit Anwendung auf die Wildunfalldaten auf die Schätzung eingegangen.

Gegeben $N(W) = n$ haben die Punkte eines Poissonprozesses die Dichte (Illian et al., 2008):

$$f_n(x_1, \ldots, x_n) \;=\; \prod_{i=1}^{n} \frac{\lambda(x_i)}{\Lambda(W)} \tag{2.2}$$

Die Dichte eines Punktes ist offenbar proportional zu seiner Intensität und die gemeinsame Dichte ein Produkt der Beiträge der einzelnen Beobachtungen. Letzteres ist gleichbedeutend mit der Unabhängigkeit der Punkte (Grimmett und Welsh, 1986).

2.3. Schätzung mit Berman-Turner-Device

Berman und Turner (1992) stellen die approximierte log-Likelihood eines Poissonprozesses mit Kovariablen in einer Form dar, die formal äquivalent ist mit der log-Likelihood in der Poissonregression bei Verwendung der log-Linkfunktion, d.h. der natürlichen Linkfunktion. Es wird dazu zunächst angenommen, dass die Intensitätsfunktion $\lambda_{\boldsymbol{\beta}}(x)$ folgende Form hat:

$$\lambda_{\boldsymbol{\beta}}(x) \;=\; \exp(\boldsymbol{Q}(x)^\top \boldsymbol{\beta}) \tag{2.3}$$

$\boldsymbol{Q}(x) = (Q_1(x), \ldots, Q_p(x))^\top$ mit Koeffizientenvektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ ist dabei ein Vektor von Kovariablen, deren Ausprägungen für alle $x \in W$ definiert sind (Berman und Turner, 1992). Die logarithmierte Intensitätsfunktion hängt dabei also linear von den Kovariablenwerten ab. Man spricht auch vom sogenannten log-Link. Obgleich der Prädiktor linear ist, können damit bei Verwendung von Spline-Basen dennoch nichtlineare Funktionen $f(\cdot) = \sum_{k=1}^{K} \theta_k B_k(\cdot)$ (Abschnitt 2.5) der Einflußgrößen modelliert werden. Betrachtet man z.B. o.B.d.A. nur eine Einflußgröße $Q_1^*(x)$ und verwendet zu deren Modellierung K Basisfunktionen so kann man $Q_k(x) = B_k(Q_1^*(x))$ und $\beta_k = \theta_k$ $(k = 1, \ldots, K)$ setzen und erhält die Schätzer $\widehat{\theta}_k$ und damit die geschätzte nonparametrische Funktion $\widehat{f}(x) = \sum_{k=1}^{K} \widehat{\theta}_k B_k(x)$.

Die Dichte eines Poissonprozesses mit Kovariablen lässt sich gemäß dem Satz von Bayes auf folgende Art multiplikativ zerlegen:

$$f(x_1, \ldots, x_n, n|\boldsymbol{\beta}) \;=\; f(x_1, \ldots, x_n|n, \boldsymbol{\beta}) \cdot f(n|\boldsymbol{\beta}) \tag{2.4}$$

Nach (2.2) ist der erste Multiplikator gegeben durch:

$$f(x_1, \ldots, x_n|n, \boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} \frac{\lambda_{\boldsymbol{\beta}}(x_i)}{\Lambda_{\boldsymbol{\beta}}(W)} \tag{2.5}$$

$f(n|\boldsymbol{\beta})$ folgt direkt aus der Definition des räumlichen Poissonprozesses (siehe Abschnitt 2.2):

$$f(n|\boldsymbol{\beta}) \;=\; \frac{\Lambda_{\boldsymbol{\beta}}(W)^n}{n!} \exp(-\Lambda_{\boldsymbol{\beta}}(W)) \tag{2.6}$$

Also gilt:

$$\begin{aligned} L(\boldsymbol{\beta}) = f(x_1, \ldots, x_n, n|\boldsymbol{\beta}) \;&=\; \left(\prod_{i=1}^{n} \frac{\lambda_{\boldsymbol{\beta}}(x_i)}{\Lambda_{\boldsymbol{\beta}}(W)} \right) \frac{\Lambda_{\boldsymbol{\beta}}(W)^n}{n!} \exp(-\Lambda_{\boldsymbol{\beta}}(W)) \\ &\propto \prod_{i=1}^{n} \lambda_{\boldsymbol{\beta}}(x_i) \exp(-\Lambda_{\boldsymbol{\beta}}(W)) \end{aligned} \tag{2.7}$$

Demnach ergibt sich die log-Likelihood zu:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \lambda_{\boldsymbol{\beta}}(x_i) - \int_W \lambda_{\boldsymbol{\beta}}(u) du \tag{2.8}$$

Das Integral $\int_W \lambda_{\boldsymbol{\beta}}(u)du$ ist analytisch nicht zugänglich und muss numerisch approximiert werden:

$$\int_W \lambda_{\boldsymbol{\beta}}(u)du \;\approx\; \sum_{j=1}^{M} w_j \lambda_{\boldsymbol{\beta}}(s_j) \tag{2.9}$$

Die s_j $(j = 1, \ldots, M)$ schließen dabei sowohl die Datenpunkte x_i als auch zusätzliche Stützstellen zur Quadratur, die sogenannten *Dummypunkte* ein. Die w_j sind Gewichte die sich aus dem verwendeten Quadraturschema ergeben. Die einzelnen Bereiche, in denen zur Berechnung des Integrals mittels numerischer Integration die Intensitätsfunktion als konstant gehandhabt wird, werden im Weiteren als *Kacheln* bezeichnet.

Setzt man dieses approximierte Integral in die log-Likelihood (2.8) ein, ergibt sich:

$$\begin{aligned} l(\boldsymbol{\beta}) &\overset{(\approx)}{=} \sum_{j=1}^{M} w_j \left\{ \frac{N_j}{w_j} \log \lambda_{\boldsymbol{\beta}}(s_j) - \lambda_{\boldsymbol{\beta}}(s_j) \right\} \\ &= \sum_{j=1}^{M} w_j \left\{ \frac{N_j}{w_j} \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta}) \right\} \quad \text{mit} \quad N_j = \mathbf{1}\{s_j \in X\} \end{aligned} \tag{2.10}$$

Es ist sofort ersichtlich, dass dies einer gewichteten Poissonregression mit Gewichten w_j und Zielgröße N_j/w_j entspricht.

Für $M \to \infty$, d.h. mit wachsender Anzahl an Stützpunkten, bzw. je nach Art des Quadraturverfahrens für wachsende Anzahl von Kacheln wird die log-Likelihood beliebig genau approximiert. In diesem Fall können die Resultate aus der Poissonregression, inbesondere Likelihoodwerte, Parameterschätzer und deren Standardfehler für den Poissonprozess mit Kovariablen übernommen werden. Die Schätzung lässt sich mit jedem Programmpaket, in dem Poissonregression implementiert ist, durchführen, gesetzt den Fall, dass nicht ganzahlige Werte für die Zielvariable zugelassen sind (Berman und Turner, 1992).

Ein weiterer praktischer Vorteil bei der Verwendung solch eines Quadraturschemas ist, dass die Ausprägungen einer Kovariable nicht an allen Lokationen in W bekannt sein müssen. Es genügt, dass sie an den Ereignispunkten und an einigen zusätzlichen Punkten gegeben sind. Letztere werden dann als Dummypunkte gewählt (Baddeley und Turner, 2005a). Allerdings müssen die Ausprägungen bei mehreren Kovariablen an denselben

Lokationen bekannt sein, oder es muss zumindest eine (genügend große) gemeinsame Untermenge an Punkten existieren, an denen die Werte aller Einflußgrößen verfügbar sind.

2.4. Generalisierte lineare Modelle

Obgleich die in der im Berman-Turner-Device aufkommende Poissonregression gewichtet ist, werden Gewichte im folgenden Abschnitt der Einfachheit halber weggelassen.

Die folgenden Ausführungen beziehen sich, soweit nicht anders gekennzeichnet auf Fahrmeir et al. (2007). Das klassische *lineare* Modell für Individuen $i = 1, \ldots, n$ mit Zielgröße y_i, Kovariablenvektor $\boldsymbol{x}_i$ und Koeffizientenvektor $\boldsymbol{\beta}$ ist gegeben durch:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}(\epsilon_i) = 0, \quad \mathbb{V}(\epsilon_i) = \sigma^2, \quad \epsilon_i \text{ i.i.d.} \tag{2.11}$$

Hierbei gilt also:

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{und} \quad \mathbb{V}(y_i|\boldsymbol{x}_i) = \sigma^2. \tag{2.12}$$

Nach dem Gauß-Markov-Theorem ist der beste lineare erwartungstreue Schätzer für $\boldsymbol{\beta}$ für dieses Problem der bekannte Kleinste-Quadrate-Schätzer (Fox, 1997). Zur Konstruktion von Konfidenzintervallen und Teststatistiken ist es günstig, zusätzlich anzunehmen, dass die ϵ_i normalverteilt sind. Damit ergibt sich

$$y_i|\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2). \tag{2.13}$$

Nimmt man verallgemeinernd an, dass die $y_i|\boldsymbol{x}_i$ unabhängig voneinander der Verteilungsfunktion einer Exponentialfamilie folgen und der Erwartungswert dieser, allgemeiner als in (2.12) über eine Funktion $h(\cdot)$ vom linearen Prädiktor $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ abhängt, gelangt man zu *generalisierten linearen* Modellen. Die Dichte der $y_i|\boldsymbol{x}_i$ lässt sich damit darstellen als:

$$f(y_i|\theta_i) \;=\; \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} - c(y_i, \phi)\right). \tag{2.14}$$

Die bei der Schätzung der räumlichen Poissonprozesse aufkommende Poissonverteilung und die Verteilungsannahme in dem generalisierten additiven Modell (siehe Abschnitt 2.5), die negative Binomialverteilung bei festem α gehören zu den Exponentialfamilien. In (2.14) heißt θ_i kanonischer Parameter und w_i ist ein Gewicht. ϕ ist ein sogenannter Dispersionsparameter. In dieser Parametrisierung ergeben sich Erwartungswert und Varianz zu:

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \mu_i = b^{'}(\theta_i), \quad \mathbb{V}(y_i|\boldsymbol{x}_i) = \sigma_i^2 = \phi b^{''}(\theta_i). \tag{2.15}$$

Offensichtlich beeinflusst ϕ nicht die Erwartungswert-, sondern nur die Varianzstruktur. Ist dieser Parameter unbekannt, lässt er sich schätzen durch:

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)}. \tag{2.16}$$

Dabei ist p die Anzahl der Parameter und $v(\cdot)$ die sogenannte Varianzfunktion, die zweite Ableitung von b nach θ, dargestellt als Funktion des Erwarungswertes.

Wie bereits weiter oben erwähnt, nimmt man nun folgende Beziehung an:

$$\mu_i = h(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \quad \text{bzw.} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{2.17}$$

$\boldsymbol{x}_i^\top \boldsymbol{\beta}$ kann dabei auch um eine für jede Beobachtung individuelle Konstante c_i, einen sogenannten *Offset* erweitert werden. Dieser kann als Regressionskoeffizienten mit Parameterwert 1 interpretiert werden. $h(\cdot)$ bezeichnet man als Responsefunktion und $g(\cdot)$ als Linkfunktion. Die sogenannte natürliche Linkfunktion ist definiert durch $g(\mu_i) = \theta_i$.

Unter Vernachlässigung der hinsichtlich θ_i konstanten Terme $c(y_i, \phi)$ lautet also die log-Likelihood für $\boldsymbol{\beta}$:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\phi}. \tag{2.18}$$

Ableiten nach $\boldsymbol{\beta}$ ergibt die Score-Funktion:

$$\begin{aligned} s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{1}{\phi} \cdot \frac{\partial (y_i\theta_i - b(\theta_i))}{\partial \theta_i} \cdot \left(1/\frac{\partial \mu_i}{\partial \theta_i}\right) \cdot \frac{\partial \mu_i}{\partial (\boldsymbol{x}_i^\top \boldsymbol{\beta})} \cdot \frac{\partial \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \\ &= \sum_{i=1}^{n} \frac{1}{\phi} \cdot (y_i - b'(\theta_i)) \cdot (1/b''(\theta_i)) \cdot h'(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \cdot \boldsymbol{x}_i = \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i \frac{d_i}{\sigma_i^2}(y_i - \mu_i), \end{aligned} \tag{2.19}$$

wobei d_i gleich $h'(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ ist.

Damit ist die Fisher-Informationsmatrix gegeben durch:

$$\begin{aligned} F(\boldsymbol{\beta}) &= \mathbb{E}\left(- \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right) = Cov(s(\boldsymbol{\beta})) = \mathbb{E}(s(\boldsymbol{\beta}) s(\boldsymbol{\beta})^\top) \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top w_i \quad \text{mit} \quad w_i = d_i^2/\sigma_i^2. \end{aligned} \tag{2.20}$$

Bezeichne $\boldsymbol{y}$ den Vektor der Zielgrößen y_i, $\boldsymbol{\mu}$ den der Erwartungswerte μ_i und

$$\boldsymbol{D} = \text{diag}(d_1, \ldots, d_n), \quad \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2), \quad \boldsymbol{W} = \text{diag}(w_1, \ldots, w_n), \tag{2.21}$$

so lassen sich s und F wie folgt darstellen:

$$s(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \quad F(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}. \tag{2.22}$$

Die Scoregleichung $s(\boldsymbol{\beta}) = \boldsymbol{0}$ lässt sich nicht analytisch, sondern nur numerisch lösen, üblicherweise iterativ mit Hilfe von Fisher-Scoring mit einem geeigneten Startwert $\widehat{\boldsymbol{\beta}}^{(0)}$:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + F^{-1}\big(\widehat{\boldsymbol{\beta}}^{(k)}\big) s\big(\widehat{\boldsymbol{\beta}}^{(k)}\big), \quad k = 0, 1, 2, \ldots \tag{2.23}$$

Dieses Verfahren lässt sich auch als sogenannte iterativ gewichtete Kleinste-Quadrate-Schätzung oder kurz IRLS-Schätzung darstellen. Dabei wird obige Gleichung in Form eines Kleinste-Quadrate-Schätzers mit Gewichtungsmatrix umgeschrieben. $\widehat{\boldsymbol{\beta}}^{(k+1)}$ ergibt sich also aus der Minimierung einer gewichteten Quadratsumme. Diese lautet:

$$|\sqrt{\boldsymbol{W}^{(k)}}(\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{X}\boldsymbol{\beta})|^2 \tag{2.24}$$

Dabei ist $\tilde{\boldsymbol{y}}^{(k)}$ ein sogenannter Arbeitsvektor mit Elementen $\tilde{y}_i(\hat{\boldsymbol{\beta}}^{(k)}) = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{(k)} + d_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)})(y_i - \widehat{\mu}_i(\boldsymbol{\beta}^{(k)}))$ mit der Prädiktion $\widehat{\mu}_i(\boldsymbol{\beta}^{(k)})$ und $\boldsymbol{W}^{(k)}$ die Gewichtungsmatrix aus (2.21) ausgewertet bei $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$. Ein möglicher Offset geht dabei nicht in $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{(k)}$ ein, sondern nur in die Berechnung von $\widehat{\mu}_i(\boldsymbol{\beta}^{(k)})$. Diese Iteration wird fortgesetzt bis ein Abbruchkriterium, z.B. $||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||/||\hat{\boldsymbol{\beta}}^{(k)}|| \leq \epsilon$ erfüllt ist.

Gemäß den asymptotischen Eigenschaften von Maximum-Likelihood-Schätzern gilt:

$$\hat{\boldsymbol{\beta}} \overset{\text{a}}{\sim} N(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}})). \tag{2.25}$$

Eine eng verwandte Methodik ist die der Quasi-Likelihood-Modelle: Die durch die sich aus der Verteilungsannahme ergebende Beziehung zwischen Erwartungswert und Varianz in (2.15) könnte in realen Daten nicht erfüllt sein, z.B. bei einer vermeintlich poissonverteilten Zielgröße, für die die geschätzte Varianz aber stark von dem geschätzten Erwartungswert abweicht. In solchen Fällen kann man eine sogenannte Quasi-Likelihoodfunktion aufstellen, die zu keiner Verteilung gehört, deren Ableitung aber die Form der Scorefunktion (2.22) hat und gleichzeitig eine realistischere Beziehung zwischen Erwartungswert und Varianz spezifiziert.

Eine Größe, die man in ähnlicher Weise, wie die Residuenquadratsumme in linearen Modellen interpretieren kann, ist die der Devianz. Diese ist gegeben durch (vgl. z.B. Wood (2006a)):

$$D \;=\; 2\big[l(\widehat{\boldsymbol{\beta}}_{\max}) - l(\widehat{\boldsymbol{\beta}})\big]\phi \;=\; \sum_{i=1}^{n} \big[y_i(\tilde{\theta}_i \,-\, \widehat{\theta}_i) \,-\, b(\tilde{\theta}_i) + b(\widehat{\theta}_i)\big]. \tag{2.26}$$

Dabei bezeichnet $l(\widehat{\boldsymbol{\beta}}_{\max})$ die Likelihood des sogenannten saturierten Modells, bei welchem jede Beobachtung einen eigenen Parameter erhält. Diese kann man berechnen, indem man in der Likelihood y_i für μ_i einsetzt und sie besitzt den größmöglich erreichbaren Likelihoodwert. $\tilde{\theta}_i$ bzw. $\widehat{\theta}_i$ sind die Parameter des saturierten bzw. des betrachteten Modells. D ist offenbar unabhängig von dem Skalenparameter ϕ. Wenn $y_i|\boldsymbol{x}_i$ einer Normalverteilung folgt, entspricht D gerade der Residuenquadratsumme. Der Absolutwert

der Devianz ist für sich genommen schwer zu interpretieren. Um die Fähigkeit eines Modells, die Varianz in den Daten aufzuklären, zu beurteilen, verwendet man deshalb üblicherweise die erklärte Devianz, welche ein Analogum zu R^2 aus der linearen Regression darstellt. Diese ist der Anteil der Devianz des Nullmodells, d.h. des Modells, das nur den Intercept und, wenn vorhanden, den Offset enthält, der durch das betrachtete Modell aufgeklärt wird.

2.5. Generalisierte additive Modelle mit Regressionssplines

Die folgenden Ausführungen beziehen sich im Wesentlichen, wenn nicht anders gekennzeichnet, auf Fahrmeir et al. (2007) und Wood (2006a). Bei klassischer linearer bzw. generalisierter linearer Regression wird angenommen, dass der Einfluss stetiger Kovariablen auf den Erwartungswert der Zielgröße bzw. auf die Linkfunktion linear ist. Das bedeutet, dass sich Letztere unabhängig von dem Niveau einer Kovariable immer um einen konstanten Wert ändert, wenn sich der Wert der Kovariable um einen konstanten Wert ändert (bei dabei konstanten Werten der restlichen Kovariablen), was eine sehr rigide Annahme ist, die häufig nicht erfüllt ist. Man möchte deshalb den Einfluss von Kovariablen durch flexiblere Funktionen modellieren. Seien $x_1, \ldots, x_p$ die Kovariablen mit linearen Einflüssen und $x_{p+1}, \ldots, x_P$ diejenigen mit nichtlinearen Einflüssen $f_{p+1}(\cdot), \ldots, f_P(\cdot)$ so lautet also der Prädiktor in diesem Fall:

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + f_{p+1}(x_{p+1}) + \ldots + f_P(x_P) \tag{2.27}$$

Man spricht dann von sogenannten *generalisierten additiven Modellen*. Eine sehr einfache Methode wäre, die Kovariablen mit Hilfe eines Polynoms zu transformieren. Um dies angemessen zu leisten, muss man aber aus inhaltlichen Überlegungen Kenntnis über die Form des Einflusses haben, was selbst dann bei einer größeren Zahl von Einflußgrößen nur sehr schwierig durchzuführen ist. Außerdem kann es bei komplizierten Formen des Einflusses sein, dass man diesen gar nicht mehr über ein Polynom ausdrücken kann. Man

benötigt also eine Methodik, die es erlaubt, nicht-lineare, glatte Einflüsse ohne vorheriges Wissen über deren Form zu schätzen. Eine naheliegende Herangehensweise, um die Rigidität von Polynomen zu überwinden, aber sich dennoch deren Glattheit zu Nutze zu machen, wäre es, den Wertebereich in Intervalle aufzuteilen und auf jedem dieser ein einzelnes Polyom vom Grad l zu schätzen. Dabei stößt man aber auf das Problem, dass benachbarte Polynome, an ihren Enden, den sogenannten *Knoten(punkten)* sich nicht nahtlos aneinanderfügen, was inhaltlich nicht sinnvoll zu interpretieren wäre. Stellt man jedoch die Forderung, dass die Polynome an den Knoten $(l-1)$-mal stetig differenzierbar sind, haben sie an den Grenzen nicht nur dieselben Werte, sondern man erhält auch eine Funktion, die auf dem ganzen Wertebereich angemessen glatt ist. Dies entspricht der Definition eines *Polynom-Splines* vom Grad l. Es können hier zwei Faktoren variert werden, zum einen der Grad l des Polynoms, zum anderen die Anzahl (und Lage) der Knoten und damit der Polynome. Ersteres steuert die Glattheit der einzelnen Polynome und also auch der Funktion, letzteres die Flexibilität mit der die Funktion auf Änderungen im Verlauf des Einflusses reagiert.

Um alle möglichen Polynom-Spline-Funktionen eines bestimmten Grades und mit einer festgelegten Anzahl von Knoten darstellen zu können, benötigt man eine Spline-Basis. Man kann zeigen, dass der Raum der Polynom-Spline-Funktionen vom Grad l mit m Knoten $\kappa_1, \ldots, \kappa_m$, wobei $\kappa_1 < \ldots < \kappa_m$, ein Vektorraum der Dimension $d = m + l - 1$ ist. Demnach lassen sich alle Polynom-Spline-Funktionen in diesem Raum als Linearkombination von d Basisfunktionen $B_i(\cdot)$ $(i = 1, \ldots, d)$, d.h. als

$$f(\cdot) = \sum_{i=1}^{d} \theta_i B_i(\cdot) \tag{2.28}$$

darstellen (Keele, 2008). Man sagt dann auch, dass die Kovariable *semiparametrisch* modelliert wurde. Hier werden zwei Typen behandelt, die Trunkierte-Potenzen- und die B-Spline-Basis, welche auch in der Arbeit verwendet wurde.

Die Trunkierte-Potenzen-Basis ist gegeben durch:

$$\begin{aligned} B_1(z) = 1, \quad B_2(z) = z, \;\; \ldots, \;\; B_{l+1}(z) = z^l, \\ B_{l+2}(z) = (z-\kappa_2)_+^l, \;\; \ldots, \;\; B_d(z) = (z-\kappa_{m-1})_+^l, \end{aligned} \tag{2.29}$$

mit

$$(z-\kappa_j)_+^l = \begin{cases} (z-\kappa_j)^l & z \geq \kappa_j \\ 0 & \text{sonst} \end{cases} \tag{2.30}$$

Eingesetzt in (2.28) erkennt man schnell, dass sich mit dieser Basis auf jedem Intervall ein Polynom vom Grad l ergibt: Bis zum zweiten Knoten ist f offensichtlich ein Polynom vom Grad l mit Koeffizienten $\theta_1, \ldots, \theta_{l+1}$. Ab diesem Knoten kommt zu θ_{l+1} additiv θ_{l+2} hinzu, sodass sich wieder ein Polynom vom Grad l ergibt, ab dem dritten Knoten verändert sich der Koeffizient des höchstens Polynoms um θ_{l+3}, usw., sodass man wie gewünscht, auf allen Intervallen individuelle Polynome vom Grad l erhält, deren Funktionswerte an den Enden gleich zu denen der Nachbarn sind. Dass die Funktion überall $(l-1)$-mal stetig differenzierbar ist, ergibt sich daraus, dass jeder Funktionswert zu einem Polynom vom Grad l gehört. Obgleich die Eigenschaften von Spline-Funktionen einfach nachzuvollziehen sind, hat die Trunkierte-Potenzen-Matrix zwei Nachteile. Die Basen sind besonders für nahe beieinanderliegende Knoten annäherend kollinear. Außerdem kann es zu Gleitkommaüberläufen in der numerischen Auswertung kommen, da die trunkierten Polynome ab dem jeweiligen Knoten multipliziert mit ihrem Koeffizienten immer in die Addition eingehen, wodurch es für betragsmäßig große Werte der Kovariablen zu sehr großen Werten in der Addition (2.28) kommt.

B-Spline-Basisfunktionen sind nur über $(l+2)$ benachbarte Knoten größer Null, sodass obiges Problem nicht auftreten kann. Allerdings ist ihre Konstruktion keineswegs zugänglich (Höllig, 2003). Sie können rekursiv definiert werden als:

$$\begin{aligned} B_j^l(z) \;&=\; \frac{z-\kappa_j}{\kappa_{j+l}-\kappa_j} B_j^{l-1}(z) \;+\; \frac{\kappa_{j+l+1}-z}{\kappa_{j+l+1}-\kappa_{j+1}} B_{j+1}^{l-1}(z) \\ \text{und} \quad B_j^0(z) &= \mathbf{1}\{z \in [\kappa_j, \kappa_{j+1})\} \end{aligned} \tag{2.31}$$

Offensichtlich ist $B_j^0(z)$ ein Polynom vom Grad 0 auf dem Intervall $[\kappa_j, \kappa_{j+1})$ und $B_j^1(z)$ besteht aus zwei zusammengesetzten Polynomen vom Grad 1 auf den Intervallen $[\kappa_j, \kappa_{j+1})$ und $[\kappa_{j+1}, \kappa_{j+2}]$. Allgemein kann man über Induktion zeigen, dass $B_j^l(z)$ entsprechend aus $(l+1)$ Polynomstücken besteht, die $(l-1)$-mal stetig differenzierbar zusammengesetzt sind, sodass die beiden Forderungen an die Spline-Funktion f erfüllt sind. Zur Implementierung der obigen Definition benötigt man außerhalb des Wertebereiches der Kovariable in negativer wie in positiver Richtung zusätzlich l Knoten. Bei nicht-äquidistanten Knoten wählt man diese dabei üblicherweise mit Abstand gleich dem zwischen den ersten beiden bzw. letzten beiden inneren Knoten.

Setzt man die Splinefunktionen in der Form (2.28) in den Prädiktor (2.27) ein, erhält man wieder einen linearer Prädiktor. Damit lässt sich dieses Modell mit Standardmethoden generalisierter linearer Regression schätzen, wobei sich die Koeffizientenschätzer der Spline-Funktionen für sich genommen nicht sinnvoll interpretieren lassen und man zur Interpretation die geschätzten Funktion $\widehat{f}_{p+1}, \ldots, \widehat{f}_P$ grafisch betrachtet.

Nimmt man an, dass die Einflüsse zweier Kovariablen z_1 und z_2 in ihrer Wirkung auf die Zielgröße bzw. den Wert der Linkfunktion interagieren, so reicht es nicht, für jede der beiden eine Funktion $\widehat{f}_1$ bzw. $\widehat{f}_2$ zu schätzen und diese aufzusummieren. Vielmehr muss man eine Funktion $f_{12}(x_1, x_2)$ für beide Kovariablen schätzen. Im Falle klassischer linearer Regression modelliert man Interaktionen zwischen Kovariablen, indem man ihre Werte multipliziert und mit einem eigenen zu schätzenden Parameter versieht. Da die Splinefunktionen auch linear in den Basisfunktionen sind, kann man diese Idee auch in diesem Fall anwenden, indem man entsprechend die paarweisen Produkte der Splinebasen betrachtet:

$$B_{jk}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_k^{(2)}(z_2), \quad j = 1, \ldots, d_1, \quad k = 1, \ldots, d_2, \tag{2.32}$$

wobei $(B_1^{(1)}, \ldots, B_{d_1}^{(1)})$ und $(B_1^{(2)}, \ldots, B_{d_1}^{(2)})$ die sogenannten *marginalen* Basen von z_1 und z_2 sind, die dieselben Formen wie im eindimensionalen Fall haben. Das sind die

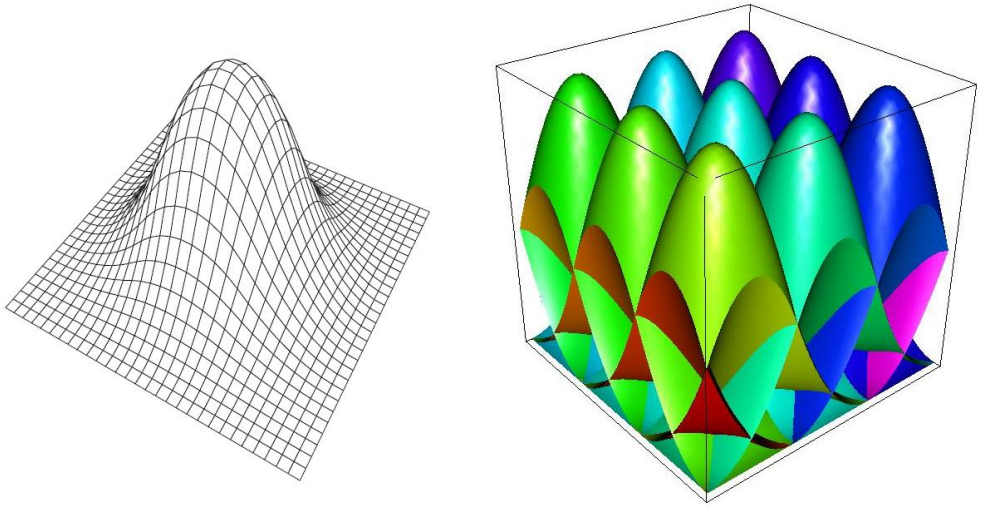

Abbildung 2.1.: links: Tensorprodukt-B-Spline-Basisfunktion zum Grad $l = 2$, rechts: entsprechende Basisfunktion mit 3 inneren Knoten

zweidimensionalen Basisfunktionen zu der Funktion f_{12}, die sich damit ergibt zu:

$$f_{12}(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} \theta_{jk} B_{jk}(z_1, z_2). \tag{2.33}$$

Hierbei spricht man von Tensorprodukt-Splines. Diese könnte man auch leicht für drei und mehr beteiligte Kovariablen formulieren. Die bereits erläuterten, numerischen Probleme der Trunkierte-Potenzen-Basis treten im zweidimensionalen Fall noch stärker auf, weshalb die B-Spline-Basis zu bevorzugen ist. In Abbildung 2.1 ist links eine Tensorprodukt-B-Spline-Basisfunktion zum Grad $l = 2$ zu sehen und rechts die entsprechende Basis mit 3 inneren Knoten.

Die Güte der Anpassung von Spline-Funktionen hängt stark von der Anzahl m der Knoten $\kappa_1, \ldots, \kappa_m$ ab. Wählt man zu wenige Knoten, so ist die Schätzung zu grob, und wichtige Änderungen im Verlauf gehen unter, wählt man zu viele, erfolgt eine zu starke Anpassung an die gegebenen Daten, die Schätzung reagiert aber flexibel auf Änderungen in der Form des Einflusses. Letztere Eigenschaft würde man gerne bewahren, aber dennoch eine den Daten angemessen glatte Schätzung erhalten. Eine gängige Möglichkeit,

die auch in der Analyse verwendet wird, ist, viele Knoten zu wählen, aber Parameter in der Modellierung mit aufzunehmen, die zu raue Funktionen bestrafen. Hierbei spricht man von *penalisierten* Splines oder *P-Splines*, die im Folgenden erläutert werden. Ein lokales Maß für die Krümmung einer Funktion - positiv oder negativ - ist ihre zweite Ableitung. Integriert man die quadrierte, zweite Ableitung über den Wertebereich, erhält man demnach ein globales Maß für die Kurvigkeit. Dieses sollte sich in einem vernünftigen Bereich befinden. Anstatt die Likelihood direkt zu maximieren, maximiert man sie unter der Bedingung, dass die Kurvigkeit der beteiligten Spline-Funktionen nicht zu groß ist. Dazu maximiert man die folgende *penalisierte* log-Likelihood (Wood, 2004):

$$l_{\text{pen}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\sum_j \lambda_j \int [f_j''(x)]^2 dx, \tag{2.34}$$

wobei $\boldsymbol{\theta}$ alle zu schätzenden Koeffizienten enthält. Für zweidimensionale (Tensorprodukt-)Splines wird für jede der beiden Kovariablen ein eigener Strafterm mit jeweiligem Glättungsparameter verwendet. Diese Strafterme ergeben sich, indem jeweils das (approximative) Mittel bzw. das Integral der Integrale über die quadrierten zweiten Ableitungen der jeweils auf die andere Variable bedingten Funktionen über alle Werte der anderen Variablen berechnet wird, für Details siehe Wood (2006b). Je größer ein λ_j ist, desto stärker wird die Kurvigkeit der entsprechenden Funktion bestraft und desto glatter ist die resultierende Schätzung. Bevor auf die datengestützte Wahl dieser Glättungsparameter eingegangen wird, soll zunächst die spezielle Art der Penalisierung bei den in der Analyse verwendeten B-Splines erläutert werden. Man kann über die Ableitungen der B-Spline-Basisfunktionen recht leicht zeigen, dass bei äquidistanten Knoten die Summe der quadrierten Differenzen k-ter Ordnung der Koeffizienten der Basisfunktionen eine gute Approximation an das Integral der quadrierten k-ten Ableitungen in (2.34) liefert. Die k-te Differenz ist rekursiv definiert als:

$$\begin{aligned}
&\Delta^1\theta_j = \theta_j - \theta_{j-1},\\
&\Delta^2\theta_j = \Delta^1\Delta^1\theta_j = \Delta^1\theta_j - \Delta^1\theta_{j-1} = \theta_j - 2\theta_{j-1} + \theta_{j-2},\\
&\quad\vdots
\end{aligned}$$

$$\Delta^k \theta_j = \Delta^{k-1}\theta_j - \Delta^{k-1}\theta_{j-1}. \tag{2.35}$$

Approximativ lässt sich damit obige penalisierte Likelihood (o.B.d.A.) für den Fall einer Spline-Funktion und verallgemeinert auf einen Strafterm, der Rauheit im Sinne der k-ten Ableitung bestraft, schreiben als

$$l_{\text{pen}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2}\lambda \sum_{j=k+1}^{d} (\Delta^k \theta_j)^2. \tag{2.36}$$

Für das Schätzverfahren ist es günstig, den (bzw. die) Strafterm(e) in der Matrixnotation $\boldsymbol{\theta}^\top K_k \boldsymbol{\theta}$ mit einer sogenannten Strafmatrix $\boldsymbol{K}_k$ auszudrücken. Zunächst bildet man eine sogenannte Differenzenmatrix $\boldsymbol{D}_k$, die multipliziert mit $\boldsymbol{\theta}$ den Vektor der $k$ten Differenzen der θ_i ergibt. Für $k = 1$ hat diese Dimension $(d-1) \times d$ und ist gegeben durch:

$$\boldsymbol{D}_1 = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \tag{2.37}$$

Differenzmatrizen höheren Ordnung lassen sich daraus rekursiv berechnen durch:

$$\boldsymbol{D}_k = \boldsymbol{D}_1 \boldsymbol{D}_{k-1}. \tag{2.38}$$

Damit ergibt sich die Strafmatrix $\boldsymbol{K}_k$ als

$$\boldsymbol{K}_k = \boldsymbol{D}_k^\top \boldsymbol{D}_k. \tag{2.39}$$

Die Idee der Bestimmung von datengesteuert optimierten Glättungsparametern wird nun für normalverteilte Fehler und aufgrund der einfacheren Notation für lediglich eine Spline-Funktion als Einflussgröße motiviert und dann auf den generalisierten Fall übertragen. Eine zu schätzende Funktion $\widehat{f}$ soll so gewählt werden, dass die Abweichungen von der wahren Funktion und die Varianz der Schätzung möglichst gering sind. Da sich diese beide Größen nicht simultan verringern lassen, verwendet man die erwartete quadratische

Abweichung, den sogenannte MSE, der sich additiv in quadratische Abweichung und Varianz zerlegen lässt. Dieser ist nur punktweise definiert, weshalb man als globales Maß den über die Beobachtungen gemittelten MSE betrachtet:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\widehat{f}(z_i)-f(z_i))^2. \tag{2.40}$$

Als empirische Approximation an diesen Erwartungswert erscheint es auf den ersten Blick sinnvoll, das Mittel der quadrierten Residuen zu betrachten. Dieses wird aber immer durch die den Daten am besten angepasste, d.h. die raueste Funktion, also für $\lambda = 0$ minimiert. Es sollen aber neue Daten bestmöglich prognostiziert werden. Da diese aber (i.d.R.) nicht zur Verfügung stehen, bildet man für alle Beobachtungen z_i die quadrierten Differenzen dieser von der geschätzten Funktion $\widehat{f}^{(-i)}(z_i)$ bei Auslassung der jeweiligen Beobachtung und erhält so das sog. Kreuzvalidierungskriterium (CV):

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{f}^{(-i)}(z_i))^2, \tag{2.41}$$

wobei y_i die normalverteilte Zielgröße ist. Es wäre sehr aufwendig, dieses Kriterium in dieser Form für verschiedene λ-Werte zu berechnen, da man das Modell jeweils n-mal fitten müsste. Es lässt sich aber zeigen, dass man es auf folgende Art umformen kann, sodass nur eine Schätzung des gesamten Modells vorliegen muss:

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\widehat{f}(z_i)}{1-h_{ii}}\right)^2. \tag{2.42}$$

h_{ii} ist dabei das i-te Diagonalelement der sogenannten Hat-Matrix $\boldsymbol{H}$, welche multipliziert mit dem Vektor der Beobachtungen $\boldsymbol{y}$ den Vektor der Prädiktionen $\widehat{\boldsymbol{y}}$ ergibt. Wenn $\boldsymbol{X}$ die Modellmatrix bezeichnet, in diesem beispielhaften Fall also die Matrix, die die Werte der Basisfunktionen von $\widehat{f}$ an den Beobachtungen enthält, ist $\boldsymbol{H}$ bei normalverteilten Fehlern gegeben durch $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Der Erwartungswert von CV ist ungefähr gleich dem mittleren quadratischen Fehler bei n Beobachtungen mit Kovariablen z_i, den man bei Vorhersage mit $\widehat{f}$ machen würde. Da die Berechnung der (Diagonalen der) Hat-Matrix numerisch aufwendig ist, ersetzt man die h_{ii} in (2.42) durch ihr Mittel, also $\mathrm{sp}(\boldsymbol{H})/n$, wodurch sich das sogenannte *generalisierte* Kreuzvalidierungskriterium (GCV)

ergibt:

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f}(z_i)}{1 - \mathrm{sp}(\boldsymbol{H})/n}\right)^2. \tag{2.43}$$

Zur Gewinnung eines äquivalenten Kriteriums für den generalisierten Fall macht man sich das Analogon des IRLS-Verfahrens aus Abschnitt 2.4 für die penalisierte Likelihood-Schätzung, das penalisierte IRLS- oder kurz P-IRLS-Verfahren zu Nutze. Bei Letzterem minimiert man in jeder Iteration äquivalent eine penalisierte gewichtete Quadratsumme. Im Falle eines *gewichteten* linearen Modells wäre eine zu (2.43) äquivalente Größe gegeben durch:

$$\frac{n|\sqrt{\boldsymbol{W}}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}|^2}{[n - \mathrm{sp}(\boldsymbol{H})]^2}. \tag{2.44}$$

Dabei ist $\boldsymbol{W}$ die Gewichtungsmatrix, $\boldsymbol{z}$ der Vektor der Zielgrößen und $\boldsymbol{H}$ die Hatmatrix - nun mit Gewichten. Wenn nun $\boldsymbol{W}$ und $\boldsymbol{z}$ die Arbeitsgewichtungsmatrix bzw. den Vektor der Arbeitsbeobachtungen im P-IRLS-Algorithmus bei Konvergenz bezeichnen, so lässt sich zeigen, dass sich die log-Likelihood von $\boldsymbol{\theta}$ in einer Umgebung von $\hat{\boldsymbol{\theta}}$ bis auf eine additive Konstante auf folgende Art approximieren lässt:

$$-\frac{1}{2}|\sqrt{\boldsymbol{W}}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\theta}|^2 \tag{2.45}$$

Berücksichtigt man nun, dass die Devianz $D(\cdot)$ bis auf eine additive Konstante gleich minus zweimal der log-Likelihood ist, erhält man schließlich gemäß (2.44) ein GCV-Kriterium für den generalisierten Fall:

$$\mathrm{GCV}_{\mathrm{gen}} = \frac{nD(\hat{\boldsymbol{\theta}})}{(n - \mathrm{sp}(\boldsymbol{H}))^2}, \tag{2.46}$$

wobei $\boldsymbol{H}$ die Hatmatrix des gewichteten linearen Modells im IRLS-Algorithmus bei Konvergenz ist. Die Minimierung von $\mathrm{GCV}_{\mathrm{gen}}$ ist besonders günstig, wenn der Skalenparameter ϕ unbekannt ist.

Für bekannten Skalenparameter ist eine Verallgemeinerung von Mallow's C_p auf den generalisierten Fall geeigneter. Mallow's C_p ist ein Informationskriterium für lineare Modelle. Es entspricht der geschätzten Quadratsumme der Abweichungen der prognostizierten von den wahren Werten und lässt sich für den *gewichteten* Fall formulieren

als:

$$C_p \;=\; |\sqrt{\boldsymbol{W}}(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\theta}})|^2/\phi \;+\; 2\,\mathrm{sp}(\boldsymbol{H}) \;-\; n. \tag{2.47}$$

Verwendet man wieder die Approximation (2.45) erhält man ein entsprechendes Kriterium für generalisierte lineare Modelle:

$$C_{p_{\mathrm{gen}}} \;=\; D(\hat{\boldsymbol{\theta}})/\phi \;+\; 2\,\mathrm{sp}(\boldsymbol{H}) \;-\; n. \tag{2.48}$$

Das in der Analyse verwendete Paket `mgcv` bietet nun zwei Möglichkeiten um die hinsichtlich eines der beiden Kriterien optimalen Glättungsparameter zu bestimmen, die „performance iteration" und die „outer iteration". Bei Ersterer werden die Kriterien für jede Iteration des P-IRLS-Verfahrens minimiert und bei Letzterer wird für jedes Test-Set der Glättungsparameter das P-IRLS-Verfahren bis zur Konvergenz iteriert und der Wert des Kriteriums bestimmt. „Performance iteration" ist rechnerisch effizienter, hat aber den entscheidenden Nachteil, dass das Verfahren nicht immer konvergiert, da Schleifen in der Iteration auftreten können. Deshalb wurde in der Analyse bis auf bei der Modellierung mit der negativen Binomialverteilung immer „outer iteration" verwendet. Aufgrund der Komplexität von GAM-Modellen mit vielen zu schätzenden Parametern und Kollinearitäten in den Spline-Basen kann es leicht dazu kommen, dass die Modellmatrix annähernd keinen vollen Spaltenrang besitzt. Daher ist eine numerisch stabile Minimierung der Kriterien sehr wichtig. In `mgcv` ist ein Verfahren implementiert, das dies gewährleistet und gleichzeitig rechnerisch effizient ist. Allerdings ist es technisch sehr aufwendig, sodass an dieser Stelle auf eine Darstellung verzichtet und auf Wood (2006a) verwiesen wird.

2.6. Maßzahlen

Die im Folgenden behandelten Maßzahlen sind Vertreter einer größeren Gruppe von Maßen, mit denen die Beziehung der Punkte eines Punktprozesses zueinander untersucht wird. Sie können auch, bei Verwendung von Simulationsmethoden in sehr vielfältiger

Weise zum Testen von Hypothesen verwendet werden, siehe z.B. Møller und Waagepetersen (2004). Aus theoretischer Sicht können die Maßzahlen vor dem Hintergrund der sogenannten Palm-Verteilungen betrachtet werden. Diese zählen allerdings zu den komplexeren Themen in der Punktprozess-Theorie und sollen daher hier nur auf intuitive Weise nähergebracht werden. Heuristisch ausgedrückt, werden Palm-Verteilungen dazu verwendet, den Punktprozess von einem typischen, d.h. zufällig ausgewählten Punkt aus zu beschreiben (Heveling, 2006). Aus theoretischer Sicht sind Palm-Charakteristiken Wahrscheinlichkeiten oder Erwartungswerte für festgelegte Eigenschaften eines Punktprozesses, bedingt darauf, dass der Prozess einen Wert im Ursprung hat (wenn dabei in allen Fällen die Beobachtung im Ursprung außen vorgelassen wird). Da allerdings die Wahrscheinlichkeit, dass sich ein Punkt exakt an einer bestimmten Stelle befindet, Null ist, ist das ein schwieriges Konzept. Bei der Schätzung verschiebt man für jeden Punkt das Punktmuster so, dass dieser im Ursprung liegt und berechnet dabei jeweils die Ausprägung des betrachteten Sachverhalts. Eine Mittelung über alle Ausprägungen ergibt schließlich eine Schätzung für die Wahrscheinlichkeit bzw. den Erwartungswert des Ereignisses, wenn der Punktprozess um einen Punkt zentriert ist und selbst nicht mitgezählt wird (Illian et al., 2008).

2.6.1. K-Funktion

Ripley's K-Funktion und die sich aus dieser ableitenden Besag's L-Funktion und Paar-Korrelationsfunktion $g(r)$ werden häufig als die wichtigsten Maßzahlen bei der Analyse von räumlichen Punktprozessen angesehen. Man möchte Aufschluss über die erwartete Anzahl $\lambda K(r)$ an Punkten in einer Entfernung r um einen typischen Punkt. Um den Einfluss der globalen Intensität λ und lokaler Dichtefluktuationen herauszufiltern, teilt man den Erwartungswert durch λ und erhält so die K-Funktion $K(r)$ (Illian et al., 2008).

Die K-Funktion ist demnach definert als

$$K(r) \;=\; \frac{1}{\lambda}\,\mathbb{E}[N(\mathbf{Y} \cap b(u,r) \setminus \{u\} \,|\, u \in \mathbf{Y}] \tag{2.49}$$

Hier würde bei sich gegenseitig anziehenden Punkten ein steiler Kurvenverlauf zu beobachten sein und entsprechend bei konkurrierendem Verhalten der Punkte ein langsamer. Für komplett unabhängige Punkte, d.h. bei Vorliegen eines homogenen Poissonprozesses hat die K-Funktion die Form (Baddeley, 2008):

$$K_{\text{pois}}(r) \;=\; \pi r^2 \tag{2.50}$$

Schätzung

Ein naiver Schätzer wäre:

$$\widehat{K}(r) \;=\; \frac{1}{\widehat{\lambda} n} \sum_{i=1}^{n} N(b(x_i, r) \setminus \{x_i\}) \tag{2.51}$$

λ wird hier und im Weiteren durch $N(W)/|W|$ geschätzt. Mit (2.51) wird nicht berücksichtigt, dass dadurch, dass das Beobachtungsfenster W beschränkt ist, Punkte nicht beobachtet werden, wenn sie sich außerhalb von W befinden. So wird die Anzahl an Punkten um einen Punkt x_i innerhalb des Radius r unterschätzt, wenn sich dieser näher als r zum Rand befindet, weil sich außerhalb von W weitere Punkte innerhalb des Radius r befinden können.

Es werden im Folgenden drei Möglichkeiten zur Korrektur dieser Randeffekte vorgestellt. Die einfachste, aber auch ineffizienteste ist der sogenannte reduced-sample-Schätzer. Der Name resultiert daher, dass bei der Schätzung jeweils nur die Punkte berücksichtigt werden, die einen größeren Abstand als r zum Rand haben und um diese damit sicher alle Punkte innerhalb des Radiuses r beobachtet werden. Er ist gegeben durch (Ripley, 1988):

$$\begin{aligned} \widehat{K}_{\text{rs}}(r) \;&=\; \frac{1}{\widehat{\lambda} m}\,\#\{(x_i, y_i) : ||x_i - y_i||_2 \le r,\, x_i \in W \ominus b(0, t_i)\} \\ \text{mit} \quad m \;&=\; N(W \ominus b(0, t_i))\} \end{aligned} \tag{2.52}$$

Der isotrope Schätzer nutzt die Eigenschaft homogener Poissonprozesse, dass sie invariant gegenüber Drehungen sind. Die Punkte y_i werden gewichtet mit dem Kehrwert des Anteils der Kreislinie eines Kreises mit Radius r und Zentrum y_i, der sich in W befindet. Wenn sich Punkte nahe am Rand befinden, ist ihr Gewicht also groß, weil sich viele Punkte in Distanz r außerhalb von W befinden könnten. In den meisten Fällen ist das Gewicht 1, nämlich dann, wenn sich der komplette Kreis in W befindet. Die Formel zur Schätzung lautet:

$$\widehat{K}_{\text{iso}}(r) \;=\; \frac{1}{\widehat{\lambda^2}|W|} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{|x_i - x_j| \leq r\} \frac{2\pi |x_i - x_j|}{\nu(\partial b(x_i, |x_i - x_j|) \cap W} \tag{2.53}$$

Hierbei ist $\partial b(a, b)$ die Kreislinie des Kreises um a mit Radius b, ν das eindimensionale Lebesgue-Maß und $\widehat{\lambda^2}$ wie auch im Weiteren $N(W)(N(W)-1)/|W|^2$. Dieser Schätzer ist für r-Werte kleiner gleich dem Umkreisradiuses von W approximativ unverzerrt, vorausgesetzt W ist konvex. Ohser (1983) führte einen Schätzer ein, der für alle Werte kleiner gleich dem Durchmesser von W approximativ unverzerrt ist, wenn W konvex ist:

$$\widehat{K}_{\text{isoohser}}(r) \;=\; \frac{1}{\widehat{\lambda^2}|W|} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{|x_i - x_j| \leq r\} \frac{2\pi |x_i - x_j|}{\nu(\partial b(x_i, |x_i - x_j|) \cap W)} \frac{|W|}{|W_{|x_i - x_j|}|} \tag{2.54}$$

Die Gewichte bei $\widehat{K}_{\text{iso}}(r)$ werden also noch mit dem Faktor $|W|/|W_{|x_i - x_j|}|$ multipliziert, wobei $W_{|x_i - x_j|}$ die Menge aller $z \in W$ ist, die von mindestens einem anderen $z_1 \in W$ in Entfernung $|x_i - x_j|$ liegen (Loh und Stein, 2004). Mit Hilfe dieses Faktors wird berücksichtigt, dass Distanzen größer als der Umkreisradius nicht überall im Beobachtungsfenster auftreten können.

Bei der sogenannten Translationskorrektur werden ähnlich wie beim isotropen Schätzer die Punkte-Paare mit der inversen Wahrscheinlichkeit, sie zu beobachten, gewichtet, allerdings auf Basis von Parallelverschiebungen und nicht Drehungen (Schladitz und Baddeley, 2000). Daher wird hier keine Isotropie vorausgesetzt, sondern lediglich Stationarität. Betrachtet man das Wertepaar (x_i, y_i), so ist das Gewicht der Anteil der Translationen von x_i innerhalb von W, zu denen y_i dabei auch in W bleibt. Dieser An-

teil lässt sich berechnen, indem man W so verschiebt, dass einmal x_i im Ursprung liegt und einmal y_i, den Schnitt dieser beiden Flächen bildet und die Fläche dieser Menge durch diejenige von W teilt. Damit ergibt sich der Translationsschätzer zu (Kühlmann-Berenzon, 2002):

$$\widehat{K}_{\text{trans}}(r) = \frac{1}{\widehat{\lambda^2}|W|} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{1}\{|x_i - x_j| \leq r\} \frac{|W|}{|W_{x_i} \cap W_{x_j}|}$$
$$\text{mit} \quad W_{x_{i/j}} = \{w + x_{i/j} : w \in W\} \tag{2.55}$$

2.6.2. L-Funktion

In heutigen Anwendungen wird zumeist die L-Funktion anstatt der K-Funktion zur Analyse von Korrelationen in Punktmustern verwendet. Sie ist definiert als

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \tag{2.56}$$

Wie in Abschnitt 2.6.1 gesehen, hat die K-Funktion im Falle eines homogenen Poissonprozesses die Form $K(r) = \pi r^2$. Teilt man dies durch π und zieht die Wurzel, so erhält man für die L-Funktion $L(r) = r$. Also müssen zur Überprüfung der Annahme eines homogenen Poissonprozesses nur Abweichungen von der Diagonalen betrachtet werden, anstatt wie bei der K-Funktion von einer parabelförmigen Kurve. Außerdem nehmen Schwankungen der K-Funktion mit steigendem r immer weiter zu. Diese werden durch die Wurzeltransformation sowohl hinsichtlich des Mittels als auch der Varianz stabilisiert bzw. können sogar unabhängig von r werden.

Schätzung

Mit einer der in Abschnitt 2.6.1 vorgestellten Methoden wird die K-Funktion geschätzt und in die theoretische Formel für die L-Funktion eingesetzt, also (Illian et al., 2008):

$$\widehat{L}(r) = \sqrt{\frac{\widehat{K}(r)}{\pi}} \tag{2.57}$$

3. Datenmaterial

3.1. Repräsentation räumlicher Daten und Koordinatensysteme

Abhängig von der beinhalteten Information sowie dem Verwendungszweck werden räumliche Daten grundsätzlich auf zwei unterschiedliche Arten repräsentiert: als Raster- und als Vektordaten (Neteler und Mitasova, 2004). Das nötige Grundwissen zu diesen beiden Typen soll im nächsten Abschnitt vermittelt werden. Unabhängig von dem verwendeten Datentyp muss es immer möglich sein, Lokationen auf der Erde eindeutig zu identifizieren. Dies leisten Koordinatensysteme. Der darauffolgende Abschnitt führt, nach einer kurzen Einführung in Kartenabbildungen, das in dieser Arbeit verwendete Gauß-Krüger-Koordinatensystem ein.

3.1.1. Raster- und Vektor-Daten-Modell

Ein Raster ist eine regelmäßige Matrix bestehend aus Werten. Im Raster-Daten-Modell können die Werte entweder Raster-Punkten oder Raster-Zellen zugeordnet werden. Im ersten Fall soll üblicherweise ein Feld mit stetigen Werten spezifiziert werden, im zweiten Fall, dem sogenannten *Image*, sind die Ausprägungen für gewöhnlich kategorial und es können auch mehrere Attribute zugewiesen werden. Die *Auflösung* eines Image ist definiert als die Länge einer Raster-Zelle, genannt *Pixel*. Das Raster-Modell hat den großen Vorteil, dass es eine sehr einfache Struktur besitzt, die die Analyse und den

Umgang mit den Daten allgemein vereinfacht. Es wird häufig für geografische Merkmale, wie die Höhe über dem Meerespiegel oder die Vegetation verwendet, eignet sich aber weniger, um geografische Strukturen abzubilden, die stark von Linien, z.B. Grenzen, abhängen.

Hierfür verwendet man besser das Vektor-Daten-Modell, das geografische Charakteristiken über Linien, Punkte oder Flächen spezifiziert. Dies geschieht über sich nicht schneidende Linien, den sogenannten *Bögen*. Diese werden über eine Reihe von Punkten, den *Eckpunkten*, spezifiziert durch Koordinaten (x, y) repräsentiert, deren Anfang- und Endpunkte als *Knoten* bezeichnet werden. Zwei aufeinanderfolgende *Eckpunkte* heißen *Bogen-Segment*. Über Identifikationsnummern können den Bögen und damit den durch sie definierten Objekte Attribute zugeordnet werden (Neteler und Mitasova, 2004). Z.B. haben Straßen in den später vorgestellten Straßen-Daten das Attribut Straßentyp.

3.1.2. Kartenabbildungen und Koordinatensysteme

Mit Hilfe von Kartenabbildungen kann die dreidimensionale Struktur der Erdoberfläche auf eine zweidimensionale Fläche, also eine Karte übertragen werden (Reimann et al., 2008). Die Abbildung geschieht allgemein auf eine sog. abwickelbare Fläche, d.h. eine Fläche, die ohne Lücken oder Verdopplungen auf einer Ebene entfaltet werden kann (Domingues, 2008). Unterschieden werden können Projektionen auf Zylinder, Kegel und Flächen, wobei die abwickelbare Fläche die Erde jeweils schneiden oder tangieren kann. Punkte bzw. Linien, an denen die Erde tangiert bzw. geschnitten wird, nennt man *Standardpunkte* bzw. *-linien*. Ist die abwickelbare Fläche um 90° gedreht, stellt die Standardlinie also einen Meridian (Zentralmeridian) der Erde da, so spricht man von einer *transversalen* Projektion. Eine Projektion kann nur entweder winkeltreu, äquidistant oder flächentreu sein und ist hinsichtlich der jeweils anderen Aspekte verzerrt. Welcher Typ gewählt wird, hängt von der jeweiligen Anwendung ab (Neteler und Mitasova, 2004). Das in der Arbeit verwendete, weiter unten beschriebene Gauß-Krüger-Koordinatensystem basiert auf einer winkeltreuen Projektion (Imhof, 1972).

Ein Koordinatensystem ist durch Festlegung des Ursprungs, der Koordinatenachsen und der Einheiten eindeutig festgelegt. Nach Erstellung der Karte über Projektion wird üblicherweise ein kartesisches Koordinatensystem verwendet. Um negative Koordinaten zu vermeiden, können zum Ursprung $(0, 0)$ in beiden Dimensionen Konstanten addiert werden, bezeichnet als *false easting* bzw. *false northing* (Neteler und Mitasova, 2004).

Gauß-Krüger-Koordinatensystem

Dieses in der Arbeit verwendete Koordinatensystem ist in Deutschland gebräuchlich. Wie bereits erwähnt, basiert es auf einer winkeltreuen, genauer einer transversalen, zylindrischen Projektion, wobei der Zylinder die Erde tangiert. Wenn die gesamte Erde auf den Zylinder projiziert würde, wäre die Verzerrung weiter weg von der Standardlinie sehr groß. Deshalb wird die Erde ausgehend vom Nullmeridian als Standardlinie in drei Grad breite Streifen zerlegt und jeder dieser Streifen unter Verwendung des mittleren Meridians, dem *Hauptmeridian*, als Standardlinie, auf die Zylinderoberfläche projiziert. So ergibt sich für jeden Streifen eine eigene Karte. Als Ursprung des Koordinatensystems wird jeweils der Schnittpunkt des Hauptmeridians mit dem Äquator gewählt. Die Entfernungen vom Ursprung werden in Metern angegeben, wobei, um negative Koordinaten zu vermeiden zum Rechtswert die Konstante 500.000 m addiert wird (false easting von 500.000 m). Zur Unterscheidung der Bezugssysteme wird den Rechtswerten eine Kennziffer vorangestellt, die ein Drittel der Gradzahl des Hauptmeridians beträgt. Für Deutschland relevant sind die Kennziffern 2,3 und 4, was also den Hauptmeridianen 6°, 9° und 12° entspricht (Olbrich et al., 2002). Obgleich man sich im Westen Bayerns nach der oben beschriebenen Prozedur auf den Hauptmeridian 9° beziehen müsste, sind die Koordinaten in dieser Arbeit nur bezogen auf den Hauptmeridian 12° angegeben, wobei die dadurch entstehende zusätzliche Verzerrung als gering angesehen werden kann. Das entsprechende Koordinatensystem wird als EPSG:31468 bezeichnet (Wikipedia, 2011). Das Institut für Statistik in München hat z.B. die Gauß-Krüger-Koordinaten $(4468973, 5334986)$. Die erste Ziffer des Rechtswertes ist eine vier, d.h. der Hauptmedian des Koordinatensystems

ist 12° und das Institut befindet sich 500.000 m − 334986 m = 165014 m davon ist westlicher Richtung. Es liegt zudem 5334986 m nördlich vom Äquator.

3.2. Wildunfalldaten

Die Daten zu den Rehunfällen stammen von der Polizei. Zur Analyse der räumlichen Wildunfallintensität lagen zunächst insgesamt Informationen zu 45821 Unfällen in Bayern vor. Von diesen konnten 10 nicht berücksichtigt werden, da sie sich nicht in dem die Grenzen Bayerns repräsentierenden Polygons (`bayernfenster`) befanden. In 21 Fällen wurden Unfälle doppelt aufgenommen, nach Entfernung derer schließlich zur Analyse 45790 Wildunfälle zur Verfügung standen, wovon 18834 im Jahr 2006 und 26956 im Jahr 2009 stattfanden.

3.3. Daten zum Wildverbiss

Diese Daten wurden auf einem Raster der Auflösung 1225 m von der bayerischen Forstverwaltung erhoben. An jedem berücksichtigten Gitterpunkt wurde die nächstgelegene geeignete Fläche zur Untersuchung der Anteile an verbissenen Setzlingen der Höhe 20-130 cm gewählt und entlang eines Transektes jeweils 75 solcher Pflanzen auf Verbiss der Leittriebe untersucht. Dabei wurde auch die Baumart erhoben. Der Anteil an Setzlingen mit Leitverbiss ist ein gängiges Maß, um die Verbissintensität einzuschätzen (Hothorn et al., in Vorb.). In der Analyse wurden nur Fichten (`verbissfichte`), Rotbuchen (`verbissbuche`) und Eichen (`verbisseiche`) betrachtet. Da die Präferenz von Rehen für die Baumarten unterschiedlich ist, wurden diese getrennt betrachtet. Die Informationen zum Wildverbiss stehen nicht an allen Lokationen zur Verfügung. Außerdem können an Lokationen, an denen sich keine Bäume einer Baumart finden, keine Anteile berechnet werden. Daher mussten die relativen Häufigkeiten der verbissenen Pflanze pro Baumart

geglättet werden mussten, um für alle Lokationen in Bayern die jeweiligen (prognostizierte) Verbissanteile zur Verfügung zu haben, vgl. Abschnitt 2.3. Die Glättung diente hier also lediglich dazu, die fehlenden Werte an den nicht-erhobenenen Lokationen zu imputieren, weshalb die geglätteteten Werte an den erhobenenen Lokationen nah an den tatsächlichen relativen Häufigkeiten liegen sollten. Demnach sollte die resultierende Oberfläche sehr rau sein. Es wurde ein adaptiver Kern verwendet, da dieser in der vorliegenden Situation geeigneter war, siehe weiter unten. Die in Davies et al. (2011) vorgestellte Methode unter Verwendung eines adaptiven Kerns ist für Kerndichteschätzung, nicht für Glättung konzipiert, lässt sich aber auch auf die Glättung der Wildverbissdaten übertragen.

Wenn x_i den Lokationen der Beobachtungen entsprechen, geschieht klassische bivariate Kerndichteschätzung nach der folgenden Formel:

$$\widehat{f}(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.1}$$

Dabei bezeichnet man h als Bandbreite und $K(\cdot)$ als Kernfunktion. Die Kernfunktion ist selbst eine stetige Dichtefunktion mit Erwartungswert $\mathbf{0}$ und endlicher Varianz (Wasserman, 2005).

In der adaptiven Kerndichteschätzung wird die fixe Bandbreite h durch eine für jede Beobachtung i individuelle Bandbreite $h_i = h_0 f(x_i)^{-1/2} \gamma^{-1}$ ersetzt. Dabei ist h_0 die sogenannte globale Bandbreite, welche steuert, wie stark die Beobachtungen generell geglättet werden und γ das geometrische Mittel der $f(x_i)^{-1/2}$, dessen Aufnahme die Abhängigkeit der h_i von der Skala der beobachteten Daten abschwächt. In der Praxis ist f nicht bekannt und muss ebenfalls geschätzt werden. Diese sogenannte Pilotdichte gewinnt man über einen Kerndichteschätzer mit fixer Bandbreite $\tilde{h}$, genannt Pilot-Bandbreite. Je höher die Dichte ist, desto kleiner werden die Bandbreiten. In der Dichteschätzung hat das den Zweck, dass die verfügbare Information besser genutzt wird, da, wenn an Lokationen die Dichte höher ist, dort mehr Punkte vorliegen, die Schätzung also flexibler sein kann. Im Rahmen der Glättung der Wildverbissdaten wird hingegen

durch die flexiblen Bandbreiten verhindert, dass in Bereichen mit weniger Punkten zu starke Schwankungen auftreten. Vielmehr orientiert sich die Glättung dort stärker an den verfügbaren Werten.

Während bei der Kerndichteschätzung der Schätzwert an der betrachteten Lokation das Mittel der Werte der Kerne an dieser Stelle ist, wird dort bei der Glättung ein mit den Kernwerten gewichtetetes Mittel der Beobachtungen y_i an Lokationen x_i berechnet (Hastie et al., 2009):

$$\widehat{f}(z) = \sum_{i=1}^{n} h^{-2} K\left(\frac{x - x_i}{h}\right) y_i \Big/ \sum_{i=1}^{n} h^{-2} K\left(\frac{x - x_i}{h}\right) \tag{3.2}$$

Wenn s_i die zu glättenden Werte an den Lokationen x_i bezeichnen, so lautet die Formel für die Glättung mit adaptiver Bandbreite also:

$$\widehat{f}(z) = \sum_{i=1}^{n} h_i^{-2} K\left(\frac{z - x_i}{h_i}\right) s_i \Big/ \sum_{i=1}^{n} h_i^{-2} K\left(\frac{z - x_i}{h_i}\right) \tag{3.3}$$

Die Schätzung wurde mit Hilfe des R-Pakets `sparr` (Davies et al., 2011) durchgeführt, wobei Abänderungen in der `bivariate.density`-Funktion gemacht werden mussten, um das für Kerndichteschätzung implementierte Verfahren auf die Glättung zu übertragen. Außerdem musste eine Änderung in der Funktion gemacht werden, um dieselbe Auflösung wie in den Originaldaten zu erhalten, da zum Zeitpunkt der Arbeit nur Schätzungen auf einem $K \times K$-Raster möglich waren.

Die Auswahl geeigneter globaler Bandbreiten für die verschiedenen Baumarten geschah durch Ausprobieren nach Augenmaß. Dazu wurden Plots prognostizierte gegen tatsächliche Werte betrachtet. Für alle drei Baumarten wurde auf diese Art die globale Bandbreite 750 gewählt, siehe Abbildung 3.1. Die Schätzung erwies sich als wenig sensitiv hinsichtlich der Wahl der Pilot-Bandbreite, diese wurde auf 2000 festgesetzt. Um zu überprüfen, ob die Verwendung von adaptiven Bandbreiten tatsächlich zu einer Verbesserung führt, wurden zum Vergleich konventionelle Glättungen mit fixen Bandbreiten 750 berechnet. Während die Prognosen dabei an den Datenpunkten zu allermeist sehr ähnlich denen bei adaptiver Glättung waren, war die geglättete Oberfläche in Gebieten, in denen über

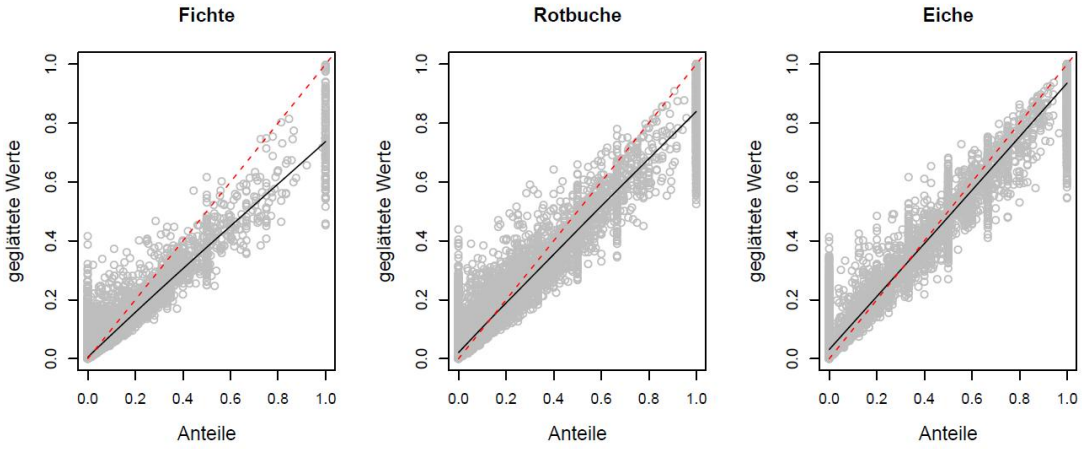

Abbildung 3.1.: mit adaptiver Glättung prognostizierte gegen wahre Anteile mit Loess-Kurve - Pilot-Bandbreite: 2000, globale Bandbreite: 750

große große Strecken keine Werte zur Verfügung standen zum Teil extrem rauh. Dabei kam es mitunter zu extremen Abweichungen weit außerhalb des Intervals $[0, 1]$.

Abbildung 3.2 zeigt die Karten für die drei Baumarten. Fichten werden offenbar im Gegensatz zu Buchen und vor allem Eichen, trotz der großen Prävalenz kaum verbissen.

3.4. CORINE[1] Land Cover-Daten

Die folgenden Ausführungen beziehen sich im Wesentlichen auf Keil et al. (2010).

Eingebunden in das europaweite, von der europäischen Umweltagentur EEA initiierte Projekt „IMAGE & CORINE Land Cover 2006" stellt „CORINE Land Cover 2006 - Germany" ein Projekt zur Erfassung und Klassifizierung der Landnutzung und -bedeckung für Deutschland dar. Für 1990 waren bereits entsprechende Daten für Deutschland erhoben worden, die bis Ende 2004 für das Jahr 2000 erstmalig aktualisiert wurden. Im

[1] **Coor**dinated **In**formation on the **E**nvironment

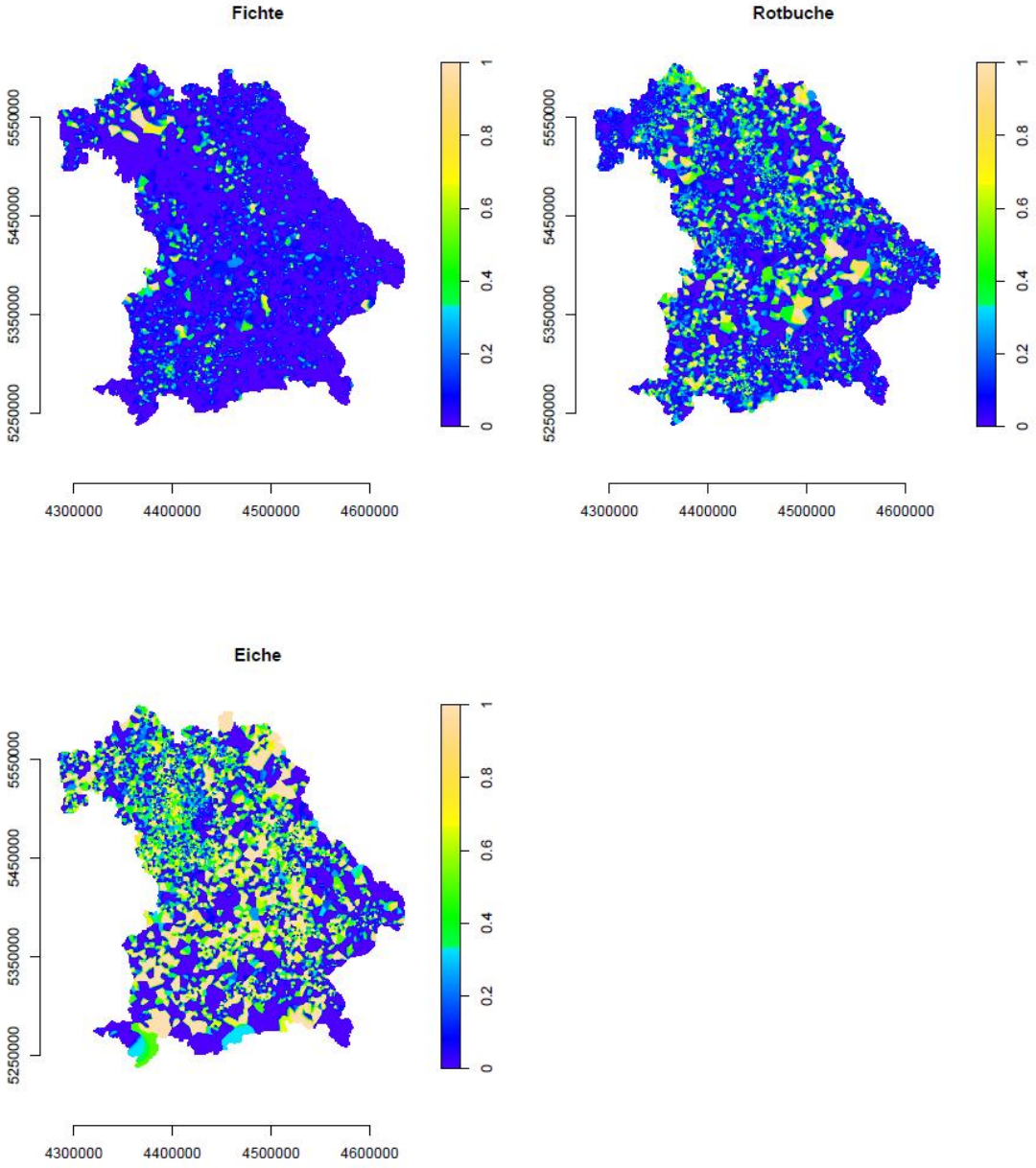

Abbildung 3.2.: mit adaptiver Glättung prognostizierte Anzahlen von verbissenen Bäumen getrennt nach Baumart

Rahmen von „CORINE Land Cover 2006 - Germany" wurde Anfang 2010 eine zweite Aktualisierung abgeschlossen, in der auch Veränderungen gegenüber dem Jahr 2000 kartiert wurden.

Das Datenmaterial basiert auf Satellitendaten im Maßstab 1:100.000, anhand derer die Landnutzung und Bodenbedeckung in Bayern in 33 (für Europa 44) Kategorien unterteilt wurde. Die einzelnen Kategorien können in hierarchischer Weise zwei übergeordneten Klassen von Kategorien zugeordnet werden, siehe Tabellen 3.1 und 3.2. Es stehen sowohl Vektor- wie auch Rasterdaten zur Verfügung, wobei sich für den vorliegenden Zweck das Rasterformat als günstiger erwies. Es stehen die Auflösungen 100 m, 250 m bzw. 1000 m zur Verfügung, wobei in dieser Arbeit die Auflösung 100 m gewählt wurde. Diese Karten liegen als TIFF-Dateien und im sogenannten ETRS89 Lambert Azimuthal Equal Area (LAEA)-Koordinaten(referenz)system vor. Mit Hilfe von GDAL einem Kommandozeilenprogramm, das Teil von FWTools ist, einer Sammlung von Open-Source-Programmen für GIS-Systeme, wurden die Koordinaten in das Gauß-Krüger-System mit Hauptmeridian 12° transformiert. Da der ursprüngliche Datensatz große Teile Europas enthielt, wurde anschließend mit Hilfe von Adobe Photoshop CS5.1 Version 12.1 ein Ausschnitt für Bayern erstellt. Um dabei die in der TIFF-Datei enthaltenen Informationen zum Koordinatensystem beizubehalten, bot sich das Photoshop Add-On Geographic Imager (Version 3.3.0) von Avenza an. Aufgrund der einfacheren Interpretation und um nicht zu viele Koeffizienten schätzen zu müssen und damit die Varianz der Prognose zu erhöhen wurden die Kategorien thematisch zu den Oberkategorien „Bebaute Flächen" (`bebaut`), „Landwirtschaftliche Flächen" (`landwirt`), „Wälder und naturnahe Flächen" (`wald`) und „Feucht- und Wasserflächen" (`sonstige`) zusammengefasst. Abbildung 3.3 zeigt die räumliche Verteilung dieser Kategorien. Offenbar nimmt landwirtschaftlich genutztes Gebiet den größten Raum ein (56,2 %), gefolgt von Wäldern und naturnahen Flächen (36,4 %). Weit abgeschlagen davon folgen bebautes Gebiet (6,0 %) und Feucht- bzw. Wasserflächen (1,2 %).

Tabelle 3.1.: berücksichtigte Kategorien aus den CORINE-Daten - I

Ebene 1	Ebene 2	Ebene 3
Bebaute Flächen	Städtisch geprägte Flächen	Durchgängig städtische Prägung
		Nicht durchgängig städtische Prägung
	Industrie-, Gewerbe- und Verkehrsflächen	Industrie- und Gewerbeflächen, öffentliche Einrichtungen
		Straßen-, Eisenbahnnetze und funktionell zugeordnete Flächen
		Hafengebiete
		Flughäfen
	Abbauflächen, Deponien und Baustellen	Abbauflächen
		Deponien und Abraumhalden
		Baustellen
	Künstlich angelegte, nicht landwirtschaftlich genutzte Grünflächen	Städtische Grünflächen
		Sport- und Freizeitanlagen
Landwirtschaftliche Flächen	Ackerflächen	Nicht bewässertes Ackerland
	Dauerkulturen	Weinbauflächen
		Obst- und Beerenobstbestände
	Grünland	Wiesen und Weiden
	Landwirtschaftliche Flächen heterogener Struktur	Komplexe Parzellenstrukturen
		Landwirtschaftlich genutztes Land mit Flächen natürlicher Bodenbedeckung von signifikanter Größe
		Land- und forstwirtschaftliche Flächen

Tabelle 3.2.: berücksichtigte Kategorien aus den CORINE-Daten - II

Ebene 1	Ebene 2	Ebene 3
Wälder und naturnahe Flächen	Wälder	Laubwälder
		Nadelwälder
		Mischwälder
	Strauch- und Krautvegatation	Natürliches Grünland
		Heiden und Moorheiden
		Wald-Strauch-Übergangsstadien
	Offene Flächen ohne / mit geringer Vegetation	Strände, Dünen und Sandflächen
		Felsflächen ohne Vegetation
		Flächen mit spärlicher Vegetation
		Gletscher und Dauerschneegebiete
Feuchtflächen	Feuchtflächen im Landesinnern	Sümpfe
		Torfmoore
Wasserflächen	Wasserflächen im Landesinnern	Gewässerläufe
		Wasserflächen

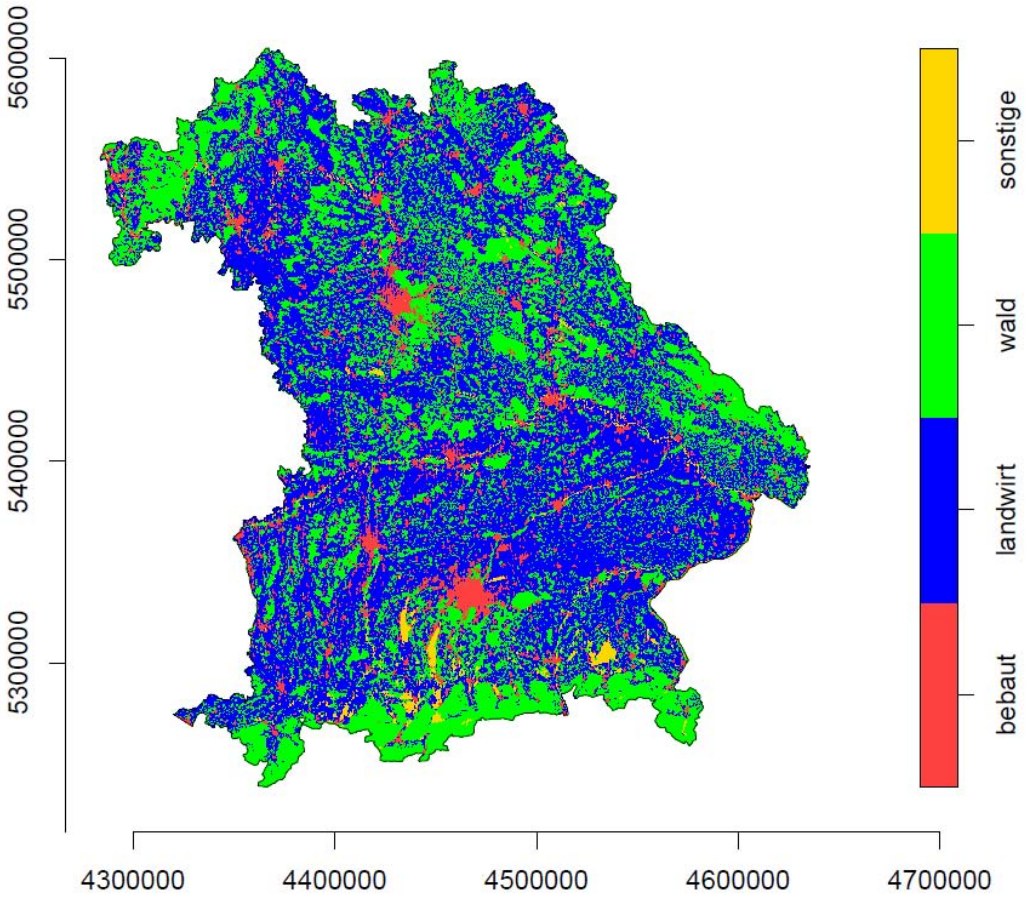

Abbildung 3.3.: CORINE Land Cover-Daten vergröbert in vier Kategorien

3.5. Straßen-Daten

Informationen zu Straßen in Bayern lagen in Form einer Shape-File im Vektor-Daten-Format vor. Das unbearbeitete Datenmaterial enthielt zu jedem der 363138 Bögen den Straßentyp in den Kategorien „Autobahn", „Bundesstraße", „Wohnstraße", „Landesstraße", „Kreisstraße", „Fernstraße" und „nicht klassifiziert".

Zur Analyse musste sich auf die Kategorien „Autobahn" (`motorway`), „Bundesstraße"

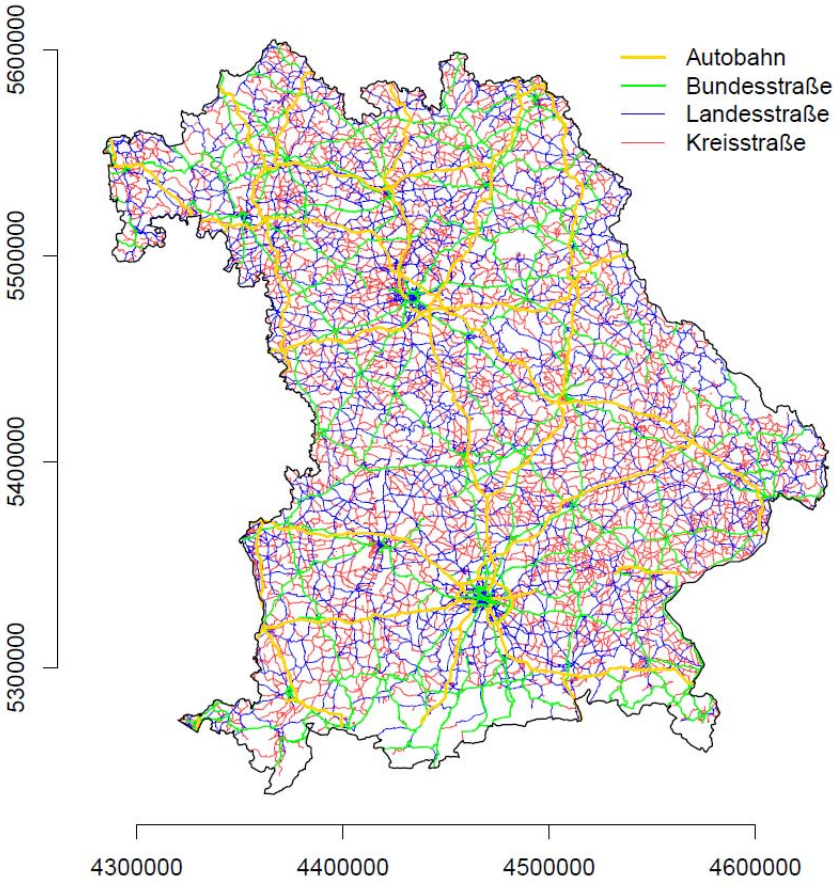

Abbildung 3.4.: Straßennetz

(`primary`), „Landesstraße" (`secondary`) und „Kreisstraße" (`tertiary`) beschränkt werden, da nur zu diesen Informationen zu Wildunfällen vorlagen. Nach zusätzlicher Beschränkung auf die Eckpunkte, die sich tatsächlich innerhalb des die Grenzen Bayerns repräsentierenden Polygons (`bayernfenster`) befanden, verblieben noch 788156 Punkte, die 73909 Bögen repräsentieren. In Abbildung 3.4 ist das Straßennetz nach Bearbeitung getrennt nach Straßentyp dargestellt. Wie zu erwarten, liegen die Straßen, bei nach Typ getrennter Betrachtung umso dichter, je weniger frequenziert dieser ist. Auffallend ist, dass es am Alpenrand kaum Straßen aus den betrachteten Kategorien gibt. Für die Analyse unter Annahme eines räumlichen Poissonprozesses müssen, wie in Abschnitt

2.3 erwähnt, die Werte der Kovariablen überall im Beobachtungsfenster gegeben sein. Anders als für die restlichen Einflußgrößen ist dies für die Kovariable `strasse` zunächst nicht der Fall, da ihre Werte für Lokationen abseits der Straßen nicht definiert sind. Um aber dennoch Werte über den ganzen Bereich hin zu erhalten, könnte man jedem Punkt den Typ der ihm am nächsten gelegenen Straße zuordnen. Zur praktischen Umsetzung hiervon wurde eine Rasterkarte erstellt und jedem Pixelmittelpunkt der Typ der ihm am nächsten gelegenen Straße zugeordnet. Durch die damit verbundene Vergröberung der Information kommt es unweigerlich dazu, dass Straßen den falschen Typen zugeordnet werden, nämlich immer dann, wenn mehrere Straßen unterschiedlichen Typs einen Pixel kreuzen. Um diese Fehlspezifikationen möglichst gering zu halten, wurde mit 250 m eine besonders feine Auflösung gewählt. Abbildung 3.5 zeigt die gesamte Karte und Abbildung 3.6 einen kleineren Ausschnitt mit eingezeichneten Straßen, wobei Fehlspezifikationen rot gekennzeichnet sind. In Bereichen, in denen die Straßen dichter liegen, kommt es häufiger zu Fehlspezifikationen. Allerdings wurden für alle Straßentypen über 90% der Straßenabschnitte richtig spezifiziert.

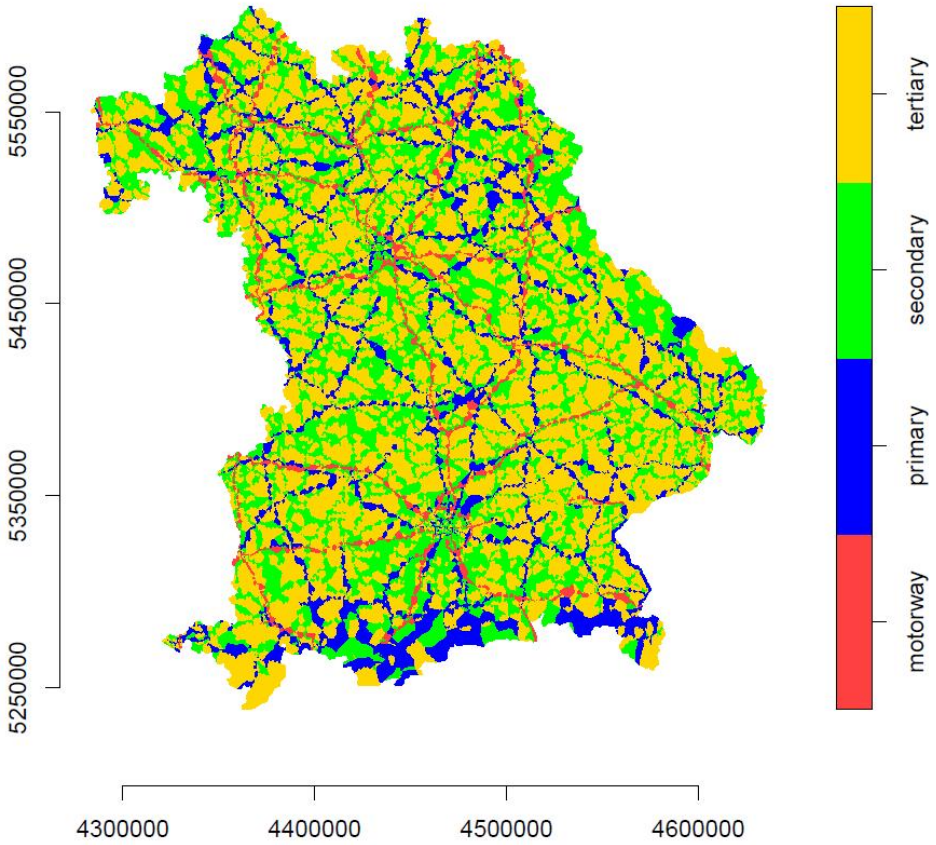

Abbildung 3.5.: Straßenkarte mit Ausprägungen an jeder Lokation

3.6. Kurvigkeit

Aus den Straßendaten wurde ein Maß für die Kurvigkeit der Straßen gewonnen. Zunächst wurden dazu auf einem Raster in jedem Pixel alle Innenwinkel zwischen $[A, B]$ und $[B, C]$ der Dreiecke, welche jeweils durch drei aufeinanderfolgende Punkte A, B und C einer Straße gegeben sind, berechnet, wobei dieselbe Auflösung und Pixellokationen wie bei den Wildverbissdaten gewählt wurde. Da größere Winkel aber weniger kurvigen Straßen

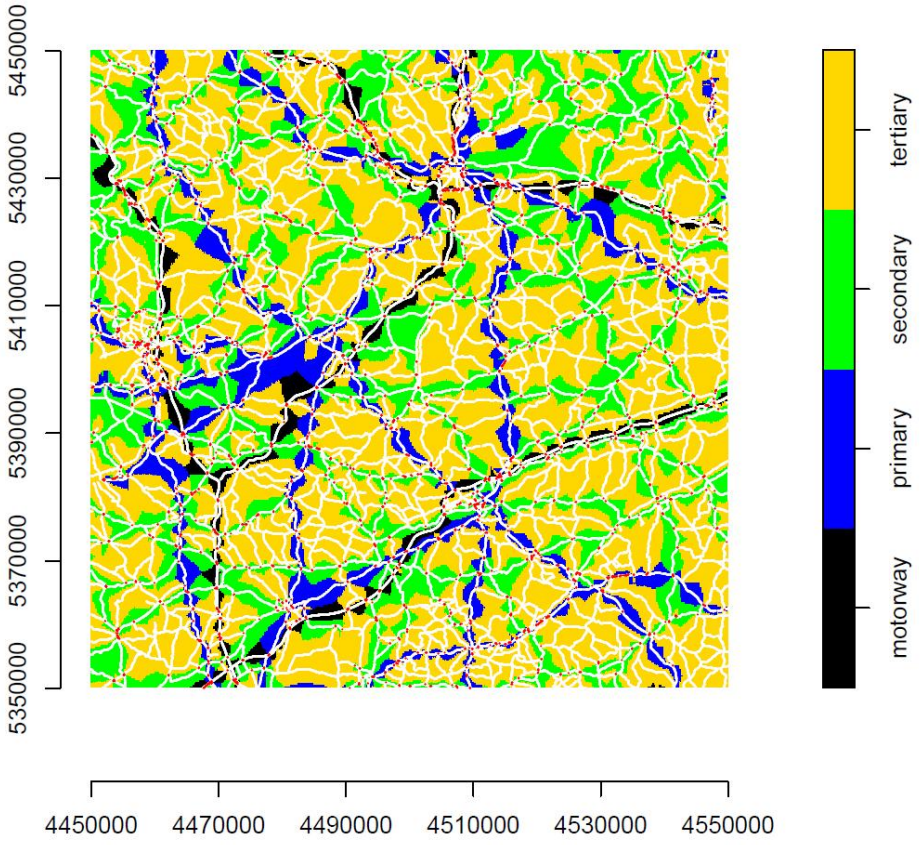

Abbildung 3.6.: Ausschnitt aus der Straßenkarte - Fehlspezifikationen sind rot gekennzeichnet

entsprechen, größere Werte des Maßes aber für kurvigere Straßen sprechen sollen, wurden die Winkel anschließend von 180° abgezogen. Da die Winkel zumeist sehr nahe an 180° lagen, manche Werte aber sehr stark davon abwichen, wurden die Werte logarithmiert, um eine gleichmäßigere Verteilung über den Wertebereich hin zu erhalten. Schließlich wurde von diesen Werten in jedem Pixel das Maximum verwendet, da sich die Werte in den Mittelwerten zu sehr glichen. Um Zufallsschwankungen auszugleichen, wurden die Werte aller Pixel, an denen Werte gegeben waren, mit den nächsten beiden solchen Pixeln gemittelt. Um Werte überall in Bayern zu erhalten, wurde abschließend analog wie bei den Straßendaten, allen Pixeln, durch die keine Straßen gingen, die Werte der jeweils nächsten Pixel, für die ein Wert verfügbar war, zugeordnet.

Makroskopisch lässt sich, wie in Abbildung 3.7 ersichtlich und wie man auch erwarten würde, kein räumlicher Trend der Kurvigkeit ausmachen.

3.7. Straßenlänge

Um eine sachdienliche Interpretation der Häufigkeit der Wildunfälle zu ermöglichen, ist es sehr wichtig, die Straßenlänge in den jeweiligen Gebieten zu berücksichtigen. Modelliert man nur die erwartete Anzahl von Wildunfällen pro Quadratmeter, wie bei den räumlichen Poissonprozessen in ihrer Grundform der Fall, so lässt man außen vor, dass sich bei sonst gleichen Bedingungen die Intensität automatisch erhöht, wenn sich die Dichte der Straßen erhöht, da Wildunfälle nur auf Straßen vorkommen. Z.B. könnte es im Extremfall vorkommen, dass in einem am Waldrand liegenden städtisch geprägten Gebiet, in dem die Verkehrsdichte sehr hoch ist, mehr Unfälle passieren, als in einem einsamen Waldstück durch das nur eine Straße führt, obwohl die Gefahr eines Wildunfalles auf dieser Straße viel höher ist, als auf einer vergleichbaren Straße in der Stadt. Diese Diskrepanz zeigte sich auch sehr deutlich in der Analyse (Kapitel 4).

Um ein Pixelimage der Straßenlängen zu erhalten, wurden die Längen der einzelnen Straßen innerhalb eines Pixels, approximiert durch die Summe der Abstände aufein-

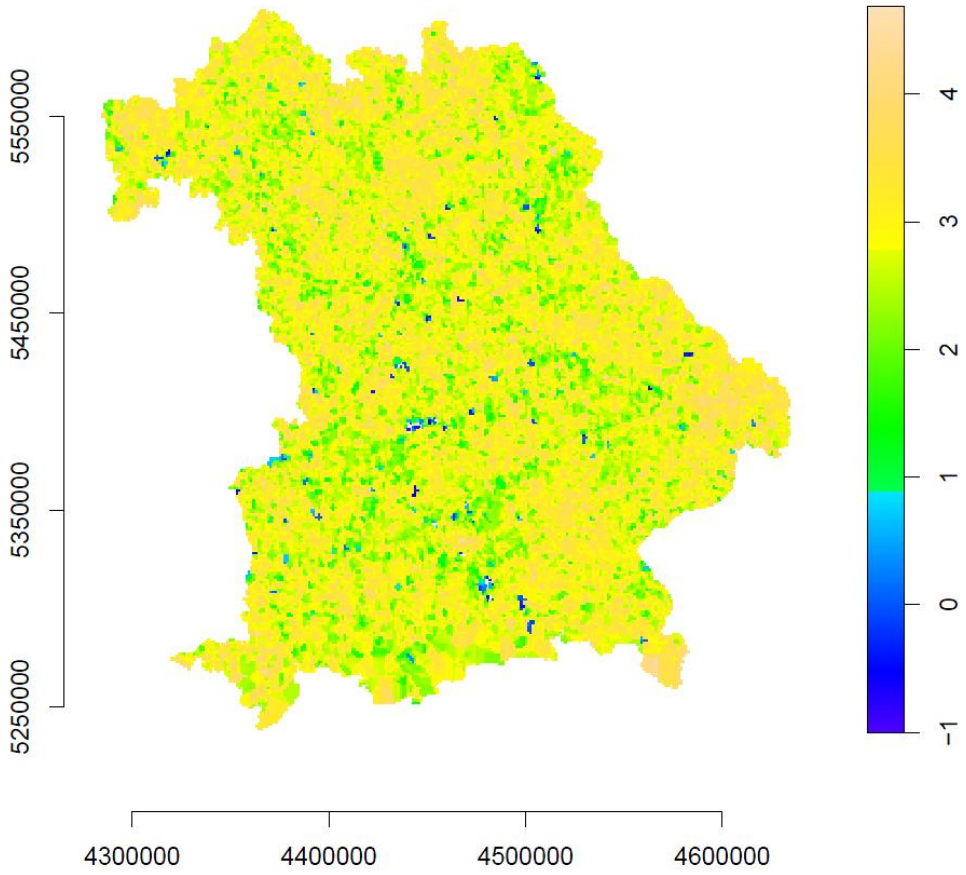

Abbildung 3.7.: Maß der Kurvigkeit - abgeschlagen kleine Werte (< -1) sind der Darstellbarkeit halber ausgeschlossen

anderfolgender Eckpunkte aufaddiert, wieder mit derselben Auflösung und bei gleichen Lokationen wie bei den Wildunfalldaten. Dazu war eine Erhöhung der Dichte der Eckpunkte nötig, da es häufig der Fall war, dass sich keine zwei Punkte einer Straße innerhalb eines Pixels befanden, obwohl die Straße durch diesen Pixel ging, oder, dass der letzte Punkt einer Straße, bevor diese den Pixel verließ, zu weit im Inneren lag, sodass man die Länge der Straße unterschätzt hätte. Wenn zwei Punkte weiter als 30 Meter auseinanderlagen, wurden zusätzliche Punkte auf der Geraden zwischen diesen eingefügt, sodass am Schluss keine zwei Punkte einer Straße weiter als 30 Meter auseinanderlagen. Bei grafischen Kontrollen, zeigte sich allerdings, dass in vielen Fällen durch das beschriebene Vorgehen, nicht die wahre Straßendichte wiedergegeben wurde, insbesondere für Bereiche, in denen wenig Straßen vorlagen. Im Extremfall kreuzte eine Straße einen sonst leeren Pixel nur durch eine Ecke, und verlief quer durch den benachbarten Pixel, sodass man bei Verwendung der beobachteten Straßenlänge völlig unterschiedliche Straßendichten in den benachbarten Pixeln erhalten hätte. Um diese zufälligen Schwankungen auszugleichen, wurde ähnlich wie bei der Kurvigkeit, für jeden Pixel das Mittel bei Einschluß der acht (bzw. am den Rändern weniger) umliegenden Pixel gebildet. Abbildung 3.8 zeigt das Ergebnis.

3.8. Nachtlichter

Die Daten zu den Nachtlichtern stammen von dem National Geophysical Data Center der USA. Sie basieren auf Satellitendaten und sind frei im Internet verfügbar. Die Helligkeit wird in Form von ganzen Zahlen im Bereich 2 bis 63 angegeben, wobei diese Werte in der Auflösung ca. 547 m verfügbar sind. Aufgrund der Vielzahl der Ausprägungen wurde diese Variable in der Analyse als stetig behandelt. Die Umwandlung in das Gauß-Krüger-System erfolgte analog wie bei den CORINE Land Cover-Daten. Abbildung 3.9 zeigt die resultierende Karte - man kann die einzelnen Städte gut als helle Flächen erkennen.

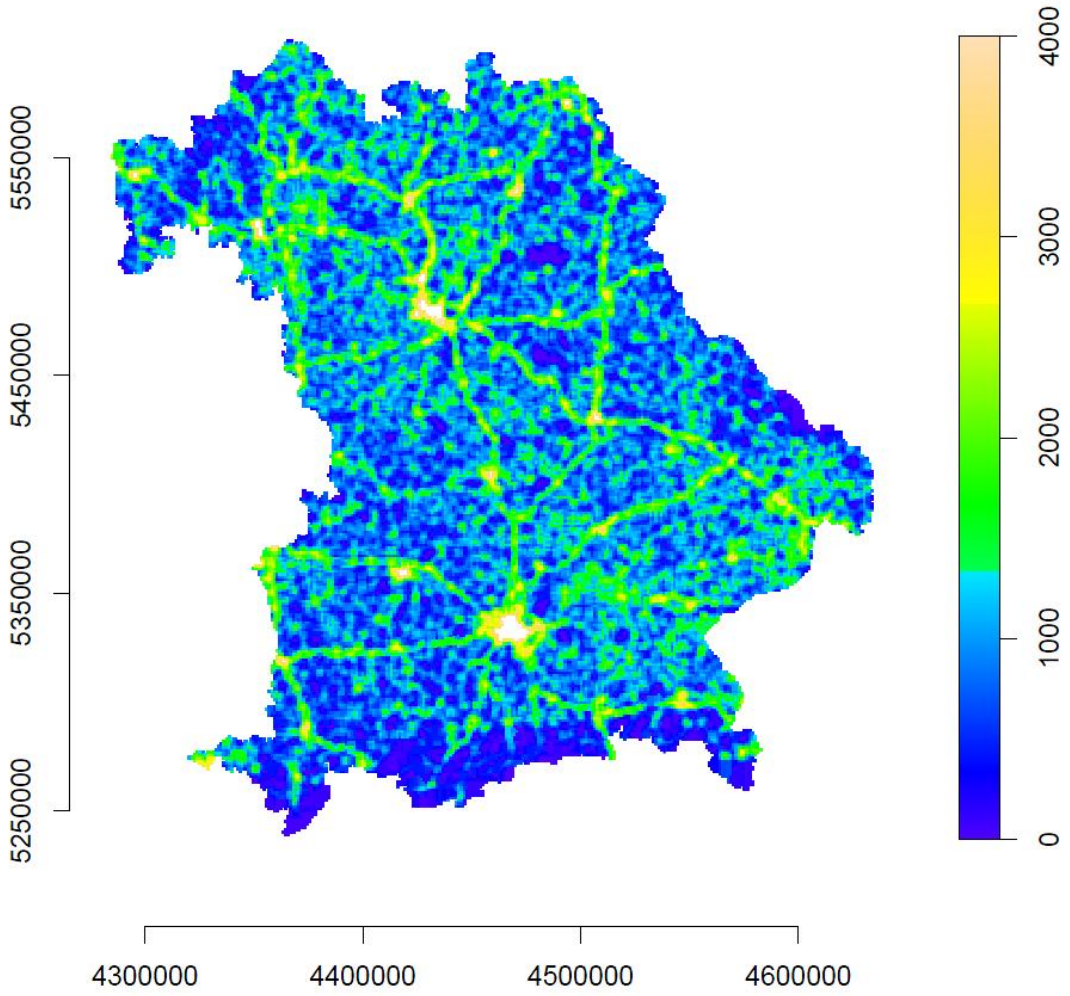

Abbildung 3.8.: approximierte Straßenlängen - abgeschlagen große Werte (> 4000) der Darstellbarkeit halber ausgeschlossen

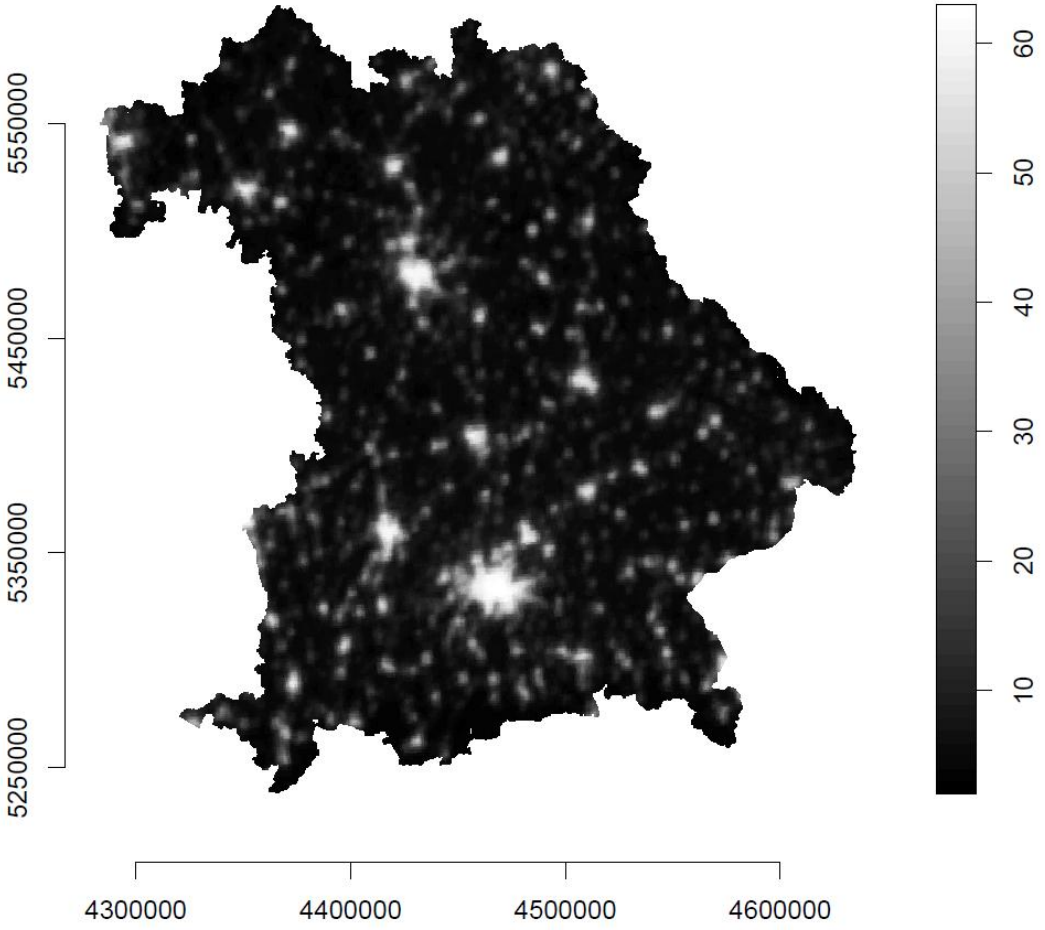

Abbildung 3.9.: Karte der Nachtlichter

4. Analyse

4.1. Modellierung als räumliche Poissonprozesse

Im Folgenden werden zuerst das genauere Vorgehen bei der Analyse und die Ergebnisse unter Annahme räumlicher Poissonprozesse beschrieben, wobei zwei verschiedene Muster an Dummypunkten verwendet wurden: Zum einen wurden die Dummypunkte auf einem regelmäßigem Gitter über Bayern hinweg gewählt, zum anderen nur auf den Straßen. Die geschätzten Einflüsse bestimmter Kovariablen unterschieden sich stark zwischen den beiden Schemata, woraus man den Rückschluss ziehen kann, dass zur Modellierung der Wildunfalldaten berücksichtigt werden muss, dass sich die Unfälle nur auf den Straßen ereignen können. Alle Analysen wurden mit der Statistik-Software-Umgebung R durchgeführt, wobei sich der verwendete Programmiercode auf der beiligenden CD-Rom befindet.

Die Analyse der Wildunfälle auf Gemeindeebene von Hothorn et al. (in Vorb.) lässt vermuten, dass sich das Risiko von Wildunfällen zwischen den Jahren 2006 und 2009 für alle Gemeinden um den gleichen Faktor erhöhte und sich die Erhöhung damit nicht von Lokation zu Lokation unterscheidet. Unter dieser Annahme ändert sich der Einfluss der Kovariablen in den beiden Jahren nicht, sodass bei nach Jahr getrennter Analyse, nur unterschiedliche Intercepts zu erwarten wären. Da die Vorhersage absoluter Anzahlen für Maßnahmen zur Wildunfallverminderung keine Bedeutung hat, sondern relative Werte genügen und da der Fokus ohnehin auf den Formen der Einflüsse der Kovariablen lag, wurden beide Jahre gemeinsam analysiert. Die Absolutwerte der Prädiktionen sind zu

interpretieren als die erwarteten Anzahlen von Wildunfällen pro Quadratmeter bzw. später Meter Straße im Laufe der Jahre 2006 und 2009.

Abbildung 4.1 zeigt eine nonparametrische Schätzung der Intensität der Wildunfälle nach der Formel $\hat{\lambda}(x) = e(x) \sum_{i=1}^{n} K(x - x_i)$. Dabei ist $K(\cdot)$ ein zweidimensionaler Gauß-Kern mit Bandbreite σ und $e(x) = \int_W K(x - u)du$ ein Korrekturgewicht zum Entgegenwirken von Randeffekten - die Faltung des Gauß-Kerns mit dem Beobachtungsfenster W (Baddeley und Turner, 2005a). Diese Schätzung erfolgt unter der Annahme, dass die zu schätzende Intensitätsfunktion die Realisation eines sogenannten Cox-Prozesses ist (Diggle, 1985). Letzterer ist ein Prozess, in dem die Intensitätsfunktion $\lambda(x)$ eine Zufallsvariable ist, gegeben eine Realisation $\lambda(x, w)$ dieser, er einem inhomogenen Poissonprozess mit eben dieser Intensitätsfunktion folgt (Schmid, 2010). Integriert man über die Schätzung auf dem Beobachtungsfenster ergibt sich die Anzahl der Beobachtungen n. Es wird also die unter einem Poissonprozess zu erwartetende Anzahl an Punkten in W gleich der tatsächlich beobachteten Anzahl gesetzt. Bei Nicht-Berücksichtigung des Faktors zur Randkorrektur, der dafür aufkommt, dass der Prozess nur in W, nicht auf ganz $\mathbb{R}^2$ beobachtet werden kann, ist die Schätzung zudem proportional zur klassischen nonparametrischen Dichteschätzung, was der Eigenschaft von inhomogenen Poissonprozessen entspricht, dass seine Dichte, gegeben die Anzahl n der Punkte proportional zur Intensitätsfunktion ist. Durch diese beiden Eigenschaften ist ein Poissonprozess bereits vollständig charakterisiert.

Die Grafik lässt erkennen, dass die Intensität der Wildunfälle in Niederbayern und in großen Teilen Oberbayerns am höchsten zu sein scheint. Auch in Mittelfranken passieren offenbar viele Wildunfälle. Auffällig ist die Anhäufung im Norden Unterfrankens. Diese Ergebnisse sind allerdings vorsichtig zu interpretieren, da die Straßendichte nicht einbezogen wurde. Dies gilt in ganz besonderem Maß für die tiefblauen Flächen am Alpenrand, da sich hier sehr wenige Straßen, insbesondere des betrachteten Typs befinden, siehe Abb. 3.4.

Die Wildunfallzahl in Relation zur Straßenlänge zu setzen ist auch vor dem Hintergrund

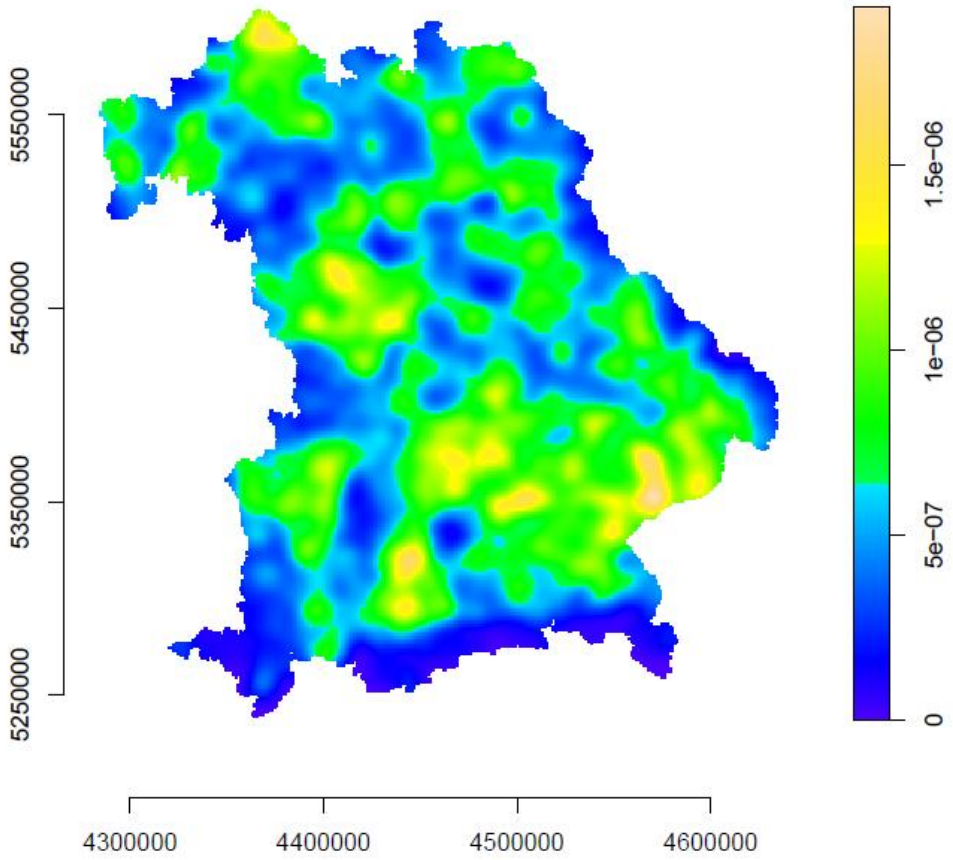

Abbildung 4.1.: nonparametrische Schätzung der Wildunfallintensität bei Verwendung eines Gaußkerns mit Bandbreite $\sigma = 5000$

wichtig, dass sich in der Datengrundlage nur der Teil der Unfälle befindet, der sich auf Autobahnen, Bundesstraßen, Landesstraßen und Kreisstraßen ereignet hat, man also bei der Modellierung der absoluten Häufigkeiten an Unfällen die Intensität automatisch unterschätzt.

Abbildung 4.2 zeigt links die geschätzte K-Funktion und rechts die geschätzte L-Funktion. Beide lassen grundsätzlich dieselbe Schlußfolgerung zu: Da sich die Kurven jeweils oberhalb derer, die man bei einem homogenen Poissonprozess erwarten würde, befinden,

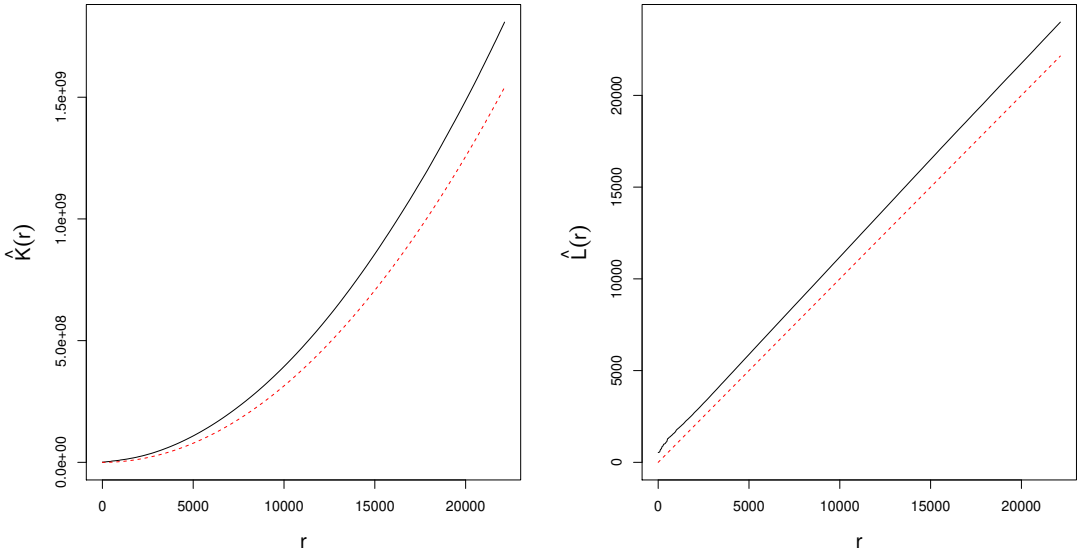

Abbildung 4.2.: geschätzte K- und L-Funktion für die Rehunfalllokationen - gestrichte rote Linien: Werte für homogenen Poissonprozess

ziehen sich die Punkte tendenziell gegenseitig eher an. Wildunfälle treten also, wie man auch erwarten würde eher gehäuft auf, als dass sie gleichmäßig über den Raum verteilt wären. Die Cluster-Bildung scheint dabei weniger auf punktueller Ebene, sondern eher auf großflächigeren Gebieten zu erfolgen, da sich die Funktionen für größere Distanzen weiter von denen unter einem homogenen Poissonprozess wegbewegen. Zur Randkorrektur wurde der reduced-sample-Schätzer verwendet, da dieser für große Datensätze schnell berechnet werden kann und man den Informationsverlust durch Nicht-Einbeziehen von Punkten (siehe Abschnitt 2.6.1) in Kauf nehmen kann.

4.1.1. Dummypunkte auf regulärem Gitter

Wie in Abschnitt 2.3 gesehen, sind die Dummypunkte zusammen mit den Datenpunkten die Stützstellen bei der Quadratur der Intensitätsfunktion. Um das Integral über Bayern hin möglichst genau zu approximieren, wäre es deshalb angebracht, die Dummy-

punkte auf einem Gitter äquidistant über das Beobachtungsfenster hin zu wählen und so den betrachteten Bereich möglichst gut abdecken zu können, da man a priori noch kein Wissen über die Form der Funktion hat. Dabei vernachlässigt man allerdings, dass sich Unfälle nur auf Straßen ereignen können und die beobachtetete Intensität deshalb von der Dichte des Straßennetzes an den jeweiligen Lokationen abhängt, wie bereits in Abschnitt 3.7 mit einem Beispiel erklärt wurde. Würden alle Unfälle auf allen Straßen zur Verfügung stehen, könnte man die Werte der bei Dummypunkten auf einem regulären Gitter resultierende Intensitätsfunktion zumindest als absolute Anzahlen an Unfällen pro Quadratmeter ungeachtet der Straßenlänge interpretieren, auch wenn dies allerdings inhaltlich weniger dienlich ist. Dies ist jedoch auch nicht möglich, da nur ein Teil der Unfälle zur Verfügung steht. Also kann man die Ergebnisse aus diesem Abschnitt lediglich interpretieren als Schätzung für die erwartete Anzahl an Wildunfällen auf den betrachteten Straßentypen pro Quadratmeter ungeachtet der Straßenlänge, welche selbstverständlich immer kleiner ist als die erwartete Anzahl auf allen Straßentypen. Möglicherweise wäre die Herangehensweise in diesem Abschnitt adäquater, wenn alle Unfälle auf allen Straßentypen zu Verfügung stünden, weil sich zum einen dann die Straßendichte von Lokation zu Lokation eventuell nur gering unterscheidet, zum anderen, weil das Straßennetz so dicht wäre, dass man die Eigenschaft räumlicher Poissonprozesse, dass sich die Punkte überall im Beobachtungsfenster befinden, als gegeben betrachten könnte.

Da mit diesem Modell die Häufigkeit von Unfällen an jedem Punkt in Bayern unabhängig davon, ob es dort Straßen gibt oder nicht geschätzt werden sollte und man in Gebieten ohne Straßen die Kurvigkeit und Länge an Straßen nicht berechnen kann, wurden diese Kovariablen hier ausgeschlossen. Vor dem Hintergrund, dass die Unfälle aber nur auf Straßen passieren und `strasse` eine wichtige Kovariable ist, bei Nichtbeachtung deren Einflusses möglicherweise die Interpretierbarkeit anderer Kovariablen gefährdet ist, sollte diese Einflussgröße dennoch Eingang in das Modell finden. Dazu wurde jedem Punkt, bzw. jedem Pixelmittelpunkt des die Kovariable repräsentierendem Pixelimages wie in Abschnitt 3.5 beschrieben der Straßentyp der nächstgelegenen Straße zugeordnet. Die

Einflussgrößen in diesem Modell sind also: Straßentyp (`strasse`), geglätteter Verbiss (`verbiss`), Landnutzung (`landnutzung`), Nachtlichter (`helligkeit`) und die Koordinaten (im Gauß-Krüger-System) (`xcoor`, `ycoor`). Um möglichst flexibel in der Modellierung zu sein und da man aufgrund der großen Fallzahl viele zu schätzende Parameter zulassen kann, wurden die (quasi-)stetigen Einflußgrößen `verbiss` und `helligkeit` als penalisierte B-Splines und (`xcoor`,`ycoor`) als penalisierte Tensorprodukt-B-Spline aufgenommen. Die Schätzung der räumlichen Poissonprozesse (oder allgemein auch räumlicher Punktprozesse) geschieht in `spatstat` (Baddeley und Turner, 2005a) mit der Funktion `ppm()`. Diese greift zur Schätzung der semiparametrisch modellierten Einflüsse auf die `gam()`-Funktion aus dem R-Paket `mgcv` (Wood, 2006a) zurück.

In der Poissonregression, auf die sich, wie gesehen, die Schätzung eines räumlichen Poissonprozesses mit Kovariablen zurückführen lässt, ist die Zielgröße nicht ganzahlig. Da sich aber in R bei Verwendung des `family`-Objektes „`poisson`" nur ganzahlige Poissonregression durchführen lässt, verwendet `spatstat` die `family` „`quasi`" mit log-Linfunktion und Varianz `mu`. Auf diese Art würde man auch eine Quasi-Poissonregression durchführen. Während bei der standardmäßigen Poissonregression angenommen wird, dass die Zielvariable gegeben die Einflußgrößen einer Poissonverteilung folgt und damit Erwartungswert und Varianz gleich sind, wird bei Quasi-Poissonregression vorausgesetzt, dass die Varianz gleich dem Erwartungswert multipliziert mit einer Konstanten, dem Skalenparameter ϕ ist, wobei letzterer aus den Daten geschätzt wird. Aus beiden Ansätzen resultieren dieselben Parameterschätzer, allerdings unterscheiden sich die Varianzen in der Quasi-Poissonregression von deren bei gewöhnlicher Poissonregression um den Faktor ϕ. Da die approximierte Likelihood des räumlichen Poissonprozesses in der Form (2.10) aber der Likelihood bei Poissonregression entspricht, müssen die geschätzten Varianzen bei Verwendung der `quasi`-Familie durch den geschätzten Skalenparameter geteilt werden, um die korrekten Varianzen zu erhalten. `spatstat` berechnet mit der Funktion `vcov.ppm` zur Ableitung der Varianz die Fisher-Informationsmatrix, die man bei Poissonregression erhalten würde, neu. Die `print`-Methode für `ppm`-Objekte greift auf diese Funktion zurück, da Standardfehler und Konfidenzintervalle mit ausgegeben

werden. Bei Verwendung von Spline-Funktionen im Prädiktor erhält man allerdings eine Fehlermeldung von der Funktion `model.frame.default` bei dem Versuch das Modell auszugeben, da die Rückgabe der in der Modell-Formel zur Spezifikation der Spline-Basen in `mgcv` verwendeten Funktionen „`s`" bzw. „`te`" nicht direkt zur Aufsetzung der Modellmatrix verwendet werden kann. Deshalb wurde in dieser Analyse mit Hilfe der `spatstat`-Funktion `getglmfit` das `gam`-Objekt, das den Fit der Quasi-Poissonregression enthält aus dem `ppm`-Objekt extrahiert und die Varianzen durch den geschätzten Skalenparameter geteilt, um die geschätzten Varianzen zu erhalten, die auch bei gewöhnlicher Poissonregression resultieren würden. Die entsprechende Funktion lautet:

```
getgam <- function(modell) {
  require(spatstat)
  gammodell <- getglmfit(modell)
  gammodell$Vp <- gammodell$Vp/gammodell$sig2
  gammodell
}
```

Wie in Abschnitt 2.3 gesehen, wird das in der Likelihood eines räumlichen Poissonprozesses enthaltene Integral bei der Schätzung numerisch integriert, wobei sich die approximierte der wahren Likelihood mit wachsender Anzahl an Stützpunkten oder Kacheln annähert. Die Berechnung der Quadraturgewichte geschieht bei Verwendung der Standardeinstellung `method = "grid"` auf folgende Weise: Das das Beobachtungsfenster einschließende kleinste Reckteck wird in ein $K \times K$ Raster von Kacheln unterteilt und dieses Schema mit den Beobachtungsfenster geschnitten. Anschließend werden die Flächen a_i $(i = 1, \ldots, K^2)$ aller Kacheln berechnet und schließlich jeder Beobachtung das Gewicht $w_j = a_i/n_i$ zugewiesen, wobei i der Index des Pixels ist, indem sich der Quadraturpunkt j befindet und n_i die Anzahl der Quadraturpunkte in a_i. Die Tatsache, dass Kacheln auch keine Quadraturpunkte enthalten können und diese Bereiche damit keinen Einfluss auf die Schätzung haben, wird sich in Abschnitt 4.1.2 durch entsprechende Wahl der Stützstellen und eine feine Kachelauflösung zu Nutze gemacht, um die Schätzung auf

Bereiche um die Straßen zu konzentrieren. K wird in der Standardeinstellung berechnet als die Hälfte der Wurzel der Quadraturpunkte abgerundet auf die nächstkleinere durch fünf teilbare Zahl. Die Flächen der Kacheln werden dabei durch die Summe der Flächen der Pixel eines das Beobachtungsfenster approximierenden Pixelimages berechnet, deren Mittelpunkte in die entsprechende Kachel fallen. Per default wird dafür eine 100×100-Pixelimage verwendet, was für ein feineres Gitter an Kacheln zu grob ist. Es wurde die Einteilung 10000×10000 verwendet, sodass für $K = 500$ im Schnitt 400 Pixel zur Berechnung der Fläche einer Kachel im Inneren verwendet werden.

Offenbar ist die Genauigkeit dieses Quadraturverfahrens abhängig von der Auflösung der Kacheln und nicht von der Anzahl an Dummypunkten, da die Schätzung auf den Kacheln konstant und die Flexibilität daher mit kleineren Kachelauflösungen wächst. Um sich zu vergewissern, dass eine genügend feine Auflösung zur Approximation der Likelihood gewählt wird, wurde das Modell mehrmals für verschieden feine Auflösungen gefittet. In allen Fällen wurde dabei ein 600×600-Gitter an Dummypunkten verwendet. In den Abbildungen 4.3 und 4.4 sind die sich ergebenden Parameterschätzer dargestellt. Die Bestimmung einer geeigneten Anzahl an Dummy-Punkten und Kacheln anhand von Likelihood-basierten Informationskriterien ist offenkundig nicht möglich.

Die Koeffizientenschätzer der kategorialen Variablen ändern sich offenbar über die verschiedenen Auflösungen kaum. Für die Schätzer der Koeffizienten der Spline-Basen ist folgendes zu beobachten: Während mit Ausnahme von `verbisseiche` die Ergebnisse für die zwei gröberen Auflösungen sehr ähnlich sind, ändern sich die Koeffizienten für die beiden feineren Gittern teilweise sehr stark, besonders deutlich zu sehen für `verbissbuche`. Es erscheint nicht plausibel, dass die abweichenden Ergebnisse dadurch erklärt werden können, dass die Likelihood noch nicht genügend genau approximiert wurde, da sich die Schätzer für die beiden niedrigen Auflösungen so stark ähneln. Als Entscheidungshilfe zwischen den beiden Auflösungen $K = 300$ und $K = 500$ wurden Lurking Variable Plots (siehe Seite 67) zu `verbiss`, `helligkeit`, sowie `xcoor` und `ycoor` erstellt. Es zeigte sich, dass die modellierten Einflüsse von `verbissbuche` und `helligkeit` in der höchsten

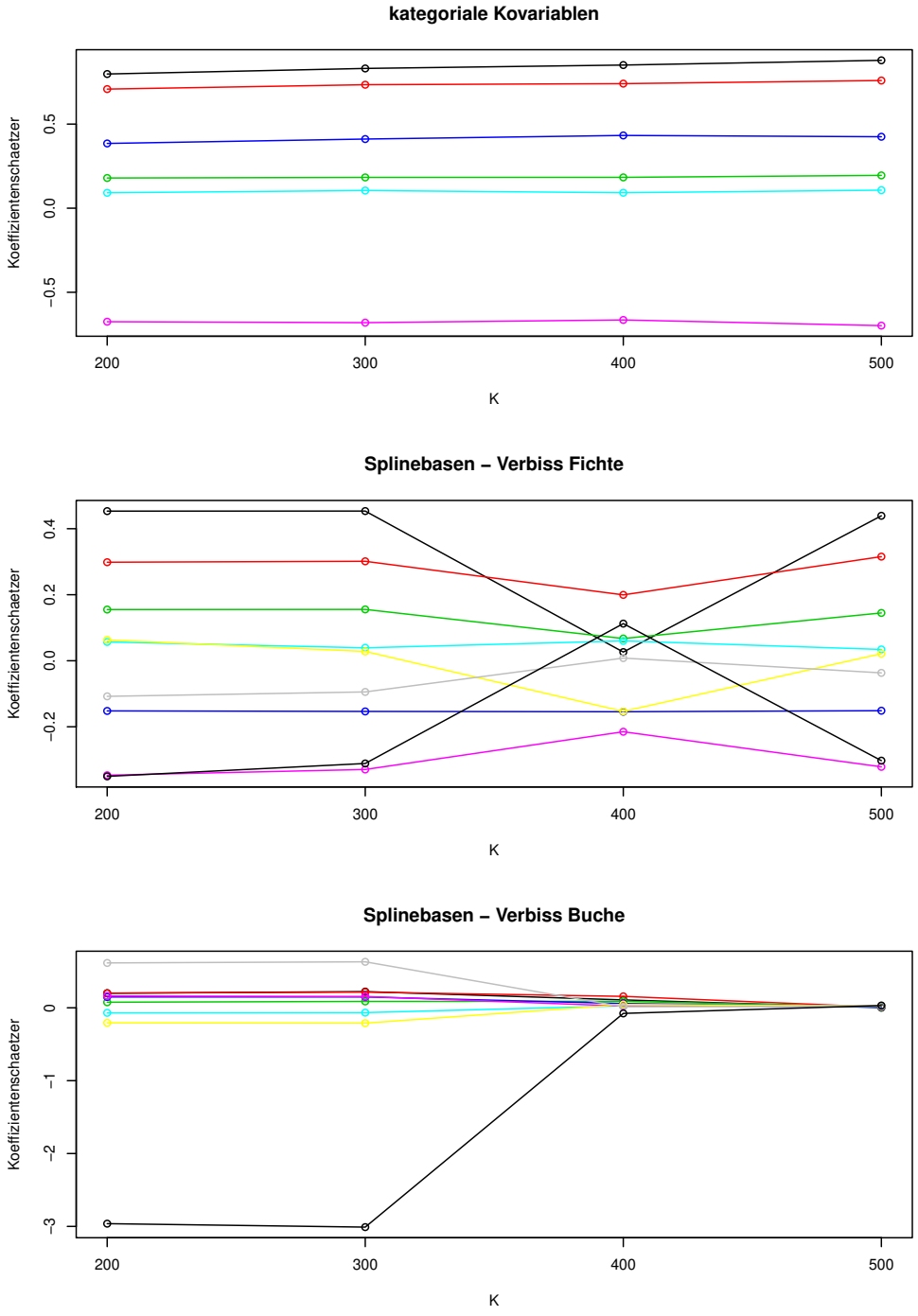

Abbildung 4.3.: Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 1

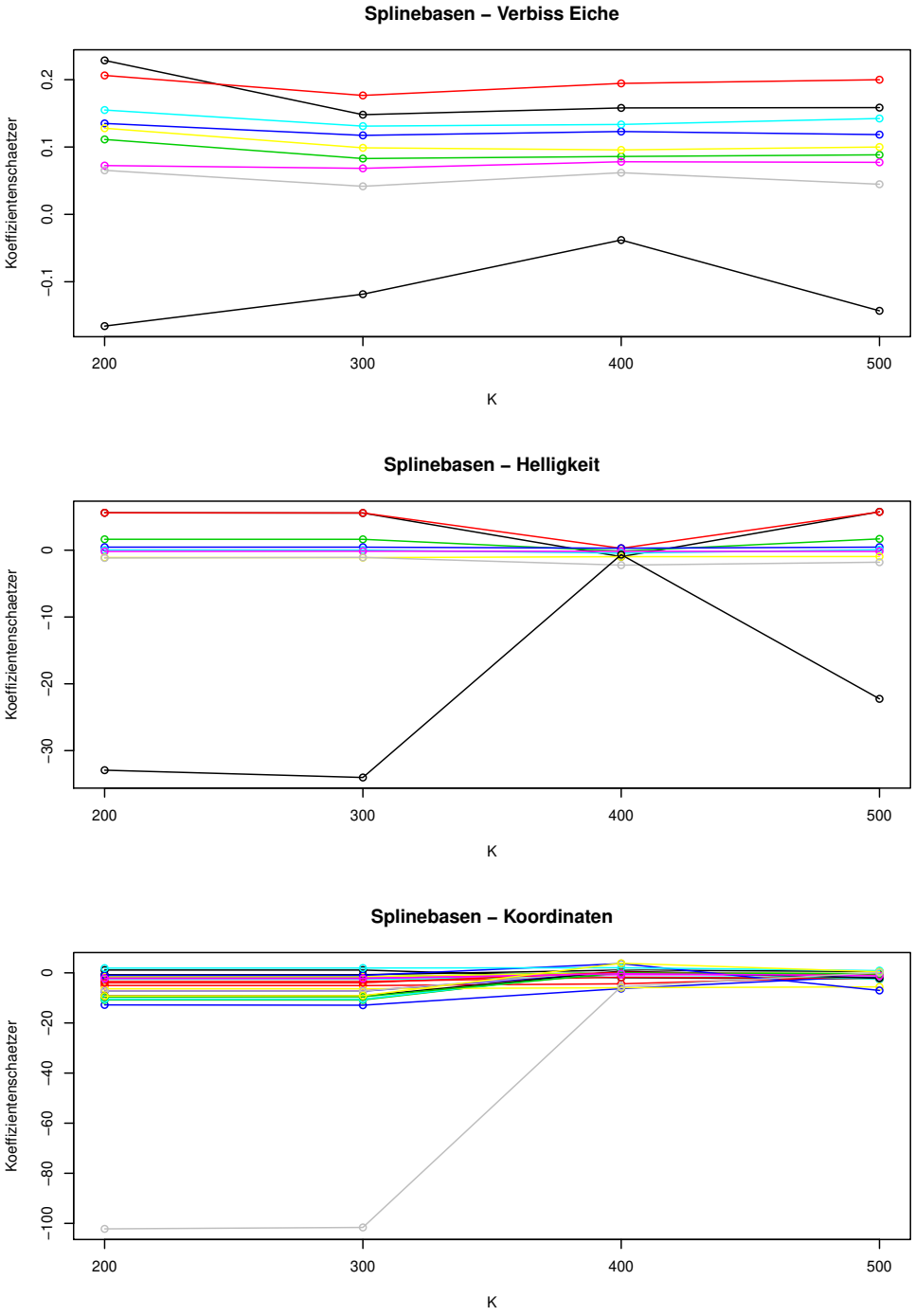

Abbildung 4.4.: Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 2

Auflösung stark von den tatsächlichen abweichen. Aus diesem Grund wurde die zweitgröbste Auflösung $K = 300$ gewählt.

Der Aufruf der Funktion `ppm` zur Schätzung des Modells lautet wie folgt:

```
> modelldummygrid <- ppm(Q, ~ strasse + landnutzung + s(verbissfichte,
    bs="ps") + s(verbissbuche, bs="ps") + s(verbisseiche, bs="ps") +
    s(helligkeit, bs="ps") + te(xcoor, ycoor, bs="ps"), covariates=
    covari, use.gam=TRUE)
```

Hierbei ist `Q` ein Objekt der Klasse `quad`, in dem alle nötigen Informationen über das Quadraturschema enthalten sind, also im Wesentlichen die Datenpunkte, die Dummypunkte und die Gewichte für die numerische Integration. Die Funktionen `s()` bzw. `te()` spezifizieren die Spline-Basen für die univariaten Splines bzw. die Tensorprodukt-Splines, wobei mit der Option `bs="ps"` festgelegt ist, dass penalisierte (Tensorprodukt-)B-Splines verwendet werden. `covari` ist ein Datensatz, der die Werte aller Kovariablen an den Dummy- und Datenpunkten enthält. `use.gam=TRUE` besagt schließlich, dass zur Schätzung der Splinefunktionen auf das R-Paket `mgcv` zurückgegriffen wird.

Der Output der `summary`-Funktion lautet nun:

```
> summary(getgam(modelldummygrid))

Family: quasi
Link function: log

Formula:
.mpl.Y ~ strasse + landnutzung + s(verbissfichte, bs = "ps") +
    s(verbissbuche, bs = "ps") + s(verbisseiche, bs = "ps") +
    s(helligkeit, bs = "ps") + te(xcoor, ycoor, bs = "ps")
<environment: 0x5972a60>
```

```
Parametric coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         -15.10765    0.03792 -398.404  < 2e-16 ***
strasseprimary        0.83089    0.03134   26.512  < 2e-16 ***
strassesecondary      0.73411    0.03031   24.221  < 2e-16 ***
strassetertiary       0.18294    0.03057    5.985 2.17e-09 ***
landnutzunglandwirt   0.41112    0.02382   17.257  < 2e-16 ***
landnutzungwald       0.10508    0.02514    4.181 2.91e-05 ***
landnutzungsonstige  -0.68051    0.07197   -9.456  < 2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Approximate significance of smooth terms:
                    edf Ref.df       F  p-value
s(verbissfichte)  7.714  8.214  12.445  < 2e-16 ***
s(verbissbuche)   8.876  8.990   5.503 1.34e-07 ***
s(verbisseiche)   6.387  7.177   9.725 2.03e-12 ***
s(helligkeit)     8.908  8.992 186.516  < 2e-16 ***
te(xcoor,ycoor)  23.981 24.000 154.860  < 2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

R-sq.(adj) =  -0.161   Deviance explained = 6.91%
GCV score = 0.75462  Scale est. = 0.75442   n = 242833
```

`.mpl.Y` ist hier also die Zielgröße in der Poissonregression. Die kategorialen Einflussgrößen `strasse` und `landnutzung` sind dummy-kodiert mit Referenzkategorien `motorway` bzw. `bebaut`. Bei Dummy-Codierung erhalten bis auf eine Referenzkategorie alle Kate-

gorien eine sogenannte Dummy-Variable. Diese nimmt den Wert 1 an, wenn die jeweilige Beobachtung in die entsprechende Kategorie fällt und 0 wenn nicht. Tritt die Referenzkategorie ein, so haben alle Dummy-Variablen den Wert 0. So erhält, ausgenommen die Referenzkategorie, jede Kategorie einen Koeffizientenschätzer, der bei Eintreten dieser Kategorie dem Unterschied zu dem Wert des Prädiktors bei Eintreten der Referenzkategorie entspricht, vorausgesetzt alle anderen berücksichtigten Variablen haben in beiden Fällen die gleichen Werte. Die Intensitätsfunktion hängt über die Exponentialfunktion vom Prädiktor ab. Das bedeutet, dass die *exponierten* Koeffizientenschätzer als multiplikative Änderung des Wertes der Intensitätsfunktion bei Eintreten einer Kategorie gegenüber dem Wert bei Eintreten der Referenzkategorie zu interpretieren sind. Z.B. wären nach diesem Modell der Interpretation der Intensitätsfunktion gemäß, im Wald pro Quadratmeter oder auch Quadratkilometer $\exp(0,10508) \approx 1,11$ mal so viele Wildunfälle wie in bebauten Gebieten zu erwarten. Die Angabe `n = 242833` entspricht hier der Anzahl der Datenpunkte plus der Anzahl der Dummypunkte, also aller „Beobachtungen", die in die Poissonregression eingingen. Die Schätzung beruht dabei auf 45701 Datenpunkten, da 94 Unfälle ausgeschlossen werden mussten, da sie sich außerhalb der die Kovariablen repräsentierenden, die Fläche Bayerns approximierenden Pixelimages befanden. Offenbar sind die geschätzten Einflüsse aller Kovariablen zum Signifikanzniveau 5% signifikant von Null verschieden. Die erklärte Devianz hat mit 6,91% einen sehr niedrigen Wert.

Abbildung 4.5 zeigt die exponierten Koeffizientenschätzer für `strasse` und `landnutzung` zusammen mit 95%-Konfidenzintervallen. Offenbar gibt es in Gebieten, in denen Autobahnen vorherrschend sind am wenigsten Unfälle. Etwas mehr geschehen auf Kreisstraßen. Am meisten passieren in der Umgebung von Bundesstraßen mit ca. 2,3 mal so vielen wie auf Autobahnen, knapp gefolgt von Landesstraße mit ca. 2,1 mal mehr Unfällen, wobei sich die Konfidenzintervalle hier überlappen.

Auf landwirtschaftlichen Flächen passieren offenbar am meisten Unfälle, gefolgt von bewaldeten und bebauten Flächen, wohingegen auf Feucht- und Wasserflächen wie zu

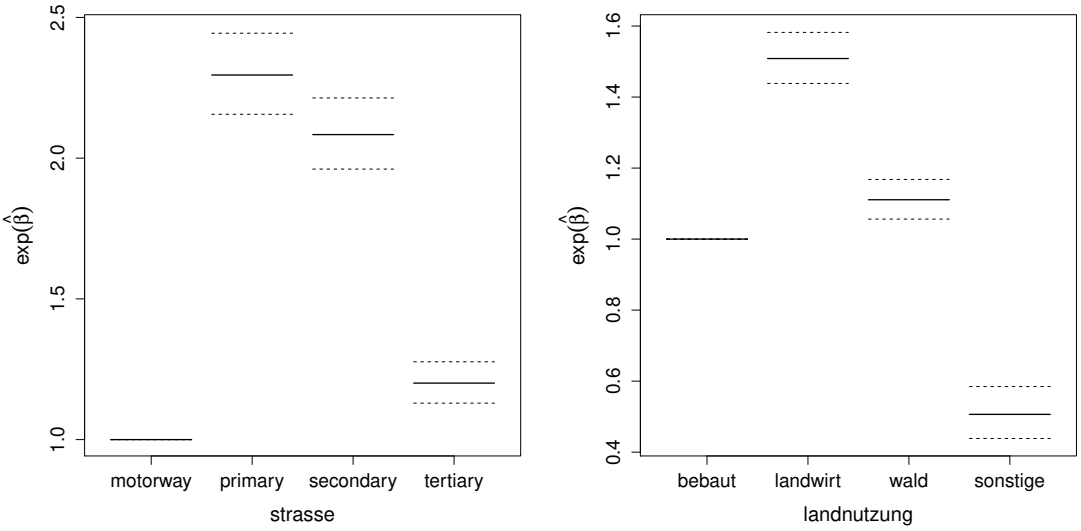

Abbildung 4.5.: exponierte Koeffizientenschätzer der kategorialen Kovariablen im Modell mit Dummypunkten auf regulärem Gitter

erwarten am wenigsten auftreten.

Die Einflüsse der exponierten Splinefunktionen sind in Abbildung 4.6 zu sehen. Dabei sind die gestrichelten Linien mittransformierte 95%-Konfidenzbänder, d.h. die gezeigten Kurven befinden sich in 95% der Fälle vollständig innerhalb dieser Grenzen. Die Splinefunktionen sind aus Identifikationsgründen um Null zentriert, d.h. dass ihre Werte gemittelt über alle Beobachtungspunkte Null ergeben. Bei sehr wenig Verbiss an Fichten sind etwas weniger Unfälle zu erwarten, im niedrigen Bereich geschehen etwas mehr. Bei einem großen Anteil verbissener Fichten sinkt die Gefahr von Wildunfällen, jedoch unterscheidet sich die geschätzte Splinefunktion für einen sehr großen Verbissanteil nicht mehr signifikant von 1. Für Rotbuche sind bei keinem bis sehr wenig Verbiss auch etwas weniger Unfälle zu erwarten. In der unteren Hälfte unterscheidet sich die Kurve nur bei etwa 0.4 signifikant von 1, in Richtung mehr zu erwartender Unfälle. Offenbar ist die Gefahr bei sehr hohen Verbissanteilen an Rotbuchen am größten. Für einen Anteil von 1 scheint das Risiko allerdings wieder zu sinken. Der Einfluss des Verbisses an Eichen

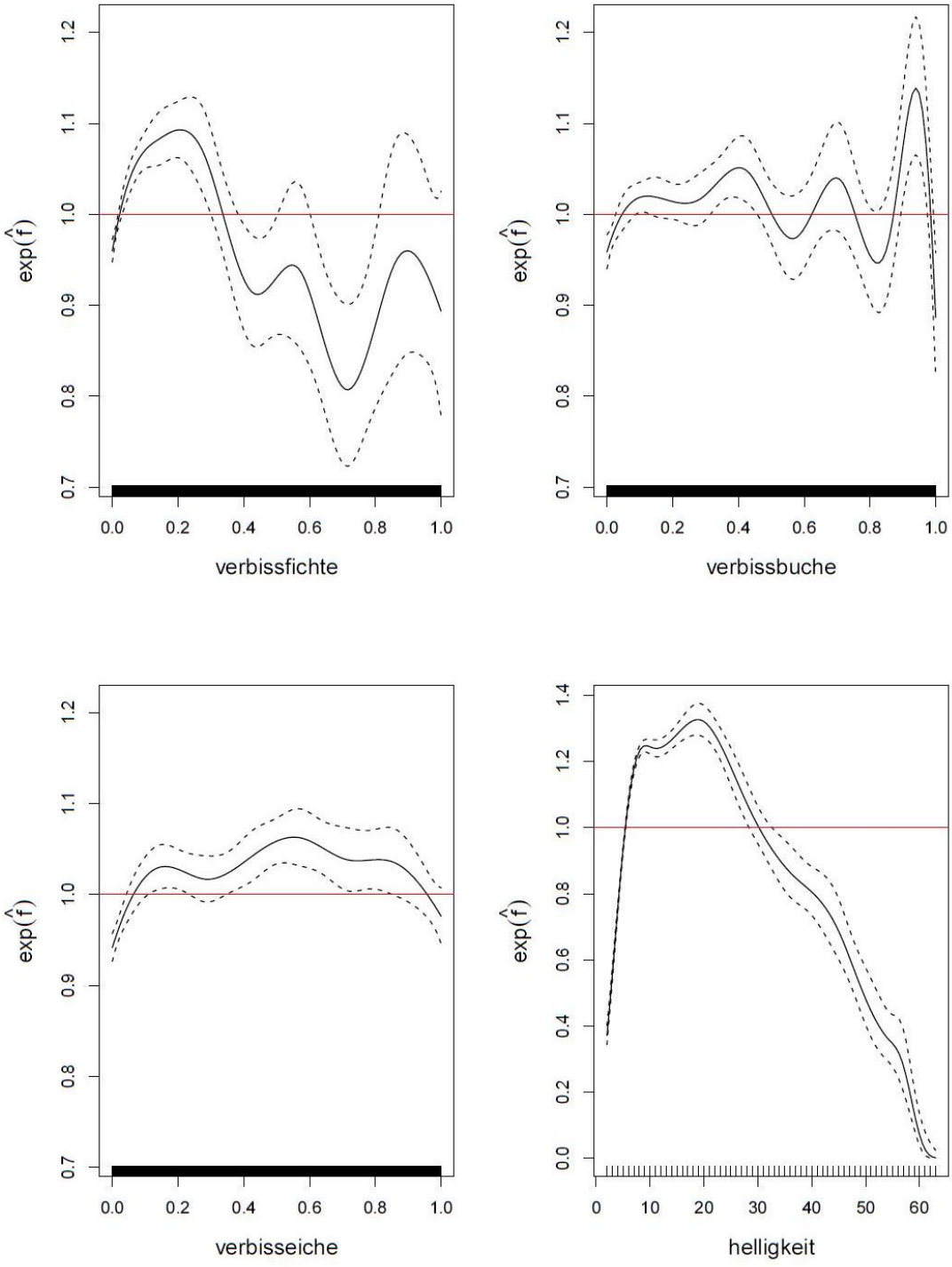

Abbildung 4.6.: exponierte Spline-Funktionen im Modell mit Dummypunkten auf regulärem Gitter

liefert bei generell geringere Stärke ein anderes Bild. Für sehr schwachen und minimal für sehr starken Verbiss ist die Gefahr kleiner, wobei sich die Funktion für sehr starken Verbiss jedoch nicht signifikant von 1 unterscheidet. Im mittleren Bereich sind dagegen etwas mehr Unfälle zu erwarten.

Ist es in einem Gebiet sehr dunkel, so sind dort flächenmäßig weit weniger Unfälle zu erwarten. Schlechtere bis mittelgute Lichtverhältnisse gehen offenbar mit mehr Unfällen einher, wobei mit steigender Helligkeit, also in dichter besiedelten Gebieten immer weniger bis kaum noch Unfälle zu erwarten sind.

Der multiplikative Einfluss der Lokation ist in Abbildung 4.7 dargestellt. Offenbar schwankt die Intensität nach Adjustierung auf die restlichen Kovariablen noch stark im Raum, woraus sich schließen lässt, dass ein großer Teil der Heterogenität nicht durch diese aufgeklärt werden kann. Den starken Abfall am Alpenrand könnte man zunächst darauf zurückzuführen, dass sich dort auch wenige Straßen des betrachteten Typs befinden. Allerdings ist dieser auch bei Einbeziehung des Straßennetzes zu beobachten, wie sich in den nachfolgenden Abschnitten zeigt.

Da die Verteilung der prädiktierten Werte ziemlich rechtsschief war und es einige sehr große Werte gab, war eine grafische Darstellung der tatsächlichen Werte nicht möglich ohne dass zu viel Aufmerksamkeit auf hohe Werte gelegt wird und der restliche Bereich dabei untergeht. Deshalb sind in Abbildung 4.8 stattdessen die Quantile der gefitteten Intensitätsfunktion dargestellt. Auf diese Art wird auch in den nachfolgenden Abschnitten vorgegangen. Die Beobachtungen auf makroskopischer Ebene, die auch schon bei der nonparametrischen Schätzung der Intensitätsfunktion (Abb. 4.1) gemacht wurden, können hier ebenfalls angestellt werden. Allerdings ist die geschätzte Oberfläche dabei weitaus rauher, da sich die Kovariablen von Lokation zu Lokation unterscheiden. Größere Orte zeichnen sich deutlich als helle Gebiete ab.

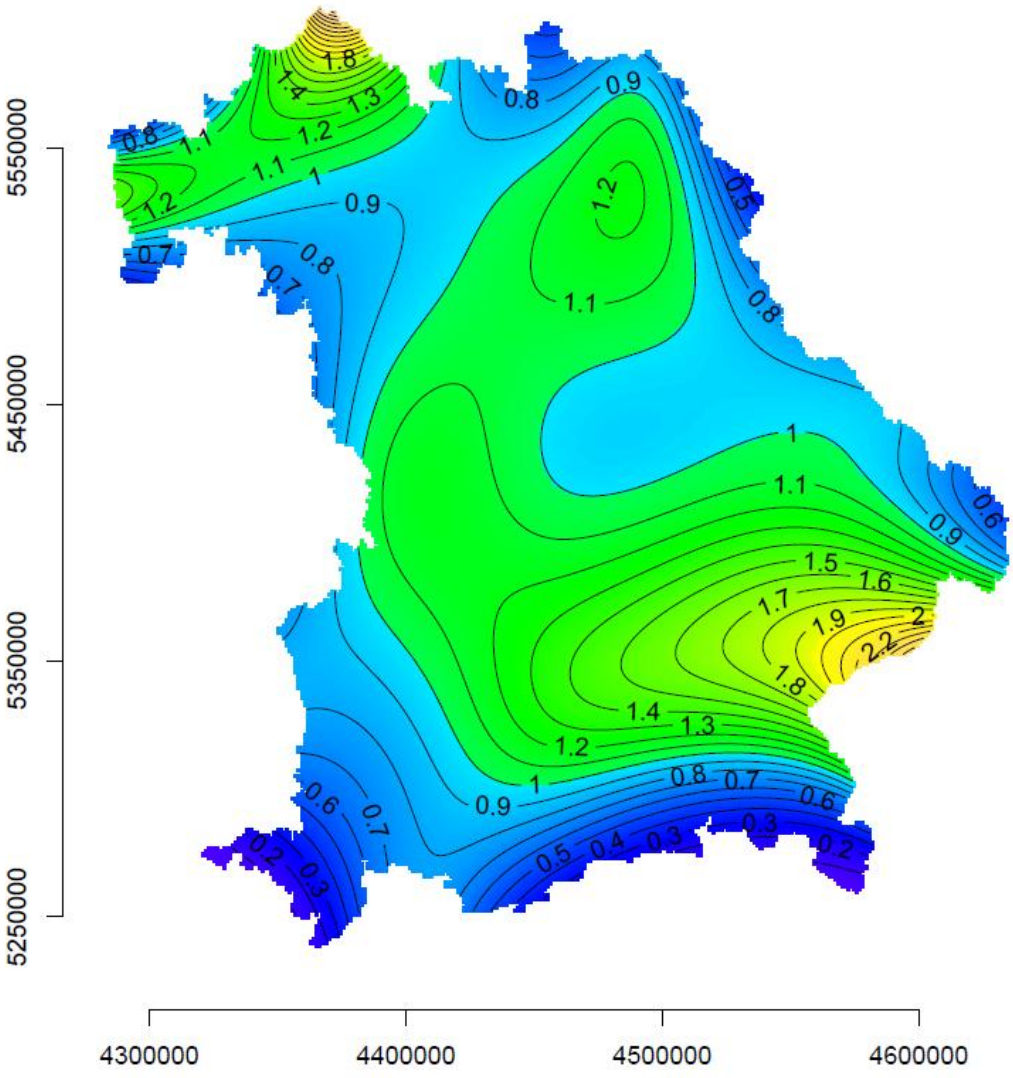

Abbildung 4.7.: Einfluss der Lokation im Modell mit Dummypunkten auf regulärem Gitter

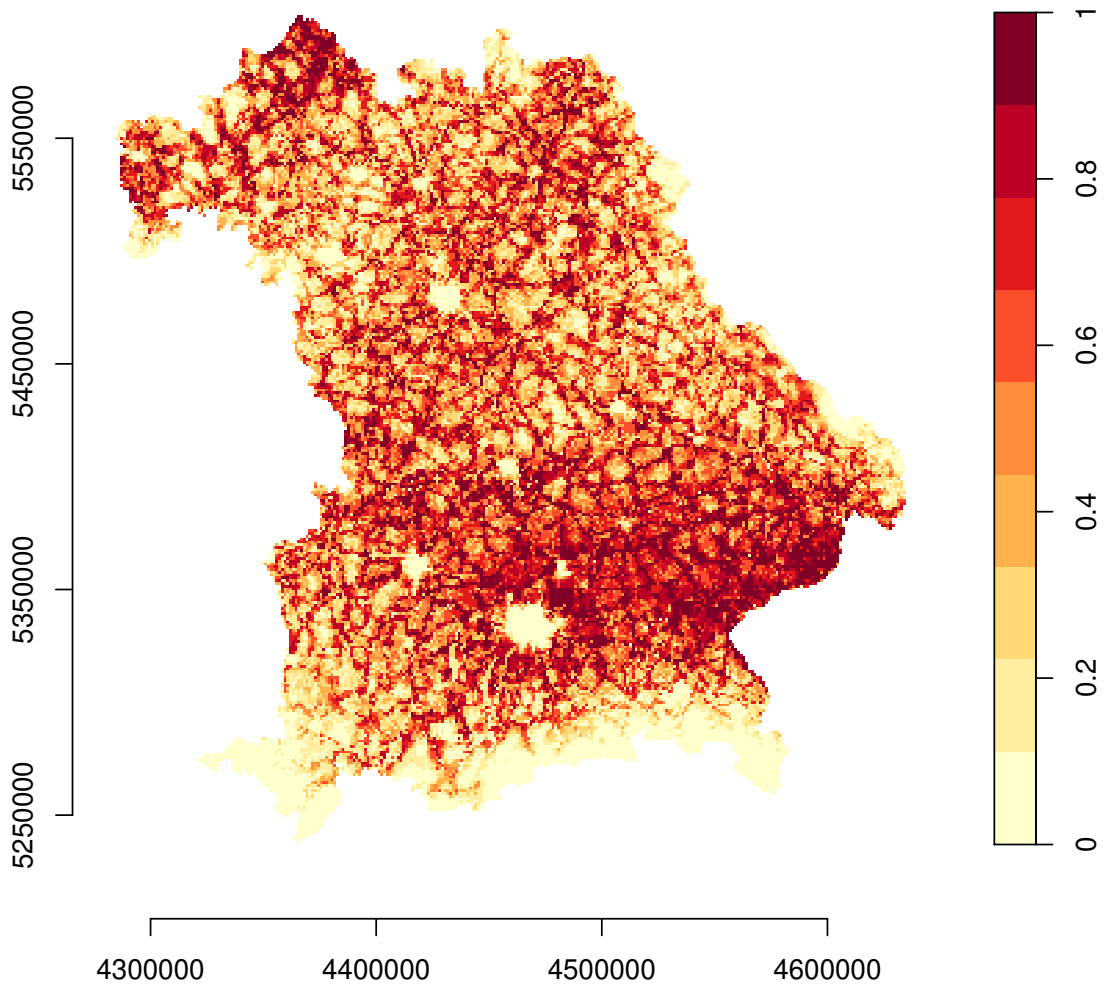

Abbildung 4.8.: Quantile des Fits des Modells bei Dummypunkten auf regulärem Gitter

Residuenanalyse

Wie in Abschnitt 2.1 gesehen, ist für einen räumlichen Punktprozess auf W mit Intensitätsfunktion $\lambda(\cdot)$ die erwartete Anzahl an Punkten, die in einen Bereich $B \subset W$ fällt, gleich $\int_B \lambda(u)du$. Demnach lässt sich das Residuum in einem Bereich B für einen räumlichen Poissonprozess $\boldsymbol{X}$ definieren als:

$$R(B) = N(\boldsymbol{X} \cap B) - \int_B \lambda_{\boldsymbol{\beta}}(u)du. \tag{4.1}$$

Ist das Modell richtig spezifiziert, so gilt $\mathbb{E}(R(B)) = 0$. Eine Möglichkeit, die geschätzten Residuen darzustellen, wäre nun, einen Imageplot der geschätzten Intensitätsfunktion zu erstellen und mit den Beobachtungen zu überlagern. Dann sollten sich in Bereichen mit hoher geschätzter Intensitätsfunktion viele Punkte befinden und in Bereichen mit niedriger wenige. Allerdings wird die Genauigkeit, mit der sich das optisch beurteilen lässt, schnell gering, wenn die Punktedichte höher wird. Eine geeignetere Möglichkeit in diesem Fall wäre es, die Differenz zwischen einer nonparametrischen Schätzung $\widehat{\lambda}(x)$ der Intensitätsfunktion, wie zu Beginn des Kapitels beschrieben und der gefitteten, $\lambda_{\widehat{\boldsymbol{\beta}}}(x)$ zu bilden. Da diese aber durch die Rauheit von $\lambda_{\widehat{\boldsymbol{\beta}}}(x)$ auch sehr rauh wäre, glättet man Letztere ebenfalls:

$$\lambda_{\widehat{\boldsymbol{\beta}}_{\text{glatt}}}(x) = e(x) \int_W K(x-u)\lambda_{\widehat{\boldsymbol{\beta}}}(u)du. \tag{4.2}$$

Hierbei sind $K(\cdot)$ und $e(x)$ wie bei der nonparametrischen Schätzung der Intensitätsfunktion. Damit ist das sogenannte geglättete Residuenfeld gegeben durch (Baddeley, 2008):

$$s(x) = \widehat{\lambda}(x) - \lambda_{\widehat{\boldsymbol{\beta}}_{\text{glatt}}}(x). \tag{4.3}$$

Abbildung 4.9 zeigt das geglättete Residuenfeld für das Modell mit Dummypunkten auf einem regulären Gitter bei Bandbreite $h = 40000$. Die Abweichungen liegen etwa zwischen -4.4×10^{-8} und 5.2×10^{-8} und sind damit im Vergleich zum allgemeinen Niveau der (nonparametrisch) geschätzten Intensität recht klein. Es scheint auch keine zunehmende Verzerrung in Nord-Süd- oder West-Ost-Richtung zu geben. Allerdings wird

die Intensität in Mittelfranken offenbar etwas überschätzt, in Nordschwaben dagegen etwas unterschätzt.

Um in linearen Modellen Auskunft darüber zu erhalten, ob der Einfluss einer Kovariable richtig spezifiziert wurde, trägt man üblicherweise die Residuen gegen die Werte der Kovariablen ab. Ist ein Trend in den Residuen zu erkennen, ist das ein Zeichen dafür, dass der modellierte Einfluss der Kovariable nicht dem wahren entspricht und die Form des Trends gibt Auskunft über die genauere Ausprägung der Misspezifikation. Diese Idee lässt sich in einfacher Weise auch auf die Residuen räumlicher Poissonprozesse (4.1) übertragen. Man spricht hier von „Lurking variable Plots":

Wenn $Q(x)$ die zu untersuchende Kovariable bezeichnet und

$$W(z) = \{x \in W : Q(x) \leq z\}. \tag{4.4}$$

Dann ist die kumulative (geschätzte) Residuenfunktion gegeben durch

$$A(z) = N(\boldsymbol{x} \cap W(z)) - \int_{W(z)} \lambda_{\widehat{\boldsymbol{\beta}}}(u)du. \tag{4.5}$$

Bei richtig spezifiziertem Einfluss gilt wieder $\mathbb{E}(A(z)) = 0$ für alle $z \in \mathbb{R}$. Der Lurking Variable Plot (LVP) ist der Plot $(z, A(z))$. Um Auskunft darüber zu erhalten, welche Bereiche in den Kovariablen besonders stark zu abweichendem Verhalten beitragen, kann man zusätzlich die Ableitung von $A(z)$ betrachten.

Die numerische Berechnung des Integrals der Intensitätsfunktion geschieht durch Aufsummierung der gewichteten Werte an den Stützstellen, d.h. den Daten- und Dummypunkten. Wenn man, wie in (4.1) und (4.5) der Fall, das Integral nur über einen Teilbereich des Beobachtungsfensters berechnen will, summiert man die Werte entsprechend nur für die Stützstellen innerhalb dieses Bereichs auf. In den Formeln geht außer des Integrals der Intensitätsfunktion nur noch die Anzahl der Datenpunkte ein. Zur Berechnung dieser kann man zunächst eine Indikatorfunktion auf den Stützstellen definieren, mit Ausprägung 1, wenn ein Punkt ein Datenpunkt ist und die Werte dieser für alle Stützstellen in dem betrachteten Bereich aufsummieren.

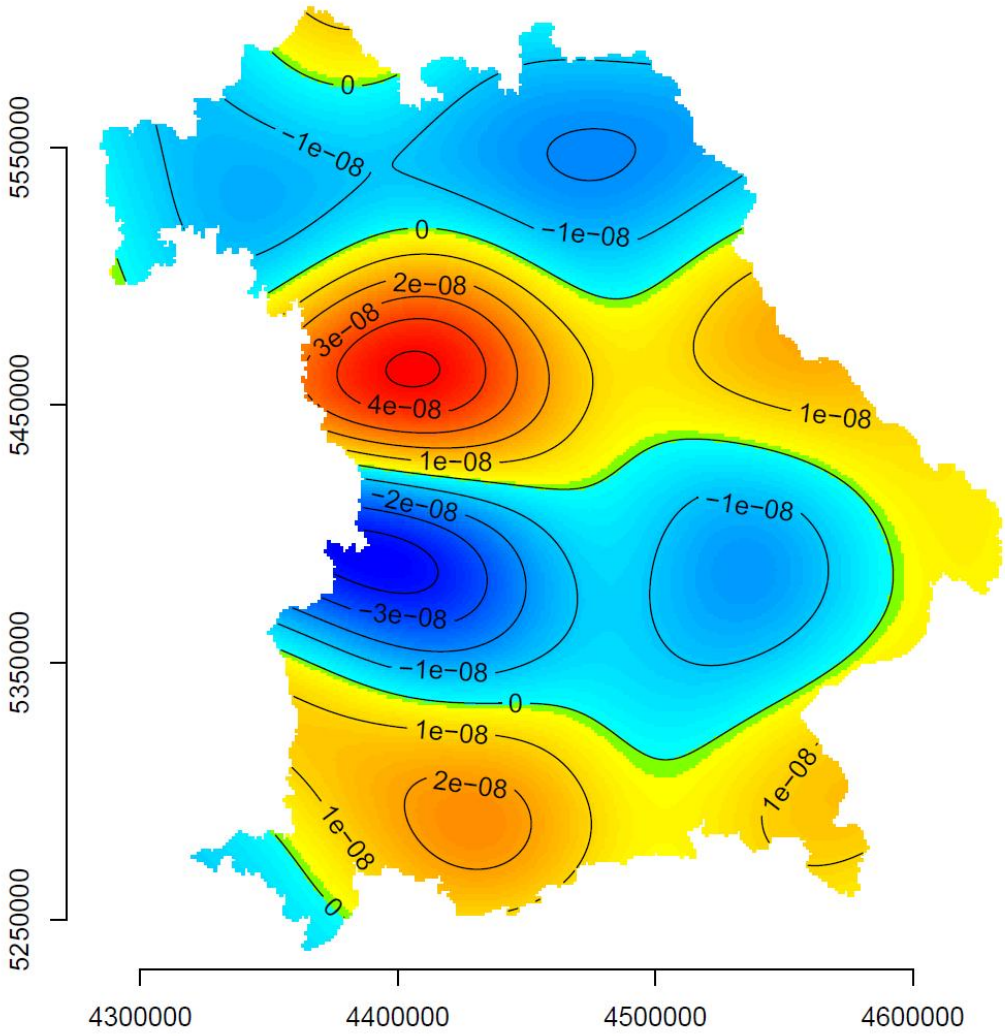

Abbildung 4.9.: geglättetes Residuenfeld für Modell mit Dummypunkten auf reguläreem Gitter

Mit diesen Ergebnissen kann man für praktische Zwecke, wie in `spatstat` geschehen, punktweise Residuen an den Stützstellen definieren:

$$\tilde{R}(s_j) = N_j - w_j \lambda_{\boldsymbol{\beta}}(s_j) \quad \text{mit} \quad N_j = \mathbf{1}\{s_j \in X\} \quad \text{und} \quad j = 1, \ldots, M \tag{4.6}$$

und zur Berechnung des (approximierten) Residuums in einem Bereich B diese für alle $s_j \in B$ aufsummieren. Zur Schätzung der Ableitung von $A(z)$ wird hierauf zunächst ein Glättungsspline angewendet und dieser abgeleitet, für Details hierzu siehe Baddeley et al. (2005b).

Abbildung 4.10 zeigt die LVPs mit zugehörigen geschätzten Ableitungen für `verbiss`. Die kumulierten Residuen zu `verbissfichte` sind für sehr geringen Verbiss stark negativ, der Verbiss wird in diesem Bereich also noch stark unterschätzt, schwanken dann jedoch ohne systematische Abweichungen um den Wert Null, was dafür spricht, dass in diesem Bereich keine groben Verzerrungen mehr zu erwarten ist. Eine Feinheit, die man aufgrund der Skaleneinteilung der x-Achse aus der Grafik kaum ablesen kann, ist dass für einen Verbissanteil von Null die Kurve in etwa bei Null liegt. Offenbar weist der Einfluss des geglätteten Verbisses an Fichten also für sehr kleine Werte ein abweichendes Verhalten auf. Das könnte insbesondere deshalb problematisch sein, da der beobachtete Verbiss in über 50% der Fälle Null ist. Daher liegt ein großer Teil der Werte, deren Einflüsse mit den Spline-Funktionen modelliert werden im untersten Bereich. Um das zu berücksichtigen, könnte man nun in diesem Bereich besonders viele Knoten wählen, um eine besonders flexible Schätzung und damit eine bessere Anpassung zu ermöglichen. Die große Anzahl an beobachteten Nullwerten lässt jedoch vermuten, dass es viele Gebiete gibt, in denen es tatsächlich keinen Verbiss an Fichten gibt, wodurch die Glättung dort mit Werten im Intervall $[0, 1]$ notwendigerweise nach oben hin verzerrt ist. Etwa 50% der geglätteten Werte sind kleiner als 0,02. Deshalb wäre es möglicherweise sinnvoll, geglättete Werte, die einen festgelegten Schwellenwert unterschreiten von vornherein auf Null zu setzen. Da die Nullwerte dann aber einem relativ großen Anteil einnähmen und, noch wichtiger offenbar in ihrem Einfluss auf die Wildunfallintensität ein verglichen mit den restlichen Anteilswerten abweichendes Verhalten aufweisen, würde es wieder zu

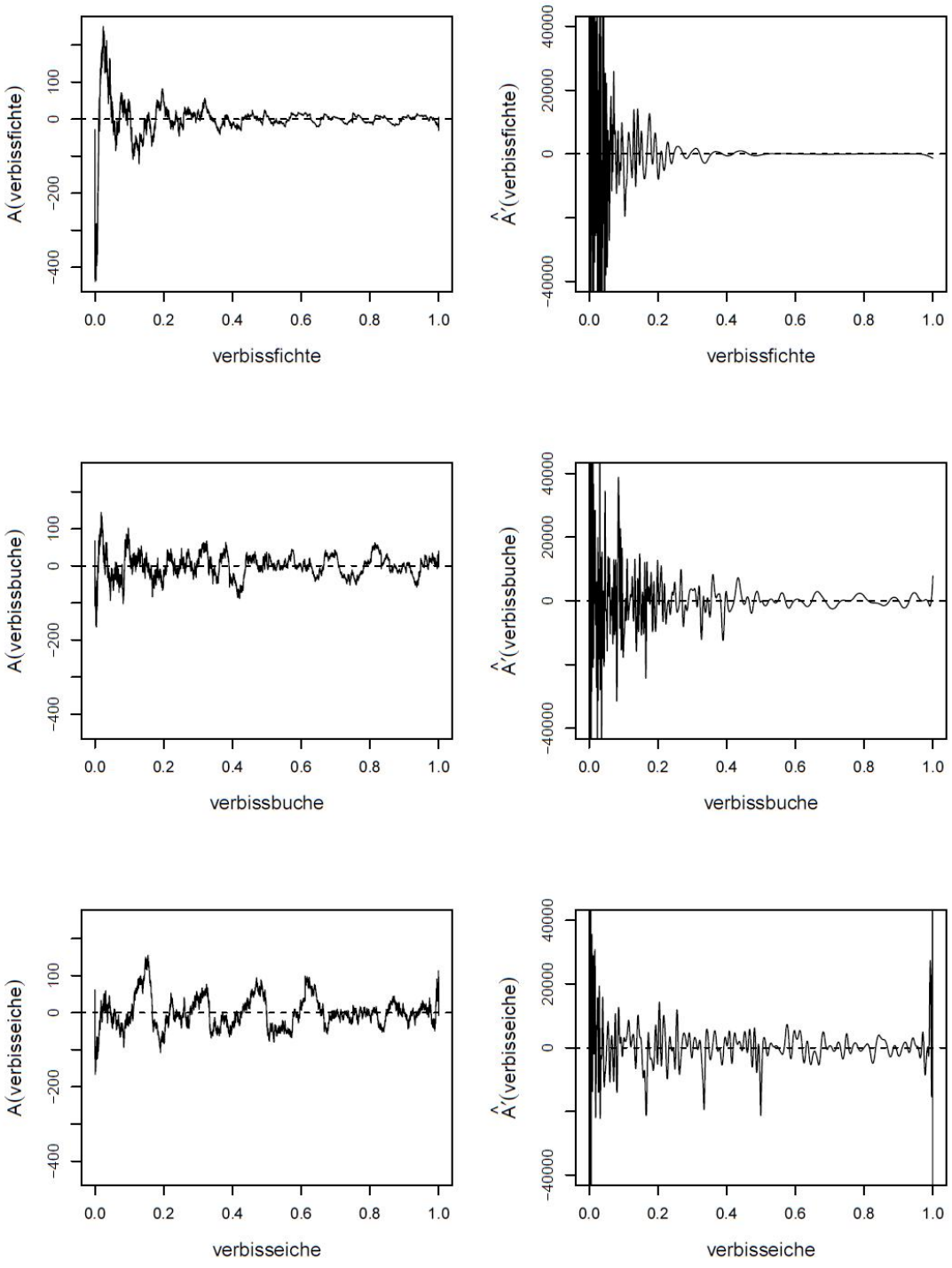

Abbildung 4.10.: Lurking Variable Plots mit Ableitungen - `verbiss`

Verzerrungen kommen, wenn man diese bei der Schätzung der Spline-Funktionen mit einschließt. Es wäre dann besser, eine Indikatorfunktion „Verbiss an Fichten (ja, nein)" mit in den Prädiktor aufzunehmen und die Verbissanteile nur in Bereichen, in denen vorausgesetzt wird, dass sich Verbiss ereignet, mit Spline-Funktionen zu modellieren. Die dabei entstehenden Verzerrungen, wenn in Gebieten, in denen der wahre Verbiss nicht Null ist, die Schätzung irrtümlich auf Null gesetzt wird, wäre wohl geringer als die, die bei non-parametrischer Schätzung des gesamten Verbisses entsteht. Eine Schwierigkeit wäre allerdings, den Schwellenwert festzulegen, ab dem Werte auf Null gesetzt werden. Hierzu könnte man evtl. Kreuzvalidierung einsetzen, wozu allerdings eine sehr große Rechenkapazität notwendig wäre.

Für `verbissbuche` und `verbisseiche` lassen sich ähnliche Beobachtungen machen, allerdings in stark abgeschwächter Form, vermutlich da der beobachtete Verbiss nur in einem guten Drittel der Fälle gleich Null war. Es sei angemerkt, dass bei `verbisseiche` für geglättete Werte von Null die Funktion im Gegensatz zu bei `verbissfichte` und `verbissrotbuche` nicht in etwa bei Null lag, sondern noch stärker negativ war. Vor dem Hintergrund, dass bei `verbisseiche` weit mehr geglättete Werte von Null vorkamen, sodass die Glättung im untersten Bereich offenbar näher an den tatsächlichen Werten lag, bekräftigt dies weiter die Vermutung, dass Werte von Null getrennt modelliert werden sollten. Für `verbisseiche` lässt sich eine zusätzliche interessante Beobachtung machen: Der LVP scheint recht systematisch in Abhängigkeit von den Verbisswerten um die Null zu schwanken. Bei genauer Betrachtung fällt auf, dass die Stellen, an denen die Kurve schlagartig ins Negative fällt, Brüchen mit kleineren Zahlen im Nenner entsprechen, besonders deutlich zu sehen für 1/2, 1/3 und 1/6, an denen die Intensität offenbar stärker überschätzt wird. Auch im Wert 1 lässt sich ein starker Abfall erkennen. Dies sind vermutlich Lokationen an denen eher wenige Eichen stehen. Dass die Intensität also für weniger Eichen stärker überschätzt wird, könnte möglicherweise ein Zeichen dafür sein, dass die zusätzliche Aufnahme der (geglätteten) absoluten Anzahlen an Eichen, bzw. evtl. auch der der anderen Baumarten, den Zusammenhang des Effektes des Verbissanteils mit dem der Wilddichte erhöhen könnte. Ein möglicher Grund dafür könnte sein,

dass wenn tendenziell weniger Eichen vorhanden sind, diese möglicherweise auch eher angefressen werden, was aber nicht zwangsläufig mit einer erhöhten Wilddichte einhergehen muss.

Die Ableitungen der LVPs liefern in diesem Fall keine neuen relevanten Informationen. In allen drei Fällen bestätigen sie das abweichende Verhalten in den unteren Bereichen. Es sind dort starke Schwankungen um den Wert Null zu beobachten, d.h. die LVPs weisen hier im Absolutwert große Steigungen auf. Der besseren Darstellbarkeit wegen wurde der Darstellungsbereich der y-Achse auf das Intervall $[-40000, 40000]$ beschränkt, da aufgrund des starken Abfalls der LVPs zu Beginn, die geschätzte Ableitung sehr klein war. Dass die Schwankungen bei `verbissfichte` für größere Werte schwächer ausfallen als bei den anderen beiden ist vermutlich darauf zurückzuführen, dass sich dort bei `verbissfichte` weniger geglättete Werte befinden. Der Plot zu `verbisseiche` zeigt noch einmal deutlich, dass die für Verzerrungen einflußreichen Stellen an den oben diskutierten Stellen liegen.

Der LVP zu `helligkeit` (Abb. 4.11) weist eine andere Struktur auf. Im unteren Bereich bis etwa zum Wert 10, indem sich etwa 70% der Beobachtungen befinden, sind starke Schwankungen hinsichtlich der Güte der Anpassung zu erkennen. Ob diese tatsächlich auf eine Fehlspezifikation des Einflusses von `helligkeit` zurückzuführen sind, oder ob sie durch nicht berücksichtigte oder fehlspezifizerte Confoundervariablen zustandekommen, lässt sich nicht sagen. Das Modell wurde noch einmal an die Daten gepasst aber mit einer erhöhten Knotenzahl der Spline-Basis zu `helligkeit`. Da sich sehr viele Beobachtungen bei `helligkeit` im unteren Bereich befinden, ergäben sich bei strikt quantilsbasierter Wahl der Knoten auch bei einer hohen Anzahl im oberen Bereich zu wenig Knoten, um eine realitätsnahe Modellierung zu ermöglichen. 70% der Werte sind kleiner als 10, der maximale Wert ist 63. Deshalb wurden in den Intervallen $[2, 10]$ und $(10, 63]$ jeweils 5 äquidistante Knoten gesetzt. Da mit `ppm()` keine beliebigen Argumente an `gam()` übergeben werden können, wurde der Funktion ein zusätzliches Argument `knots` hinzugefügt, in das die Knoten in der für `gam()` üblichen Weise als Liste übergeben werden können.

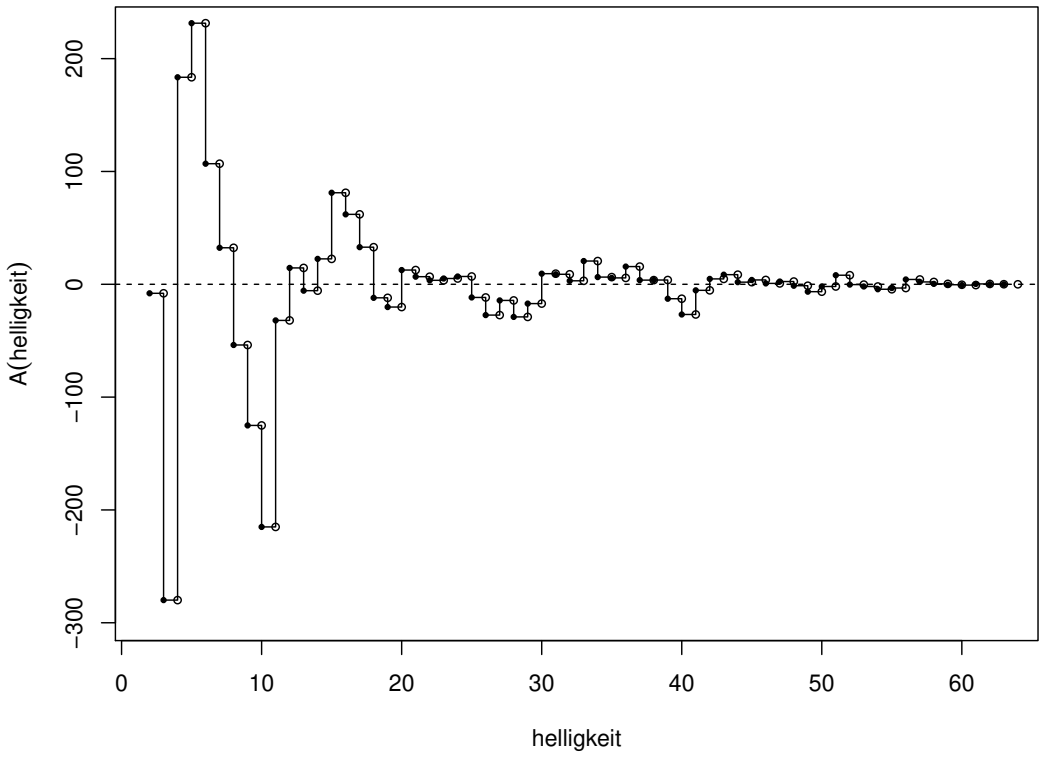

Abbildung 4.11.: Lurking Variable Plots - `helligkeit`

Die abgeänderte Funktion wurde `ppmgamknots()` benannt.

Der aus diesem Modell resultierende Lurking Variable Plot zu `helligkeit` ist in Abbildung 4.12 zu sehen. Die Schwankungen im Bereich $[2, 10]$ sind nun offenbar weitaus weniger stark ausgeprägt. Ein kleiner Nachteil besteht darin, dass die Intensität im obersten Bereich von `helligkeit` etwas unterschätzt wird. Allerdings änderte die flexiblere Modellierung kaum etwas an den gefitteten Werten. Diese korrelierten zwischen den beiden Modellen mit etwa dem Wert 0,984 und waren im flexibleren Modell im Schnitt um ca. den Faktor 0,9997 geringer. Auch die Parameterschätzer der kategorialen Kovariablen änderten sich kaum, weshalb sie nicht noch einmal dargestellt werden. In Abbildung 4.13 sind die exponierten geschätzten Splinefunktionen zu dem Modell mit erweiterter Knotenzahl zu sehen. Durch die flexiblere Modellierung wird offensichtlich, dass sich in

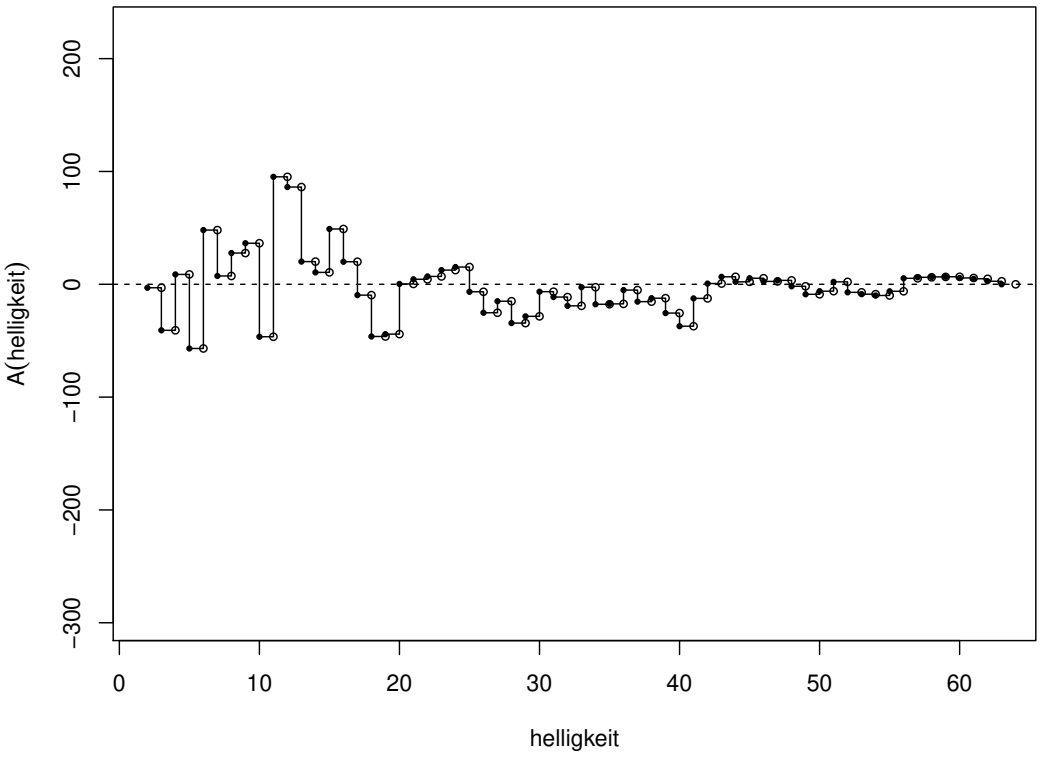

Abbildung 4.12.: Lurking Variable Plots - `helligkeit` bei erhöhter Knotenzahl

sehr dunklen Gebieten offenbar kaum mehr Wildunfälle ereignen. Bei den Einflüssen des Verbisses ist nur für `verbissbuche` eine nennenswerte Änderung zu verzeichnen: Die Funktion ist glatter und die Erhöhung um fast das 1,15-fach bei sehr großem Verbissanteil ist auf etwa das 1,05-fache gesunken. Außerdem ist die Funktion für einen Anteil von 1 nicht mehr signifikant von 1 verschieden.

Obgleich in dieser Arbeit die Analyse der Art des Einflusses der Kovariablen im Vordergrund stand, wurde dennoch eine Einschätzung für die Prognosequalität des Modells gewonnen, vor allem um aufzuzeigen, dass der Erklärwert der Kovariablen bei Weitem nicht ausreicht, um kleinflächige Gebiete zu bestimmen, an denen die Gefahr von Wildunfällen besonders hoch ist. Dazu wurden zunächst bei der Schätzung des Modells eine Reihe von Gegenden in Bayern ausgelassen, in Abbildung 4.14 gelb dargestellt. Für einen

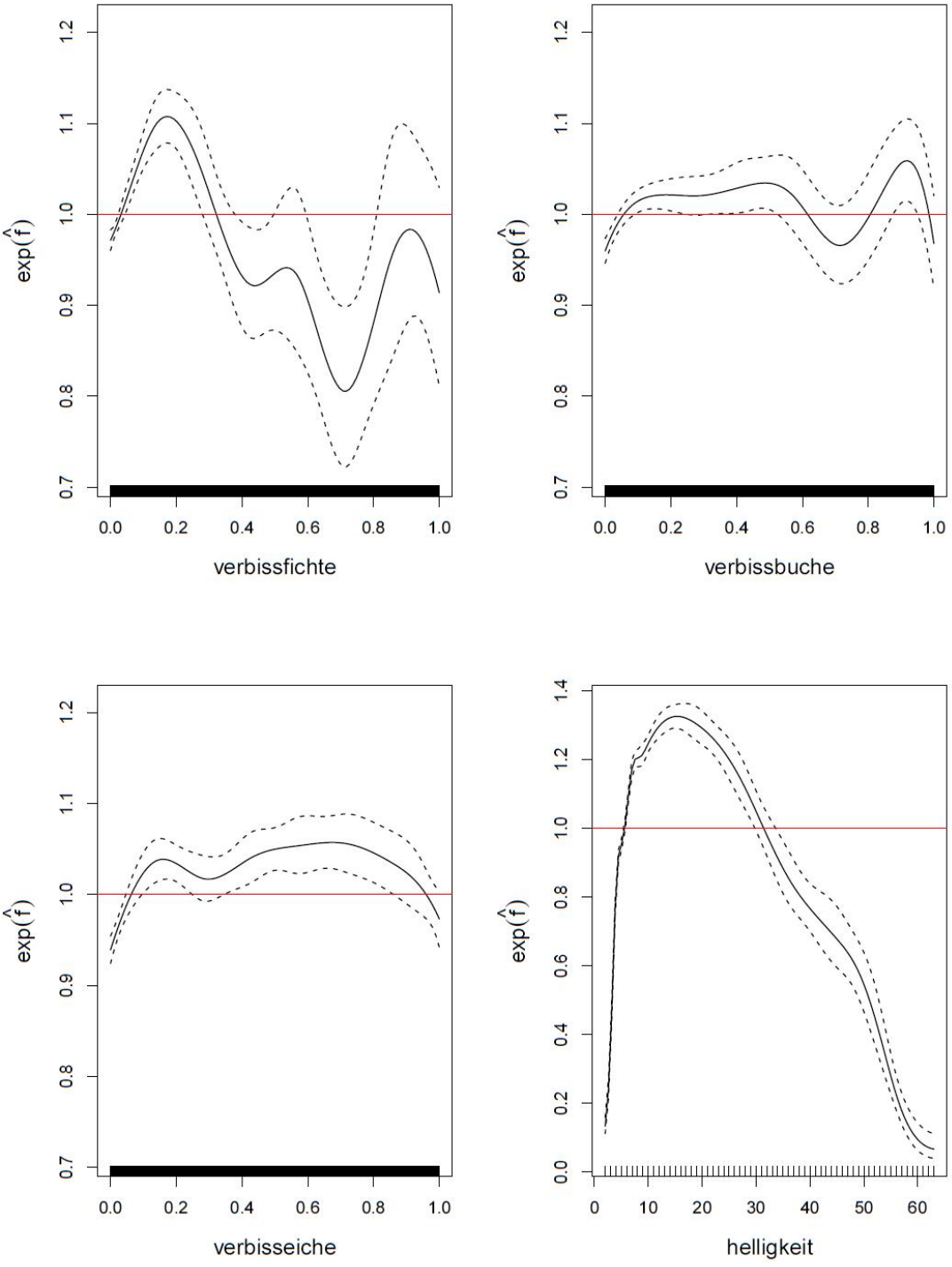

Abbildung 4.13.: exponierte Spline-Funktionen bei erhöhter Knotenzahl für `helligkeit`

Großteil der Kovariablen sind die Werte auf dem Raster der Wildverbissdaten gegeben. Da sich damit die potentiellen Werte der Intensitätsfunktion im Bereich eines Pixels auf diesem Raster ohnehin weniger unterscheiden würden, bot es sich an, die Werte der Intensitätsfunktion ebenfalls auf diesem Raster zu prognostizieren. Außerdem ist die Rasterung so fein, dass die Varianz der Anzahlen an Unfällen auf diesen groß ist, sodass die Fähigkeit des Modells, besonders risikobehaftete Gebiete zu entdecken, gut abgeschätzt werden kann. Um die vorhergesagte Anzahl an Wildunfällen in den 1225 m $\times$ 1225 m-Pixeln zu erhalten, wurden diese Werte anschließend mit 1225^2 multipliziert und mit den tatsächlichen Anzahlen verglichen. Zu Beachten ist, dass die prognostizierten und wahren Anzahlen nicht unabhängig sind, da durch die Aufnahme der Lokation bereits für die räumliche Abhängigkeit Rechnung getragen wird. Abbildung 4.15 links zeigt nun deutlich, wie wenig genau mit dem Modell die wahren Anzahlen vorhergesagt werden können. Die prognostizierten Werte steigen in Abhängigkeit der wahren Anzahlen nur sehr schwach an. Während bis zu einer Anzahl von 2 Unfällen die prognostizierten Werte zumindest im Mittel noch stetig ansteigen, können selbst sehr große Werte anhand der Prognose offenbar kaum von kleineren unterschieden werden. Dass dennoch keine Verzerrung zu erwarten ist, zeigt Abbildung 4.15 rechts. Im Mittel unterscheiden sich prognostizierte und wahre Werte kaum. Der Rangkorrelationskoeffizient nach Spearman zwischen beobachteten und prognostizierten Werte hatte den Wert 0,262.

Es ist allgemein schwierig, den Beitrag einer Kovariable zur Aufklärung der Heterogenität allein über die grafische Betrachtung der Einflüsse einzuschätzen, insbesondere für Kovariablen, deren Einflüsse als Splinefunktionen modelliert wurden. Es bietet sich daher an, das Modell wiederholt unter Auslassung einer Kovariablen zu schätzen und zu messen, wieviel der durch das Modell erklärten Heterogenität bereits durch die einzelnen Submodelle aufgeklärt ist, das heißt wie verzichtbar eine Kovariable in dem vollen Modell wäre. Ein Maß für die durch Anpassung eines Modells erklärte Heterogenität ist die in Abschnitt 2.4 vorgestellte erklärte Devianz. Deshalb wurden für jede Kovariable die erklärte Devianz des entsprechenden Submodells durch die des vollen Modells geteilt. In diesem Bruch kürzt sich die bei der Berechnung der erklärten Devianz jeweils im Nenner

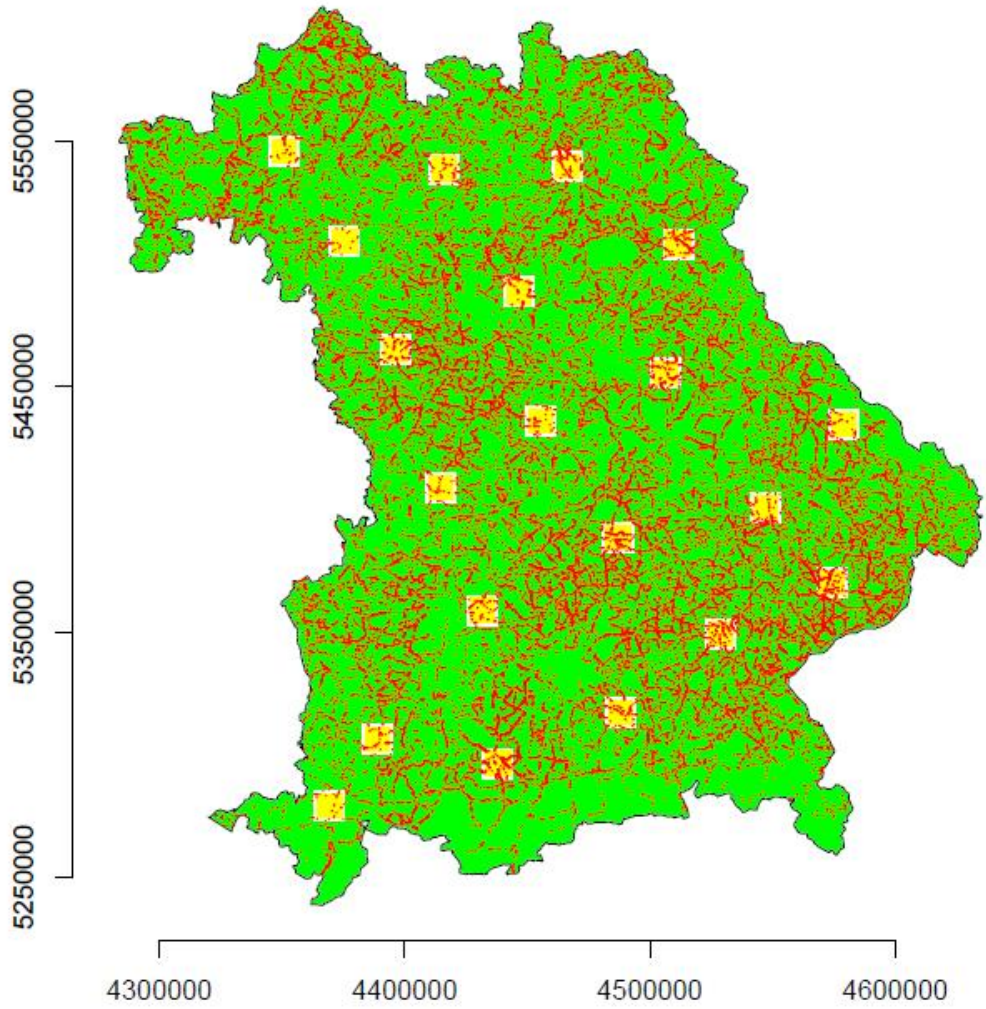

Abbildung 4.14.: Einteilung in Train- (grün) und Testdaten (gelb) zusammen mit den Wildunfällen (rot)

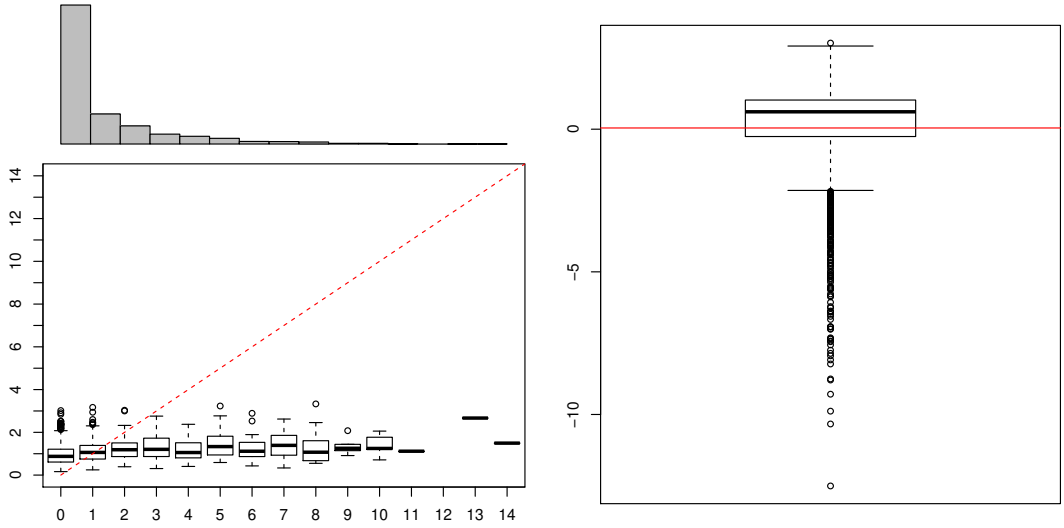

Abbildung 4.15.: links: prognostizierte gegen wahren Anzahlen, rechts: Differenz zwischen prognostizierten und wahren Anzahlen - rote Linie: Mittelwert

stehende Devianz des Nullmodells heraus. Damit kann man die Werte auch einfacher interpretieren, als Anteil der durch das volle Modell aufgeklärten Devianz, die bereits durch das Submodell aufgeklärt würde.

In Abbildung 4.16 sind die Ergebnisse für die betrachteten Kovariablen, angefangen mit dem geringsten Erklärwert der Reihe nach dargestellt. Am wenigsten an der erklärten Heterogenität scheint die Auslassung von `verbiss` zu ändern, das Submodell trägt bereits einen Anteil von ca. 99,6% an der erklärten Heterogenität. Den stärksten Einfluss haben offenbar `strasse` und `coor` mit etwa 70% Anteil an der erklärten Heterogenität. Dass die Lokation verglichen mit den anderen Kovariablen eine derart große Rolle spielt, zeigt auf, dass es offenbar noch wichtige andere nicht-beobachtete bzw. -berücksichtigte räumliche Kovariablen gibt.

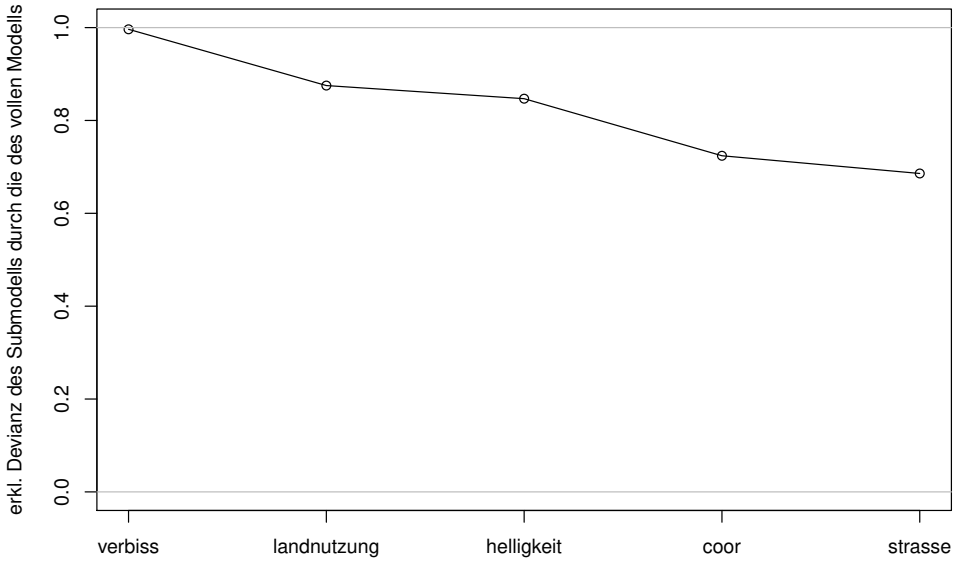

Abbildung 4.16.: Erklärwerte bei Auslassung der jeweiligen Kovariablen

4.1.2. Dummypunkte auf den Straßen

Die numerische Approximation des Integrals der Intensitätsfunktion berechnet sich mit der Standardeinstellung `method = "grid"` wie aus der Beschreibung in vorherigem Abschnitt ersichtlich, als Summe der Mittel der Werte der Intensitätsfunktion an den Stützstellen in je einer Kachel mal der Fläche der Kachel. Befindet sich keine Stützstelle innerhalb einer Kachel, so geht die entsprechende Region auch nicht in die Berechnung des Integrals ein. Die Wahl der Stützstellen und die Auflösung der Kacheln legen also auch den Bereich fest, in dem vorausgesetzt wird, dass Ereignisse stattfinden können. Im Falle der Wildunfälle wären das die Straßen. Aus theoretischer Sicht könnte man bei einer sehr feinen Auflösung und extrem vielen Dummypunkten nur auf den Straßen die Flächen dieser genügend genau approximieren, sodass man die gefittete Intensität tatsächlich als erwartete Anzahl an Wildunfällen auf einem Quadratmeter Straße an einem Punkt auf einer Straße interpretieren könnte. Dies ist aber weder rechnerisch realisierbar, noch ist die Interpretation inhaltlich dienlich, da man bei Einbeziehung der Straßenstruktur

lieber eine Schätzung für die Anzahl an Unfällen pro Meter Straße hätte.

Wählt man die Auflösung bei Dummypunkten auf den Straßen genügend fein, so ist die Quadraturregion bereits so stark um die Straßen konzentriert, dass man die Einflüsse der Kovariablen als Einflüsse auf die Wildunfallintensität im Straßenbereich interpretieren kann. Allerdings entziehen sich die Absolutwerte der Prädiktionen als erwartete Anzahl von Wildunfällen pro Quadratmeter im Bereich der Straßen einer sinnvollen Interpretation. Multipliziert man die so gefittete Intensität jedoch mit der Fläche der entstandenen Quadraturregion und teilt dann durch die Straßenlänge insgesamt, erhält man wie gewünscht eine Schätzung für die Anzahl an Wildunfällen pro Straßenmeter. Die Kachelauflösung muss so fein sein, dass sowohl die Likelihood als auch das Straßennetz genügend genau approximiert wird. Sind diese Bedingungen erfüllt, so ändern sich bei weitere Verfeinerung der Kachelstruktur die Koeffizientenschätzer bzw. bei den Splinefunktionen zumindest der geschätzte Einfluss nur noch wenig. Um dies zu überprüfen wurde das Modell, wie auch schon in Abschnitt 4.1.1 unter Verwendung unterschiedlicher Kachelauflösungen gefittet. Dabei war generell eine feinere Auflösung nötig, da sich in einer grafischen Beurteilung zeigte, dass das Straßennetz mit den in Abschnitt 4.1.1 verwendeten Auflösungen nur ungenau approximiert würde, sodass sich der Bereich über den integriert wird zu wenig von Lokation zu Lokation unterscheiden und damit die Straßenstruktur wenig Berücksichtigung finden würde. Die Koeffizientenschätzer wären damit denen in Abschnitt 4.1.1 ähnlich. Das zur Berechnung der Flächen der Kacheln verwendete Pixelimage wurde mit 15000×15000 Pixeln noch feiner gewählt, um für die kleineren Kachelauflösungen aufzukommen. Zur Gewinnung der Dummypunkte wurde aus dem zur Berechnung der Straßenlängen erstellten Punktmuster zufällig 788156 Punkte ausgewählt, also genau so viele wie in dem ursprünglichen Straßendatensatz vorhandene Eckpunkte. Hätte man das ursprüngliche Punktmuster verwendet, wäre es zu Verzerrungen gekommen, da in diesem die Punkte, abhängig von der Kurvigkeit oder dem Straßentyp unterschiedlich weit auseinander liegen. In dem verfeinerten Punktmuster hingegen, liegen zwei aufeinanderfolgende Punkte der Konstruktion nach höchstens 30 Meter auseinander.

Hier kamen zusätzlich die Kovariablen `kurvigkeit` und `laengen` hinzu, welche, wie auch die anderen (quasi-)stetigen Kovariablen als penalisierte B-Splines modelliert wurden. Bei `helligkeit` wurden dieselben Knoten wie in Abschnitt 4.1.1 gesetzt, um erneut eine flexiblere Modellierung im unteren Bereich zu gewährleisten. Im Fall von `strasse` kann sich eine in Abschnitt 2.3 beschriebene günstige Eigenschaft des Berman-Turner-Devices zu Nutze gemacht werden: Da die Dummypunkte Straßenpunkte sind und die Straßentypen an den Lokationen der Wildunfälle bekannt sind, muss man nicht auf das Rasterimage zurückgreifen, dessen Pixelwerte den Typen der nächstgelegenen Straßen entsprechen.

Abbildungen 4.17 und 4.18 zeigen die geschätzten Koeffizientenschätzer für die unterschiedlichen Auflösungen.

Einen deutlich sichtbaren, jedoch weitgehend abflachenden Trend gibt es nur bei den Schätzern zu `landnutzung`, im ersten Plot hell-, dunkelblau und rosa. Bei genauerer Betrachtung ist zu erkennen, dass die Koeffizientenschätzer zu `strasse` beim Übergang von der gröbsten zur zweitgröbsten Einteilung noch absinken, dann jedoch einen leicht steigenden Trend aufweisen, der mit feinerer Einteilung auch nicht abzuflachen scheint. Das Ergebnis für die Koeffizientenschätzer der Splinebasen ist teils weniger eindeutig. Bei `verbissfichte`, `verbissbuche` und `coor` ist ein relativ konstanter Verlauf zu beobachten. Für `verbisseiche` scheinen sich die geschätzten Koeffizienten mit feinerer Auslösung einander anzunähern. Allerdings zeigte eine Betrachtung der zugehörigen geschätzen Splinefunktionen, dass sich diese zwar für die unterschiedlichen Auflösungen noch leicht ändern, die Unsicherheit in der Schätzung dabei jedoch sehr groß ist, sodass das das Wissen über den genauen Verlauf der Funktion ohnehin sehr gering ist. Obgleich für `helligkeit` die Betrachtung der Koeffizientenschätzer in der höchsten Auflösung ein leicht abweichendes Ergebnis vermuten lässt, unterscheiden sich die geschätzten Splinefunktionen optisch nur sehr gering. Ähnliches lässt sich über `kurvigkeit` sagen: Die Koeffizientschätzer sind zwar teilweise sehr unterschiedlich, jedoch betreffen die entsprechen Abweichungen in den Splinefunktionen nur den untersten Bereich, in dem die

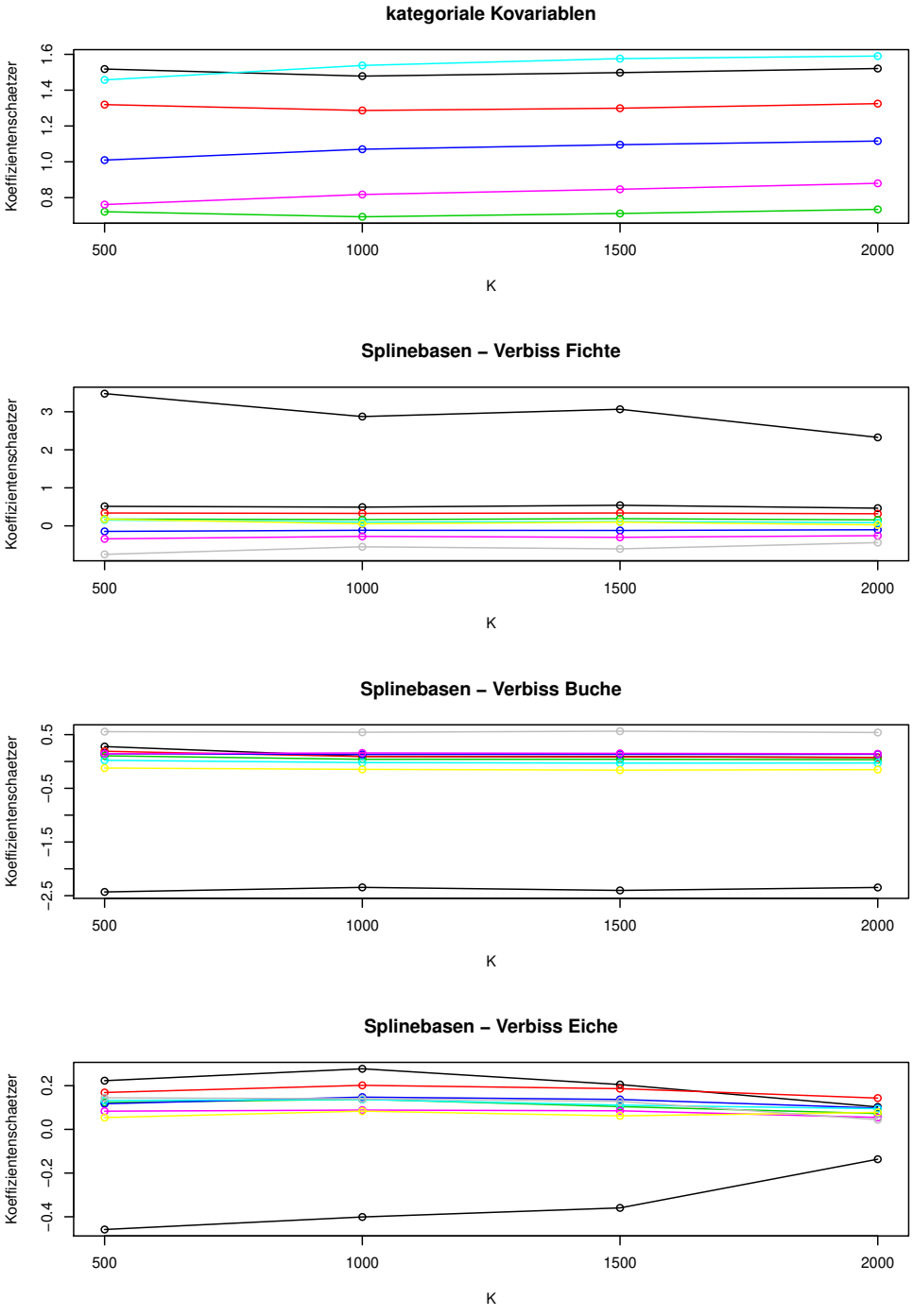

Abbildung 4.17.: Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 1

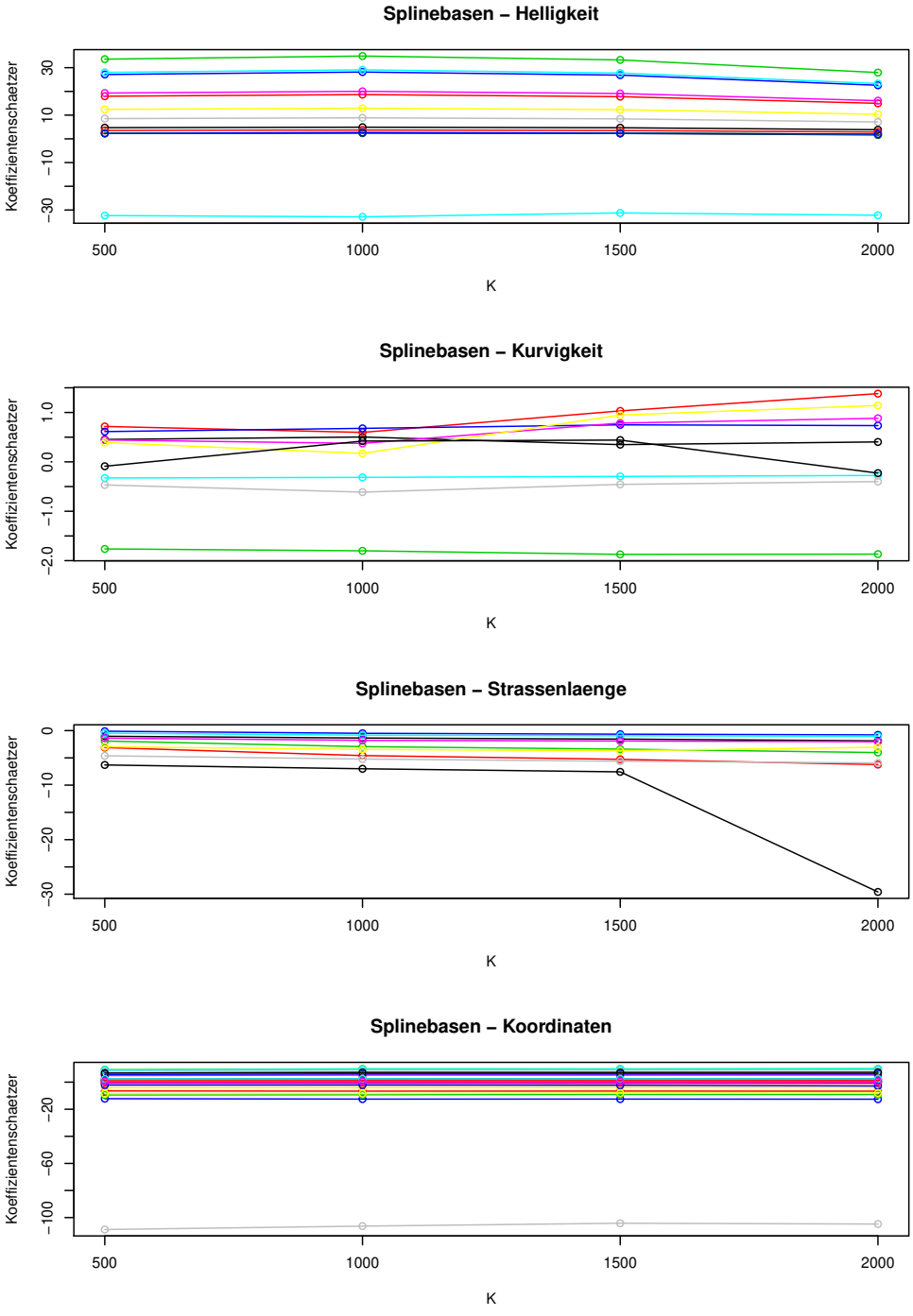

Abbildung 4.18.: Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 2

Konfidenzbänder aufgrund weniger Beobachtungen inflationär groß sind. Obwohl bis auf die feinste Auflösung, in der ein geschätzter Koeffizient eindeutig von den restlichen Ergebnis abweicht, für `laengen` bei Betrachtung der Koeffizientenschätzer nur wenig Unterschiede zu beobachten sind, stellt diese die einzige semiparametrisch geschätzte Kovariable dar, in deren geschätzten Splinefunktionen bis zur höchsten Auflösung hin ein Trend zu beobachten ist. Dieser betrifft den Bereich sehr geringer Straßendichte, in dem das geschätzte Risiko mit feinerer Auflösung weiter ansteigt. Während für $K = 500$ das geschätzte Risiko am linken Rand noch ca. 1.7 so groß wie im Mittel ist, wächst es in der höchsten Auflösung $K = 2000$ bis auf das 3.5-fache an.

Aufgrund letztere Beobachtung wurde die höchste Auflösung gewählt. Anstatt, um ein Abflachen des Trends bei `laengen` und `strasse` zu erreichen, das Gitter der Kacheln noch feiner zu machen, wurde jedoch weiter unten eine binäre Regression auf den Dummy- und Datenpunkten durchgeführt, wobei als Ereignis angesehen wurde, wenn ein Punkt ein Datenpunkt war. Es gab Indizien dafür, dass dieses Modell geeigneter ist, da eine noch stärkere Konzentrierung auf die Straßen erfolgt, Details folgen in Abschnitt 4.2.

Die Schätzung mit Hilfe von `ppmgamknots()` geschieht sehr ähnlich wie mit `ppm()` in Abschnitt 4.1.1:

```
> modelldummyroads <- ppmgamknots(Q, ~ strasse + landnutzung +
    s(verbissfichte, bs="ps") + s(verbissbuche, bs="ps") +
    s(verbisseiche, bs="ps") + s(helligkeit, bs="ps", k=14) +
    s(kurvigkeit, bs="ps") + s(laengen, bs="ps") + te(xcoor,
    ycoor, bs="ps"), covariates=covari, use.gam=TRUE,
    knots=list(helligkeit = knotenhelligkeit))
```

Wie bereits allgemein beschrieben, wird der Vektor der Knoten `knotenhelligkeit` für `helligkeit` dabei mit dem Argument `knots` übergeben.

Der Aufruf der summary-Funktion liefert:

```
> summary(getgam(modelldummyroads))

Family: quasi
Link function: log

Formula:
.mpl.Y ~ strasse + landnutzung + s(verbissfichte, bs = "ps") +
    s(verbissbuche, bs = "ps") + s(verbisseiche, bs = "ps") +
    s(helligkeit, bs = "ps", k = 14) + s(kurvigkeit, bs = "ps") +
    s(laengen, bs = "ps") + te(xcoor, ycoor, bs = "ps")
<environment: 0x33e2718>

Parametric coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -14.54178    0.05049 -288.02   <2e-16 ***
strasseprimary        1.52087    0.03469   43.85   <2e-16 ***
strassesecondary      1.32481    0.03409   38.86   <2e-16 ***
strassetertiary       0.73450    0.03438   21.36   <2e-16 ***
landnutzunglandwirt   1.11589    0.02344   47.60   <2e-16 ***
landnutzungwald       1.59039    0.02472   64.33   <2e-16 ***
landnutzungsonstige   0.87992    0.07182   12.25   <2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Approximate significance of smooth terms:
                    edf Ref.df       F  p-value
s(verbissfichte)  8.747  8.930  11.500  < 2e-16 ***
```

```
s(verbissbuche)   8.938  8.996   5.077 6.91e-07 ***
s(verbisseiche)   6.584  7.364   7.215 5.20e-09 ***
s(helligkeit)    12.771 12.951  54.604  < 2e-16 ***
s(kurvigkeit)     7.071  7.549  69.504  < 2e-16 ***
s(laengen)        6.706  6.979  24.640  < 2e-16 ***
te(xcoor,ycoor)  23.983 24.000 117.424  < 2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

R-sq.(adj) =  -0.034   Deviance explained = 7.12%
GCV score = 0.3223  Scale est. = 0.32227   n = 831439
```

Auch hier sind alle geschätzen Einflüsse zum α-Niveau 5% signifikant von Null verschieden. Die erklärte Devianz ist mit 7,12% zwar etwas größer als im vorherigen Abschnitt, aber immer noch gering. Es konnten fünf Unfälle mehr als in Abschnitt 4.1.1 verwendetet werden, die sich dort außerhalb des Straßenimages befanden. Die Fallzahl betrug also 45706.

Die Ergebnisse der kategoriellen Kovariablen `strasse` und `landnutzung` (Abb. 4.19) unterscheiden sich stark von denen bei Wahl der Dummypunkte auf einem regulären Gitter in Abschnitt 4.1.1. Die Rangfolge der Häufigkeiten an Unfällen nach Straßentyp ist dieselbe wie im vorherigen Abschnitt, aber die Abstände zur Referenzkategorie und der Kategorien untereinander sind viel größer. Laut diesem Modell passieren auf Bundesstraßen etwa 4,5 mal so viele Unfälle wie auf Autobahnen, auf Landesstraßen sind es knapp vier mal so viele. Gut doppelt mal mehr Unfälle ereignen sich unter Voraussetzung der Gültigkeit des Modells auf Kreisstraßen, in vorherigen Abschnitt waren es hier zum Vergleich nur etwa 1,2 mal mehr. Am stärksten betroffen von Wildunfällen sind offenbar Straßen in bewaldeten Gebieten mit fast fünf Mal mehr Unfällen als auf Straßen auf bebauten Flächen. Für Straßen durch landwirtschaftliches Gebiet werden etwa zwei Drittel mal weniger als für solche durch bewaldetes und drei mal mehr als für

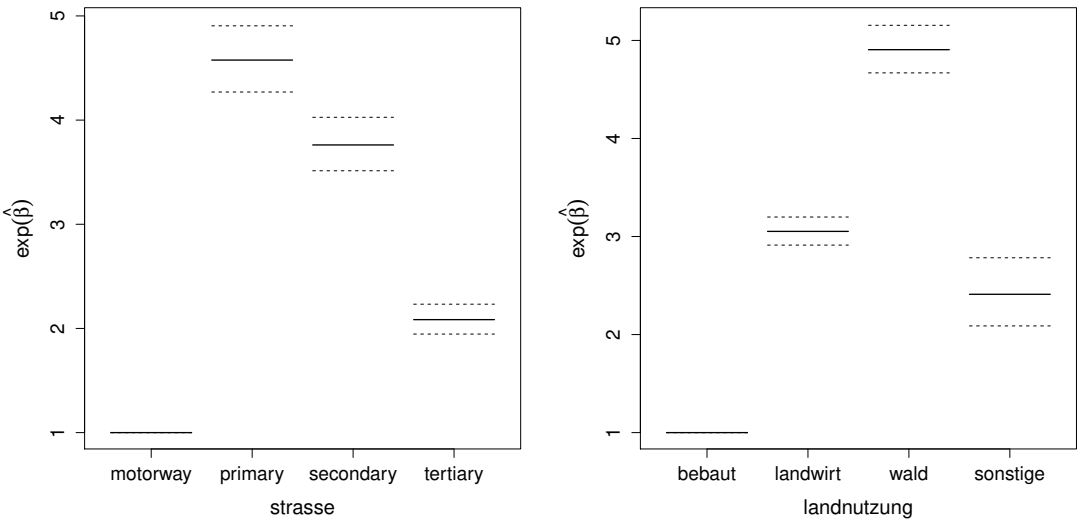

Abbildung 4.19.: exponierte Koeffizientenschätzer der kategorialen Kovariablen im Modell mit Dummypunkten auf Straßen

solche durch bebautes Gebiet prognostiziert. Etwas weniger ereignen sich auf Straßen im Bereich von Feucht- und Wasserflächen, wie auch zuvor unter der Prämisse eines korrekten Modells.

Abbildung 4.20 zeigt die geschätzten Splinefunktionen dieses Modells. Die geschätzten Einflüsse der auch schon im Modell mit Dummypunkten auf einem regulärem Gitter enthaltenen semiparametrisch modellierten Kovariablen weisen im Gegensatz zu `strasse` und `landnutzung` eine große Ähnlichkeit zwischen den beiden Modellen auf, sodass nur auf relevante Unterschiede eingegangen wird. Bei Vergleich der geschätzten Splinefunktion zu `helligkeit` mit der bei erhöhter Knotenzahl aus Ausschnitt 4.1.1 fällt lediglich auf, dass im Bereich mittelstarker Helligkeit die Unfallgefahr hier etwas weniger stark erhöht ist und die Funktion am linken Rand nun annäherend die Null erreicht. Die grundsätzlich Form des Einflusses bleibt jedoch die gleiche. Der Einfluss des Maßes zur Kurvigkeit scheint nicht besonders stark zu sein. Offenbar sind auf besonders geraden Straßen weniger Unfälle zu erwarten, wobei die Unsicherheit hierbei recht groß ist. Für

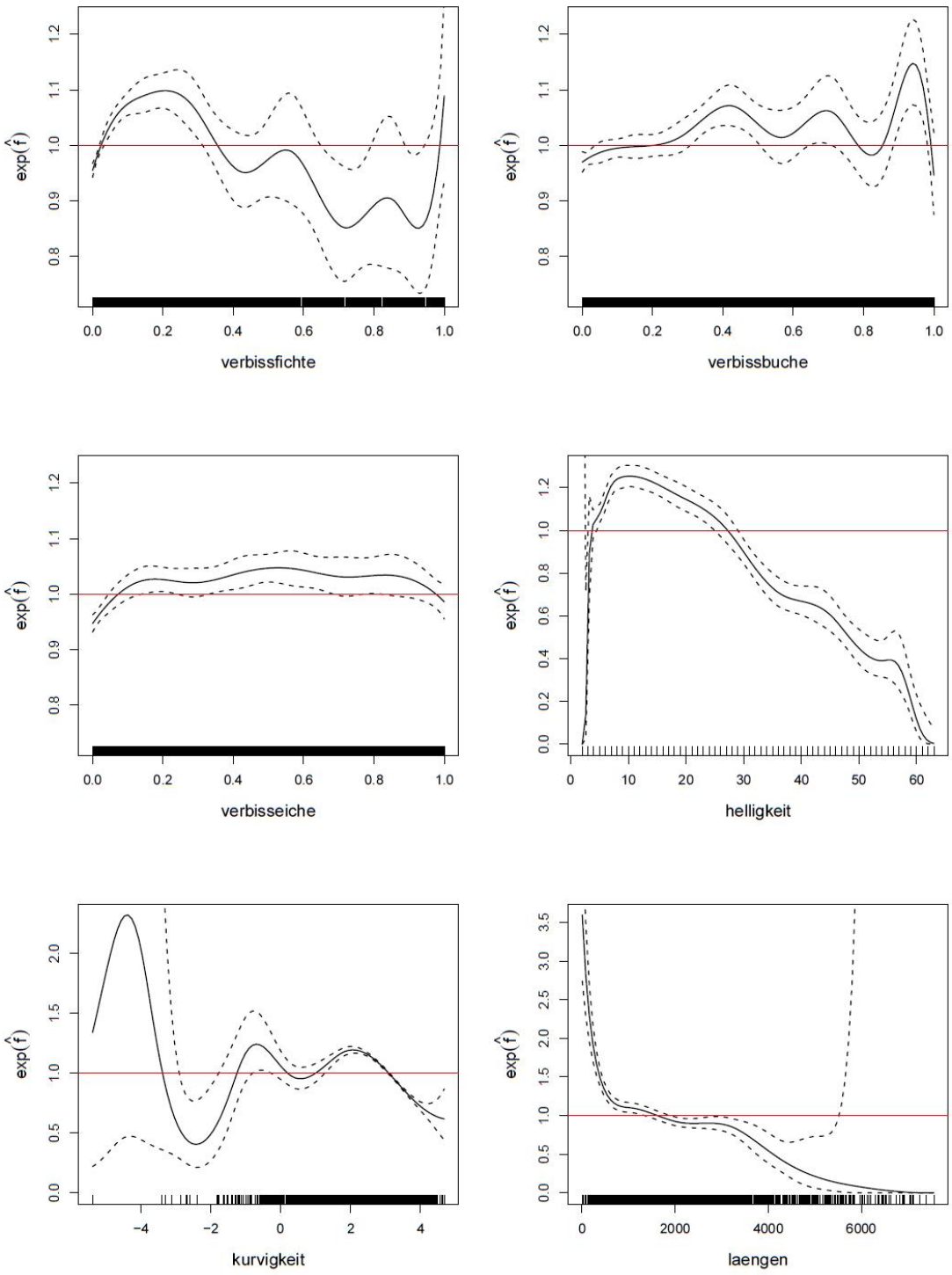

Abbildung 4.20.: exponierte Spline-Funktionen im Modell mit Dummypunkten auf Straßen

geringere bis stärkere Kurvigkeit ist die Gefahr von Wildunfällen bei unregelmäßigem Verlauf der Splinekurve etwas größer und nimmt für sehr starke Kurvigkeit wieder ab. Natürlich ist dieses Ergebnis vorsichtig zu interpretieren, da ein anderes Maß zur Kurvigkeit etwas andere Ergebnisse liefern könnte. Mit steigender Straßendichte sinkt die Gefahr für Unfälle deutlich. Auffällig sind der starke Anstieg für sehr geringe Dichte und das beobachtete schnelle Abfallen am rechten Rand bei allerdings inflationär breitem Konfidenzband.

Der Einfluss der Lokation (Abbildung 4.21) unterscheidet sich nicht stark von dem in vorherigem Abschnitt. Der in Letzterem beobachtete starke Abfall der Intensität am Alpenrand ist auch nach Konzentrierung der Schätzung um die Straßen zu beobachten. Eine Feinheit, die sich aufgrund der begrenzten Zahl an Konturlinien beim direkten Vergleich mit Grafik 4.7 nicht erkennen lässt, ist, dass dieser Abfall am äußersten unteren Rand weit weniger stark ist, sodass sich sehr wohl ein Effekt der Berücksichtigung der Straßen zeigt. In Niederbayern sind offenbar auch hier besonders viele Unfälle zu erwarten, allerdings fällt der Unterschied geringer aus, wenn die Straßenstruktur berücksichtigt wird. In der nördlichen Hälfte werden tendenziell etwas weniger Unfälle prognostiziert als in der südlichen. Allgemein ist der prognostizierte Einfluss der Lokation etwas glatter, sodass offenbar mehr Heterogenität aufgeklärt werden konnte.

Abbildung 4.22 zeigt nun die Quantile der mit der Quadraturregion multiplizierten und durch die Straßenlänge insgesamt geteilten gefitteten Intensitätsfunktion. Diese Transformation war nötig, um wie oben beschrieben, eine Schätzung für die Anzahl der Wildunfälle pro Straßenmeter zu erhalten. Um eine bessere optische Beurteilung zu ermöglichen, wurde ein Prognosewert für jeden Pixel gewonnnen, ungeachtet ob durch diesen eine Straße geht oder nicht. Dazu wurden zunächst Prädiktionen für die Pixel erstellt, durch die tatsächliche Straßen führen und anschließend allen Pixeln ohne Wert jeweils der des nächstgelegenen, durch den eine Straße geht, zugeordnet. Allgemein ist die gefittete Intensitätsfunktion weniger glatt als die bei Dummypunkten auf einem regulärem Gitter. Die Gründe hierfür sind vermutlich, dass die Kovariablen `kurvigkeit` und `laengen`

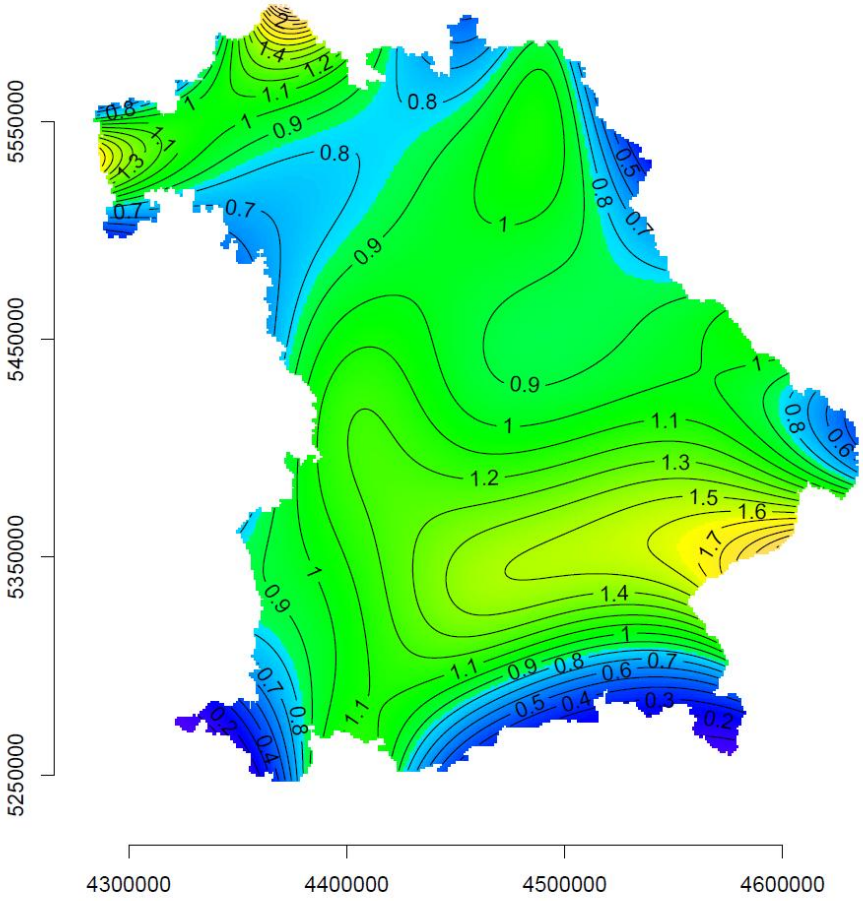

Abbildung 4.21.: Einfluss der Lokation im Modell mit Dummypunkten auf Straßen

zusätzlich aufgenommen wurden und dass die Koeffizientenschätzer der kategorialen Einflußgrößen im Absolutbetrag weiter voneinander und von der Referenzkategorie entfernt liegen. So lassen sich die Autobahnen z.B. teils gut als helle Linien erkennen. Weitere Unterschiede sind, dass am Alpenrand, besonders im westlichen Bereich und im Nordwesten Unterfrankens mehr, in Niederbayern dagegen mit Ausnahme des nördlichen Bereichs weniger Unfälle prognostiziert werden.

Trotzdem die Schätzung hier auf Bereiche um die Straßen konzentriert ist und sich die absoluten Werte der Prädiktionen daher zunächst nicht sinnvoll interpretieren lassen, können Lurking Variable Plots wie in Abschnitt 4.1.1 angewendet werden, da bei einem richtig spezifizierten Modell die Anzahlen an Unfällen in den integrierten Bereichen dennoch im Schnitt korrekt vorhergesagt werden. Die Ähnlichkeit der geschätzten Splinefunktionen zu `verbiss` mit denen aus Abschnitt 4.1.1 lässt bereits vermuten, dass sie sich nicht stark hinsichtlich ihrer Nähe zum tatsächlichen Einfluss von `verbiss` unterscheiden. Dies wird dadurch bekräftigt, dass sich auch die entsprechenden Lurking Variable Plots nicht wesentlich unterscheiden, sodass diese aus Gründen der Knappheit nicht dargestellt werden. Ein feiner Unterschied in der Ableitung des LVP zu `verbissbuche` war, dass diese bei einem Verbissanteil von 1 eine starke Schwankung nach oben aufwies, was ein Anzeichen dafür ist, dass dort die Intensität unterschätzt wird, sodass das Abfallen der Splinefunktion am rechten Rand vorsichtig interpretiert werden sollte. Der Abfall könnte jedoch auch damit zusammenhängen, dass, wie in Abschnitt 4.1.1 bereits diskutiert, wenn sich wenige Buchen an einer Lokation befinden, diese vermehrt angefressen werden könnten ohne dass die Rehpopulationen an dieser Lokation besonders groß ist.

Der LVP zu `helligkeit` (Abb. 4.23, unten links) weist auf zwei strukturelle Unterschiede hin: Offensichtlich ist der Einfluss im Bereich für Werte kleiner als zehn besser an die Daten angepasst, da die Schwankungen hier wesentlich kleiner sind. Des Weiteren wird die Intensität im obersten Bereich nicht mehr unterschätzt. Generell schwanken die Residuen recht unregelmäßig um den Wert Null, sodass mit keinen bedeutenden Verzerrungen zu rechnen ist. Der Einfluss von `kurvigkeit` scheint auch im Wesentlichen

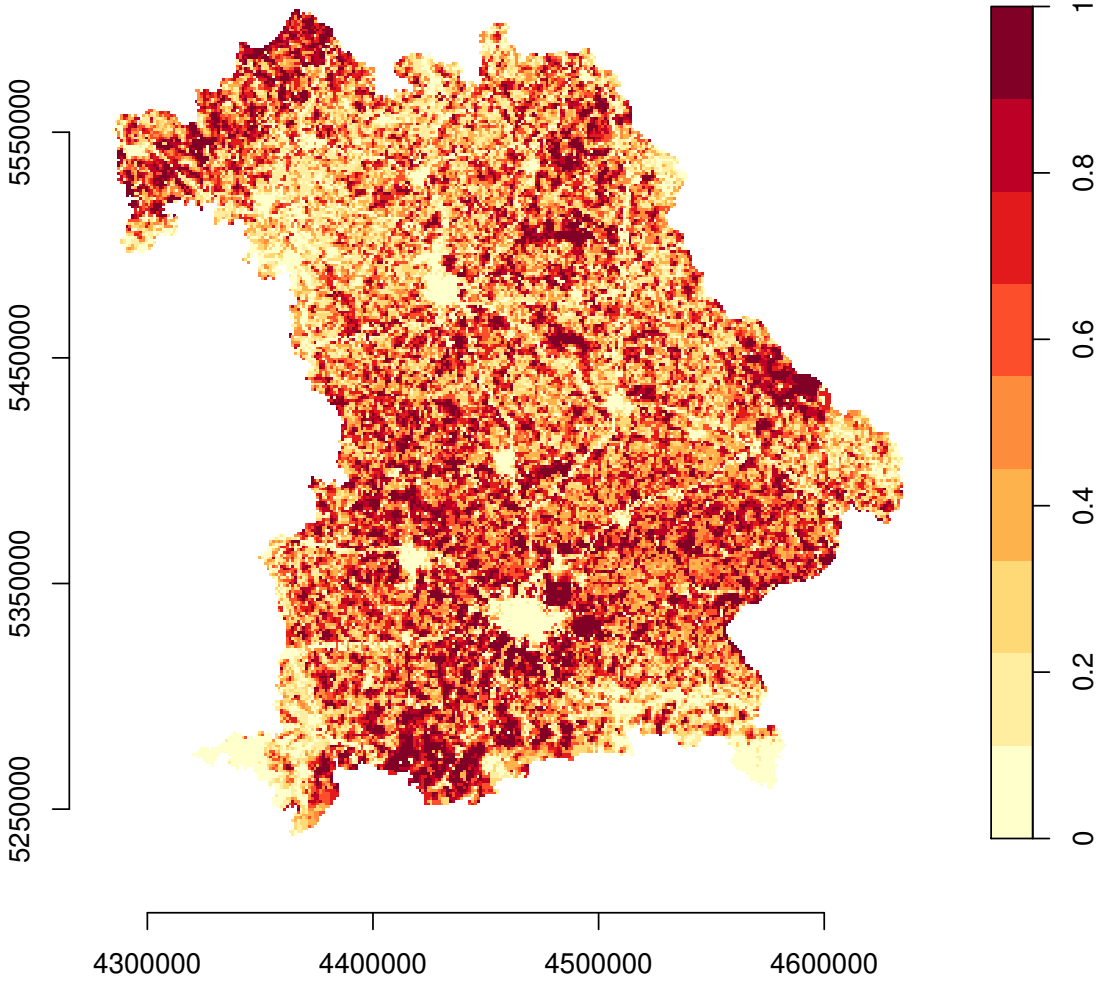

Abbildung 4.22.: Quantile des Fits des Modells bei Dummypunkten auf Straßen

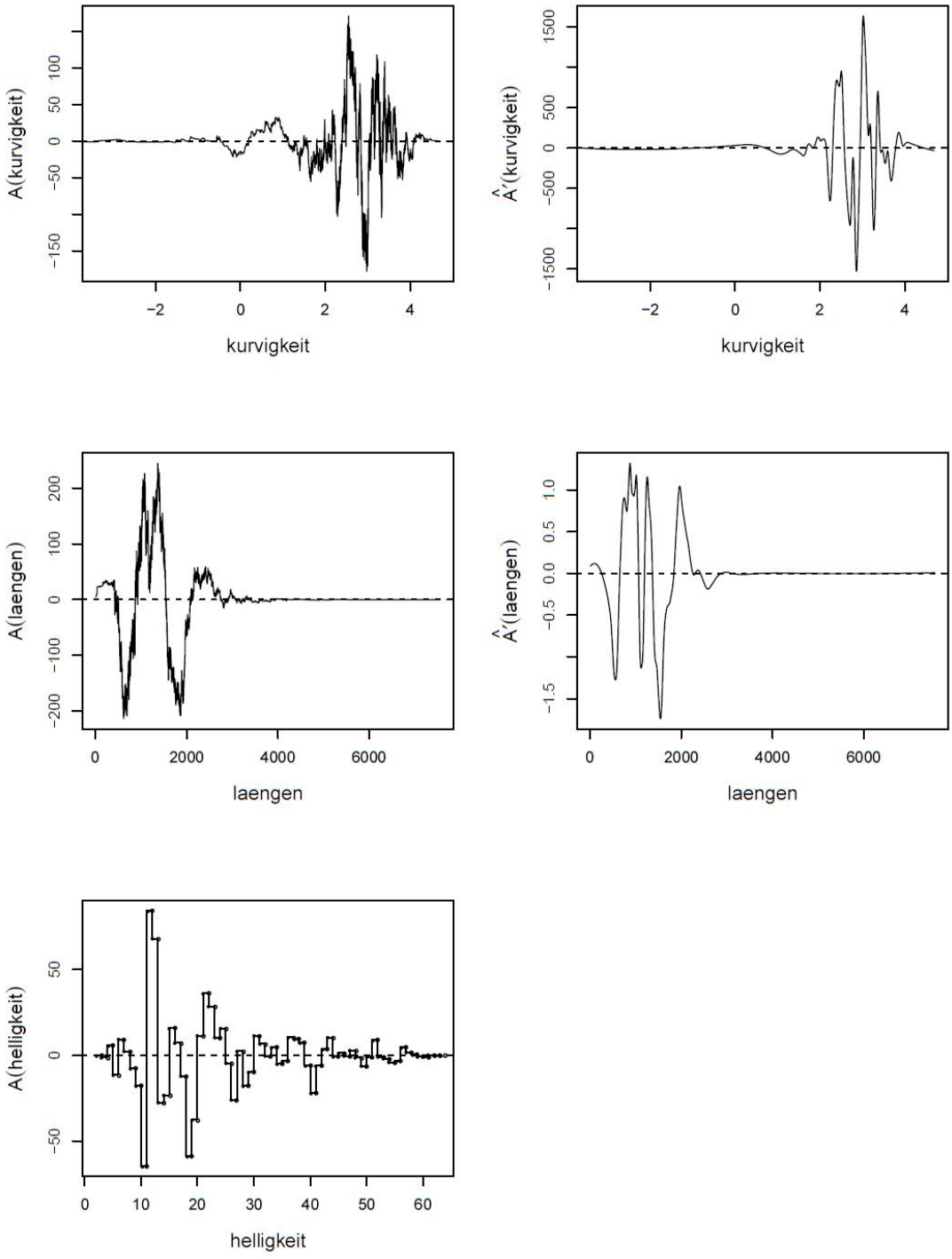

Abbildung 4.23.: Lurking Variable Plots mit Ableitungen

richtig spezifiziert zu sein - lediglich im Bereich von etwa -0.5 bis 1.5, in dem sich ohnehin nur sehr wenige Beobachtungen befinden, sind systematische Abweichungen im LVP (Abb. 4.23, oben) zu erkennen. Bei Betrachtung des LVPs zu `laengen` (Abb. 4.23, Mitte links) kann man solche Aussagen nicht machen. Die Perioden der Schwankungen sind hier zu groß, um noch von zufälligen Schwankungen zu sprechen. Die Form der Abweichungen könnte man sich damit erklären, dass sich das Straßennetz in seltenen Fällen nicht vollständig mit dem Punktmuster der Wildunfälle deckte. So kam es vor, dass Straßen scheinbar im Nichts endeten oder komplett fehlten. Dass sich dort aber in Wirklichkeit Straßen befinden, ergibt sich daraus, dass sich an diesen Stellen Wildunfälle ereigneten. Das führt zu einem zum Teil starken Übergewicht von Unfallpunkten im Verhältnis zu Straßenpunkten, also zu einer vermeintlich hohen Intensität an Wildunfällen in Bereichen vermeintlich sehr geringer Straßendichten. Das ist vermutlich für den starken Anstieg der geschätzten Splinefunktion zu `laengen` am äußerten linken Rand verantwortlich. Dass sich der LVP am linken Rand im positiven Bereich befindet, deutet darauf hin, dass der Anstieg eigentlich noch stärker sein müsste, um den Daten gerecht zu werden. Der geschätzte Koeffizient der ersten Basisfunktion hat also einen sehr großen Wert. Dies führt zu alternierend und abflachend positiven und negativen Verzerrungen im weiteren Kurvenverlauf, da sich die Splinebasen überlappen und der jeweils nächste Koeffizient den starken Einfluss der vorherigen gewichteten Basisfunktion neutralisieren muss, wodurch es aber gleichzeitig anschließend zu einer Verzerrung in die jeweils gegensätzliche Richtung kommt. Trotz diesen Abweichungen vom tatsächlichen Einfluss wurden an diesem Punkt keine Maßnahmen ergriffen, da es keine inhaltliche Begründung dafür gibt, dass die Tatsache, dass sich in seltenen Fällen Unfälle nicht auf Straßen aus dem verfügbaren Straßennetz befanden auch zu Verzerrungen im geschätzten Einfluss bei den restlichen Kovariablen führt und da die zugehörigen LVPs auch auf keine Abweichungen hindeuten, die damit erklärt werden könnten. Die geschätzten Ableitungen der LVPs zu `kurvigkeit` und `laengen` (Abb. 4.23, oben rechts bzw. Mitte rechts) liefern hier keine neue Einsichten.

Für dieses Modell wurde analog wie im vorherigem Abschnitt die Prognosegüte ab-

geschätzt. Der einzige Unterschied ist, dass hier, gemäß der Interpretation des Fits zum Vergleich mit der Prognose die Anzahl an Wildunfällen pro Straßenmeter berechnet wurde und dass demnach Rasterquadrate, durch die keine Straßen gingen, nicht berücksichtigt wurden. Als tatsächliche Werte wurden die Anzahlen an Unfällen pro Straßenmeter in den Rasterquadraten berechnet, wobei für die Straßenlängen hier nicht die geglätteten (siehe Abschnitt 3.7), sondern auf jedem Rasterquadrat auf dem die Prädiktion erfolgen sollte, die tatsächlichen verwendet wurden. Das Ergebnis (Abbildung 4.24) ist dem im vorherigen Abschnitt sehr ähnlich. Auch hier ist nur für kleinere Werte ein schwacher Anstieg der Prognose mit den tatsächlichen Werten zu erkennen. Rasterquadrate mit keinen Wildunfällen, die fast die Hälfte der Beobachtungen ausmachen, weichen etwas von diesem Trend ab, weil für diese die absoluten Anzahlen von Wildunfällen und die relativen Anzahlen pro Straßenmeter gleich sind, sodass hier hinsichtlich der konstruierten Vergleichsgröße nicht zwischen weniger und stärker risikobehafteten Gebieten unterschieden wird. Im oberen Bereich ist kein Anstieg der Prognosewerte zu erkennen, sodass offenbar auch mit diesem Modell stark gefährdete Gebiete nicht detektiert werden können. Allerdings ist die Verzerrung, wie in der rechten Grafik in Abbildung 4.24 zu sehen, auch hier zu vernachlässigen. Der Wert des Korrelationskoeffizient nach Spearman betrug 0,257.

In Übereinstimmung mit dem vorigen Abschnitt wurden auch wieder alle Submodelle unter Auslassung je einer Kovariable berechnet und der Bruch der erklärten Devianzen dieser und der des vollen Modells als Maß für den Erklärwert der Kovariablen im vollen Modell verwendet, siehe Abbildung 4.25. Auch hier stellte `verbiss` die Kovariable mit dem geringsten Erklärwert dar. 98,9% der mit dem vollen Modell erklärten Abweichungen können auch mit dem Submodell ohne diese Kovariable erklärt werden. Aber auch `laengen`, `kurvigkeit` und `helligkeit` scheinen für sich genommen, global eine geringe Rolle im vollen Modell zu spielen. Die geringere Bedeutung von `helligkeit` scheint damit zusammenhängen, dass nun, bei der Modellierung von Unfällen pro Straßenmeter, `landnutzung` eine entscheidenen Beitrag zur Erklärung der Heterogenität liefert. `strasse` ist auch hier am bedeutensten. Dass die Koordinaten einen etwas kleineren

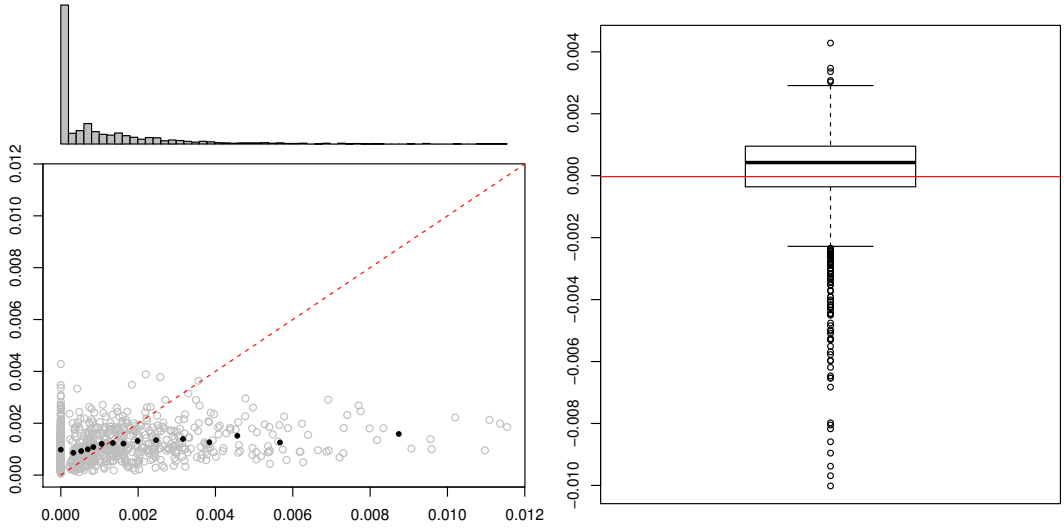

Abbildung 4.24.: links: prognostizierte gegen wahren Anzahlen, rechts: Differenz zwischen prognostizierten und wahren Anzahlen - rote Linie: Mittelwert

Anteil an der Erklärung der Devianz haben, kann man als Zeichen dafür werten, dass die Kovariablen hier eine größere Bedeutung als im Modell zur Prognose der absoluten Anzahlen haben. Bei der Interpretation ist zu beachten, dass eine Variable, die global wenig zur Erklärung der Devianz im vollen Modell beiträgt, an einzelnen Stellen dennoch von großer Bedeutung sein kann. Ein zweiter Punkt ist, dass man die auf diese Art gemessene Bedeutung von Kovariablen nur im Bezug auf das volle Modell interpretieren kann. Z.B. könnte möglicherweise die geringe Bedeutung von `helligkeit`, wie bereits erwähnt, zum Teil dadurch erklärt werden, dass hier `landnutzung` eine entscheidende Rolle spielte und `helligkeit` und `landnutzung` sich in ihrem Einfluss auf die Intensität überschneiden.

4.2. Modellierung durch logistisches Regressionsmodell

Die Vorgehensweise aus dem vorherigen Abschnitt, die Dummypunkte nur auf den Straßen zu wählen und die Kachelauflösung so fein zu wählen, dass nur Bereiche in direkter

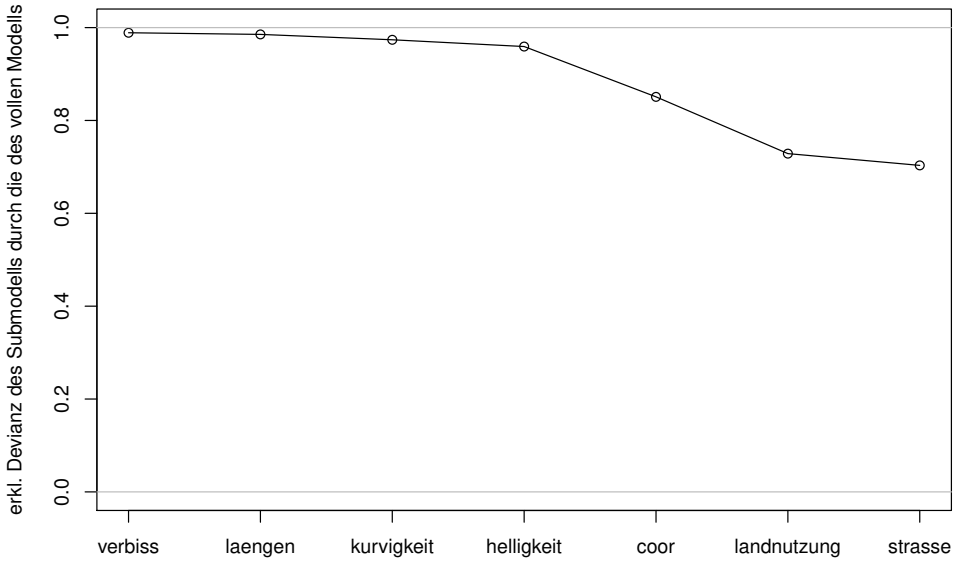

Abbildung 4.25.: Erklärwerte bei Auslassung der jeweiligen Kovariablen

Umgebung der Straßen in die Schätzung eingehen, stellte lediglich ein Bedarfsmittel dar, die Intensität der Wilfunfälle trotz der Tatsache, dass diese sich nur auf Straßen ereignen können, im Rahmen räumlicher Poissonprozesse zu modellieren. Ein direkterer Ansatz wäre, von vornherein die Anzahl an Wildunfällen pro Straßenmeter zu modellieren, was zu einem Poissonprozess auf einem Netzwerk linearer Linien führen würde. Die Theorie hierzu ist allerdings noch sehr jung, sodass in `spatstat` zwar eine Funktion `lppm()` existiert, mit der dies prinzipiell möglich ist, die Quellen hierzu allerdings noch nicht veröffentlicht wurden, bzw. unzugänglich sind. Zudem ist die Implementation nach Aussage eines Autors nur temporär und eignet sich nicht für sehr große Datensätze. Wenn man das Straßennetz ausreichend genau approximieren möchte, müsste man sehr viele lineare Abschnitte konstruieren, sodass die Analyse rechnerisch wohl sehr aufwendig würde. Aus diesen Gründen wurde dieses Vorgehen nicht weiter verfolgt.

Zur Bildung der Information „Anzahl an Wildunfällen pro Straßenmeter" auf einem Straßenabschnitt benötigt man die Anzahl der Unfälle auf diesem und seine Länge. Die-

se Information ist gleichbedeutend damit, für jeden Punkt auf dem Straßenabschnitt zu wissen, ob an diesem ein Unfall stattgefunden hat, oder nicht. Für diese Betrachtungsweise ist es auch nicht notwendig, das Straßennetz in Abschnitte zu unterteilen. Das ist allerdings ein theoretisches Konzept, da die Anzahl der möglichen Punkte auf den Straßen unbegrenzt ist. Für praktische Zwecke reicht es jedoch aus, eine begrenzte Teilmenge dieser Punkte zu betrachten. Die im vorherigen Abschnitt verwendeten Dummypunkte waren Straßenpunkte, die so konstruiert wurden, dass sie möglichst äquidistant zueinander sind. Diese Punkte wurden zusammen mit den Lokationen der Wildunfälle als Zielvariable in einer logistischen Regression verwendet, mit den Ereignissen „kein Unfall" bzw. „Unfall". Die Interpretierbarkeit des Fits aus diesem Modell, die Wahrscheinlichkeit, dass ein Punkt die Ausprägung „Unfall" hat, bzw. etwas salopp formuliert, die erwartete relative Häufigkeit von Punkten mit der Ausprägung „Unfall" ist inhaltlich nicht befriedigend. Multipliziert man die gefitteten Werte jedoch analog zum vorherigen Abschnitt mit der Anzahl der Punkte insgesamt und teilt durch die Straßenlänge insgesamt, erhält man wieder die erwartete Anzahl an Wildunfällen pro Straßenmeter. Für größere Datensätze nähern sich die Parameterschätzer zu den Kovariablen aus diesem Modell und die aus dem räumlichen Poissonprozess mit Dummypunkten auf den Straßen immer weiter an, wenn in ersterem die Anzahl an Punkten gegen unendlich geht und in letzterem die Kachelauflösung gegen Null und die Anzahl an Dummypunkten gegen unendlich geht . Dies wird in Anhang A nachvollzogen, wobei sich der Nachweis in Teilen auf Warton und Shepherdy (2009) bezieht. Die Annäherung ist auch gegeben, wenn die Logarithmusfunktion als Linkfunktion verwendet wird.

Für dieses Modell wurden dieselben Kovariablen wie im vorherigen Abschnitt verwendet, auch um einen direkten Vergleich der Koeffizientenschätzer zwischen den beiden Modellen zu ermöglichen. Es wurden auch einmal probehalber alle Straßenpunkte, die zur Berechnung der Straßenlängen verwendet wurden, eingeschlossen, sodass mit gut 2 Millionen über 2.5 mal so viele Straßenpunkte verwendet wurden. Damit sollte überprüft werden, ob die Anzahl der Straßenpunkte bereits ausreicht, oder ob sich die Koeffizientenschätzer noch stark ändern. Die Schätzer sowohl der kategorialen Kovarialben

als auch der Splinefunktionen änderten sich zwischen den unterschiedlichen Anzahlen an Straßenpunkten fast ausschließlich nur minimal. Die bedeutendste Änderung in der höheren Auflösung war, dass bei `laengen` die Funktion am linken Rand mit knapp über 6 im Vergleich zu gut 5,5 etwas höher war.

Da in diesem Modell kein Skalenparameter geschätzt werden musste, wurde gemäß der Standardeinstellung von `gam()` als Gütemaß bei der Optimierung der Glättungsparameter der Splinefunktionen die Verallgemeinerung von Mallow's C_p auf den generalisierten Fall - $C_{p_{\text{gen}}}$ - verwendet (siehe Abschnitt 2.5). Dadurch ergaben sich zum Teil völlig andere Glättungsparameter als in Abschnitt 4.1.2, sodass sich auch die geschätzten Einflüsse der semiparametrisch aufgenommenen Kovariablen - für `verbissfichte` und `verbissbuche` sehr stark - veränderten. Bis auf bei `laengen` waren die Glättungsparameter in allen Fällen größer, sodass die Funktionen glatter waren. Dies würde zum einen den direkten Vergleich mit dem Ergebnis aus dem Poissonprozess erschweren, zum anderen ist die Optimierung der Glättungsparameter bei `verbiss` aufgrund des geringen Einflusses ohnehin mit einer großen Varianz verbunden. Deshalb wurden die optimierten Glättungsparameter aus dem Modell aus vorherigem Abschnitt verwendet. Um dennoch sicherzugehen, dass dies nicht zu einer Verschlechterung des Fits führt, wurden die beiden Modelle anhand eines Analogons zu den Lurking Variable Plots (siehe Abschnitt 4.1.1) verglichen. Anstatt wie bei den räumlichen Poissonprozessen, die Differenzen der Anzahlen an Wildunfällen und der integrierten Intensitätsfunktion für Stützpunkte jeweils innerhalb des Bereichs mit Kovariablenwert $\leq z$ zu betrachten, wurden die Differenzen ersterer und der Summe der Wahrscheinlichkeiten für die Unfallokationen und Straßenpunkte in den Bereichen gebildet. Diese müssten für korrekt spezifizierte Modelle ebenfalls im Schnitt Null sein. Dabei gab es nur für `verbissfichte`, `verbissbuche` und `helligkeit` erkennbare Unterschiede, die darauf hinwiesen, dass das Modell mit Glättungsparametern aus Abschnitt 4.1.2 den Daten etwas mehr entspricht. Die Analogie des Konzepts dieses Modells mit dem aus dem vorherigen Abschnitt bestätigend, unterschieden sich die jeweiligen LVPs zwischen den beiden Modellen bei gleichen Glättungsparametern nicht nennenswert.

Der Aufruf der gam-Funktion lautet wie folgt:

```
modellbinaersmoothroads <-  gam(unfall ~ strasse + landnutzung +
    s(verbissfichte, bs="ps") + s(verbissbuche, bs="ps") +
    s(verbisseiche, bs="ps") + s(helligkeit, bs="ps", k=14) +
    s(kurvigkeit, bs="ps") + s(laengen, bs="ps") + te(xcoor, ycoor,
    bs="ps"), data=covari, family="binomial", knots=
    list(helligkeit = knotenhelligkeit), sp=smoothparamroads)
```

Neu sind hier die binäre Zielvariable `unfall` und die Angaben `family="binomial"` sowie `sp=smoothparamroads`. Mit Letzterer werden die Glättungsparameter aus dem Poissonprozess mit Dummypunkten auf den Straßen - `smoothparamroads` - übergeben und `family="binomial"` besagt, dass als Exponentialfamilie die Bernoulliverteilung verwendet wird, wobei dabei der logit-Link aus der logistischen Regression bereits die Standardeinstellung ist.

Die Anwendung der `summary`-Funktion auf `modellbinaersmoothroads` liefert:

```
> summary(modellbinaersmoothroads)
Loading required package: splines

Family: binomial
Link function: logit

Formula:
unfall ~ strasse + landnutzung + s(verbissfichte, bs = "ps") +
    s(verbissbuche, bs = "ps") + s(verbisseiche, bs = "ps") +
    s(helligkeit, bs = "ps", k = 14) + s(kurvigkeit, bs = "ps") +
    s(laengen, bs = "ps") + te(xcoor, ycoor, bs = "ps")
```

```
Parametric coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.78044    0.05104 -113.25   <2e-16 ***
strasseprimary       2.01036    0.03523   57.07   <2e-16 ***
strassesecondary     1.78049    0.03454   51.55   <2e-16 ***
strassetertiary      1.17532    0.03482   33.76   <2e-16 ***
landnutzunglandwirt  1.25308    0.02373   52.80   <2e-16 ***
landnutzungwald      1.76092    0.02518   69.95   <2e-16 ***
landnutzungsonstige  1.03981    0.07399   14.05   <2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Approximate significance of smooth terms:
                    edf Ref.df  Chi.sq  p-value
s(verbissfichte)  8.738  8.925   84.47 1.91e-14 ***
s(verbissbuche)   8.932  8.996   45.24 8.31e-07 ***
s(verbisseiche)   6.512  7.295   56.52 1.04e-09 ***
s(helligkeit)    12.753 12.943  748.65  < 2e-16 ***
s(kurvigkeit)     7.038  7.529  646.43  < 2e-16 ***
s(laengen)        6.696  6.968  328.52  < 2e-16 ***
te(xcoor,ycoor)  23.981 24.000 2877.17  < 2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

R-sq.(adj) =  0.0312   Deviance explained = 7.43%
UBRE score = -0.60566  Scale est. = 1         n = 831439
```

Die geschätzten Einflüsse sind hier ebenfalls alle signifikant von Null verschieden und die erklärte Devianz wieder etwas größer.

Es sei darauf hingewiesen, dass man hier formal den multiplikativen Einfluss einer Kovariable auf die Wahrscheinlichkeit, dass ein Punkt ein Unfallpunkt ist und damit auf die dazu (approximativ) proportionale Intensität nicht berechnen könnte, da im Gegensatz zu dem räumlichen Poissonprozess mit der log-Linkfunktion die logit-Linkfunktion verwendet wird. Allerdings sind die prognostizierten Wahrscheinlichkeiten bei der logistischen Regression sehr klein und für Werte Nahe bei Null haben die link-Funktionen fast die gleichen Werte.

Die kategorialen Kovariablen weisen von der Rangfolge der Kategorien und deren relativen Abstände zueinander her dasselbe Bild auf (Abbildung 4.26). Insbesondere für `strasse` sind die absoluten Abstände jedoch größer. Zum Beispiel ist die Gefahr von Wildunfällen hier auf Bundesstraßen ca. 7,5 mal größer als auf Autobahnen, während sie nach dem Modell aus vorherigem Abschnitt nur gut 4,5 mal so hoch war. Hinsichtlich `landnutzung` sind die Unterschiede zwischen den Modellen weit weniger ausgeprägt. Als Beispiel sei genannt, dass sich die geschätzte Gefahr von Wildunfällen auf Straßen im Wald im Vergleich zu solchen in bebautem Gelände von etwas unter 5 auf knapp 6 mal so viele erhöhte. Diese Ergebnisse sind in Übereinstimmung mit dem, was in Abschnitt 4.1.2 beobachtet wurde: Dort drifteten die Parameterschätzer zu `strasse` und `landnutzung` mit feiner werdender Kachelauflösung auseinander, wobei der Trend bei `landnutzung` allerdings in der höchsten Auflösung schon weitestgehend abgeflacht war. Dass hier die geschätzten Koeffizienten größer sind, kann man als Anzeichen dafür werten, dass die Approximation besser greift, dass also die Ergebnisse dieses Modells bereits näher an denen bei expliziter Modellierung der Anzahlen an Unfällen pro Straßenmeter sind.

Der einzige auffallende Unterschied bei den semiparametrisch geschätzten Einflüssen (Abb. 4.27) ist, dass die Funktion für `laengen` im Bereich sehr geringer Straßendichte noch weiter nach oben geht. Dieser Anstieg ist aber wie im vorherigen Abschnitt bereits im Rahmen der Lurking Variable Plots diskutiert, höchstwahrscheinlich nicht mit dem wahren Sachverhalt vereinbar. Vermutlich fallen durch die Konzentrierung auf die Straßenpunkte anstatt nur auf den Bereich um die Straßen, künstliche Übergewichte von

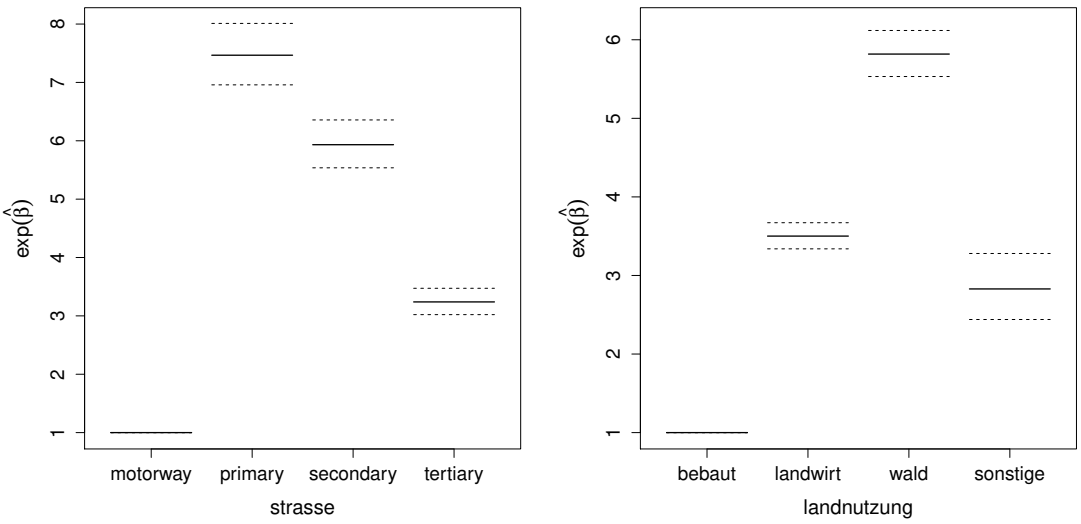

Abbildung 4.26.: exponierte Koeffizientenschätzer der kategorialen Kovariablen im logistischen Modell

Unfallpunkten noch stärker ins Gewicht.

Der Einfluss der Lokation, dargestellt in Abbildung 4.28 ist praktisch identisch zu dem aus Abschnitt 4.1.2.

Auch der Imageplot zum Modellfit (Abb. 4.29) ist praktisch identisch mit dem zum Modellfit des Modells mit Dummypunkten auf den Straßen. Direktes Vergleichen zeigte, dass die gefitteten Werte hier im Schnitt etwa 1,07 mal so groß waren. Die Korrelation der gefitteten Werte zwischen den beiden Modellen betrug ca. 0,992. In Abbildung 4.30 sind sie gegeneinander geplottet. In nur drei Fällen (rot gekennzeichnet) waren größere Abweichungen zwischen den Prädiktionen aus den beiden Modellen zu beobachten.

Abbildung 4.31 zeigt wieder die Brüche der erklärten Devianzen der Submodelle unter Auslassung je einer Kovariable und der des vollen Modells. Die Unterschiede im Vergleich zum vorherigen Abschnitt sind minimal. Bei genaueren Hinsehen, bzw. bei direktem Vergleich der Anteile, ist ersichtlich, dass `laengen`, `kurvigkeit` und `helligkeit`

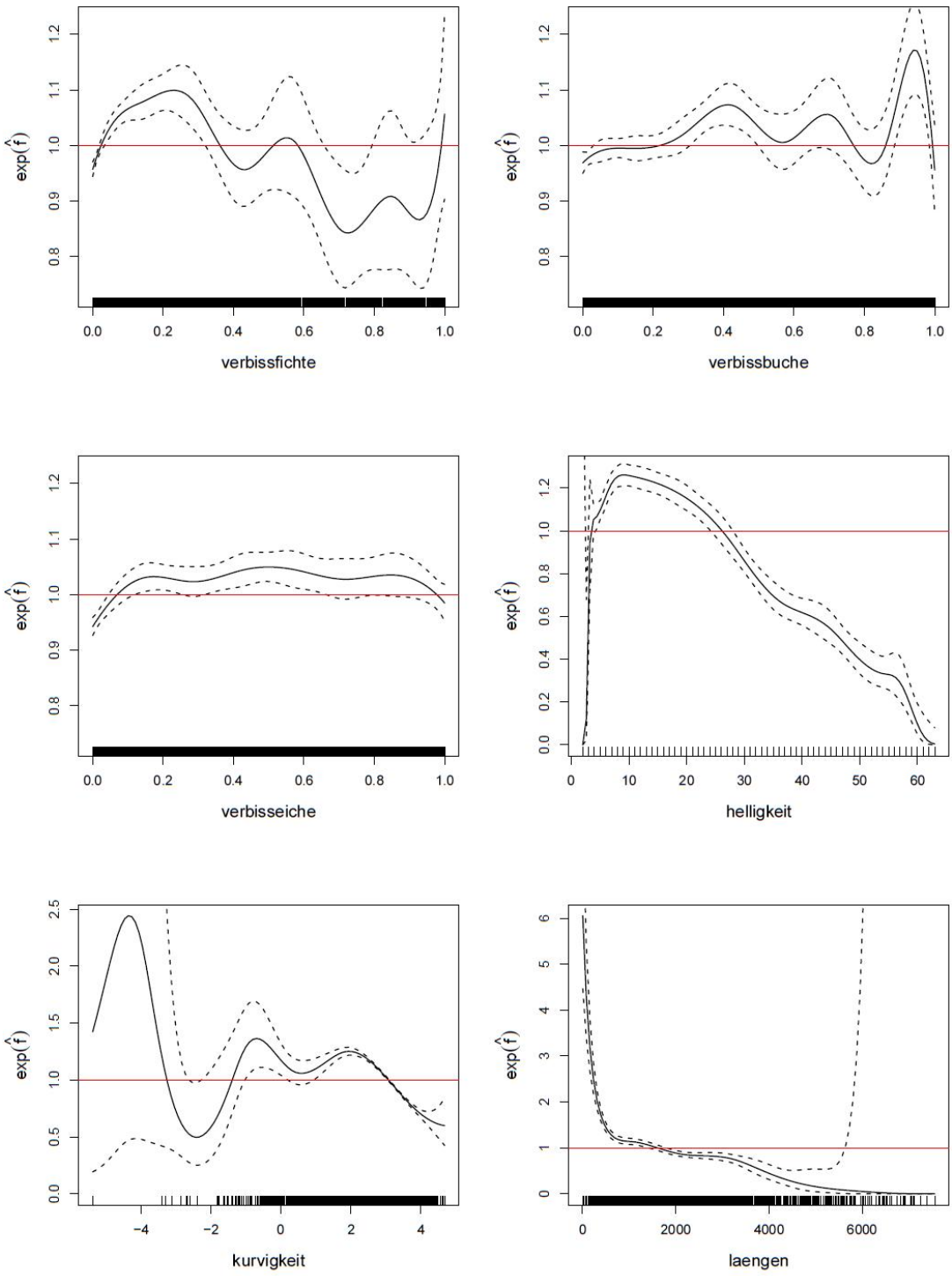

Abbildung 4.27.: exponierte Spline-Funktionen im logistischen Modell

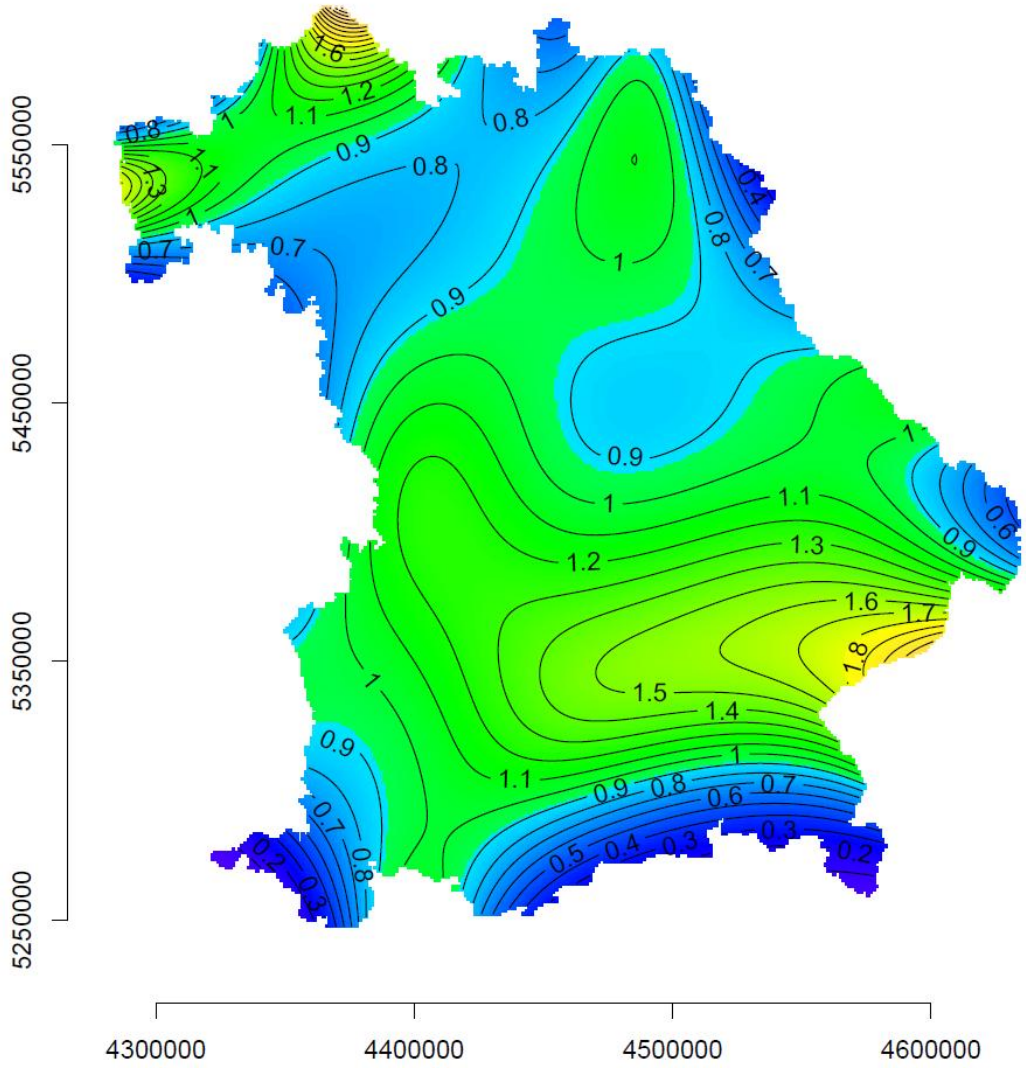

Abbildung 4.28.: Einfluss der Lokation im logistischen Modell

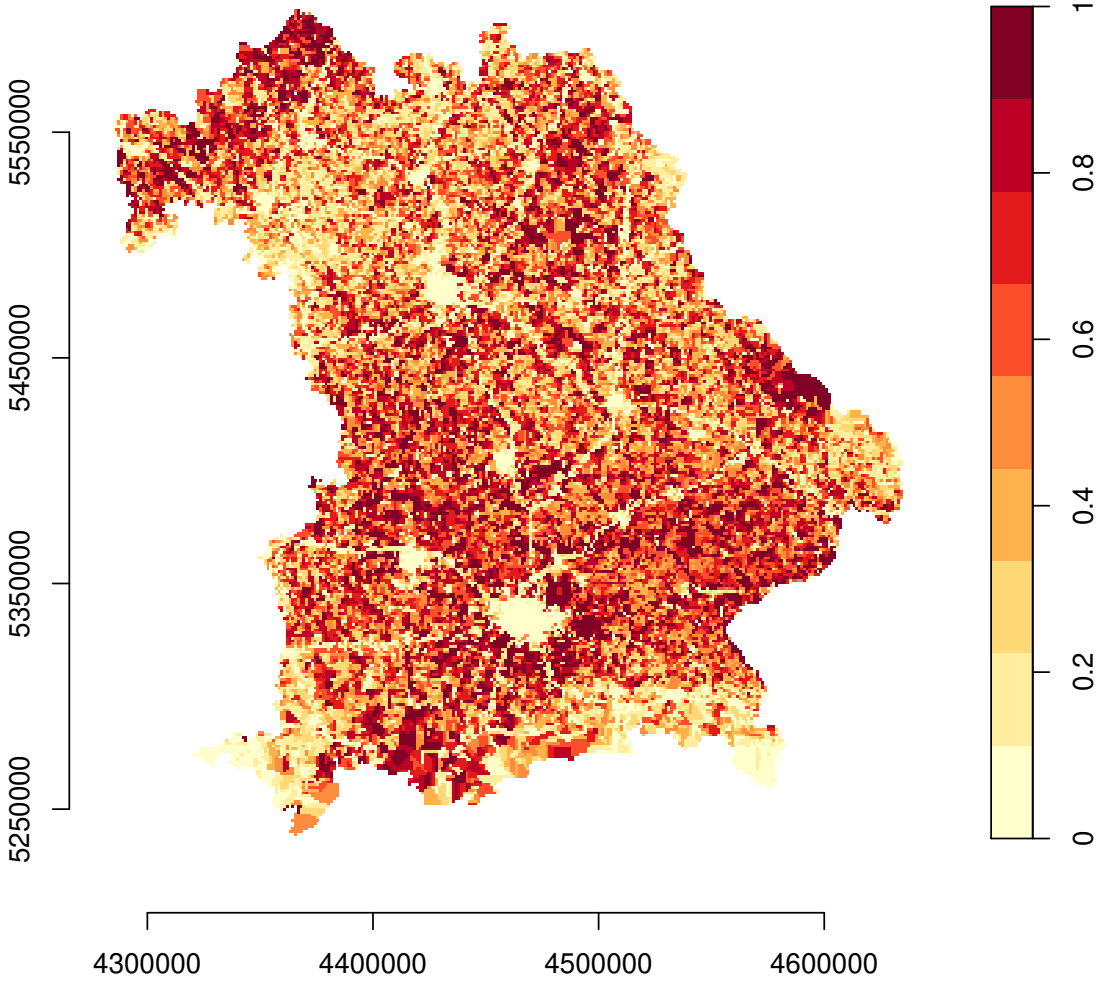

Abbildung 4.29.: Quantile des Fits des logistischen Modells

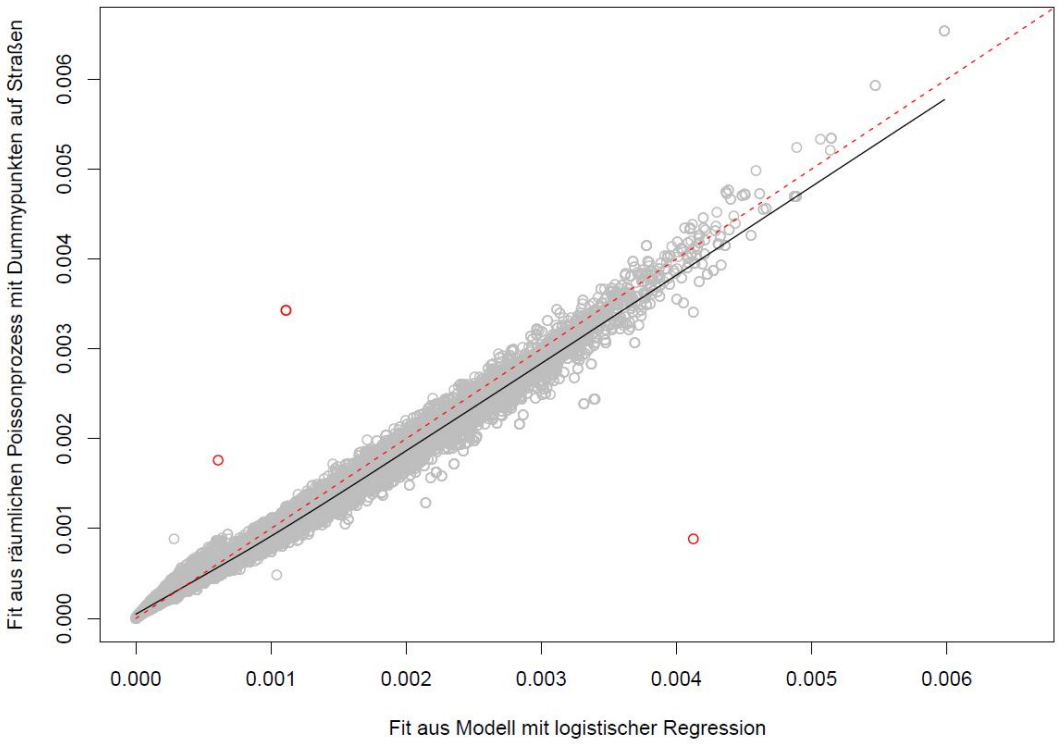

Abbildung 4.30.: Vergleich des Fits zwischen Poissonprozess mit Dummypunkten auf den Straßen und logistischem Modell

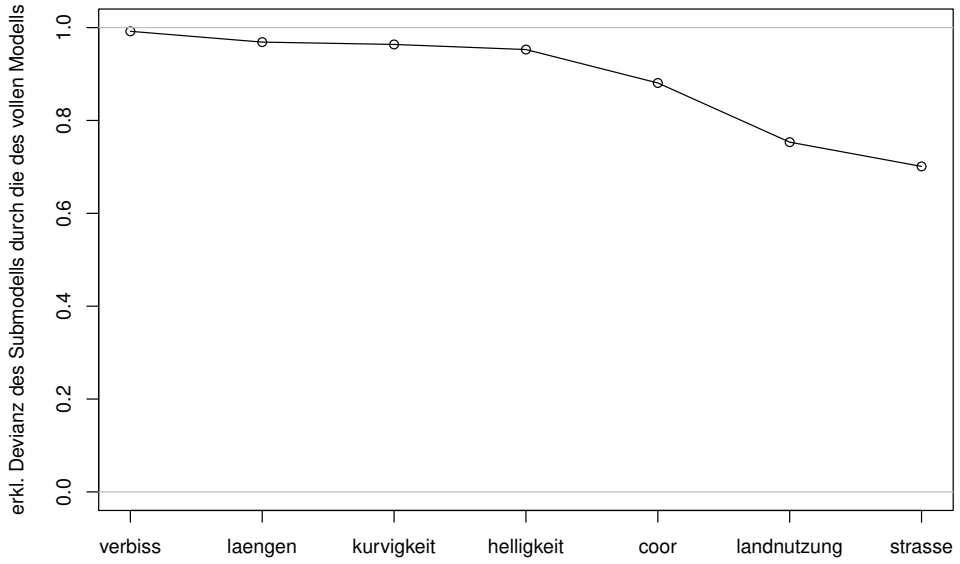

Abbildung 4.31.: Erklärwerte bei Auslassung der jeweiligen Kovariablen

etwas mehr an der Erklärung der Devianz teilhaben, allerdings sind die Unterschiede, insbesondere bei `helligkeit` sehr gering. Dagegen scheinen `coor` und `landnutzung` etwas ein Einfluss verloren zu haben. Die Rangfolge hat sich aber, wie man auch aufgrund der approximativen Analogie erwarten würde, nicht geändert.

4.3. Modellierung mit negativer Binomialverteilung

Die beiden betrachteten Möglichkeiten, die Unfälle auf den Straßen zu modellieren, der Poissonprozess mit Dummypunkten auf den Straßen und die Modellierung der Straßenpunkte und Unfalllokationen als Zielgröße in einem logistischen Regressionsmodell, stellten insbesondere bei der logistischen Regression, in der keine künstliche Konzentrierung um die Straßen notwendig war, eine gute Approximation an die tatsächliche Modellierung der Unfälle pro Straßenmeter dar. Die Modellierung letzterer Größe über einen räumlichen Poissonprozess war allerdings aus den in Abschnitt 4.2 genannten Gründen

nicht möglich. Ein intuitiver Weg, wäre nun die Anzahlen pro Straßenmeter auf einem besonders feinen Gitter direkt als Zielgröße in einem generalisierten additiven Modell zu verwenden. Dies wäre allerdings problematisch, da sich auf einem genügend feinen Raster gleichzeitig sehr viele Nullzählungen ergäben und daher eine Berechnung der Anzahl der Wildunfälle pro Straßenmeter zu einer Anhäufung bei Null führen würde. Dies hätte zum einen zur Folge, dass die entstehende Zielgröße keiner gängigen Verteilung mehr folgen würde und zum anderen käme es zu einem großen Informationsverlust, da bei den Nullzählungen nicht unterschieden würde zwischen Rasterquadraten mit geringer und großer Straßendichte. Deshalb wurden die Anzahlen in den Rasterquadraten als Zielgröße verwendet und die logarithmierten Straßenlängen in den Prädiktor mit aufgenommen. Durch Verwendung der Logarithmusfunktion als Linkfunktion wirken die Kovariablen auf diese Art über die Exponentialfunktion multiplikativ auf die erwartete Anzahl an Unfällen pro Straßenmeter ein, wenn man die logarithmierte Straßenlänge auf die andere Seite der Gleichung zur Erwartungswertstruktur bringt. Aufgrund der vielen Nullwerte, weniger größeren und einiger sehr großer Zählungen kam es zu einer starken Überdispersion. Es erscheint inhaltlich nicht sinnvoll, dass Rasterquadrate mit Null-Zählungen hinsichtlich der Zielgröße (in ihrer Interaktion mit den Kovariablen) ein anderes Verhalten aufweisen. Deshalb bot sich auch nicht die Verwendung eines Zero-Inflated- oder eines Hurdle-Modells an. Bei Ersterem wird angenommen, dass die Dichte der Zielvariable eine Mischung aus einer Punktmasse bei Null und einer Zählverteilung, etwa der Poissonverteilung ist, bei Letzterem, dass es zwei Prozesse gibt, einen der die Nullen und einen der die positiven Werte erzeugt, sodass die Dichte aus zwei Komponenten besteht, eine die die Wahrscheinlichkeit für Nullwerte angibt und eine die die Wahrscheinlichkeiten für größere Werte durch eine bei 1 linkstrunkierte Zähldichte bestimmt, für Details siehe Zeileis et al. (2008). Um Überdispersion in Zähldaten mit einer Exponentialfamilie im Rahmen generalisierter additiver Modelle zu berücksichtigen bietet sich die negative Binomialverteilung an. Diese hat im Vergleich zur Poissonverteilung einen zusätzlichen Parameter α, welcher gewährleistet, dass die Varianz größer als der Erwartungswert ist (Hilbe, 2011). Dass die negative Binomialverteilung für festes

α eine einparametrische Exponentialfamilie ist, lässt sich beinahe unmittelbar an ihrer Dichtefunktion (vgl. (4.7)) ablesen. Die Kovariablen `verbissfichte`, `verbissbuche`, `verbisseiche`, `kurvigkeit` und `laengen` sind alle auf einem Raster der Auflösung 1225 m verfügbar. Mit der selben Argumentation wie bei der Abschätzung der Prognosegüte in den vorherigen Modellen wurden die Anzahlen an Unfällen pro Straßenmeter auch auf diesem Raster berechnet. Außerdem ist der mit dieser minimalen Aggregation verbundene Informationsverlust ohnehin sehr gering. Vielmehr ergibt sich der praktische Vorteil, dass die Fallzahl in der Regression weitaus geringer ist - ca. 30.000 im Vergleich zu den vorherigen Modellen mit gut 800.000. Dadurch wird der rechnerische Aufwand weitaus geringer, sodass auch aufwendige Methoden zur besseren Berücksichtigung der räumlichen Korrelation als die bloße Aufnahme der Lokation als Tensorprodukt-Spline in den Prädiktor denkbar sind. Hierauf wird weiter unten noch einmal näher eingegangen.

Eine in der Regression gängige Parametrisierung der negativen Binomialverteilung ist gegeben durch:

$$f(y_i|\mu_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{4.7}$$

Bei einer ebenfalls üblichen Parametrisierung, die auch `mgcv` verwendet, wird das Inverse des Heterogenitätsparameters α, also $\alpha^* = \frac{1}{\alpha}$ verwendet. Während obige Form für eine Interpretation wenig zugänglich ist, erschließt sich das Konzept der Modellierung der Überdispersion wenn man bedenkt, dass diese Verteilung eine stetige Mischung aus Poissonverteilungen ist. Genauer gesagt folgt y_i einer negativen Binomialverteilung in der obigen Notation, wenn $y_i \sim \text{Po}(\mu_i u)$ und $u \sim \gamma(\frac{1}{\alpha}, \frac{1}{\alpha})$, wobei $\gamma(\cdot)$ die Gammaverteilung bezeichnet. Da damit $\mathbb{E}(u) = 1$ und $\mathbb{V}(u) = \alpha$ gilt, wird deutlich, dass je größer der sogenannte Heterogenitätsparameter α, desto heterogener die Mischung der Poissonverteilungen, bzw. desto größer die Varianz im Vergleich zum Mittelwert. Es gilt:

$$\mathbb{E}(y_i|\mu_i) = \mu_i \quad \text{und} \quad \mathbb{V}(y_i|\mu_i) = \mu_i + \alpha\mu_i^2. \tag{4.8}$$

Eine Besonderheit bei der Optimierung stellt dar, dass ein zusätzliche Parameter - α - geschätzt werden muss. In `mgcv` geschieht dies bei „performance iteration" innerhalb und

bei „outer iteration" außerhalb der Schleife der P-IRLS-Iterationen. Da bei Letzterem das Modell sehr oft geschätzt werden muss, um einen α-Wert zu erhalten, der genügend genau am Optimum liegt, wurde hier „performance iteration" verwendet. Bei dieser Option geschieht die Optimierung in `mgcv` auf folgende Weise: Der Benutzer übergibt zwei Grenzen, innerhalb derer er den optimalen α^*-Wert vermutet. Dann wird nach jeder P-IRLS-Iteration, derjenige α^*-Wert bestimmt, für den, gegeben die Parameterschätzer der geschätzte Skalenparameter $\widehat{\phi}$, siehe Abschnitt 2.4, (approximativ) den Wert 1 hat. In der nächsten Iteration wird dann dieser α^*-Wert verwendet. Die Bestimmung des α^*-Wertes für den $\widehat{\phi} \overset{(\approx)}{=} 1$ geschieht dabei wie folgt. Zunächst wird ausgehend von dem aktuellen α^*-Wert dieser solange halbiert bzw. verdoppelt, bis der damit verbundene $\widehat{\phi}$-Wert kleiner bzw. größer 1 ist oder die vom Benutzer übergebenen Grenzen überschritten werden. In letzterem Fall wird die überschrittene Grenze als optimaler α^*-Wert verwendet. Sonst wird mit Hilfe von Bisektion (siehe z.B. Lange (2010)) innerhalb des so gewonnen Intervalls nach dem optimalen Wert gesucht bis der resultierende Skalenparameter nicht weiter als etwa $1,5 \times 10^{-8}$ von 1 entfernt liegt.

Die betrachteten Kovariablen in dem vorliegenden Abschnitt sind dieselben wie in den vorherigen beiden. Natürlich wurden in der Regression nur die Rasterquadrate verwendet, durch die Straßen liefen. Die Werte zu `helligkeit` liegen wie beschrieben auf einem Raster der Auflösung ca. 547 m vor. Da sich diese nur wenig von ihren Nachbarn unterscheiden war keine Mittelung der jeweils in ein Rasterquadrat der mit 1225 m gröberen Auflösung fallenden Pixelwerte nötig. Es wurden also die Originalwerte verwendet, die sich jeweils an den Mittelpunkten der Rasterquadrate der gröberen Auflösung befanden. Für `strasse` wurde jedem Pixel der Straßentyp zugeordnet, von dem sich am meisten Punkte aus dem zur Berechnung der Straßenlängen erstellten feineren Straßennetz innerhalb dem jeweiligen Pixel befanden. Dies lässt sich damit rechtfertigen, dass die Straßenpunkte für die unterschiedlichen Straßentypen in etwa diesselben Abstände haben, sodass die Anzahl der Punkte ein gutes Maß für die Straßenlänge darstellt. Für `landnutzung` wäre ein solches Vorgehen nicht angebracht gewesen, da sich die Ausprägungen über den Raum sehr schnell ändern und deshalb Ungenauigkeiten entstanden

wären, hätte man einem Pixel einheitlich eine Ausprägung zugeordnet. Deshalb war ein Vorgehen nötig, bei dem die flächenmäßigen Anteile der unterschiedlichen Kategorien in den einzelnen Pixeln Berücksichtigung finden. Da die Anteile sich zu eins aufsummieren und daher notwendigerweise negativ korreliert sind, musste eine Kategorie in der Modellierung außen vorbleiben. Dafür wurden landwirtschaftliche Gebiete gewählt, da diese flächenmäßig den größten Raum einnehmen und ihre Anteile damit mit denen der anderen Kategorien am stärksten korrelieren. Für die restlichen drei Kategorien wurde wie folgt vorgegangen: Es wurde jeweils eine Indikatorvariable eingeführt, die angibt, ob die jeweilige Kategorie überhaupt in den einzelnen Pixeln vorkommt und positive Anteile jeweils als penalisierter B-Spline modelliert. Für die Kategorie `sonstige` war allerdings weder der Koeffizientenschätzer zu der Indikatorvariable signifikant noch der über den penalisierten B-Spline geschätzte Einfluss. Um diese Kategorie aber dennoch im Modell zu behalten, aber auf zu große Ungenauigkeiten in der Schätzung zu verzichten, wurde für diese Kategorie lediglich die Indikatorvariable im Modell gelassen. Natürlich hätte man auch die Anteile inklusive Nullwerte als Splines modellieren können, hätte dabei aber auf Information verzichtet, da die Nullwerte in allen Fällen einen großen Anteil einnehmen, weil die Kategorien nicht in allen Gebieten vorkommen. Außerdem könnten die Nullwerte in ihren Einfluss auf die Wildunfallintensität ein abweichendes Verhalten aufweisen. Ein zusätzlicher Vorteil, die Anteile in die Modellierung aufzunehmen, ist eine Verfeinerung der Information bzw. die implizite Aufnahme sonst unbeachteter Konstrukte. Z.B. ging in den vorherigen Abschnitten nur die Präsenz oder Absenz von Wald ein. Allerdings stellt die Länge an Waldrändern einen wichtigen Faktor für geeignete Wildhabitate dar, vergleiche z.B. Hothorn et al. (in Vorb.). Durch die Aufnahme der Anteile von Wald, findet der Effekt „Waldrand" erst Berücksichtigung, bei nur minimaler Aggregation. Es wäre auch möglich, die Länge des Waldrand direkt als Einflußgröße aufzunehmen. Dazu müsste man allerdings den Einfluss einer weiteren Kovariable schätzen, was zu unnötigen Ungenauigkeiten führen würde. Ähnlich könnte man auch für bebautes Gebiet argumentieren.

Wie auch schon im vorherigen Abschnitt, war hier kein Skalenparameter zu schätzen,

weshalb bei der Bestimmung der optimalen Glättungsparameter wieder das Analogon zu Mallow's C_p verwendet wurde. Es resultierten in allen Fällen sehr glatte Funktionen, z.B. war der geschätzte Einfluss bei `verbissbuche` praktisch linear. Um wieder eine bessere Vergleichbarkeit zu ermöglichen, wurden die Glättungsparameter aus dem Modell aus vorherigem Abschnitt verwendet, sofern möglich. Diejenigen zu den Splinefunktionen von `landnutzung` wurden mit dem Analogon zu Mallow's C_p optimiert. Hier war zur Beurteilung der Auswirkung der Festsetzung der Glättungsparameters keine wie in Abschnitt 4.2 geschehene Analyse mit Hilfe eines Analogums zu Lurking Variable Plots möglich, da sich die geschätzten Residuen in einem Negativ-Binomial-Modell der obigen Form nicht zu Null aufsummieren (Cameron und Windmeijer, 1996). Deshalb wurden die Modelle hinsichtlich ihrer Prognosegüte verglichen, zu deren Abschätzung wie bei dem Poissonprozess mit Dummypunkten auf den Straßen vorgegangen wurde. Es wurden jeweils wieder die Korrelationskoeffizienten nach Spearman berechnet. Diese unterschieden sich mit 0,2998 für das Modell mit bis auf die der neu hinzugekommmenen Splinefunktionen fixierten und 0,3018 für das Modell mit vollständig über das Analogon von Mallow's C_p bestimmten Glättungsparametern kaum, wobei man allerdings beachten muss, dass, wie in Abschnitt 4.1.1 bereits beschrieben, die wahre Prognosegüte, bei vollständiger Unabhängigkeit von Trainings- und Testdatensatz anders wäre und sich demnach auch die Prognosegüten etwas stärker zwischen den beiden Modellen unterscheiden könnten. Die Unterschiede sind allerdings vermutlich sehr gering, was auch durch die sehr hohe Korrelation zwischen den Prognosewerten von 0,998 bekräftigt wird. Die prognostizierten Werte verhielten sich ähnlich zu denen der räumlichen Poissonprozesse bzw. des logistischen Modells. Auch hier wurde Lokationen mit sehr hoher Wildunfalldichte nicht erkannt. Allerdings war der Anstieg zwischen den prognostizierten und den tatsächlichen Werten im unteren Bereich größer, was sich auch in dem höheren Korrelationskoeffizienten widerspiegelt. Dabei muss man allerdings beachten, dass hier durch die Modellierung der Anteile der Landnutzungskategorien eine detailliertere Information übergeben wurde und die Ausprägungen der Zielgröße auf dem selben Raster wie einige Kovariablen und die zu prognostizierenden Werte gegeben waren.

Das Modell wird mit folgendem Aufruf geschätzt:

```
modellnegbinsmoothroads <- gam(counts ~ strasse + landnutzbebautind +
    landnutzwaldind + landnutzsonstigeind + s(landnutzbebaut, bs="ps",
    by=landnutzbebautind) + s(landnutzwald, bs="ps", by=landnutzwaldind)
    + s(verbissfichte, bs="ps") + s(verbissbuche, bs="ps") +
    s(verbisseiche, bs="ps") + s(helligkeit, bs="ps", k=14) +
    s(kurvigkeit, bs="ps") + s(laengen, bs="ps") + te(xcoor, ycoor,
    bs="ps") + offset(loglaengen), data=daten, optimizer="perf", family=
    negbin(theta=c(0.2,1.1)), knots=list(helligkeit = knotenhelligkeit),
    sp=smoothparamnegbin)
```

`landnutzbebautind`, `landnutzwaldind` und `landnutzsonstigeind` sind die oben beschriebenen Indikatorvariablen zu `landnutzung`. In den Funktionen `s()` zur Aufsetzung der Splinebasen werden für die Landnutzungskategorien mit dem Argument `by` die jeweiligen Indikatorvariablen übergeben. Damit wird angegeben, dass die Glättung mit der jeweiligen Indikatorvariable multipliziert wird, was gerade der oben beschriebenen Modellierung entspricht. `optimizer="perf"` besagt, dass „performance iteration" verwendet werden soll und `family=negbin(theta=c(0.2,1.1))`, dass die `negbin`-Familie eingesetzt und dabei innerhalb der Grenzen 0,2 und 1,1 nach dem optimalen α^*-Wert gesucht werden soll. Letzteres Intervall hat sich als ausreichend groß für die Suche erwiesen.

Die `summary`-Funktion lieferte nun:

```
> summary(modellnegbinsmoothroads)

Family: Negative Binomial(0.754)
Link function: log
```

```
Formula:
counts ~ strasse + landnutzbebautind + landnutzwaldind + landnutzsonstigeind +
    s(landnutzbebaut, bs = "ps", by = landnutzbebautind) + s(landnutzwald,
    bs = "ps", by = landnutzwaldind) + s(verbissfichte, bs = "ps") +
    s(verbissbuche, bs = "ps") + s(verbisseiche, bs = "ps") +
    s(helligkeit, bs = "ps", k = 14) + s(kurvigkeit, bs = "ps") +
    s(laengen, bs = "ps") + te(xcoor, ycoor, bs = "ps") + offset(loglaengen)

Parametric coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         -8.155419   0.045578 -178.935   <2e-16 ***
strasseprimary       1.270156   0.045610   27.848   <2e-16 ***
strassesecondary     1.213566   0.042565   28.511   <2e-16 ***
strassetertiary      0.732641   0.042290   17.324   <2e-16 ***
landnutzbebautind   -0.896369   0.049063  -18.270   <2e-16 ***
landnutzwaldind      0.458053   0.025685   17.833   <2e-16 ***
landnutzsonstigeind -0.002078   0.037631   -0.055    0.956
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

Approximate significance of smooth terms:
                                       edf Ref.df       F  p-value
s(landnutzbebaut):landnutzbebautind  1.900  2.333 176.267  < 2e-16 ***
s(landnutzwald):landnutzwaldind      3.131  3.775  41.336  < 2e-16 ***
s(verbissfichte)                     8.991  9.000   3.771 9.20e-05 ***
s(verbissbuche)                      8.995  9.000   2.029   0.0323 *
s(verbisseiche)                      8.051  8.639   1.585   0.1167
s(helligkeit)                       12.993 13.000  25.654  < 2e-16 ***
s(kurvigkeit)                        7.995  8.033  30.501  < 2e-16 ***
```

```
s(laengen)                              6.835  7.073   7.147 1.24e-08 ***
te(xcoor,ycoor)                        23.999 24.000  45.419  < 2e-16 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1

R-sq.(adj) =   0.19   Deviance explained = 20.4%
UBRE score = 1.0029  Scale est. = 1         n = 31095
```

Der kleine α^*-Wert von 0,754 weist auf starke Überdispersion hin, wenn man bedenkt, dass das Negativ-Binomial-Modell für $\alpha^* \to \infty$ (bzw. $\alpha \to 0$) gegen das Poisson-Modell strebt, in welchem bekanntlich die Varianz gleich dem Erwartungswert ist. Die erklärte Devianz ist mit 20,4% weit größer als in den vorherigen Modellen. Die Werte kann man allerdings nicht vergleichen, da zwischen den Modellen Unterschiede bestehen. Neben der bereits erwähnten detaillierteren Information bei `landnutzung` muss man auch bedenken, dass die Kovariablenwerte immer auf einem Raster gegeben und damit über kleine Strecken konstant sind. Das bedeutet, dass sie sich nicht so flexibel ändern, wie die der sich bei (approximativ) tatsächlich punktueller Betrachtung der Intensität ergebende binäre Zielgröße aus vorherigen Abschnitten. In vorliegendem Modell ist die Zielgröße ebenfalls auf einem Raster gegeben, sodass diese Diskrepanz nicht auftritt. Gleichzeitig enthalten die Werte der Zielgröße aber durch die leichte Agregierung mehr Information. Der Einfluss von `verbisseiche` ist in diesem Modell nicht mehr signifikant. Auch der von `landnutzungsonstigeind` ist wie bereits erwähnt nicht signifikant. Es wurde versuchsweise auch einmal ein Modell mit $\alpha^* = 1$ angepasst. In diesem Fall ist die negative Binomialverteilung gleich der geometrischen Verteilung (Hilbe, 2011). Dabei war die erklärte Devianz mit 21,2% etwas größer und der AIC-Wert etwas kleiner. Die Formen der Spline-Funktionen änderten sich aber nur minimal. Der Einfluss von `verbisseiche` war dabei schwach signifikant.

Bei den Koeffizientenschätzern zu `strasse` (Abb. 4.32 links) hat sich die Rangfolge nicht verändert, allerdings sind die Abstände zwischen den Kategorien weit geringer. Das hängt

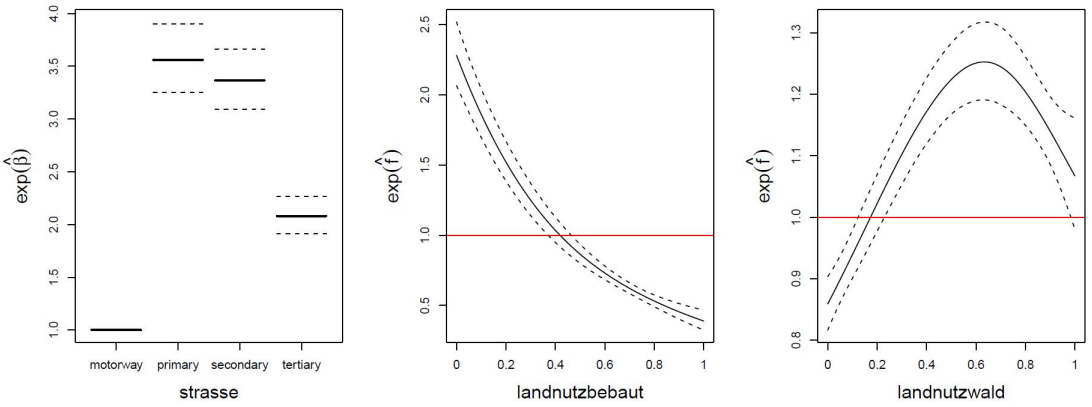

Abbildung 4.32.: exponierte Koeffizientenschätzer zu `strasse` und exponierte Splinefunktionen zu `landnutzung`

vermutlich damit zusammen, dass hier der zugeordnete Straßentyp nicht in allen Fällen dem tatsächlichen entsprach, weil die Kovariablen auf einem Raster angegeben wurden, sodass dem Ergebnis aus vorherigem Abschnitt weit mehr Vertrauen geschenkt werden kann. Die genauen Werte der exponierten Splinefunktionen zu `landnutzung` (Abb. 4.32 Mitte bzw. rechts) lassen sich interpretieren als der multiplikative Einfluss des Anteils der jeweiligen Kategorie in einem Rasterquadrat auf die erwartete Anzahl an Wildunfällen pro Straßenmeter, wenn diese mindestens einmal in dem Rasterquadrat vorkommt. Dies ist zwar zunächst nicht sehr sachdienlich, die Formen der Funktionen bieten aber dennoch Aufschluss über Zusammenhänge die bisher verborgen blieben. Je größer die Dichte an bebautem Gebiet, desto geringer ist offenbar die Intensität der Wildunfälle. Für bewaldetes Gebiet steigt die Gefahr zunächst an, erreicht bei knapp 70% ihren Höhenpunkt und fällt dann wieder etwas ab. Das bedeutet also, dass in Waldrandgebieten das Risiko größer ist als innerhalb der Wälder, oder nur schwach bewaldeten Gebieten.

Eine Beobachtung, die man bei Betrachtung der Splinefunktionen der anderen auch schon zuvor semiparametrisch modellierten Kovariablen (Abb. 4.33) machen kann, ist, dass der starke Anstieg am linken Rand bei `laengen` wesentlich abgeschwächt ist. Das

Konfidenzband enthält für sehr geringe Straßendichte sogar den Wert eins. Durch die leichte Aggregierung haben also offenbar die punktuell auftretenden, künstlichen, extremen Übergewichte an Unfallpunkten (siehe Abschnitt 4.1.2) an Einfluss verloren. Für `verbiss` sind keine wesentlichen Unterschiede zu beobachten, bei `verbisseiche` ist der Abfall an den Rändern schwächer ausgeprägt. Allgemein sind die Konfidenzbänder breiter. Das hängt vermutlich damit zusammen, dass die Zielgröße zwar mehr Information birgt, die Fallzahl aber mit 31095 wesentlich geringer ist als etwa bei Modellierung mit logistischer Regression mit 831439, obgleich natürlich bei letzterer die Fallzahl durch die große Anzahl an Straßenpunkten künstlich erhöht wurde. Die Schätzung würde bei besserer Berücksichtigung der räumlichen Korrelation (siehe weiter unten) wohl präziser werden.

Im Einfluss der Lokation 4.34 sind keine strukturellen Unterschieden zu den Einflüssen in den vorherigen Abschnitten zu erkennen.

Der Modellfit (Abb. 4.35) ist verblüffend ähnlich zu denen aus den vorangegangenen beiden Abschnitten. Bei Abtragen der gefitteten Werte gegen die des logistischen Modells (Abb. 4.36) sind jedoch leichte Diskrepanzen zu erkennen: Während für kleinere Werte der Fit aus der logistischen Regression tendenziell größer ist, überholt das Modell mit negativer Binomialverteilung das logistische Modell mit steigender Intensität zunehmend. Die Unterschiede sind aber im Mittel gering und die Korrelation der Werte mit 0,839 recht hoch. Im Schnitt waren die Werte aus dem Modell mit negativer Binomialverteilung um 0,98 mal kleiner.

Abbildung 4.37 zeigt schließlich wieder die Ergebnisse der Analyse der Erklärwerte der Kovariablen anhand der Submodelle unter Auslassung je einer Kovariablen. Die Unterschiede zum vorherigen Abschnitt in der Rangfolge sind, dass `verbiss` und `laengen`, sowie `strasse` und `landnutzung` die Plätze gewechselt haben. Ersteres ist vermutlich damit zu erklären, dass der vermeintlich starke Anstieg der Intensität für sehr geringe Straßendichte stark abgemildert war, sodass der Einfluss von `laengen` abnimmt. Allerdings haben beide Kovariablen einen sehr geringen Erklärwert im vollen Modell.

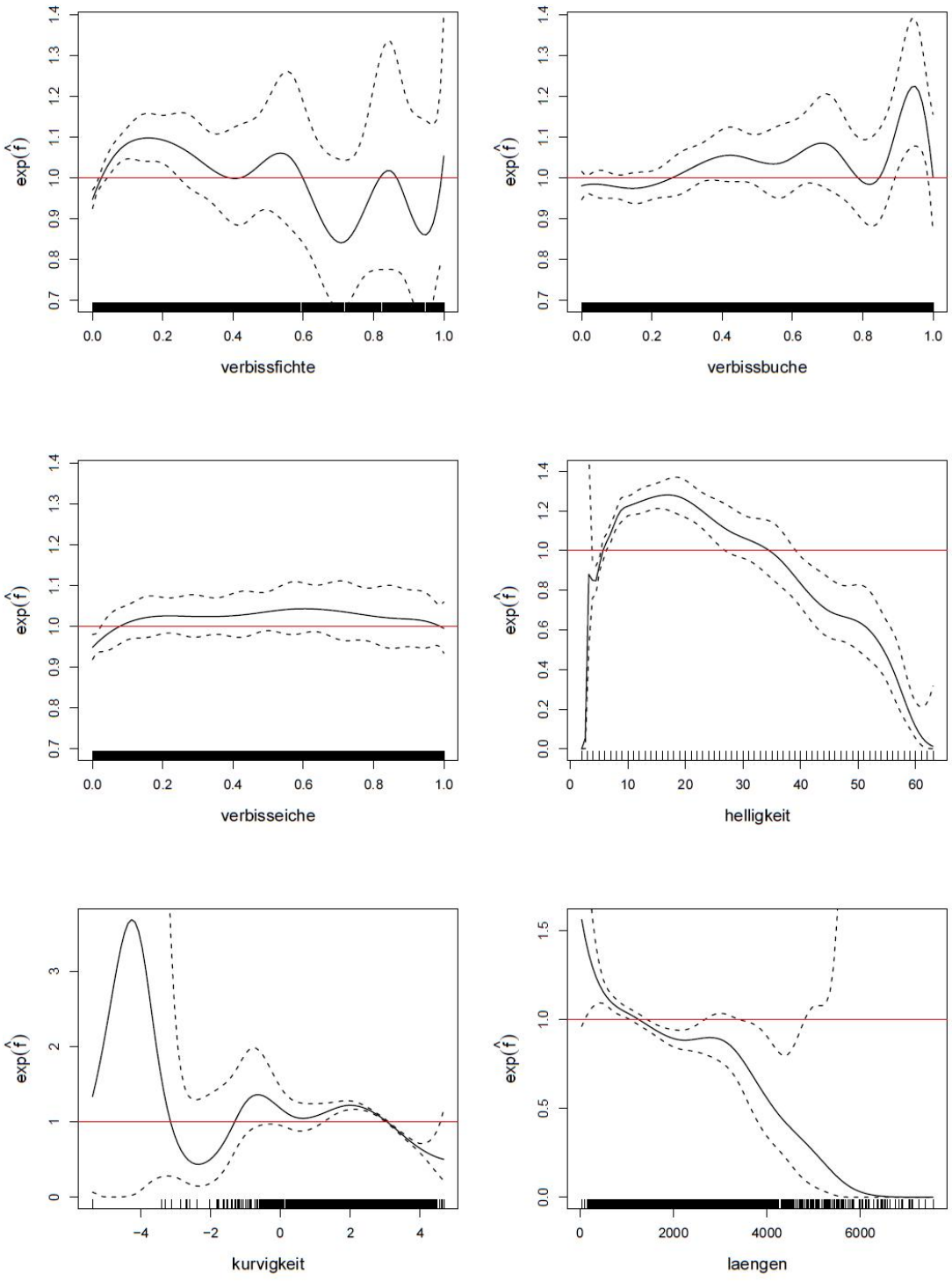

Abbildung 4.33.: exponierte Spline-Funktionen im Modell mit negativer Binomialverteilung

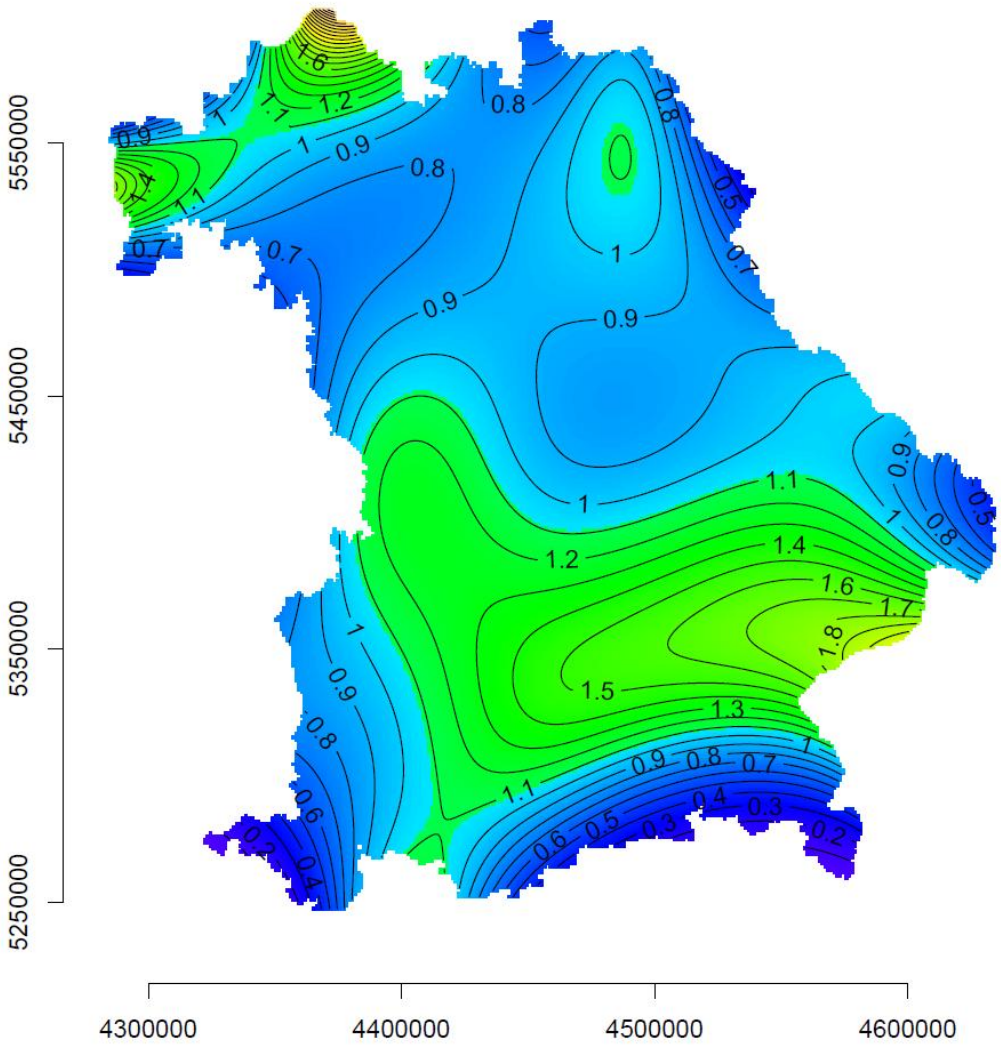

Abbildung 4.34.: Einfluss der Lokation im Modell mit negativer Binomialverteilung

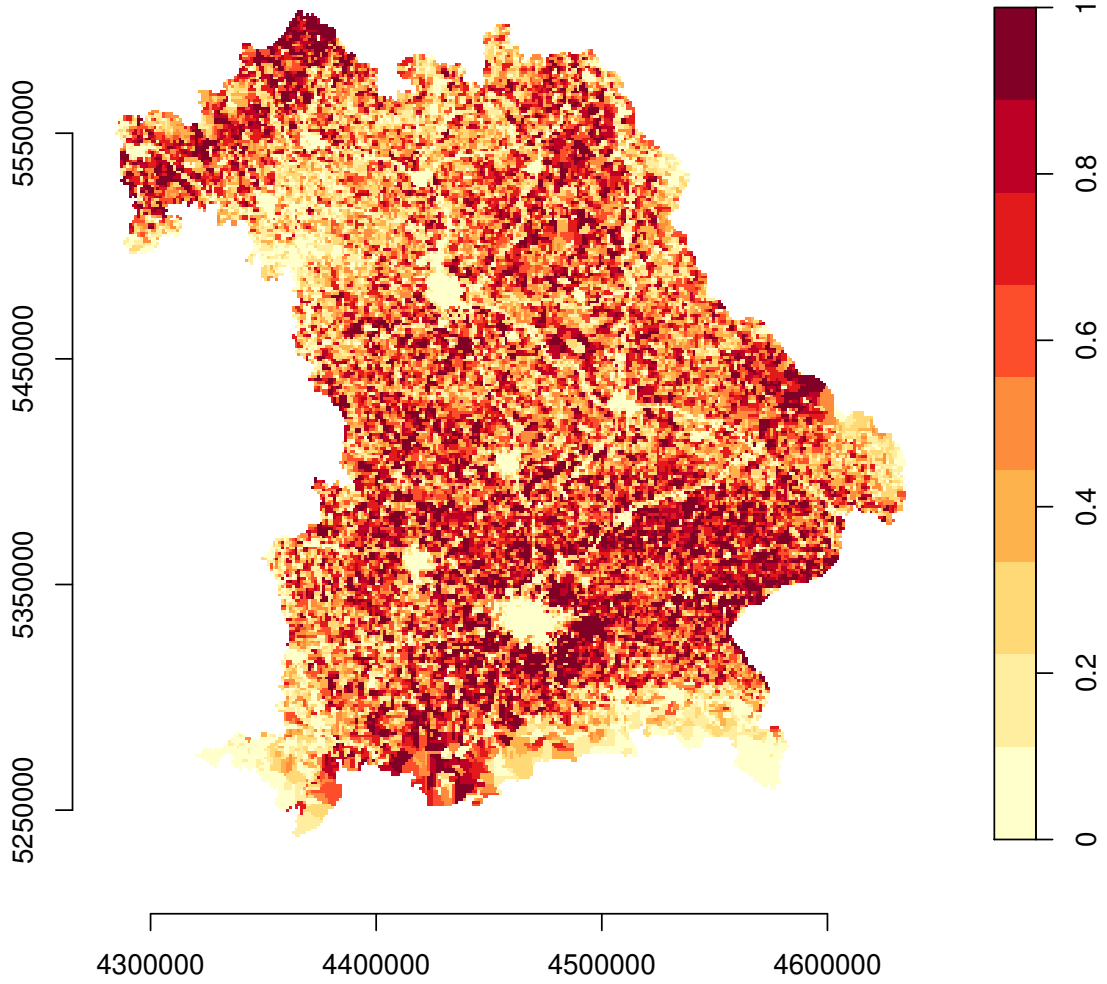

Abbildung 4.35.: Quantile des Fits des Modells mit negativer Binomialverteilung

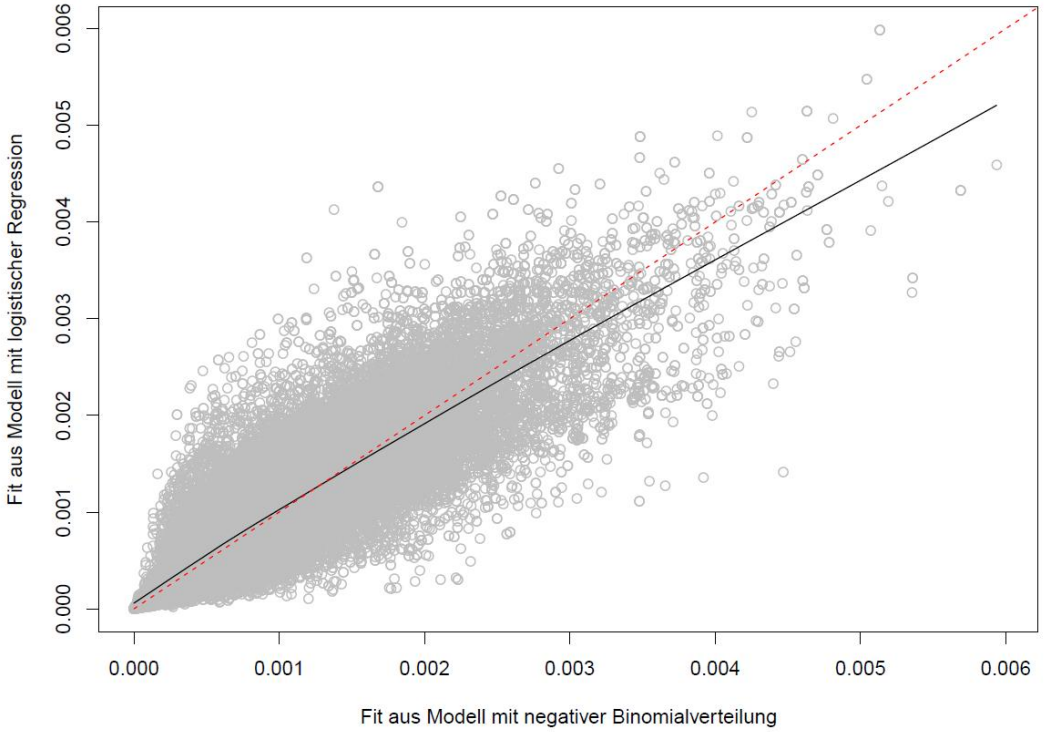

Abbildung 4.36.: Vergleich des Fits zwischen logistischem Modell und Modell mit negativer Binomialverteilung

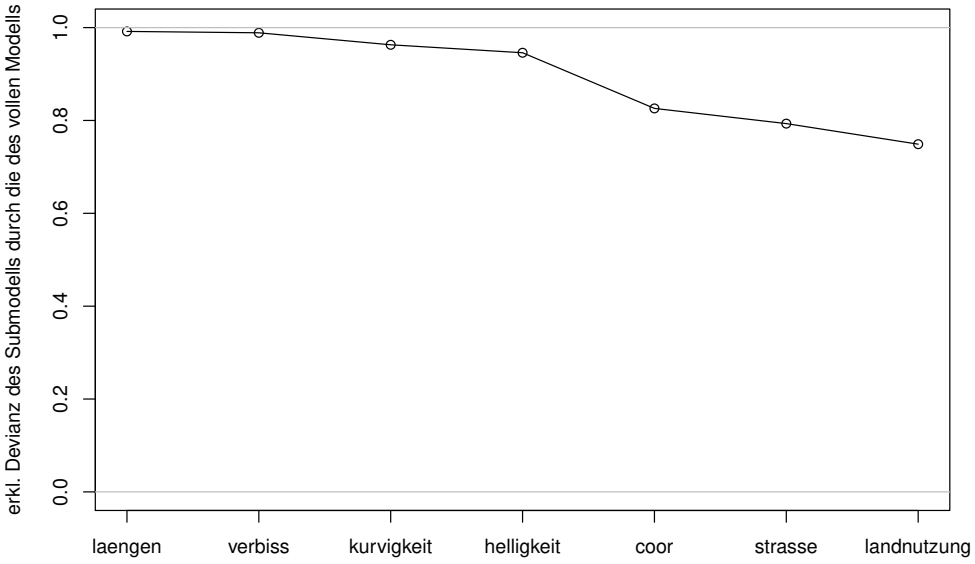

Abbildung 4.37.: Erklärwerte bei Auslassung der jeweiligen Kovariablen

Der verhältnismäßig geringere Einfluss von `strasse` kann, wie oben bereits beschrieben, höchstwahrscheinlich mit der Vergröberung der Information erklärt werden.

Wie bereits weiter oben erwähnt, ist durch die geringere Fallzahl die Möglichkeit gegeben, die räumliche Korrelation auf geeignetere Weise zu berücksichtigen als lediglich die Lokation als Tensorproduktspline in den Prädiktor mit aufzunehmen. Da die Werte der Zielgröße sich hier im Gegensatz zu der Modellierung mit einem räumlichen Poissonprozess auf einem Gitter befinden, würde sich ein generalisiertes Kriging-Modell, siehe z.B. Diggle und Ribeiro Jr. (2007), unter Annahme etwa der negativen Binomialverteilung für die Zielvariable anbieten. Dabei stößt man aber auf das Problem, dass die Fallzahl immer noch zu groß ist, um so ein Modell frequentistisch oder bayesianisch zu schätzen. Zur Likelihood-Schätzung muss auf Monte Carlo-Ziehungen zurückgegriffen werden, da zur direkten numerischen Maximierung der Likelihood ein n-dimensionales Integral gelöst werden müsste. Bei der Monte Carlo-Methode muss lediglich aus der n-dimensionalen multivariaten Normalverteilung des Gauß-Zufallsfeldes an den Daten-

punkten gezogen werden, was aber bei der vorliegenden Fallzahl ebenfalls kaum möglich ist. Für die bayesianische Herangehensweise ist das nicht nötig, man kann auch aus den vollbedingten Dichten der multivariaten Normalverteilung ziehen, was sich aber vom Zeitaufwand her auch nicht realisieren ließe, da damit pro Zug des Parametervektors ca. 30.000 Einzelzüge nötig wären. Wenn man zur Ausdünnung der Ziehungen nur jede hundertste verwenden würde, wären es schon 3 Millionen. Gotway und Stroup (1997) präsentieren eine Version der generalisierten Schätzgleichungen für räumliche Daten, die sich vermutlich besser für den vorliegenden Fall eignen würde. Die generalisierten Schätzgleichungen wurden ursprünglich für den Fall entwickelt, in dem die Zielvariable aus vielen Realisationen eines mehrdimensionalen Vektors abhängiger Zufallsvariablen besteht, sodass sich die Korrelationen zwischen der Zielgröße gegeben die Einflußgrößen (iterativ in Abwechslung mit den Koeffizienten) schätzen lassen. Bei räumlichen Daten existiert nur eine Realisation eines Zufallsvektors der Länge gleich der Anzahl der Beobachtungen. Deshalb ist ein solches Vorgehen hier nicht möglich. Stattdessen schätzt man zunächst, wie in diesem Abschnitt geschehen, ein gewöhnliches generalisiertes (bzw. additives) Modell unter Vernachlässigung der räumlichen Korrelation. Dann verwendet man die standardisierten Residuen aus diesem Modell, um anhand deren empirischen Variograms ein Modell für die Korrelation zu gewinnen. Mit Hilfe von Letzterem stellt man dann die $n \times n$ Korrelationsmatrix auf und löst unter Verwendung der entsprechenden Kovarianzmatrix die Schätzgleichungen. Dazu muss die $n \times n$-Kovarianzmatrix nur einmal invertiert werden. Dieses Modell eignet sich gut, wenn man Auskunft darüber möchte, wie die Kovariablen den mittleren Response, also in diesem Fall die Intensität der Wildunfälle beeinflussen, aber weniger, wenn der Fokus darauf liegt, Prädiktionen zu machen, da die Korrelationsstruktur recht ad hoc geschätzt wird (Diggle und Ribeiro Jr., 2007). Für die generalisierten Schätzgleichungen ist der mittlere Response wie in einer Querschnittsstudie zu interpretieren, also als Mittel über alle hypothetischen Einheiten, die die jeweilige Kovariablenkombination teilen (Diggle et al., 2002). Da die räumliche Korrelation auf diese Art bereits berücksichtigt wird, würde man die Lokation nicht mehr als Kovariable mit aufnehmen.

4.4. Analyse der Zeitpunkte

Zur Untersuchung der Anzahl der Wildunfälle in Abhängigkeit von Jahres- und Tageszeit war es nicht möglich, die Intensität für die einzelnen Tage nur anhand der Zeitpunkte der Unfälle an den jeweils betrachteten Tagen zu schätzen, da über einen einzelnen Tag hin zu wenig Unfälle stattfinden, um eine gleichzeitig flexible und stabile Schätzung zu erlauben. Ein derartiges Vorgehen ist auch nicht nötig, weil man sich die Gegebenheit zu Nutze machen kann, dass die Intensität an einem Tageszeitpunkt über mehrere Tage ähnliche Werte annimmt.

Würde man die zeitliche Intensität an einem Tag zum Zeitpunkt t nur anhand der Unfälle an diesem Tag schätzen, geschähe dies unter Auslassung eines Gewichts zur Randkorrektur nach der folgenden Formel

$$\hat{\lambda}(t) = \sum_{i=1}^{n} K(t - t_i), \tag{4.9}$$

wobei t_i der i-te von n Unfällen an einem bestimmten Tag und $K(\cdot)$ ein univariater Gaußkern mit Bandbreite σ ist. Dabei wird angenommen, dass das Punktmuster der Zeitpunkte die Realisation eines eindimensionalen Coxprozesses ist. Letzterer ist ganz analog zu dem in Abschnitt 4.1 beschriebenen räumlichen Coxprozess, ein Prozess mit zufälliger Intensitätsfunktion, gegeben eine Realisation dieser, er einem eindimensionalen inhomogenen Poissonprozess folgt (Diggle, 1985). Letzterer ist analog zu einem räumlichen inhomogenen Poissonprozess dadurch definiert, dass die Punkte in einem Intervall Poissonverteilt sind, mit Erwartungswert gleich dem Integral seiner Intensitätsfunktion über dieses Intervall und dass die Anzahl der Punkte in zwei sich nicht überlappenden Intervallen unabhängig voneinander sind (Chandler und Scott, 2011).

Zur Einbringung der Information über die Wildunfallzeitpunkte aus k benachbarten Tagen bei der Schätzung der Intensität an einem Tag d ($d \in \{1, \ldots, 365\}$) wurden diese mit Gewichten $w_i < 1$ ($i = 1, \ldots, k$), umso kleiner, je entfernter die Tage von dem eigentlich betrachteten Tag liegen, bei der Schätzung mit obiger Formel hinzugezogen.

Die so geschätzte Intensität ist (im Schnitt) um einen konstanten Faktor größer als die tatsächliche. Man erhält richtig geschätzte Werte, wenn man die Schätzung anschließend durch die Summe der Gewichte der benachbarten Tage plus eins für den betrachteten Tag teilt. Dazu nehme man als vereinfachendes Beispiel an, dass nur die zwei direkt benachbarten Tage mit einem Gewicht von jeweils 0,5 zusätzlich aufgenommen werden und man davon ausgeht, dass sich die zeitliche Intensität über diese Tage nicht ändert. Dann haben die Zeitpunkte der drei Tage zwar eine homogene Verteilung, sodass sich die Intensität schätzen lässt, aber es werden zu viele Zeitpunkte betrachtet. Genauer kommen durch den Einschluss der Zeitpunkte des vergangenen Tages im Schnitt 0,5 mal mehr hinzu. Dasselbe gilt für den nächsten Tag. Also verwendet man jeweils im Schnitt doppelt so viele Zeitpunkte (0,5 + 0,5 + 1) wie man sonst erwarten würde, sodass sich durch Teilen durch zwei (im Schnitt) eine korrekte Schätzung der Intensität ergibt. Die Gewichte werden verwendet, da sich die Intensität in Wirklichkeit über den jeweils betrachteten Zeitraum leicht ändert und auf diese Art den Zeitpunkten an näheren Tagen mehr Vertrauen geschenkt wird.

Es wurden 30 benachbarte Tage eingeschlossen, wobei die Gewichte basierend auf der Dichtefunktion der Normalverteilung mit Mittelwert Null und Standardabweichung fünf gewählt wurden. Genauer kamen sie dadurch zu stande, dass die Werte der Dichtefunktion an den Punkten 1 bis 15 durch den Wert bei Null geteilt wurden, sodass sie vom ersten Tag ausgehend langsam abfallen und kleiner als eins sind. Abbildung 4.38 zeigt die Gewichte.

Da sich bei Betrachtung eines Tages außerhalb des durch Tagesanfang und -ende gegebenen Zeitintervalls keine Punkte befinden, würde die Schätzung an den Rändern notwendigerweise abfallen. Um dies zu verhindern, wurden zur Schätzung der Intensität an einem Tag die jeweils zurückliegenden und nachfolgenden Tage angeschlossen, sodass sich die Schätzung an den Tagesgrenzen weitgehend bündig aneinanderfügt. Selbstverständlich wurde das jeweils auch für die heruntergewichteten, benachbarten Tage gemacht. Dabei trat das Problem auf, dass es zu Beginn bzw. am Ende eines Jahres keine Daten

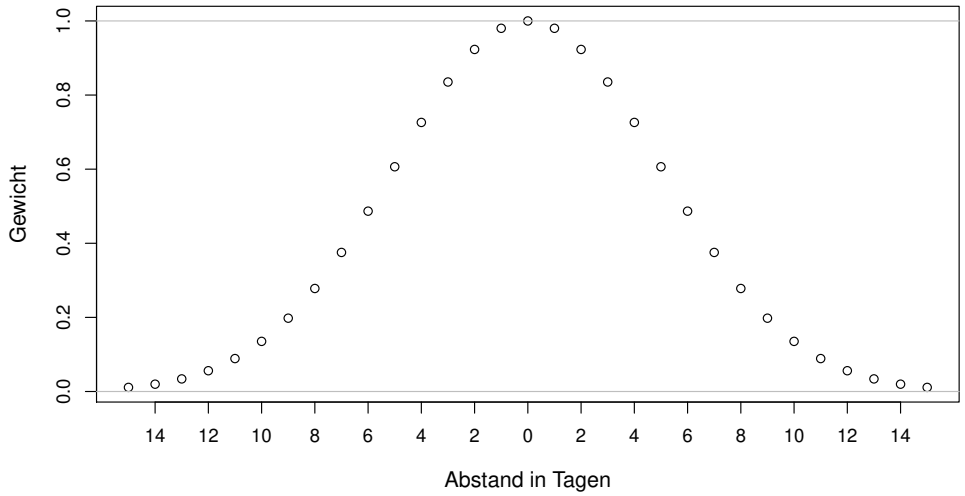

Abbildung 4.38.: Gewichte für benachbarte Tage aus $N(0, 5)$-Verteilung bei Schätzung der zeitlichen Intensität

zu dem vorherigen bzw. nachfolgenden Tag gibt, weshalb die (gewichteten) Zeitpunkte der vorherigen bzw. nachfolgenden Tage entsprechend hochgewichtet wurden, wenn sich zur Schätzung der Intensität an einem Tag verwendete Zeitpunkte am ersten bzw. letzten Tag des Jahres befanden. Es wurde die `density.default`-Funktion verwendet, die zur Auswahl einer geeigneten Bandbreite Silverman's Daumenregel (Silverman, 1986, S. 48, Gl. (3.31)) benutzt. Um eine flexiblere Schätzung der Intensität über den Tagesverlauf zu erhalten, wurde diese Bandbreite durch fünf geteilt. Die Gewichte konnten dabei über das Argument `weights` übergeben werden, da keine Kerndichteschätzung berechnet wird, wenn sich die Gewichte nicht zu eins aufsummieren, z.B. könnte man mit `density.default` auch eine gewöhnliche Intensitätsschätzung nach (4.9) durchführen, wenn alle Gewichte auf den Wert eins gesetzt werden.

Abbildung 4.39 zeigt die Ergebnisse für die beiden Jahre, wobei jeweils die roten Linien Sonnenauf- bzw. untergang, die gelben der bürgerlichen Dämmerung und die weißen der nautischen Dämmerung entsprechen, siehe z.B. Thompson und Thompson (2005). Es sind kaum Unterschiede zwischen den beiden Jahren zu erkennen. Offenbar passieren die

allermeisten Unfälle in der Dämmerung. Tagsüber ist die Gefahr von Wildunfällen vernachlässigbar gering. In der Abendämmerung geschehen tendentiell mehr Unfälle, wobei das Risiko etwa zwischen Mitte April bis Anfang August am größten ist. Zu dieser Zeit wurden auch mehr Unfälle in den späteren Nachtstunden registiert. Zu erwähnen sind außerdem die verringerten Anzahlen in der Morgendämmerung im Februar und Juni. Allgemein fluktuiert die Intensität über das Jahr recht stark. Der einzige auffallende Unterschied zwischen den beiden Jahren ist, dass 2006 zwischen Mitte Juni und Mitte Juli in der Abendämmerung verhältnismäßig weniger Unfälle stattfanden. Möglicherweise hängt dies mit der Fußballweltmeisterschaft in Deutschland zu dieser Zeit zusammen. Bei der Interpretation der absoluten Anzahlen ist hier, wie auch zuvor bei der räumlichen Analyse, zu beachten, dass sich nur Unfälle auf Autobahnen, Bundestraßen, Landstraßen und Kreisstraßen in dem zur Verfügung gestandenen Datensatz befanden. Bei Mittelung über alle Werte der Pixelimages der geschätzten zeitlichen Intensität für 2006 bzw. 2009 ergaben sich die Werte 3,909 bzw. 4,641. Die tatsächlichen Anzahlen an Unfällen pro Stunde in den beiden Jahren betrugen 3,902 bzw. 4,630. Man kann also davon ausgehen, dass die Schätzung im Mittel das richtige Ergebnis liefert.

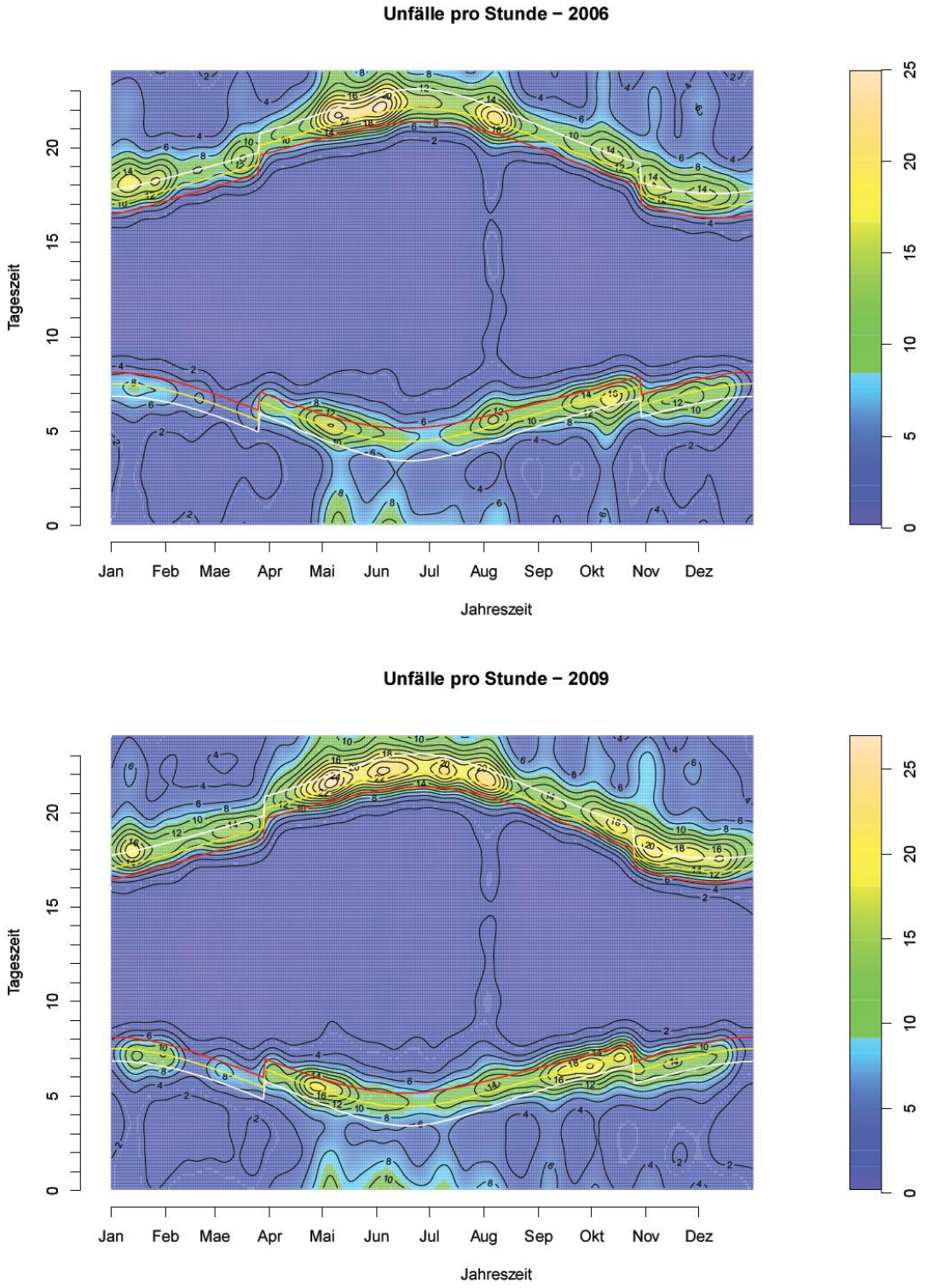

Abbildung 4.39.: geschätzte zeitliche Intensität der Rehunfälle in den Jahren 2006 und 2009

5. Zusammenfassung, Ausblick und Vergleich mit anderen Ergebnissen

Im Folgenden sollen nun die Ergebnisse aus den vorangegangenen Abschnitten in diskutierender Weise zusammengefasst, und insbesondere, soweit möglich mit denen von Hothorn et al. (in Vorb.) und Kaldhusdal (2011) verglichen werden. Wie schon in der Einleitung erwähnt, modellieren diese die Anzahl der Wildunfälle auf Gemeindeebene. Diese Analysen erfolgen auch für das Gebiet Bayern und die stetigen Kovariablen werden ebenfalls semiparametrisch in die Modelle aufgenommen. Dabei wurde angenommen, dass die Anzahlen gegeben die Kovariablenwerte der Poissonverteilung folgen. In den genannten Arbeiten wurden allerdings zusätzliche Kovariablen, wie Temperatur oder die Präsenz bzw. Absenz von Rotwildhegegemeinschaften berücksichtigt und andere hier verwendete Einflußgrößen wie die Nachtlichter oder die Kurvigkeit fanden dort keinen Eingang, sodass kein strikter Vergleich, insbesondere des allgemeinen Modellfits möglich ist. Außerdem betrachteten die genannten Arbeiten auch Rot- und Damwild. Allerdings sind diese in Bayern an weniger als 1% der Wildunfälle beteiligt Hothorn et al. (in Vorb.).

Generell konnte durch die eingeschlossenen Kovariablen auf lokaler Ebene wenig an der bestehenden Heterogenität aufgeklärt werden. Bis auf bei der Modellierung mit der negativen Binomialverteilung waren die erklärten Devianzen immer kleiner als 10%, bei der Modellierung als logistischer Regressionsmodell mit 7,4% am größten. Auffällig war,

dass kleine Gebiete mit besonders hoher beobachteter Wildunfalldichte mit den Modellen nicht detektiert werden konnten. Auch wenn bei der Modellierung mit der negativen Binomialverteilung sich die Struktur der Kovariablen, die auf einem feinen Raster gegeben waren, zu Nutze gemacht wurde, könnte man den relativ hohen Wert der erklärten Devianz bei dieser Modellierung (20,4%) auch teilweise darauf zurückführen, dass hier die beobachtete Überdispersion mit vereinzelt sehr großen Anzahlen bereits in der Modellierung berücksichtigt wird und so die starken Abweichungen durch diese „Ausreißer" weniger ins Gewicht fallen. Auf Gemeindeebene konnten Hothorn et al. (in Vorb.) die Anzahlen an Wildunfällen pro Straßenmeter hingegen mit hoher Genauigkeit vorhersagen. Das könnte auch teilweise mit der in Abschnitt 4.3 diskutierten anderen Informationen aus den Kovariablen bei Aggregierung zusammenhängen. Denkbar wäre eine der dem Vorgehen in Abschnitt 4.3 ähnliche leichte Aggregierung unter Berücksichtigung der räumlichen Korrelation über räumliche generalisierte Schätzgleichungen nach Gotway und Stroup (1997). Vermutlich hängt das lokale Risiko von Wildunfällen von sehr vielen, miteinander interagierenden Faktoren ab, die nicht mit den betrachteten Kovariablen abgedeckt werden können. Außerdem sind räumliche Kovariablen notwendigerweise häufig miteinander korreliert. Vor diesem Hintergrund und mit der Möglichkeit sehr viele, möglicherweise annäherend redundante Kovariablen aus dem Internet herunterzuladen, könnte man evtl. auch ein dimensionsreduzierende Verfahren aus der Regression, z.B. penalisierte Kleinste-Quadrate anwenden, um die Prädiktion zu verbessern. Das Hauptziel dieser Analyse war jedoch die Untersuchung der Form des Einflusses der Kovariablen.

Hothorn et al. (in Vorb.) führen in ihrer Arbeit den sogenannten „deer-vehicle collision index" ein. Dieser entspricht in ihrem Modell dem geschätzten multiplikativen Einfluss von Klima, Landnutzung, Verbissintensität und der geografischen Lage auf die Wildunfallintensität an einer bestimmten Lokation. In der Arbeit wird dieser grafisch als Karte dargestellt. In der hier durchgeführten Analyse wurde das Klima nicht mit einbezogen. Dennoch wurde eine analoge Karte basierend auf den drei anderen Kovariablen für den räumlichen Poissonprozess mit Dummypunkten auf den Straßen erstellt. Für

das logistische Modell könnte man das, wie in Abschnitt 4.3 diskutiert, aus formaler Sicht nicht machen, wobei sich allerdings in etwa dasselbe Ergebnis ergeben hätte, zumal bis auf bei `strasse` die Einflüsse zwischen den beiden Modellen sehr ähnlich waren. Bei Vergleich der Karte zu der Analyse aus dieser Arbeit und der von Hothorn et al. (in Vorb.) ergab sich makroskopisch dasselbe Bild. Aber es waren auch lokal Unterschiede zu erkennen, die sich oft in Bereichen von Rotwildhegegemeinschaften befanden. In Hothorn et al. (in Vorb.) wurde hinsichtlich der Präsenz bzw. Absenz Letzterer adjustiert, deren Vorhandensein einen negativen Effekt auf die Intensität hatte. Wichtig ist auch, dass hier keine Klimavariablen berücksichtigt wurden, welche einen sehr starken Einfluss auf die Intensität haben. Ein erwähnenswerter Unterschied war noch, dass hier in dem südlich an München angrenzenden Bereich viel mehr Unfälle prognostiziert wurden. Allerdings konnten sehr oft auch feinere Strukturen mit beiden Modellen beobachtet werden. Ein allgemeiner Unterschied war, dass die multiplikativen Änderungen hier viel stärker schwankten, so befand sich das 90%-Quantil bereits bei gut 5 mal erhöhtem Risiko, während bei Hothorn et al. (in Vorb.) der höchste Wert etwa 2,8 mal erhöhtes Risiko war. Das kann man aber damit erklären, dass hier bei `landnutzung` aufgrund der Wahl der Referenzgkategorie `bebaut` die Koeffizienten alle positiv waren, sodass kleinere Werte bei Betrachtung des Einflusses der drei Kovariablen selten auftreten können.

Die Einflußgrößen Landnutzung, Verbissintensität und die geografischen Koordinaten finden sich auch in Hothorn et al. (in Vorb.) und Kaldhusdal (2011). Allerdings wurden diese Konstrukte bis auf die geografischen Koordinaten auf andere Arten berücksichtigt. Z.B. unterscheidet Kaldhusdal (2011) hinsichtlich der verbissenen Bäume Nadelbäume und Laubbäume bzw. Tannen, während hier Fichten, Rotbuchen und Eichen unterteilt wurden. Auch aus diesem Grund können nur grobe Vergleich gemacht werden.

Die Verbissintenisität hatte in Übereinstimmung mit Hothorn et al. (in Vorb.) und Kaldhusdal (2011) die geringste Wichtigkeit zur Aufklärung der Heterogenität. Der Einfluss von `verbissfichte` war für geringe Anteile leicht negativ, stieg dann bis etwa 0,25 zu seinem Maximum mit ca. 1,1 mal erhöhter Intensität an und fiel für größere Anteile

tendentiell ab, wobei das Konfidenzband in diesem Bereich aufgrund weniger Beobachtung sehr groß war. Bei Hothorn et al. (in Vorb.) und Kaldhusdal (2011) waren für Nadelbäume aufgrund der Aggregierung auf Gemeindeebene und der damit verbundenen geringeren Volatilität die Anteile kaum größer als 0,25. In diesem Bereich ergaben sich aber, wenn man von Abweichungen im Bereich sehr weniger Beobachtungen und den aufgrund anderer Knoten in den Spline-Basen flexibleren Funktionen absieht, dieselben Ergebnisse. `verbissbuche` wurde in den beiden Arbeiten nicht expliziert berücksichtigt. In dieser Arbeit lag der Einfluss für geringen Verbiss bis etwa 0,25 nahe bei eins, erreichte dann bei ca. 0,4 mit 1,1 sein vorläufiges Maximum. Für sehr großen Verbissanteil nahe bei eins war der Einfluss mit 1,2-fach erhöhter Intensität jedoch am größten. `verbisseiche` fand bei Hothorn et al. (in Vorb.) in Kombination mit Tannen und bei Kaldhusdal (2011) zusammen mit Laubbäumen und Tannen Eingang. Bei Ersteren zeigte sich nur bei sehr geringem Anteil ein negativer Effekt, der allerdings aufgrund weniger Beobachtungen in diesem Bereich nicht verlässlich ist. Bei Letzterem war der Einfluss nur für sehr großem Verbissanteil leicht erhöht. Hier hatte `verbisseiche` in Übereinstimmung mit diesen Ergebnissen einen sehr geringen Effekt. Die Lurking Variable Plots wiesen darauf hin, dass der Einfluss z.B. bei `verbissfichte` für sehr geringe Verbissanteile ein abweichendes Verhalten aufweist. Außerdem lieferten sie Hinweise darauf, dass der Verbissanteil von der Intensität der jeweiligen Baumart abhängig sein könnte, sodass es möglicherweise sinnvoll wäre, die Intensitäten der Baumarten ebenfalls in den Prädiktor mit aufzunehmen, damit der Effekt des Verbissanteiles besser mit dem der Wilddichte übereinstimmt.

Straßentyp hatte den größten Einfluss zur Erklärung der Devianz im vollen Modell. Eine Ausnahme bildete hier das Ergebnis bei der Modellierung mit der negativen Binomialverteilung, das aber aufgrund der Vergröberung der Information weniger verlässlich ist. Dieses Ergebnis bekräftigt die Notwendigkeit, den Straßentyp, wie in Hothorn et al. (in Vorb.) auch bei Analysen auf Gemeindeeben zu berücksichtigen. Die Rangfolge, absteigend hinsichtlich der Gefahr von Wildunfällen, war dabei: Bundesstraßen, Landstraßen, Kreisstraßen und Autobahnen, wobei zum Teil große Unterschiede bestanden.

Hothorn et al. (in Vorb.) merken ebenfalls an, dass die Gefahr von Wildunfällen auf Autobahnen gering war.

`landnutzung` hatte hier den zweitgrößten Einfluss, was mit dem Ergebnis von Hothorn et al. (in Vorb.) übereinstimmt, dass diese nach den klimatischen Bedingungen die wichtigste Einflussgröße darstellt. Der Einfluss dieser Kovariable war sehr verschieden zwischen dem räumlichen Poissonprozess bei Dummypunkten auf einem regulären Gitter auf der einer Seite und dem bei Dummypunkten auf den Straßen bzw. dem logistischen Modell auf der anderen. Dieses Ergebnis deutet darauf hin, dass die Verwendung eines gewöhnlichen räumlichen Poissonprozesses zur Untersuchung der Intensität der Wildunfälle offenbar weniger geeignet ist. Allerdings könnte die Diskrepanz auch weniger stark ausgeprägt sein, wenn die Unfälle auf allen Straßen verwendet werden und damit die Dichte der Unfälle viel größer und weniger abhängig von der Straßenstruktur ist. Für den räumlichen Poissonprozess mit Dummypunkten auf den Straßen und das logistische Regressionsmodell war die Intensität auf Straßen im Wald mit Abstand am größten, danach folgten Straßen in landwirtschaftlichen Gebieten und solche im Bereich von Feucht- und Wasserflächen. Straßen in bebautem Gebiet wiesen das geringste Risiko auf. Für das Modell bei Annahme der negativen Binomialverteilung konnten auch Anteile an bebautem und bewaldeten Gebiet betrachtet werden. Diese Ergebnisse können in gewissen Maß mit denen von Hothorn et al. (in Vorb.) und Kaldhusdal (2011) verglichen werden, da diese ebenfalls Anteile der jeweiligen Kategorien betrachteten, mit dem Unterschied, dass dort keine Indikatorvariablen für das Vorhandensein der einzelnen Kategorien verwendet wurden. Zum Vergleich der Werte werden die exponierte Koeffizienten der Indikatorvariablen mit den exponierten Splinefunktionen aus dieser Arbeit multipliziert. Der in den beiden Arbeiten festgestellte annäherend lineare und negative Einfluss des Anteils an bebautem Gebiet kann hier ebenfalls bestätigt werden, wobei sich für Hothorn et al. (in Vorb.) auch der beobachtete Wertebereich von etwa 0,2 bis ungefähr 1 der exponierten Splinefunktion mit dem Ergebnis aus dieser Arbeit deckt. Hothorn et al. (in Vorb.) und Kaldhusdal (2011) betrachten auch die Länge an Waldrand pro Flächeneinheit und stellen fest, dass die Gefahr von Wildunfällen in Gebieten mit

mehr Waldrand tendenziell größer ist, mit Ausnahme von solchen mit sehr viel Waldrand, in denen die Intensität wieder abnimmt. Ähnliches wurde hier auch beobachtet: Die Gefahr stieg für Rasterquadrate mit größerem Waldanteil bis etwa 0,7 linear an und fiel dann wieder, ebenfalls linear, ab. Zu Wasser- und Feuchtflächen waren weder der Koeffizient zu der Indikatorvariable noch der Einfluss der geschätzten Splinefunktion signifikant, wobei Ersterer auch nach Auslassen des semiparametrischen Teils nicht signifikant wurde.

Die räumliche Lokation kam in der Rangfolge des Einflusses, wie auch in Hothorn et al. (in Vorb.), bereits nach der Landnutzung, was ein Anzeichen dafür ist, dass es noch wichtige räumliche, unberücksichtigte oder unbeobachtete Einflußgrößen gibt und auch ein Anreiz dazu ist, viele weitere Kovariablen zu sammeln, um die Prädiktion zu verbessern. In dieser Analyse war das Risiko in Niederbayern und im Norden Unterfrankens am größten, während es am Alpenrand stark abfiel. Letzteres wurde in den anderen beiden Arbeiten nicht beobachtet, was sich vermutlich darauf zurückführen lässt, dass sich in diesem Bereich fast alle Gemeinden in Rotwildhegegemeinschaften befinden. Interessanterweise ist derartiges für die Rotwildhegegemeinschaften im Norden Unterfrankens nicht zu beobachten, vielmehr ist hier der Einfluss stärker positiv. Ähnliches gilt für die Rotwildhegegemeinschaften in der Oberpfalz. Das könnte darauf hinweisen, dass der Einfluss der Indikatorvariable zu den Rotwildhegegemeinschaften räumlich nicht stationär ist. Wenn man berücksichtigt, dass der Einfluss der räumlichen Lage in den drei Fällen nicht exakt auf die gleiche Art geschätzt wurde und, dass hier etwas andere Kovariablen verwendet wurden, scheint es zwischen den drei Ergebnissen, neben den bereits erwähnten, keine bedeutenden Unterschiede zu geben.

Die drei Kovariablen `laengen`, `kurvigkeit` und `helligkeit` zeigten nach `verbiss` den geringsten Einfluss. Für sehr dunkle Gebiete ist die Gefahr offenbar sehr gering, steigt dann für etwas hellere Gebiete im Vergleich zum Mittel bis auf etwa das 1,2-fache an und nimmt dann in beinahe linearer Weise wieder ab, bis in sehr hellen Gebieten kaum noch Unfälle zu erwarten sind. Bei `kurvigkeit` war ein etwas komplexerer und schwäche-

rer Zusammenhang zu erkennen, wobei ein anderes Maß für die Kurvigkeit auch etwas andere Ergebnisse liefern könnte. Im Bereich sehr gerader Straßen war die Intensität geringer, für mittlere bis stärkere Kurvigkeit war sie bei unstetigem Verlauf der exponierten Splinefunktion etwas größer und nahm dann für starke Kurvigkeit wieder ab. Die Betrachtung des Lurking Variable Plots zu `laengen` zeigte, dass dem geschätzten Einfluss hier, insbesondere für geringe Straßenlängen wenig Vertrauen geschenkt werden kann, was auf vereinzelte künstliche Ungleichgewichte zwischen Unfalllokationen und Punkten aus dem Straßennetz zurückgeführt wurde. Die exponierte Splinefunktion stieg für geringe Straßendichte stark an und fiel für Bereiche mit dichter liegenden Straßen immer weiter ab, wobei Ersteres bei der Modellierung mit negativer Binomialverteilung nur in sehr abgeschwächter Form beobachtet wurde und das Konfidenzband dabei den Wert eins einschloß.

Die Analyse der Zeitpunkte der Unfälle ergab, dass sich die allermeisten Unfälle in der Dämmerung, und dabei tendenziell eher abends ereignen. Zwischen den beiden Jahren 2006 und 2009 gab es kaum Unterschiede. Die Verteilung der Unfälle über den Tag hinweg ändert sich über das Jahr bisweilen offensichtlich recht stark.

Abschließend lässt sich festhalten, dass die Ergebnisse in Einklang mit denen stehen, die man auch bei Modellierung der Wildunfalldaten auf Gemeindeebene erhält, man aber über die Lokalität der Kovariablen zusätzliche Erkenntnisse gewinnen kann. Allerdings ließ sich die Intensität der Wildunfälle auf punktueller Ebene nur schwer prognostizieren, wobei insbesondere Lokationen mit besonders hoher Intensität nicht detektiert werden konnten.

Literaturverzeichnis

A. Baddeley (2008). *Analysing spatial point patterns in R.* CSIRO and University of Western Australia. URL `http://www.csiro.org/files/files/piph.pdf`.

A. Baddeley, R. Turner (2005a). 'Spatstat: an R package for analyzing spatial point patterns'. *Journal of Statistical Software*, **12**(6):1–42. URL `www.jstatsoft.org`. ISSN 1548-7660.

A. Baddeley, R. Turner, J. Møller, M. Hazelton (2005b). 'Residual analysis for spatial point processes'. *Journal of the Royal Statistical Society, series C*, **67**:617–666.

M. Berman, T. R. Turner (1992). 'Approximating Point Process Likelihoods with GLIM'. *Applied Statistics*, **41**:31–38.

R. Bivand, C. Rundel (2011). *rgeos: Interface to Geometry Engine - Open Source (GEOS).* URL `http://CRAN.R-project.org/package=rgeos`. R package version 0.1-12.

A. C. Cameron, F. A. G. Windmeijer (1996). 'R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization'. *Journal of Business & Economic Statistics*, **14**(2):209–220.

R. Chandler, M. Scott (2011). *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences.* John Wiley & Sons, Chichester.

T. M. Davies, M. L. Hazelton, J. C. Marshall (2011). 'sparr: Analyzing Spatial Relative Risk Using Fixed and Adaptive Kernel Density Estimation in R'. *Journal of Statistical Software*, **39**(1):1–14. URL `http://www.jstatsoft.org/v39/i01/`.

P. Diggle (1985). 'A Kernel Method for Smoothing Point Process Data'. *Applied Statistics (Journal of the Royal Statistical Society, series C)*, **34**:138–147.

P. J. Diggle, P. Heagerty, K. Y. Liang, S. L. Zeger (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

P. J. Diggle, P. J. Ribeiro Jr. (2007). *Model-based Geostatistics*. Springer, New York.

J. C. Domingues (2008). *Lacroix and the calculus*. Birkhäuser, Basel.

L. Fahrmeir, T. Kneib, S. Lang (2007). *Regression - Modelle, Methoden und Anwendungen*. Springer, Berlin.

J. Fox (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage Publications, London, New Delhi.

C. Gaetan, X. Guyon (2010). *Spatial Statistics and Modeling*. Springer, New York.

C. A. Gotway, W. W. Stroup (1997). 'A generalized linear model approach to spatial data analysis and prediction'. *Journal of Agricultural, Biological and Environmental Statistics*, **2**:157–178.

G. Grimmett, D. Welsh (1986). *Probability: An Introduction*. Oxford University Press, Oxford.

T. Hastie, H. Tibshirani, J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

M. Heveling (2006). *Bijective point maps, point-stationarity and characterization of Palm measures*. Dissertation, Universität Karlsruhe.

J. M. Hilbe (2011). *Negative Binomial Regression*. Cambridge University Press, Cambridge.

K. Höllig (2003). *Finite Element Methods with B-Splines*, Bd. 26 von *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.

T. Hothorn, R. Brandl, J. Müller (in Vorb.). 'Large-scale Model-based Assessment of Deer-vehicle Collision Risk'.

J. Illian, A. Penttinen, H. Stoyan, D. Stoyan (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Chichester.

E. Imhof (1972). *Thematische Kartographie. Lehrbuch der Allgemeinen Geographie. Band 10*. Walter de Gruyter, Berlin, New York.

A. Kaldhusdal (2011). *Modellierung von Umwelt- und Klimaeinflüssen auf die Intensität von Wildunfällen*. Bachelorarbeit, Ludwig-Maximilians-Universität München.

L. J. Keele (2008). *Semiparametric Regression for the Social Sciences*. John Wiley & Sons, Chichester.

M. Keil, M. Bock, T. Esch, A. Metz, S. Nieland, A. Pfitzner (2010). *CORINE Land Cover Aktualisierung 2006 für Deutschland. Abschlussbericht zu den F+E Vorhaben UBA FKZ 3707 12 200 und FKZ 3708 12 200*. Deutsches Zentrum für Luft- und Raumfahrt e.V., Deutsches Fernerkundungsdatenzentrum Oberpfaffenhofen. URL `http://www.corine.dfd.dlr.de`.

T. H. Keitt, R. Bivand (2011). *rgdal: Bindings for the Geospatial Data Abstraction Library*. URL `http://CRAN.R-project.org/package=rgdal`. R package version 0.7-1.

S. Kühlmann-Berenzon (2002). *Tree influence on understory vegetation: an edge correction and a conditional model*. Lizentiatsarbeit, Universität Göteborg.

K. Lange (2010). *Numerical Analysis for Statisticians*. Springer, New York.

N. J. Lewin-Koh, R. Bivand (2011). *maptools: Tools for reading and handling spatial objects*. URL `http://CRAN.R-project.org/package=maptools`. R package version 0.8-10.

J. M. Loh, M. L. Stein (2004). 'Bootstrapping a spatial point process'. *Statistica Sinica*, **14**:69–101.

J. Møller, R. P. Waagepetersen (2004). *Statistical inference and simulation for spatial point processes*. Chapman and Hall, Boca Raton.

J. Møller, R. P. Waagepetersen (2007). 'Modern Statistics for Spatial Point Processes'. *Scandinavian Journal of Statistics*, **34**(4):643–684.

M. Neteler, H. Mitasova (2004). *Open Source GIS: A GRASS-GIS Approach. 2nd Edition*. Kluwer Academic Publishers, Boston.

E. Neuwirth (2011). *RColorBrewer: ColorBrewer palettes*. URL `http://CRAN.R-project.org/package=RColorBrewer`. R package version 1.0-5.

J. Ohser (1983). 'On estimators for the reduced second moment measure of point processes'. *Math. Operationsforsch. u. Statist., ser. statist*, **14**:63–71.

G. Olbrich, M. Quick, J. Schweikart (2002). *Desktop Mapping: Grundlagen und Praxis in Kartographie und GIS*. Springer, Berlin.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

C. Reimann, P. Filzmoser, R. G. Garrett, R. Dutter (2008). *Statistical data analysis explained: applied environmental statistics with R*. John Wiley & Sons, Chichester.

B. D. Ripley (1988). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge.

K. Schladitz, A. J. Baddeley (2000). 'A third order point process characteristic'. *Scandinavian Journal of Statistics*, **27**:657–671.

V. Schmid (2010). 'Räumliche Statistik'. Vorlesungsskript.

J. T. Schnute, N. Boers (2010). *PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools*. URL `http://CRAN.R-project.org/package=PBSmapping`. R package version 2.61.9.

B. W. Silverman (1986). *Density Estimation*. Chapman and Hall, London.

R. B. Thompson, B. F. Thompson (2005). *Astronomy Hacks*. O'Reilly Media, Inc., Sebastopol.

D. I. Warton, L. Shepherdy (2009). 'Presence only data, logistic regression and Poisson point processes'. Techn. Ber., School of Mathematics and Statistics, University of Sydney. URL `http://www.maths.unsw.edu.au/sites/default/files/ippannals_0.pdf`. Eingereicht bei *Annals of Applied Statistics*.

L. Wasserman (2005). *All of statistics: a concise course in statistical inference*. Springer, New York.

Wikipedia (2011). 'Gauß-Krüger-Koordinatensystem — Wikipedia, Die freie Enzyklopädie'. `http://de.wikipedia.org/wiki/Gau%C3%9F-Kr%C3%BCger-Koordinatensystem`. [Online; Stand 12. Juni 2011].

S. N. Wood (2004). 'Stable and efficient multiple smoothing parameter estimation for generalized additive models'. *Journal of the American Statistical Association*, **99**:673–686.

S. N. Wood (2006a). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

S. N. Wood (2006b). 'Low rank scale invariant tensor product smooths for generalized additive mixed models'. *Biometrics*, **62**(4):1025–1036.

A. Zeileis, C. Kleiber, S. Jackman (2008). 'Regression Models for Count Data in R'. *Journal of Statistical Software*, **27**(8):1–14. URL `http://www.jstatsoft.org/v27/i08/`.

Abbildungsverzeichnis

2.1. links: Tensorprodukt-B-Spline-Basisfunktion zum Grad $l = 2$, rechts: entsprechende Basisfunktion mit 3 inneren Knoten 17

3.1. mit adaptiver Glättung prognostizierte gegen wahre Anteile mit Loess-Kurve - Pilot-Bandbreite: 2000, globale Bandbreite: 750 33
3.2. mit adaptiver Glättung prognostizierte Anzahlen von verbissenen Bäumen getrennt nach Baumart 34
3.3. CORINE Land Cover-Daten vergröbert in vier Kategorien 38
3.4. Straßennetz 39
3.5. Straßenkarte mit Ausprägungen an jeder Lokation 41
3.6. Ausschnitt aus der Straßenkarte - Fehlspezifikationen sind rot gekennzeichnet 42
3.7. Maß der Kurvigkeit - abgeschlagen kleine Werte (< -1) sind der Darstellbarkeit halber ausgeschlossen 44
3.8. approximierte Straßenlängen - abgeschlagen große Werte (> 4000) der Darstellbarkeit halber ausgeschlossen 46
3.9. Karte der Nachtlichter 47

4.1. nonparametrische Schätzung der Wildunfallintensität bei Verwendung eines Gaußkerns mit Bandbreite $\sigma = 5000$ 50
4.2. geschätzte K- und L-Funktion für die Rehunfalllokationen - gestrichte rote Linien: Werte für homogenen Poissonprozess 51
4.3. Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 1 56

4.4. Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 2 57
4.5. exponierte Koeffizientenschätzer der kategorialen Kovariablen im Modell mit Dummypunkten auf regulärem Gitter 61
4.6. exponierte Spline-Funktionen im Modell mit Dummypunkten auf regulärem Gitter 62
4.7. Einfluss der Lokation im Modell mit Dummypunkten auf regulärem Gitter 64
4.8. Quantile des Fits des Modells bei Dummypunkten auf regulärem Gitter . 65
4.9. geglättetes Residuenfeld für Modell mit Dummypunkten auf reguläreem Gitter 68
4.10. Lurking Variable Plots mit Ableitungen - `verbiss` 70
4.11. Lurking Variable Plots - `helligkeit` 73
4.12. Lurking Variable Plots - `helligkeit` bei erhöhter Knotenzahl 74
4.13. exponierte Spline-Funktionen bei erhöhter Knotenzahl für `helligkeit` . . 75
4.14. Einteilung in Train- (grün) und Testdaten (gelb) zusammen mit den Wildunfällen (rot) 77
4.15. links: prognostizierte gegen wahren Anzahlen, rechts: Differenz zwischen prognostizierten und wahren Anzahlen - rote Linie: Mittelwert 78
4.16. Erklärwerte bei Auslassung der jeweiligen Kovariablen 79
4.17. Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 1 82
4.18. Koeffizientenschätzer bei $K \times K$-Gitter an Kacheln - Teil 2 83
4.19. exponierte Koeffizientenschätzer der kategorialen Kovariablen im Modell mit Dummypunkten auf Straßen 87
4.20. exponierte Spline-Funktionen im Modell mit Dummypunkten auf Straßen 88
4.21. Einfluss der Lokation im Modell mit Dummypunkten auf Straßen 90
4.22. Quantile des Fits des Modells bei Dummypunkten auf Straßen 92
4.23. Lurking Variable Plots mit Ableitungen 93
4.24. links: prognostizierte gegen wahren Anzahlen, rechts: Differenz zwischen prognostizierten und wahren Anzahlen - rote Linie: Mittelwert 96
4.25. Erklärwerte bei Auslassung der jeweiligen Kovariablen 97

4.26. exponierte Koeffizientenschätzer der kategorialen Kovariablen im logistischen Modell . . . 103
4.27. exponierte Spline-Funktionen im logistischen Modell . . . 104
4.28. Einfluss der Lokation im logistischen Modell . . . 105
4.29. Quantile des Fits des logistischen Modells . . . 106
4.30. Vergleich des Fits zwischen Poissonprozess mit Dummypunkten auf den Straßen und logistischem Modell . . . 107
4.31. Erklärwerte bei Auslassung der jeweiligen Kovariablen . . . 108
4.32. exponierte Koeffizientenschätzer zu `strasse` und exponierte Splinefunktionen zu `landnutzung` . . . 117
4.33. exponierte Spline-Funktionen im Modell mit negativer Binomial- verteilung 119
4.34. Einfluss der Lokation im Modell mit negativer Binomialverteilung . . . 120
4.35. Quantile des Fits des Modells mit negativer Binomialverteilung . . . 121
4.36. Vergleich des Fits zwischen logistischem Modell und Modell mit negativer Binomialverteilung . . . 122
4.37. Erklärwerte bei Auslassung der jeweiligen Kovariablen . . . 123
4.38. Gewichte für benachbarte Tage aus $N(0, 5)$-Verteilung bei Schätzung der zeitlichen Intensität . . . 127
4.39. geschätzte zeitliche Intensität der Rehunfälle in den Jahren 2006 und 2009 129

A. Nachweis approximativer Äquivalenz von räumlichem Poissonprozess mit Dummypunkten auf den Straßen und binärem (logistischem) Modell

Es reicht zu zeigen, dass unter den auf Seite 98 genannten Bedingungen die approximierte Likelihood des Poissonprozesses und die Likelihood des logistischen Regressionsmodells approximativ dieselbe Form haben, wenn man davon absieht, dass der Intercept bei dem räumlichen Poissonprozess immer größer und der beim logistischen Regressionmodell immer kleiner wird. Das allgemeine Niveau der Prognose wird ohnehin im Nachhinein durch die Normierung angepasst.

Die approximierte log-Likelihood $l_{\text{approxpois}}(\boldsymbol{\beta})$ für einen räumlichen Poissonprozess aus dem Berman-Turner-Device hat nach Abschnitt 2.3 die Form:

$$\begin{aligned} l_{\text{approxpois}}(\boldsymbol{\beta}) &= \sum_{j=1}^{M} w_j \left\{ \frac{N_j}{w_j} \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta}) \right\} \\ &= \sum_{j=1}^{M} N_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - \sum_{j=1}^{M} w_j \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta}) \end{aligned} \tag{A.1}$$

Für eine große Fallzahl n ist das approximierte Integral $\sum_{j=1}^{M} w_j \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta})$ nach der Eigenschaft von räumlichen Punktprozessen, dass die Fläche unter der Intensitätsfunktion der erwarteten Anzahl der Punkte entspricht ungefähr gleich n. Das bedeutet, dass,

wenn die Kachelauflösung A gleichzeitig gegen Null geht, d.h. $A \to 0$ und damit $w_j \to 0$ für alle j der Intercept β_0 gegen unendlich gehen muss, vorausgesetzt die restlichen Koeffizienten ändern sich nicht. Um die Abhängigkeit des Intercepts von der Kachelauflösung zu betonen wird dieser im Weiteren mit β_{0A} bezeichnet. $\boldsymbol{\beta}^*$ bzw. $\boldsymbol{Q}^*(s_j)$ seien außerdem der Koeffizienten- bzw. Kovariablenvektor ohne den Intercept. Es gilt also:

$$l_{\text{approxpois}}(\boldsymbol{\beta}) \approx \sum_{j=1}^{M} N_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - n = \sum_{j=1}^{M} N_j \{\beta_{0A} + \boldsymbol{Q}^*(s_j)^\top \boldsymbol{\beta}^*\} + C. \tag{A.2}$$

N_j ist gemäß der Konstruktion des Berman-Turner-Devices gleich 1, wenn s_j ein Datenpunkt und gleich 0, wenn s_j ein Dummypunkt ist. Damit entspricht N_j gerade der Zielvariable Y_j in der logistischen Regression, wenn die Dummypunkte als Straßenpunkte gewählt werden, also

$$l_{\text{approxpois}}(\boldsymbol{\beta}) \approx \sum_{j=1}^{M} Y_j \beta_{0A} + \sum_{j=1}^{M} Y_j \boldsymbol{Q}^*(s_j)^\top \boldsymbol{\beta}^* + C. \tag{A.3}$$

Die log-Likelihood in der logistischen Regression ist gegeben durch:

$$l_{\text{logist}}(\boldsymbol{\beta}) = \sum_{j=1}^{M} Y_j \log(p_j) + (1 - Y_j) \log(1 - p_j). \tag{A.4}$$

Wenn die Anzahl der Straßenpunkte gegen unendlich geht, folgt $p_j \to 0$ und daraus $\log(1 - p_j) \approx -p_j$. Außerdem gilt für großes n analog zu der Argumentation bei dem räumlichen Poissonprozess: $\sum_{j=1}^{M} p_j \approx n$. Setzt man nun die Responsefunktion aus der logistischen Regression $h(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$ in (A.4) ein, lassen sich damit folgende Approximationen durchführen:

$$\begin{aligned}
l_{\text{logist}}(\boldsymbol{\beta}) &= \sum_{j=1}^{M} Y_j \log\left(\frac{\exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta})}\right) - (1 - Y_j) \log(1 + \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta})) \\
&= \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{Q}(s_j)^\top \boldsymbol{\beta})) = \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} + \log(1 - p_j) \\
&\approx \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - p_j = \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - \sum_{j=1}^{M} p_j \\
&\approx \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} - n = \sum_{j=1}^{M} Y_j \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta} + C
\end{aligned} \tag{A.5}$$

Da $p_j \to 0$, gilt $p_j/(1-p_j) \approx p_j$, also $\log(p_j) \approx \log[p_j/(1-p_j] = \boldsymbol{Q}(s_j)^\top \boldsymbol{\beta}$. Damit muss außerdem β_0 gegen minus unendlich gehen. β_0 wird daher mit β_{0M} bezeichnet und $\boldsymbol{\beta}^*$ und $\boldsymbol{Q}^*(s_j)$ seien wie oben. In dieser Notation lautet die approximative Likelihood:

$$l_{\text{logist}}(\boldsymbol{\beta}) \approx \sum_{j=1}^{M} Y_j \beta_{0N} + \sum_{j=1}^{M} Y_j \boldsymbol{Q}^*(s_j)^\top \boldsymbol{\beta}^* + C. \tag{A.6}$$

Bis auf die divergenten Intercepttterme haben also (A.3) und (A.6) unter den genannten Bedingungen dieselbe Form, sodass sich approximativ gleiche Parameterschätzer für die Kovariablen und nach Multiplikation mit Integrationsfläche bzw. Anzahl Punkten und Teilen durch die Straßenlänge dieselben Prädiktionen ergeben. Die Grenzfälle existieren allerdings aufgrund der Divergenz der Intercepts nicht. Man beachte, dass auch obiges Ergebnis resultieren würde, wenn man die Logarithmusfunktion als Linkfunktion verwenden würde: Setzt man die Approximation $\log(1-p_j) \approx -p_j$ bereits in (A.4) ein und bedenkt, dass da $p_j \to 0 \quad \sum_{j=1}^{M} Y_j p_j$ ungefähr gleich Null ist, gelangt man direkt zu (A.6).

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, 15. November 2011 _____

Roman Hornung