Räumliche Logit-Modelle der individuellen Verkehrsmittelwahl mit Berücksichtigung von Clustereffekten

Claudia Czado[1]
Sergiy Prokopenko[1]
Thomas W. Zängler[2]

Technische Universität München
[1] Fakultät für Mathematik
[2] Fakultät für Wirtschaftswissenschaften

1. Einführung

Im Zusammenhang mit der wachsenden Verkehrsnachfrage in städtischen Gebieten ist es wichtig das Verkehrsmittelwahlverhalten von Individuen zu verstehen, um adäquate Fördermaßnahmen für den öffentlichen Verkehr entwerfen zu können. Eine Möglichkeit das individuelle Verkehrsmittelwahlverhalten zu untersuchen ist durch die Durchführung von Studien zur Mikroanalyse des Mobilitätsverhaltens mittels Mobilitätstagebüchern gegeben. Bei der Erstellung von Mobilitätsbüchern ist darauf zu achten, dass eine Vielzahl von Merkmalen der zurückgelegten Wege in Betracht gezogen wird. Insbesondere sollten dabei haushaltsbezogene, personenbezogene und wegspezifische Faktoren berücksichtigt werden. Es stellt sich dabei auch die verkehrswissenschaftliche Frage:

> Gibt es räumliche Strukturen, in denen nach Berücksichtigung aller bekannten Einflussfaktoren, der öffentliche Verkehr über- oder unterdurchschnittlich benutzt wird?

Mobilitätstagebücher können zur Beantwortung dieser Fragestellung herangezogen werden, da diese häufig Information über räumliche Strukturen, wie Verkehrszellen,

PLZ-Bezirke, städtische Strukturen, Bezirke und Bundesländer enthalten. In diesem Manuskript beschränken wir uns auf Wege, die im öffentlichen Verkehr (ÖV) oder im motorisierten Individualverkehr (IV) zurückgelegt wurden. Damit ist es notwendig, statistische Modelle zu erstellen und anzupassen, die eine binäre (ÖV oder IV) Zielvariable enthalten, unter Berücksichtigung von

- Einflussfaktoren (Kovariablen),
- räumlicher Information aggregiert über räumliche Einheiten und
- zusätzliche Heterogenität (Clustereffekte).

Zusätzliche Heterogenität entsteht dann, wenn wichtige Einflussfaktoren auf die Zielvariable vernachlässigt werden, was bei der Komplexität des individuellen Verkehrsmittelwahlverhaltens nicht auszuschließen ist.

Das Manuskript ist wie folgt gegliedert: Im Abschnitt 2 werden die theoretischen Grundlagen für binäre Regressionsmodelle gelegt und im Abschnitt 3 auf räumliche Logit-Modelle mit Clustereffekten erweitert. Wegen der Komplexität der statistischen Modelle können Standardschätzverfahren wie Maximum Likelihood nicht mehr durchgeführt werden und moderne computerintensive Verfahren wie Markov Chain Monte Carlo (MCMC) Verfahren kommen zum Einsatz. Diese wurden im Rahmen der Dissertation von S. Prokopenko (Prokopenko, 2004) entwickelt und implementiert. Im Abschnitt 4 werden diese Verfahren zur Inferenz auf Münchener Mobilitätsdaten angewandt, die von der Infratest Verkehrsforschung und der TU München erhoben wurden (Zängler 2000). Insbesondere wurde nach der Modellwahl das gewählte Modell, das sowohl nichtlineare Regressionseffekte und Wechselwirkungen, als auch räumliche und Clustereffekte enthält, interpretiert. Abschließend werden die Ergebnisse zusammengefasst und einen Ausblick auf anstehende wissenschaftliche Erweiterungen gegeben.

2. Binäre Regressionsmodelle

Um die Abhängigkeit der individuellen Verkehrsmittelwahlentscheidung zwischen Individualverkehr (IV) und öffentlichen Verkehr (ÖV) von haushaltsbezogenen, personenbezogenen und wegspezifischen Faktoren zu untersuchen, kann man binäre Regressionsmodelle wie das Logit- oder das Probit-Modell verwenden. Eine Einführung in diese Modelle gibt z.B. das Buch von Kleinbaum (2002), während Collett (2002) eine umfassendere Darstellung enthält.

Für das Logit-Modell ist die Wahrscheinlichkeit p_i den i-ten Weg mit IV zurückzulegen gegeben durch

$$p_i = \frac{\exp\left(x_i^t \alpha\right)}{1 + \exp\left(x_i^t \alpha\right)} \qquad (1),$$

wobei $x_i^t \alpha = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_p x_{ip}$. Hierbei sind $\alpha_0, \alpha_1, \ldots, \alpha_p$ zu schätzende Regressionsparameter und $x_{i1}, \ldots, x_{ip}$ sind die zu dem i-ten Weg zugehörigen Kovariablen. Für das Probit-Modell ersetzt man (1) durch

$$p_i = \phi\left(x_i^t \alpha\right) \quad (2),$$

mit $\phi(\theta)$ als die Verteilungsfunktion einer standard-normalverteilten Zufallsvariable. In diesem Manuskript beschränken wir uns auf Logit-Modelle, da es nur für große Datensätze möglich ist, zwischen (1) und (2) zu unterscheiden. Für Logit-Modelle betrachtet man häufig auch das logarithmierte Chancenverhältnis (log odds), welches durch

$$\theta_i = \log\left(\frac{p_i}{1 - p_i}\right) = x_i^t \alpha \qquad (3)$$

definiert ist. Insbesondere ist das logarithmierte Chancenverhältnis linear in den Regressionsparametern. Bei der Wahl der Kovariablen kann man sowohl nichtline-

are Funktionen der Einflussfaktoren als auch Wechselwirkungen verwenden. Im nächsten Abschnitt wird das Logit-Modell (1) um räumliche und Clustereffekte erweitert.

3. Räumliche Logit-Modelle mit Clustereffekten

3.1 Räumliche Logit-Modelle

Um die räumliche Information, die in den Postleitzahlen (PLZ) enthalten sein könnte, zu verwenden, wird das Logit-Modell (1) mit Hilfe der logarithmierten Chancenverhältnisse erweitert:

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right) = x_i^t \alpha + b_{j(i)},$$

wobei $j(i)$ der PLZ-Bezirk der Person ist, die den i-ten Weg unternimmt. Für jeden der J PLZ-Bezirke wird ein räumlicher Effekt b_j eingeführt.

Ein häufig verwendetes statistisches Modell für räumlich aggregierte Daten ist das intrinsic conditional autoregressive (CAR) Modell von Besag und Green (1993). Dieses modelliert die Verteilung eines räumlichen Effektes, gegeben alle anderen räumlichen Effekte als Normalverteilung. Hierbei sind Erwartungswert und Varianz allein bestimmt durch die räumlichen Nachbarn. Insbesondere gilt:

$$b_j \left| b_{j'}, j' \neq j \sim N\left(\frac{1}{N_j}\sum_{j \sim j} b_{j'}, \frac{\tau^2}{N_j}\right)\right.$$

$j \sim j \Leftrightarrow j$ und j' sind benachbart, $N_j =$ Anzahl der Nachbarbezirke von PLZ j

Hier bedeutet "$x \sim N(\mu, \sigma^2)$", dass x normalverteilt ist mit Erwartungswert μ und Varianz σ^2.

Es ist jedoch zu bemerken, dass die gemeinsame Verteilung der räumlichen Effekte im CAR Modell nicht existiert. Daneben bestimmt τ^2 sowohl die bedingte Varianz als auch die räumliche Abhängigkeitsstruktur zwischen den räumlichen Effekten. Aus diesen Gründen schlagen Pettitt, Weir und Hart (2002) ein räumliches Modell mit einem zusätzlichen Parameter ϕ vor, das sowohl eine gemeinsame Verteilung der räumlichen Effekte, als auch eine flexiblere räumliche Abhängigkeitsstruktur erlaubt. Prokopenko (2004) erweitert dieses Modell nochmals, um im Grenzfall das intrinsic - CAR Modell zu erlauben. Wir folgen diesem Ansatz hier und verwenden

$$b_j \middle| b_{j'}, j' \neq j \sim N\left(\frac{\phi}{1+|\phi| N_j} \sum_{j \sim j'} b_{j'}, \frac{(1+|\phi|)\tau^2}{1+|\phi| N_j} \right),$$

wobei ϕ die Stärke der räumlichen Abhängigkeit misst. Ferner gilt keine räumliche Abhängigkeit falls $\phi = 0$. Für $\phi \to \infty$ erhält man das intrinsic - CAR Modell.

3.2 Räumliche Logit-Modelle mit Clustereffekten

Zusätzliche Heterogenität, die durch das Weglassen von Einflussfaktoren entstehen kann, kann man durch die Hinzunahme von Clustereffekten modellieren. Dafür seien $c_m, m = 1, \ldots, M$ unabhängige normal verteilte Clustereffekte mit Erwartungswert Null und Clustervarianz τ_c^2 definiert. Die logarithmierten Chancenverhältnisse werden nun nochmals erweitert zu

$$\theta_i = x_i^t \alpha + b_{j(i)} + c_{m(i)},$$

wobei $m(i)$ den zugehörigen Clustereffekt von Weg i bezeichnet.

Insgesamt sind nun in diesem hierarchischen Modell die Regressionsparameter α, die räumlichen Effekte $b_j, j = 1, \ldots, I$ mit Hyperparametern τ^2 und ϕ und die Clustereffekte $c_m, m = 1, \ldots, M$ mit Clustervarianz σ_c^2 zu schätzen. Eine Maximum Likeli-

hood Schätzung müsste durch Maximierung einer hochdimensionalen Integralfunktion erfolgen, was sowohl analytisch als auch numerisch nicht durchführbar ist. Daher verwenden wir hier einen Bayesianischen Ansatz (siehe z.B. Berry 1996) mit Markov Chain Monte Carlo (MCMC) Methoden für die Parameterschätzung. Beim Bayesianischen Ansatz wird Vorwissen mit Hilfe einer a priori Verteilung über die Parameter mit verwendet, das man aus Vorstudien und/ oder Expertenwissen erhalten kann. Das Wissen über die Parameter wird nach dem Beobachten der Daten mit Hilfe der a posteriori Verteilung angepasst. Im Falle von nicht informativen a priori Verteilungen entspricht der Modus der a posteriori Verteilung dem Maximum Likelihood Schätzer. In komplexen statistischen Modellen kann man jedoch die a posteriori Verteilung nicht bestimmen. Dennoch ermöglichen die MCMC Verfahren approximativ eine beliebig große Stichprobe von der a posteriori Verteilung zu ziehen. Diese kann dann verwendet werden, um den Modus, die Dichte und den Intervallschätzer zu schätzen. Chib (2001) gibt eine Einführung und Zusammenfassung von MCMC Verfahren. Prokopenko (2004) hat für räumliche Logit-Modelle mit und ohne Clustereffekte entsprechende MCMC Algorithmen zur Parameterschätzung entwickelt, getestet und implementiert. Diese Modelle werden nun auf Daten, die das individuelle Verkehrsmittelwahlverhalten untersuchen, angewendet. Diese Anwendung wird im nächsten Kapitel beschrieben.

4. Anwendung auf das individuelle Verkehrsmittelwahlverhalten

4.1 Daten

Die hier verwendeten Daten sind Teil der Studie "Mobilität '97", in der das individuelle Mobilitätsverhalten der bayerischen Wohnbevölkerung abgebildet wird (Zängler 2000). In diesem Beitrag beschränken wir uns auf die Wohnbevölkerung in der Landeshauptstadt München und auf Wege, die im motorisierten Individualverkehr (IV) und im öffentlichen Verkehr (ÖV) zurückgelegt wurden. Wege zu Fuß und mit dem Fahrrad bleiben somit unberücksichtigt. Die verwendete Teilstichprobe enthält 1.375 Wege von 296 Personen aus 167 Haushalten.

Abbildung 1 zeigt die geographische Aufschlüsselung der verwendeten Anzahl der Wege bzgl. der PLZ-Bezirke.

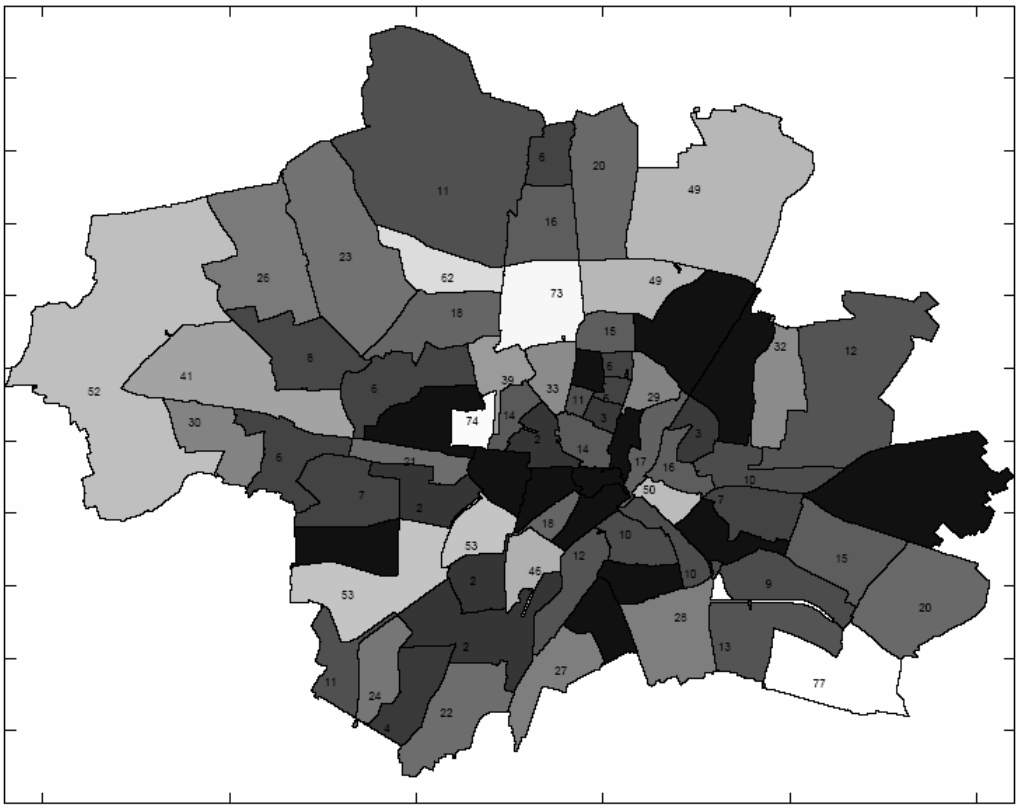

Abb.1: Anzahl der Wege nach PLZ-Bezirken in München

Hier interessiert uns besonders der räumliche Effekt der PLZ-Bezirke unter Berücksichtigung von weiteren Wegmerkmalen, um Gebiete mit über- bzw. unterdurchschnittlicher IV-Benutzung aufzudecken.

Dafür untersuchen wir zunächst, welche Faktoren (Kovariablen) Einfluss auf das individuelle Verkehrsmittelwahlverhalten haben, das durch die binäre Zielvariable

$$y_i = \begin{cases} 1 & \text{Weg i mit IV} \\ 0 & \text{Weg i mit ÖV} \end{cases} \qquad i = 1,\ldots,1.375$$

charakterisiert wird.

Folgende mögliche Kovariablen wurden verwendet:

- Personenbezogene Variablen: Einkommen, Pkw-Zugang, Besitz einer ÖV-Netzkarte, Geschlecht und Alter,
- Haushaltsbezogene Variable: Haushaltstyp,
- Wegbezogene Variablen: Wochenende (ja/nein), Entfernung, Weg allein/gemeinsam zurückgelegt und Tageszeit.

Eine genaue Aufschlüsselung der Daten bzgl. der Kovariablen ist in den Tabellen 1 und 2 gegeben.

Wir möchten anmerken, dass in den Tabellen 1 und 2 nur die Wegmerkmale aufgeführt sind, die sich als statistisch signifikant in den verwendeten Modellen erwiesen haben. Eine ausführliche Datenbeschreibung mit allen gemessenen Wegmerkmalen findet man in Zängler (2000).

Tabelle 1: Zurückgelegte Wege klassifiziert bzgl. personen- und haushaltsbezogenen Wegemerkmalen

Covariable type	Variable	Levels	Number of Trips Using			Most frequently obs. value
			Individual Transport	Public Transport	Total	
PERSON related	PERSONAL INCOME	*NO INCOME (< 200 DM)*	24	31	55	0
		MIDDLE (200 – 3000 DM)	475	193	668	1
		HIGH (> 3000 DM)	521	131	652	0
	USAGE	*MAIN USER*	731	100	831	1
		SECONDARY USER	213	99	312	0
		NOT USER	76	156	232	0
	NET CARD	*YES*	235	247	482	0
		NO	785	108	893	1
	SEX	*MALE*	549	172	721	1
		FEMALE	471	183	654	0
	AGE	metric (quadratic, normalized with Splus function *poly(age,2)*)				median 42 years
HOUSE-HOLD related	HOUSEHOLD TYPE	*SINGLE*	156	125	281	0
		SINGLE PARENT	84	10	94	0
		NOT SINGLE	780	220	1000	1

Tabelle 2: Zurückgelegte Wege klassifiziert bzgl. wegbezogenen Merkmalen

Covariable type	Variable	Levels	Number of Trips Using			Most frequently obs. value
			Individual Transport	Public Transport	Total	
TRIP related	DAY TYPE	*WORK DAY*	595	297	892	1
		WEEKEND	425	58	483	0
	DISTANCE	*SHORT* (≤ 3.5 *km*)	294	71	365	0
		MIDDLE ($3.6 - 21.5$ *km*)	571	257	828	1
		FAR (> 21.5 *km*)	155	27	182	0
	WAY ALONE	*ALONE*	507	267	774	1
		NOT ALONE	513	88	601	0
	DAY TIME	*DAY (6 a.m. - 9 p.m.)*	905	336	1241	1
		NIGHT (9 p.m. - 6 a.m.)	115	19	134	0
TOTAL			1020	355	1375	

4.2 Statistische Modelle

Es werden nun zwei statistische Modelle untersucht, eines nur mit räumlichen Effekten (Model 1) und das andere mit sowohl räumlichen als auch Clustereffekten (Model 2). Dies ist nur eine kleine Auswahl von Modellen, die von Prokopenko (2004) untersucht worden sind. Jedoch haben sich diese beiden Modelle als die geeignetsten herausgestellt (siehe auch Abschnitt 4.3). Wir verwenden:

Modell 1: Logit Modell mit räumlichen Effekten

$$\theta_i := \text{logit}(p_i) = x_i^t \alpha + b_{j(i)}$$

Modell 2: Logit Modell mit räumlichen und Clustereffekten

$$\theta_i := \text{logit}(p_i) = x_i^t \alpha + b_{j(i)} + c_{m(i)}$$

Bei den Regressionseffekten α wurden sowohl nichtlineare (beim Alter) als auch Wechselwirkungen zu gelassen. Genaue Spezifizierungen sind im Kapitel 6 von Prokopenko (2004) zu finden. Für den Clustereffekt im Modell 2 wurde die Anzahl der Wege eines Haushalts in 5 Gruppen aufgeteilt, eine genaue Definition ist in Tabelle 3 gegeben.

Tabelle 3: Definition der Clustereffekte für Modell 2.

Cluster	Description		Total
1st	Households which conducted	≥ 23 trips	275 trips
2nd	Households which conducted	16 – 22 trips	296 trips
3rd	Households which conducted	12 – 15 trips	250 trips
4th	Households which conducted	8 – 11 trips	275 trips
5th	Households which conducted	≤ 7 trips	279 trips

Für beide Modelle wurden 25.000 MCMC Iterationen nach einer Anfangsphase (burnin) der MCMC Algorithmen von Prokopenko (2004) durchgeführt und jede 25. Iteration wurde für die Parameterschätzung verwendet. Dies bedeutet, dass die entsprechende a posteriori Verteilung mit einer Stichprobe von 1.000 Beobachtungen geschätzt wurde.

4.3 Modellvergleich

Als Maß für die Anpassungsgüte werden verwendet:

- ein nach räumlichen Strukturen gewichtetes quadratisches Abstandsmaß zwischen den beobachteten Anteil minus den durch das Modell geschätzten Anteil der IV-Wege innerhalb eines PLZ-Bezirkes,

- das Bayesianische Devianz Informationskriterium (Spiegelhalter et. al. 2002), welches ein Analog zum Akaike Informationskriterium (Akaike 1977) ist.

Das gewichtete quadratische Anstandsmaß, bezeichnet mit A, betrachtet die Anpassung auf der PLZ-Bezirksebene allein, während das Devianz Informationskriterium, bezeichnet mit DIC, auch einen Strafterm für die Komplexität des verwendeten statistische Modells beinhaltet. Das Abstandsmaß A wurde gewählt, da es in diesem Anwendungsproblem besonders um die Anpassung der räumlichen Strukturen geht.

Eine präzise Skala für die Bewertung von A bzw. DIC existiert nicht, aber Modelle mit kleineren Werten sind Modellen mit größeren Werten vorzuziehen.

Das quadratische Abstandsmaß A ist wie folgt definiert:

$$A := \sum_{j=1}^{74} n_j \left(p_j^{empir} - p_j^{estim} \right)^2 ,$$

wobei n_j die Anzahl der Wege im PLZ j, p_j^{empir} der beobachtete Anteil der IV-Wege im PLZ j und p_j^{estim} der durch das Modell geschätzte Anteil ist. Dieser wird durch

$$p_j^{estim} = \frac{1}{n_j R} \sum_{i:j(i)=j} \sum_{r=1}^{R} \frac{\exp\left(\theta_i^r\right)}{1+\exp\left(\theta_i^r\right)}$$

ermittelt, wobei θ_i^r das mit Hilfe der $r-$ten approximativen Beobachtung von der a posteriori Verteilung geschätzten logarithmierten Chancenverhältnis ist, d.h. θ_i^r ist durch

$$\theta_i^r = \begin{cases} x_i^t \alpha^r + b_{j(i)}^r & \text{Model 1} \\ \\ x_i^t \alpha^r + b_{j(i)}^r + c_{m(i)}^r & \text{Model 2} \end{cases}$$

gegeben. Hier bezeichnen α^r, b^r und c^r die $r-$ten Beobachtungen der gemeinsamen a posteriori Verteilung für α, b und c.

Wie bereits bemerkt, besitzt das DIC die zwei Komponenten $\overline{D(\theta)}$ zur Modellanpassung und p_D zur Modellkomplexität. Die Modellanpassung erfolgt über die Devianz, ein häufig verwendetes Anpassungsmaß für generalisierte lineare Modelle (GLM), zu denen die Logit-Modelle gehören. GLM's und deren Devianz werden im Buch von McCullagh und Nelder (1989) behandelt. Die Modellkomplexität wird durch die Anzahl der effektiven Parameter p_D geschätzt. Diese ergibt sich wiederum durch die Differenz zwischen durchschnittlicher Devianz minus der Devianz, ausgewertet an den entsprechenden Mittelwerten der beobachteten MCMC Stichprobe der a posteriori Verteilung. Mathematisch ausgedrückt erhält man damit

$$DIC = \overline{D(\theta)} + p_D ,$$

wobei

$$\overline{D(\theta)} = \frac{1}{R}\sum_{r=1}^{R} D\left(\theta^r\right) \text{ mit } \theta^r = \left(\theta_1^r \ldots \theta_n^r\right)$$

und

$$D\left(\theta^r\right) = -2\sum_{i=1}^{n}\left[y_i\theta_i^r + \log\left(1+\exp\left\{\theta_i^r\right\}\right)\right]$$

Schließlich ergibt sich für die Anzahl der effektiven Parameter

$$p_D = \overline{D(\theta)} - D\left(\overline{\theta}\right) \text{ mit } \overline{\theta} = \left(\overline{\theta_1}, \ldots, \overline{\theta_n}\right) ,$$

wobei

$$\overline{\theta_i} = \begin{cases} x_i^t\overline{\alpha} + \overline{b}_{j(i)} & \text{Model 1} \\ \\ x_i^t\overline{\alpha} + \overline{b}_{j(i)} + \overline{c}_{m(i)} & \text{Model 2} \end{cases}$$

und

$$\bar{\alpha} = \frac{1}{R}\sum_{r=1}^{R}\alpha^r, \ \bar{b}_j = \frac{1}{R}\sum_{r=1}^{R}b_j^r \text{ und } \bar{c}_m = \frac{1}{R}\sum_{r=1}^{R}c_m^r$$

für $j = 1,\ldots,J$ und $m = 1,\ldots,M$.

Die sich für diesen Datensatz ergebenden Werte sind in Tabelle 4 zusammengefasst. Danach ist Modell 2 dem Modell 1 vorzuziehen. Dies bedeutet, dass die Hinzunahme des Clustereffektes die Anpassung verbessert.

Tabelle 4: Anpassungsgüte von Modell 1 und Modell 2 für die Verkehrsmittelwahlverhaltensdaten.

Model	1	2
D	2.35	0.95
DIC	845.76	772.48
p_D	58.81	67.26
$\overline{D}$	786.95	705.22

4.4 Modellinterpretation

Wir interpretieren nun Modell 2 bzgl. des Einflusses von Wegmerkmalen, räumlichen und Clustereffekten auf die dichotome Variable zum Verkehrmittelwahlverhalten (IV vs. ÖV). Da Modell 2 sowohl räumliche als auch Clustereffekte enthält, basieren die geschätzten Wahrscheinlichkeiten für IV Wege neben dem Einfluss der Wegemerkmale ebenfalls auf räumlichen und Clustereffekten. Wir wählen beispielhaft den PLZ-Bezirk 81377 und den 5. Clustereffekt (Haushalte mit höchstens 7 zurückgelegten Wegen) aus.

Bei der Schätzung der Wahrscheinlichkeit, IV in Abhängigkeit von bestimmten Wegmerkmalen zu benutzen, wird bei allen anderen Einflussfaktoren die häufigste Ausprägung verwendet (siehe Tabellen 2 und 3). Abbildung 2 zeigt die geschätzten

Wahrscheinlichkeiten für IV in Abhängigkeit vom Alter eines männlichen KFZ Hauptnutzers mit mittleren Einkommen und keiner ÖV Netzkarte. Er lebt im PLZ-Bezirk 81377 in einem Nicht-Singlehaushalt mit höchstens 7 zurückgelegten Wegen. Dabei wird ein allein zurückgelegter Weg an einem Werktag (Mo-Fr) und während des Tages (6:00 bis 19:00 Uhr) mit mittlerer Distanz zugrunde gelegt. In diesem Fall sieht man, dass die Wahrscheinlichkeit einer IV Verwendung nicht linear von Alter abhängt: die Wahrscheinlichkeit steigt bis zu einem Alter von 58 Jahren und fällt dann wieder. Die gestrichelten Linien in der Abbildung geben 90% punktweise Bayesianische Konfidenzintervalle an. Die Festsetzung von anderen Ausprägungen der Wegmerkmale verändert nur additiv die Struktur auf der Ebene der logarithmierten Chancenverhältnisse.

Wir wenden uns nun der Interpretation von Wechselwirkungen zu. Es werden 2 Wechselwirkungen gezeigt. In Abb. 3 wird die Wechselwirkung auf IV Benutzung von Geschlecht und KFZ-Nutzungsgrad gezeigt.

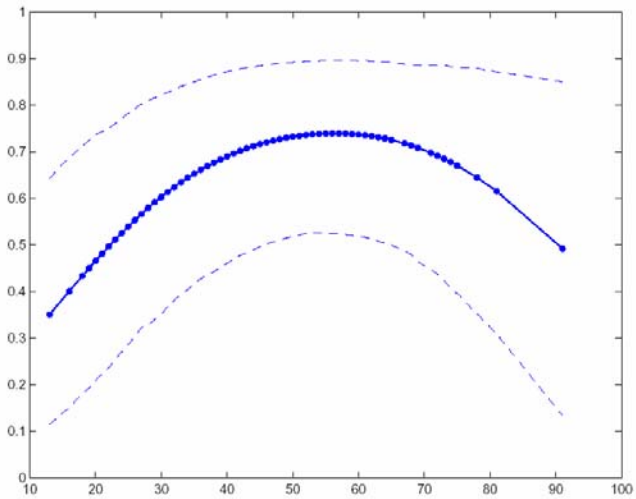

Abb. 2: Geschätzte Wahrscheinlichkeiten für IV nach Alter (PLZ 81377, Haushalt ≤ 7 Wege, alle andere Wegmerkmale haben die häufigste Ausprägung)

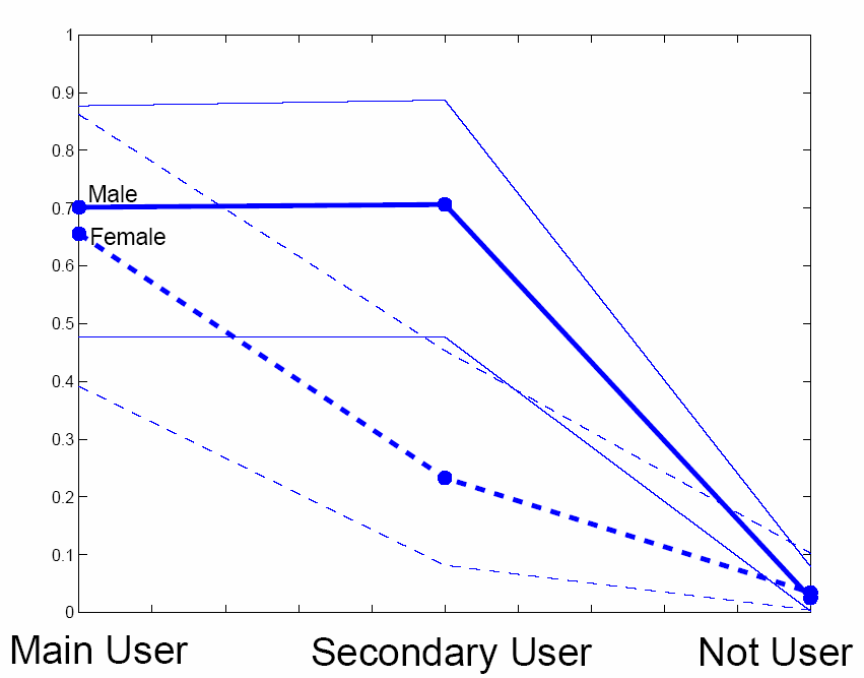

Abb. 3: Geschätzte Wahrscheinlichkeiten für IV nach Geschlecht und KFZ-Nutzungsgrad (PLZ 81377, Haushalt $\leq$ 7 Wege, alle anderen Wegmerkmale haben die häufigste Ausprägung)

Hier zeigt sich, dass Männer, die gemeinsam mit einer oder mehreren anderen Personen einen Pkw in Absprache benutzen, den IV wesentlich häufiger als Frauen benutzen, auf die das gleiche Merkmal zutrifft.

Als zweites Beispiel besprechen wir nun die Wechselwirkung von Geschlecht und Tageszeit auf die Wahrscheinlichkeit der IV Benutzung (vgl. Abb. 4). Es zeigt sich, dass während des Tages Frauen und Männer in etwa gleich häufig den IV wählen. Im Gegensatz dazu wählen Frauen während der Nacht den IV wesentlich häufiger als Männer.

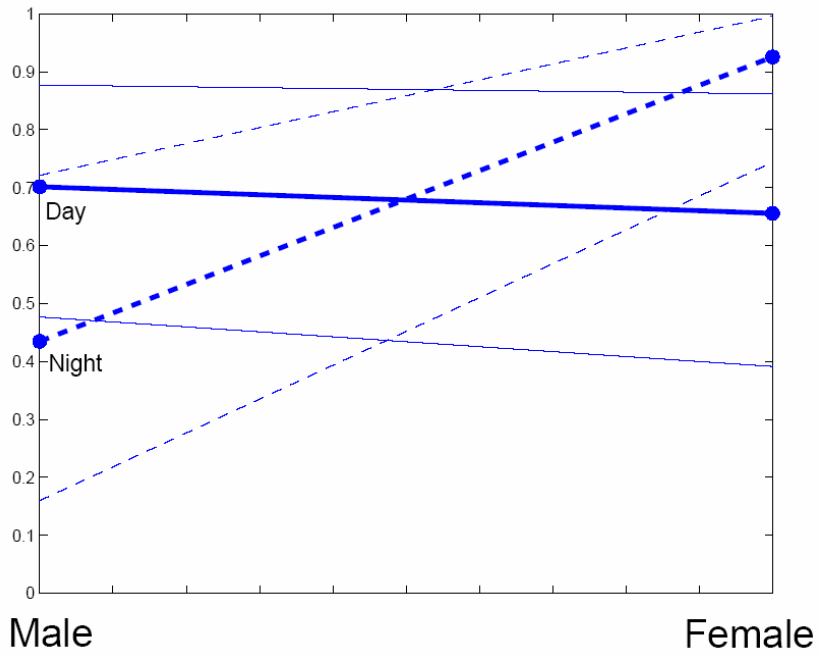

Abb. 4: Geschätzte Wahrscheinlichkeiten für IV nach Geschlecht und Tageszeit (PLZ 81377, Haushalt ≤ 7 Wege, alle andere Wegmerkmale haben die häufigste Ausprägung)

Prokopenko (2004) zeigt, dass noch weitere statistisch signifikante Wechselwirkungen zwischen folgenden Merkmalen bestehen:

- Netzkartenbesitz und Tagestyp (Wochentags/Wochenende),
- Netzkartenbesitz und Weg allein oder gemeinsam unternommen,
- Benutzungsgrad und Weg allein oder gemeinsam unternommen,
- Benutzungsgrad und Tageszeit,
- Benutzungsgrad und Distanz.

Deren Interpretation wird in Prokopenko (2004) im Abschnitt 6.4 ausgeführt.

Die Clustereffekte können wie folgt beschrieben werden. Es besteht die Tendenz, dass die Clustereffekte negativ werden, wenn sich die Clusternummer erhöht. Dies

impliziert eine niedrige Verwendung von IV in Haushalten mit weniger zurückgelegten Wegen. Details findet man in Abschnitt 6.4 von Prokopenko (2004).

Nun kommen wir zu unserer Anfangsfragestellung zurück:

Gibt es PLZ-Bezirke, in denen nach Berücksichtigung von Wegmerkmalen IV über- bzw. unterdurchschnittlich verwendet wird?

Um diese Frage zu beantworten, verwenden wir Bayesianische Vertrauensintervalle (credible intervals) für die räumlichen Effekte. Ein solches $100(1-\alpha)$% Bayesianische Vertrauensintervall für die räumlichen Effekte ergibt sich als Intervall zwischen den empirischen 100 x $\alpha/2$% Quantil und 100 x $(1-\alpha/2)$% Quantil der a posteriori Verteilung der entsprechenden räumlichen Effekte. Damit klassifizieren wir die PLZ-Bezirke wie folgt:

Falls das 90% Bayesianische Vertrauensintervall für $b_{j(i)}$

- 0 enthält, ist die IV-Benutzung im PLZ-Bezirk j durchschnittlich,
- größer als 0 ist, ist die IV-Benutzung im PLZ-Bezirk j überdurchschnittlich,
- kleiner als 0 ist, ist die IV-Benutzung im PLZ-Bezirk j unterdurchschnittlich.

Bei unserem Datensatz können wir damit 7 (17) PLZ-Bezirke als Bezirke mit überdurchschnittlicher (unterdurchschnittlicher) IV-Benutzung charakterisieren. Es ist natürlich denkbar, dass dieser über- bzw. unterdurchschnittliche Benutzungsgrad IV von der räumlichen Verteilung der U- und S-Bahnhaltestellen erklärt werden kann.

Tabelle 5: Zusammenhang zwischen über- bzw. unterdurchschnittlicher IV-Benutzung und räumlicher Verteilung von U- und S-Bahnhaltestellen in München.

	With U- or S-stops	**Without U- or S-stops**
90% *CI* over 0	2	5
90% *CI* below 0	11	6

In Tabelle 5 wird ein derartiger möglicher Zusammenhang untersucht. Dabei zeigt sich, dass der IV-Benutzungsgrad nicht allein durch das U- und S-Bahnnetz erklärt werden kann, da es 2+6 PLZ-Bezirke gibt, deren IV-Benutzungsgrad nicht durch die räumliche Verteilung des U- und S-Bahnnetz erklärt werden kann.

Abb. 5: PLZ-Bezirke und U-/ S-Bahn-Netz in München (weiß = überdurchschnittliche IV-Benutzung, dunkelgrau = unterdurchschnittliche IV-Benutzung)

In Abb. 5 werden die PLZ-Bezirke von München zusammen mit dem U-/S-Bahn-Netz gezeigt. Damit können einige auffällige PLZ-Bezirke von Tabelle 5 durch das Vorhandensein der U- bzw. S-Bahnhaltestellen auf den PLZ-Grenzen erklärt werden. Für das abweichende Verkehrsmittelwahlverhalten die übrigen PLZ-Bezirken konnte mit dem vorgestellten Verfahren keine Erklärung gefunden werden.

5. Zusammenfassung und Ausblick

In diesem Manuskript wurde eine allgemeine Klasse von statistischen Modellen für binäre Regressionsdaten mit räumlichen und Clustereffekten eingeführt und auf Daten zum Mobilitätsverhalten der Münchener angewandt. Danach konnte ein nichtlinearer Effekt des Alters und Wechselwirkungen, wie zum Beispiel zwischen Geschlecht und KFZ-Benutzungsgrad auf die Wahrscheinlichkeit IV zu benutzen, aufgedeckt werden. Es stellte sich auch heraus, dass PLZ-Bezirke existieren, in denen der Individualverkehr über- oder unterdurchschnittlich trotz Berücksichtigung von haushalts-, personen- und wegbezogenen Merkmalen und der räumlichen Verteilung von Münchener U- und S-Bahnhaltestellen. Andere Anwendungen im Bereich von Mikroverkehrsstudien sind möglich.

Es ist zu bemerken, dass diese Klasse von Modellen für die Parameterschätzung computerintensive MCMC Verfahren benötigt, die nicht mit Standardsoftware angepasst werden können. Diese entwickelte S. Prokopenko in seiner Dissertation (Prokopenko, 2004), sie wurden auch für diesen Beitrag eingesetzt. Er entwickelte diese Modelle darüber hinaus auf Probit-Modelle und individuelle Clustereffekte innerhalb eines Clusters. Für die Zukunft ist die Erweiterung auf nominale Zielvariablen, damit man auch Fuß- und Radwege berücksichtigen kann, geplant. Für die Münchener Teilstichprobe von „Mobilität '97“ hat die Sekundäranalyse von Ehrlich (2002) bereits nominelle Regressionsmodelle ohne räumliche und Clustereffekte untersucht.

Danksagung:
Die Autoren möchten sich bei allen Personen und Institutionen bedanken, die diesen Beitrag ermöglicht haben. Insbesondere dankt C. Czado der Deutschen Forschungsgemeinschaft (DFG) für die Unterstützung durch den SFB 387 „Statistische Analyse diskreter Strukturen“, S. Prokopenko für die Förderung durch das Graduiertenkolleg „Angewandte Algorithmische Mathematik“ an der TU München und T.W. Zängler der BMW AG, München, für die Finanzierung der Datenerhebung und die weitere Datennutzung für sekundäranalytische Zwecke.

Literatur

- Akaike, H. (1977) An entropy maximization principle, Applications of statistics (Krishnaiah, P.R. ed.) North Holland, Amsterdam, 27-41.

- Berry, D.A. (1996) Statistics - A Bayesian Perspective, Boston: Duxbury Press.

- Besag, J. und P. Green (1993) Spatial statistics and Bayesian computation (with discussion). J. R. Statist. Soc B, 55, 25-37.

- Chib, S. (2001) Markov Chain Monte Carlo methods: Computation and Inference. In Handbook of Econometrics, Vol. 5, 381-394, North Holland.

- Collett, D. (2002) Modelling Binary Data, 2nd Edition, London: Chapman & Hall.

- Ehrlich, K. (2002) Regressionsverfahren mit nominaler Zielvariable, Diplomarbeit an der Fakultät Mathematik, Technische Universität München.

- Kleinbaum, D. G. (1994) Logistic Regression: A self-learning Text, Berlin: Springer Verlag.

- McCullagh, P. und J. Nelder (1989). Generalized Linear Models, 2nd Edition. London: Chapman & Hall.

- Pettitt, A., I. Weir und A. Hart (2002) Conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. Statistics and Computing 12, 353-367.

- Prokopenko, S. (2004) Hierarchical binary spatial regression models with cluster effects, Doktorarbeit an der Fakultät Mathematik, Technische Universität München.

- Zängler, T. (2000) Mikroanalyse des Mobilitätsverhaltens in Alltag und Freizeit, Berlin: Springer Verlag.