



# KLEINRÄUMIGE EXTRAPOLATION VON UMFRAGEDATEN

NAMEN:	ALEXANDER LANGE, KAI HUSMANN
MATR. NR.:	21426614, 20707176
STUDIENGANG:	ANGEWANDTE STATISTIK
MAIL:	ALEXANDER.LANGE@UNI-GOETTINGEN.DE KAI.HUSMANN@FORST.UNI-GOETTINGEN.DE
KURS:	STATISTISCHES PRAKTIKUM
KURSLEITER:	PROF.DR. THOMAS KNEIB
LEHRSTUHL:	STATISTIK
FAKULTÄT:	WIRTSCHAFTSWISSENSCHAFTEN
ABGABEDATUM:	30. SEPTEMBER 2016

## INHALTSVERZEICHNIS

<b>1 Einleitung</b>	<b>1</b>
<b>2 Material und Methoden</b>	<b>2</b>
2.1 Daten . . . . .	2
2.1.1 Parametrisierungsstichprobe . . . . .	2
2.1.2 Melderegister . . . . .	7
2.1.3 Zensus . . . . .	8
2.2 Statistische Methoden . . . . .	9
2.2.1 Modell . . . . .	9
2.2.2 Modellwahl . . . . .	10
2.2.3 Konfidenzintervalle . . . . .	11
2.2.4 Kreuzvalidierung . . . . .	12
2.2.5 Validierung . . . . .	12
<b>3 Ergebnisse</b>	<b>12</b>
3.1 Modell . . . . .	12
3.2 Modellwahl . . . . .	12
3.3 Reklassifizierung . . . . .	13
3.4 Kreuzvalidierung . . . . .	14
3.5 Validierung . . . . .	16
3.6 Extrapolation . . . . .	18
<b>4 Diskussion</b>	<b>18</b>
<b>5 Fazit</b>	<b>19</b>
<b>Literatur</b>	<b>21</b>
<b>Anhang</b>	<b>22</b>

## ABBILDUNGSVERZEICHNIS

1	Häufigkeit der Kategorienausprägungen der endogene Variablen in der Parameterisierungsstichprobe . . . . .	3
2	Kontur Plot der absoluten Anzahl der Gruppenbeobachtungen zur Meinung zu Stuttgart 21 in drei Gruppen. Quelle der Hintergrundgrafik: REF: Google Maps . . . . .	4
3	Kontur Plot der absoluten Anzahl der Gruppenbeobachtungen zur Wohnzufriedenheit in fünf Gruppen. Quelle der Hintergrundgrafik: REF: Google Maps . . . . .	5
4	Anteile zur Meinung zu Stuttgart 21 nach Stadtbezirken. . . . .	6
5	Anteile der Meinung zu Stuttgart 21 nach Stadtteilen. . . . .	7
6	Illustration geschätzte gegen wahre Anteile für Gauss-Krüger Informationen extrapoliert auf Melderegister mit zwei und drei Klassenmodell mit 95% Quantilen	18
7	Vergleich der extrapolierten Gesamtanteile für Stuttgart mit allen geschätzten Modellen und beiden Extrapolationsdateien mit den wahren Anteilen, sowie den Anteilen der Stichprobe mit 95% Quantilen zur Meinung zu Stuttgart 21 . . . . .	19
8	Vergleich der extrapolierten Gesamtanteile für Stuttgart mit allen geschätzten Modellen und beiden Extrapolationsdateien mit den Anteilen der Stichprobe mit 95% Quantilen zur Bewertung der Wohngegend . . . . .	20
9	Häufigkeit der Kategorienausprägungen der exogenen Variablen in der Parameterisierungsstichprobe. . . . .	22
10	Anteile der Bewertung der Wohngegend nach Stadtbezirken. . . . .	23
11	Anteile der Bewertung der Wohngegend nach Stadtteilen. . . . .	24

## TABELLENVERZEICHNIS

1	Erhobene sozioökonomische und geographische Variablen der Parameterisierungsstichprobe und deren Anzahl der Ausprägungen sowie vermutete Modellierung im additiven Modell. . . . .	2
2	Sozioökonomische und geographische Variablen des Melderegisters und deren Anzahl der Ausprägungen. . . . .	8
3	Erhobene sozioökonomische und geographische Variablen der Gebäude- und Wohnungszählung im Rahmen des Zensus und deren Anzahl der Ausprägungen. . . . .	8
4	Vergleich der step AIC Ergebnisse zwischen den Modellen . . . . .	13
5	Reklassifizierung der Meinung zu Stuttgart 21 . . . . .	14
6	Reklassifizierung der Bewertung der Wohngegend . . . . .	14
7	Kreuzvalidierung der Meinung zu Stuttgart 21 nach einzelnen Klassen . . . . .	15
8	Kreuzvalidierung der Bewertung der Wohngegend nach einzelnen Klassen . . . . .	16
9	Vergleich der mittleren quadratischen Abweichung (MSE) und der Überdeckungswahrscheinlichkeit bei allen Prognosen aus den geschätzten Modellen und den beiden Extrapolationsdateien für die Meinung zu Stuttgart 21 . . . . .	17

### EINLEITUNG

Die Grundgesamtheit dieser Unteruchung ist die Bevölkerung Stuttgarts. Fragestellungen: Wie ist die Wohzufriedenheit in Stuttgart? Wie ist die Meinung zu Stuttgart 21? Kleinräumige Extrapolation test

## MATERIAL UND METHODEN

### DATEN

Insgesamt liegen für die Analysen drei Umfragen mit unterschiedlichen Stichprobenumfängen vor. Die kleinste Datei enthält Angaben zur Bewertung der Wohngegend, der Meinung zu Stuttgart 21 sowie weitere sozioökonomische Kovariablen, die zur Erklärung der beiden abhängigen Variablen dienen sollen. Sie wird im Folgenden als Parametrisierungsstichprobe bezeichnet. Die Parametrisierungsumfrage ist eine Stichprobe von der die Grundgesamtheit für eine Validierung nicht zur Verfügung steht. Alle Modellqualitätskriterien müssen demnach entweder an der Stichprobe selbst oder an einer anderen Erhebung entwickelt werden. Die beiden anderen Umfragen haben jeweils einen deutlich größeren Stichprobenumfang. An diesen Umfragen werden die parametrisierten Modelle angewendet und die Meinung zu Stuttgart 21 sowie die Wohnzufriedenheit somit kleinräumig extrapoliert. Einige Variablen unterscheiden sich in ihren Ausprägungen zwischen den Umfragen. Zur Vereinheitlichung der Dateien mussten einige Gruppenausprägungen demnach umkodiert werden. Die Umkodierungen können in der digital anhängenden Datei *Aufbereitung\_Stuttgart21.R* nachvollzogen werden.

### PARAMETRISIERUNGSTICHPROBE

Mit den Datensätzen der Parametrisierungsstichprobe (Tabelle 1) werden die Modelle für die kleinräumige Extrapolation parametrisiert. Bei dieser Umfrage handelt es sich um eine Befragung aus dem Jahr 2015 zur Lebensqualität der Einwohner Stuttgarts bei der unter anderem die Bewertung der Wohnsituation und die Meinung zu Stuttgart 21 abgefragt wurden [Landeshauptstadt Stuttgart, 2015]. Insgesamt standen 8 sozioökonomische Variablen und Angaben zur räumlichen Lage zur Verfügung. Von jedem Datensatz waren stetige räumliche Lage als Gauss-Krüger Geokoordinate sowie die diskrete räumliche Lage im Stadtteil und Stadtbezirk bekannt.

TABELLE 1: ERHOBENE SOZIOÖKONOMISCHE UND GEOGRAPHISCHE VARIABLEN DER PARAMETERISIERUNGSTICHPROBE UND DEREN ANZAHL DER AUSPRÄGUNGEN SOWIE VERMUTETE MODELLIERUNG IM ADDITIVEN MODELL.

Anzahl Beobachtungen: 3.143

Variable	Anzahl Ausprägungen	Modellierung
Bewertung Wohngegend	6	Geordnet Kategorial
Meinung Stuttgart 21	6	Geordnet Kategorial
Personenanzahl im Haushalt	5	Nicht Parametrisch
Monatliches Netto Haushaltseinkommen	6	Nicht Parametrisch
Altersklasse Befragter	6	Nicht Parametrisch
Geschlecht	2	Parametrisch
Familienstand	4	Parametrisch
Nationalität	2	Parametrisch
Stadtbezirk	23	Markov-Zufallsfeld
Stadtteil	142	Markov-Zufallsfeld
Gauß-Krüger		Tensorprodukt-Splines

In Tabelle 1 sind nicht nur die Anzahlen der Ausprägungen der Variablen, sondern auch die vermuteten Formen der Einflüsse der Kovariablen auf die abhängigen Variablen nach visueller Einschätzung aufgelistet. Es ist ersichtlich, dass alle nominal skalierten Variablen, wie z.B. die Nationalität, als parametrisch und dass alle kardinal skalierte Variablen, wie z.B. die Altersklasse des Befragten, als nicht-parametrisch modelliert werden sollten. Laut [Fahrmeir et al., 2009, p.9] sind diese beobachteten Zusammenhänge typisch für eine Regressionsanalyse. In Anlehnung an [Fahrmeir et al., 2013, p. 503 ff. & p. 524 ff.] wird der kontinuierliche räumliche Effekt als Tensor Produkt und die diskreten räumlichen Effekte durch eins Markov-Zufallsfeld **ist es ein MRF oder ein GMRF das sollten wir klären und einheitlich bezeichnen** im additiven Regressionsmodell berücksichtigt.

Für die Auswahl der geeigneten Regressionsmethode und der Ergebnisinterpretation ist es hilfreich, das Verhältnis der Häufigkeiten der Kategorienausprägungen der abhängigen Variable zu kennen und seltene Ereignisse zu identifizieren. Während die meisten befragten Personen ihre Wohngegend mit *gut* oder *sehr gut* bewertet haben, treten Beobachtungen mit *schlechter* oder *sehr schlechter* Einschätzung relativ selten auf (Tabelle 1). Im Vergleich sind die Verhältnisse der Gruppenhäufigkeiten zur Meinung zu Stuttgart 21 ausgeglichener. Die *neutrale* Haltung ist etwa halb so häufig vertreten wie die *zustimmende* Haltung. Bei beiden Variablen wurden die wenigen, für die Modellierung irrelevanten, Kategorien *Keine Angabe* entfernt. Den Variablen kann in bei der Gruppenausprägung eine Rangfolge, jedoch kein Intervall unterstellt werden. Es handelt sich demnach in beiden Fällen um ordinal Skalierte Daten.



ABBILDUNG 1: HÄUFIGKEIT DER KATEGORIENAUSPRÄGUNGEN DER ENDOGENE VARIABLEN IN DER PARAMETERISIERUNGSTICHPROBE.

Das amtliche, nach Stadtteilen oder Stadtbezirken aufgelöste Ergebnis der Volksabstimmung zu Stuttgart 21 von 2011 kann dem Internetauftritt der Stadt entnommen werden [Stuttgart, 2011]. Es bietet sich dadurch eine zusätzliche Möglichkeit zur Modellevaluierung an, indem die Modellierungsergebnisse mit den tatsächlichen Ergebnissen verglichen werden. Da bei der Abstimmung mit Sicherheit nur die beiden Kategorien (*Zustimmung* und *Ablehnung*) unterschieden werden können, wurden die Gruppenausprägungen der Parameterisierungsstichprobe neu zusammengefasst. In den Rohdaten wurden noch 6 Gruppen unterschieden. Es wurde eine Neugruppierung in drei Gruppenausprägungen vorgenommen (Tabelle 1). Dafür wurden jeweils die Gruppen *sehr gut* und *gut* zu *Zustimmung* und *schlecht* und *sehr schlecht* zu *Ablehnung* zusammenge-

fasst. Falls nur *Zustimmung* und *Ablehnung* für die Modellierung berücksichtigt werden sollen, reduziert sich der Stichprobenumfang auf 2377 Beobachtungen. Dadurch bleibt die Möglichkeit erhalten eine multinomial verteilte abhängige Variable zu modellieren und trotzdem eine Validierung für zwei Klassen vorzunehmen. Für die exogen in die Analyse einfließenden Variablen sind detailliertere Informationen zu den Häufigkeiten der Ausprägungen im Anhang verfügbar (Abbildung 9). **bei der abb. Einkommen müsste es < 900 statt 900 sein**

Da in dieser Arbeit ein Schwerpunkt auf der Analyse unterschiedlicher räumlicher Effekte liegt, vergleicht dieser Abschnitt alle drei räumlichen Effekte in Bezug zu den beiden endogenen Variablen. Abbildung 2 zeigt die absolute Häufigkeit der Beobachtungen der Meinung zu Stuttgart 21 in kontinuierlicher räumlicher Lage. Zur besseren Übersicht wurden nicht alle Beobachtungen geplottet, sondern Beobachtungsdichten über bivariate normalverteilte Kerndichteschätzer mit festem Abstand für jede Richtungen ermittelt [Wickham, 2009] sowie [Venables and Ripley, 2002]. Um die Hintergrundkarte einbinden zu können wurden die Gauß-Krüger Koordinaten in Dezimalgrad umgerechnet. Da absolute Dichten dargestellt werden, ist die Dichte der Kategorien nicht nur von der Anzahl der Kategorien selbst, sondern auch von der Einwohnerdichte beeinflusst. Des Weiteren wird die Dichte von der Ausschöpfung, also der Anzahl beantworteter Befragungsbögen beeinflusst. Wegen der hohen Einwohnerdichte im Innenstadtbereich sind dort die Beobachtungsdichten aller 3 Klassen tendenziell höher als in den Randbezirken. Des Weiteren ersichtlich ist, dass einige Bereiche, wie das Naturschutzgebiet *Rotwildpark* im Westen oder der *Schurwald* im Osten, aufgrund ihrer geographischen Beschaffenheit oder Landnutzungsform nicht oder relativ dünn besiedelt sind.

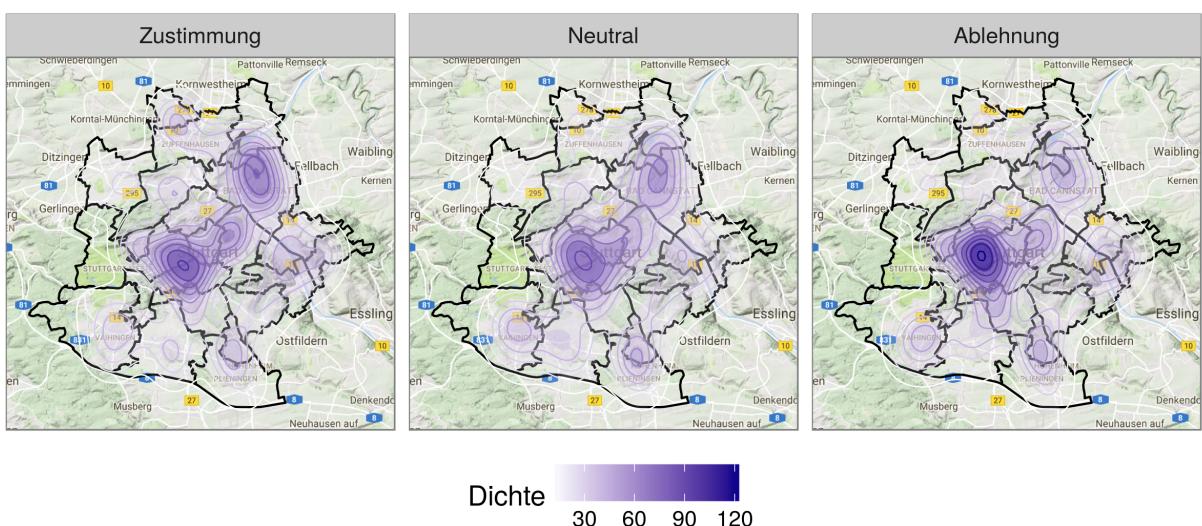


ABBILDUNG 2: KONTUR PLOT DER ABSOLUTEN ANZAHL DER GRUPPENBEOBACHTUNGEN ZUR MEINUNG ZU STUTTGART 21 IN DREI GRUPPEN. QUELLE DER HINTERGRUNDGRAFIK: REF: GOOGLE MAPS

Die *Zustimmung* zeigt offensichtliche räumliche Muster. Im Zentrum und im Nordosten ist sie höher als im Rest der Stadt. Der räumliche Trend der *Ablehnung* ist schwächer ausgeprägt.

Es zeigt sich jedoch, dass der Bereich der Innenstadt, sowie die südlichen Stadtbereiche etwas höhere Dichten bei der *Ablehnung* aufweisen. Die Beobachtungen der Kategorie *neutral* sind eher gleichmäßig über die Stadt verteilt.

Abbildung 3 zeigt die Dichte der Beobachtungen der fünf Kategorien zur Bewertung der Wohngegend. Hier zeigt sich ein deutlich ausgeprägteres räumliches Muster als bei der Meinung zu Stuttgart 21. Die Beobachtungen der Klasse *sehr gut* häufen sich sehr stark im Innenstadtbereich und im Süden. Die Kategorie *gut* verteilt sich relativ homogen über das gesamte Stadtgebiet mit einer etwas stärkeren Konzentration in der Innenstadt und im Nordosten. Bei der Klasse *neutral* zeigt sich eine stärkere Konzentration auf den Osten und Nordosten der Stadt. Praktisch alle *schlechten* und *sehr schlechten* Bewertungen sind deutlich abgegrenzt im Osten und Nordosten lokalisiert. Hierbei ist zu erwähnen, dass der Anteil der Personen, die ihre Wohngegend mit *schlecht* oder *sehr schlecht* bewertet haben sehr gering ist Abbildung 1.

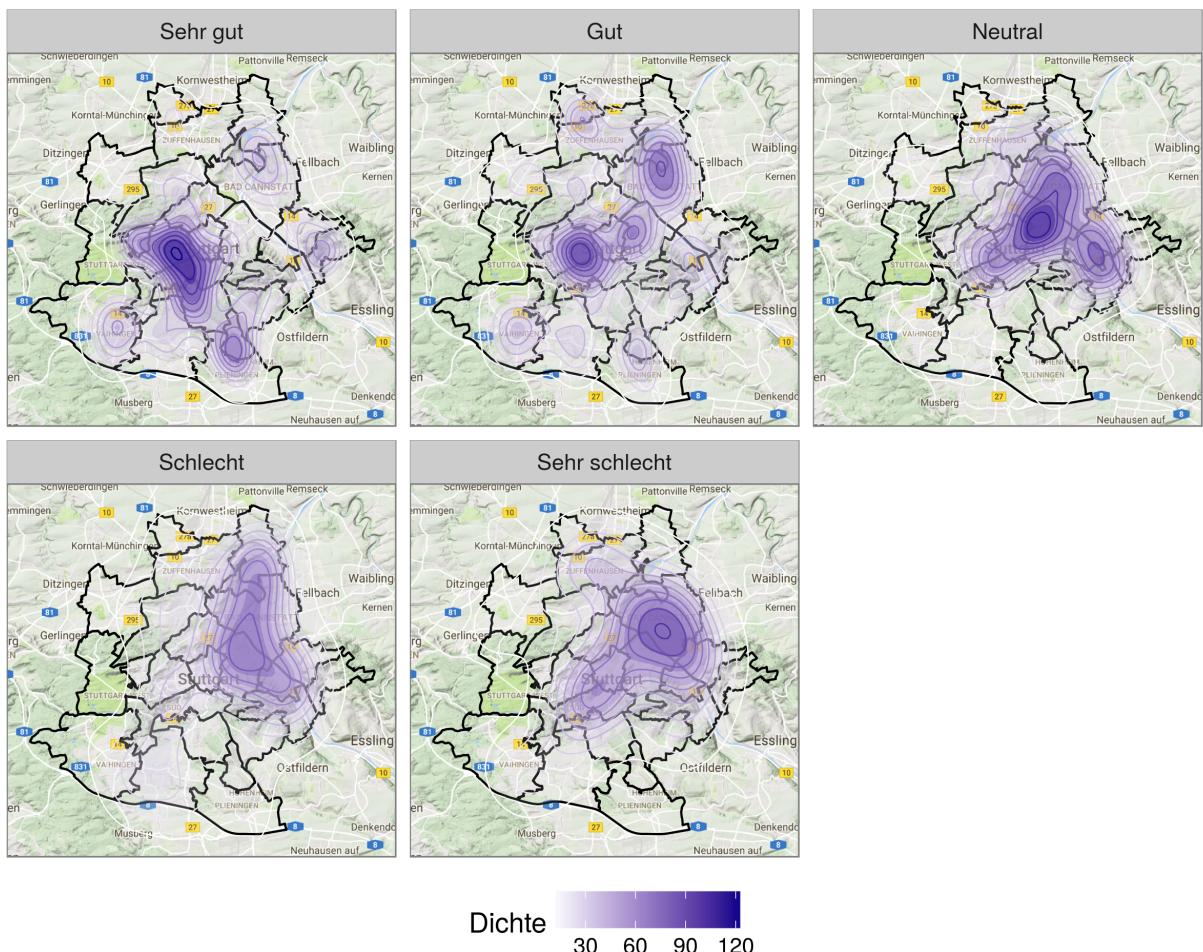


ABBILDUNG 3: KONTUR PLOT DER ABSOLUTEN ANZAHL DER GRUPPENBEOBACHTUNGEN ZUR WOHNZUFRIEDENHEIT IN FÜNF GRUPPEN. QUELLE DER HINTERGRUNDGRAFIK: REF: GOOGLE MAPS

Im folgenden werden die diskreten räumlichen Informationen auf Stadtbezirksebene beschrieben. Es werden 23 Stadtbezirke unterschieden. Im Gegensatz zur stetigen Beobachtungsdichte werden

die Beobachtungen nach Regionen aggregiert dargestellt. Dies hat den Vorteil, dass eine relative Anteilsdarstellung möglich wird (Abbildung 4). Die absoluten Häufigkeiten sind jedoch nicht dargestellt. Analog zur stetigen Darstellung (Abbildung 2) ist auch hier zu sehen, dass die Bürger des Nordostens eine positivere Meinung zu Stuttgart 21 haben als die Bürger aus dem Süden. Die *neutrale* Klasse hat in allen Bezirken einen geringeren Anteil und es ist kein räumliches Muster erkennbar. Die entsprechenden Anteilsgrafiken mit fünf Klassen für die Bewertung der Wohngegend sind im Anhang verfügbar (Abbildung 10). Wie in Abbildung 3 bereits angedeutet, zeigen sich negative Wohngebietseinschätzungen vor allem im Nordosten.

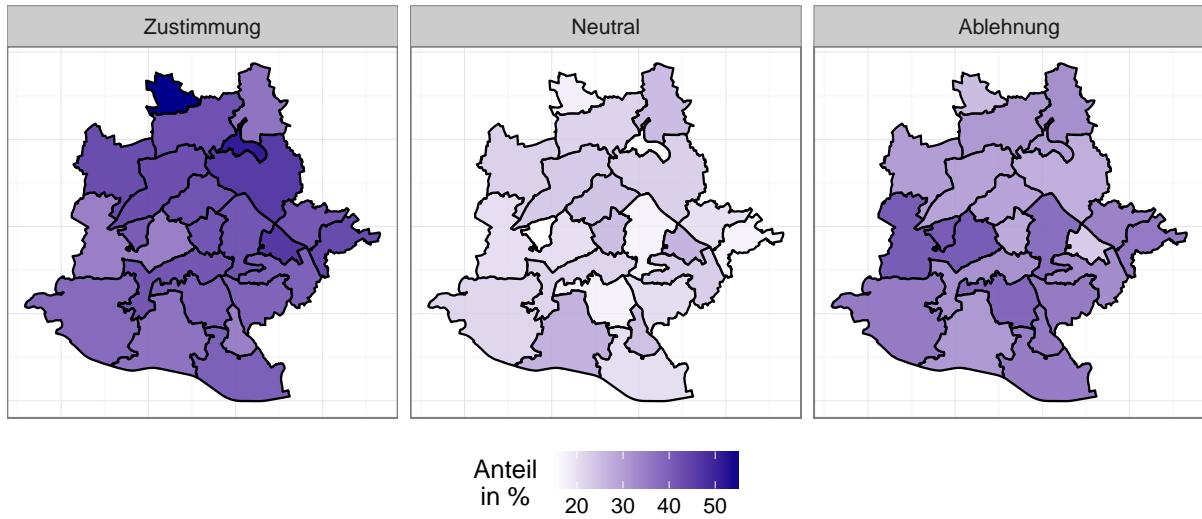


ABBILDUNG 4: ANTEILE ZUR MEINUNG ZU STUTTGART 21 NACH STADTBEZIRKEN.

Die dritte und letzte untersuchte räumliche Aggregationsebene ist die Stadtteilebene. Abbildung 5 zeigt die räumliche Verteilung von *Zustimmung*, *Ablehnung* und *neutraler* Haltung. Wie aus Tabelle 1 hervorgeht, ist die Stadtteilebene deutlich feiner Auflöst als die Bezirksebene. Dies führt in der Abbildung dazu, dass einige Stadtteile mit geringer Gesamteinwohneranzahl in einer oder mehreren Klassen keine Beobachtungen zeigen. Es gibt im Umkehrschluss auch Stadtteile, in denen eine Klasse zu 100 % vertreten ist. Außerdem gibt es in dieser Aggregationsebene sogar Stadtteile ohne jede Beobachtung, wie z. B. das Benzviertel im Innenstadtbereich oder die bereits angesprochenen Lagen im Westen.

Diese Stadtteile werden in den Diagrammen schwarz dargestellt. Wegen der feineren Auflösung ergibt sich ein mosaikartiges, visuell schwerer interpretierbares Bild. In keiner der drei Klassen lässt sich eine klare Struktur oder ein räumliches Muster erkennen. Die Anteile auf Stadtteilebene zu der Bewertung der Wohngegend sind im Anhang verfügbar 11. Hier zeigt sich ein ähnlich schwer differenzierbares Muster wie bei der Meinung zu Stuttgart 21. Wegen der deutlichen Unterschiede in den Anteilen der fünf Gruppen sind die Farbskalen nicht einheitlich, sondern unterscheiden sich in den Diagrammen.

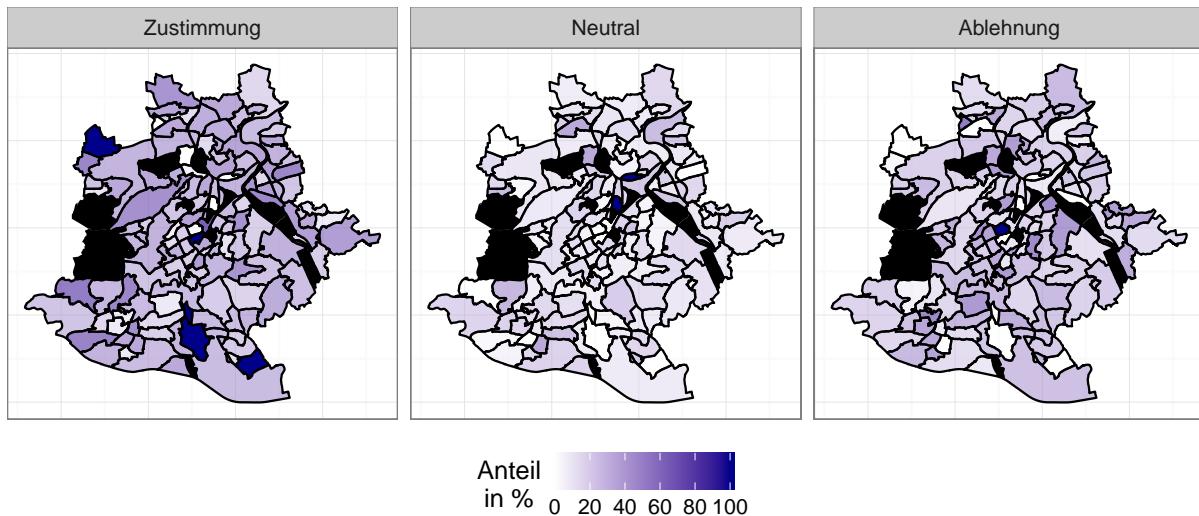


ABBILDUNG 5: ANTEILE DER MEINUNG ZU STUTTGART 21 NACH STADTTEILEN.

#### MELDEREGISTER

Die erste Datei an der die Modelle Anwendung finden sollen ist eine personenbezogene Auswertung aus dem Melderegister Stuttgarts vom 31.01.2011. Der Auszug umfasst alle volljährige Einwohner Stuttgarts außer Bewohner von Anstalten und Pflegeheimen. Mit einem Stichprobenumfang von 470.190 Bürgern liegt der Melderegisterauszug also sehr nah an der Grundgesamtheit von 573.104 Bürgern, die 2011 mit Hauptwohnsitz in Stuttgart gemeldet waren ([REF Stat. Bundesamt](#)). Insgesamt wurden 8 sozioökonomische Variablen erhoben. Bei allen Datensätzen liegt der Wohnsitz als kontinuierliche Gauss-Krüger Geokoordinate vor. Um eine kleinräumige Extrapolation mit diskreten räumlichen Informationen vornehmen zu können, wurden die Stadtteil- und Stadtbezirksinformationen an die Datensätze des Melderegisters angehängt. Hierzu wurden die Stadtteil- und Stadtbezirkspolygone, welche uns von der Stadt Stuttgart für die Analyse zur Verfügung gestellt wurden, über eine räumliche Abfrage mit den Melderegisterdatensätzen verknüpft. Die geographische Verknüpfung wurde mit dem freien Geoinformationssystem QGIS ([REF QGIS](#)) durchgeführt. Die Projektdatei mit der Geoabfrage (*Geographische\_Abfrage.QGS*) sowie die Shapefiledateien (*Stadtteile\_netto.SHP*) liegen dieser Arbeit digital bei. Wie in den Tabellen 1 und 2 ersichtlich, eignen sich nicht alle Variablen zur Extrapolation, da nicht alle Variablen in jeder Umfrage erhoben wurden. Es ergibt sich ein Überschneidungsbereich der fünf sozioökonomischen Variablen *Altersklasse Befragter*, *Geschlecht*, *Nationalität*, *Familienstand* und *Personenzahl im Haushalt*. **IN TAB 2 BRAUCHEN WIR DIE SPALTE MODLLEIRUNG EIGENTLICH NICHT MEHR ODER? ICH HABE DIE VARIABLEN Z.T. UMBEBENANNT, DAMIT SIE GLEICH HEIßEN WIE IN TAB 1. AUßerdem sind ja nach dem Umkodieren bei der Altersklasse nur noch 6 Auspr. Ich würde das in der Tab. 2 anpassen was meinst du?**

TABELLE 2: SOZIOÖKONOMISCHE UND GEOGRAPHISCHE VARIABLEN DES MELDEREGISTERS UND DEREN ANZAHL DER AUSPRÄGUNGEN.

Anzahl Beobachtungen: 470.190

Variable	Modellierung	Anzahl Ausprägungen
Altersklasse Befragter	Nicht Parametrisch	14
Geschlecht	Parametrisch	2
Nationalität	Parametrisch	2
Familienstand	Parametrisch	4
Personenzahl im Haushalt	Nicht Parametrisch	5
Wohndauer	Nicht Parametrisch	3
ALG II Quote	Nicht Parametrisch	9
Ein/Zweifamilienhäuser	Nicht Parametrisch	8
Gauß-Krüger	Tensorprodukt-Splines	

## ZENSUS

Im Rahmen der bundesweiten Volkszählung von 2011 wurde in Stuttgart eine Gebäude- und Wohnungszählung durchgeführt. Diese Datei umfasst 380.238 Bürger Datensätze (Tabelle 3). Da bei der Zählung auch für die Fragestellung dieser Arbeit relevante sozioökonomische Variablen erhoben wurden, eignet sich diese Umfrage ebenfalls zur kleinräumige Extrapolation. Die diskreten geographischen Angaben wurden analog zum Vorgehen beim Melderegister per geographischer Abfrage ergänzt. Für die kleinräumige Extrapolation kommen die gleichen soziökonomischen Variablen wie beim Melderegister in Frage. Wie in Tab. 2.: Habe die Namen angepasst. Anzahl Personenzahl und Alter umkodiert, also auch in Tabelle ändern? Wir könnten sogar überlegen, die nicht verwendeten Variablen zu löschen. Wir gehen ja im Text nicht mehr auf diese ein (gilt auch für Tab 2)

TABELLE 3: ERHOBENE SOZIOÖKONOMISCHE UND GEOGRAPHISCHE VARIABLEN DER GEBÄUDE- UND WOHNUNGSZÄHLUNG IM RAHMEN DES ZENSUS UND DEREN ANZAHL DER AUSPRÄGUNGEN.

Anzahl Beobachtungen: 380.238

Variable	Modellierung	Mögliche Ausprägungen
Altersklasse Befragter	Nicht Parametrisch	9
Geschlecht	Parametrisch	2
Nationalität	Parametrisch	2
Familienstand	Parametrisch	4
Personenzahl im Haushalt	Nicht Parametrisch	6
Wohnfläche	Nicht Parametrisch	24
Stellung Beruf	Parametrisch	9
Beamter	Parametrisch	2
Gebäudetyp	Parametrisch	10
Gebäudenutzung	Parametrisch	2
Gauß-Krüger	Tensorprodukt-Splines	

## STATISTISCHE METHODEN

### MODELL

Bei allen endogenen Variablen handelt es sich um kategoriale Variablen, von denen die Ausprägungswahrscheinlichkeiten modelliert werden sollen. Da Ausprägungswahrscheinlichkeiten für diese kategorialen Responses nur im Wertebereich zwischen 0 und 1 sinnvoll interpretierbar sind und keine der Response-Variablen Normalverteilung aufweist, sind einfache Regressionsmodelle ohne Transformationsfunktion ungeeignet. Für die Fragestellung dieser Arbeit kommen nur Regressionsmodelle in Frage, die eine Modellierung von Wahrscheinlichkeiten im genannten Wertebereich sicherstellen. Lineare generalisierte Modelle ermöglichen die Transformation des linearen Prädiktors

$$\eta_i = \mathbf{x}'_i \boldsymbol{\beta}$$

durch eine Response-Funktion  $h$ , wobei die Wahl von  $h$  von der Fragestellung und der Skalierung der Response-Variable abhängt.  $h$  kann jede monoton ansteigende Funktion sein. Zur Sicherstellung des gewünschten Wertebereichs zwischen 0 und 1 wird der lineare Prädiktor bei kategorialen Regressionen häufig über eine Logit Response-Funktion transformiert. Bei einer binären Response ergibt sich nach [Fahrmeir et al., 2013, p. 270 f.] folgende Transformation des linearen Prädiktors **weißt du, wie man die Formel nummeriert?**

$$h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

Nach einer Logit-Transformation folgen die Daten häufig einer Binomialverteilung. Ein weiterer Vorteil von GLM ist, dass diese spezifischen Verteilungsannahmen berücksichtigt werden können. Bei ordinalskalierter Response kann die Reihenfolge der Ausprägung im generalisierten Modell explizit berücksichtigt werden [Fahrmeir et al., 2013, p.334 ff.]. Im kumulativen Modell wird angenommen, dass für jeden Datensatz eine kontinuierliche (unbekannte) latente Variable  $u_i$  existiert, deren Ausprägung die Kategoriezugehörigkeit bestimmt. Unter der Annahme, dass diese latente Variable über den linearen Prädiktor

$$u_i = -\mathbf{x}'_i \tilde{\boldsymbol{\beta}} + \varepsilon$$

vorhergesagt werden kann, existieren für die latenten Variablen geordnete Schwellenwerte  $\theta_r$ . Damit die Schwellenwerte identifizierbar werden, hat der lineare selbst keinen Intercept. Eine Beobachtung zeigt die Ausprägung  $r$ , wenn ihre latente Variable zwischen  $\theta_{r-1}$  und  $\theta_r$  liegt. Über die kumulative Verteilungsfunktion  $F$  des Fehlers  $\varepsilon_i$  um die latente Variable  $u_i$  kann die Wahrscheinlichkeit jeder Ausprägung für jede Beobachtung als

$$P(y_i = r) = F(\theta_r + \mathbf{x}'_i \tilde{\boldsymbol{\beta}}) - F(\theta_{r-1} + \mathbf{x}'_i \tilde{\boldsymbol{\beta}})$$

berechnet werden. Wenn das in der Literatur vorgeschlagene kumulative Logit Modell gewählt wird [Fahrmeir et al., 2013, p. 334], berechnet sich die kumulative Wahrscheinlichkeit einer Be-

obachtung  $y_i$  kleiner oder gleich der geordneten Ausprägung  $r$  zu sein als

$$P(y_i \leq r) = \frac{\exp(\theta_r + \mathbf{x}'_i \tilde{\boldsymbol{\beta}})}{\exp(1 + \exp(\theta_r + \mathbf{x}'_i \tilde{\boldsymbol{\beta}}))}$$

Die exogenen Variablen, mit denen das Modell für die räumliche Extrapolation parametrisiert werden soll, sind sehr heterogen. Es kommen sowohl nominal als auch kardinalskalierte sozioökonomischen Variablen in Frage (Abbildung 1, Tabelle 1). Während ein linearer Prädiktor zur Beschreibung des Zusammenhangs zwischen ordinalskalierte Variablen und einer kategorielten Response-Variablen ausreichend ist, zeigen kardinal skalierte Variablen oft einen nichtlinearen Zusammenhang [Fahrmeir et al., 2009, p. 9]. Die in generalisierten linearen Modellen vorausgesetzte lineare Abhängigkeit zwischen allen transformierten Kovariablen und der Response-Variablen  $\eta$  ist demnach für die Daten dieser Arbeit zu unflexibel. Aufgrund dieser Ausgangslage haben flexiblere Modelle in dieser Arbeit also voraussichtlich Vorteile gegenüber reinen linearen Modellen. Durch generalisierte additive Modelle (GAM) können lineare und nichtlineare Effekte sehr flexibel in einem Modell verbinden werden. Im Unterschied zu generalisierten linearen Modellen sind die einzelnen Terme des Prädiktors nicht auf lineare Funktionen beschränkt, sondern können auch durch semiparametrische Splines nichtlinear modelliert werden. Zur Modellierung der nichtlinearen, also der nichtparametrischen Terme des Prädiktors, wurden penarisierte Basissplines verwendet [Eilers and Marx, 1996]. Diese haben den Vorteil, dass

Analog zu den generalisierten linearen Modellen wird der Prädiktor beim GAM ebenfalls über die Response Logit Response-Funktion transformiert. Im Falle einer binären Response wird die Ausprägungswahrscheinlichkeit über eine Logit Response-Funktion Je nachdem, ob die Ausprägungswahrscheinlichkeiten für eine Binäre oder geordnet kategoriale Response modelliert werden

`ocat, gam` [Wood et al., 2016]

Kurze Erläuterung GAM und logit,... Zusammengesetzt aus parametrischen Effekten und Splines. Pseudobeobachtungen. Wahl des Glättungsparameters (Wood 2000).

## MODELLWAHL

Basierend auf den vorhergegangenen Analysen zum Beamten- und Eigenheimanteils in Stuttgart, die uns zur Verfügung standen wurde eine Funktion zur schrittweisen AIC [Akaike, 1981] Berechnung programmiert, die zur Identifikation der geeignetsten Kovariablenkombination dient (siehe digitaler Anhang `stepAIC.R`). In der Funktion werden mithilfe der `gam()` Funktion des `mgcv` Paketes [Wood, 2011] generalisierte additive Modelle mit unterschiedlichen Kovariablen erstellt und deren AIC berechnet. Vor dem Aufruf der Funktion müssen die abhängige Variable, die Verteilungsannahme des Regressionsmodells, die Gewichtungen der Einzelbeobachtungen und unveränderliche Kovariablen definiert werden. Außerdem muss eingeschätzt werden, welche der veränderlichen Kovariablen parametrisch oder semiparametrisch als Spline in das Modell eingehen. Es wird zunächst der AIC des einfachsten, nur aus den fest vorgegebenen Kovariablen bestehenden Modells berechnet. In Iteration eins werden alle veränderlichen Kovariablen einzeln nacheinander in die Modellformel aufgenommen und es wird jeweils ein GAM erstellt sowie

dessen AIC berechnet. Die Kovariablen gehen entsprechend der vorigen Eingabe parametrisch oder semiparametrisch ein. Falls die Hinzunahme mindestens einer Kovariable in Iteration eins zu einer Reduktion des AIC führt, wird diejenige Kovariable, welche zu dem Modell mit dem kleinsten AIC führt zur Modellformel hinzugefügt. Falls das Modell nur mit den festen Modellbestandteilen bereits den geringsten AIC zeigt ist die Modellwahl folglich in Iteration eins bereits beendet.

Andernfalls setzt sich das Ausgangsmodell für Iteration zwei aus den festen Kovariablen und einer weiteren Kovariable zusammen. In Iteration zwei werden wie zuvor alle verbleibenden Kovariablen zunächst nacheinander zur aktuellen Modellformel hinzugefügt. Wenn die Kovariable mit dem geringsten AIC gefunden ist (falls diese existiert und das Modell aus Iteration eins nicht bereits das geeignetste ist), werden alle veränderbaren Kovariablen in Iteration zwei nochmals nacheinander eliminiert. Das Modell mit dem geringsten AIC bildet das Ausgangsmodell der nächsten Iteration. Dies wird wiederholt bis in einer Iteration kein Modell mit einem geringeren AIC als in der vorigen Iteration parametrisiert werden kann. Um die Laufzeit der Funktion zu begrenzen, wurde auf die Analyse von Wechselwirkungen zwischen den Kovariablen verzichtet. Wechselwirkungen können jedoch als unveränderliche Modellbestandteile eingehen.

### KONFIDENZINTERVALLE

Die Intervalle zu den Punktschätzungen liefern zusätzliche wichtige Informationen, da die Punktschätzungen alleine keine Angaben zur Unsicherheit beinhalten. Das Konfidenzintervall gibt an, mit welcher Wahrscheinlichkeit die unbekannte wahre Punktschätzung von einer Konfidenzregion überdeckt wird. Mit dem Vertrauensintervall erhält man die Relation der Unsicherheit zur Punktschätzung und somit ein weiteres Modellgütemerkmal [Fahrmeir et al., 2013, p. 471]. Des weiteren wurden die Intervalle genutzt um Überdeckungswahrscheinlichkeiten des Volksabstimmungsergebnisses [Stuttgart, 2011] zu berechnen. Grundsätzlich gibt es die Möglichkeiten die Intervalle entweder aus Modelleigenschaften abzuleiten oder Bootstrap-Intervalle durch wiederholte zufällige Modellreparametrisierungen und Punktschätzungen zu berechnen. Bei ersterem Vorgehen unterliegen die Intervallberechnungen, je nach gewählter Berechnungsmethode, asymptotischen Eigenschaften oder Verteilungsannahmen. So wird zur analytischen Berechnung von Score-Konfidenzintervallen [Held, 2008, p. 64 ff.], welche den Vorteil haben, dass sie invariant gegenüber eindeutigen Parametertransformationen sind, beispielsweise die Fischer-Information benötigt. Sie sind deshalb oft nur schwer analytisch zu berechnen [Held, 2008, p. 74]. Andere Intervalle, wie beispielsweise das Wald-Konfidenzintervall, sind obligatorisch symmetrisch [Held, 2008, p. 60] und werden der wahren Datenverteilung daher oft nicht gerecht. Bei beiden Intervallen hängt die tatsächliche Überdeckungswahrscheinlichkeit dea unbekannten wahren Parameter bei binärer Response Variable zudem von der Parameterausprägung sowie der Beobachtungsanzahl der Modellstichprobe ab und entspricht deshalb in ungünstigen Kombinationen nicht der nominalen, also der gewünschten Überdeckungswahrscheinlichkeit [Lawrence D. Brown, 2001]. Bootstrap Konfidenzintervalle hingegen können für jede Form der statistischen Inferenz berechnet werden [DiCiccio and Efron, 1996]. Da in dieser Arbeit Modelle mit unterschiedlichen Verteilungsannahmen und unterschiedlichen räumlichen Effekten verglichen werden, bietet es sich an auf die verteilungsunabhängigen Bootstrap-Intervallschätzungen zurückzugreifen. Von Interesse sind weniger die Intervalle um die Modellparameter als vielmehr die Intervalle um die

Punktschätzungen, also die einzelnen Hochrechnungen. Dementsprechend wurden Bootstrap-Konfidenzintervalle für jede einzelne Punktschätzung berechnet. Zur Berechnung der Intervalle wurde jedes Modell 1.000 mal mit einer Zufallsstichprobe (*Ziehen-mit-Zurücklegen*) aus der Parameterisierungsstichprobe reparametrisiert. Um den Einfluss des Stichprobenumfangs zu eliminieren, enthielt jede Stichprobe die tatsächliche Anzahl der Beobachtungen. Mit jedem neu-parametrisierten Modell wurde eine kleinräumige Extrapolation durchgeführt. Die Hochrechnungsergebnisse wurden genutzt um die arithmetischen Mittel, die Mediane sowie untere und obere 95 % Perzentile der Punktschätzungen zu berechnen.

### KREUZVALIDIERUNG

Für die Modellerstellung der additiven Modelle lag eine Stichprobe von 3.143 Beobachtungen vor. Informationen zur Grundgesamtheit dieser Stichprobe lagen nicht vor. Die Qualität der parametrisierten Modelle ließ sich folglich nicht anhand einer Grundgesamtheit validieren, sondern musste an der Stichprobe selbst eingeschätzt werden. Die erstellten Modelle wurden an anderen Dat3en als den Parametrisierungsdaten angewendet. Aus diesem Grunde ist war vor allem die Modellqualität außerhalb der Parametrisierungsdaten interessant. Aus diesem Grunde wurde eine *Leave-One-Out* Kreuzvalidierung durchgeführt [Fahrmeir et al., 2013, p. 149], in welcher jeweils eine Beobachtung zufällig entfernt wurde. Mit den verbliebenen Datensätzen wurde das additive Modell, welches sich zuvor im schrittweisen AIC Vergleich als am vorteilhaftesten herausgestellt hat, neu parametrisiert und die entfernte Beobachtung mit diesem Modell vorhergesagt. Dies wurde so lange wiederholt bis jeder Datensatz einmal entfernt (und geschätzt) wurde. Mit diesen Daten ließ sich eine Statistik zur Reklassifikation der Beobachtungen außerhalb der Stichprobe erstellen.

### VALIDIERUNG

Ziel der Validierung ist es, die prognostizierten Anteile aus dem gewählten Modell mit den wahren Anteilen aus der Volksabstimmung [Stuttgart, 2011] zu vergleichen, um eine Aussage über die Qualität der geschätzten Prognosemodelle geben zu können. Die Validierung erfolgt auf Stadtteil- und Bezirksebene, sowie für das Gesamtergebnis der Stadt Stuttgart. Dazu werden insbesondere zwei statistische Gütemaße verwendet.

Bei der Wahl des Schätzers geht es zum einen darum, einen möglichst erwartungstreuen als auch effizienten Schätzer zu finden. Als geeignetes Gütemaß hat sich die mittlere quadratische Abweichung erwiesen, da sie sowohl die Varianz, als auch die quadrierte Verzerrung berücksichtigt. Zudem hat ein konsistenter Schätzer die Eigenschaft, dass die mittlere quadratische Abweichung bei unendlich groß werdender Stichprobe gen Null konvergiert [Georgii, 2009, p. 201]. Ein weiteres Kriterium ist die Überdeckungswahrscheinlichkeit. Sie gibt an, mit welcher Wahrscheinlichkeit das geschätzte Konfidenzintervall den wahren Wert enthält. Erwartet wird hier, dass die Überdeckungswahrscheinlichkeit dem Konfidenzniveau entspricht. Mögliche größere Abweichungen können durch die Approximation einer diskreten Verteilung durch eine stetige Verteilung resultieren, was z.B. oft bei der Approximation der Binomial- durch die Normalverteilung kommt [Lawrence D. Brown, 2001, p. 102].

## ERGEBNISSE

### MODELL

#### MODELLWAHL

Die Modellwahl mit Hilfe der Funktion `stepAIC` hat eine Zusammensetzung von Kovariablen geliefert, welche das Modell mit dem besten Fit und der geringsten Komplexität liefert. Da der räumliche Effekt stets als fester Bestandteil aufgenommen wurde, listest Tabelle 4 zu jedem Modell das AIC für ein vollständiges Geoadditives Modell, einem Modell nur mit räumlichem Effekt und ein Modell bestehend aus allen Kovariablen außer dem räumlichem Effekt auf. Dadurch lässt sich die relative Informationsqualität des jeweiligen räumlichen Effekts untersuchen. Für die Meinung zu Stuttgart 21 im drei Klassen Fall lässt sich erkennen, dass das Geoadditive Modell mit den Gauss-Krüger Informationen und den Stadtteil Informationen die höchste relative Qualität aufweist. Bei dem Modell mit Bezirken als räumlichem Effekt scheint dieser die Qualität des Modells in Abhängigkeit von der Komplexität nicht zu verbessern. Insgesamt deutet das AIC für den drei Klassen Fall auf das Geoadditive Modell mit den Gauss-Krüger Informationen als qualitativ höchstes Modell hin. Beim zwei Klassen Fall schneidet das Geoadditive Modell in jedem Fall am besten ab, wobei hier das Modell mit Stadtteilen als räumlichem Effekt die höchste Qualität laut des AIC aufweist.

TABELLE 4: VERGLEICH DER STEP AIC ERGEBNISSE ZWISCHEN DEN MODELLEN

Meinung zu Stuttgart 21			
	Geoadditives Modell	Modell ohne räumlichem Effekt	Modell nur mit räumlichem Effekt
Drei Kl. Bezirke Stadtteile	Gauss-Krüger	6379,345	6382,654
	Bezirke	6455,284	6455,284
	Stadtteile	6428,6	6510,838
Zwei Kl. Bezirke Stadtteile	Gauss-Krüger	3114,143	3268,483
	Bezirke	3115,858	3271,325
	Stadtteile	3079,12	3163,597
Bewertung der Wohngegend			
	Geoadditives Modell	Modell ohne räumlichem Effekt	Modell nur mit räumlichem Effekt
Gauss-Krüger Bezirke Stadtteile	7051,606	7318,815	7062,108
	Bezirke	7175,601	7197,003
	Stadtteile	8252,374	8750,801

Für die Bewertung der Wohngegend als endogene Variable zeigt sich noch ein deutlicherer Unterschied zwischen Schätzungen mit räumlichem Effekt und ohne räumlichem Effekt. Modelle mit räumlichem Effekt zeigen einen deutlich niedrigeren AIC, so ist der Unterschied bei den Gauss-Krüger und Bezirksinformationen zwischen dem Geoadditiven Modell und dem Modell nur mit räumlichem Effekt sehr viel geringer als zwischen Geoadditivem Modell und Modell ohne räumlichem Effekt. Dies weißt darauf hin, dass der räumliche Effekt einen hohen Erklärungsgehalt für die Ausprägung der abhängigen Variable liefert. Die gegenteilige Situation findet man bei der Meinung zu Stuttgart 21 vor. Insgesamt signalisiert das AIC die höchste Qualität für das Geoadditive Modell mit Gauss-Krüger Informationen bei der Bewertung der

Wohngegend.

#### REKLASSIFIZIERUNG

Die Ergebnisse der Reklassifizierung zur Meinung zu Stuttgart 21 (Tabelle 5) zeigen, dass die Erfolgsquote im drei Klassenmodell zwischen 44% und 50% liegt und im zwei Klassenmodell zwischen 55% und 62% liegt. Zum Vergleich mit einem reinen Zufallsmodell, dass im drei Klassenmodell eine Erfolgswahrscheinlichkeit von  $1/3$  und im zwei Klassenmodell von  $1/2$  hat, weisen die geschätzten Modelle eine höhere Erfolgsquote auf. Auch bei einem Vergleich mit Tabelle 1 zeigt sich, dass die geschätzten Modelle besser abschneiden, als ein triviales Wählen der immer gleichen Klasse.

TABELLE 5: REKLASSIFIZIERUNG DER MEINUNG ZU STUTTGART 21

		Geoadditives Modell	Modell ohne räumlichem Effekt	Modell nur mit räumlichem Effekt
Drei Kl.	Gauss-Krüger	0,4918	0,4716	0,4732
	Bezirke	0,4726	0,4719	0,4726
	Stadtteile	0,451	0,4685	0,4449
Zwei Kl.	Gauss-Krüger	0,6193	0,6104	0,5524
	Bezirke	0,6079	0,6104	0,5515
	Stadtteile	0,6282	0,6052	0,6099

Des weiteren ist zu sehen, dass das Geoadditive Modell in fast allen Fällen die höchste Erfolgsquote aufweist. Abweichungen bestehen im drei Klassenmodell mit Stadtteilen als räumlichem Effekt und im zwei Klassenmodell mit Bezirken als räumlichem Effekt. Außerdem ist zu beachten, dass im drei Klassenmodell das Modell nur mit räumlichem Effekt besser abschneidet als das Modell ohne räumlichem Effekt, während sich für zwei Klassen diese Situation umgekehrt hat.

Für die Reklassifikation der Bewertung der Wohngegend (Tabelle 6) ergibt sich eine Erfolgsquote zwischen 40% und 50%. Damit ist auch hier eine deutliche Verbesserung gegenüber reinem Raten oder dauerhaftem wählen einer Klasse gegeben.

TABELLE 6: REKLASSIFIZIERUNG DER BEWERTUNG DER WOHNGEEND

		Geoadditives Modell	Modell ohne räumlichem Effekt	Modell nur mit räumlichem Effekt
Gauss-Krüger	0,4922	0,4461	0,4896	
Bezirke	0,4701	0,4461	0,4621	
Stadtteile	0,4046	0,4204	0,4347	

Für die Modelle mit den kontinuierlichen Gauss-Krüger Informationen und den Bezirken als räumlichem Effekt hat das Geoadditive Modell die höchste Erfolgsrate, wohingegen für die Stadtteilinformationen das Modell nur mit räumlichem Effekt die beste Reklassifizierung aufweist. Insgesamt schneidet das Geoadditive Modell für beide endogene Variablen und alle drei möglichen Klassenanzahlen am besten ab. Außer bei dem zwei Klassenmodell zur Meinung zu Stuttgart 21 weisen beim Geoadditivem Modell die Gauss-Krüger Informationen als räumliche Effekte die höchste und die Stadtteile als räumliche Effekte die niedrigste Erfolgsrate auf. Da die

Modelle ohne- oder nur mit räumlichem Effekt in der AIC-Untersuchung und Reklassifizierung in den meisten Fällen schlechter Abschnitten, wurden die Ansätze ohne- oder nur mit räumlichem Effekt nicht weiter verfolgt.

## KREUZVALIDIERUNG

Die Kreuzvalidierung wurde vorgenommen, um die Vorhersagequalität außerhalb der eigenen Stichprobe der Modelle einzuschätzen. Die Ergebnisse zur Meinung zu Stuttgart 21 (Tabelle 7) zeigen eine ähnliche Erfolgsquote wie die innerhalb der Stichprobe vorgenommene Reklassifizierung (Tabelle 5). Für das drei Klassenmodell liegt der Anteil der erfolgreich Klassifizierten Beobachtungen bei ca. 49% und im zwei Klassenmodell zwischen 60% und 63%. Im drei Klassen Fall kann das Modell mit Stadtteilen als räumlichem Effekt ca. 3,5% mehr Beobachtungen richtig zuweisen als in der Reklassifizierung und das Modell mit Bezirken als räumlichem Effekt ca. 2%, womit es insgesamt auch die höchste Erfolgsrate aufweist. Die Erfolgsquote mit den Gauss-Krüger Informationen ist in etwa gleich geblieben, ähnlich wie die Modelle im zwei Klassen Fall.

TABELLE 7: KREUZVALIDIERUNG DER MEINUNG ZU STUTTGART 21 NACH EINZELNEN KLASSEN

Drei Klassen										
		Gauss-Krüger			Bezirke			Stadtteile		
					Geschätzte Klasse					
Wahre Klasse	1	1	2	3	1	2	3	1	2	3
	2	0,756	0	0,244	0,754	0	0,246	0,746	0	0,254
	3	0,673	0	0,327	0,670	0	0,330	0,663	0	0,337
Klassifikation Modell		0,4905			0,4928			0,4893		
Insgesamt										
Zwei Klassen										
		Gauss-Krüger			Bezirke			Stadtteile		
					Geschätzte Klasse					
Wahre Klasse	1	1	2		1	2		1	2	
	2	0,747	0,253		0,732	0,268		0,748	0,252	
Klassifikation Modell		0,538			0,545			0,516		
Insgesamt		0,6193			0,6079			0,6286		

Zudem lässt sich anhand von Tabelle 5 die Trefferquote der einzelnen Klassen entnehmen. Die Tabelle zeigt die Anteile der Zuordnung von Beobachtungen der jeweiligen Klasse auf die eigene und jede der anderen Klassen. Die Hauptdiagonale liefert die Anteile der korrekt zugeordneten Beobachtungen. Die Einträge auf den Nebendiagonalen geben damit an, wie viele Beobachtungen einer Klasse relativ zur Gesamtzahl der Beobachtungen in eine falsche Klasse zugeordnet wurden. Markant ist, dass der zweiten Klasse in keinem Modell Beobachtungen zugeordnet wurden, weder falsch noch richtig. Mehrheitlich wurden diese der ersten Kategorie zugewiesen. Zudem ist die erste Klasse am öftesten richtig zugeordnet worden, allerdings wurden ihr auch die meisten falsch klassifizierten Beobachtungen zugeordnet. Dies gilt für alle drei Modelle. Im zwei Klassen Fall wurde das Problem der nicht geschätzten zweiten Klasse durch ihr auslassen umgangen. Allerdings findet sich auch hier ein deutlich häufigeres falsches zuordnen der zweiten Klasse in

die erste als anders herum.

TABELLE 8: KREUZVALIDIERUNG DER BEWERTUNG DER WOHNGEgend NACH EINZELNEN KLASSEN

	Gauss-Krüger					Bezirke					Stadtteile					
						Geschätzte Klasse										
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	
Wahre Klasse	1	0,445	0,555	0	0	0	0,387	0,613	0	0	0	0,495	0,478	0,019	0	0,008
	2	0,254	0,746	0	0	0	0,252	0,748	0	0	0	0,300	0,671	0,023	0	0,006
	3	0,149	0,845	0	0	0	0,153	0,847	0	0	0	0,142	0,771	0,079	0	0,008
	4	0,141	0,859	0	0	0	0,141	0,859	0	0	0	0,099	0,474	0,349	0	0,078
	5	0,114	0,886	0	0	0	0,086	0,914	0	0	0	0,031	0,275	0,556	0	0,138
Klassifik. Modell Insgesamt		0,4918				0,4704									0,4594	

Tabelle 6 ist analog zu Tabelle 5 aufgebaut. Das Ergebnis der Kreuzvalidierung für die Bewertung der Wohngegend ist, dass die Erfolgsquote zwischen ca. 46% und 49% liegt. Die Modelle mit Gauss-Krüger Informationen und Bezirken als räumliche Effekte haben vergleichbar wie in der Reklassifizierung abgeschnitten. Das Modell mit den Stadtteilen hingegen weiß eine ca 5% höhere Genauigkeit auf. Insgesamt jedoch behält das Modell mit den Gauss-Krüger Informationen die höchste Erfolgsrate. Auffällig ist auch hier, dass eine Klasse in keinem Modell Beobachtungen zugewiesen bekommen hat und zwei weitere Klassen in zwei von drei Modellen keine Beobachtungen zugewiesen bekommen haben. Abbildung 1 gibt dabei einen Hinweis auf die Ursache dieses Phänomens, denn die Kategorien *sehr gut* und *gut* sind überproportional häufig im Datensatz vertreten im Vergleich zu den Klassen *schlecht* und *sehr schlecht*, weshalb schon durch die relative Häufigkeit die Wahrscheinlichkeit dass eine Beobachtungen in einer der letzten drei Kategorien ist, sehr klein wird. Bei der Bewertung der Wohngegend wurde am häufigsten die zweite Klasse sowohl richtig als auch falsch geschätzt.

Durch das relativ hohe Überschätzen einer Klasse bei beiden endogenen Variablen, sowie das totale Unterschätzen einiger Klassen, wird in der weiteren Validierung und Extrapolation auf das Vorhersagen einzelner Beobachtungen verzichtet. Stattdessen wird in einem bestimmten Gebiet über die Wahrscheinlichkeiten der Beobachtungen aggregiert und ein Anteil geschätzt, um auch geringere Wahrscheinlichkeiten für Beobachtungen zu berücksichtigen.

## VALIDIERUNG

Die Ergebnisse der Validierung sind aus Tabelle 9 ersichtlich. Die Tabelle zeigt die mittlere Quadratische Abweichung und Überdeckungswahrscheinlichkeit der prognostizierten Anteile aus den Modellen, bezogen auf die wahren Anteile aus der Volksabstimmung. Da in dieser Volksabstimmung nur Informationen zu zwei Klassen verfügbar sind, wird erwartet, dass das zwei Klassenmodell die beiden Kategorien signifikant besser prognostiziert. Tabelle 9 ist so aufgebaut, dass die linke Spalte angibt um welche Kombination aus Modell, räumlichem Effekt und Datei es sich handelt. Die mittlere Spalte zeigt die mittlere quadratische Abweichung für die beiden Kategorien der Volksabstimmung und die rechte Spalte zeigt die Überdeckungswahrscheinlichkeit ebenfalls für diese Kategorien. Damit gibt die erste Reihe z.B. die statistischen Gütemaße für

das zwei Klassenmodell mit Gauss-Krüger Informationen als räumlichem Effekt aggregiert auf Bezirksebene mit dem Melderegister an.

TABELLE 9: VERGLEICH DER MITTLEREN QUADRATISCHEN ABWEICHUNG (MSE) UND DER ÜBERDECKUNGSWAHRSCHEINLICHKEIT BEI ALLEN PROGNOSSEN AUS DEN GESCHÄTZTEN MODellen UND DEN BEIDEN EXTRAPOLATIONSDATEIEN FÜR DIE MEINUNG ZU STUTTGART 21

		MSE		Überdeckungswk.	
		Zustimmung	Ablehnung	Zustimmung	Ablehnung
Gauss-Krüger	Bez.	U	0,04	0,749	1
		Z	0,116	0,557	0,391
	Sadtt.	U	0,461	5,708	0,954
		Z	0,813	4,415	0,553
3 Kl. Bezirke	Bez.	U	0,041	0,756	1
		Z	0,117	0,562	0,522
	Stadt.	U	0,482	5,678	0,934
		Z	0,835	4,38	0,567
Stadtteile	Bez.	U	0,032	0,862	1
		Z	0,148	0,552	0,478
	Stadt.	U	0,646	6,367	0,947
		Z	1,078	4,336	0,66
Gauss-Krüger	Bez.	U	0,312	0,312	0,826
		Z	0,152	0,152	0,522
	Stadt.	U	2,694	2,679	0,649
		Z	1,581	1,569	0,46
2 Kl. Bezirke	Bez.	U	0,312	0,312	0,826
		Z	0,153	0,153	0,565
	Stadt.	U	2,642	2,645	0,636
		Z	1,527	1,513	0,433
Stadtteile	Bez.	U	0,524	0,524	0,043
		Z	0,172	0,172	0,652
	Stadt.	U	2,813	2,797	0,848
		Z	1,758	1,746	0,693
					0,713

Zuerst erfolgt ein Vergleich zwischen den beiden Modellierungsmöglichkeiten. Für die Kategorie *Ablehnung* zeigt die Tabelle die erwarteten Ergebnisse, sowohl die mittlere quadratische Abweichung als auch die Überdeckungswahrscheinlichkeit ist durchweg für alle Kombinationen besser im zwei Klassenmodell als im drei Klassenmodell. Die Überdeckung im drei Klassen Fall ist so gering, dass der wahre Wert nahezu nie enthalten ist. Für die Kategorie *Zustimmung* ergibt sich ein komplett konträres Ergebnis. Hier ist die mittlere quadratische Abweichung bei jeder Kombination im zwei Klassenmodell deutlich geringer als im drei Klassenmodell. Auch die Überdeckungswahrscheinlichkeit ist insbesondere bei den Ergebnissen zum Melderegister signifikant höher. Zu besseren Einordnung dieser Ergebnisse wurde Abbildung 6 erstellt. Dort ist zu sehen, dass die prognostizierten Anteile des drei Klassenmodells sehr nahe bei den wahren Anteilen liegen, wobei Punkte auf der 45 Grad Linie Beobachtungen signalisieren bei denen der wahre und der prognostizierte Anteil nahezu gleich sind. Beim zwei Klassenmodell ist zu erkennen, dass die wahren Anteile deutlich überschätzt werden, sowohl auf Bezirks- als auch auf Stadtteilebene.

Innerhalb des drei Klassenmodells lässt sich anhand von Tabelle 9 nicht eindeutig bestimmen,

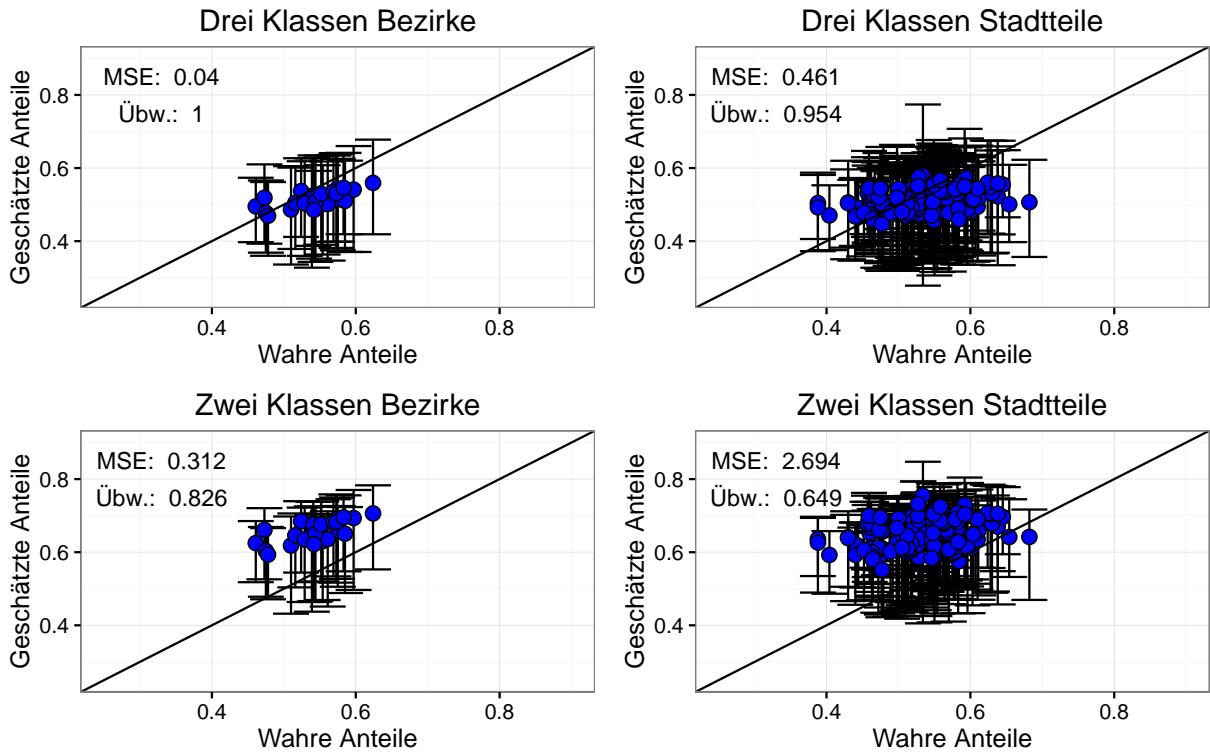


ABBILDUNG 6: ILLUSTRATION GESCHÄTZTE GEGEN WAHRE ANTEILE FÜR GAUSS-KRÜGER INFORMATIONEN EXTRAPOLIERT AUF MELDEREGISTER MIT ZWEI UND DREI KLASSENMODELL MIT 95% QUANTIELEN

welcher räumliche Effekt das beste Modell liefert, da die größten Unterschiede sowohl bei dem MSE und der Überdeckung zwischen den Dateien auftreten, was auch darauf hinweisen kann, dass der Melderegister näher mit der Volksabstimmung übereinstimmt. Beim Melderegister liefert das Modell mit Stadtteilen als räumlichem Effekt das beste Modell und auf Bezirksebene und auf Stadtteilebene das Modell mit Gauss-Krüger Informationen, wobei die Unterschiede eher marginal sind. Insgesamt kann das zwei Klassenmodell aus bestes Prognosemodell verworfen werden, da das drei Klassenmodell die *Zustimmung* und folglich auch die *Ablehnung* deutlich genauer bestimmen kann.

## EXTRAPOLATION

## DISKUSSION

Die beiden Datensätze zur Grundgesamtheit stammen aus einer Melderegister mit 470.190 Beobachtungen und dem Zensus mit 380.238 Beobachtungen. Im Verhältnis zu den Grundgesamtheiten dieser Größenordnung sind 3143 Beobachtungen in der Stichprobe relativ gering, was eine gewisse Unsicherheit für die Extrapolation mit sich bringt [...].

Weiterhin ist zu beachten, dass Informationen zu dem monatlichen Netto Haushaltseinkommen in beiden Grundgesamtheiten fehlen und somit die Variable nicht für die Prognose verwendet werden kann. Auch war eine denkbare Erstellung von Proxy-Variablen nicht möglich. Eine genaue Auflistung der enthaltenen Variablen aus den Grundgesamtheiten ist im Anhang verfügbar. Die Arbeit zielt darauf ab, die Meinung der Befragten zu dem Projekt Stuttgart 21 und die Zu-

## FAZIT

---

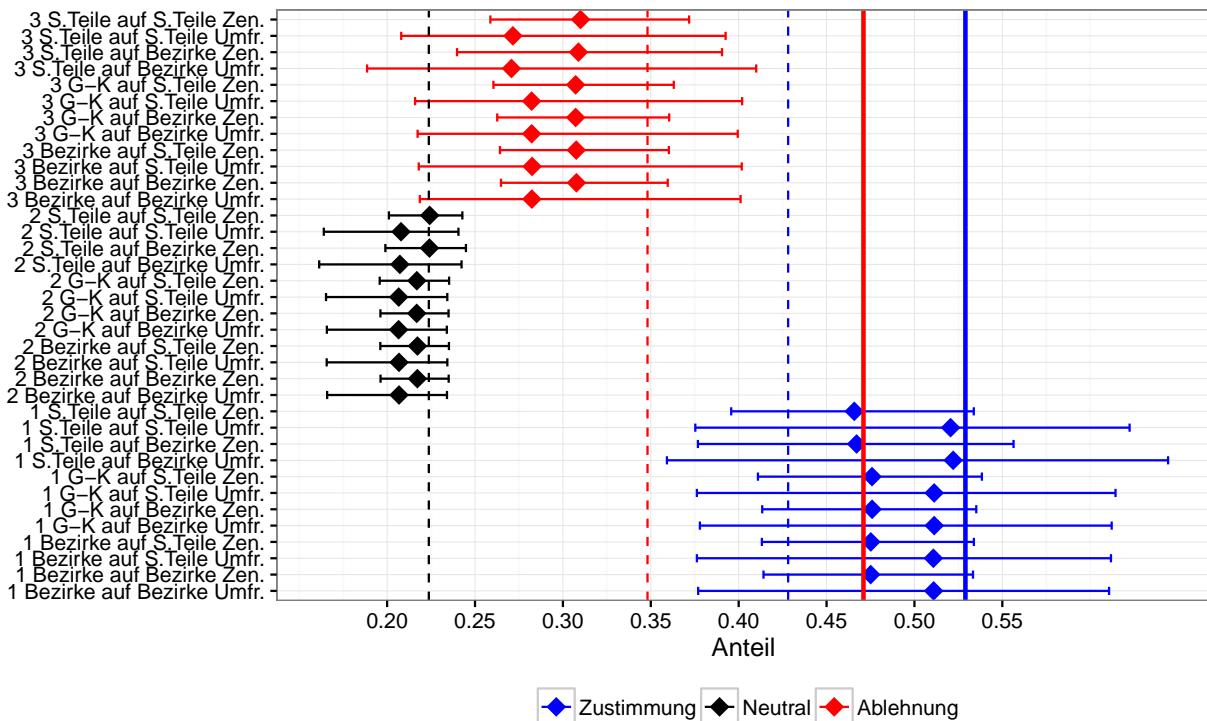


ABBILDUNG 7: VERGLEICH DER EXTRAPOLIERTEN GESAMTANTEILE FÜR STUTTGART MIT ALLEN GESCHÄTZTEN MODELLEN UND BEIDEN EXTRAPOLATIONSDATEIEN MIT DEN WAHREN ANTEILEN, SOWIE DEN ANTEILEN DER STICHPROBE MIT 95% QUANTILEN ZUR MEINUNG ZU STUTTGART 21

friedenheit mit der Wohngegend der Befragten auf die Grundgesamtheit zu extrapoliieren. Daher ist es sinnvoll die Ausprägungen dieser Variablen genauer zu untersuchen.

Insgesamt ist zu sagen, dass die Gauss-Krüger Informationen ein relativ klaren Einblick in die Verteilung der Beobachtungen in der jeweiligen Klasse geben. Bei den diskreten räumlichen Informationen könnte die grobe Aufteilung auf Bezirksebene zu einem Underfitting und die sehr feine Aufteilung auf Stadtteilebene zu einem Overfitting führen [...]. Zudem ist zu vermuten, dass die räumlichen Informationen einen stärkeren Effekt auf die Bewertung der Wohngegend haben, als auf die Meinung zu Stuttgart 21.

Ein  $B$  von 1.000 erwies sich als ausreichend, da sich die Mittelwerte und Quantile bei einer Erhöhung kaum noch änderten.

## FAZIT

---

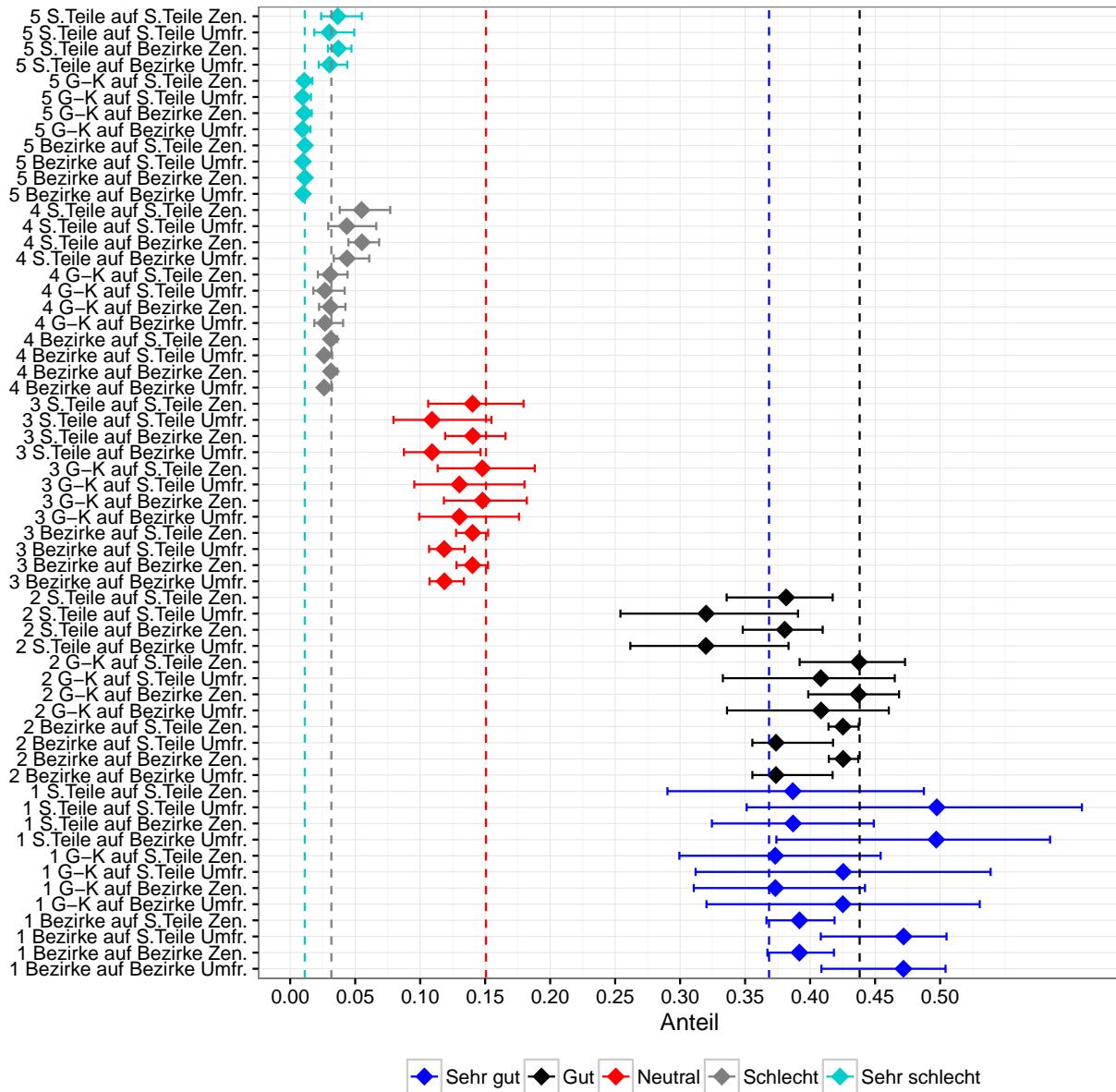


ABBILDUNG 8: VERGLEICH DER EXTRAPOLIERTEN GESAMTANTEILE FÜR STUTTGART MIT ALLEN GESCHÄTZTEN MODELLEN UND BEIDEN EXTRAPOLATIONSDATEIEN MIT DEN ANTEILEN DER STICHPROBE MIT 95% QUANTILEN ZUR BEWERTUNG DER WOHNGEgend

## LITERATUR

- [Akaike, 1981] Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, pages 3–14.
- [DiCiccio and Efron, 1996] DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.
- [Eilers and Marx, 1996] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.
- [Fahrmeir et al., 2009] Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression. Statistik und ihre Anwendungen*. Springer Berlin Heidelberg.
- [Fahrmeir et al., 2013] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer, Dordrecht.
- [Georgii, 2009] Georgii, H.-O. (2009). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter.
- [Held, 2008] Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes*. Spektrum Akad. Verl., Heidelberg.
- [Landeshauptstadt Stuttgart, 2015] Landeshauptstadt Stuttgart (2015). Erste Ergebnisse der Stuttgarter Bürgerumfrage 2015 - Stadt Stuttgart.
- [Lawrence D. Brown, 2001] Lawrence D. Brown, T. Tony Cai, A. D. (2001). Interval estimation for a binomial proportion. *Statistical Science*.
- [Stuttgart, 2011] Stuttgart, S. (2011). <http://www.stuttgart.de/volksabstimmung>.
- [Venables and Ripley, 2002] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.
- [Wickham, 2009] Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- [Wood, 2011] Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.
- [Wood et al., 2016] Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of The American Statistical Association*, pages 1–45. arXiv: 1511.03864.

ANHANG

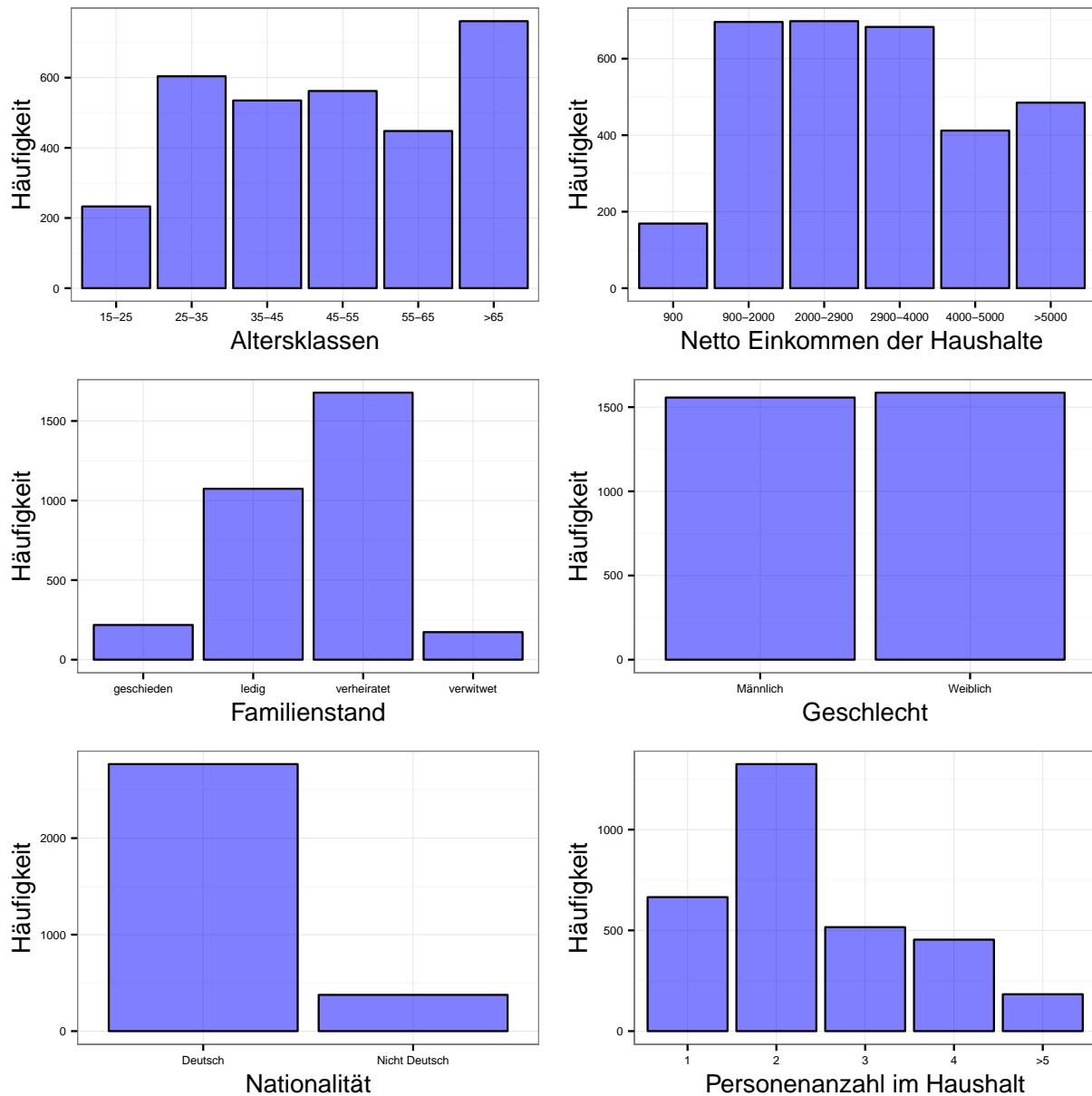


ABBILDUNG 9: HÄUFIGKEIT DER KATEGORIENAUSPRÄGUNGEN DER EXOGENEN VARIABLEN IN DER PARAMETERISIERUNGSSTICHPROBE.

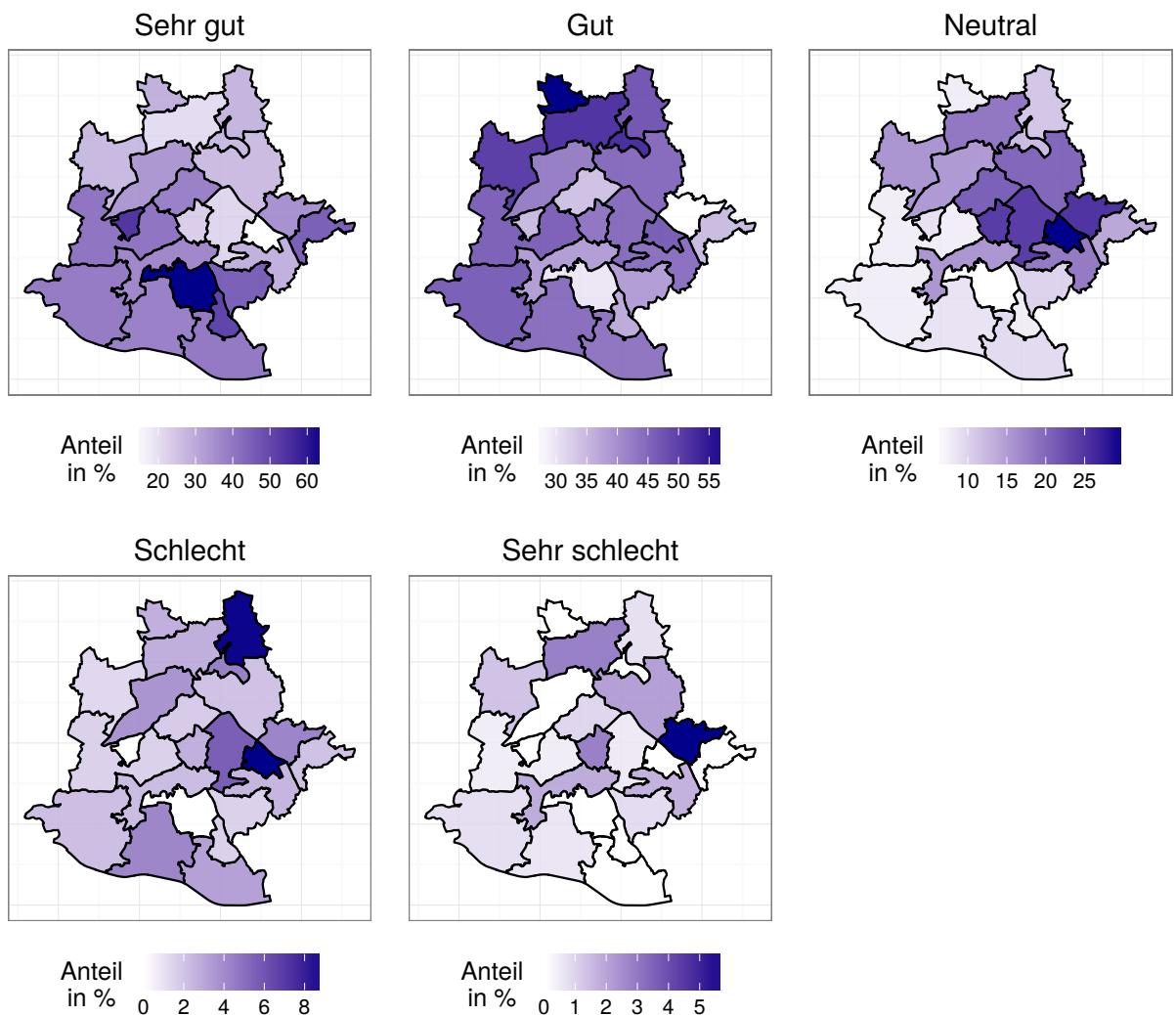


ABBILDUNG 10: ANTEILE DER BEWERTUNG DER WOHNGEGEND NACH STADTBEZIRKEN.

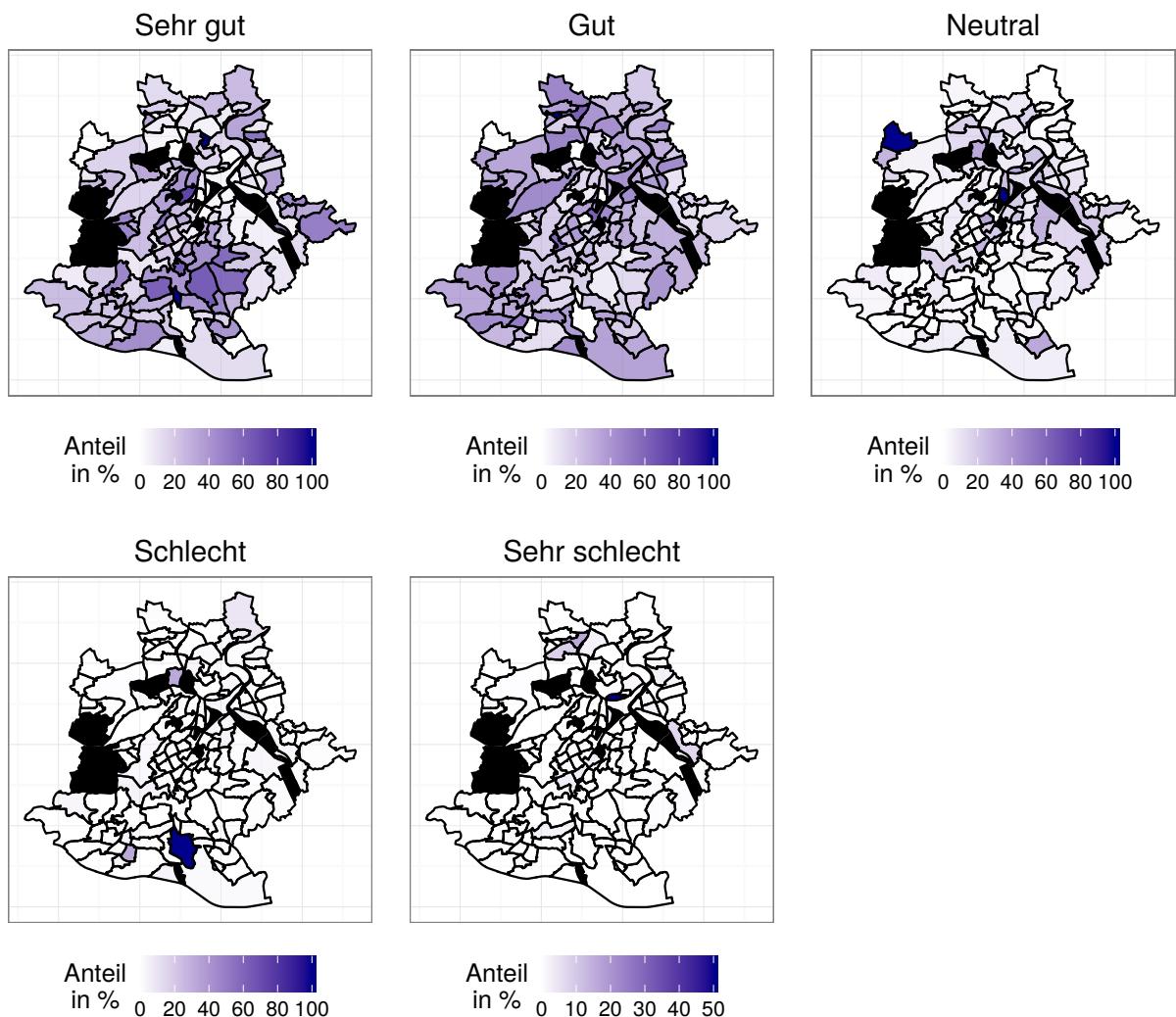


ABBILDUNG 11: ANTEILE DER BEWERTUNG DER WOHNGEGEND NACH STADTTEILEN.

Hiermit versichere ich, dass ich die vorliegende Hausarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle wörtlich oder sinngemäß den Schriften anderer entnommenen Stellen habe ich unter Angabe der Quellen kenntlich gemacht. Dies gilt auch für beigelegte Zeichnungen, Skizzen, bildliche Darstellungen und dergleichen.

Mir ist bewusst, dass ich mich im Falle einer unbeabsichtigten oder vorsätzlichen Missachtung durch den fehlerhaften Umgang mit Quellen unter Umständen strafbar mache und die vorliegende Hausarbeit mit nicht ausreichend bewertet wird.

Göttingen, den  
Unterschrift

Hiermit erlaube ich, dass meine Arbeit auf Betrug und falsche, sowie fehlende Zitate auch online geprüft wird.

Mir ist bewusst, dass ich mich im Falle einer unbeabsichtigten oder vorsätzlichen Missachtung durch den fehlerhaften Umgang mit Quellen unter Umständen strafbar mache und die vorliegende Hausarbeit mit nicht ausreichend bewertet wird.

Göttingen, den  
Unterschrift