



KLEINRÄUMIGE EXTRAPOLATION VON UMFRAGEDATEN

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

INHALTSVERZEICHNIS

1 Einleitung	1
2 Material und Methoden	2
2.1 Daten	2
2.1.1 Parametrisierungsstichprobe	2
2.1.2 Zensus	6
2.2 Statistische Methoden	7
2.2.1 Modell	7
2.2.2 Modellwahl	7
2.3 Evaluierung	8
2.4 Validierung	8
3 Ergebnisse	8
3.1 Validierung	8
3.2 Extrapolation	8
4 Fazit	8
Literatur	12
Anhang	13

ABBILDUNGSVERZEICHNIS

1	Endogene Variablen der Parameterisierungsstichprobe	3
2	Gauss Krüger Informationen Stuttgart 21	4
3	Gauss Krüger Informationen Bewertung Wohngegend	5
4	Anteile in Bezirken Stuttgart 21	6
5	Anteile in Stadtteilen Stuttgart 21	7
6	Validierung Stuttgart 21	10
7	Extrapolation aller Modelle	11
8	Anteile in Bezirken Bewertung Wohngegend	15
9	Anteile in Stadtteilen Bewertung Wohngegend	16

TABELLENVERZEICHNIS

1	Erhobene sozioökonomische und geographische Variablen der Parameterisierungsstichprobe und deren Anzahl der Ausprägungen sowie vermutete Modellierung im additiven Modell.	2
2	Validierung	9
3	Grundgesamtheit Bürgerumfrage	13
4	Grundgesamtheit Zensus	13

1 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

2 MATERIAL UND METHODEN

2.1 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. Die Datei wird im Folgenden als Parametrisierungsstichprobe bezeichnet. Bei der Parametrisierungsumfrage handelt es sich folglich um eine Stichprobe von der die Grundgesamtheit für eine Validierung nicht zur Verfügung steht. Alle Modellqualitätskriterien müssen entweder an der Stichprobe selbst oder an den beiden separat erhobenen Umfragen. Diese beiden anderen Umfragen haben jeweils einen deutlich größeren Stichprobenumfang. An diesen werden die parametrisierten Modelle angewendet und die Meinung zu Stuttgart 21 sowie die Wohnzufriedenheit somit kleinräumig extrapoliert. Die Bürgerumfrage von 20?? liegt mit einem Stichprobenumfang von 470.190 Befragten relativ nah an der Grundgesamtheit von 573.104 Bürgern, die 2011 mit Hauptwohnsitz in Stuttgart gemeldet waren (Stat. Bundesamt). Beim Zensus (Stuchjahr 2011) wurden 380.238 Bürger befragt.

2.1.1 PARAMETRISIERUNGSTICHPROBE

Mit den Datensätzen der Parametrisierungsstichprobe (Tabelle 1) werden die Modelle zur kleinräumigen Extrapolation parametrisiert. Insgesamt beinhaltet die Umfrage 8 sozioökonomische Variablen und Angaben zur räumlichen Lage. Von jedem Datensatz sind die stetige räumliche Lage als Gauß-Krüger Geokoordinate sowie die diskrete räumliche Lage in 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 der Tabelle sind nicht nur die Kategorienanzahlen der Variablen, sondern auch die vermuteten Formen der Einflüsse der Kovariablen auf die abhängigen Variablen aufgelistet. Diese wurden durch visuelle Darstellung aller Kovariablen über den abhängigen Variablen ermittelt. Nach

visueller Einschätzung ergab sich, dass alle nominal skalierte Variablen, wie z.B. die Nationalität, als parametrisch und dass alle kardinal skalierte Variablen, wie z.B. die Altersklasse des Befragten, offensichtlich als nicht parametrisch modelliert werden sollten. Diese beobachteten Zusammenhänge finden sich häufig in Regressionsmodellen [Fahrmeir et al., 2009, p. 9]. In Anhang an ZITAT Fahrmeier (2013) S. 503-510 wird der räumliche Effekt entweder kontinuierlich als Tensor Produkt (p. 503-510) oder diskret als Gauß-Markov Zufallsfeld (p. 524-527) modelliert. Für die Auswahl der geeigneten Regressionsmethode ist es hilfreich, das Verhältnis der Häufigkeiten der Kategorienausprägungen der abhängigen Variable zu kennen uns seltene Ereignisse zu identifizieren. Während die meisten befragten Personen ihre Wohngegend mit *gut* oder *sehr gut* bewertet haben, Ereignisse mit schlechterer Einschätzung folglich deutlich seltener auftreten, sind die Gruppenhäufigkeiten der Antworten zum Projekt Stuttgart 21 näherungsweise gleichverteilt. In beiden Fällen wurden die wenigen, für die Modellierung irrelevanten Kategorien 'Keine Angabe' entfernt. NOCH DIE K.A. ENTFERNEN! In beiden Fällen kann den Gruppen eine Rangfolge, jedoch kein Intervall Abstand unterstellt werden. Es handelt sich demnach jeweils um ordinal Skalierte Daten.

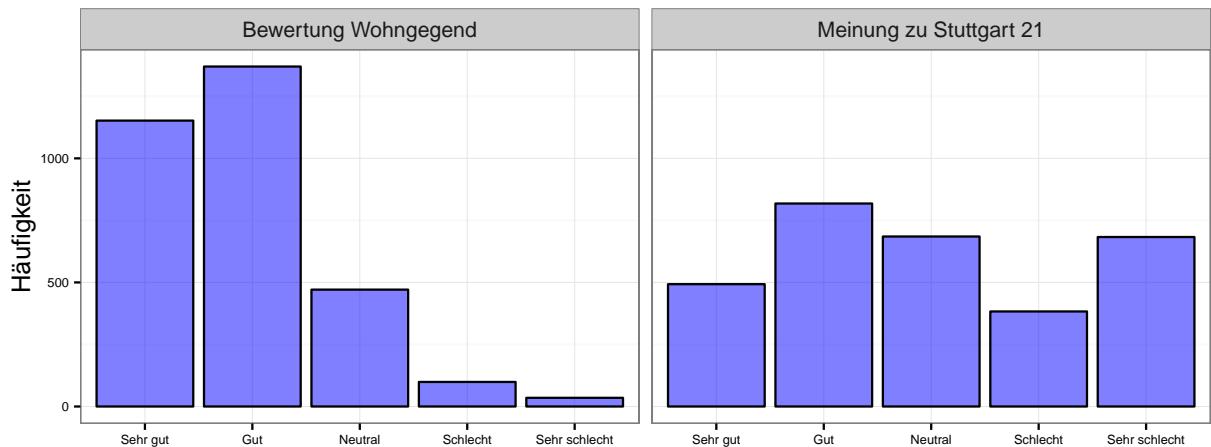


ABBILDUNG 1: ENDOGENE VARIABLEN DER PARAMETERISIERUNGSSTICHPROBE.

Das amtliche, nach Stadtteilen oder Stadtbezirken aufgelöste Ergebnis der Bürgerumfrage zu Stuttgart 21 von 2011 kann dem Internetauftritt der Stadt entnommen werden [Stuttgart, 2011]. Es bietet eine zusätzliche Möglichkeit zur Modellevaluierung. Da hier nur 3 Kategorien (Zustimmung, Enthaltung, Ablehnung) ICH VERSTEHE NICHT GANZ VOHER DIE GRUPPE ENTHALTUNG KOMT unterschieden wurden, wurden auch die Daten dieser Arbeit neu gruppiert. Es wurde, ergänzend zur Einteilung in 5 Kategorien (Abbildung 1, links), eine Neugruppierung in 3 (Zustimmung, Neutral, Ablehnung) sowie in 2 Gruppen (Zustimmung, Ablehnung) vorgenommen. Für die exogen in die Analyse einfließenden Variablen sind detailliertere Informationen zu den Häufigkeiten der Ausprägungen im Anhang verfügbar (REF ABB A.1).

Da in dieser Arbeit ein Schwerpunkt auf die unterschiedlichen räumlichen Effekte gelegt werden soll, vergleicht dieses Kapitel alle drei räumlichen Effekte in Relation zu den beiden endogenen Variablen.

Aus Abbildung 2 ist die absolute Häufigkeit der Beobachtungen für jede der drei Kategorien zur Meinung zu Stuttgart 21 ersichtlich. Zuerst ist erkenntlich wo die meisten Personen befragt wur-

den, nämlich in der Nähe des Innenstadt Bereichs. Hier findet sich eine Häufung der Beobachtung zu allen drei Kategorien, wohingegen die Randbezirke im allgemeinen weniger Beobachtungen aufweisen. Außerdem ist zu sehen, dass auf Grund der geographischen Beschaffenheit der Region in einigen Bereichen mit weniger Beobachtungen zu rechnen ist z.B. beim Wald im Westen und der hügeligen Landschaft im Osten.

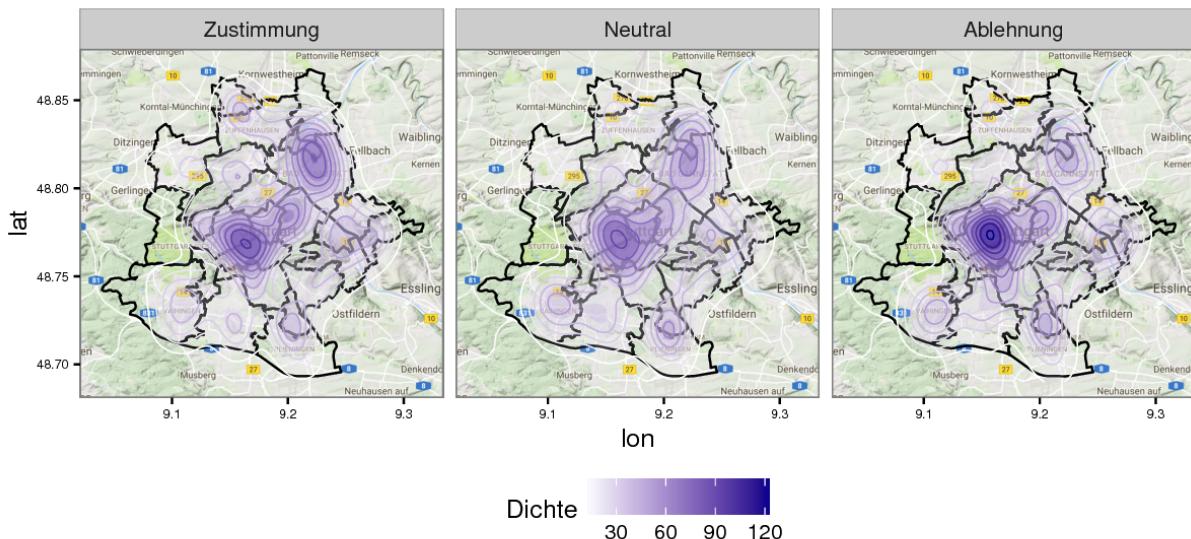


ABBILDUNG 2: GAUSS KRÜGER INFORMATIONEN STUTTGART 21

Zu der spezifischen Häufung der Beobachtungen innerhalb der Klassen ist ein leichter Trend zu erkennen, sodass für die Kategorie *Zustimmung* eine größere Häufung der Beobachtungen im Nordosten zu erkennen ist. Für die Kategorie *Ablehnung* ist eher noch eine stärkere Konzentration auf den Bereich Südwestlich der Innenstadt zu erkennen und einige kleinere Häufung im Süden Stuttgarts. Die Beobachtungen der Kategorie *neutral* sind eher gleichmäßig über die Stadt verteilt.

Abbildung 3 zeigt die Absolute Verteilung der Beobachtungen für die fünf Kategorien zur Bewertung der Wohngegend. Hier zeigt sich ein deutlich heterogenes Bild als bei der Meinung zu Stuttgart 21. Die Beobachtungen der Klasse *Sehr gut* häufen sich sehr stark im Innenstadt Bereich und im Süden. Die Kategorie *gut* verteilt sich über das gesamte Stadtgebiet mit einer stärkeren Konzentration in der Innenstadt und im Nordosten. Für die Klasse *Neutral* zeigt sich schon eine stärkere Konzentration auf den Osten und Nordosten der Stadt im Vergleich zu den beiden vorherigen Klassen. Am interessantesten ist hier die Lokalisierung der Personen die ihre Wohngegend mit *schlecht* oder *sehr schlecht* bewertet haben. Hier finden sich nahezu alle Beobachtungen im Osten und Nordosten Stuttgarts.

Allerdings ist auch zu beachten, dass der Anteil der Personen, die ihre Wohngegend mit *schlecht* oder *sehr schlecht* bewertet haben sehr gering ist im Vergleich zu allen Befragten wie aus Abbildung 1 ersichtlich.

Als nächstes können die diskreten räumlichen Informationen auf Stadtbezirksebene untersucht werden. Dazu wurde Abbildung 4 erstellt. Hier sind die Anteile der drei Klassen für die jeweiligen Bezirke zu erkennen. Zu sehen ist, dass der Anteil der Personen die das Projekt Stuttgart 21

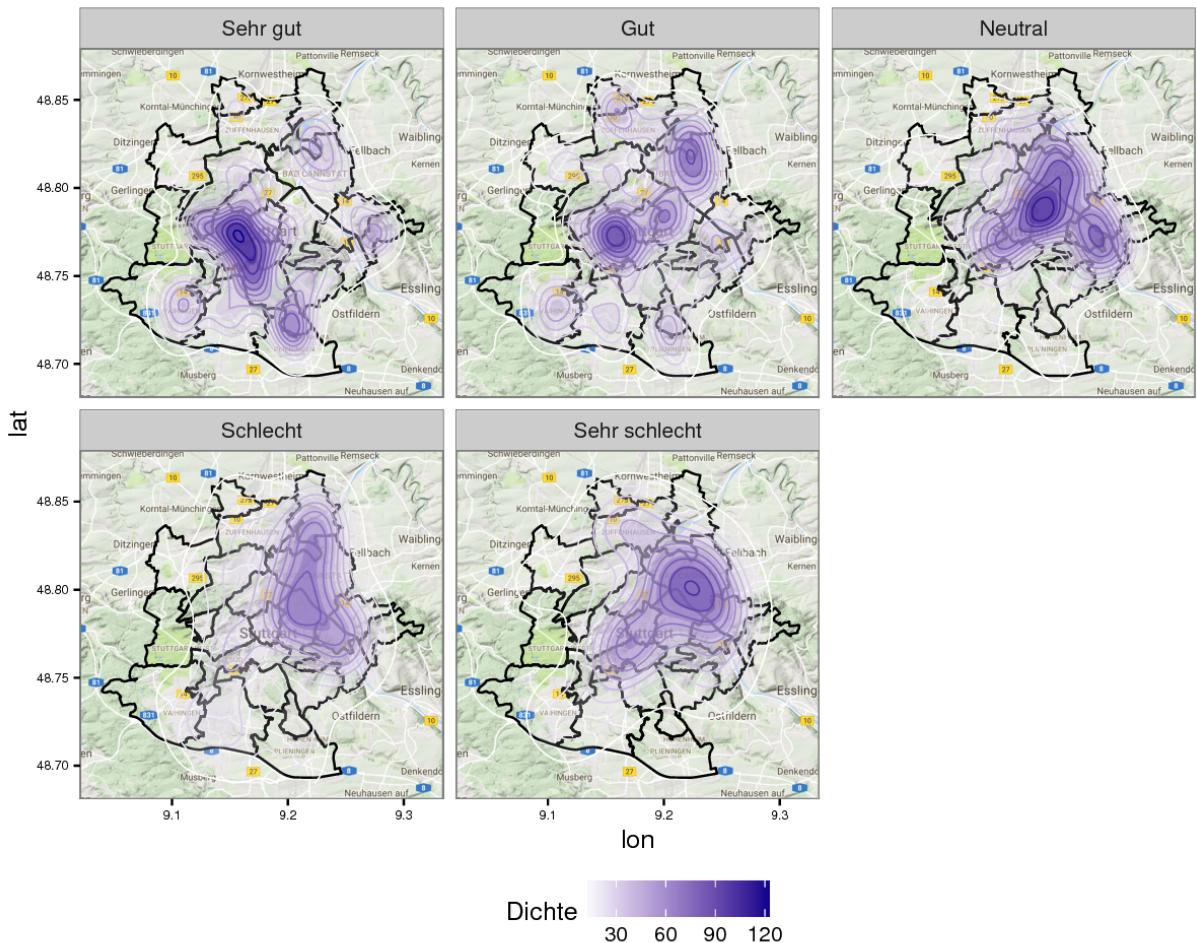


ABBILDUNG 3: GAUSS KRÜGER INFORMATIONEN BEWERTUNG WOHNGEGEND

positiv bewerten in allen Stadtbezirken relativ hoch ist, wobei der Anteil in den nordöstlichen Bezirken etwas höher ist. Die neutrale Klasse hat in allen Bezirken einen eher geringeren Anteil und es ist kein deutliches Muster von höheren oder niedrigeren Anteilen erkennbar. Die Klasse *Ablehnung* hat etwas höhere Anteile in den mittleren und südlichen Bezirken. Insgesamt deckt sich das Bild der Anteile auf Bezirksebene mit den absoluten Häufigkeiten aus Abbildung 2. Die Anteile auf Bezirksebene mit fünf Klassen für die Bewertung der Wohngegend sind im Anhang verfügbar. Zu beachten ist hier, dass die Anteile auf fünf verschiedenen Skalen dargestellt werden, da der Anteil der beiden negativen Klassen zu gering ist im Verhältnis zu den positiven Klassen, um sie anders darzustellen. Aber auch hier deckt sich die Verteilung der hohen und niedrigen Anteile mit den Absoluten Anzahlen an Beobachtungen aus Abbildung 3.

Die letzte räumliche Information ist der Anteil der jeweiligen Klasse pro Stadtteil. Wie schon aus Tabelle 1 bekannt, besitzt die Stadtteilebene eine wesentlich feinere Aufteilung, als die Bezirksebene. Dadurch ist es möglich, dass in einzelnen Stadtteilen keine Beobachtungen einer Klasse auftauchen. Damit kann die Ansicht auf Stadtteilebene zu einer Überschätzung der Bedeutung einer bestimmten Klasse in einem Stadtteil führen. Die Anteile der drei Klassen zu der Meinung zu Stuttgart 21 sind aus Abbildung 5 zu entnehmen.

Die schwarz eingefärbten Flächen signalisieren Stadtteile in denen von keiner der drei Klassen eine

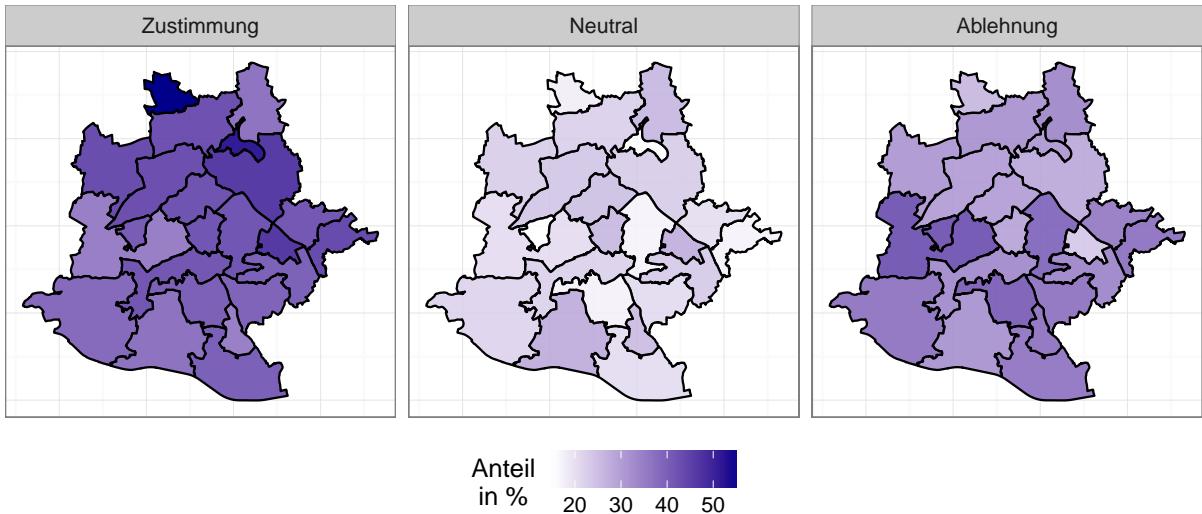


ABBILDUNG 4: ANTEILE IN BEZIRKEN STUTTGART 21

Beobachtung vorhanden sind. Wie bereits erwähnt handelt es sich dabei vor allem um Regionen die auf Grund ihrer geographischen Beschaffenheit kaum Einwohner haben (z.B. Wald). Hier fügt sich das Bild der Anteile nicht ganz in den bisherigen Verlauf der räumlichen Statistiken ein. In keiner der drei Klassen lässt sich klar eine Struktur oder räumliche Korrelation der Stadtteile beobachten. Die Anteile auf Stadtteilebene zu der Bewertung der Wohngegend sind im Anhang verfügbar, wobei die Plots wieder mit eigenen Skalen illustriert wurden. Hier zeigt sich ein ähnlich unklares Muster wie bei der Meinung zu Stuttgart 21.

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.

2.1.2 ZENSUS

Die beiden Datensätze zur Grundgesamtheit stammen aus einer Bürgerumfrage 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 Zufriedenheit mit der Wohngegend der Befragten auf die Grundgesamtheit zu extrapoliieren. Daher ist es sinnvoll die Ausprägungen dieser Variablen genauer zu untersuchen. DIES IST ALLES

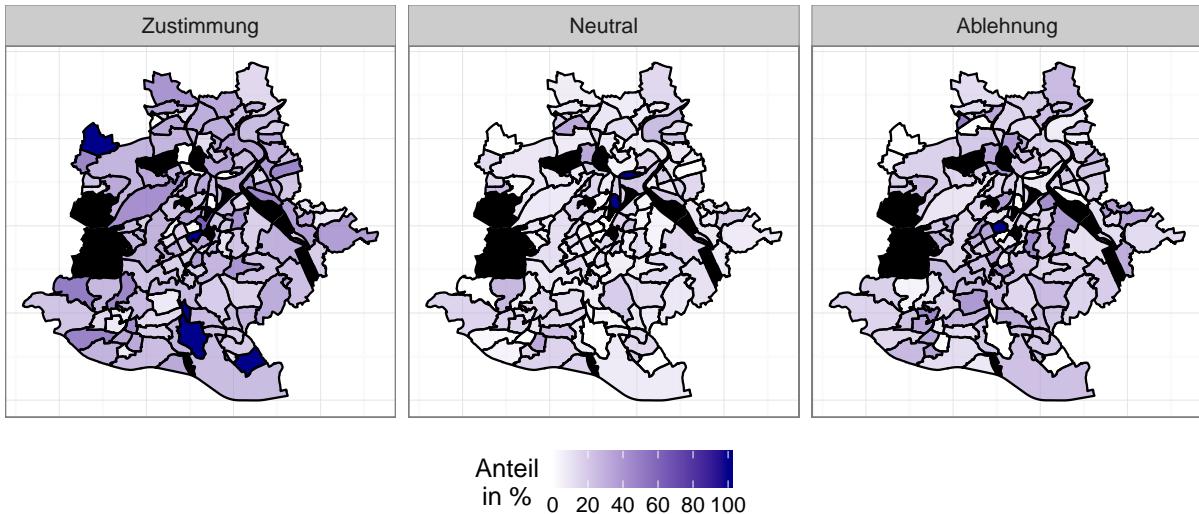


ABBILDUNG 5: ANTEILE IN STADTTEILEN STUTTGART 21

DISKUSSION; GANZER ABSATZ MUSS IN DISK. KAPITEL

2.2 STATISTISCHE METHODEN

2.2.1 MODELL

Kurze Erläuterung GAM und logit,... Zusammengesetzt aus parametrischen Effekten und Splines

2.2.2 MODELLWAHL

Basierend auf den vorhergegangenen Analysen des Beamten- und Eigenheimanteils in Stuttgart wurde die R Funktion `stepAIC()` zur schrittweisen AIC (CITE AKAIKE) Berechnung programmiert, die zur Identifikation der geeigneten Kovariablenkombination dient. In der Funktion werden mithilfe der `gam()` Funktion des `mgcv` CITE WOOD 2011 Paketes 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 1 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 Kovariablen in Iteration 1 zu einer Reduktion des AIC führt, wird diejenige Kovariablen, welche zum 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 1 bereits beendet.

Andernfalls setzt sich das Ausgangsmodell für Iteration 2 aus den festen Kovariablen und einer weiteren Kovariable zusammen. In Iteration 2 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 1 nicht bereits das geeignete ist), werden alle veränderbaren Kovariablen in Iteration 2 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.

2.3 EVALUIERUNG

2.4 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 des Prognosemodells geben zu können. Die Validierung erfolgt auf Stadtteil- und Bezirksebene, sowie für das Gesamtergebnis von Stuttgart. Dafür 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 [Georgii, 2009, p. 201]. Ein weiteres Kriterium ist die Überdeckungswahrscheinlichkeit.

3 ERGEBNISSE

3.1 VALIDIERUNG

3.2 EXTRAPOLATION

4 FAZIT

TABELLE 2: VALIDIERUNG

		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
	Stadtteile	U	0.482	5.678	0.934
		Z	0.835	4.38	0.567
Stadtteile	Bez.	U	0.022	0.903	
		Z	0.081	0.593	
	Stadtteile	U	0.453	6.69	
		Z	0.61	4.641	
Gauss-Krüger	Bez.	U	0.312	0.312	0.826
		Z	0.152	0.152	0.522
	Stadtteile	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
	Stadtteile	U	2.642	2.645	0.636
		Z	1.527	1.513	0.433
Stadtteile	Bez.	U	0.468	0.468	
		Z	0.201	0.201	
	Stadtteile	U	3.741	3.723	
		Z	1.906	1.893	

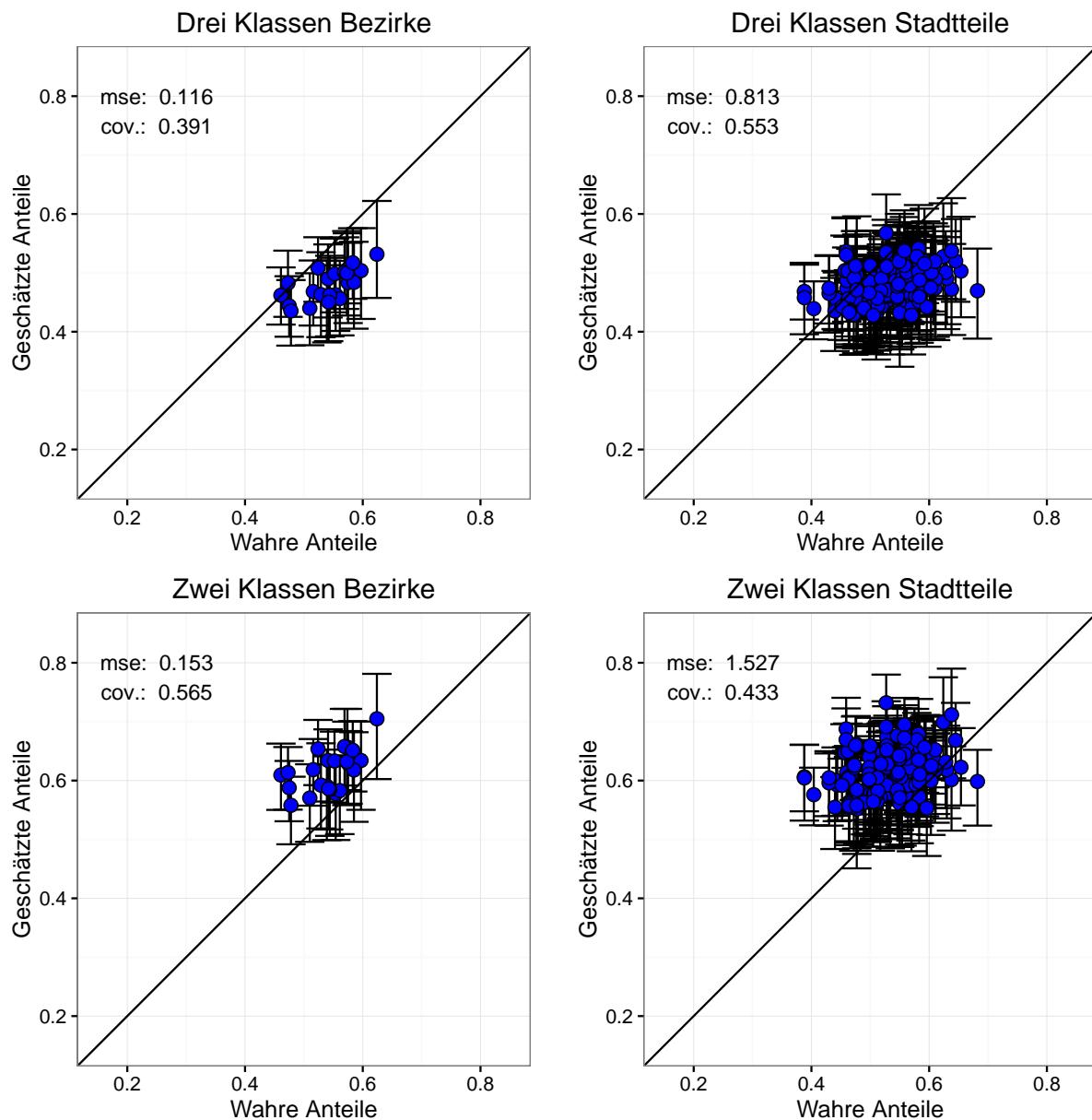


ABBILDUNG 6: VALIDIERUNG STUTTGART 21

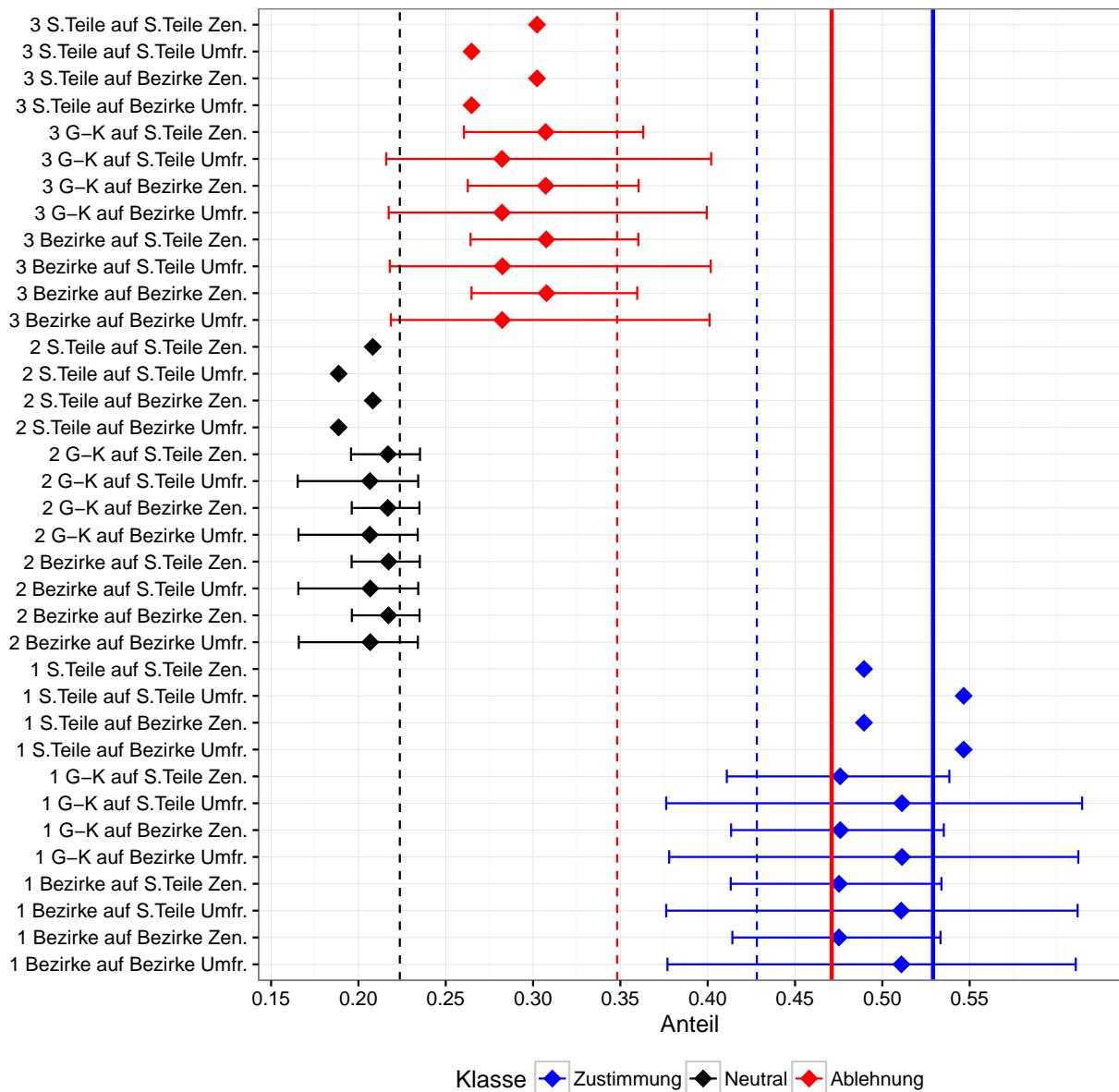


ABBILDUNG 7: EXTRAPOLATION ALLER MODELLE

LITERATUR

- [Fahrmeir et al., 2009] Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression. Statistik und ihre Anwendungen*. Springer Berlin Heidelberg.
- [Georgii, 2009] Georgii, H.-O. (2009). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter.
- [Stuttgart, 2011] Stuttgart, S. (2011). <http://www.stuttgart.de/volksabstimmung>.

ANHANG

TABELLE 3: GRUNDGESAMTHEIT BÜRGERUMFRAGE

Anzahl Beobachtungen: 470.190

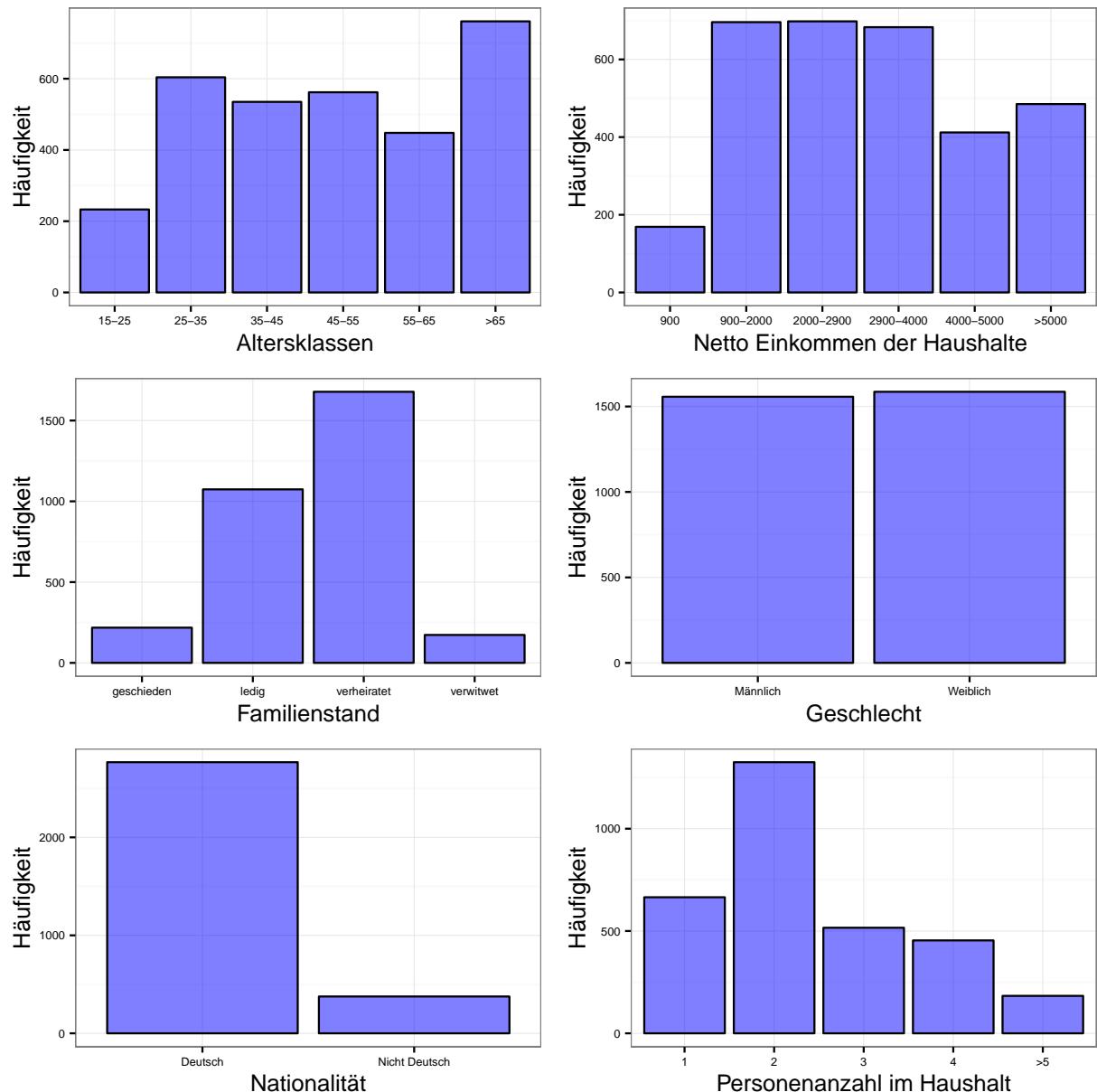
Variable	Modellierung	Mögliche Ausprägungen
Altersklasse	Nicht Parametrisch	14
Geschlecht	Parametrisch	2
Nationalität	Parametrisch	2
Familienstand	Parametrisch	4
Haushaltsgröße	Nicht Parametrisch	5
Wohndauer	Nicht Parametrisch	3
ALG II Quote	Nicht Parametrisch	9
Ein/Zweifamilienhäuser	Nicht Parametrisch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

TABELLE 4: GRUNDGESAMTHEIT ZENSUS

Anzahl Beobachtungen: 380.238

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

ANHANG



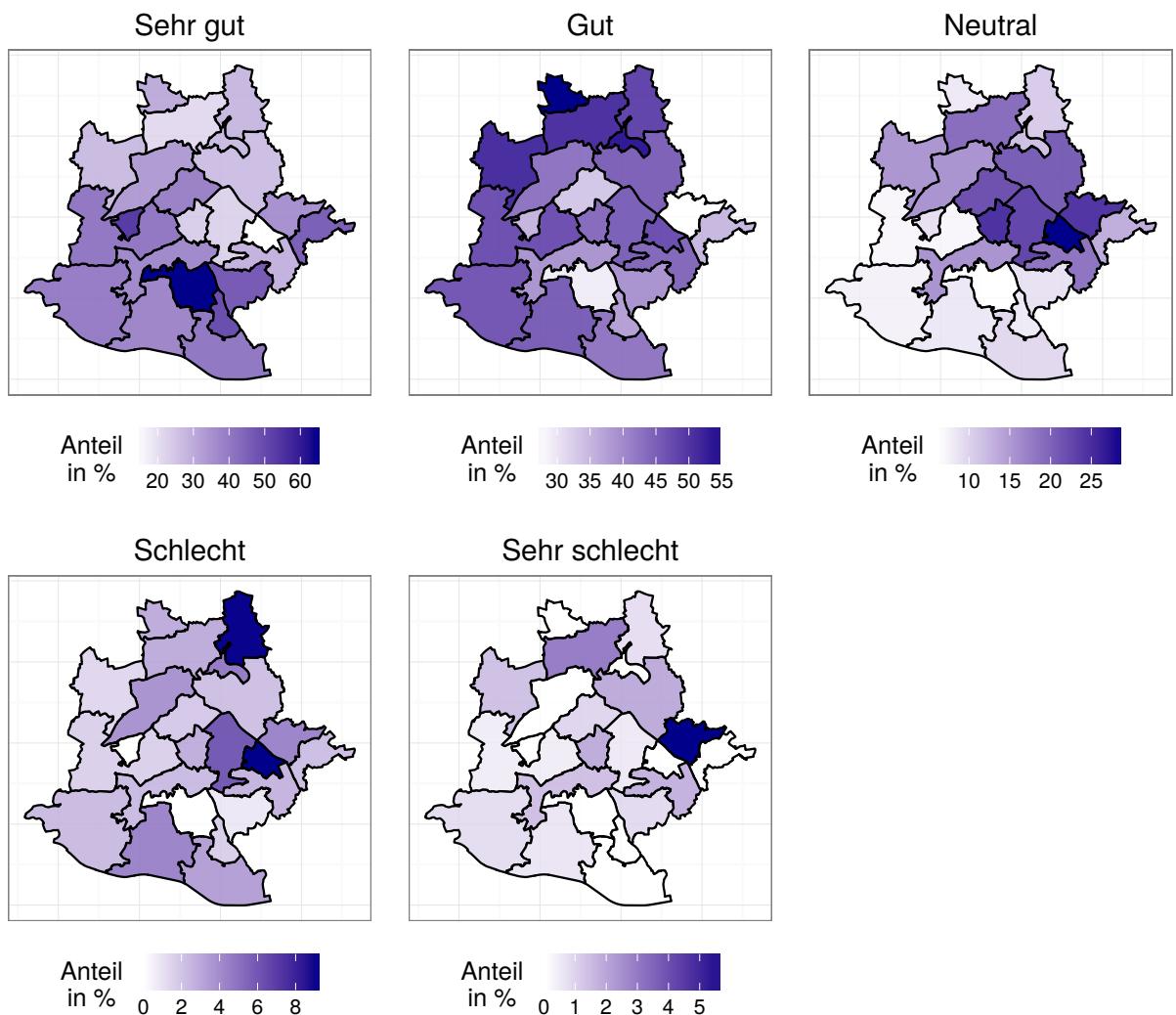


ABBILDUNG 8: ANTEILE IN BEZIRKEN BEWERTUNG WOHNGEgend

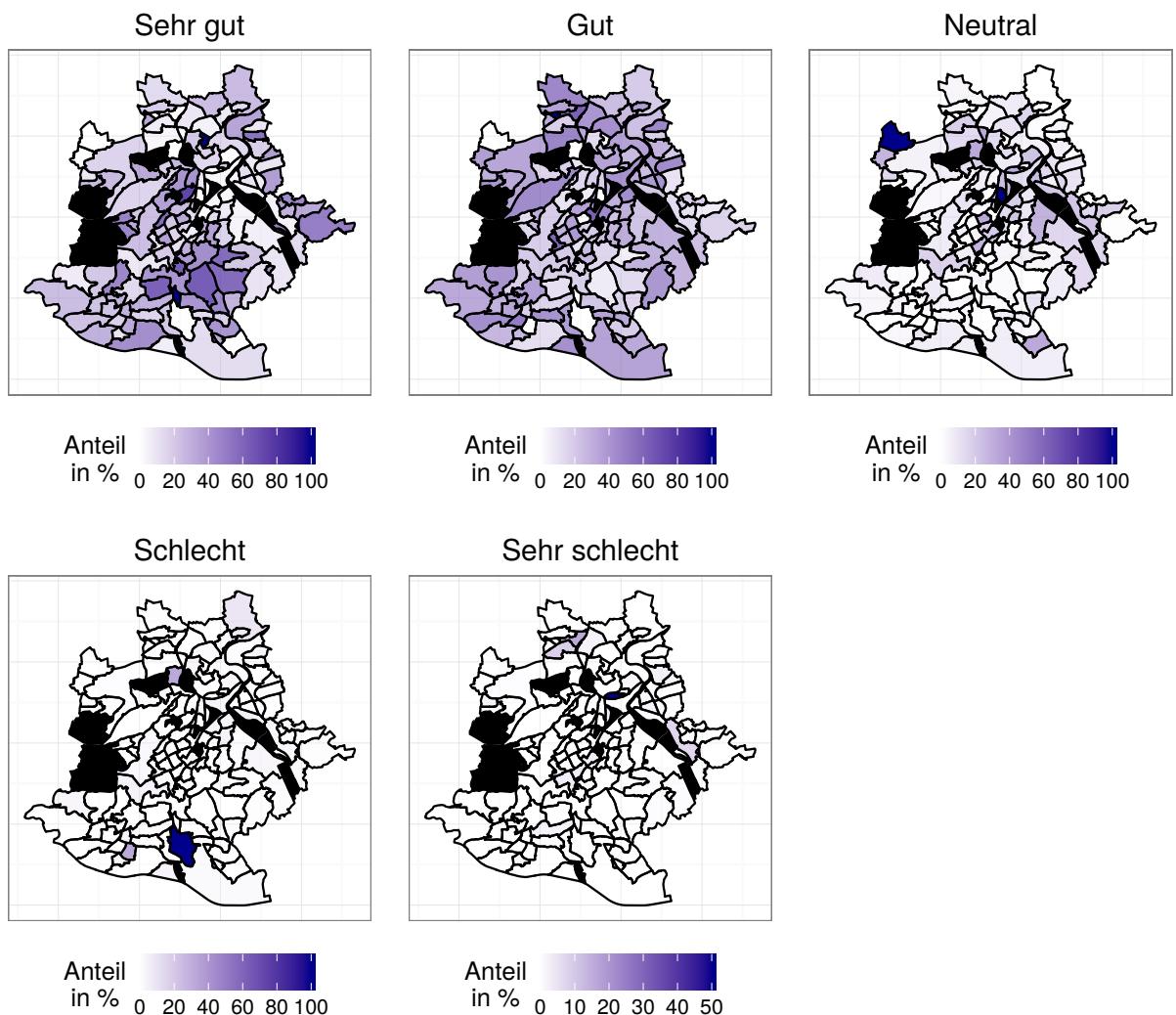


ABBILDUNG 9: ANTEILE IN STADTTEILEN BEWERTUNG WOHNGEgend

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