



# 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

<b>1 Einleitung</b>	<b>1</b>
<b>2 Material und Methoden</b>	<b>2</b>
<b>3 Daten</b>	<b>2</b>
3.1 Deskriptive Statistik . . . . .	2
3.2 Räumliche Effekte . . . . .	5
<b>4 Statistische Methoden</b>	<b>8</b>
4.1 Modell . . . . .	8
4.2 Modellwahl . . . . .	8
4.3 Evaluierung . . . . .	9
<b>5 Ergebnisse</b>	<b>9</b>
<b>6 Fazit</b>	<b>9</b>
<b>Literatur</b>	<b>10</b>
<b>Anhang</b>	<b>11</b>

## ABBILDUNGSVERZEICHNIS

1	Endogene Variablen . . . . .	3
2	Gauss Krüger Informationen Stuttgart 21 . . . . .	5
3	Gauss Krüger Informationen Bewertung Wohngegend . . . . .	6
4	Anteile in Bezirken Stuttgart 21 . . . . .	7
5	Anteile in Stadtteilen Stuttgart 21 . . . . .	7
6	Anteile in Bezirken Bewertung Wohngegend . . . . .	13
7	Anteile in Stadtteilen Bewertung Wohngegend . . . . .	14

## TABELLENVERZEICHNIS

1	Datensatz . . . . .	2
2	Validierung . . . . .	9
3	Grundgesamtheit Bürgerumfrage . . . . .	11
4	Grundgesamtheit Zensus . . . . .	11

1 EINLEITUNG

## 2 MATERIAL UND METHODEN

## 3 DATEN

Insgesamt liegen für die Analysen drei Umfragen vor. Die kleinste Datei enthält Angaben zur Bewertung der Wohngegend, der Meinung zu Stuttgart 21 sowie weitere sozioökonomische Kovariablen, die zur Erklärung der beiden unabhängigen Variablen dienen. Diese Datei wird im Folgenden als Stichprobe bezeichnet. Die Bevölkerung Stuttgarts bildet die Grundgesamtheit dieser Untersuchung. Da die Zensus Umfrage von 20?? und die Bürgerumfrage von 20?? diese Grundgesamtheit sehr nah kommen, wird angenommen, dass diese beiden Umfragen jeweils die Grundgesamtheit repräsentieren.

### 3.1 DESKRIPTIVE STATISTIK

In diesem Kapitel soll der Stichprobendatensatz vorgestellt werden, sowie ein kurzer Einblick in die beiden Datensätze zur Grundgesamtheit der Bevölkerung von Stuttgart gegeben werden. Anhand des Stichprobendatensatzes soll im Verlauf der Arbeit das Modell zur Extrapolation geschätzt werden und mit Hilfe der Datensätze zur Grundgesamtheit die Häufigkeiten der abhängigen Variablen prognostiziert werden.

Tabelle 1 gibt einen Überblick über den Inhalt der Stichprobe. Wie zu erkennen ist, handelt es sich dabei um strukturell sozioökonomische Variablen. Zudem enthält der Datensatz drei verschiedene räumliche Informationen zu den Beobachtungen, auf die im nächsten Kapitel ausführlicher eingegangen wird.

TABELLE 1: DATENSATZ

Anzahl Beobachtungen: 3.143

Variable	Mögliche 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

Die rechte Spalte zeigt an, wie die einzelnen Variablen in das zu schätzende Modell einfließen sollen. Dabei ist zu erkennen, dass nominal skalierte Variablen, wie z.B. Nationalität als parametrisch und metrisch skalierte Variablen wie z.B. die Altersklasse der Befragten als nicht parametrisch modelliert werden sollen [Fahrmeir et al., 2009, p. 9]. Genauere Erläuterungen zur Modellierung der Variablen finden sich im Kapitel Methodik. 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.

Dazu wurde Abbildung 1 erstellt. Sie zeigt die Häufigkeiten der einzelnen Ausprägungen der beiden endogenen Variablen. Wie schon zuvor aus Tabelle 1 ersichtlich, besitzen beide Variablen sechs mögliche Realisationen, wobei die Klasse *keine Angabe* keine Informationen über die Meinung der Befragten liefert und diese Beobachtungen daher aus dem Datensatz entfernt werden müssen. Damit bleiben fünf mögliche Klassen zur Modellierung übrig.

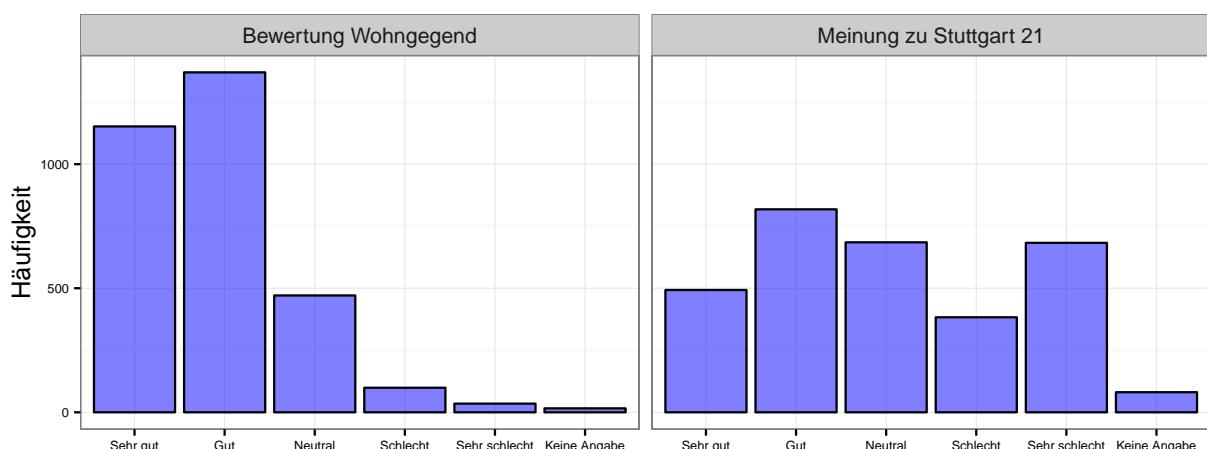


ABBILDUNG 1: ENDOGENE VARIABLEN

Aus Abbildung 1 ist zudem ersichtlich, dass die Verteilungen der Beiden Variablen sehr verschieden sind für die gleichen Antwortmöglichkeiten. Während die meisten befragten Personen ihre Wohngegend mit *gut* oder *sehr gut* bewertet haben, sind die Antworten zum Projekt Stuttgart 21 eher gleichmäßig verteilt.

Abschließend ist die Entscheidung gefallen die Bewertung der Wohngegend mit fünf Klassen zu Modellieren, um eine möglichst genau Prognose der Meinung der Bevölkerung zu erreichen. Zudem sind keinerlei Daten zur Bewertung der Wohngegend für die Grundgesamtheit verfügbar, womit auch keine Rücksicht auf eine mögliche Validierung des Modells genommen werden muss. Hingegen ist eine Validierung für die Meinung zu Stuttgart 21 mit den Ergebnissen der Volksabstimmung aus dem Jahre 2011 möglich [Stuttgart, 2011]. Allerdings bietet die Volksabstimmung nur Informationen für maximal drei Kategorien (*Zustimmung*, *Ablehnung*, *Enthaltung*), weshalb die weitere Analyse ebenfalls mit drei Kategorien erfolgen soll. Dafür werden die Kategorien *sehr gut* und *Gut* zusammengefasst zu *Zustimmung* und die Klassen *schlecht* und *sehr schlecht* zu *Ablehnung*. Die Klasse *neutral* soll als mittlere Kategorie erhalten bleiben. Für die exogen in die Analyse einfließenden Variablen sind detailliertere Informationen zu den Häufigkeiten der

---

## DATEN

---

Ausprägungen im Anhang verfügbar.

### 3.2 RÄUMLICHE EFFEKTE

Da in dieser Arbeit ein besonderer Schwerpunkt auf die unterschiedlichen räumlichen Effekte gelegt werden soll, vergleicht dieses Kapitel alle drei räumlichen Effekte für beide endogene Variablen. Zunächst wird der stetige räumliche Effekt mit den Gauss-Krüger Informationen behandelt. Anschließend folgen die diskreten räumlichen Informationen auf Bezirks und Stadtteilebene.

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 wurden, 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.

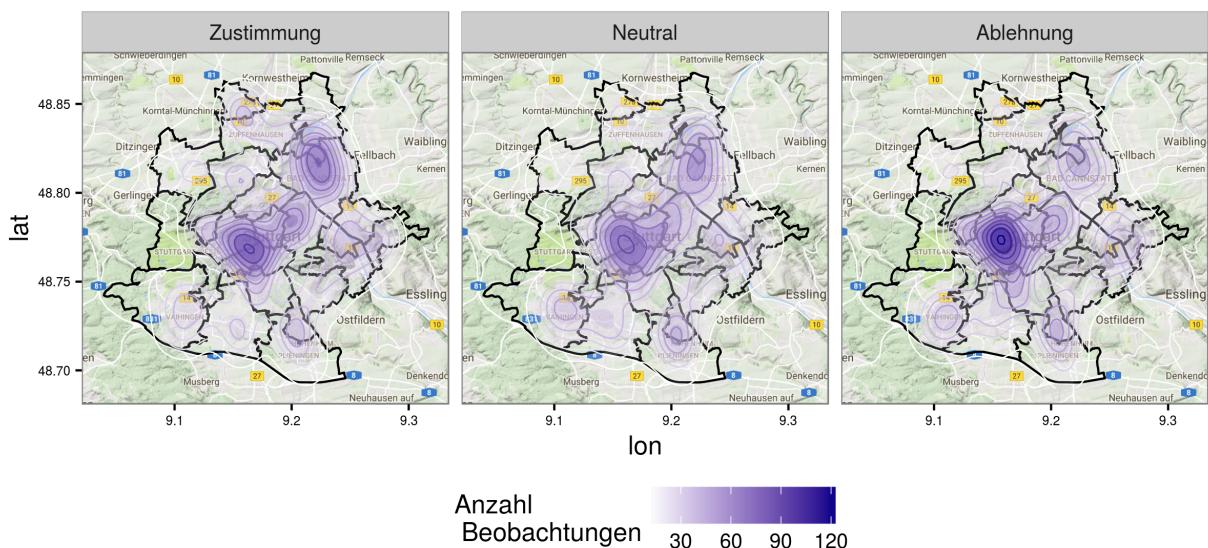


ABBILDUNG 2: GAUSS KRÜGER INFORMATIONEN STUTTGART 21

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 ist 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.

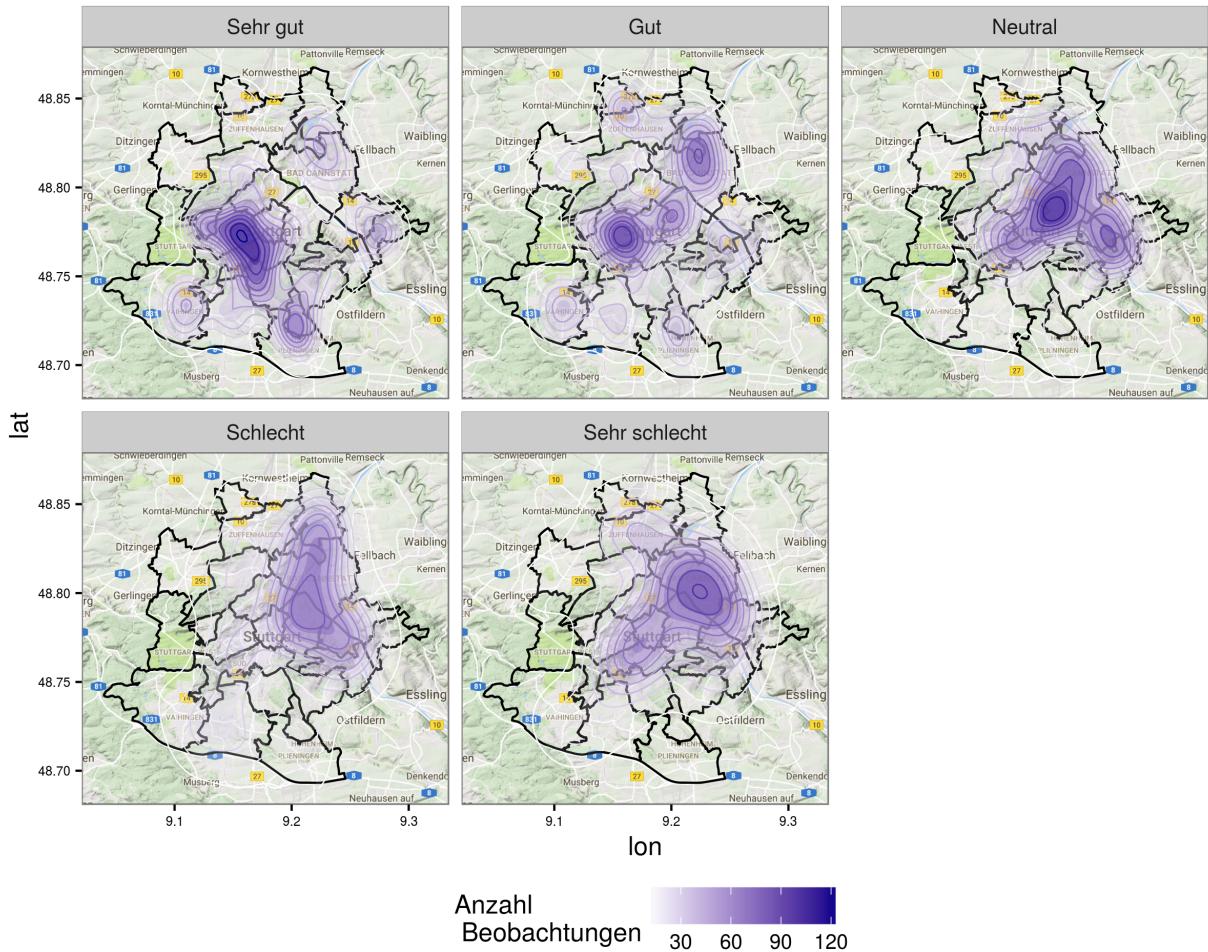


ABBILDUNG 3: GAUSS KRÜGER INFORMATIONEN BEWERTUNG WOHNGESEND

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 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.

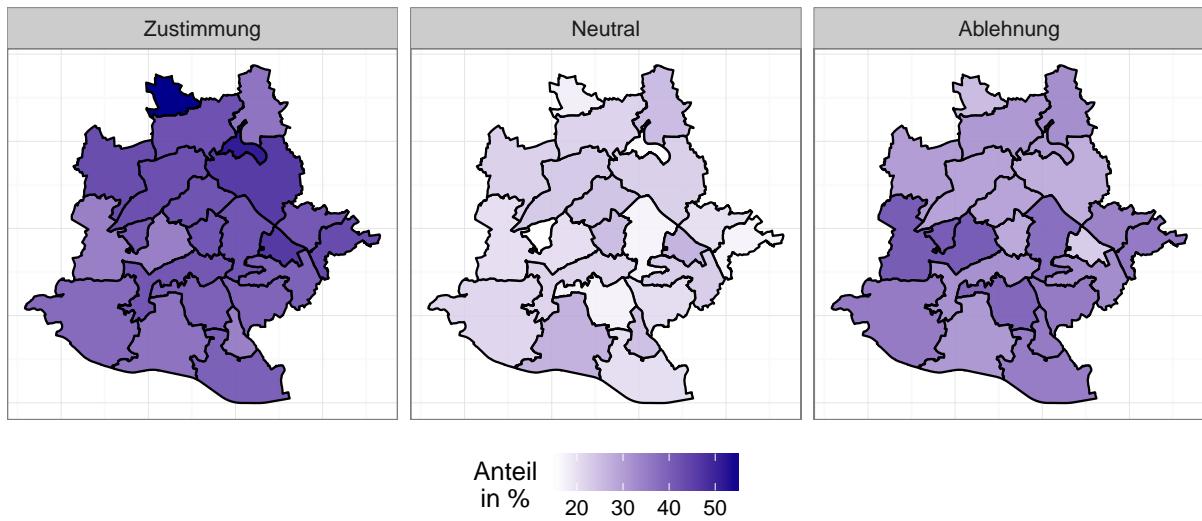


ABBILDUNG 4: ANTEILE IN BEZIRKEN STUTTGART 21

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.

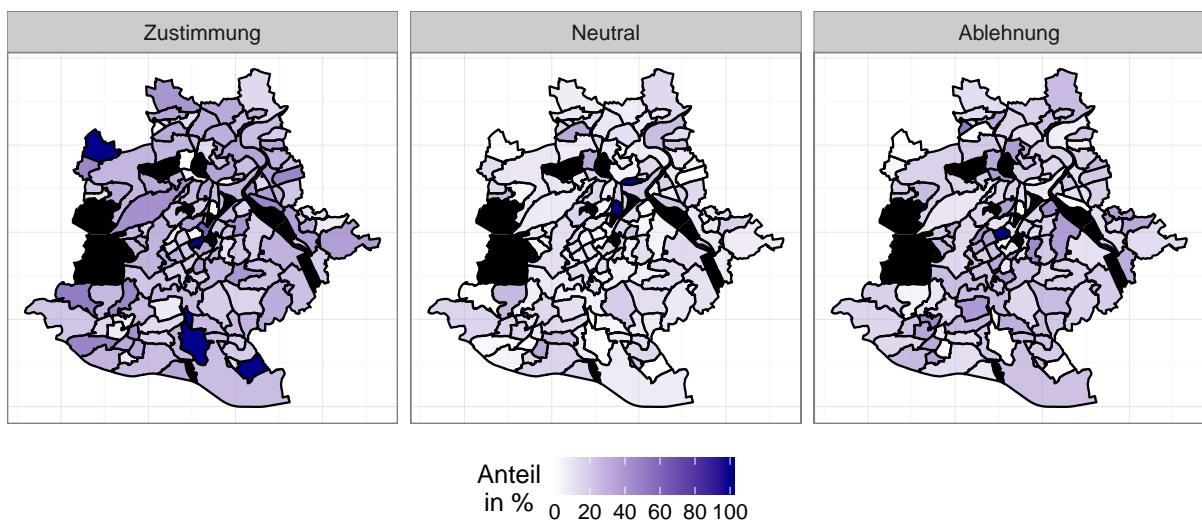


ABBILDUNG 5: ANTEILE IN STADTTEILEN STUTTGART 21

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

## 4 STATISTISCHE METHODEN

### 4.1 MODELL

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

### 4.2 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 Kovariable, 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.

### 4.3 EVALUIERUNG

## 5 ERGEBNISSE

TABELLE 2: VALIDIERUNG

		MSE		Überdeckungswk.	
		Zustimmung	Ablehnung	Zustimmung	Ablehnung
Gauss-Krüger	Bez.	U	0.04	0.749	1
		Z	0.116	0.557	0.043
	Sadtt.	U	0.461	5.708	0.954
		Z	0.813	4.415	0.139
3 Kl. Bezirke	Bez.	U	0.041	0.756	0.553
		Z	0.117	0.562	0.02
	Stadtteile	U	0.482	5.678	0.934
		Z	0.835	4.38	0.139
Gauss-Krüger	Bez.	U	0.023	0.903	0.567
		Z	0.081	0.593	0.027
	Stadtteile	U	0.453	6.69	
		Z	0.61	4.641	
2 Kl. Bezirke	Bez.	U	0.337	0.337	
		Z	0.164	0.164	
	Stadtteile	U	2.87	2.87	
		Z	1.65	1.65	
Bezirke	Bez.	U	0.337	0.337	
		Z	0.164	0.164	
	Stadtteile	U	2.87	2.85	
		Z	1.65	1.64	

## 6 FAZIT

### LITERATUR

[Fahrmeir et al., 2009] Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression. Statistik und ihre Anwendungen.* Springer Berlin Heidelberg.

[Stuttgart, 2011] Stuttgart, S. (2011). <http://www.stuttgart.de/volksabstimmung>.

ANHANG

TABELLE 3: GRUNDGESAMTHEIT BÜRGERUMFRAGE

Anzahl Beobachtungen: 470.190

<b>Variable</b>	<b>Modellierung</b>	<b>Mögliche Ausprägungen</b>
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	

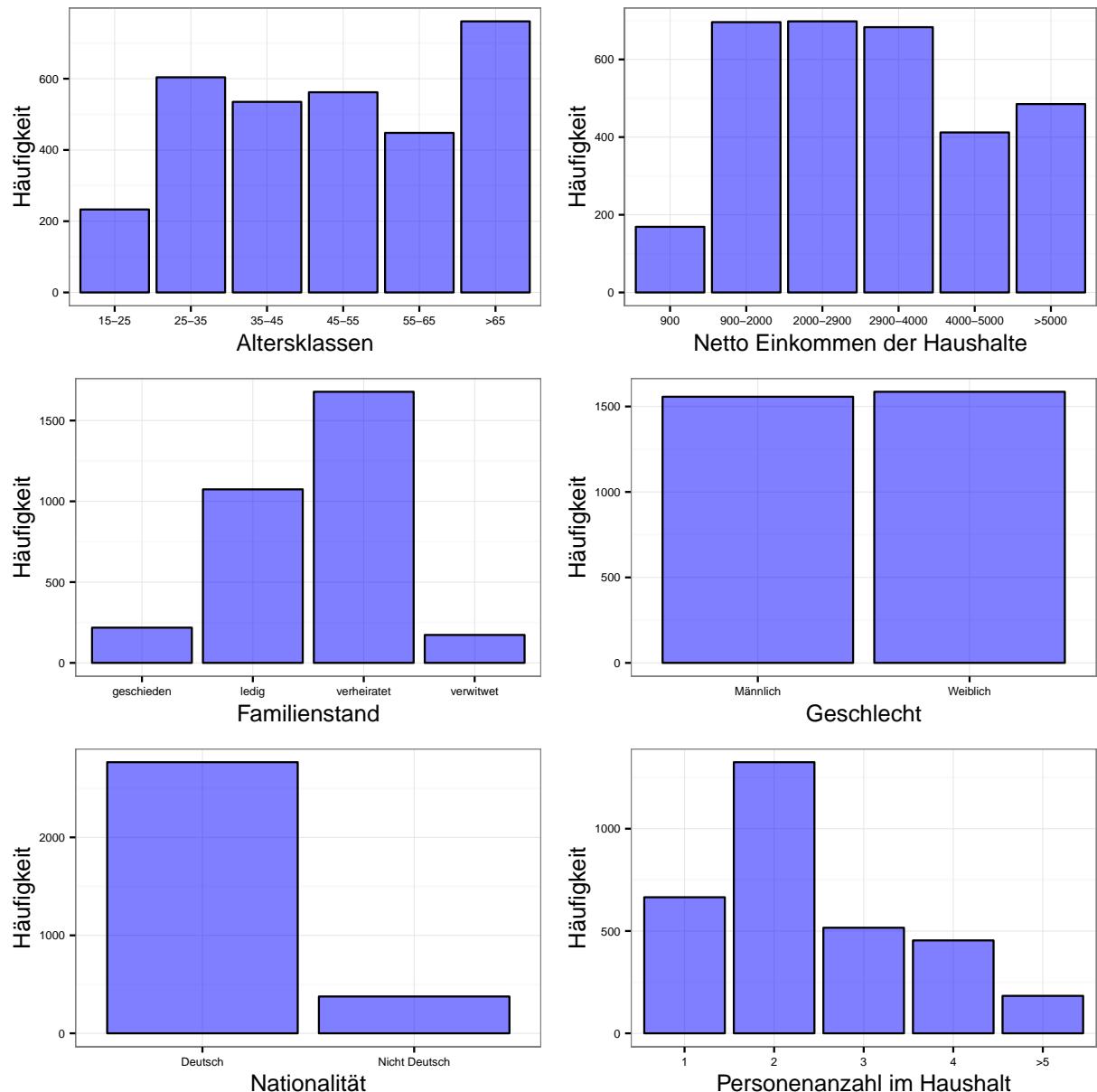
TABELLE 4: GRUNDGESAMTHEIT ZENSUS

Anzahl Beobachtungen: 380.238

<b>Variable</b>	<b>Modellierung</b>	<b>Mögliche Ausprägungen</b>
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

---



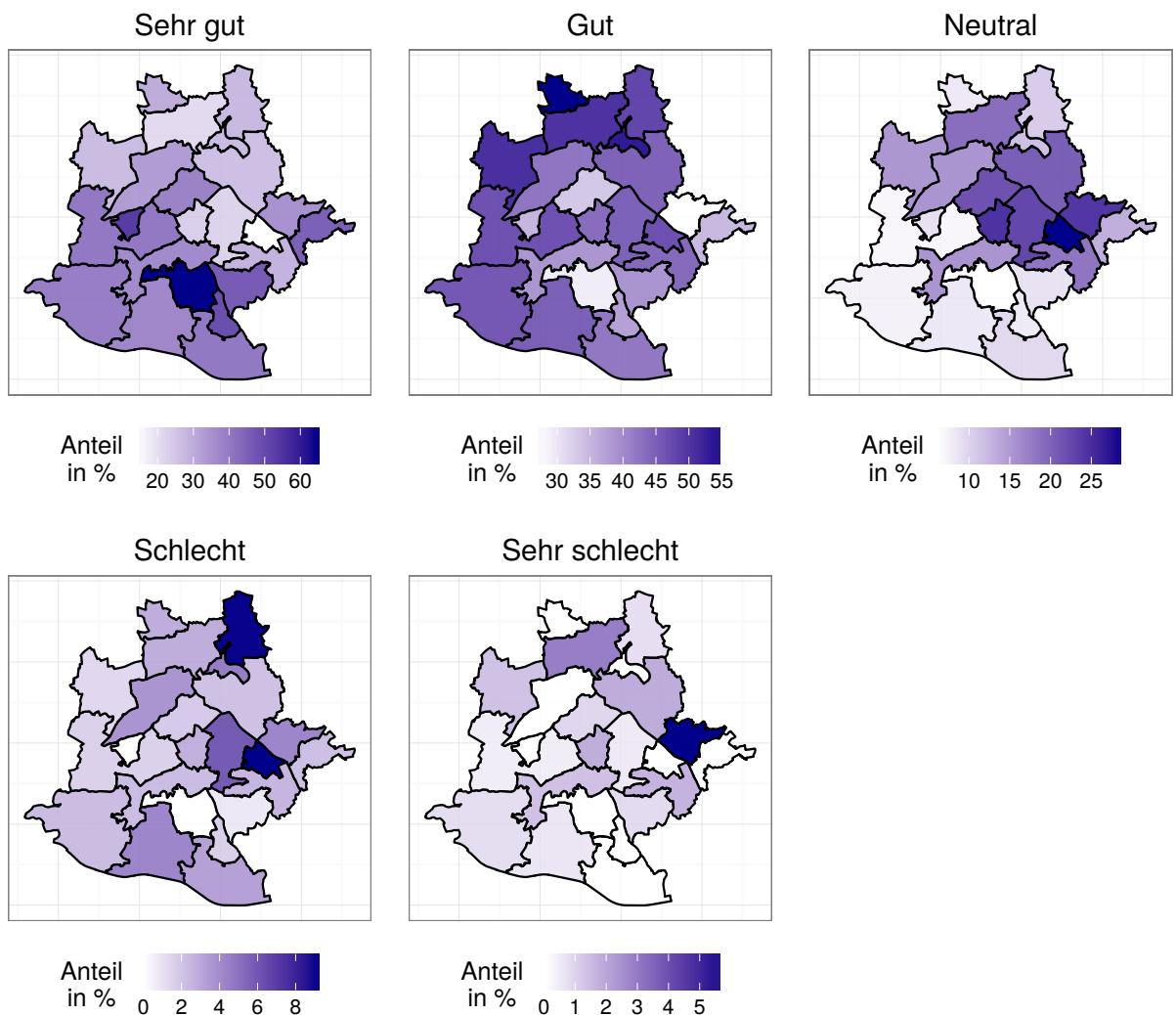


ABBILDUNG 6: ANTEILE IN BEZIRKEN BEWERTUNG WOHNGEgend

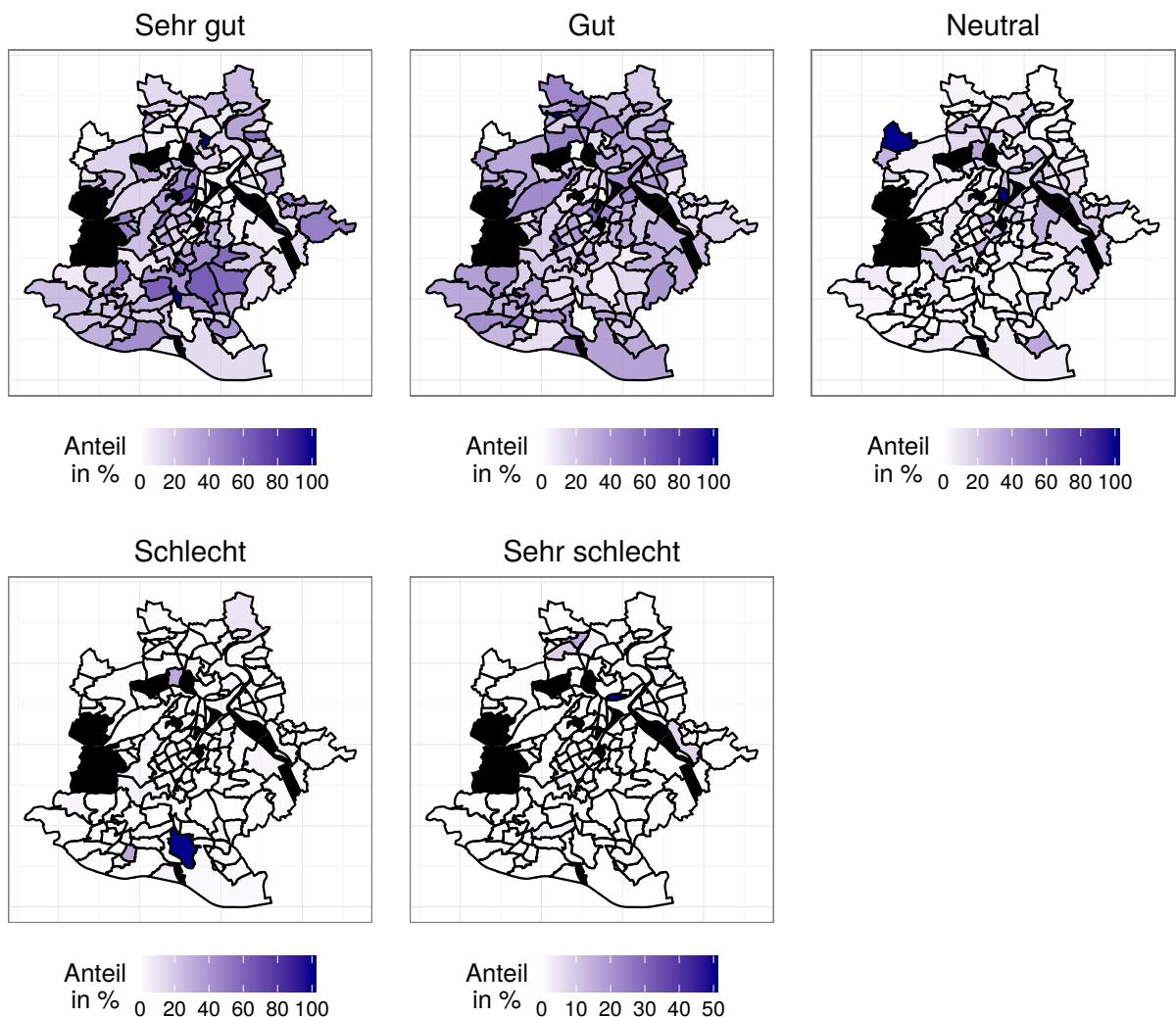


ABBILDUNG 7: ANTEILE IN STADTTEILEN BEWERTUNG WOHNGEgend

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