

Bachelorarbeit

Geodatenbasierte Modellierung der Wohnlagen in München

Institut für Statistik
Ludwig-Maximilians-Universität München

Vanessa Kleisch

München, den 21.08.2025



Zur Erlangung des akademischen Grades Bachelor of Science (B. Sc.)
Betreut durch Prof. Dr. Göran Kauermann

Eidesstattliche Erklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

München, 21.08.2025

.....
(Unterschrift)

Zusammenfassung

Die Wohnlage zählt zu den zentralen Einflussfaktoren bei der Mietpreisbildung. Insbesondere in München spielt sie eine entscheidende Rolle, allerdings sind die Muster hinter der offiziellen Einteilung unklar. Vor diesem Hintergrund untersucht diese Arbeit, ob die offizielle Einteilung der Münchner Wohnlagen durch messbare, geodatenbasierte Standortfaktoren erklärt werden kann, ob sich diese Zusammenhänge zwischen zentralen und nicht-zentralen Gebieten unterscheiden und ob eine zuverlässige Modellierung der Wohnlage möglich ist. Methodisch werden hierfür zwei multinomiale Generalisierte Additive Modelle auf Basis eines hochaufgelösten Geodatensatzes für zentrale und nicht-zentrale Lagen geschätzt. Die Ergebnisse belegen, dass die Zusammenhänge hochgradig nicht-linear sind. Die zentrale Erkenntnis ist die starke Kontextabhängigkeit der Standortfaktoren. Merkmale wie z.B. der Straßentyp, an dem ein Wohnobjekt liegt, die im Zentrum stark positiv wirken, haben in nicht-zentralen Lagen einen negativen oder irrelevanten Einfluss. Die entwickelten Modelle erreichen für alle Wohnlagekategorien eine hohe Vorhersagegüte, insbesondere nach der statistischen Bereinigung, welche die Sensitivität für seltene Wohnlagenkategorien entscheidend verbessert. Zusammenfassend zeigt die Arbeit, dass die Treiber der Wohnlagenqualität nicht universell sind, sondern stark von der Makrolage abhängen, und liefert ein flexibles Framework zur Quantifizierung dieser komplexen räumlichen Zusammenhänge.

Inhaltsverzeichnis

1 Einleitung	1
2 Datengrundlage	2
2.1 Wohnlagen des Mietspiegel Münchens	2
2.2 Daten von infas360	4
2.3 Kombination der Datensätze	4
2.4 Datenvorverarbeitung	6
3 Deskriptive Analyse	9
3.1 Verteilung der Wohnlagen	9
3.2 Verteilung der Einflussgrößen	11
3.2.1 Verteilung der Straßentypen vor und nach der Gruppierung	11
3.2.2 Verteilung der Distanzvariablen	13
3.2.3 Verteilung der Indexvariablen	18
4 Modellierung	22
4.1 Die Multinomialverteilung	22
4.2 Das Generalisierte Additive Modell	22
4.3 Multinomiale GAMs	23
4.4 Modellierung der Wohnlagen	24
5 Effekte der Einflussvariablen	27
5.1 Interpretation des Effekts des Straßentyps	27
5.2 Interpretation der Effekte der metrischen Variablen	30
5.2.1 Interpretation der Effekte der Distanzvariablen	30
5.2.2 Interpretation der Effekte der Indexvariablen	40
5.2.3 Überblick über die nicht-linearen Effekte	46
5.2.4 Bewertung der Konkurvität	48
5.2.5 Alternatives Modell der zentralen Wohnlagen	50
6 Analyse der Fehlklassifikationen	54
6.1 Modellevaluierung	54
6.2 Bereinigung der Vorhersagen	56
6.3 Modellevaluierung mit bereinigten Vorhersagen	57
6.4 Analyse der räumlichen Muster	58
6.4.1 Struktur der unbereinigten Vorhersagen	59
6.4.2 Struktur der bereinigten Vorhersagen	62
7 Fazit und Ausblick	66
A Anhang	V
A.1 Plots	V
A.2 Tabellen	VIII
B Elektronischer Anhang	XII

Abbildungsverzeichnis

1	Wohnlagen der Stadt München	2
2	Räumlicher Anteil der Wohnlagen in München	3
3	Räumlicher Anteil der zentralen Wohnlagen in München	3
4	Räumlicher Anteil der Wohnlagen außerhalb in München	4
5	Verteilung der Wohnlagen im rohen Datensatz	5
6	Wohnlagen der Wohnobjekte im Datensatz	5
7	Verteilung der Wohnlagen im finalen Datensatz	9
8	Verteilung der zentralen Wohnlagen im finalen Datensatz	10
9	Verteilung der Wohnlagen außerhalb im finalen Datensatz	10
10	Verteilung der Straßentypen vor der Gruppierung: zentrale Wohnlagen . . .	11
11	Verteilung der Straßentypen nach der Gruppierung: zentrale Wohnlagen . .	12
12	Verteilung der Straßentypen vor der Gruppierung: Wohnlagen außerhalb .	12
13	Verteilung der Straßentypen nach der Gruppierung: Wohnlagen außerhalb .	13
14	Verteilung der Distanz zum Bahnhof nach zentralen Wohnlagen	13
15	Verteilung der Distanz zum Bahnhof nach nicht-zentralen Wohnlagen . . .	14
16	Verteilung der Distanz zur U-Bahn nach zentralen Wohnlagen	14
17	Verteilung der Distanz zur U-Bahn nach nicht-zentralen Wohnlagen	15
18	Verteilung der Distanz zur Bushaltestelle nach zentralen Wohnlagen	15
19	Verteilung der Distanz zur Bushaltestelle nach nicht-zentralen Wohnlagen .	16
20	Verteilung der Distanz zum Mittelzentrum nach zentralen Wohnlagen . . .	16
21	Verteilung der Distanz zum Mittelzentrum nach nicht-zentralen Wohnlagen	17
22	Verteilung der Distanz zum Unterzentrum nach zentralen Wohnlagen . . .	17
23	Verteilung der Distanz zum Unterzentrum nach nicht-zentralen Wohnlagen	18
24	Verteilung des ÖPNV-Index nach zentralen Wohnlagen	18
25	Verteilung des ÖPNV-Index nach nicht-zentralen Wohnlagen	19
26	Verteilung des Nahversorgungsindex nach zentralen Wohnlagen	19
27	Verteilung des Nahversorgungsindex nach nicht-zentralen Wohnlagen . . .	20
28	Verteilung des Hauspreisindex nach zentralen Wohnlagen	20
29	Verteilung des Hauspreisindex nach nicht-zentralen Wohnlagen	21
30	Odds-Ratios und 95%-Konfidenzintervalle: Effekt des Straßentyps auf die zentralen Lagen	27
31	Odds-Ratios und 95%-Konfidenzintervalle: Effekt des Straßentyps auf die Lagen außerhalb	29
32	Partieller Effekt der Distanz zum Bahnhof auf die zentrale Wohnlage . . .	31
33	Partieller Effekt der Distanz zum Bahnhof auf die Wohnlage außerhalb . .	32
34	Partieller Effekt der Distanz zur U-Bahn auf die zentrale Wohnlage	33
35	Partieller Effekt der Distanz zur U-Bahn auf die Wohnlage außerhalb . . .	34
36	Partieller Effekt der Distanz zur Bushaltestelle auf die zentrale Wohnlage .	35
37	Partieller Effekt der Distanz zur Bushaltestelle auf die Wohnlage außerhalb	36
38	Partieller Effekt der Distanz zum Mittelzentrum auf die zentrale Wohnlage	37
39	Partieller Effekt der Distanz zum Mittelzentrum auf die Wohnlage außerhalb	38
40	Partieller Effekt der Distanz zum Unterzentrum auf die zentrale Wohnlage	39
41	Partieller Effekt der Distanz zum Unterzentrum auf die Wohnlage außerhalb	40

42	Partieller Effekt des ÖPNV-Index auf die zentrale Wohnlage	41
43	Partieller Effekt des ÖPNV-Index auf die Wohnlage außerhalb	42
44	Partieller Effekt des Nahversorgungsindex auf die zentrale Wohnlage	43
45	Partieller Effekt des Nahversorgungsindex auf die Wohnlage außerhalb	44
46	Partieller Effekt des Hauspreisindex auf die zentrale Wohnlage	45
47	Partieller Effekt des Hauspreisindex auf die Wohnlage außerhalb	46
48	Korrelation der metrischen Variablen im Modell der Lagen außerhalb	48
49	Korrelation der metrischen Variablen im Modell der zentralen Lagen	49
50	Vergleich der Odds-Ratios der beiden zentralen Modelle	50
51	Vergleich der partiellen Effekte von <code>distanz_unterzentrum</code> auf die zentrale gute Lage	51
52	Vergleich der partiellen Effekte von <code>distanz_unterzentrum</code> auf die zentrale beste Lage	51
53	Vergleich der partiellen Effekte von <code>nahversorgungs_index</code> auf die zentrale gute Lage	52
54	Vergleich der partiellen Effekte von <code>opnv_index</code> auf die zentrale beste Lage	52
55	Karte der Fehlklassifikationen aus dem Modell der zentralen Wohnlagen . .	59
56	Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Nordwesten	60
57	Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Nordosten	60
58	Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Südosten	61
59	Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Südwesten	61
60	Karte der bereinigten Fehlklassifikationen aus dem Modell der zentralen Wohnlagen	62
61	Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Nordwesten	63
62	Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Nordosten	64
63	Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Südosten	64
64	Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Südwesten	65
65	Ausschnitt aus der interaktiven Karte mit bereinigten Vorhersagen	65
66	Karte aller Lokationen mit zentraler Wohnlage	V
67	Karte aller Lokationen: Nordwesten	VI
68	Karte aller Lokationen: Nordosten	VI
69	Karte aller Lokationen: Südosten	VII
70	Karte aller Lokationen: Südwesten	VII

Tabellenverzeichnis

1	Beschreibung der Ausprägungen von straßentyp	6
2	Beschreibung der Variablen des finalen Datensatzes	7
3	Vergleich der effektiven Freiheitsgrade und Chi-Quadrat-Werte für die Smooth-Terme der beiden GAMs	47
4	Konkurvitätswerte der Smooth-Terme in Bezug auf die Variable distanz-bahnhof im Modell der zentralen Lagen	49
5	Konfusionsmatrix und Gütekennzahlen des GAM der zentralen Lagen. Accuracy: 89,67 %, n = 4220	55
6	Konfusionsmatrix und Gütekennzahlen des GAM der Lagen außerhalb. Accuracy: 78,15 %, n = 89.189	55
7	Konfusionsmatrix und Gütekennzahlen des GAM der zentralen Lagen mit Bereinigung. Accuracy: 87,18 %, n = 4220	57
8	Konfusionsmatrix und Gütekennzahlen des GAM der Lagen außerhalb mit Bereinigung. Accuracy: 73,77 %, n = 89.189	58
9	Flächenmäßiger Anteil der vordefinierten Wohnlagen der Stadt München .	VIII
10	Verteilung der vordefinierten Wohnlagen im finalen Datensatz	VIII
11	Modelloutput des GAM für die zentralen Lagen. Deviance explained = 70.3 %, n = 4220	IX
12	Modelloutput des GAM für die Lagen außerhalb. Deviance explained = 43.6 %, n = 89184	X
13	Modelloutput des angepassten GAM für die zentralen Lagen (ohne distanz-bahnhof). Deviance explained = 66.8 %, n = 4220	XI

1 Einleitung

Die Wohnlage, welche die Qualität eines Wohnobjekts in Abhängigkeit von dessen Umgebung angibt, zählt zu den zentralen Einflussfaktoren bei der Mietpreisbildung in deutschen Großstädten. Insbesondere in München, einer der teuersten Städte Deutschlands, spielt sie eine entscheidende Rolle bei der Bestimmung der ortsüblichen Vergleichsmiete, die im qualifizierten Mietspiegel der Stadt ausgewiesen wird (vgl. Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen n. d.). Mieter können durch diesen beispielsweise feststellen, ob ihre Miete nach einer vorgeschlagenen Erhöhung über dem örtlichen Vergleichswert liegt und diese gegebenenfalls ablehnen (vgl. Landeshauptstadt München 2025a).

Für die Stadt München setzt sich die Wohnlage unter anderem aus den von der Geschäftsstelle des Gutachterausschusses für Grundstückswerte der Landeshauptstadt München bereitgestellten Wohnlagen zusammen (vgl. Landeshauptstadt München 2025b, S. 45). Doch während die Einteilung in Wohnlagen weitreichende finanzielle Konsequenzen hat, bleibt die Frage, auf welchen quantitativen und räumlichen Mustern diese Einteilung beruht. Lassen sich die Grenzen zwischen den verschiedenen Lagen durch messbare Standortfaktoren wie die Nähe zu Infrastruktur oder die Anbindung an den Nahverkehr erklären? Ergeben sich unterschiedliche Zusammenhänge zwischen zentralen Wohnlagen und solchen die am Rande der Stadt liegen?

Die vorliegende Bachelorarbeit widmet sich genau diesen Fragen, indem sie Generalisierte Additive Modelle und hochauflöste Geodaten zu einzelnen Wohnobjekten in ganz München nutzt, um die Struktur der eingeteilten Wohnlagen zu entschlüsseln. Als Datengrundlage dient ein umfangreicher Geodatensatz, der unter anderem präzise Distanzmaße zur Infrastruktur enthält. Der Datensatz und die angewandten Methoden werden dabei detailliert vorgestellt. Ein weiteres Ziel dieser Arbeit ist es zudem, diejenigen Lokationen zu untersuchen, deren Wohnlage nicht mithilfe der entwickelten Modelle korrekt zugeordnet werden kann.

Um diese Forschungsziele systematisch zu verfolgen, ist die vorliegende Arbeit wie folgt aufgebaut. Kapitel 2 legt die Datengrundlage dar und beschreibt die notwendigen Schritte der Datenvorverarbeitung. Daran anknüpfend erfolgt in Kapitel 3 eine deskriptive Analyse des Datensatzes, um erste Eindrücke über die räumlichen Muster zu gewinnen und potenziell problematische Größen zu identifizieren. Kapitel 4 widmet sich der Methodik und stellt die theoretischen Grundlagen sowie die praktische Umsetzung der Generalisierten Additiven Modelle vor. Die detaillierte Interpretation der Modelleffekte für die einzelnen Einflussvariablen ist Gegenstand von Kapitel 5. In Kapitel 6 werden die Modelle auf die Daten angewandt und Vorhersagen über die Wohnlagen der einzelnen Wohnobjekte erstellt, welche anschließend analysiert werden, um potenziell räumliche Muster zu identifizieren. Abschließend werden die zentralen Erkenntnisse in Kapitel 7 in einem Fazit zusammengefasst und ein Ausblick auf die weiterführende Forschung gegeben.

2 Datengrundlage

2.1 Wohnlagen des Mietspiegel Münchens

Die empirische Analyse dieser Arbeit basiert auf der Kombination zweier Datensätze. Einer davon umfasst georeferenzierte Informationen zu den Wohnlagen in München, die im Jahr 2023 erhoben und bereitgestellt wurden. Die Ausweisung der einzelnen Wohnlagen erfolgte auf Basis der Bodenrichtwerte sowie relevanter Merkmale des Wohnumfelds. Es wurde einem Gebiet dabei eine von sechs möglichen Wohnlagen zugeordnet.

Die beste Lage umfasst besonders hochwertige Teilräume des Stadtgebiets, etwa zentrale oder zentrumsnahe Areale mit ausgeprägtem Prestige oder traditionellen Villenvierteln. Die gute Lage ist durch eine Kombination aus ruhiger Wohnatmosphäre, städtebaulicher Qualität, etwa in größeren Neubauarealen, sowie einer soliden Infrastruktur und einem positiven Image gekennzeichnet. Als durchschnittliche Lage gelten jene Bereiche, die weder durch besondere Standortvorteile noch durch offensichtliche Defizite auffallen und sich damit im mittleren Spektrum der Lagequalität einordnen lassen. Ergänzend erfolgt eine zusätzliche Differenzierung nach zentralen und peripheren Lagen, da sich die Makrolage (die Position des Wohnobjekts innerhalb des Stadtgebiets) in früheren Analysen als signifikanter Einflussfaktor für den Mietpreis erwiesen hat (vgl. Landeshauptstadt München 2023, S. 13). In Abbildung 1 ist diese beschriebene Einteilung des Münchner Stadtgebiets dargestellt.

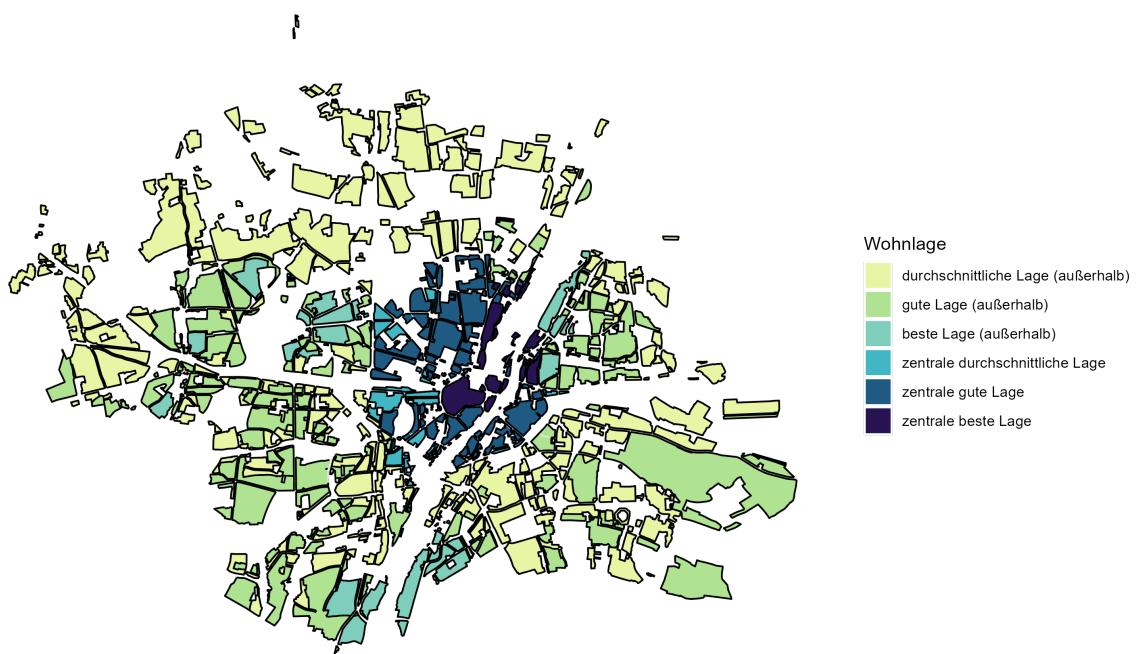


Abbildung 1: Wohnlagen der Stadt München

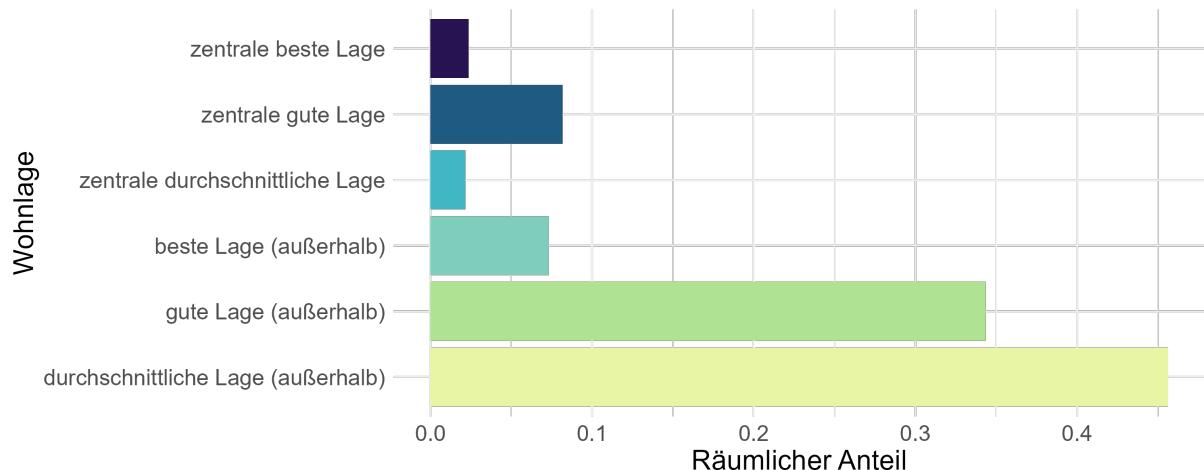


Abbildung 2: Räumlicher Anteil der Wohnlagen in München

Abbildung 2 zeigt zusätzlich den flächenmäßigen Anteil der verschiedenen Wohnlagen im Verhältnis zur eingeteilten Gesamtfläche. Es wird deutlich, dass der überwiegende Teil der Fläche auf nicht-zentrale Wohnlagen entfällt. Insgesamt werden 87,3 % des Stadtgebiets dieser Eigenschaft zugeordnet.

Die folgenden beiden Abbildungen schlüsseln diese Verteilung für die zentralen und nicht-zentralen Lagen detaillierter auf.

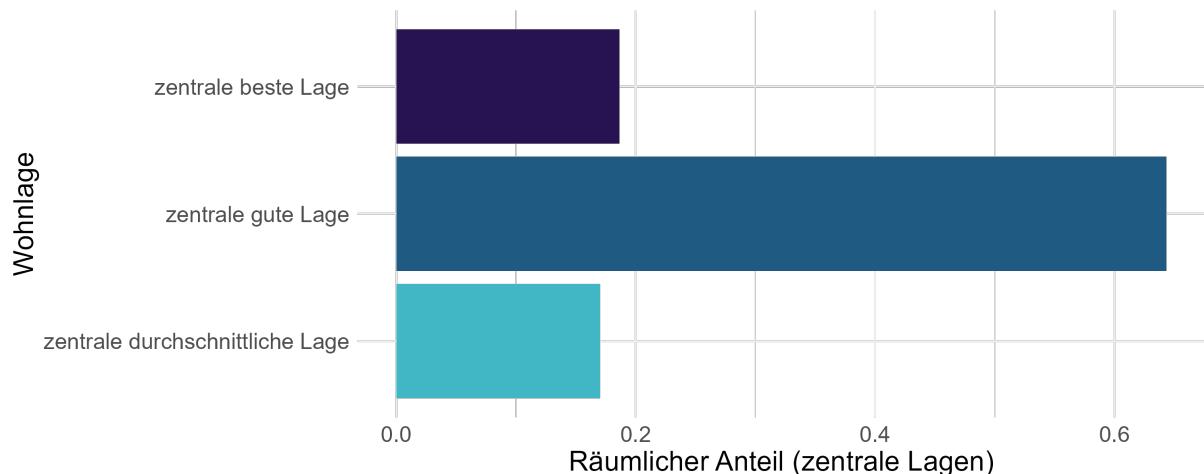


Abbildung 3: Räumlicher Anteil der zentralen Wohnlagen in München

Innerhalb der zentralen Lagen, dargestellt in Abbildung 3, wird der Großteil des Gebiets der Kategorie „zentrale gute Lage“ zugeordnet. Der räumliche Anteil der anderen beiden Kategorien ist in etwa gleich groß.

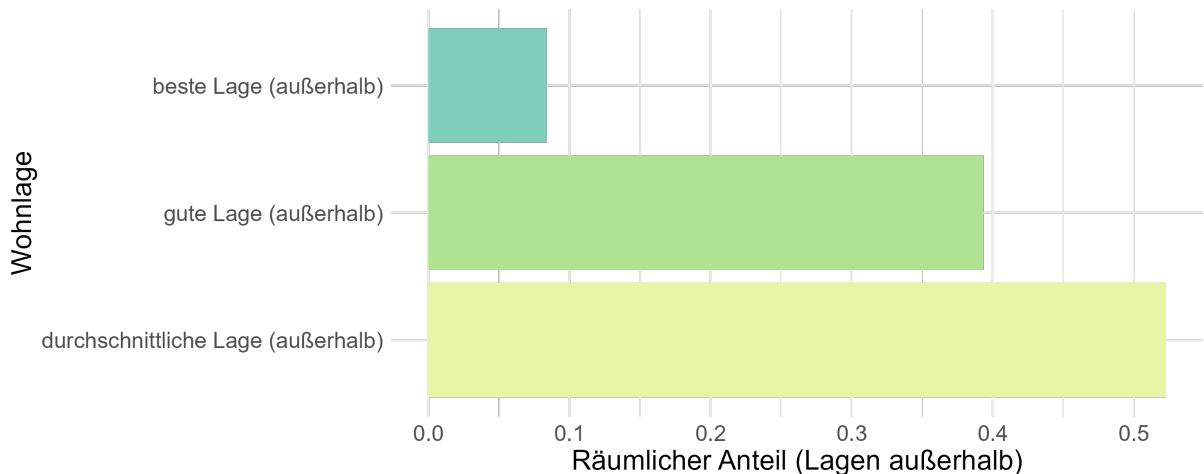


Abbildung 4: Räumlicher Anteil der Wohnlagen außerhalb in München

Abbildung 3 wiederum visualisiert die Anteile innerhalb der nicht-zentralen Gebiete. Hier dominiert flächenmäßig die „durchschnittliche Lage außerhalb“, dicht gefolgt von der „guten Lage außerhalb“. Der räumliche Anteil der „besten Lage außerhalb“ ist mit unter 10 % am geringsten.

2.2 Daten von infas360

Bei dem zweiten Datensatz handelt es sich um den von infas360 bereitgestellten sogenannten CASA-Datensatz. Er kombiniert geo-basierte Gebäudedaten, offizielle Statistiken und viele weitere Informationen aus anderen öffentlichen und privaten Quellen. Der Datensatz enthält die exakten Koordinaten (Breiten- und Längengrad) einzelner Wohnobjekte sowie zahlreiche weitere Variablen mit objektbezogenen Informationen. In dieser Analyse wird nur eine Teilmenge dieser betrachtet. Insgesamt befinden sich Angaben zu 162698 Wohnobjekten im CASA-Datensatz.

2.3 Kombination der Datensätze

Für die empirische Analyse wurden beide Datensätze über die gemeinsamen räumlichen Referenzen verknüpft. Die im CASA-Datensatz enthaltenen Koordinaten ermöglichen eine präzise Zuordnung jeder Beobachtung zu einer der definierten Wohnlagen. Es kann allerdings 25.237 Wohnobjekten keine Wohnlage zugeordnet werden, da diese sich außerhalb der von der Stadt München klassifizierten Gebiete befinden. Durch die Kombination beider Datenquellen entsteht ein strukturierter Datensatz, auf dessen Basis nun der Einfluss ausgewählter Kovariablen auf die Wohnlage untersucht werden kann.

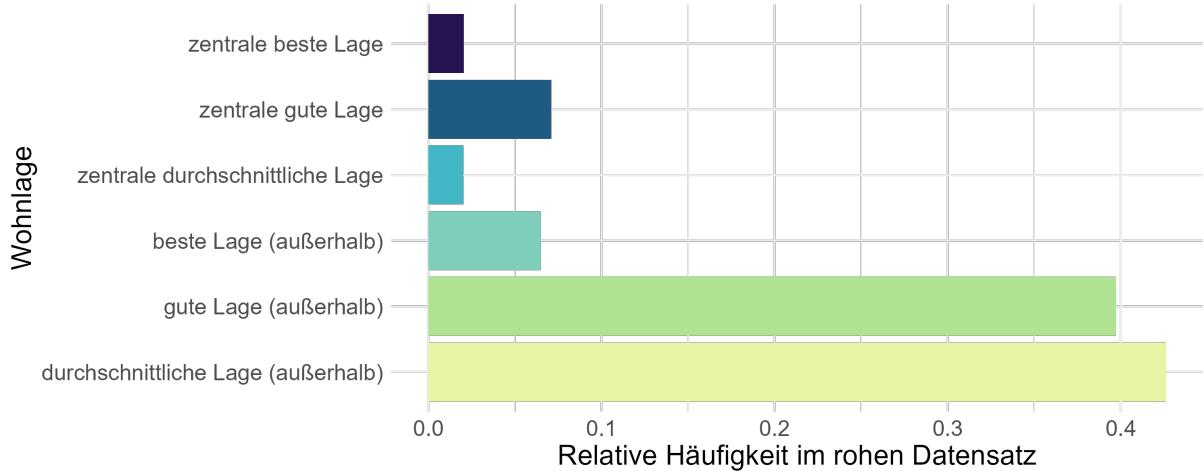


Abbildung 5: Verteilung der Wohnlagen im rohen Datensatz

Abbildung 5 zeigt zunächst die Verteilung der im zusammengeführten Datensatz zugewiesenen Wohnlagen. Dabei ist ersichtlich, dass diese nicht exakt den Flächenanteilen der einzelnen Wohnlagengebiete entspricht, jedoch grundsätzlich jede Wohnlage angemessen repräsentiert wird.

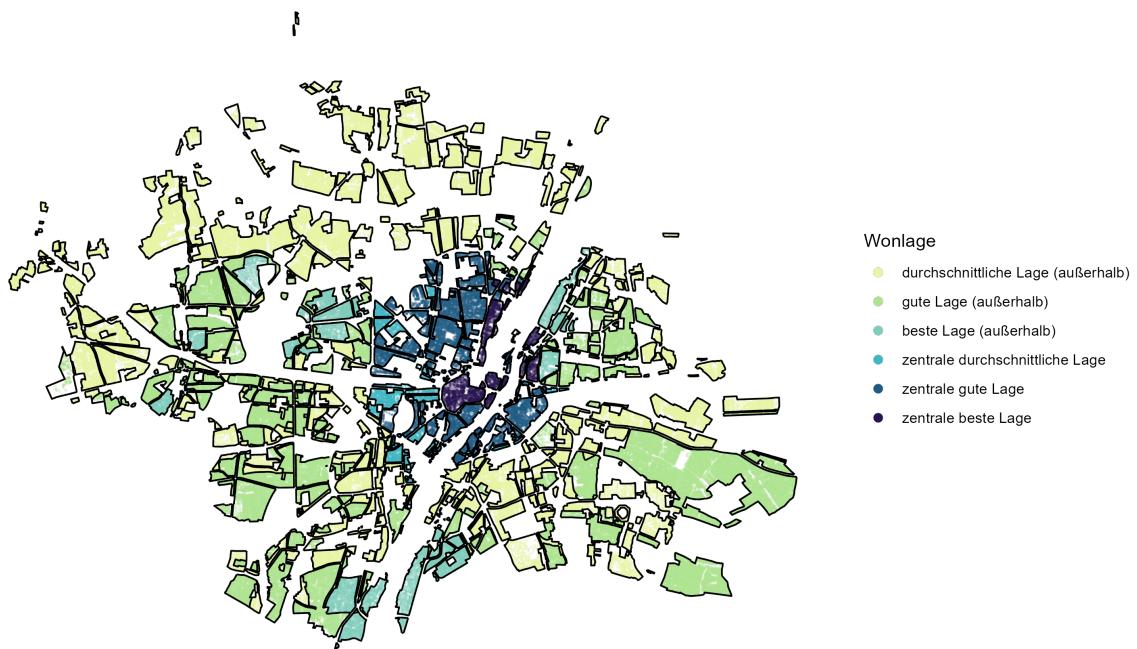


Abbildung 6: Wohnlagen der Wohnobjekte im Datensatz

In Abbildung 6 sind zusätzlich alle Wohnobjekte mit zuordenbarer Wohnlage als Punkte dargestellt. Es ist ersichtlich, dass der Datensatz fast alle klassifizierten Gebiete in München gut abdeckt, da kaum weiße Flächen innerhalb der Grenzen auftreten.

2.4 Datenvorverarbeitung

Eine sorgfältige Datenvorverarbeitung ist eine zentrale Voraussetzung für leistungsfähige statistische Modelle. Im Folgenden werden daher alle Schritte der Datenaufbereitung dokumentiert, die notwendig sind, um eine zuverlässige Grundlage für die nachfolgenden Modellierungen zu schaffen. Alle nun beschriebenen Änderungen wurden an dem kombinierten Datensatz durchgeführt.

Zunächst wurden die 25.237 Lokationen ohne zuordenbare Wohnlage sowie vier außerhalb Münchens liegende Objekte ausgeschlossen. Zusätzlich werden für diese Analyse alle Wohnobjekte ausgeschlossen, die nicht rein privat genutzt werden. Dadurch fallen weitere 37.877 Beobachtungen weg.

Im nächsten Schritt wurden die Variablen entfernt, die entweder eine unverhältnismäßig hohe Menge an fehlenden Werten besitzen oder für die Analyse der Wohnlage nicht relevant sind. Dazu gehören genaue Adressdaten (z.B. Hausnummer), Variablen, welche die Eigenschaften der Gebäude beschreiben (z.B. Aussehen, Alter, Höhe, etc.) und sozioökonomische Größen (z.B. Arbeitslosenquote der unmittelbaren Gegend, Altersstruktur, etc.). Schließlich werden alle Beobachtungen entfernt, die bei den restlichen ausgewählten Größen mindestens einen fehlenden Wert besitzen. Dieser Schritt stellt sicher, dass der gleiche Datensatz für die deskriptive Analyse und die Modellierung verwendet wird.

Unter den übrigen Variablen befindet sich **straßentyp**, welche den Straßentyp beschreibt, an dem die Lokation liegt. Tabelle 1 zeigt die absoluten Häufigkeiten im Datensatz.

Ausprägung	Absolute Häufigkeit	Relative Häufigkeit
Bundesstraße	453	0,355 %
Kreisstraße	1148	0,898 %
Hauptverkehrsstraße	12014	9,403 %
Sammelstraße	27049	21,168 %
Anliegerstraße/ Wohnstraße	77314	60,506 %
Verkehrsberuhigter Bereich	5785	4,527 %
Fußgängerzone	204	0,160 %
Fußweg	3813	2,984 %

Tabelle 1: Beschreibung der Ausprägungen von **straßentyp**

Die Verteilung der einzelnen Ausprägungen weist ein deutliches Ungleichgewicht auf. Besonders selten vorkommende Kategorien wie „Bundesstraße“ oder „Fußgängerzone“ stellen hierbei eine Herausforderung dar. Sie können zu Konvergenzproblemen in Klassifikationsmodellen führen oder eine Überschätzung der Effekte der entsprechenden Straßentypen verursachen. Um diesen Problemen entgegenzuwirken und die Stabilität der Modellierung zu garantieren, wurde eine sinnvolle Gruppierung der ursprünglichen Kategorien vorgenommen. Die Straßentypen wurden dabei in vier übergeordnete Kategorien zusammengefasst:

- **Hauptstraße** (umfasst „Bundesstraße“, „Kreisstraße“ und „Hauptverkehrsstraße“),

- **Sammelstraße**,
- **Wohnstraße** (bestehend aus „Anliegerstraße / Wohnstraße“ und „verkehrsberuhigter Bereich“),
- **Fußgängerbereich** (inklusive „Fußgängerzone“ und „Fußweg“).

Durch diese Aggregation ist eine robuste statistische Auswertung möglich, ohne die inhaltliche Differenzierung vollständig aufzugeben.

Durch die beschriebenen Verarbeitungsschritte wurde der finale Datensatz erstellt. Dieser umfasst insgesamt 93.409 Beobachtungen und 10 Variablen. Eine detaillierte Beschreibung der enthaltenen Variablen ist in Tabelle 2 dargestellt.

Name	Beschreibung	Ursprung
Wohnlage	Offizielle Wohnlage der Adresse, sechs Ausprägungen	Manuelle Einteilung der Stadt München
straßentyp	Straßentyp, an dem die Adresse liegt, 4 Kategorien	infas 360 GmbH, auf Basis von TomTom & OSM
distanz_bahnhof	Entfernung zum nächsten Bahnhof (in Metern)	infas 360 GmbH
distanz_ubahn	Entfernung zur nächsten U-Bahn (in Metern)	infas 360 GmbH
distanz_bushaltestelle	Entfernung zur nächsten Bushaltestelle (in Metern)	infas 360 GmbH
distanz_mittelzentrum	Entfernung zum nächsten Mittelzentrum (in Metern)	infas 360 GmbH
distanz_unterzentrum	Entfernung zum nächsten Unterzentrum (in Metern)	infas 360 GmbH
opnv_index	Index zur ÖPNV-Verfügbarkeit (0-100, sehr schlecht bis sehr gut),	infas 360 GmbH, Berechnung aus Distanzvariablen
naheversorgungs_index	Index zur Nahversorgung	infas 360 GmbH
hauspreis_index	Hauspreisindex des Kreises (100 = Kreisdurchschnitt)	infas 360 GmbH, ImmoScout 24, Berechnung durch Small Area Statistic-Verfahren

Tabelle 2: Beschreibung der Variablen des finalen Datensatzes

Aus dem finalen Datensatz werden zwei kleinere Datensätze gebildet, die jeweils ausschließlich zentrale bzw. nicht-zentrale Wohnlagen enthalten. Ersterer umfasst 4.220 Beobachtungen, letzterer 89.189. Diese Aufteilung ist erforderlich, da die Modellierung getrennt nach zentralen und nicht-zentralen Lagen erfolgt, um Unterschiede und Gemeinsamkeiten zwischen beiden Gruppen präziser herauszuarbeiten. Frühere Analysen haben bereits einen

signifikanten Einfluss dieser Differenzierung belegt (vgl. Landeshauptstadt München 2025b, S. 48).

Zudem wird die räumliche Verortung der einzelnen Beobachtungen in der Spalte `geometry` gespeichert. Es handelt sich hierbei um eine sogenannte Simple Feature Column, eine spezielle Listenstruktur aus dem `sf`-Paket. Jedes Element dieser Liste repräsentiert die Geometrie einer einzelnen Beobachtung, in diesem Fall eine Punktgeometrie. Jeder Punkt ist durch ein Tupel numerischer X- und Y-Koordinaten definiert, welches die exakte Position im Raum beschreibt. Die Interpretation dieser Koordinaten ist durch das zugrundeliegende Koordinatenreferenzsystem (KRS) festgelegt, welches hier ETRS89 / UTM Zone 32N ist. Dieses System gewährleistet, dass alle räumlichen Berechnungen und Darstellungen auf einer konsistenten und geodätisch korrekten Basis erfolgen (vgl. Pebesma 2018).

3 Deskriptive Analyse

Bevor die eigentliche Modellierung erfolgt, wird zunächst eine deskriptive Analyse der verwendeten Daten durchgeführt. Ziel dieses Kapitels ist es, einen Überblick über die Verteilung zentraler Variablen zu geben und potenzielle Auffälligkeiten zu identifizieren. Die gewonnenen Erkenntnisse dienen als Grundlage für die spätere Modellbildung und Interpretation der Ergebnisse.

3.1 Verteilung der Wohnlagen

Um eine aussagekräftige und robuste Modellierung der Wohnlagen in München zu gewährleisten, ist zunächst zu prüfen, ob alle sechs Kategorien ausreichend im finalen Datensatz vertreten sind.

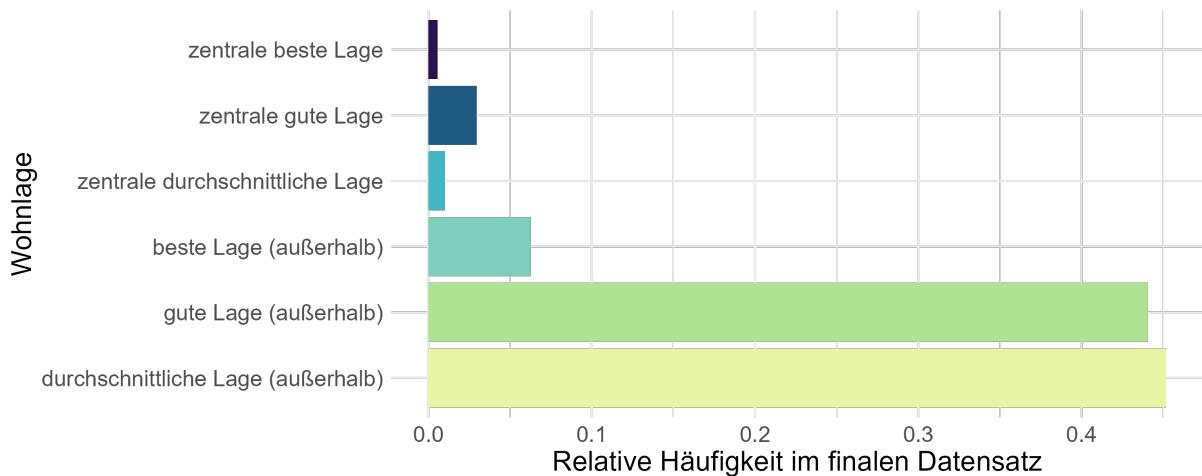


Abbildung 7: Verteilung der Wohnlagen im finalen Datensatz

Abbildung 7 zeigt die relativen Häufigkeiten der verschiedenen Wohnlagentypen. Da zentrale und nicht-zentrale Lagen in der Analyse separat modelliert werden, ist es sinnvoll, zusätzlich die Verteilungen für beide Gruppen getrennt zu betrachten. Dies ist insbesondere relevant, da der Flächenanteil der zentralen Lagen in München insgesamt deutlich geringer ausfällt als der der nicht-zentralen Lagen und die Verhältnisse somit besser eingeschätzt werden können.

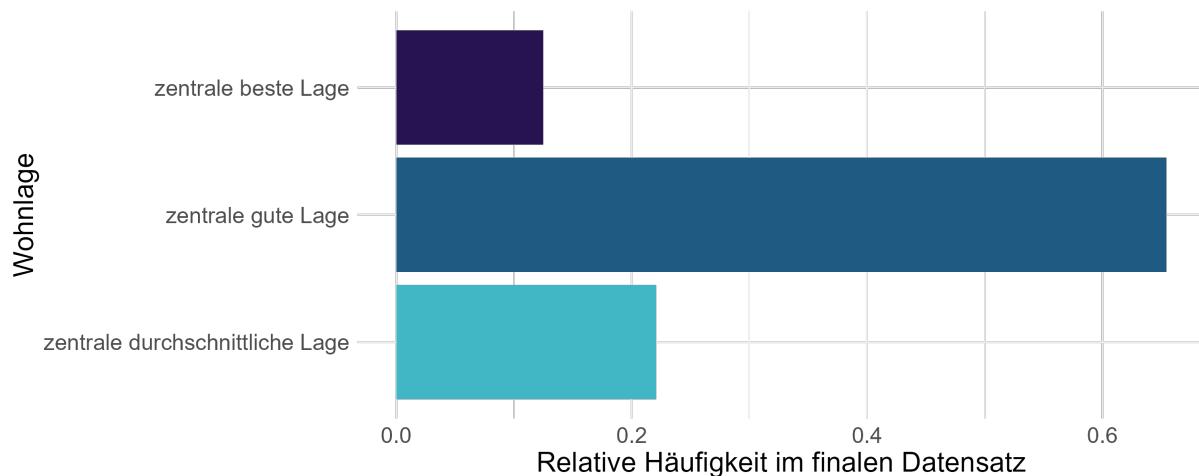


Abbildung 8: Verteilung der zentralen Wohnlagen im finalen Datensatz

Abbildung 8 zeigt die relative Häufigkeitsverteilung der zentralen Wohnlagen im finalen Datensatz. Mit 65,4 % stellen Objekte in guter Lage die größte Gruppe dar, gefolgt von durchschnittlichen Lagen mit 22,1 %. Die besten Lagen treten mit 12,5 % am seltensten auf. Der Abstand zwischen der häufigsten und der seltensten Kategorie beträgt 52,9 Prozentpunkte, was auf eine deutliche asymmetrische Verteilung hinweist.

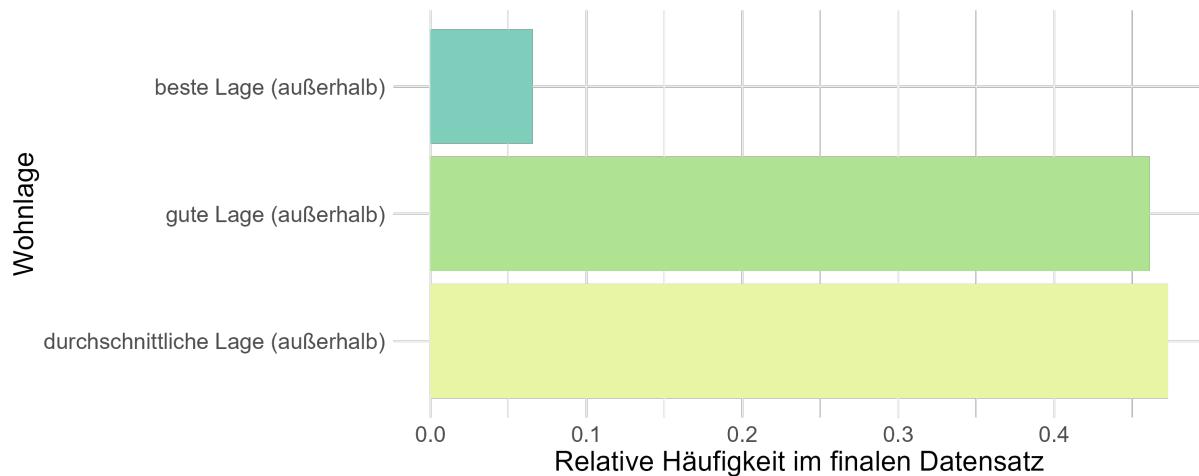


Abbildung 9: Verteilung der Wohnlagen außerhalb im finalen Datensatz

Auch bei den nicht-zentralen Lagen zeigt sich eine ungleichmäßige Verteilung. Durchschnittliche Lagen stellen mit 47,3 % die größte Gruppe dar, gefolgt von guten Lagen mit 46,1 %. Die besten Lagen kommen deutlich seltener vor und erreichen lediglich einen Anteil von 6,6 %.

3.2 Verteilung der Einflussgrößen

3.2.1 Verteilung der Straßentypen vor und nach der Gruppierung

Im folgenden Abschnitt werden die Verteilungen der Variable `straßentyp` vor und nach der vorgenommenen Gruppierung visualisiert und miteinander verglichen. Dabei erfolgt eine getrennte Betrachtung zentraler und nicht zentraler Lagen, da auf Basis dieser Unterscheidung separate Modelle entwickelt werden.

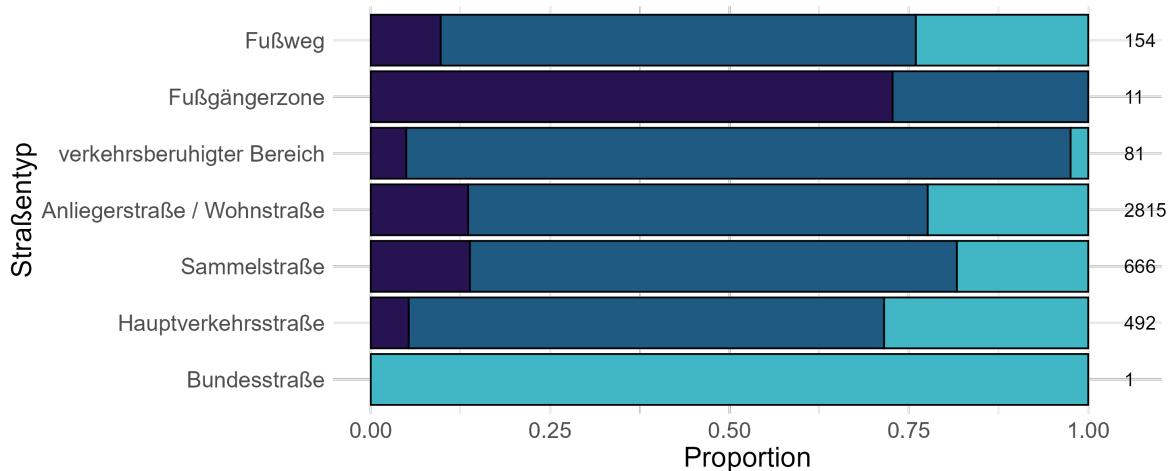


Abbildung 10: Verteilung der Straßentypen vor der Gruppierung: zentrale Wohnlagen

Abbildung 10 zeigt zunächst die Verteilung der ursprünglichen Ausprägungen der Variable `straßentyp` für zentral gelegene Lagen. Die horizontale Balkengrafik stellt die relativen Häufigkeiten (Proportionen) der einzelnen Straßentypen dar, wobei zusätzlich die absoluten Fallzahlen am rechten Rand jedes Balkens angegeben sind. Auch hier ist die Problematik der Kategorien „Bundesstraße“, „Fußgängerzone“ und „verkehrsberuhigter Bereich“ gut erkennbar. Sie kommen sehr selten vor und treten in Kombination mit bestimmten Wohnlagen teilweise gar nicht auf. Die Kategorie „Kreisstraße“ lässt sich generell keiner zentral gelegenen Lokation zuordnen, dementsprechend kann in einem Modell dafür kein Effekt geschätzt werden.

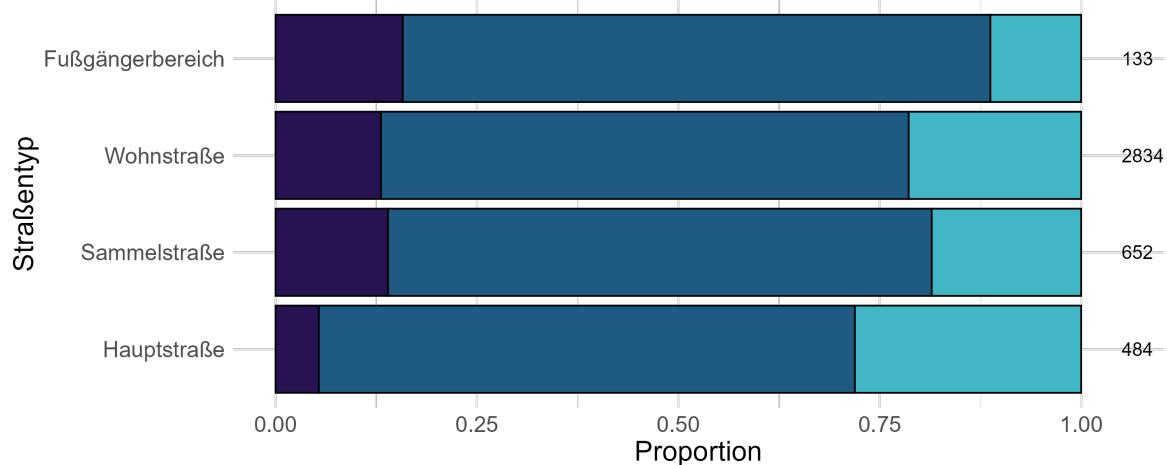


Abbildung 11: Verteilung der Straßentypen nach der Gruppierung: zentrale Wohnlagen

In Abbildung 11 ist die Verteilung der gruppierten Straßentypen dargestellt. Auffällige Ungleichgewichte, wie sie in der ursprünglichen Kategorisierung sichtbar waren, treten nun nicht mehr auf. Zwar geht eine solche Gruppierung stets mit einem gewissen Informationsverlust einher, jedoch ist sie in diesem Fall notwendig, um überhaupt eine robuste und verlässliche Modellschätzung zu ermöglichen (vgl. Zumel und Mount 2019). Die Aggregation stellt somit einen methodisch sinnvollen Kompromiss zwischen inhaltlicher Differenzierung und statistischer Umsetzbarkeit dar.

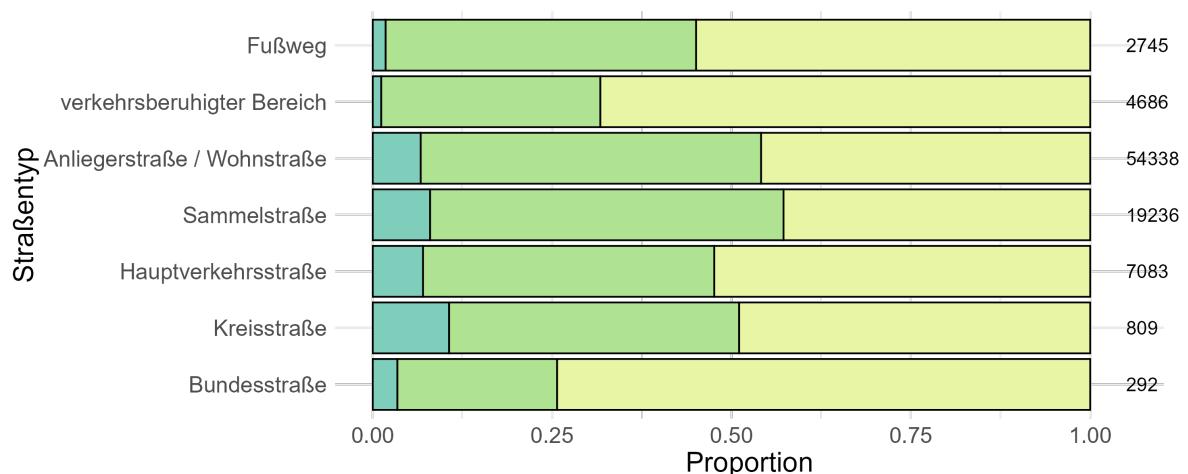


Abbildung 12: Verteilung der Straßentypen vor der Gruppierung: Wohnlagen außerhalb

Ähnliche Ungleichgewichte zeigen sich für die nicht-zentralen Lagen in Abbildung 12. Besonders problematisch sind hier die Kategorien „Fußgängerzone“, „Bundesstraße“ und „Kreisstraße“, die gar nicht oder nur in sehr geringer Fallzahl vertreten sind. Zwar weist die Kategorie „verkehrsberuhigter Bereich“, anders als bei den zentralen Lagen, in diesem Teil der Stichprobe eine ausreichende Anzahl an Beobachtungen auf, um potenziell als eigenständige Gruppe bestehen zu bleiben, allerdings würde dies die Vergleichbarkeit der

Effekte zwischen zentralen und nicht-zentralen Lagen einschränken, da die Gruppenzusammensetzung in den Modellen dann uneinheitlich wäre. Somit wurde auch hier die gleiche Gruppierungsstrategie angewandt. Die folgende Abbildung zeigt das Ergebnis dieser Transformation.

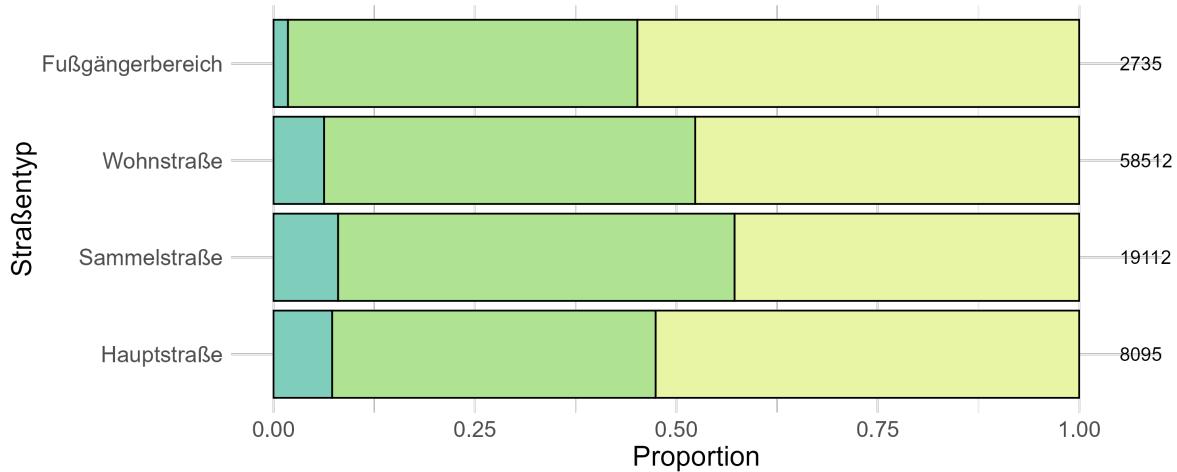


Abbildung 13: Verteilung der Straßentypen nach der Gruppierung: Wohnlagen außerhalb

3.2.2 Verteilung der Distanzvariablen

Im Folgenden wird die Verteilung verschiedener Distanz-Variablen aus dem finalen Datensatz analysiert. Die Variablen geben an, wie viele Meter die Wohnobjekte von bestimmten Orten entfernt liegen. Zur Veranschaulichung werden Dichte-Plots verwendet, bei denen die X-Achse die Distanz in Metern und die Y-Achse die geschätzte Wahrscheinlichkeitsdichtefunktion darstellt. Die Werte auf der Y-Achse basieren auf einer Kernel-Dichteschätzung und geben an, wie wahrscheinlich bestimmte Distanzen innerhalb einer betrachteten Kategorie auftreten (vgl. Wickham et al. 2025, S. 98). Hohe Dichtewerte weisen dabei auf eine größere Häufung von Beobachtungen in diesem Distanzbereich hin.

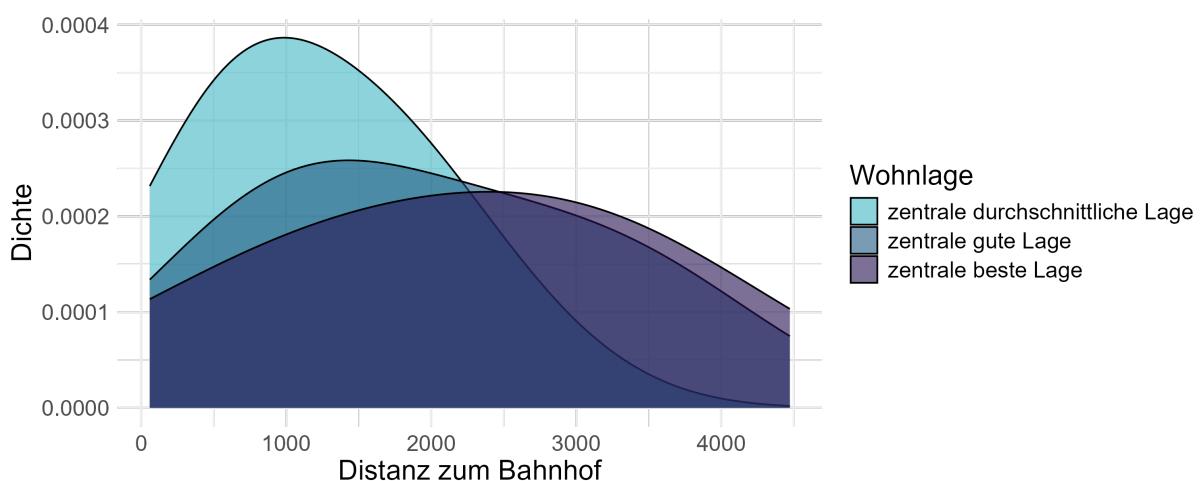


Abbildung 14: Verteilung der Distanz zum Bahnhof nach zentralen Wohnlagen

Abbildung 14 zeigt die Verteilung der Distanz zum Bahnhof für die drei Kategorien zentraler Wohnlagen. Bei der zentralen durchschnittlichen Lage zeigt sich dabei ein klarer Peak um etwa 1000 Meter, was darauf hinweist, dass viele dieser Wohnobjekte etwa so weit vom nächsten Bahnhof entfernt sind. Die Lokationen mit zentraler guter und bester Lage sind gleichmäßiger über mittlere Distanzen verteilt. Insgesamt zeigt sich, dass mit steigender Qualität der zentralen Lage die Wohnobjekte im Mittel etwas weiter vom nächsten Bahnhof entfernt liegen.

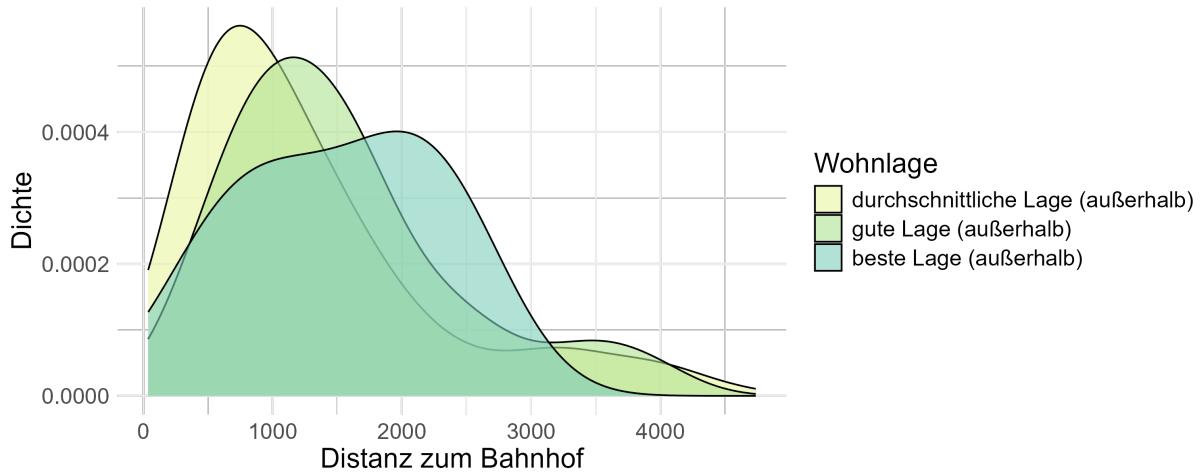


Abbildung 15: Verteilung der Distanz zum Bahnhof nach nicht-zentralen Wohnlagen

Ein ähnlicher Trend ist auch in Abbildung 15 für die nicht-zentralen zu beobachten. Es ist allerdings festzustellen, dass sehr wenige Lokationen mit sehr großen Distanzen zum nächsten Bahnhof unter den besten Lagen zu finden sind.

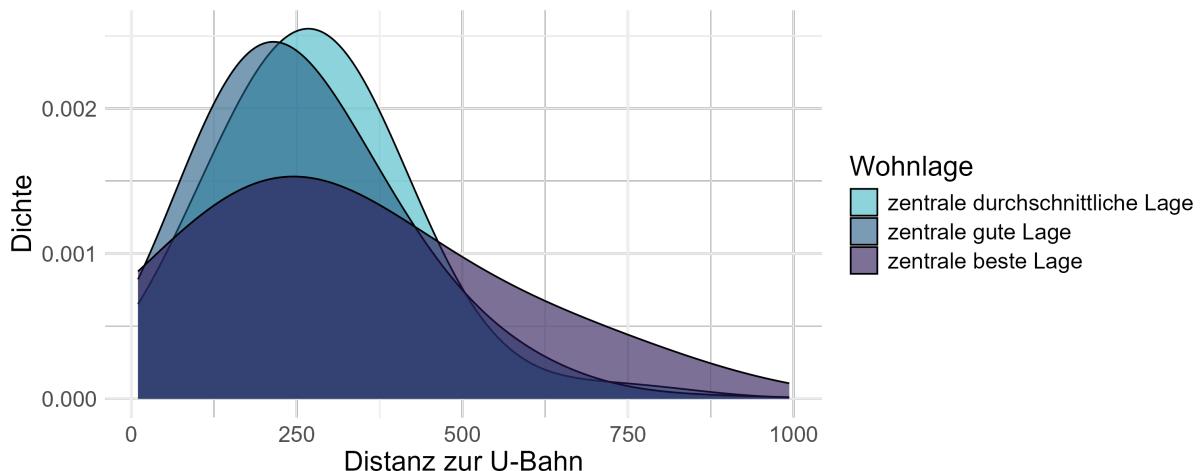


Abbildung 16: Verteilung der Distanz zur U-Bahn nach zentralen Wohnlagen

Zusätzlich wird nun die Verteilung der Distanzen zur nächstgelegenen U-Bahn-Haltestelle untersucht. Wie anhand der X-Achse in Abbildung 16 zu erkennen ist, liegt die größte

gemessene Distanz hier bei knapp unter einem Kilometer. Im Gegensatz dazu beträgt die maximale Distanz zum nächsten Bahnhof 4470 Meter. Dieser Unterschied erklärt sich unter anderem dadurch, dass München über rund 100 U-Bahn-Haltestellen verfügt, welche tendenziell eher zentral konzentriert sind (vgl. Landeshauptstadt München 2025).

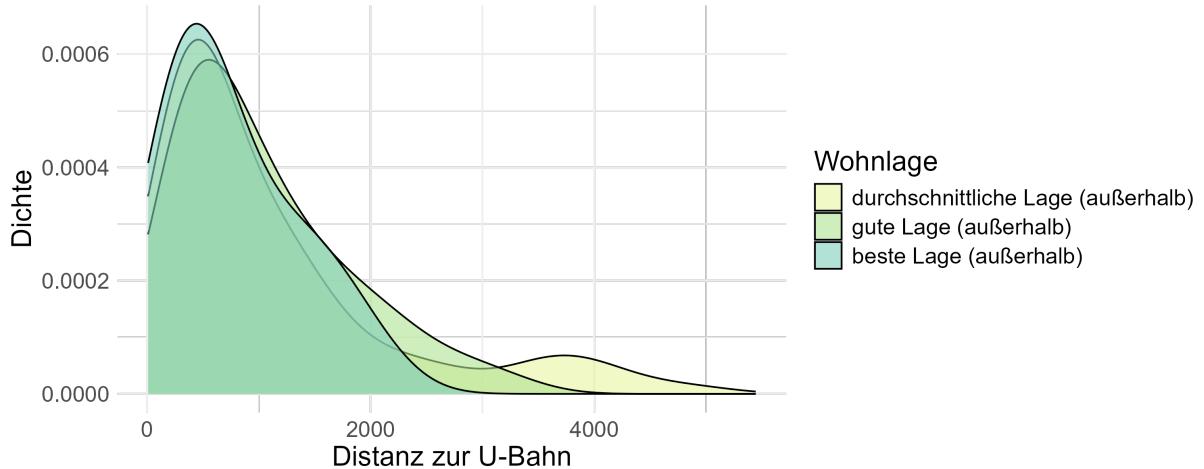


Abbildung 17: Verteilung der Distanz zur U-Bahn nach nicht-zentralen Wohnlagen

Dies zeigt sich auch bei Betrachtung der Distanzverteilung zur nächstgelegenen U-Bahn-Haltestelle für Standorte außerhalb des Zentrums. Abbildung 17 verdeutlicht, dass viele dieser Standorte, insbesondere solche in durchschnittlichen Lagen, vergleichsweise weit von einer U-Bahn-Haltestelle entfernt sind. Zwischen den verschiedenen Wohnlagen treten generell jedoch nur geringe Unterschiede auf.

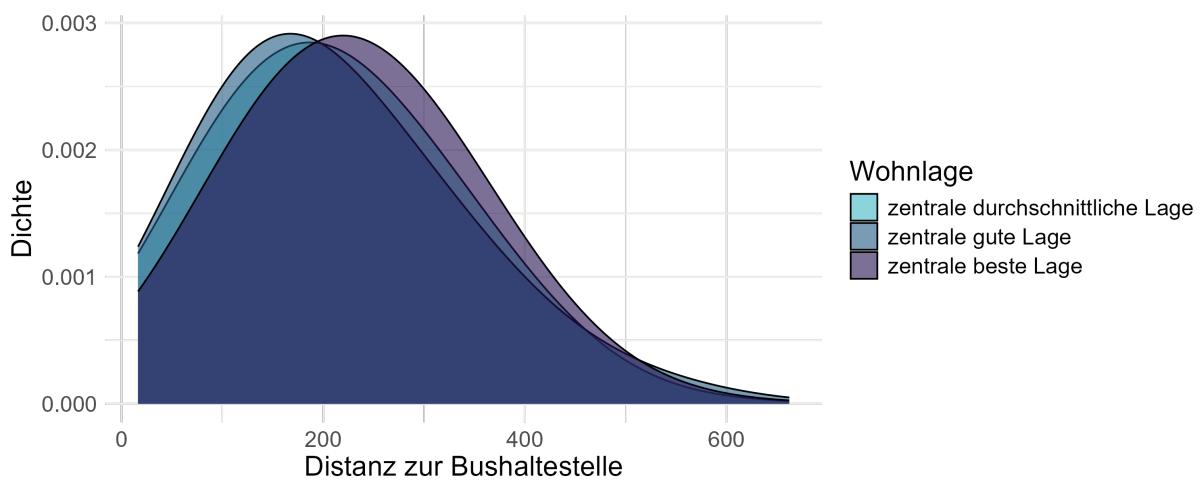


Abbildung 18: Verteilung der Distanz zur Bushaltestelle nach zentralen Wohnlagen

Ebenfalls nur geringfügige Unterschiede zwischen den drei Wohnlagenkategorien zeigen die Verteilungsdichten für die Distanz zur nächsten Bushaltestelle bei den zentralen Lagen (Abbildung 18).

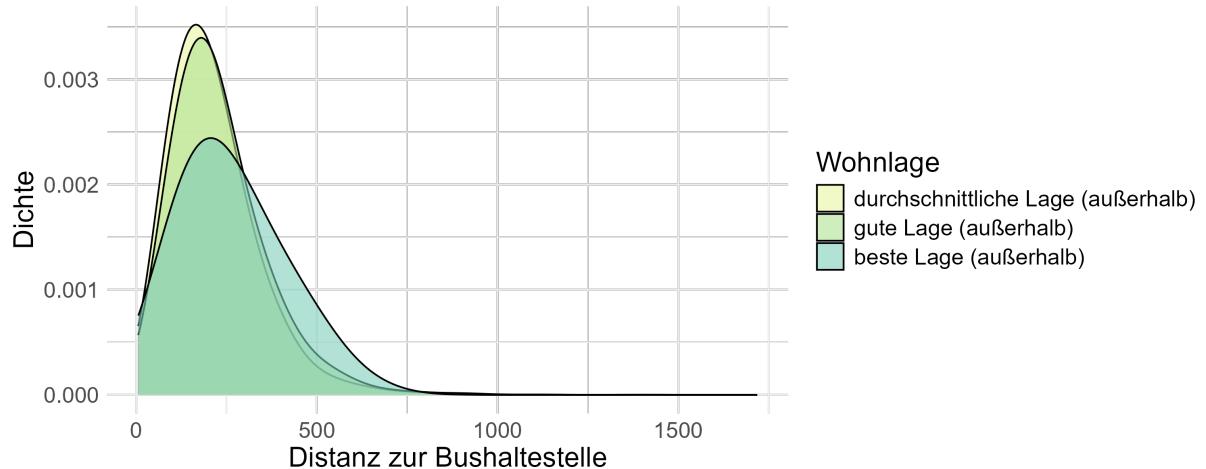


Abbildung 19: Verteilung der Distanz zur Bushaltestelle nach nicht-zentralen Wohnlagen

Ein differenzierteres Bild ergibt sich hingegen für die nicht-zentralen Lagen in Abbildung 19. Hier hebt sich die Verteilung der besten Lage außerhalb von den anderen beiden Kategorien etwas ab, da sie eine geringere Dichte bei den niedrigeren Distanzen aufweist. Zudem zeigt die Verteilung der durchschnittlichen Lage einen langen rechten Rand, was bedeutet, dass einige wenige Objekte dieser Kategorie sehr große Distanzen zur nächsten Haltestelle aufweisen.

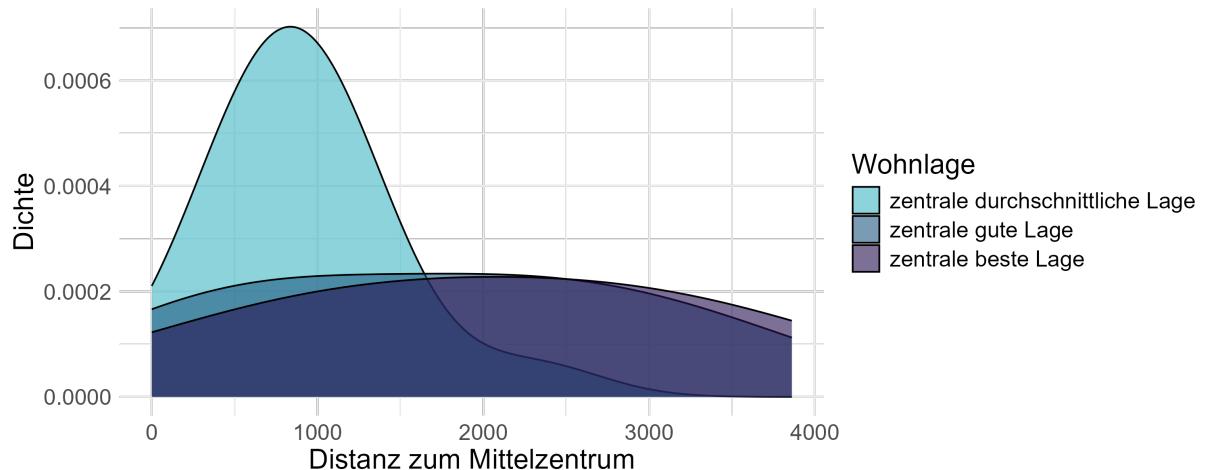


Abbildung 20: Verteilung der Distanz zum Mittelzentrum nach zentralen Wohnlagen

Zusätzlich wird bei der Bestimmung der Wohnlage auch die Nähe zu Einkaufsmöglichkeiten berücksichtigt. Für die Analyse dieses Zusammenhangs werden unter anderem die Variablen `distanz_mittelzentrum` und `distanz_unterzentrum` betrachtet. Diese geben die Entfernung zu Einkaufsbereichen unterschiedlicher Größe an. Befindet sich eine Lokation direkt in einem solchen Bereich, beträgt der Wert der jeweiligen Distanzvariable 0, da kein zusätzlicher Weg dorthin zurückgelegt werden muss. Abbildung 20 zeigt zunächst die Verteilung der Distanz zum Mittelzentrum für zentrale Lagen. Es ergeben sich insgesamt

nur geringe Unterschiede zwischen den guten und besten Lagen. Die durchschnittlichen Lagen stechen dagegen mit einer hohen Dichte für geringe Distanzen hervor.

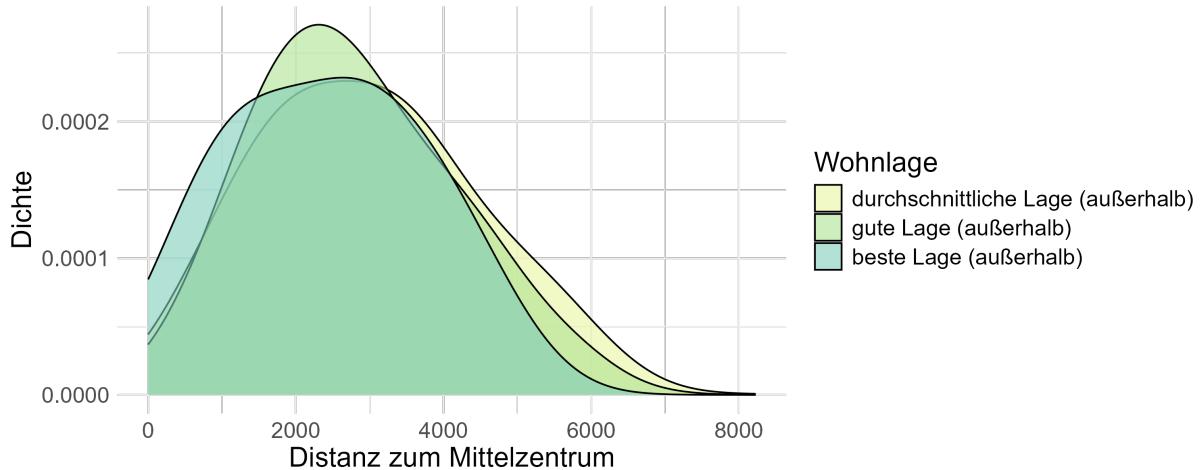


Abbildung 21: Verteilung der Distanz zum Mittelzentrum nach nicht-zentralen Wohnlagen

Für die nicht-zentralen Lagen zeigt sich ein deutlich abweichendes Verteilungsmuster, wie in Abbildung 21 dargestellt. Die Dichteverteilungen der verschiedenen Wohnlagen sind relativ ähnlich zueinander. Zudem treten hier vereinzelt deutlich größere Distanzen auf.

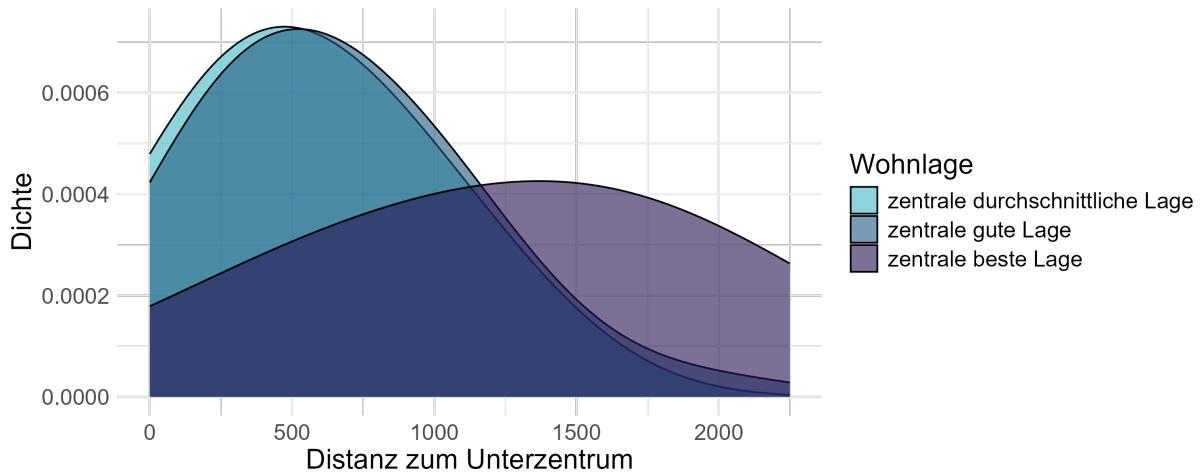


Abbildung 22: Verteilung der Distanz zum Unterzentrum nach zentralen Wohnlagen

Abschließend werden nun die Dichteverteilungen der Distanz zum nächstgelegenen Unterzentrum betrachtet. Aus Abbildung 22 ist erkennbar, dass die Verteilungen für die zentrale gute und zentrale durchschnittliche Lage nahezu deckungsgleich sind. Beide Gruppen weisen eine sehr ähnliche Konzentration von Beobachtungen bei relativ kurzen Distanzen auf. Im Gegensatz dazu hebt sich die Verteilung für die zentrale beste Lage klar von den beiden anderen ab. Insgesamt sind die Entfernungswerte für diese Kategorie im Schnitt größer.

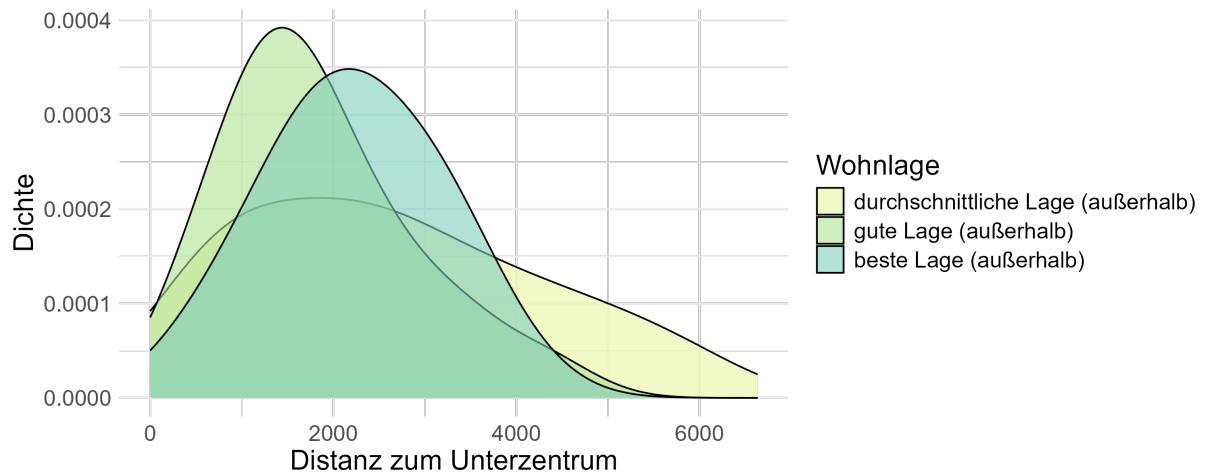


Abbildung 23: Verteilung der Distanz zum Unterzentrum nach nicht-zentralen Wohnlagen

Anders verhält es sich dagegen bei den nicht-zentralen Lagen (Abbildung 23). Dort sind im Mittel die durchschnittlichen Lagen am weitesten vom nächsten Unterzentrum entfernt. Die guten Lagen sind im Durchschnitt am nächsten. Bei den besten Lagen ergeben sich am häufigsten moderate Distanzen.

3.2.3 Verteilung der Indexvariablen

Nun werden ergänzend die Verteilung zweier Indexvariablen untersucht, die eine zusammenfassende und erweiterte Form der zuvor betrachteten Distanzvariablen darstellen. Diese Indizes fassen mehrere Aspekte räumlicher Erreichbarkeit zusammen und verdichten sie zu einem leicht interpretierbaren Maß, das sich besonders für den Vergleich unterschiedlicher Standorte eignet.

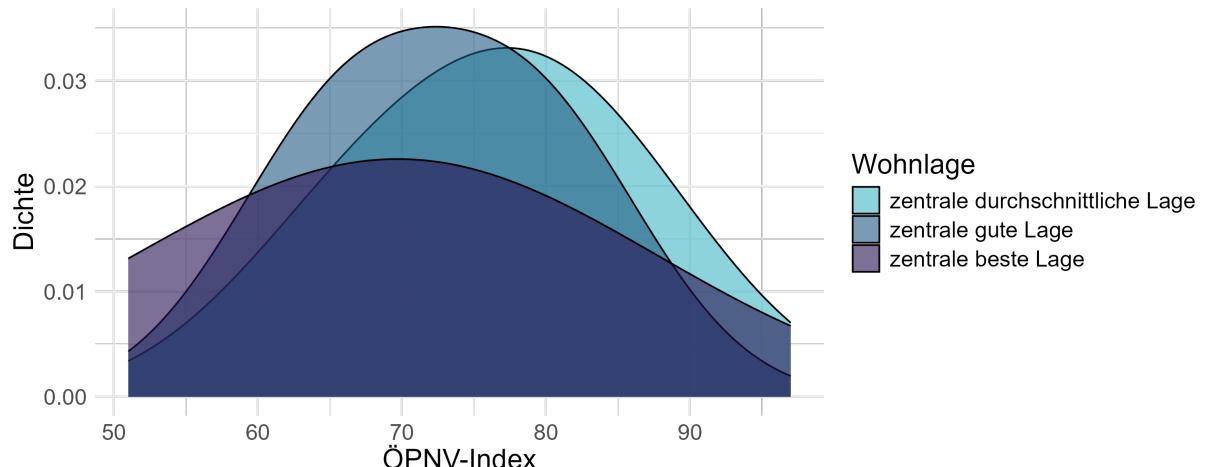


Abbildung 24: Verteilung des ÖPNV-Index nach zentralen Wohnlagen

Einer dieser Indizes ist der ÖPNV-Index, welcher aus den diversen Nahverkehr-Distanzvariablen errechnet wird. Sein Wertebereich ist von 0 bis 100 definiert, wobei eine bessere

ÖPNV-Erreichbarkeit durch einen höheren Wert konnotiert wird. In Abbildung 22 ist die Verteilung dieses Index für die zentralen Lagen dargestellt. Es treten keine Werte unter 50 auf, was folglich bedeutet, dass keine der Lokationen einen besonders schlechten ÖPNV-Anschluss hat.

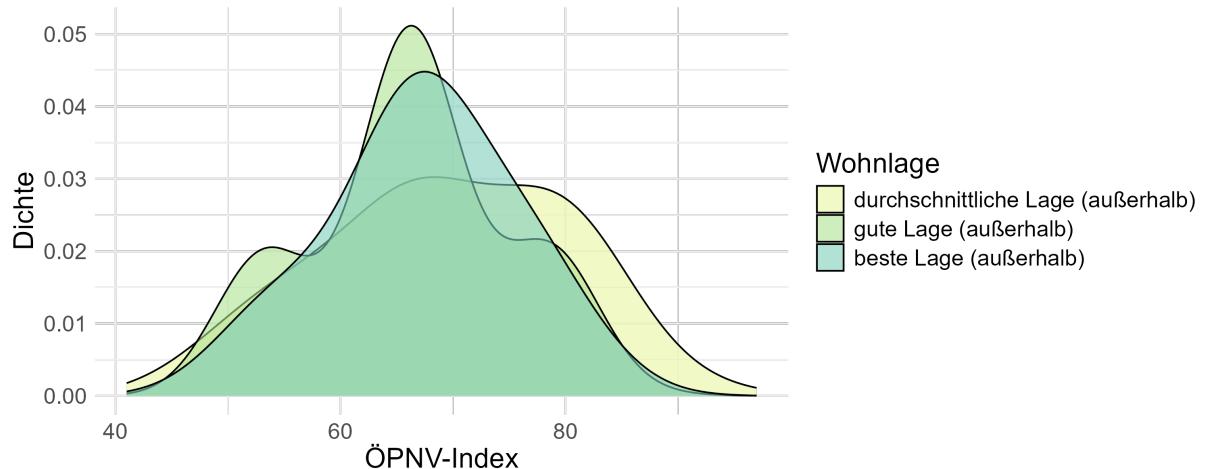


Abbildung 25: Verteilung des ÖPNV-Index nach nicht-zentralen Wohnlagen

Im Gegensatz dazu, wie auch in Abbildung 23 erkennbar, treten bei den nicht-zentralen Lagen einige niedrigere Werte auf. Ebenfalls ist die Dichte bei besonders guter ÖPNV-Erreichbarkeit insgesamt geringer.

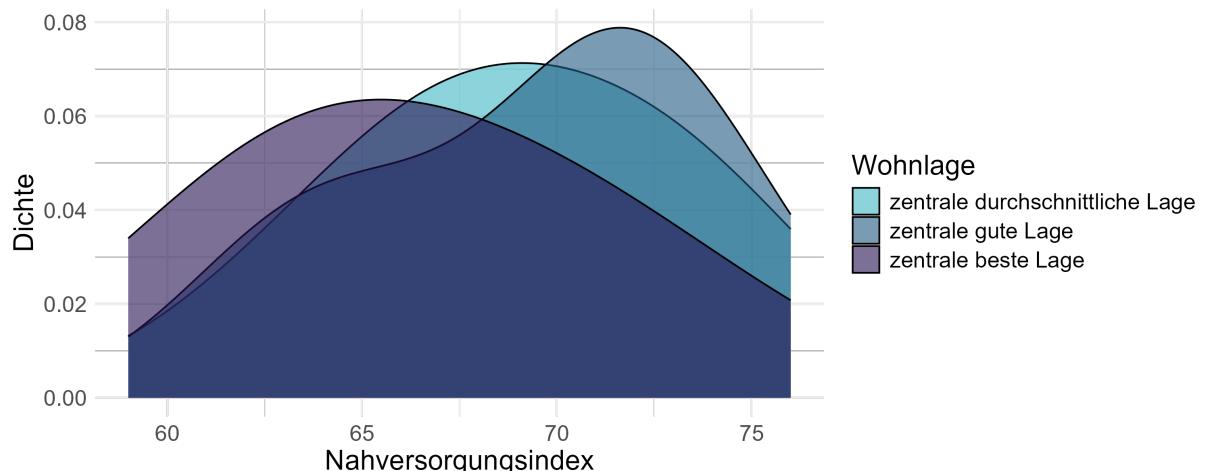


Abbildung 26: Verteilung des Nahversorgungsindex nach zentralen Wohnlagen

Ein weiterer für diese Analyse berücksichtigter Index ist der Nahversorgungsindex. In Abbildung 26 ist seine Verteilung für zentrale Lagen dargestellt. Dieser Index bewertet, wie gut die Bevölkerung an einem Standort mit Waren und Dienstleistungen in kurzer Entfernung vom Wohnort versorgt ist. Ein höherer Wert steht dabei für eine bessere Versorgung. Grundlage für seine Berechnung sind unter anderem die Variablen `distanz_mittelzentrum` und

`distanz_unterzentrum`, die die Entfernung zu Einkaufsbereichen unterschiedlicher Größe messen. Im Schnitt haben die zentralen guten Lagen somit die beste Nahversorgung.

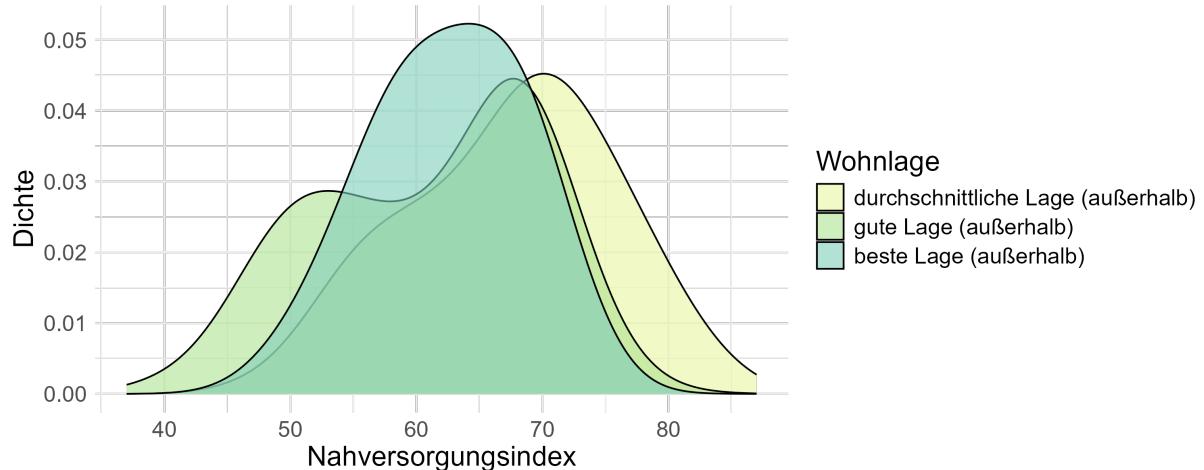


Abbildung 27: Verteilung des Nahversorgungsindex nach nicht-zentralen Wohnlagen

Abbildung 27 zeigt im Vergleich dazu die Verteilung des Nahversorgungsindex für die nicht-zentralen Lagen. Die Ränder der Dichteverteilung fallen hierbei deutlich breiter aus, was darauf hinweist, dass sowohl besonders gute als auch vergleichsweise schlechte Versorgungsniveaus häufiger auftreten.

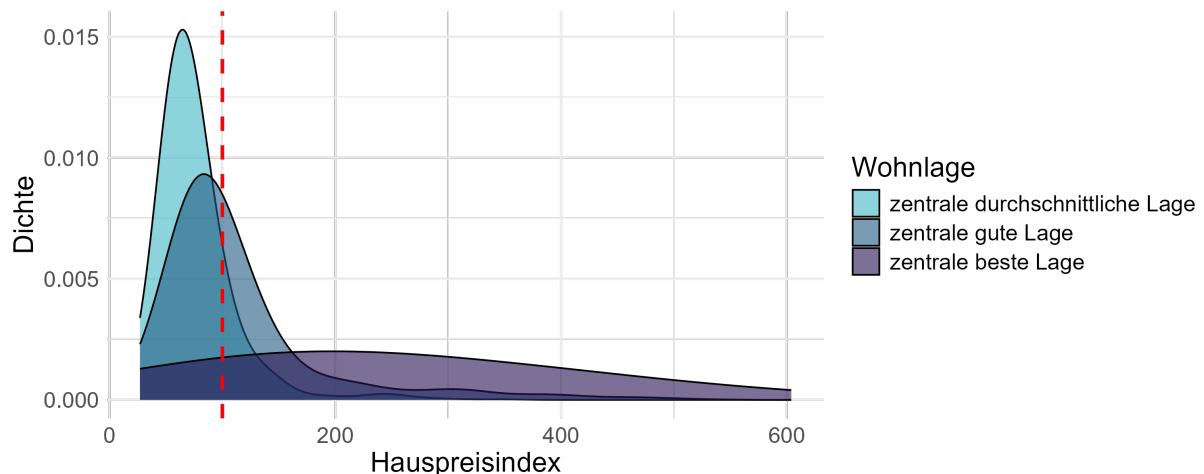


Abbildung 28: Verteilung des Hauspreisindex nach zentralen Wohnlagen

Eine weitere zentrale Größe bei der Bestimmung der Wohnlage einer Lokation ist deren Preisniveau. In dieser Analyse wird dies durch die Variable `hauspreisindex` abgebildet. Ein Wert von 100 entspricht dabei exakt dem durchschnittlichen Hauspreis im jeweiligen Kreis. Werte oberhalb von 100 weisen auf ein überdurchschnittliches Preisniveau hin, während Werte unterhalb von 100 auf günstigere Lagen im Vergleich zum Kreisdurchschnitt schließen lassen. Abbildung 28 zeigt dabei die Verteilung des Hauspreisindex für verschiedene zentrale

Wohnlagentypen. Erwartungsgemäß weisen die zentralen durchschnittlichen Lagen den höchsten Anteil an Wohnobjekten mit einem Preisniveau nahe dem Kreisdurchschnitt auf, was sich in einer schmalen, stark ausgeprägten Dichtespitze um Werte leicht unter 100 widerspiegelt. Im Gegensatz dazu gibt es einige Lokationen mit zentraler bester Lage, die stark von dem Kreisdurchschnitt nach oben abweichen.

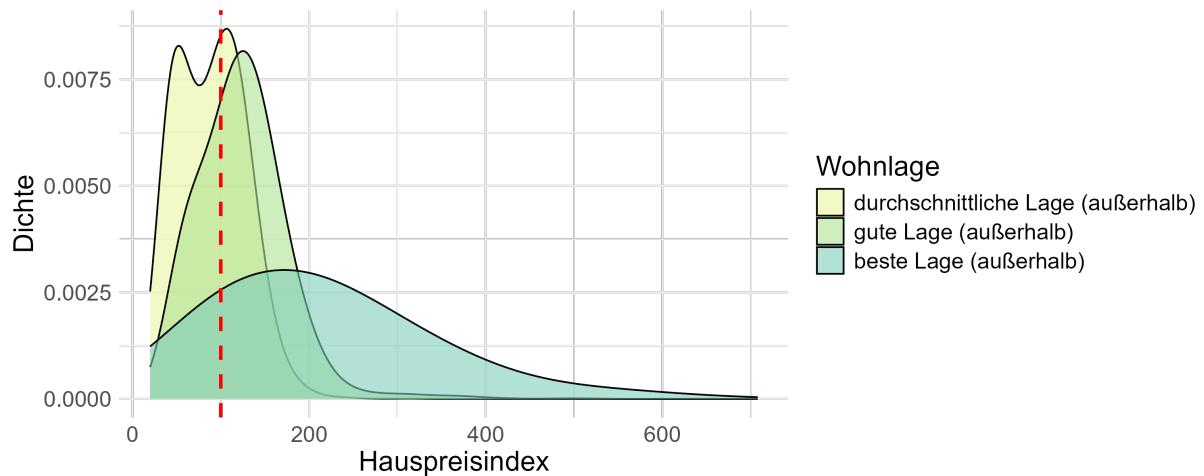


Abbildung 29: Verteilung des Hauspreisindex nach nicht-zentralen Wohnlagen

Bei den nicht-zentralen Lagen, wie auch in Abbildung 29 dargestellt, ist ein ähnlicher Trend zu erkennen.

4 Modellierung

Das folgende Kapitel erläutert das methodische Vorgehen bei der Modellierung der Wohnlagen in München. Ziel ist es, die eingesetzten Verfahren nachvollziehbar darzustellen und deren Eignung zur Vorhersage der Wohnlage im Stadtgebiet München zu begründen. Im weiteren Verlauf dient **Wohnlage** als abhängige Variable, deren Ausprägung durch unabhängige Kovariablen modelliert und vorhergesagt wird.

4.1 Die Multinomialverteilung

Für die Modellierung einer kategorialen Zielgröße wie **Wohnlage** ist die Multinomialverteilung von zentraler Bedeutung. Dabei wird angenommen, dass jede Ausprägung der Zielvariable einer von C möglichen Kategorien zugeordnet werden kann und dass die Beobachtungen gemäß einer multinomialen Verteilung realisiert werden.

Sei $\mathbf{y} = (y_1, \dots, y_C)'$ ein Vektor, der eine einzelne Beobachtung beschreibt. Dabei werden die Komponenten y_1 bis y_C wie folgt definiert:

$$y_i = \begin{cases} 1 & \text{falls Kategorie } i \text{ zutrifft,} \\ 0 & \text{sonst} \end{cases} \quad \text{für } i = 1, \dots, C. \quad (1)$$

Weiter sei $\boldsymbol{\pi} = (\pi_1, \dots, \pi_C)'$ ein Vektor mit den dazugehörigen Erfolgswahrscheinlichkeiten, wobei π_i die Wahrscheinlichkeit für das Auftreten der Kategorie i dieser Beobachtung angibt.

Wird dieser Versuch nun m Mal unabhängig wiederholt, so ergibt sich daraus der Zählvektor $\mathbf{y_m} = (y_{1,m}, \dots, y_{C,m})'$ als Anzahl der Beobachtungen je Kategorie. Damit folgt die Verteilung des Zählvektors y_m der Multinomialverteilung:

$$\mathbf{y_m} \sim \text{Multinomial}(m, \boldsymbol{\pi}).$$

Für die weitere Modellierung wird die Multinomialverteilung als Verteilungsklasse verwendet (vgl. Fahrmeier et al. 2013, S. 327).

4.2 Das Generalisierte Additive Modell

Generalisierte Additive Modelle (GAMs) stellen eine flexible, semi-parametrische Erweiterung der Generalisierten Linearen Modelle (GLMs) dar. Ihre Stärke liegt darin, dass sie die strikte Annahme eines linearen Zusammenhangs zwischen den Prädiktorvariablen und der Zielgröße lockern und somit auch nicht-lineare Effekte modellieren können (vgl. Yee 2015, S. 81). Ähnlich zu den GLMs erlauben GAMs zudem die Annahme einer nicht-normalverteilten Zielvariablen (vgl. Fahrmeier et al. 2013, S. 52).

Das Kernstück eines GAMs ist der additive Prädiktor η_i . Dieser setzt sich aus einer Summe von linearen Funktionen für einige Kovariablen und nicht-linearen, sogenannten Glättungsfunktionen (smooth functions) für andere Kovariablen zusammen. Für eine einzelne Beobachtung $i = 1, \dots, n$ hat der Prädiktor die allgemeine Form:

$$\begin{aligned}\eta_i := & \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \\ & + f_1(x_{i,k+1}) + \cdots + f_{z-k}(x_{iz}).\end{aligned}\quad (2)$$

In dieser Struktur wird für die Kovariablen x_1 bis x_k ein rein linearer Einfluss auf den Prädiktor angenommen, der durch die Koeffizienten β_1 bis β_k quantifiziert wird. Für die Kovariablen x_{k+1} bis x_z wird hingegen ein potenziell nicht-linearer Zusammenhang modelliert, der durch die Glättungsfunktionen f_1 bis f_{z-k} abgebildet wird (vgl. Fahrmeier et al. 2013, S. 56).

Die Beziehung zwischen dem Erwartungswert der Zielvariable $E(y_i) = \mu_i$ und dem additiven Prädiktor η_i wird über eine Link-Funktion $g(\cdot)$ hergestellt, sodass $g(\mu_i) = \eta_i$ gilt (vgl. Fahrmeier et al. 2013, S. 270).

4.3 Multinomiale GAMs

Bei der Modellierung einer kategorialen Zielgröße Y mit den Klassen $\{1, \dots, c, c+1\}$ durch ein generalisiertes additives Modell wird konkret wie folgt vorgegangen.

Zunächst wird eine beliebige Kategorie als Referenz ausgewählt. Im Folgenden sei dies die Klasse $c+1$. Für die restlichen Kategorien wird anschließend jeweils ein Prädiktor η analog zu Gleichung (2) definiert. Um nun die Wahrscheinlichkeit für eine beliebige Klasse $r \in \{1, \dots, c\} = R$ einer Beobachtung $i = 1, \dots, n$ zu modellieren, kann folgende Gleichung verwendet werden:

$$P(Y_i = r) = \pi_{r,i} = \frac{\exp(\eta_{r,i})}{1 + \sum_{s \in R} \exp(\eta_{s,i})}, \quad r \in \{1, \dots, c\} = R. \quad (3)$$

Die Wahrscheinlichkeit der Referenzkategorie $c+1$ ergibt sich dann aus den Wahrscheinlichkeiten der restlichen Kategorien:

$$P(Y_i = c+1) = \pi_{c+1,i} = 1 - \pi_{1,i} - \dots - \pi_{c,i} = \frac{1}{1 + \sum_{s \in R} \exp(\eta_{s,i})}. \quad (4)$$

Die dazugehörige Link-Funktion, welche die sogenannten Log Odds modelliert, ist somit wie folgt definiert:

$$g(\pi_{r,i}) = \eta_{r,i} = \ln \left(\frac{\pi_{r,i}}{\pi_{c+1}} \right). \quad (5)$$

Dadurch kann die Beziehung zwischen dem Erwartungswert $E(\mathbf{y}) = (\pi_1 \dots \pi_{c+1})'$ einer Beobachtung Y_i und dem linearen Prädiktor $\eta_{i,r}$ hergestellt werden. Insgesamt wird jede Kategorie $r \in R = \{1, \dots, c\}$ relativ zur ausgewählten Referenzkategorie $c+1$ modelliert (vgl. Fahrmeier et al. 2013, S. 329f).

4.4 Modellierung der Wohnlagen

Im Folgenden werden die Wohnlagen der Standorte im Datensatz durch zwei separate multinomiale Generalisierte Additive Modelle modelliert. Dabei wird jeweils ein Modell für die zentralen Lagen (durchschnittliche, gute und beste zentrale Lage) und ein Modell für die nicht-zentralen Lagen (durchschnittliche, gute und beste Lage außerhalb) geschätzt. Diese Vorgehensweise ermöglicht es, den Einfluss der einzelnen Kovariablen zu quantifizieren und Vorhersagen für die Wohnlagen der einzelnen Lokationen zu treffen.

Beide Modelle berücksichtigen die in Tabelle 2 aufgeführten Kovariablen. Da für die metrischen Variablen ein nicht-linearer Zusammenhang mit der Wohnlage vermutet wird, werden für diese, analog zu Kapitel 4.2, Glättungsfunktionen verwendet. Als Referenzkategorien dienen die „durchschnittliche zentrale Lage“ bzw. die „durchschnittliche Lage außerhalb“, da die übrigen Kategorien eine qualitative Steigerung abbilden. Für beide Modelle werden die Prädiktoren wie folgt definiert:

$$\begin{aligned} \eta_{gut,i} := & \beta_{0,gut} + \beta_{1,gut} \cdot I\{\text{straßentyp}_i = \text{Sammelstraße}\} \\ & + \beta_{2,gut} \cdot I\{\text{straßentyp}_i = \text{Wohnstraße}\} \\ & + \beta_{3,gut} \cdot I\{\text{straßentyp}_i = \text{Fußgängerbereich}\} \\ & + f_{1,gut}(\text{distanz_bahnhof}_i) + f_{2,gut}(\text{distanz_mittelzentrum}_i) \\ & + f_{3,gut}(\text{opnv_index}_i) + f_{4,gut}(\text{distanz_unterzentrum}_i) \\ & + f_{5,gut}(\text{hauspreis_index}_i) + f_{6,gut}(\text{distanz_ubahn}_i) \\ & + f_{7,gut}(\text{distanz_bushaltestelle}_i) + f_{8,gut}(\text{nahversorgungs_index}_i) \end{aligned} \quad (6)$$

und

$$\begin{aligned} \eta_{beste,i} := & \beta_{0,beste} + \beta_{1,beste} \cdot I\{\text{straßentyp}_i = \text{Sammelstraße}\} \\ & + \beta_{2,beste} \cdot I\{\text{straßentyp}_i = \text{Wohnstraße}\} \\ & + \beta_{3,beste} \cdot I\{\text{straßentyp}_i = \text{Fußgängerbereich}\} \\ & + f_{1,beste}(\text{distanz_bahnhof}_i) + f_{2,beste}(\text{distanz_mittelzentrum}_i) \\ & + f_{3,beste}(\text{opnv_index}_i) + f_{4,beste}(\text{distanz_unterzentrum}_i) \\ & + f_{5,beste}(\text{hauspreis_index}_i) + f_{6,beste}(\text{distanz_ubahn}_i) \\ & + f_{7,beste}(\text{distanz_bushaltestelle}_i) + f_{8,beste}(\text{nahversorgungs_index}_i). \end{aligned} \quad (7)$$

Für die kategoriale Einflussvariable **straßentyp** werden jeweils drei Koeffizienten geschätzt, wobei die „Hauptstraße“ als gemeinsame Referenzkategorie dient.

Zur Schätzung der beiden Modelle eignet sich die Funktion **gam** aus dem R-Paket **mgcv** (vgl. Wood 2017b, S. 58) und wird daher verwendet.

Für diese Analyse werden die einzelnen nicht-linearen Funktionsterme des Modells durch Splines approximiert. Im Kern ist ein Spline eine hochflexible Funktion, die aus mehreren einfachen Polynomstücken zusammengesetzt ist. Diese Stücke werden an bestimmten Punkten, den sogenannten „Knoten“, so nahtlos miteinander verbunden, dass eine durchgehend

glatte Kurve entsteht. Konkret kommen für die Modellierung der zentralen Wohnlagen Thin Plate Regression Splines mit einer Basisdimension von $k = 10$ und für die Modellierung der nicht-zentralen Lagen die recheneffizienteren und robusteren Cubic Regression Splines mit einer Basisdimension von $k = 11$ zum Einsatz (vgl. Wood 2017b, S.290). Zur Schätzung der Glättungsparameter dieser Splines, welche die Flexibilität bzw. „Kurvigkeit“ der einzelnen Funktionsterme steuern, wird auf Basis des erweiterten Fellner-Schall-Algorithmus die penalisierten Likelihood (penalized likelihood) maximiert. Dieser Prozess findet einen optimalen Kompromiss zwischen einer zu starren Anpassung, auch genannt Underfitting, und einer zu flexiblen Anpassung an das Datenrauschen, auch genannt Overfitting. Das Verfahren ist für diese Aufgabe besonders geeignet, da es recheneffizient arbeitet und die Schätzung auch für komplexe Modelle robust ermöglicht (vgl. Wood 2017a). Es ist zusätzlich zu beachten, dass die Zielvariable **Wohnlage** mit den Klassen $\{\text{durchschnittlich}, \text{gut}, \text{beste}\}$ die Class-Labels $\{0, 1, 2\}$ besitzen muss (vgl. Wood 2017b, S.170).

Die Wahrscheinlichkeiten der verschiedenen Klassen von **Wohnlage** können nach dieser Schätzung der Modelle analog zu Kapitel 4.3 berechnet werden:

$$\begin{aligned} P(Wohnlage_i = R) &= \pi_{i,r} = \frac{\exp(\eta_{r,i})}{1 + \sum_{s \in R} \exp(\eta_{s,i})} \\ &= \frac{\exp(\eta_{r,i})}{1 + \exp(\eta_{\text{gut},i}) + \exp(\eta_{\text{beste},i})} \end{aligned} \quad (8)$$

wobei $r \in \{\text{gut}, \text{beste}\} = R$. Konkret gilt:

$$P(Wohnlage_i = \text{gut}) = \pi_{\text{gut},i} = \frac{\exp(\eta_{\text{gut},i})}{1 + \exp(\eta_{\text{gut},i}) + \exp(\eta_{\text{beste},i})} \quad (9)$$

und

$$P(Wohnlage_i = \text{beste}) = \pi_{\text{beste},i} = \frac{\exp(\eta_{\text{beste},i})}{1 + \exp(\eta_{\text{gut},i}) + \exp(\eta_{\text{beste},i})}. \quad (10)$$

Für die Referenz „durchschnittlich“ folgt dann:

$$\begin{aligned} P(Wohnlage_i = \text{durchschnittlich}) &= \pi_{\text{durchschnittlich},i} = 1 - \pi_{\text{gut},i} - \pi_{\text{beste},i} \\ &= \frac{1}{1 + \sum_{s \in R} \exp(\eta_{s,i})} \\ &= \frac{1}{1 + \exp(\eta_{\text{gut},i}) + \exp(\eta_{\text{beste},i})}. \end{aligned} \quad (11)$$

Die Verteilungsannahme ist somit in beiden Modellen für die Lokationen 1,..., n gegeben durch:

$$\text{Wohnlage}_i \mid \mathbf{x}_i \sim \text{Multinomial} \left(n, \begin{pmatrix} \pi_{i,\text{durchschnittlich}} \\ \pi_{i,\text{gut}} \\ \pi_{i,\text{beste}} \end{pmatrix} \right). \quad (12)$$

Für jedes Wohnobjekt i kann folglich nach Erstellung der Modelle der Wahrscheinlichkeitsvektor π_i berechnet werden. Die Wohnlage mit dem jeweils höchsten Wert bzw. der höchsten Wahrscheinlichkeit wird dann als Vorhersage für i verwendet.

5 Effekte der Einflussvariablen

In diesem Kapitel werden zunächst die geschätzten Effekte der zuvor vorgestellten GAMs interpretiert, um die Bedeutung der einzelnen Einflussgrößen für die Wohnlagen Münchens herauszuarbeiten. Zudem wird die Gültigkeit dieser Interpretationen mit einer Korrelationsanalyse und einem alternativen Modellierungsansatz bestätigt.

5.1 Interpretation des Effekts des Straßentyps

Im Folgenden werden die Effekte der kategorialen Variable **straßentyp** auf die Wohnlage analysiert. Hierfür werden die vom Modell geschätzten Koeffizienten $\beta_{1,gut}$ bis $\beta_{3,gut}$ und $\beta_{1,beste}$ bis $\beta_{3,beste}$, welche die Veränderung der Log-Quoten (Log-Odds) beschreiben, mithilfe der inversen Link-Funktion, in diesem Fall der Exponentialfunktion, in Odds-Ratios (OR) umgerechnet (vgl. Kapitel 4).

Ein Odds-Ratio ist das Verhältnis zweier Quoten und lässt sich oft intuitiver interpretieren als der reine Koeffizient β . Ein OR von 1 bedeutet, dass kein Effekt vorliegt, während Werte größer oder kleiner als 1 auf einen positiven bzw. negativen Einfluss der jeweiligen Kategorie im Vergleich zur Referenzkategorie hindeuten (vgl. Agresti 2007, S. 111).

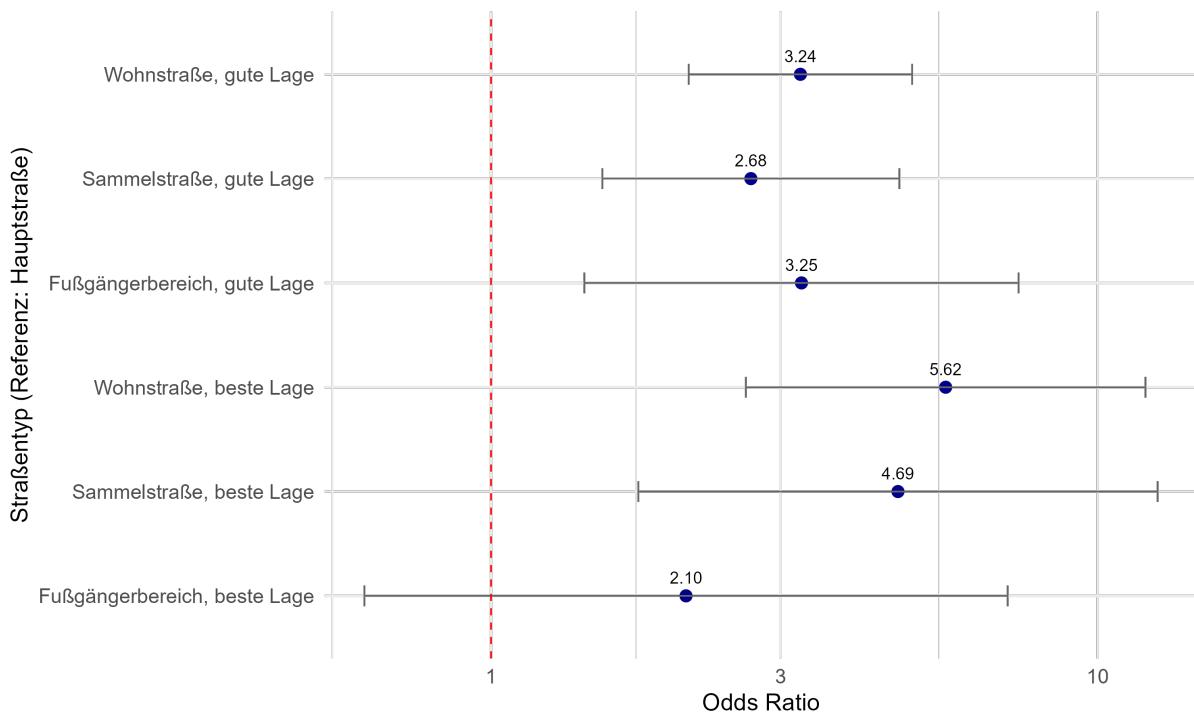


Abbildung 30: Odds-Ratios und 95%-Konfidenzintervalle: Effekt des Straßentyps auf die zentralen Lagen

Abbildung 30 stellt die Effekte des Straßentyps als Odds-Ratios aus dem multinomialen GAM für zentrale Lagen dar. Jeder Punkt repräsentiert das geschätzte OR für eine bestimmte Kombination aus **straßentyp** und **Wohnlage**. Die horizontalen grauen Linien

zeigen die zugehörigen 95%-Konfidenzintervalle, welche die statistische Unsicherheit der Schätzung abbilden (vgl. Fahrmeier et al. 2016, S. 364). Die vertikale rote Linie bei einem Odds Ratio von 1 markiert die Referenzlinie, bei der kein Effekt vorliegt. Die Grafik zeigt somit, dass der Straßentyp einen überwiegend signifikanten Einfluss auf die Einordnung einer zentralen Wohnlage als „gut“ oder „beste“ hat. Die Y-Achse listet die verschiedenen Vergleichsgruppen auf, wobei die „Hauptstraße“ als Referenzkategorie für den Straßentyp dient.

Nun sollen die konkreten Effekte interpretiert werden. Dabei ist zu beachten, dass die folgenden Aussagen nur jeweils unter Konstanthaltung aller anderen Variablen im Modell (*ceteris paribus*) und relativ zur durchschnittlichen Lage gelten.

Der stärkste signifikante Effekt zeigt sich bei der „Wohnstraße, beste Lage“ (OR = 5,62). Für eine Immobilie an einer Wohnstraße sind die Odds bzw. Chance, in einer zentralen besten Lage (anstatt einer zentralen durchschnittlichen Lage) zu sein, 5,62-mal so hoch wie für eine Immobilie an einer Hauptstraße.

Ein ebenfalls sehr starker Effekt ist für die „Sammelstraße, beste Lage“ (OR = 4,69) zu beobachten. Die Chance, in einer zentralen besten Lage (anstatt einer zentralen durchschnittlichen Lage) zu sein, ist für Standorte an einer Sammelstraße immer noch fast 4,69-mal so hoch wie an einer Hauptstraße.

Auch für die Kategorie „gute Lage“ zeigen sich signifikante, positive Effekte. Für die „Wohnstraße, gute Lage“ (OR = 3,24) sind die Odds auf eine zentrale gute Lage (statt einer zentralen durchschnittlichen) 3,24-mal so hoch wie für die Referenzkategorie „Hauptstraße“ (vgl. Fahrmeier et al. 2013, S.331).

Im Gegensatz dazu ist der Effekt für den „Fußgängerbereich, beste Lage“ (OR = 2,10) nicht statistisch signifikant. Obwohl der Punktschätzer auf einen mehr als verdoppelten Effekt hindeutet, schließt das Konfidenzintervall den Wert 1 mit ein. Es gibt somit keinen statistisch belastbaren Nachweis, dass sich die Chance auf eine zentrale beste Lage in einem Fußgängerbereich signifikant von der in einer Hauptstraße unterscheidet. Dies ist vor allem auf die geringen Stichprobengrößen der Kategorien „Fußgängerbereich“ und „beste Lage“ zurückzuführen.

Zusammenfassend lässt sich festhalten, dass fast alle untersuchten Straßentypen die Wahrscheinlichkeit für eine bessere Wohnlageneinstufung im Vergleich zur Hauptstraße signifikant erhöhen, wobei Wohnstraßen den insgesamt größten positiven Einfluss haben.

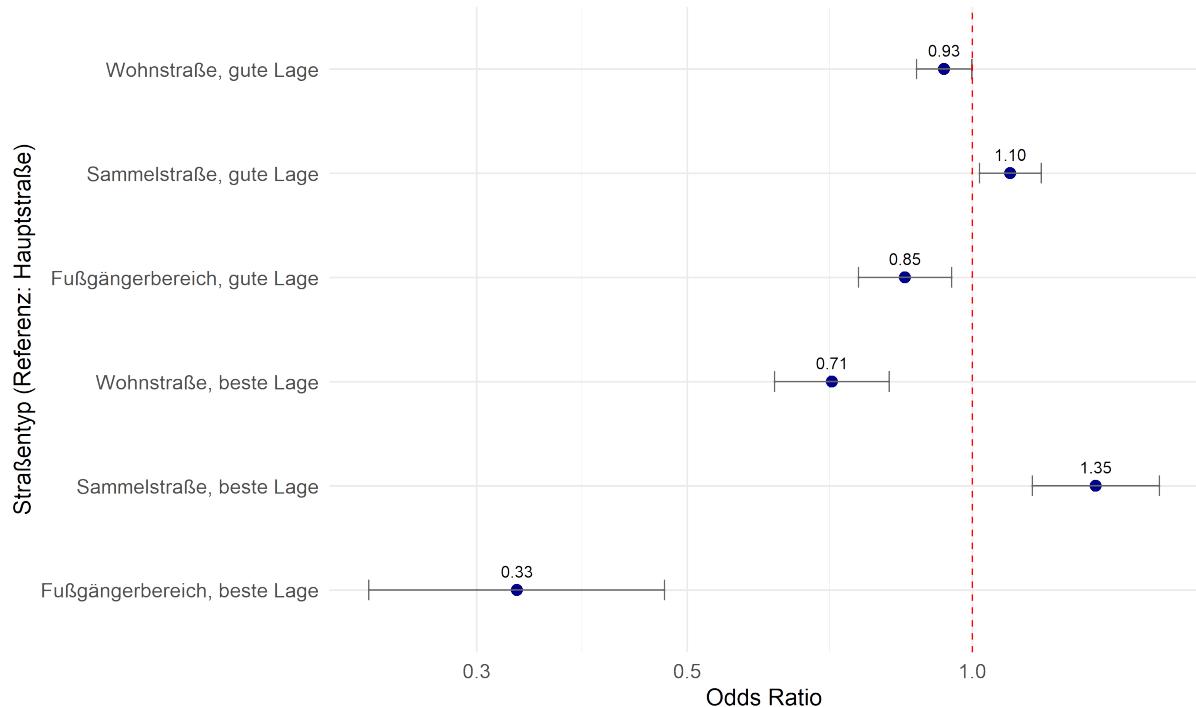


Abbildung 31: Odds-Ratios und 95%-Konfidenzintervalle: Effekt des Straßentyps auf die Lagen außerhalb

Abbildung 31 visualisiert die Effekte des Straßentyps aus dem GAM für die nicht-zentralen Lagen. Dargestellt sind erneut die Odds-Ratios (OR) und die zugehörigen 95%-Konfidenzintervalle. Die Interpretation erfolgt analog zum vorherigen Modell. Die Referenzkategorie für den Straßentyp ist wieder die Hauptstraße, und die Effekte beziehen sich auf den Vergleich zur durchschnittlichen Lage außerhalb.

Alle dargestellten Effekte sind statistisch signifikant und zeigen, dass der Straßentyp auch außerhalb des Zentrums einen entscheidenden, aber im Vergleich zu den zentralen Lagen völlig anderen Einfluss hat. Bei den folgenden Interpretationen werden erneut alle anderen Variablen im Modell konstant gehalten (*ceteris paribus*).

Der stärkste positive Effekt ist für die „Sammelstraße, beste Lage“ (OR = 1,35) zu beobachten. Die Chance, in einer besten Lage außerhalb (anstatt einer durchschnittlichen Lage außerhalb) zu sein, ist für Standorte an einer Sammelstraße also um 35 % höher als für Standorte an einer Hauptstraße.

Im starken Kontrast dazu steht der stärkste negative Effekt des Modells, der sich für den „Fußgängerbereich, beste Lage“ (OR = 0,33) zeigt. Für ein Wohnobjekt in einem Fußgängerbereich sind die Odds, in einer besten Lage außerhalb (statt einer „durchschnittlichen Lage außerhalb) zu sein, um 67 % geringer als für eine Immobilie an einer Hauptstraße.

Auch für die Kategorie „gute Lage außerhalb“ zeigen sich signifikant negative Effekte. Für die „Wohnstraße, gute Lage“ (OR = 0,93) sind die Odds auf eine „gute Lage außerhalb“

(statt einer durchschnittlichen) um 7 % geringer als in der Referenzkategorie Hauptstraße. Allerdings liegt die obere Grenze des Konfidenzintervalls ganz knapp unter 1, daher ist die Evidenz für diesen negativen Effekt als eher schwach einzustufen.

Zusammenfassend lässt sich festhalten, dass in nicht-zentralen Gebieten, anders als im Zentrum, Wohnstraßen und Fußgängerbereiche die Wahrscheinlichkeit für eine bessere Wohnlageneinstufung im Vergleich zur Hauptstraße signifikant verringern. Lediglich die Sammelstraße weist durchgehend einen moderaten, positiven Einfluss auf.

5.2 Interpretation der Effekte der metrischen Variablen

Da die metrischen Variablen in den beiden Modellen durch flexible Glättungsfunktionen (Splines) abgebildet werden, ist ihr Effekt nicht durch einen einzelnen Koeffizienten beschreibbar. Stattdessen werden sogenannte partielle EffektpLOTS verwendet, um den funktionalen Zusammenhang zwischen einer metrischen Variable und der jeweiligen Zielgröße des Modells zu visualisieren.

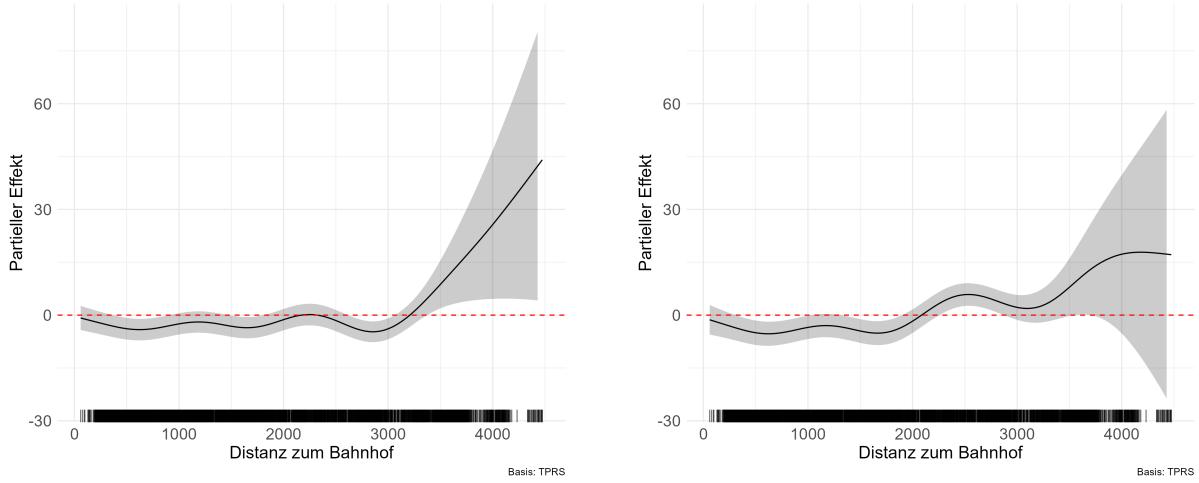
Die Y-Achse dieser Plots stellt den Einfluss der Variable auf der Skala des Prädiktors (hier der Log-Odds-Skala) dar und zeigt konkret, wie sich die Log-Odds für eine gute bzw. beste Lage im Vergleich zur Referenzkategorie „durchschnittliche Lage“ verändern. Eine direkte Umrechnung dieser Achse in Odds-Ratios, wie es bei **straßentyp** möglich war, ist hier nicht sinnvoll. Der Grund dafür ist, dass der Effekt einer metrischen Variable nicht konstant ist, sondern sich mit ihrem Wert verändert. Die Kurve zeigt genau diese Veränderung, während alle anderen Variablen im Modell konstant gehalten werden (Ceteris-Paribus-Bedingung).

Obwohl die Achse selbst auf der Log-Odds-Skala verbleibt, ist es dennoch möglich, eine interpretierbare Odds-Ratio für den Vergleich zweier spezifischer Punkte auf der Kurve zu berechnen. Hierfür wird die Differenz der Y-Werte (der Log-Odds) an zwei gewählten Punkten auf der X-Achse gebildet und anschließend exponentiert. So lässt sich beispielsweise quantifizieren, um wie viel sich die Chance auf eine bessere Wohnlage verändert, wenn die Distanz zum Bahnhof von 500 auf 3000 Metern ansteigt. Die primäre Interpretation der Plots konzentriert sich jedoch auf die qualitative Form der dargestellten Zusammenhänge.

Die Kurve im partiellen EffektpLOT ist die geschätzte Glättungsfunktion, deren Form die Art des Zusammenhangs offenbart. Das graue Band um die Linie ist das 95 %-Konfidenzintervall und visualisiert die statistische Unsicherheit. Die vertikalen schwarzen Linien an der X-Achse geben den Variablenwert jeder im Datensatz vorhandenen Beobachtung an (vgl. Clark 2024).

5.2.1 Interpretation der Effekte der Distanzvariablen

Zunächst werden die nicht-linearen Effekte der fünf Distanzvariablen auf die zentralen und nicht-zentralen Wohnlagen interpretiert.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 32: Partieller Effekt der Distanz zum Bahnhof auf die zentrale Wohnlage

Abbildung 32 zeigt den Einfluss der Distanz zum nächsten Bahnhof auf die Chance, eine gute bzw. beste zentrale Lage im Vergleich zu einer durchschnittlichen zentralen Lage als Wohnobjekt zu besitzen.

Für den Effekt auf die zentrale gute Lage (Abbildung 32a) zeigt sich ein leichter wellenförmiger Verlauf. Eine unmittelbare Nähe zum Bahnhof wirkt sich zunächst negativ auf die Chance einer guten Lage (im Vergleich zu einer durchschnittlichen Lage) aus. Im mittleren Distanzbereich ist dies ebenfalls der Fall. Allerdings ist die Steigung der Kurve häufig nicht signifikant von Null verschieden, da das Konfidenzintervall in diesem Bereich häufig die Nulllinie schneidet. Dies bedeutet, dass eine Entfernungssänderung in diesem Intervall keinen statistisch belastbaren Effekt hat. Erst ab einer Distanz von ca. 3200 Metern steigt die Chance auf eine gute zentrale Lage an. Es wird eine sehr hohe Steigung für die Glättungsfunktion geschätzt, allerdings sind hier die breiten Konfidenzintervalle zu beachten, die auf eine erhöhte Unsicherheit in diesem Intervall hindeuten.

Ein ähnliches Muster zeigt sich für den Effekt auf die zentrale beste Lage (Abbildung 32b). Hier ist die Chance auf diese Wohnlage im Vergleich zur Referenz „durchschnittliche zentrale Lage“ bei kürzeren Distanzen ebenfalls negativ. Ab einer mittleren Entfernung von ca. 2100 Metern steigt dagegen die Chance für eine beste Lage. Für besonders weite Strecken ab ca. 3000 Metern ist dieser Effekt jedoch aufgrund des breiten Konfidenzintervalls nicht signifikant.

Zusammenfassend lässt sich sagen, dass für zentrale Wohnlagen die direkte Bahnhofsnahe tendenziell ein Nachteil ist, vermutlich aufgrund negativer externer Effekte wie Lärm.

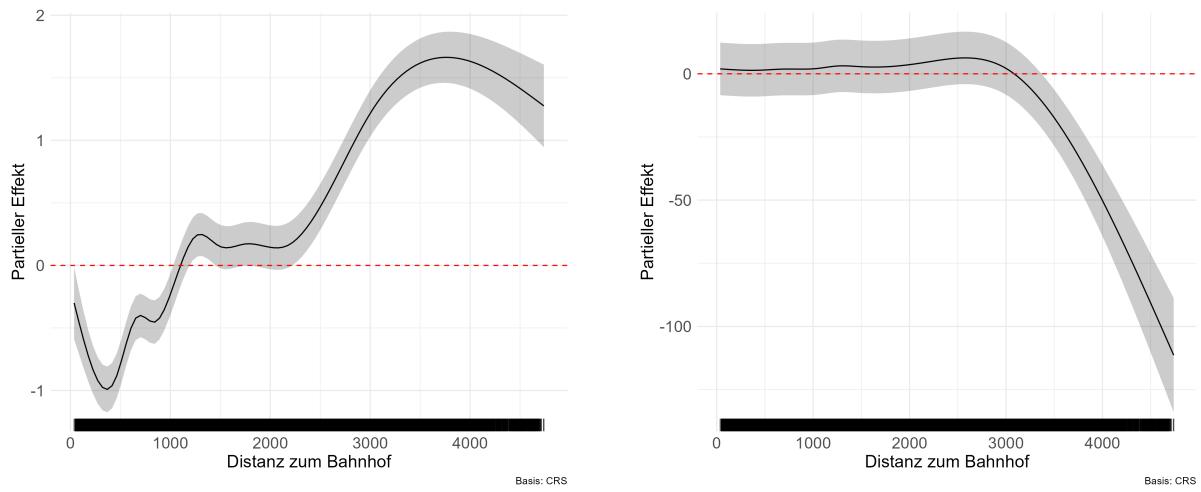
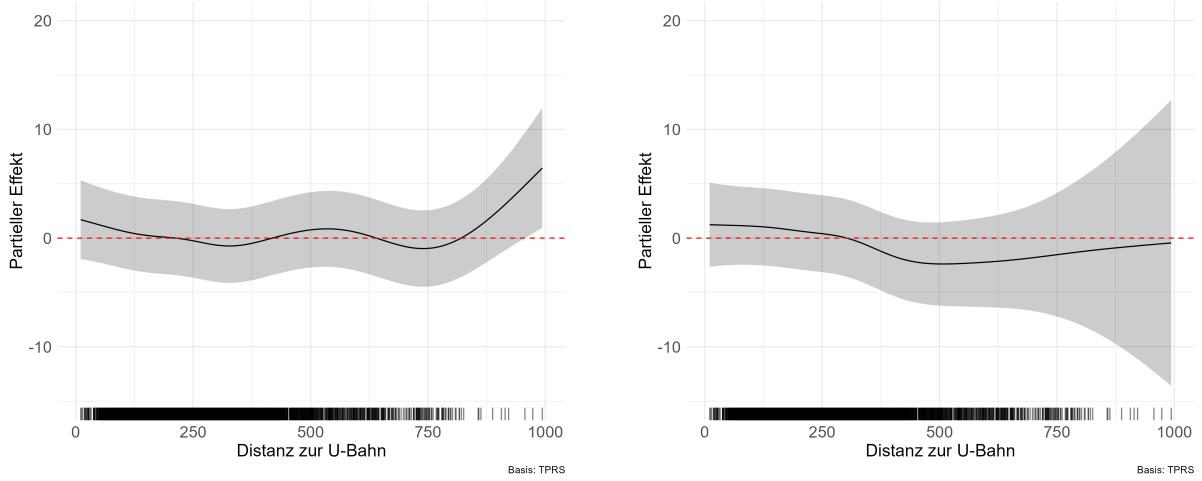


Abbildung 33: Partieller Effekt der Distanz zum Bahnhof auf die Wohnlage außerhalb

Ein noch differenzierteres und teilweise gegensätzliches Bild ergibt sich bei der Betrachtung der nicht-zentralen Lagen in Abbildung 33. Ähnlich wie im Zentrum ist auch hier die unmittelbare Bahnhofsnahe für eine gute Lage im Vergleich zu einer durchschnittlichen Lage ein signifikanter negativer Faktor. Ab ca. 1050 Metern Entfernung steigt dagegen die Chance auf eine gute Lage im Vergleich zu einer durchschnittlichen.

Für die beste Lage außerhalb zeigt sich dagegen ein anderes Muster. Für Distanzen bis ca. 3100 Metern ist kein signifikanter Unterschied zu der Referenz „durchschnittliche Lage außerhalb“ festzustellen. Für Distanzen in diesem Bereich steigt oder sinkt somit die Chance auf eine beste Lage außerhalb nicht signifikant. Erst ab einer Entfernung von ca. 3100 Metern reduziert sich die Chance auf eine beste Lage außerhalb stark. Man beachte die unterschiedliche Skalierung der Y-Achse.

Insgesamt lässt sich festhalten, dass die optimale Distanz zum Bahnhof stark von der Makrolage abhängt. Während in zentralen Lagen eine größere Distanz vorteilhaft sein kann, existiert für Spitzenlagen außerhalb des Zentrums eine klare Obergrenze, ab der die Lagequalität durch eine zu große Entfernung zur Infrastruktur leidet.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

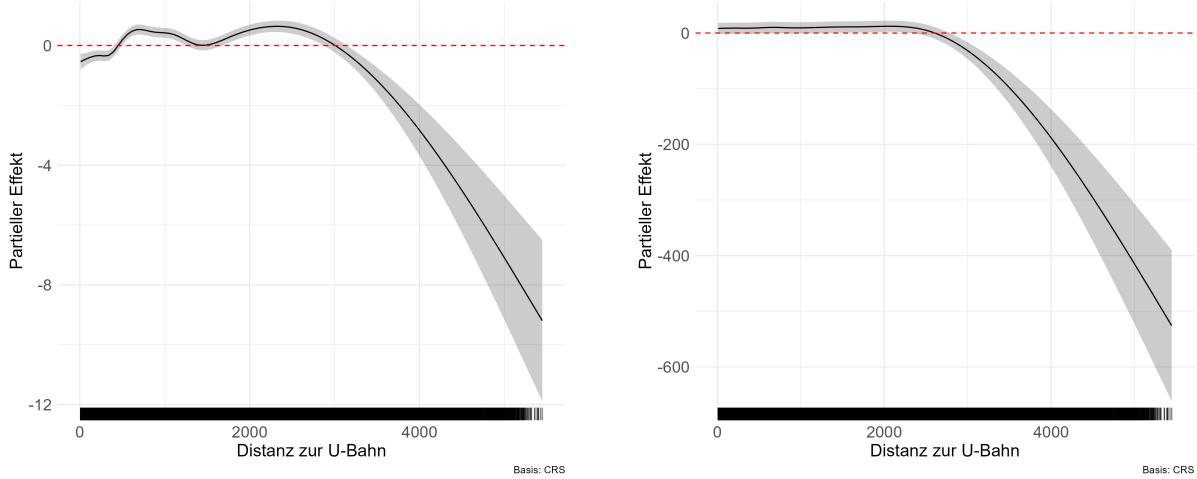
Abbildung 34: Partieller Effekt der Distanz zur U-Bahn auf die zentrale Wohnlage

Abbildung 34 zeigt nun den partiellen Effekt der Distanz zur nächsten U-Bahn-Station auf die zentralen Wohnlagen.

Für die gute Lage (a) zeigt sich ein leicht wellenförmiger, um die Nulllinie schwankender Verlauf. Auffällig ist hier, dass der Einfluss dieser Variable über weite Teile ihres Wertebereichs statistisch nicht signifikant ist, da das graue 95%-Konfidenzband die rote Nulllinie fast vollständig einschließt. Dieses Ergebnis ist plausibel, da sich bereits in der deskriptiven Analyse (vgl. Abbildung 16) gezeigt hat, dass die Verteilungen der U-Bahn-Distanzen der zentralen durchschnittlichen und der zentralen guten Lage nahezu identisch sind. Auf dieser Basis kann das Modell kaum einen robusten Unterschied zwischen den beiden Gruppen finden.

Für die beste Lage (b) lässt sich über den gesamten Wertebereich kein statistisch signifikanter Effekt der U-Bahn-Distanz nachweisen. Obwohl die Kurve tendenziell einen negativen Verlauf im mittleren Bereich ab ca. 280 Metern Entfernung andeutet, ist die statistische Unsicherheit so groß, dass ein Nulleffekt nicht ausgeschlossen werden kann.

Insgesamt lässt sich also festhalten, dass die Distanz zur nächsten U-Bahn-Station im Modell für die zentralen Wohnlagen ein eher schwacher Prädiktor ist, da keine eindeutigen Strukturen und Unterschiede zwischen den Wohnlagen festgestellt werden konnten.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

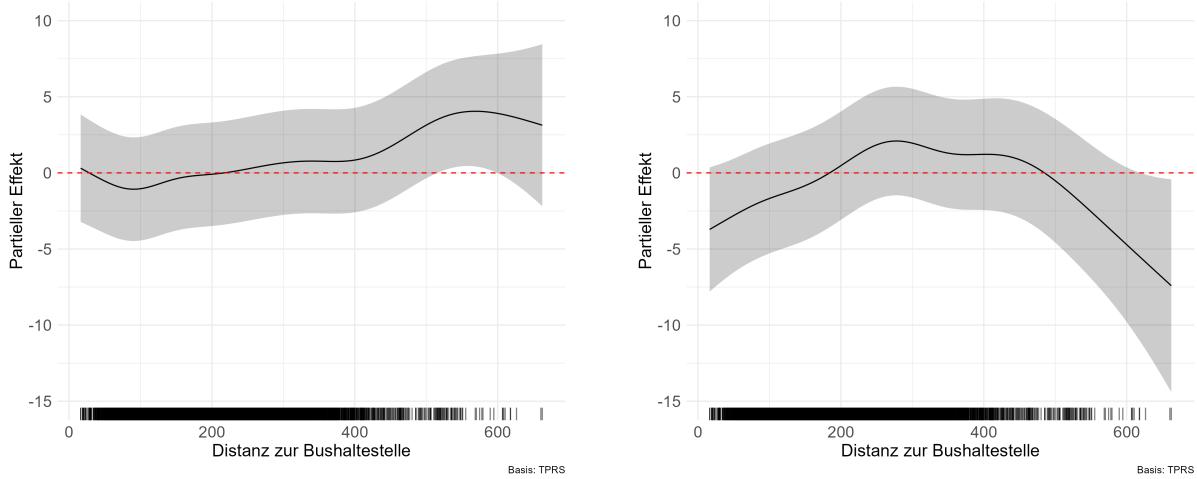
Abbildung 35: Partieller Effekt der Distanz zur U-Bahn auf die Wohnlage außerhalb

Im Gegensatz dazu ergibt sich für die nicht-zentralen Lagen in Abbildung 35 ein wesentlich eindeutigeres Bild.

Für die gute Lage (a) zeigt sich bei einer Distanz von bis zu ca. 500 Metern zunächst ein statistisch signifikanter, aber in seiner Stärke sehr geringer negativer Effekt auf die Chance für eine gute Lage außerhalb im Vergleich zu einer durchschnittlichen. Im Bereich von 500-3000 Metern steigt die Chance leicht und größtenteils signifikant. Insgesamt schwankt die geschätzte Kurve im Bereich 0-3000 Metern nahe um die Nulllinie, was bedeutet, dass eine Veränderung der U-Bahn-Distanz die Chance auf eine gute Lage (im Vergleich zu einer durchschnittlichen Lage) nur minimal beeinflusst. Allerdings reduziert sich für Distanzen ab ca. 3000 Metern die Chance auf eine gute Lage im Vergleich zu einer durchschnittlichen stark.

Bei dem partiellen Effektposten für die beste Lage (b) ist es zunächst wichtig, die Unterschiede in der Skalierung der Y-Achse zu beachten. Für Distanzen bis ca. 2500 Metern ergibt sich ein positiver Effekt für die Chance einer „besten Lage außerhalb“ im Vergleich zu einer durchschnittlichen. Jedoch schneidet auch hier das Konfidenzintervall größtenteils die Nulllinie, wodurch die statistische Relevanz dieses Einflusses durchaus fraglich ist. Jenseits dieser Schwelle wird eine zunehmende Entfernung zur U-Bahn jedoch zu einem sehr starken negativen Faktor. Die Skalierung der Y-Achse verdeutlicht, dass der negative Einfluss hier um ein Vielfaches stärker ist als bei der guten Lage.

Die U-Bahn-Anbindung ist somit vor allem als Abgrenzungsmerkmal für die besten Lagen essenziell. Obwohl sich in der deskriptiven Analyse auch für nicht-zentralen Lagen wenige Unterschiede im Bereich geringer Distanzen gezeigt haben, konnte das Modell die verschiedenen Wohnlagen trotzdem besser voneinander abgrenzen. Möglicherweise ist dies hier auf die höhere Stichprobengröße zurückzuführen.



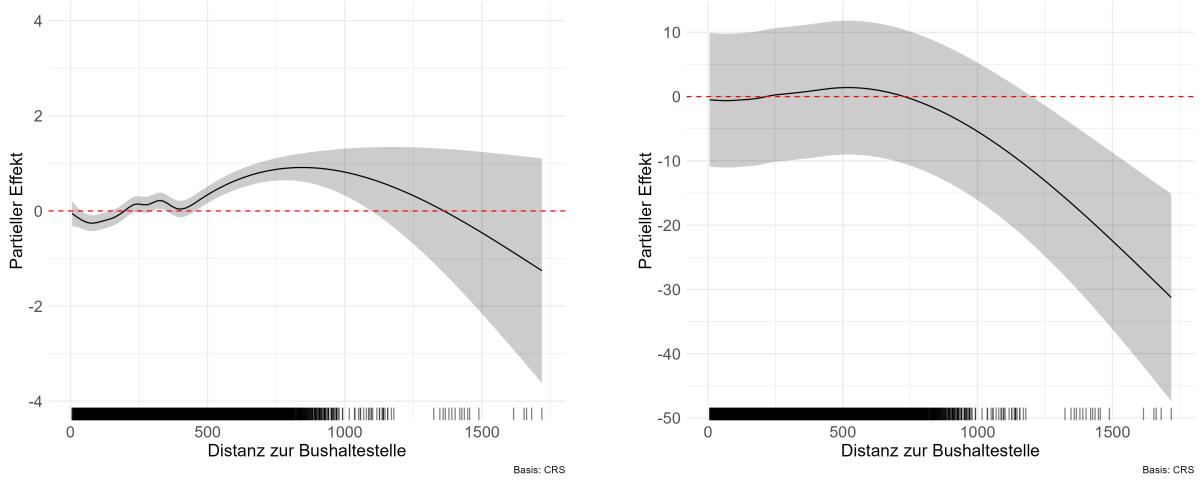
(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 36: Partieller Effekt der Distanz zur Bushaltestelle auf die zentrale Wohnlage

Die Abbildung 36 zeigt nun den partiellen Effekt der Distanz zur nächsten Bushaltestelle auf die zentralen Wohnlagen.

Fast über den gesamten Wertebereich ist der Einfluss der Bushaltestellen-Distanz statistisch nicht signifikant. Sowohl für die gute Lage (a) als auch für die beste Lage (b) schließt das graue Konfidenzintervall die rote Nulllinie mit ein. Obwohl die Kurven leichte, wellenförmige Tendenzen andeuten, sind diese Schwankungen statistisch nicht von einem Nulleffekt zu unterscheiden.

Dieses Ergebnis ist plausibel und wird durch die deskriptive Analyse gestützt. Ähnlich wie bei der U-Bahn-Distanz zeigten auch hier die Dichteverteilungen (vgl. Abbildung 18 und 19) für die durchschnittliche, gute und beste zentrale Lage eine fast identische Struktur. Wenn die Verteilung eines Merkmals zwischen den Kategorien kaum variiert, ist es für das statistische Modell schwierig, einen robusten und signifikanten Effekt zu identifizieren. Somit ist die Distanz zur nächsten Bushaltestelle kein entscheidendes Kriterium im zentralen Stadtgebiet zur Abgrenzung von guten oder besten Lagen von durchschnittlichen Lagen. Die gute Erreichbarkeit durch Haltestellen scheint hier eine generelle Eigenschaft der Stadt München zu sein und stellt für zentrale Lagen kein besonderes Qualitätsmerkmal dar.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

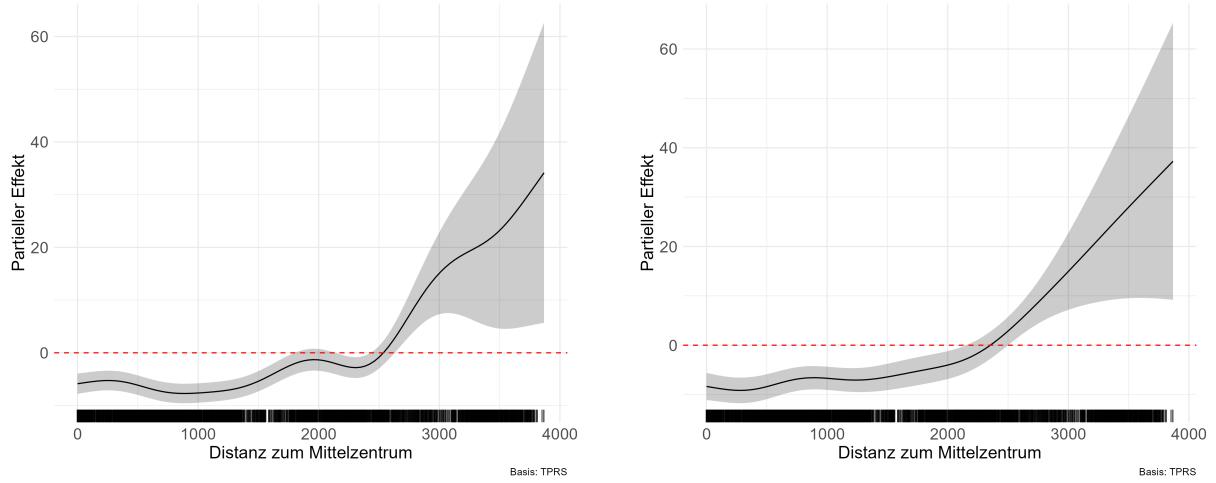
Abbildung 37: Partieller Effekt der Distanz zur Bushaltestelle auf die Wohnlage außerhalb

Für die nicht-zentralen Lagen ergibt sich in Abbildung 37 wiederum ein anderes Bild.

Zunächst ist für die gute Lage (a) bis ca. 250 Metern ein leicht negativer, geschätzter Effekt festzustellen. Eine unmittelbare Nähe zu einer Bushaltestelle senkt also die Chance auf eine gute Lage im Vergleich zu einer durchschnittlichen leicht. Im Gegensatz dazu erhöht sich die Chance für eine gute Lage im Intervall von ca. 250 bis 1370 Metern, wobei die Chance am größten für Distanzen um die 860 Meter ist. Ab einer Entfernung von ca. 1370 Metern ist wieder ein negativer Effekt festzustellen. Es ist jedoch zu beachten, dass ab ca. 1220 Metern Entfernung das Konfidenzband die Nulllinie beinhaltet und zunehmend breiter wird. Diese Unsicherheit entsteht vor allem durch die wenigen Ausprägungen in diesem Intervall, welche an den schwarzen Linien an der X-Achse der Grafik erkennbar sind.

Für die beste Lage (b) ist hingegen bis zu einer Entfernung von ca. 750 Metern kein Unterschied zu der durchschnittlichen Lage festzustellen, da die geschätzte Funktion nah an der Nulllinie liegt. Ab dieser Schwelle beginnt sich jedoch die Chance für die gute Lage für größere Entfernungen immer weiter zu erhöhen. Ab einer Distanz von ca. 1220 Metern ist der Effekt trotz weniger Beobachtungen sogar signifikant. Dies ist vor allem auf den langen Rand der Dichte der durchschnittlichen Lage außerhalb und den kurzen Rand der Dichte der besten Lage außerhalb zurückzuführen. Man beachte auch den Unterschied in der Skalierung der Y-Achse.

Im Gegensatz zum Zentrum, wo die Bushaltestellendichte hoch und ihr Einfluss gering ist, spielt die Erreichbarkeit in nicht-zentralen Lagen eine etwas wichtigere Rolle.



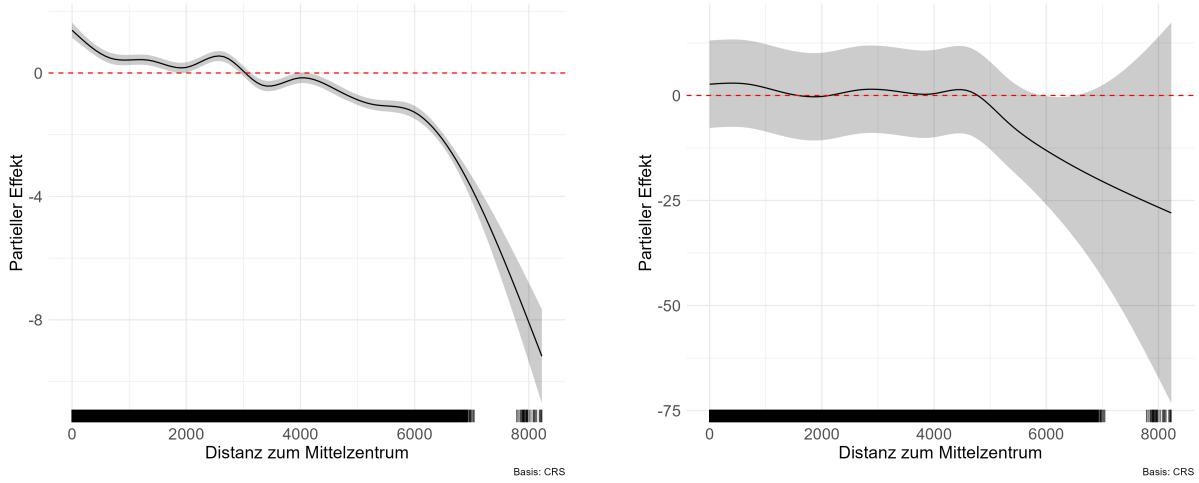
(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 38: Partieller Effekt der Distanz zum Mittelzentrum auf die zentrale Wohnlage

Die Abbildung 38 zeigt den partiellen Effekt der Distanz zum Mittelzentrum auf die zentralen Wohnlagen. Der Zusammenhang verläuft für die beiden Zielkategorien „gute zentrale Lage“ (a) und „beste zentrale Lage“ (b) sehr ähnlich.

Über einen weiten Bereich von 0 bis ca. 2.500 Metern ist der Effekt signifikant negativ. Dies deutet darauf hin, dass die unmittelbare Nähe zum Mittelzentrum die Chance für die gute und beste zentrale Lage im Vergleich zur zentralen durchschnittlichen Lage senkt. Jenseits dieser Schwelle kehrt sich der Zusammenhang jedoch um, wobei der Effekt mit zunehmender Entfernung stärker wird. Beim partiellen Effektplot der guten Lage ist etwas mehr Schwankung in der Glättungsfunktion zu beobachten.

Zusammenfassend deuten die Ergebnisse darauf hin, dass die begehrtesten Standorte nicht jene mit der geringsten, sondern mit einer moderaten Distanz zum Mittelzentrum sind.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 39: Partieller Effekt der Distanz zum Mittelzentrum auf die Wohnlage außerhalb

Für die nicht-zentralen Wohnlagen (Abbildung 39) zeigt sich ein fundamental anderer Zusammenhang als im Zentrum der Stadt.

Beim partiellen Effektpplot der guten Lage außerhalb (a) ist zu erkennen, dass die Glättungsfunktion für niedrige und moderate Entferungen bis zu ca. 2000 Metern im niedrigen positiven Bereich liegt. Folglich wird hier die Chance für eine gute Lage im Vergleich zu einer durchschnittlichen leicht erhöht eingeschätzt. Ab einer Entfernung von 4000 Metern wird dagegen die Distanz zum nächsten Mittelzentrum zunehmend zu einem negativen Faktor. Das besonders dünne Konfidenzband in diesem partiellen Effektpplot spricht zudem für eine hohe Sicherheit der Schätzung.

Diese scheint für die beste Lage außerhalb nicht vorhanden zu sein, da in Abbildung 39 (b) ein sehr breites Konfidenzintervall geschätzt wird. Für Distanzen bis zu ca. 4500 Metern wird die Chance für eine beste Lage außerhalb nicht signifikant verschieden von einer durchschnittlichen Lage außerhalb eingeschätzt. Ab dieser Schwelle fällt die Glättungskurve jedoch drastisch, wodurch ein starker negativer Effekt für die beste Lage geschätzt wird. Man beachte allerdings erneut das breite 95%-Konfidenzband, welches die Nulllinie beinhaltet.

Zusammenfassend ist festzustellen, dass im Gegensatz zu den zentralen Lagen, wo eine gewisse Distanz zum Zentrum vorteilhaft ist, für nicht-zentrale Lagen die Anbindung an ein Mittelzentrum als deutlich wichtiger eingestuft wird.

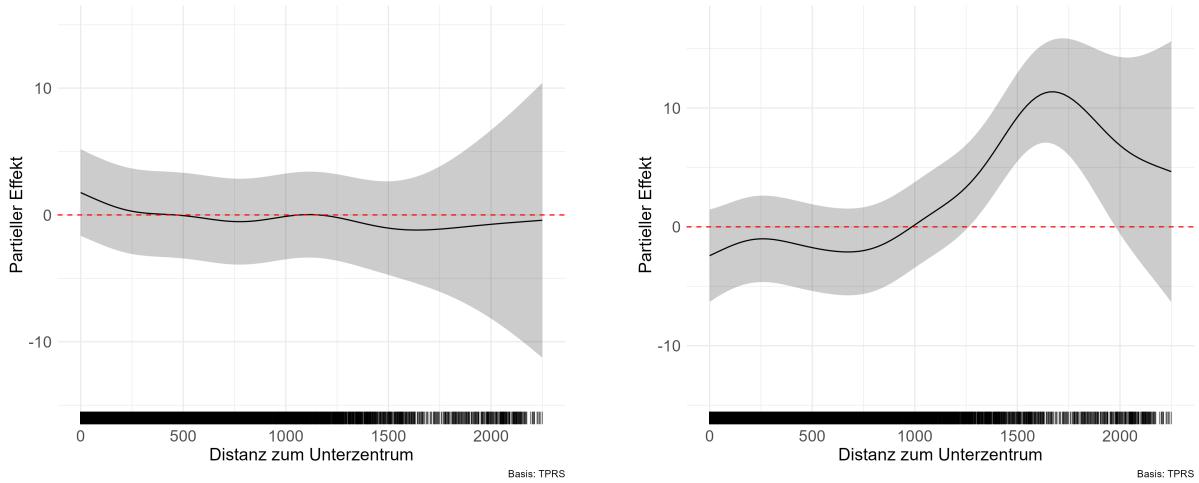
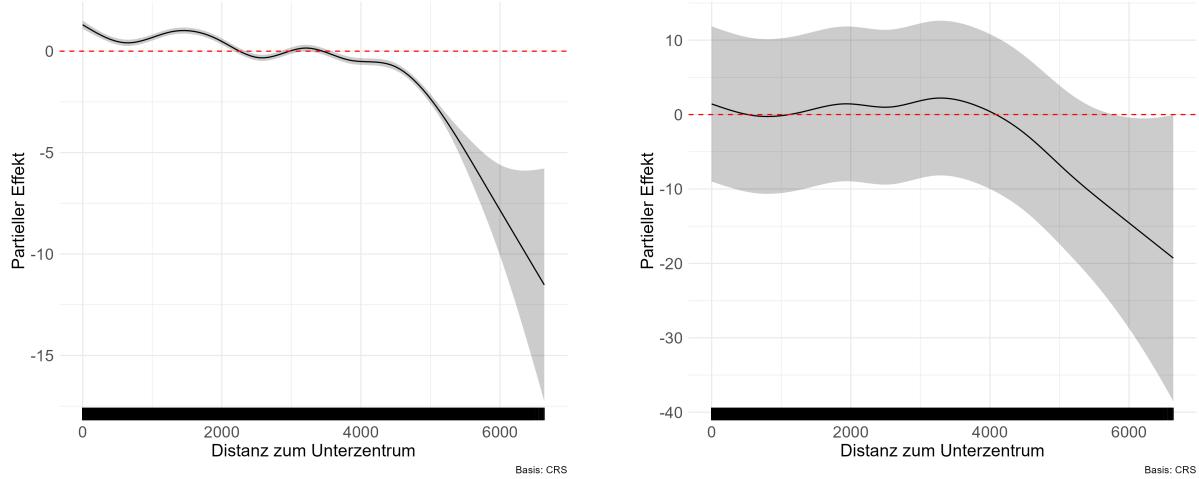


Abbildung 40: Partieller Effekt der Distanz zum Unterzentrum auf die zentrale Wohnlage

Abbildung 40 zeigt den partiellen Effekt der Distanz zum nächstgelegenen Unterzentrum auf die zentralen Wohnlagen.

Für die gute Lage (a) verläuft die Effektkurve leicht wellenförmig um die Nulllinie, wobei das breite Konfidenzband durchgehend den Wert Null einschließt. Dies deutet darauf hin, dass die Distanz zum Unterzentrum keinen statistisch belastbaren Einfluss darauf hat, ob eine Lage als gut oder durchschnittlich eingestuft wird.

Ein etwas anderes Muster zeigt sich für die beste Lage (b). Hier ist der Effekt bei sehr kurzen Distanzen zunächst negativ, wobei auch hier das Konfidenzband um die Glättungsfunktion die Nulllinie einschließt. Bei Entfernungen von mehr als 1000 Metern steigt die Chance für eine gute Lage im Vergleich zu einer durchschnittlichen. Man beachte jedoch die erhöhte Unsicherheit für die weitesten Distanzen.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 41: Partieller Effekt der Distanz zum Unterzentrum auf die Wohnlage außerhalb

Abschließend wird der partielle Effekt der Distanz zum nächsten Unterzentrum der nicht-zentralen Wohnlagen in Abbildung 41 betrachtet.

Für die gute Lage (a) ist ein komplexer und in mehreren Bereichen signifikanter Zusammenhang zu erkennen. Bei einer Entfernung von bis zu ca. 2100 Metern wird ein leicht positiver Effekt auf die Chance der guten Lage im Vergleich zur durchschnittlichen Lage geschätzt. Im Intervall bis ca. 3700 Metern schwankt die Kurve leicht um die Nulllinie. Das besonders schmale Konfidenzband in diesem Bereich deutet auf eine hohe Schätzgenauigkeit hin. Ab der Schwelle von 3700 Metern wird ein mit höheren Entfernungen zunehmender negativer Effekt auf die Chance der guten Lage geschätzt. Allerdings erhöht sich für besonders weite Distanzen die Schätzungsgenauigkeit.

Ein anderes Muster zeigt sich für die beste Lage (b). Hier ist der Effekt für fast alle Distanzen nicht signifikant. Die Glättungsfunktion schwankt bei einer Distanz von bis zu ca. 4000 Metern um die Nulllinie, was bedeutet, dass kein relevanter Unterschied zwischen der besten und durchschnittlichen Lage bei dieser Variablen festgestellt werden kann. Ab dieser Schwelle wird dagegen ein zunehmend stark negativer Effekt geschätzt, was auf eine Obergrenze der tolerierbaren Distanz hindeutet. Die Unsicherheit stammt hier möglicherweise auch aus der vergleichbar geringen Stichprobengröße der besten Lage außerhalb.

5.2.2 Interpretation der Effekte der Indexvariablen

In diesem Abschnitt werden nun die partiellen Effekte der drei Indexvariablen interpretiert. Die Interpretation erfolgt dabei analog zu Kapitel 5.2.1. Es wird somit der Effekt nach wie vor relativ zur Referenz „durchschnittliche Lage“ und unter Konstanthaltung der restlichen Größen geschätzt.

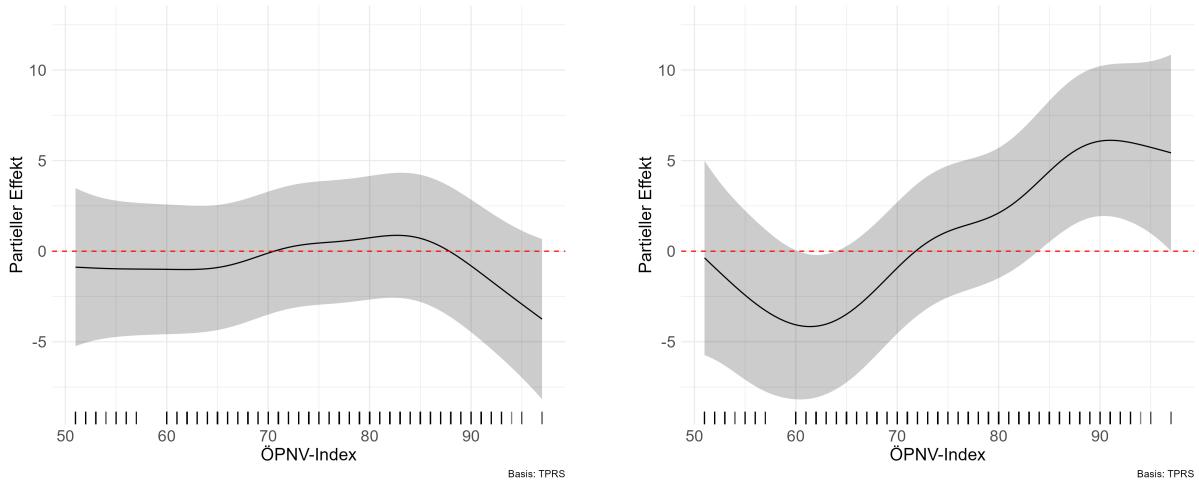


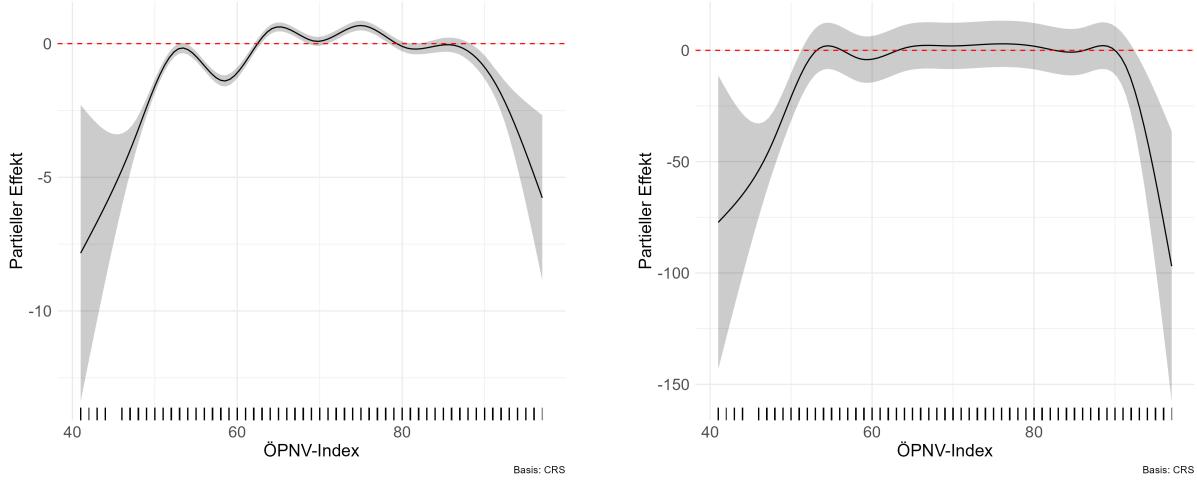
Abbildung 42: Partieller Effekt des ÖPNV-Index auf die zentrale Wohnlage

Abbildung 42 visualisiert zunächst den partiellen Effekt des ÖPNV-Index auf die zentralen Wohnlagen.

Für die gute Lage (a) verläuft die Effektkurve über fast den gesamten Wertebereich sehr nah an der Nulllinie. Lediglich für besonders gute Anbindungen ab einem Indexwert von ca. 90 wird ein negativer Effekt auf die Chance für die gute Lage geschätzt. Man beachte zudem die durch das Konfidenzintervall gezeigte erhöhte Unsicherheit. Insgesamt scheint der Unterschied zwischen den ÖPNV-Anbindungen der guten und durchschnittlichen zentralen Lagen eher gering zu sein.

Ein anderes Bild zeigt sich für die beste Lage (b). Bei einem vergleichbar niedrigen ÖPNV-Index (ca. < 70) reduziert sich die Chance auf eine beste Lage im Vergleich zu einer durchschnittlichen. Ab diesem Schwellenwert kehrt sich der Zusammenhang um, somit wirkt sich dann eine gute ÖPNV-Anbindung positiv auf die Chance für eine gute Lage aus. Man beachte allerdings auch hier die erhöhte Unsicherheit über diesen Effekt.

Generell ist festzustellen, dass keine zentral liegende Lokation eine schlechte ÖPNV-Anbindung hat, da der niedrigste Indexwert größer als 50 ist und der mögliche Wertebereich des ÖPNV-Index [0, 100] beträgt. Die Unterschiede zwischen den verschiedenen zentralen Wohnlagen scheinen klein zu sein, weshalb das Modell den Zusammenhang nur mit erhöhter Unsicherheit schätzen kann.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

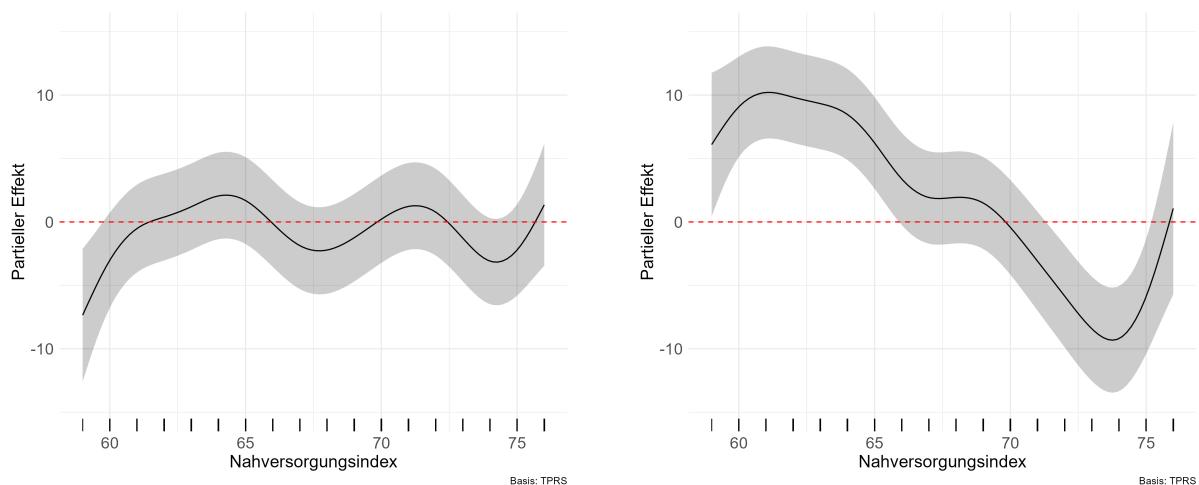
Abbildung 43: Partieller Effekt des ÖPNV-Index auf die Wohnlage außerhalb

Für die nicht-zentralen Lagen (Abbildung 43) zeigt der ÖPNV-Index dagegen ein gänzlich anderes Muster als im Zentrum.

Für die Kategorie „gute Lage außerhalb“ lässt sich erkennen, dass ein niedriger ÖPNV-Index die Chance einer entsprechenden Einstufung deutlich verringert. Mit zunehmender Erreichbarkeit steigt der Effekt jedoch an und liegt im mittleren Bereich des Index überwiegend über null, was auf eine höhere Wahrscheinlichkeit für die Klassifikation als „gute Lage“ hindeutet. Ab sehr hohen Indexwerten von ca. 85 flacht der Effekt hingegen wieder ab, was bedeutet, dass sich die Chance für eine gute Lage statt einer durchschnittlichen reduziert.

Ein sehr ähnlicher Zusammenhang ist ebenfalls für die beste Lage (b) zu beobachten. Allerdings werden die negativen Effekte für sehr hohe und sehr niedrige Indexwerte als sehr stark geschätzt. Im mittleren Bereich liegt die Glättungsfunktion nah an der Nulllinie, was bedeutet, dass in diesem Intervall kein signifikanter Unterschied zwischen der durchschnittlichen und besten Lage außerhalb festgestellt werden kann.

Insgesamt scheint der ÖPNV-Index für die nicht-zentralen Wohnlagen ein besserer Prädiktor zu sein.



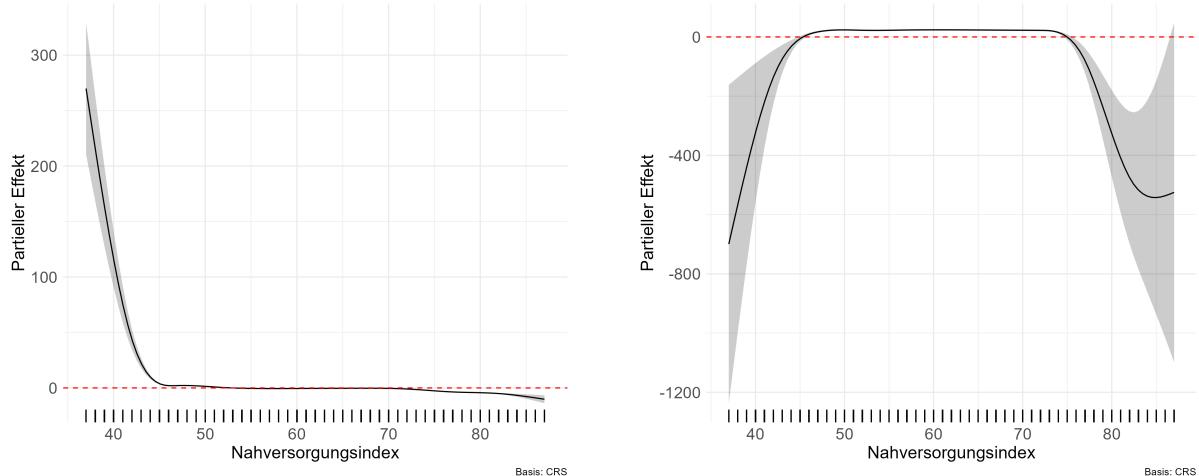
(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 44: Partieller Effekt des Nahversorgungsindex auf die zentrale Wohnlage

Nun werden die partiellen Effektplots des Nahversorgungsindex auf die zentralen Wohnlagen betrachtet (Abbildung 44).

Für die Kategorie gute zentrale Lage zeigt sich zunächst ein negativer Effekt bei vergleichsweise niedrigen Indexwerten, sodass eine geringe Versorgung die Chance einer entsprechenden Einstufung reduziert. Mit zunehmender Nahversorgung kehrt sich der Effekt jedoch um und liegt im unteren mittleren Bereich des Index über Null, was auf eine erhöhte Chance für die Klassifikation als „gute Lage“ hinweist. Auffällig ist, dass der Zusammenhang nicht linear verläuft, sondern wellenförmige Schwankungen um Null herum aufweist.

Für die Kategorie beste zentrale Lage zeigt sich ein anderes Muster. Hier ist der Effekt bei vergleichbar niedrigen bis mittleren Werten des Nahversorgungsindex zunächst stark positiv, sodass eine gute, aber noch nicht übermäßig ausgeprägte Nahversorgung die Chance einer Einstufung als beste Lage deutlich erhöht. Ab einem Indexwert von etwa 60 nimmt der Effekt jedoch sukzessive ab und wird ab 70 schließlich negativ, sodass eine sehr hohe Dichte an Nahversorgungseinrichtungen die Chance für beste Lagen reduziert. Dieses Ergebnis lässt darauf schließen, dass beste Lagen eher durch eine ausgewogene, aber nicht übermäßige Nahversorgung gekennzeichnet sind, während eine zu hohe Versorgungsdichte möglicherweise mit stärker verdichteten und damit weniger exklusiven Wohnumfeldern einhergeht.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

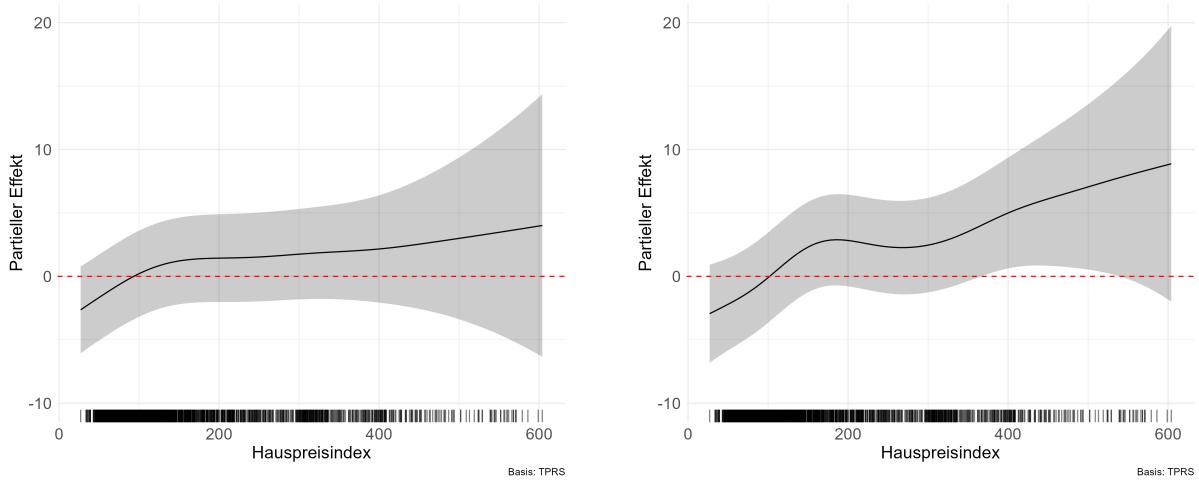
Abbildung 45: Partieller Effekt des Nahversorgungsindex auf die Wohnlage außerhalb

Die partiellen Effekte des Nahversorgungsindex auf die Wohnlagen außerhalb, erkennbar in Abbildung 45, zeigen ein stark ausgeprägtes Muster, das sich deutlich von den zuvor betrachteten zentralen Lagen unterscheidet.

Für die gute Lage im Vergleich zur Referenz durchschnittlich ist der Effekt bei sehr niedrigen Werten des Nahversorgungsindex extrem positiv. Dies bedeutet, dass eine sehr geringe Nahversorgung die Chance einer Einstufung als gute Lage (statt durchschnittliche Lage) massiv erhöht. Bereits ab einem Indexwert von etwa 45 fällt der Effekt jedoch abrupt ab und nähert sich null, sodass eine bessere Nahversorgung außerhalb kaum mehr mit einer guten Lage assoziiert ist. Ab einem Indexwert von ca. 75 wird ein signifikant negativer Effekt geschätzt.

Für die Kategorie beste Lage zeigt sich ein umgekehrtes Bild. Bei vergleichsweise niedrigen Werten des Nahversorgungsindex (< 45) ist der Effekt sehr stark negativ, sodass die Chance, als beste Lage eingestuft zu werden, hier äußerst gering ist. Ab dieser Schwelle verläuft die Glättungsfunktion über der Nulllinie. Im Intervall der mittleren Indexwerte wird die Chance für die beste Lage als erhöht eingeschätzt. Ab der Schwelle 75 beginnt wieder ein zunehmender, stark negativer Effekt. Man beachte hier allerdings die erhöhte Unsicherheit.

Zusammenfassend lässt sich sagen, dass der Nahversorgungsindex eine geeignete Variable ist, um die nicht-zentralen guten und besten Lagen von den durchschnittlichen abzugrenzen.



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

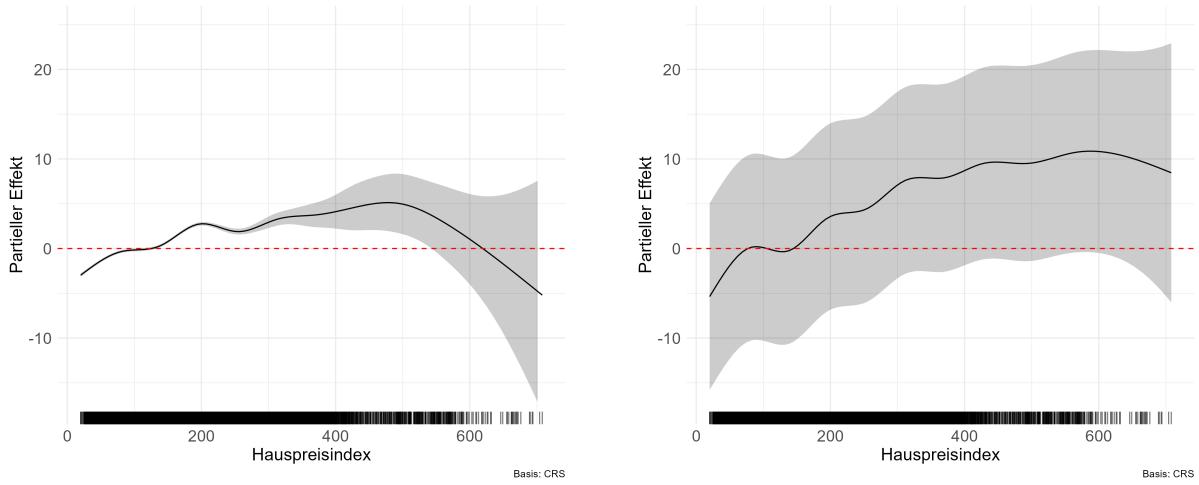
Abbildung 46: Partieller Effekt des Hauspreisindex auf die zentrale Wohnlage

Abbildung 46 zeigt den partiellen Effekt des Hauspreisindex auf die zentralen Wohnlagen.

Aufgrund der großen Unterschiede bei den Dichteverteilungen der besten und guten Lage zur durchschnittlichen Lage (vgl. Abbildung 27) wäre für diese Variable ein eindeutiger hauptsächlich positiver Effekt zu erwarten. Allerdings ist dies hier nur bedingt der Fall.

Man kann gut erkennen, dass sowohl für die gute als auch die beste Lage bis genau 100 ein negativer Effekt geschätzt wird. Diese Schwelle ergibt sich unter anderem aus der Definition des Hauspreisindex, bei dem ein Wert von 100 genau dem Kreisdurchschnitt entspricht. Hat ein Wohnobjekt einen Indexwert von kleiner als 100, bedeutet dies somit, dass es im Verhältnis zu seiner Umgebung günstiger ist. Ab dieser Schwelle verläuft die Glättungsfunktion über der Nulllinie, folglich steigt nun die Chance auf eine gute bzw. beste Lage statt einer durchschnittlichen. Der Effekt ist dabei für die zentrale beste Lage etwas stärker.

Es wird allerdings hier für die beiden gesamten Funktionen eine erhöhte Unsicherheit geschätzt. Dies ist vermutlich auf die relativ geringe Anzahl an zentralen Lagen mit besonders hohen Indexwerten und den eher geringen Unterschieden zwischen den Wohnlagen für niedrige Indexwerte zurückzuführen (vgl. Abbildung 27).



(a) Effekt auf „gute Lage“ (gegen durchschnittlich) (b) Effekt auf „beste Lage“ (gegen durchschnittlich)

Abbildung 47: Partieller Effekt des Hauspreisindex auf die Wohnlage außerhalb

Abschließend wird der partielle Effekt des Hauspreisindex auf die nicht-zentralen Wohnlagen betrachtet. Auch hier wäre intuitiv ein eindeutiger Zusammenhang aufgrund von Abbildung 28 zu erwarten.

Ähnlich wie bei den zentralen Lagen findet sich auch hier der Schwellenwert von 100 wieder. Vor diesem Wert wird für die gute und beste Lage erneut ein negativer Effekt geschätzt.

In dem Intervall von 100 bis ca. 620 kehrt sich dieser Zusammenhang für die guten Lagen (a) erneut um. Allerdings nimmt die anfänglich sehr niedrige Unsicherheit mit größeren Indexwerten immer weiter zu. Für sehr hohe Indexwerte wird erneut ein zunehmender negativer Effekt auf die Chance für die gute Lage geschätzt.

Für die besten Lagen wird dagegen ab dem Schwellenwert von 100 ein durchweg positiver Effekt geschätzt. Man beachte allerdings hier die hohe Unsicherheit.

5.2.3 Überblick über die nicht-linearen Effekte

Um die adäquate Wahl und die Eigenschaften der zwei geschätzten Generalisierten Additiven Modelle zu überprüfen, fasst Tabelle 3 die zentralen Kenngrößen der Smooth-Terme zusammen. Diese diagnostische Übersicht dient zwei Hauptzielen: Erstens, die Notwendigkeit des flexiblen GAM-Ansatzes gegenüber einem starren linearen Modell zu validieren, und zweitens, die Komplexität und die relative Wichtigkeit der einzelnen Prädiktoren zwischen den Modellen für zentrale und nicht-zentrale Lagen zu vergleichen.

Variable	Zentrale Lagen		Lagen außerhalb	
	EDF	Chi-Quadrat	EDF	Chi-Quadrat
distanz_bahnhof_gut	7.677	105.07	9.689	784.4
distanz_bahnhof_beste	7.735	91.48	9.846	1088.5
distanz_ubahn_gut	7.229	56.50	9.435	564.0
distanz_ubahn_beste	4.122	29.95	9.812	490.3
distanz_bushaltestelle_gut	7.006	78.34	7.999	176.1
distanz_bushaltestelle_beste	5.772	68.22	7.010	320.9
distanz_mittelzentrum_gut	8.348	215.50	9.855	2189.8
distanz_mittelzentrum_beste	5.937	75.15	8.281	1366.8
distanz_unterzentrum_gut	5.077	50.32	8.733	2680.5
distanz_unterzentrum_beste	7.660	142.13	7.470	901.8
opnv_index_gut	5.007	43.93	8.934	1349.3
opnv_index_beste	5.757	54.40	8.287	903.0
nahversorgungs_index_gut	8.473	342.62	9.875	6053.5
nahversorgungs_index_beste	8.223	186.64	7.573	663.5
hauspreis_index_gut	3.962	138.87	7.430	7734.3
hauspreis_index_beste	5.499	146.07	9.035	3904.2

Tabelle 3: Vergleich der effektiven Freiheitsgrade und Chi-Quadrat-Werte für die Smooth-Terme der beiden GAMs

Die Tabelle listet für jede metrische Variable und jede der beiden Zielkategorien (*_gut* und *_beste*) zwei Kennzahlen auf. Die effektiven Freiheitsgrade (EDF) dienen als Maß für die Nicht-Linearität der geschätzten Zusammenhänge. Ein minimal möglicher Wert von 1 würde bedeuten, dass ein rein linearer Zusammenhang vorliegt (vgl. Clark 2022). Die Chi-Quadrat-Werte können als ein Indikator für die statistische Effektstärke interpretiert werden, wobei ein höherer Wert auf einen stärkeren Einfluss der Variable auf die Wohnlage hindeutet (vgl. Wood 2017b, S.303). Diese beiden Werte werden jeweils für die Modelle der zentralen und nicht-zentralen Lagen direkt gegenübergestellt, um Unterschiede in den Wirkungsweisen der Prädiktoren aufzuzeigen.

Aus der Tabelle lassen sich folgende wesentliche Schlüsse ziehen:

- 1. Die Wahl des GAM-Ansatzes ist gerechtfertigt:** Es ist ersichtlich, dass alle EDF-Werte deutlich über 1 liegen. Dies belegt für sämtliche Variablen in beiden Modellen, dass ihre Zusammenhänge mit der Wohnlage hochgradig nicht-linear sind. Die Entscheidung für einen flexiblen GAM-Ansatz wird durch diese Ergebnisse somit klar validiert. Ein einfaches lineares Modell hätte diese komplexen Muster nicht abbilden können.
- 2. Unterschiedliche Treiber der Wohnlagenqualität:** Bei der Betrachtung der

Effektstärken (Chi-Quadrat-Werte) zeigen sich klare Unterschiede zwischen den Makrolagen. In den zentralen Lagen sind vor allem der Nahversorgungsindex (Chi-Quadrat = 342.62) und die Distanz zum Mittelzentrum (Chi-Quadrat = 215.50) die dominantesten Prädiktoren für die Abgrenzung einer guten Lage von einer durchschnittlichen. In den Lagen außerhalb ist der Einfluss des Hauspreisindex (Chi-Quadrat = 7734.3) und des Nahversorgungsindex (Chi-Quadrat = 6053.5) um ein Vielfaches stärker, was auf eine andere Priorisierung von Standortfaktoren in nicht-zentralen Gebieten hindeutet.

- 3. Höhere Komplexität außerhalb des Zentrums:** Die Zusammenhänge in den nicht-zentralen Lagen sind tendenziell noch komplexer als im Zentrum, was anhand der durchgehend höheren EDF-Werte erkennbar ist.

5.2.4 Bewertung der Konkurvität

Von Konkurvität spricht man, wenn ein geglätteter Term in einem Modell durch einen oder mehrere der anderen geglätteten Terme angenähert werden kann. Sie kann als eine Verallgemeinerung der Kollinearität betrachtet werden und führt zu ähnlichen Interpretationsproblemen. Dabei kann aufgrund der Ähnlichkeit nicht mehr eindeutig bestimmt werden, welche der betroffenen Variablen ein Effekt zugeordnet werden soll (vgl. Wood 2017b, S. 29f).

Im Folgenden soll zunächst die Korrelationsstruktur der metrischen Kovariablen miteinander verglichen werden, um einen ersten Eindruck über potenziell problematische Größen zu gewinnen. Dafür wird der Pearson-Korrelationskoeffizient zwischen den verschiedenen metrischen Kovariablen berechnet und in einer Korrelations-Heatmap dargestellt. Der Koeffizient gibt die Stärke des linearen Zusammenhangs an und besitzt einen Wertebereich von $[-1, 1]$ (vgl. Fahrmeier et al. 2016, S.128).

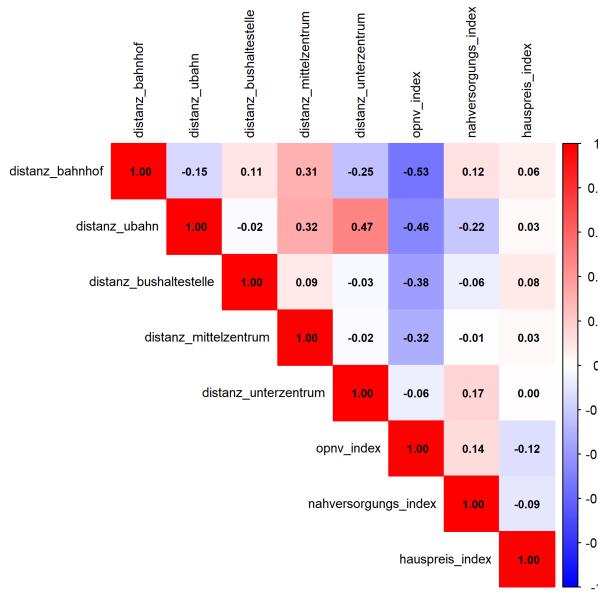


Abbildung 48: Korrelation der metrischen Variablen im Modell der Lagen außerhalb

Abbildung 48 zeigt zunächst die Korrelations-Heatmap der metrischen Kovariablen aus dem Datensatz der nicht-zentralen Lagen. Dort sind keine auffälligen Werte festzustellen.

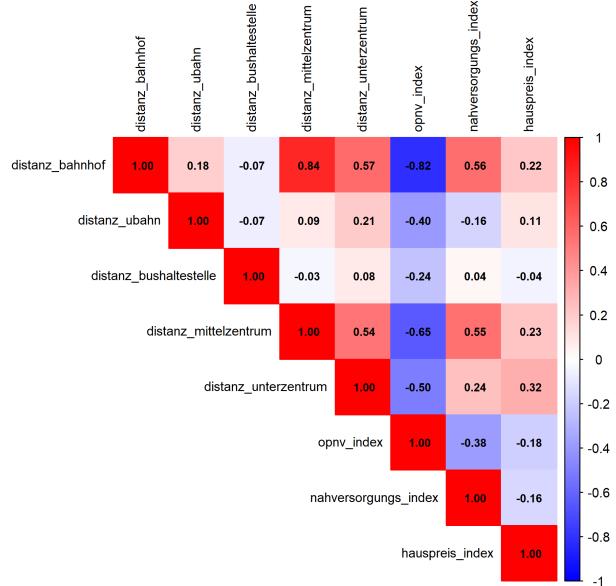


Abbildung 49: Korrelation der metrischen Variablen im Modell der zentralen Lagen

Dagegen sind in Abbildung 49 in der Korrelations-Heatmap für die zentralen Lagen gleich mehrere problematische Werte erkennbar. Die Variable `distanz_bahnhof` erweist sich hier als die essentielle problematische Größe.

Zum einen besteht eine sehr hohe negative Korrelation mit dem ÖPNV-Index ($r = -0.82$). Zum anderen korreliert sie sehr stark positiv mit der Distanz zum Mittelzentrum ($r = 0.84$).

Zur weiteren Überprüfung wurde zudem eine Konkurvitätsanalyse durchgeführt, deren zentrale Ergebnisse in folgender Tabelle zu sehen sind.

Geglätteter Term	Konkurvität mit $s(distanz_bahnhof)$
$s(opnv_index)$	0.875
$s(distanz_mittelzentrum)$	0.804
$s(distanz_unterzentrum)$	0.535
$s(nahversorgungs_index)$	0.534
$s(hauspreis_index)$	0.144
$s(distanz_ubahn)$	0.133
$s(distanz_bushaltestelle)$	0.090

Tabelle 4: Konkurvitätswerte der Smooth-Terme in Bezug auf die Variable `distanz_bahnhof` im Modell der zentralen Lagen

Auch hier hat sich die Distanz zum Bahnhof als problematische Größe erwiesen. Sowohl der ÖPNV-Index (0,875) als auch die Distanz zum Mittelzentrum (0,804) haben eine starke Abhängigkeit zur Bahnhofsdistanz. Dies belegt, dass die Information der Bahnhofsdistanz bereits in hohem Maße in diesen beiden anderen Prädiktoren enthalten ist.

Folglich stellt sich die Frage, ob die bisherigen im Modell geschätzten Effekte möglicherweise verzerrt sind. Um diese Hypothese zu überprüfen, wird das GAM für die zentralen Lagen ohne die Größe `distanz_bahnhof` nochmals neu geschätzt und mit dem vollständigen Modell verglichen.

5.2.5 Alternatives Modell der zentralen Wohnlagen

Zur Überprüfung einer möglichen Verzerrung werden im Folgenden einige Grafiken zur Interpretation der Effekte im vollständigen zentralen Modell und im zentralen Modell ohne `distanz_bahnhof` miteinander verglichen.

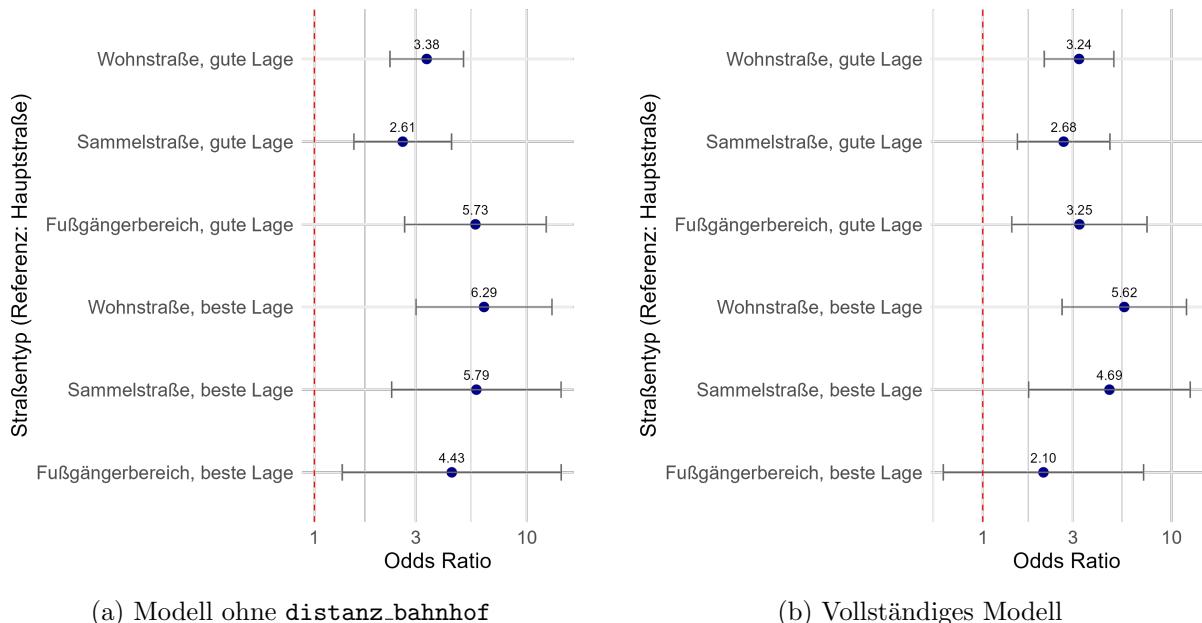


Abbildung 50: Vergleich der Odds-Ratios der beiden zentralen Modelle

In Abbildung 50 ist zunächst der Odds-Ratio von der Variable `straßentyp` der beiden zentralen Modelle im Vergleich dargestellt. Es ist festzustellen, dass sich kein Effekt drastisch verändert hat oder sich vom Positiven ins Negative gewandelt hat. Lediglich die Odds-Ratios von der Kategorie „Fußgängerbereich“ wurden im vorherigen Modell niedriger geschätzt.

Auch bei den Effekten der meisten metrischen Variablen konnten kaum Unterschiede zwischen den beiden Modellen festgestellt werden. Lediglich bei folgenden partiellen Effektplots ist ein kleiner Unterschied zu erkennen. Die Interpretation erfolgt analog zu den Kapiteln 5.2.1 und 5.2.2.

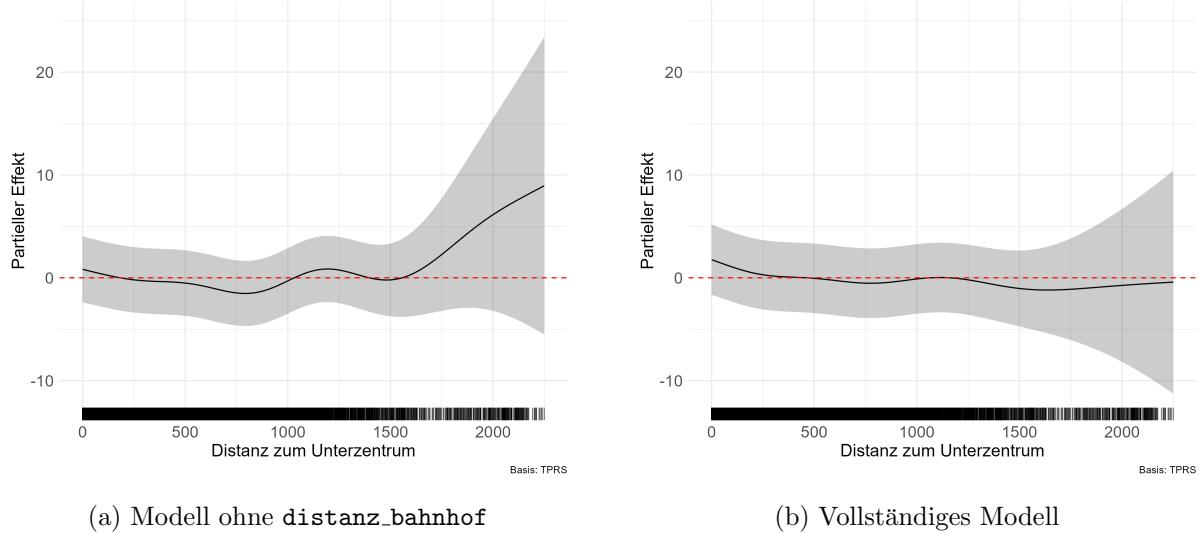


Abbildung 51: Vergleich der partiellen Effekte von `distanz_unterzentrum` auf die zentrale gute Lage

In Abbildung 51 ist erkennbar, dass im Modell ohne `distanz_bahnhof` (b) ab einer Distanz von über 1500 Metern zum nächsten Unterzentrum eine erhöhte und mit höheren Entfernung steigende Chance für die zentrale gute Lage im Vergleich zur zentralen durchschnittlichen Lage geschätzt wird. Beim vollständigen Modell ist in diesem Intervall ein leicht negativer Effekt geschätzt worden. Es ist jedoch zu beachten, dass beide dieser Effekte nicht signifikant sind, da die Konfidenzbänder die Nulllinie einschließen.

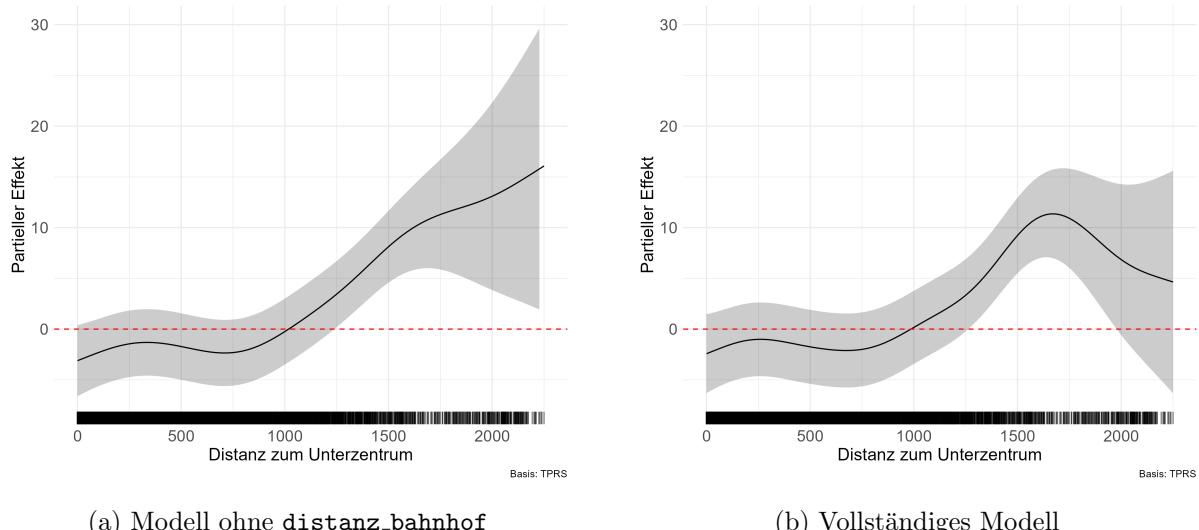
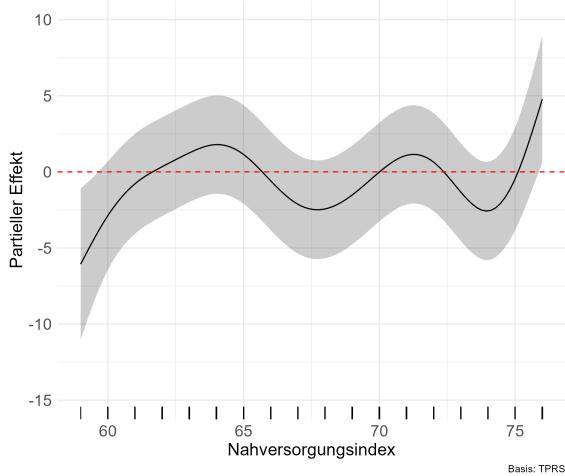
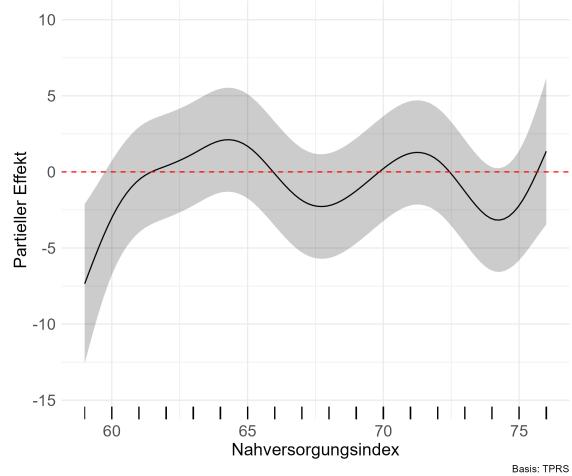


Abbildung 52: Vergleich der partiellen Effekte von `distanz_unterzentrum` auf die zentrale beste Lage

Auch die Glättungsfunktion für die beste Lage hat sich für größere Distanzen zum Unterzentrum etwas geändert. In Abbildung 52 (a) ist zu beobachten, dass ab einer Distanz von ca.

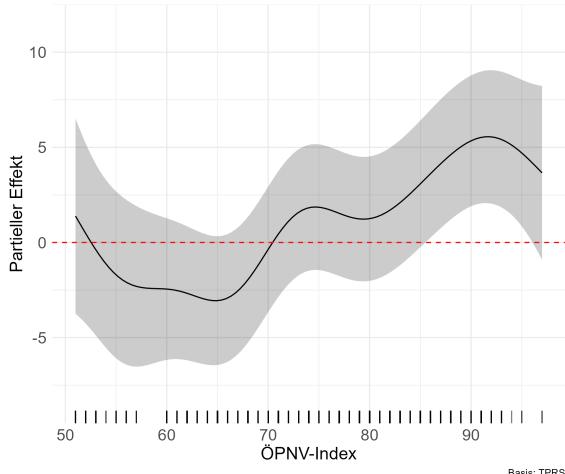
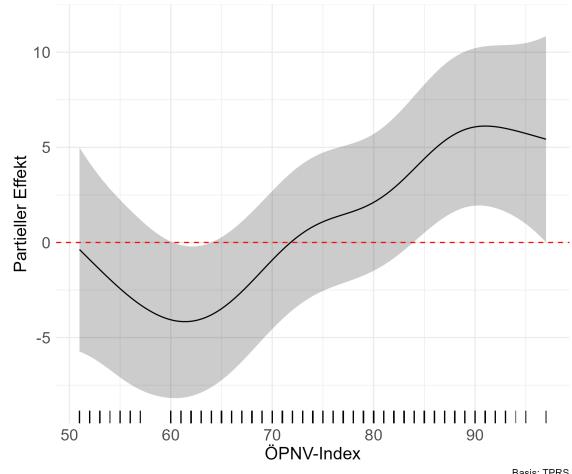
1000 Metern die Chance auf eine zentrale beste Lage steigt und für größere Entfernungen immer höher wird. Im vollständigen Modell wird dieser positive Effekt ab ca. 1630 Metern wieder schwächer.

(a) Modell ohne `distanz_bahnhof`

(b) Vollständiges Modell

Abbildung 53: Vergleich der partiellen Effekte von `nahversorgungs_index` auf die zentrale gute Lage

In Abbildung 53 werden die partiellen Effekte auf die gute zentrale Lage der Variable `nahversorgungs_index` verglichen. Dort sind die Unterschiede zwischen den Modellen ebenfalls sehr gering. Es wird lediglich im Modell ohne `distanz_bahnhof` für sehr hohe Nahversorgungsindexwerte ab 75 ein postiver Effekt auf die Chance für die gute Lage statt durchschnittliche Lage geschätzt. Im vollständigen Modell war dieser Effekt zum Großteil negativ.

(a) Modell ohne `distanz_bahnhof`

(b) Vollständiges Modell

Abbildung 54: Vergleich der partiellen Effekte von `opnv_index` auf die zentrale beste Lage

Abschließend wird in Abbildung 54 der partielle Effekt auf die zentral beste Lage des ÖPNV-Index verglichen. Dabei ist der Verlauf und Trend der Schätzfunktionen bei beiden Modellen fast identisch, allerdings schwankt die Kurve bei dem Modell ohne `distanz_bahnhof` etwas mehr und schätzt die Effekte teilweise etwas schwächer ein.

Insgesamt lässt sich sagen, dass sich die Einflüsse und Effekte der Kovariablen nicht stark verändert haben. Es ist somit kaum möglich, von einer schwerwiegenden Verzerrung durch `distanz_bahnhof` zu sprechen. Selbst die Effekte der Variablen, die eine hohe Korrelation aufwiesen, sind in etwa gleich geblieben.

Aufgrund der kleinen Unterschiede und der besseren Vergleichbarkeit zum Modell für die nicht-zentralen Lagen wird in der weiteren Analyse mit dem vollständigen Modell weitergearbeitet.

6 Analyse der Fehlklassifikationen

Neben der Interpretation der einzelnen Effekte ist die Bewertung der prädiktiven Güte des Modells ein zentraler Bestandteil der Analyse. Dieser Schritt ist entscheidend, um zu beurteilen, wie gut das Modell die tatsächlichen Wohnlagen in der Praxis klassifizieren kann und wo seine Stärken und Schwächen liegen. In diesem Kapitel wird daher die Vorhersageleistung der Modelle anhand von Konfusionsmatrizen und davon abgeleiteten Kennzahlen evaluiert. Darüber hinaus werden die räumlichen Muster der Fehlklassifikationen untersucht, um systematische Fehlerquellen zu identifizieren.

6.1 Modellevaluierung

Zunächst werden die GAMs für die zentralen und nicht-zentralen Lagen auf die beiden jeweiligen Trainingsdatensätze dieser Modelle angewandt.

Zur Bewertung der Güte dieser Vorhersagen dienen die in Tabelle 5 und 6 dargestellten Konfusionsmatrizen sowie die davon abgeleiteten Gütekennzahlen. Die Konfusionsmatrix stellt die vom Modell vorhergesagten Wohnlagen den tatsächlichen, wahren Wohnlagen gegenüber. Die Werte auf der Hauptdiagonalen zeigen die Anzahl der korrekten Klassifikationen, während die Werte außerhalb der Diagonalen die Fehlklassifikationen darstellen. Darunter werden zwei wichtige, klassenspezifische Gütemaße aufgeführt:

- **Sensitivität:** Gibt an, wie gut das Modell eine bestimmte Klasse erkennt. Eine hohe Sensitivität für die Klasse „beste Lage“ bedeutet beispielsweise, dass ein Großteil der tatsächlich besten Lagen auch als solche vorhergesagt wird. Der Wertebereich der Sensitivität beträgt $[0, 1]$ (vgl. James et al. 2021, S. 149).
- **Spezifität:** Gibt an, wie gut das Modell die anderen Klassen von einer bestimmten Klasse unterscheiden kann. Eine hohe Spezifität für „beste Lage“ bedeutet, dass Objekte, die nicht die beste Lage haben, vom Modell auch nur selten fälschlicherweise als „beste Lage“ klassifiziert werden. Der Wertebereich der Spezifität beträgt $[0, 1]$. (vgl. James et al. 2021, S. 149).

Ebenfalls wird die Accuracy, welche das Verhältnis der korrekt klassifizierten Beobachtungen im Verhältnis zu allen Beobachtungen darstellt, mit angegeben.

		Wahre Wohnlage		
Vorhergesagte Wohnlage	durchschnittlich	gut	beste	
durchschnittlich	750	96	26	
gut	165	2598	65	
beste	17	67	436	
Gütekennzahlen pro Klasse				
Sensitivität	0,8047	0,9410	0,8273	
Spezifität	0,9629	0,8424	0,9773	

Tabelle 5: Konfusionsmatrix und Gütekennzahlen des GAM der zentralen Lagen. Accuracy: 89,67 %, n = 4220

In Tabelle 5 ist zunächst das GAM der zentralen Wohnlagen evaluiert worden. Mit einer Gesamtgenauigkeit (Accuracy) von 89,67 % zeigt das Modell eine sehr hohe prädiktive Güte. Die hohen Werte auf der Hauptdiagonalen der Konfusionsmatrix bestätigen, dass ein Großteil der Beobachtungen korrekt zugeordnet wird.

Eine detaillierte Analyse der Fehlklassifikationen zeigt zudem ein plausibles Muster. Wenn das Modell einen Fehler macht, prognostiziert es überwiegend die jeweils benachbarte Kategorie. So wird eine durchschnittliche Lage fälschlicherweise am häufigsten als gut klassifiziert (165 Fälle), aber nur sehr selten als beste (17 Fälle). Umgekehrt wird eine beste Lage am häufigsten fälschlicherweise als gut eingeordnet (65 Fälle).

Die Leistungsunterschiede zwischen den Klassen lassen sich unter anderem durch deren unterschiedliche Häufigkeit im Datensatz erklären. Die dominante Klasse „gute zentrale Lage“ wird mit einer sehr hohen Sensitivität (0,9410) exzellent erkannt. Die selteneren Klassen „durchschnittliche zentrale Lage“ und „beste zentrale Lage“ weisen erwartungsgemäß eine geringere Sensitivität, aber dafür eine exzellente Spezifität auf. Da das Modell bei der Vorhersage dieser seltenen Klassen vorsichtiger ist, vermeidet es zuverlässig, eine solche Lage fälschlicherweise zuzuordnen.

		Wahre Wohnlage		
Vorhergesagte Wohnlage	durchschnittlich	gut	beste	
durchschnittlich	33621	7426	507	
gut	8484	32876	2148	
beste	81	842	3204	
Gütekennzahlen pro Klasse				
Sensitivität	0,7970	0,7990	0,5469	
Spezifität	0,8312	0,7787	0,9889	

Tabelle 6: Konfusionsmatrix und Gütekennzahlen des GAM der Lagen außerhalb. Accuracy: 78,15 %, n = 89.189

Die Tabelle 6 fasst die Vorhersagegüte des Modells für die nicht-zentralen Lagen zusammen. Die Accuracy liegt mit 78,15 % zwar auf einem guten Niveau, ist jedoch merklich geringer als bei den zentralen Lagen (89,67 %). Dies deutet darauf hin, dass die Abgrenzung der Wohnlagen außerhalb des Zentrums anhand der verwendeten Einflussvariablen etwas schwieriger ist.

Die Analyse der Konfusionsmatrix zeigt, dass die Kategorien „durchschnittlich“ und „gut“, welche eine ähnliche Häufigkeit im Datensatz aufweisen, insgesamt gut klassifiziert werden. Ebenfalls wird, wie bei den zentralen Lagen, bei einer Fehlklassifikation meist die benachbarte Wohnlagekategorie vorhergesagt.

Besonders auffällig dagegen sind die Werte für die seltenste Wohnlage „beste Lage“. Die Sensitivität von 0,5469 bedeutet, dass das Modell nur in etwa jedes zweite Wohnobjekt mit „bester Lage außerhalb“ auch als solches korrekt klassifiziert. Im Gegenzug ist die Spezifität mit 0,9889 extrem hoch. Das Modell macht also so gut wie nie den Fehler, eine durchschnittliche oder gute Lage fälschlicherweise als beste zu deklarieren. Daraus kann man schließen, dass das Modell selten diese Wohnlage vorhersagt.

Insgesamt scheinen beide Modelle voreingenommen gegenüber den Mehrheitsklassen „zentrale gute Lage“, „durchschnittliche Lage außerhalb“ und „gute Lage außerhalb“ zu sein. Dies ist problematisch, da es in diesem Sachzusammenhang gleichermaßen wichtig ist, die seltenen Kategorien zu erkennen. Eine Möglichkeit, um diesem Problem entgegenzuwirken, wird im folgenden Kapitel vorgestellt.

6.2 Bereinigung der Vorhersagen

Wie bereits in Kapitel 4.2 eingeführt, berechnet ein multinomiales GAM für jede Beobachtung die Wahrscheinlichkeiten für die Zugehörigkeit zu den einzelnen Zielkategorien, wobei sich diese pro Beobachtung zu 1 aufsummieren. Diese Vorhersagen werden jedoch von der Häufigkeitsverteilung der Klassen in den Trainingsdaten beeinflusst. Ein Maß für diese Voreingenommenheit (Bias) sind die vom Modell implizit gelernten Prior-Wahrscheinlichkeiten, welche dem Durchschnitt der vorhergesagten Wahrscheinlichkeiten $\bar{\pi}$ über den gesamten Datensatz entsprechen. Die folgenden zwei Vektoren zeigen diese empirischen Prioren für die Modelle der zentralen und nicht-zentralen Lagen.

$$P(\text{Wohnlage})_{\text{biased, zentral}} = \begin{pmatrix} \bar{\pi}_{\text{durchschnittlich}} \\ \bar{\pi}_{\text{gut}} \\ \bar{\pi}_{\text{beste}} \end{pmatrix} = \begin{pmatrix} 0,2208531 \\ 0,6542654 \\ 0,1248815 \end{pmatrix} \quad (13)$$

$$P(\text{Wohnlage})_{\text{biased, außerhalb}} = \begin{pmatrix} \bar{\pi}_{\text{durchschnittlich}} \\ \bar{\pi}_{\text{gut}} \\ \bar{\pi}_{\text{beste}} \end{pmatrix} = \begin{pmatrix} 0,47298987 \\ 0,46131818 \\ 0,06569196 \end{pmatrix} \quad (14)$$

Es wird deutlich, dass die in den Daten häufiger vertretenen Klassen, wie die „gute zentrale Lage“ oder die „durchschnittliche Lage außerhalb“, vom Modell systematisch bevorzugt werden.

Um diesen Bias zu korrigieren und eine fairere Klassifikation zu ermöglichen, wird eine Anpassung der Wahrscheinlichkeiten vorgenommen. Dabei wird der empirische Prior durch einen uniformen Prior ersetzt, der von einer theoretischen Gleichverteilung der Klassen ausgeht:

$$P(\text{Wohnlage})_{\text{angepasst}} = \begin{pmatrix} \bar{\pi}_{\text{durchschnittlich}} \\ \bar{\pi}_{\text{gut}} \\ \bar{\pi}_{\text{beste}} \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}. \quad (15)$$

Die Re-Kalibrierung der ursprünglichen Vorhersage für eine Lokation erfolgt dann für jede Klasse k nach der folgenden Transformationsformel, welche den alten Prior entfernt und den neuen einsetzt. Dabei ist $P(\text{Wohnlage} = k | \mathbf{x}_i)_{\text{biased}}$ die vom Modell vorhergesagte Wahrscheinlichkeit für die Klasse k einer Lokation $i = 1, \dots, n$:

$$\begin{aligned} P(\text{Wohnlage} = k | \mathbf{x}_i)_{\text{angepasst}} &= P(\text{Wohnlage} = k | \mathbf{x}_i)_{\text{biased}} \cdot \frac{P(\text{Wohnlage} = k)_{\text{angepasst}}}{P(\text{Wohnlage} = k)_{\text{biased}}} \\ &= P(\text{Wohnlage} = k | \mathbf{x}_i)_{\text{biased}} \cdot \frac{\bar{\pi}_{k, \text{angepasst}}}{\bar{\pi}_{k, \text{biased}}} \end{aligned} \quad (16)$$

wobei $k \in \{\text{durchschnittlich}, \text{gut}, \text{beste}\}$ (vgl. Tian et al. 2020). Diese Formel wird nun auf die geschätzten Klassenwahrscheinlichkeiten aus beiden Modellen angewendet.

6.3 Modellevaluierung mit bereinigten Vorhersagen

Die folgenden Tabellen evaluieren analog zu Abschnitt 5.4.1 die Modelle mit den bereinigten Vorhersagen.

Wahre Wohnlage			
Vorhergesagte Wohnlage	durchschnittlich	gut	beste
durchschnittlich	829	262	20
gut	62	2355	12
beste	41	144	495
Gütekennzahlen pro Klasse			
Sensitivität	0,8895	0,8530	0,9393
Spezifität	0,9142	0,9493	0,9499

Tabelle 7: Konfusionsmatrix und Gütekennzahlen des GAM der zentralen Lagen mit Bereinigung. Accuracy: 87,18 %, n = 4220

Tabelle 7 zeigt zunächst die Ergebnisse des Modells für die zentralen Lagen nach der vorgenommenen Wahrscheinlichkeits-Anpassung. Die Gesamtgenauigkeit (Accuracy) ist

mit 87,18 % weiterhin auf einem sehr hohen Niveau, aber leicht gesunken im Vergleich zum ursprünglichen Modell (vorher 89,67 %). Dieser leichte Rückgang ist zu erwarten und der Preis für ein faireres, weniger voreingenommenes Modell.

Die entscheidende Veränderung zeigt sich bei den klassenspezifischen Gütemaßen, insbesondere bei den selteneren Klassen. Die Fähigkeit des Modells, die seltenen besten zentralen Lagen korrekt zu erkennen, hat sich dramatisch verbessert. Die Sensitivität ist von 0,8273 auf 0,9393 gestiegen. Das Modell übersieht folglich deutlich weniger der wertvollen Spitzenlagen. Auch bei der zentralen durchschnittlichen Lage gab es eine deutliche Steigerung der Sensitivität von 0,8047 auf 0,8895. Im Gegenzug ist die Sensitivität für die häufigste Klasse „gute zentrale Lage“ von 0,9410 auf 0,8530 gesunken. Das Modell ist nun vorsichtiger, die dominante Klasse vorherzusagen.

		Wahre Wohnlage		
Vorhergesagte Wohnlage	durchschnittlich	gut	beste	
durchschnittlich	32682	6561	182	
gut	7369	28172	738	
beste	2135	6411	4939	
Gütekennzahlen pro Klasse				
Sensitivität	0,7747	0,6847	0,84298	
Spezifität	0,8565	0,8313	0,8974	

Tabelle 8: Konfusionsmatrix und Gütekennzahlen des GAM der Lagen außerhalb mit Bereinigung. Accuracy: 73,77 %, n = 89.189

Im Vergleich dazu zeigt Tabelle 8 die Ergebnisse des Modells für die nicht-zentralen Lagen nach der Wahrscheinlichkeits-Anpassung. Die Gesamtgenauigkeit ist auch hier mit 73,77 % leicht gesunken (vorher 78,15 %).

Ebenfalls hat sich die Fähigkeit, die seltene „beste Lage“ zu erkennen, erheblich verbessert. Die Sensitivität ist von einem sehr niedrigen Wert von 0,5469 auf 0,8430 gestiegen. Das Modell findet nun die überwiegende Mehrheit der Spitzenlagen. Im Gegenzug ist die Sensitivität für die beiden häufigeren Klassen deutlich gesunken (von ca. 0,79 auf 0,68 für „gut“ und 0,77 für „durchschnittlich“). Die Spezifität ist für die Klasse „beste Lage“ mit 0,8974 außerdem etwas schwächer als zuvor, bleibt aber nach wie vor auf einem exzellenten Niveau.

6.4 Analyse der räumlichen Muster

Ebenfalls ist in diesem Zusammenhang die Frage interessant, welche Regionen Münchens viele oder wenige falsch klassifizierte Lokationen enthalten und ob sich dabei räumliche Muster ergeben. Gebiete mit hoher Fehlklassifikationsrate könnten möglicherweise nicht zu der ihnen zugeschriebenen Wohnlage passen. Dafür werden im Folgenden alle falsch klassifizierten Wohnobjekte auf die Wohnlagenkarte der Stadt München abgebildet. Die

Farbe der Punkte entspricht dabei der vorhergesagten Wohnlage. Zur besseren visuellen Betrachtung wird die Stadt in jeweils fünf Karten geteilt.

6.4.1 Struktur der unbereinigten Vorhersagen

Zunächst soll die räumliche Struktur der unbereinigten Vorhersagen analysiert werden.



Abbildung 55: Karte der Fehlklassifikationen aus dem Modell der zentralen Wohnlagen

Die Fehlklassifikationskarte des Modells der zentralen Wohnlage (Abbildung 55) offenbart ein klares, systematisches Muster der Fehlentscheidungen. Es ist eine auffällige Häufung von Fehlklassifikationen in den Gebieten der seltenen Klassen zu erkennen. Viele Punkte liegen in den Arealen der besten zentralen Lage (dunkelviolett) und der durchschnittlichen zentralen Lage (hellblau). Dies bestätigt die Erkenntnisse aus der Konfusionsmatrix. Das ursprüngliche Modell ist voreingenommen (biased) und neigt stark dazu, die dominante Klasse „gute Lage“ vorherzusagen. Zudem ist erkennbar, dass in einigen Gebieten keine oder sehr wenige Lokationen falsch eingeordnet werden. An anderen Stellen finden sich dagegen einige räumliche Cluster wieder. Oft passieren Fehler auch direkt an der Grenze zu einer anderen Wohnlagekategorie.

Für die nicht-zentralen Lagen geben folgende vier Grafiken die Fehlklassifikationsstruktur wieder:



Abbildung 56: Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Nordwesten



Abbildung 57: Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Nordosten

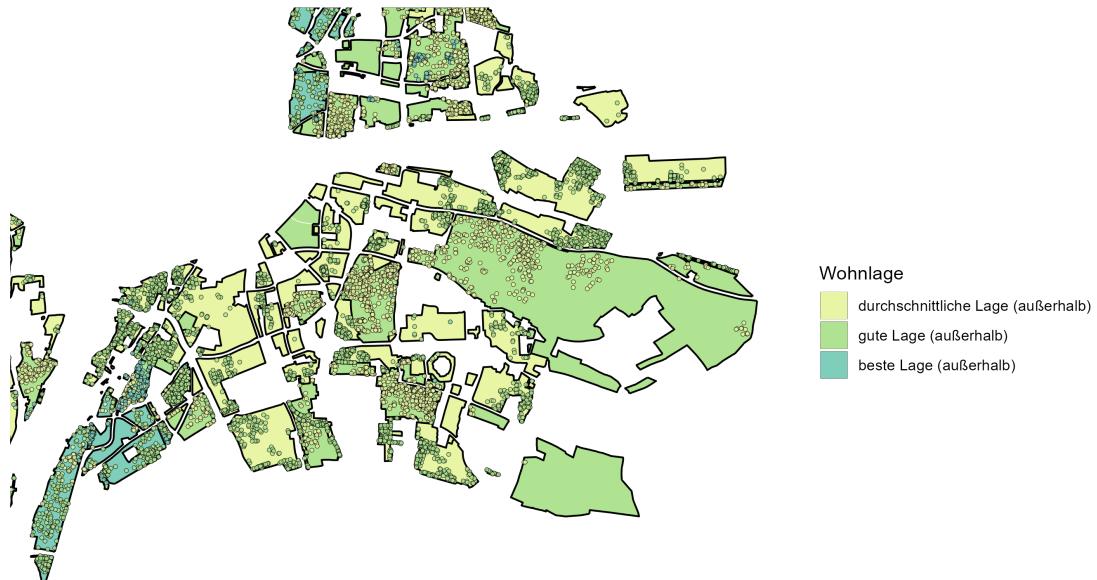


Abbildung 58: Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Südosten



Abbildung 59: Karte der Fehlklassifikationen aus dem Modell der Wohnlagen außerhalb: Südwesten

Insgesamt ist festzustellen, dass einige Gebiete vor allem im Nordwesten und Nordosten von München keine fehlklassifizierten Lokationen beinhalten. In anderen Stadtteilen bilden sich dagegen auch hier, ähnlich wie bei den zentralen Lagen, räumliche Cluster. Ebenfalls ist zu erkennen, dass die beste Lage außerhalb aufgrund des Bias im Modell selten richtig vorhergesagt wurde, da viele Fehler in den Arealen (türkise Flächen) dieser Lage passieren, aber wenige zentrale beste Lagen in den anderen Gebieten falsch klassifiziert werden.

6.4.2 Struktur der bereinigten Vorhersagen

Des Weiteren wird nun die Struktur der Fehlklassifikationen der Modelle mit bereinigten Vorhersagen betrachtet.

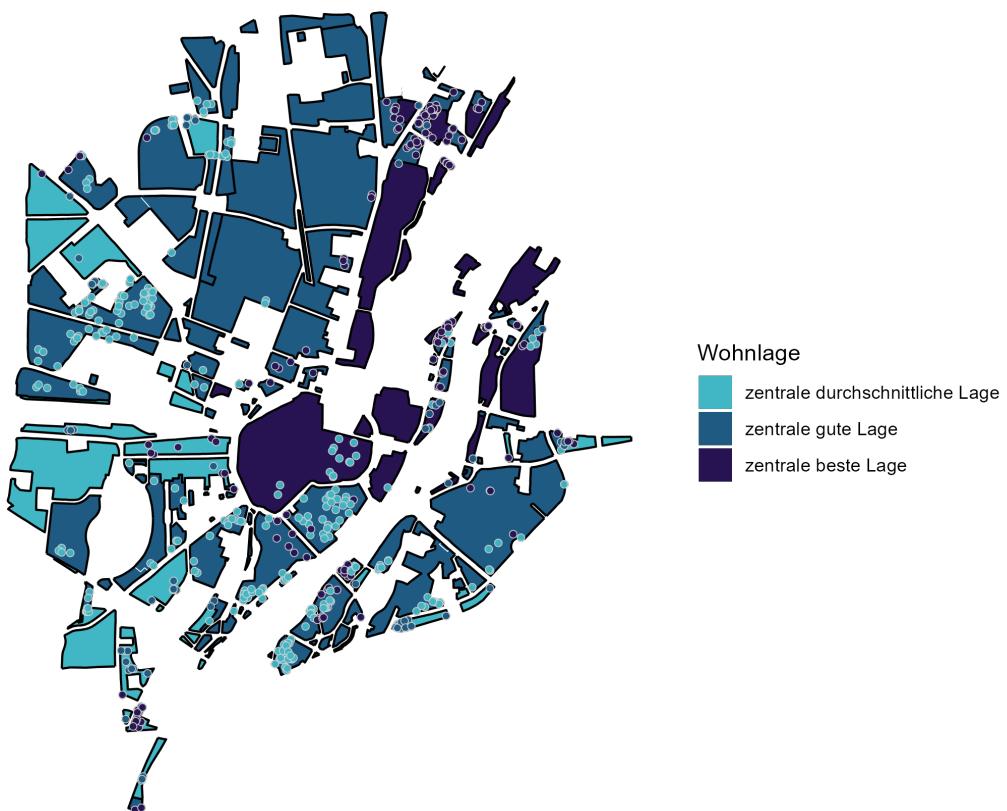


Abbildung 60: Karte der bereinigten Fehlklassifikationen aus dem Modell der zentralen Wohnlagen

Durch die Abbildung 60 sind einige interessante Unterschiede zu den Fehlklassifikationen ohne bereinigte Vorhersagen festzustellen. Die systematische Häufung von Fehlern in den Arealen der seltenen Klassen ist nun deutlich reduziert. Das Modell ist jetzt wesentlich besser in der Lage, auch „beste“ und „durchschnittliche“ Lagen korrekt zu identifizieren. Die verbleibenden Fehler sind nicht mehr das Ergebnis eines systematischen Bias, sondern konzentrieren sich nun stärker auf die geografischen Grenzen zwischen den verschiedenen

Wohnlagenzonen. Insgesamt sind auch, wie zu erwarten, generell mehr Fehler zu erkennen, da mit der Bereinigung die Accuracy gesunken ist (vgl. Kapitel 6.3).

Des Weiteren soll nun die räumliche Struktur der fehlerhaften Vorhersagen des nicht-zentralen Modells mit Bereinigung dargestellt werden.

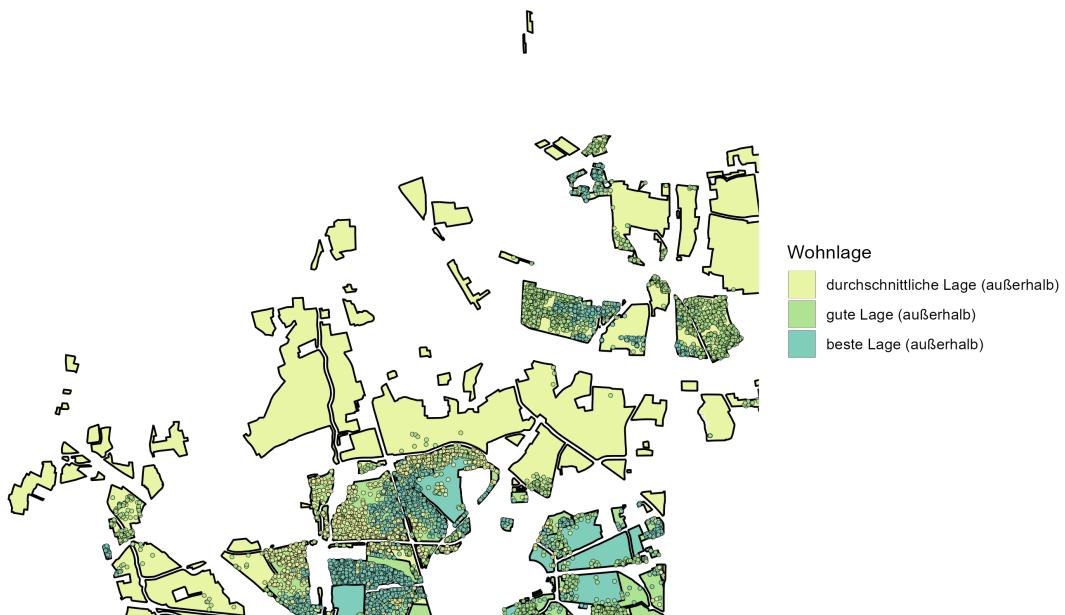


Abbildung 61: Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Nordwesten



Abbildung 62: Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Nordosten



Abbildung 63: Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Südosten



Abbildung 64: Karte der bereinigten Fehlklassifikationen aus dem Modell der nicht-zentralen Wohnlagen: Südwesten

Um eine detailliertere und fundiertere räumliche Analyse möglich zu machen, befinden sich interaktive Karten im elektronischen Anhang. Sie beinhalten alle korrekt und falsch klassifizierten Lokationen, sowie diverse Straßennamen und Häuserumrisse. Folgende Abbildung zeigt einen kleinen Ausschnitt der interaktiven Karte mit den bereinigten Vorhersagen.

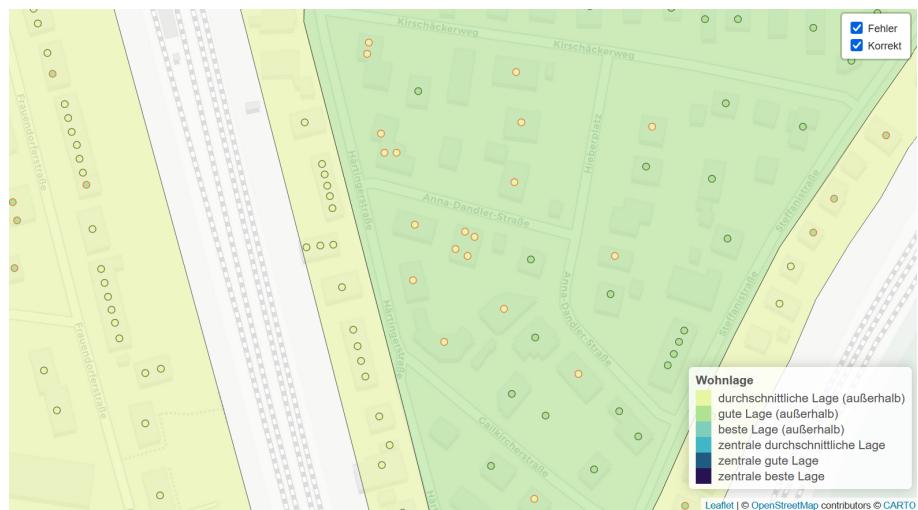


Abbildung 65: Ausschnitt aus der interaktiven Karte mit bereinigten Vorhersagen

7 Fazit und Ausblick

Die vorliegende Arbeit untersuchte die Bestimmungsfaktoren von Wohnlagenqualität in München, wobei ein besonderer Fokus auf den Unterschieden zwischen zentralen und nicht-zentralen Gebieten lag. Die methodische Grundlage bildeten zwei Generalisierte Additive Modelle, deren Eignung bestätigt wurde. Die durchweg hohen effektiven Freiheitsgrade für alle metrischen Prädiktoren und diversen partiellen Effektpplots belegten, dass die Zusammenhänge hochgradig nicht-linear sind und der flexible GAM-Ansatz gegenüber einem rein linearen Modell klar überlegen war.

In dieser Arbeit konnten einige klare Zusammenhänge zu wichtigen Standortfaktoren herausgearbeitet und quantifiziert werden. Mit Variablen wie dem Nahversorgungsindex oder dem Hauspreisindex können gute und beste Lagen von durchschnittlichen Lagen abgegrenzt werden. Die starken Unterschiede zwischen den zentralen und nicht-zentralen Lagen haben sich als wichtige Erkenntnis dieser Arbeit herausgestellt. Einige Faktoren, die sich positiv auf die Wohnlagequalität für zentrale Wohnlagen auswirkten, zeigten einen vollkommen gegensätzlichen Zusammenhang zu den nicht-zentralen Wohnlagen. Beispielsweise erhöht sich die Chance für eine gute oder beste zentrale Lage c.p., wenn ein Wohnobjekt direkt an einem Fußgängerbereich liegt. Dagegen reduzierte sich die Chance für eine gute oder beste Lage außerhalb, wenn dies der Fall ist. Insgesamt deuten die Ergebnisse somit auf unterschiedliche Präferenzen hin.

Durch die beiden GAMs lassen sich die überwiegende Mehrheit der im Datensatz enthaltenen Lokationen korrekt klassifizieren, was für einen starken Zusammenhang der im Modell verwendeten Kovariablen mit den Wohnlagen spricht. Der bisher nur indirekt über Variablen wie `distanz_bahnhof` und `straßentyp` erfasste Faktor „Lärm“ spielt höchstwahrscheinlich ebenfalls eine Rolle beim Einteilen der Wohnlagen.

Methodisch erwies sich die im Analyseprozess durchgeführte Bereinigung der mit den Modellen vorgenommenen Vorhersagen der Wohnlagen als wertvoll. Die ursprünglichen, rein auf den Trainingsdaten basierenden Modelle zeigten eine deutliche Voreingenommenheit gegenüber den häufigeren Wohnlagenkategorien wie zum Beispiel der zentralen guten Lage. Durch die Anpassung der Prior-Wahrscheinlichkeiten konnte vor allem die Sensitivität für die selteneren, aber wichtigen besten Lagen verbessert werden. Dieser Schritt ermöglicht es, zu praktisch relevanten und fairen Vorhersagemodellen zu gelangen, auch wenn er zu einem leichten Rückgang der Gesamtgenauigkeit führte.

Ein oberflächliche Analyse der räumlichen Position der Fehlklassifikationen hat ergeben, dass sich bei den zentralen und nicht-zentralen Lagen einige räumliche Cluster bilden. Während sich in manchen Regionen keine Fehlklassifikationen finden lassen, häufen sie sich in anderen Bereichen der Stadt.

Eine genauere Betrachtung und Analyse auffälliger Regionen wäre ein interessantes Thema für zukünftige weiterführende Analysen.

A Anhang

A.1 Plots

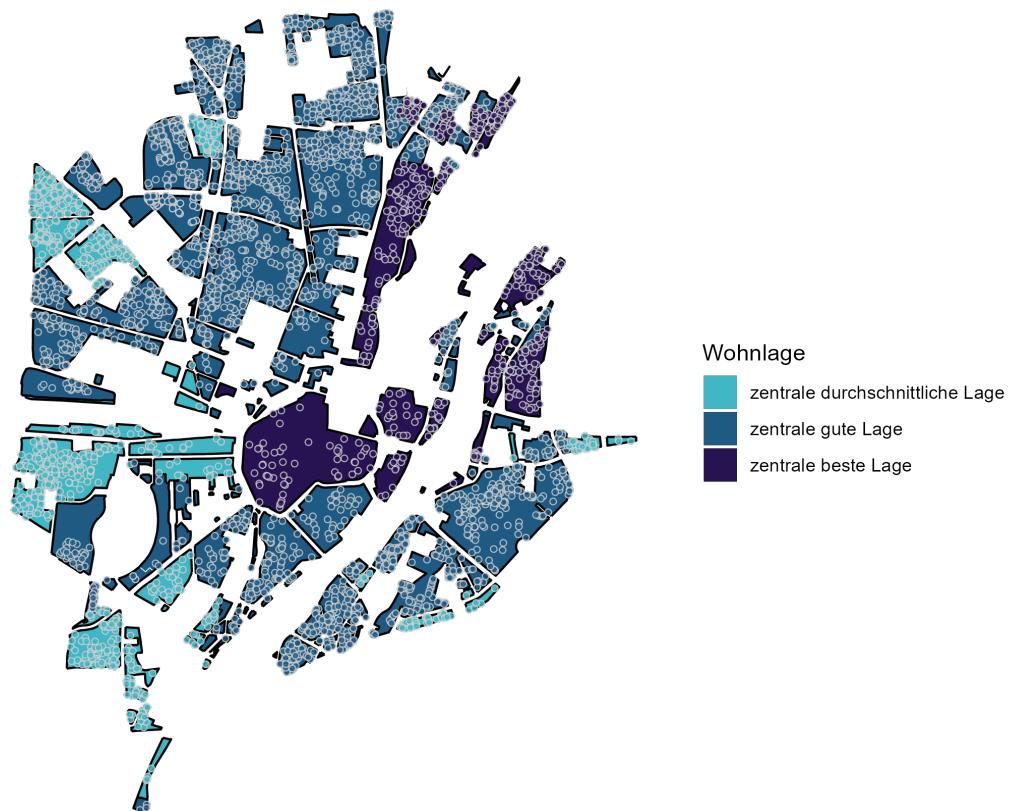


Abbildung 66: Karte aller Lokationen mit zentraler Wohnlage

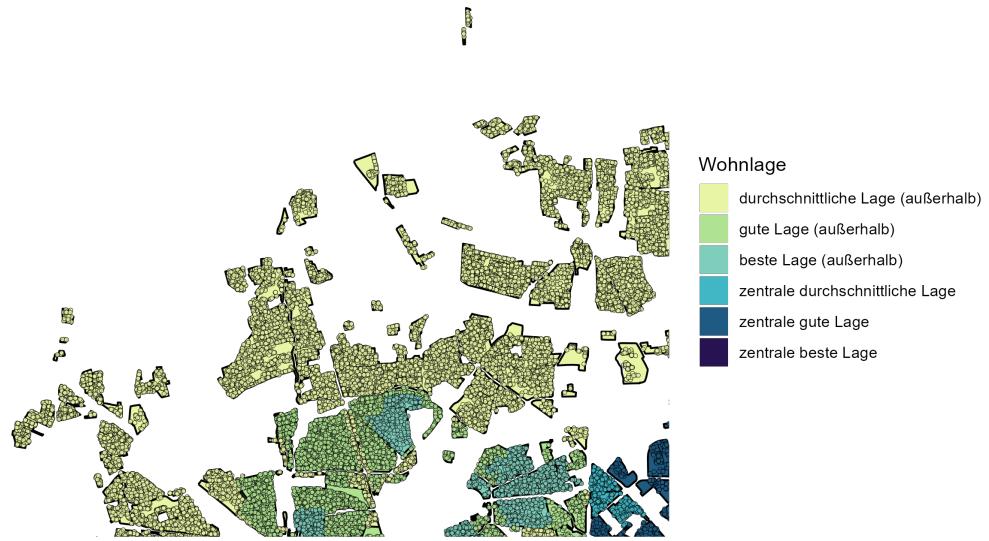


Abbildung 67: Karte aller Lokationen: Nordwesten



Abbildung 68: Karte aller Lokationen: Nordosten

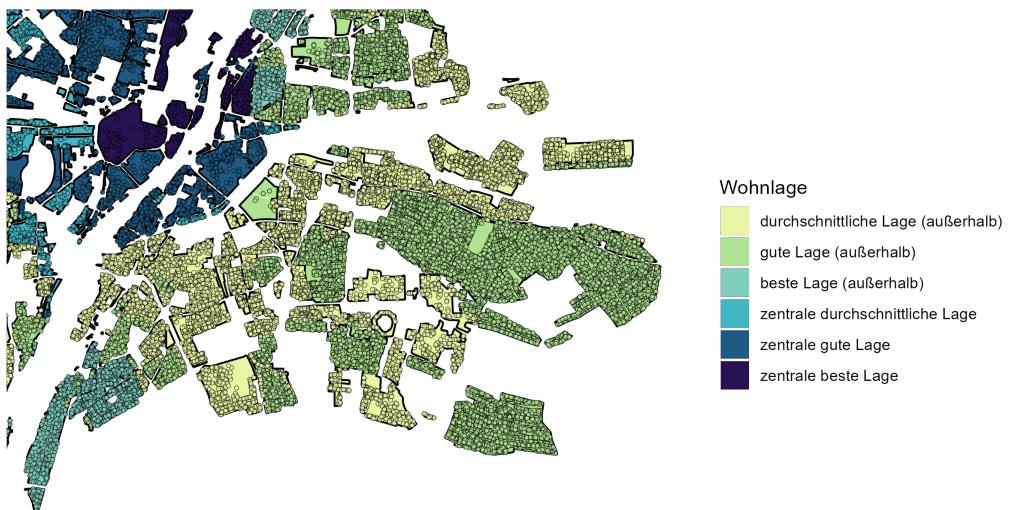


Abbildung 69: Karte aller Lokationen: Südosten

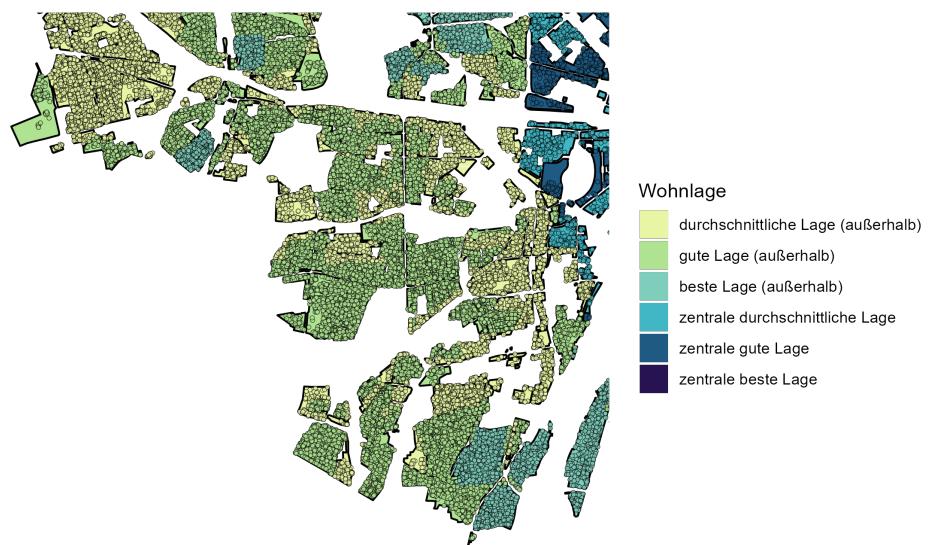


Abbildung 70: Karte aller Lokationen: Südwesten

A.2 Tabellen

Wohnlage	Anteil an der Gesamtfläche
durchschnittliche Lage (außerhalb)	45.60 %
gute Lage (außerhalb)	34.40 %
beste Lage (außerhalb)	7.33 %
zentrale durchschnittliche Lage	2.16 %
zentrale gute Lage	8.17 %
zentrale beste Lage	2.37 %

Tabelle 9: Flächenmäßiger Anteil der vordefinierten Wohnlagen der Stadt München

Wohnlage	Absolute Häufigkeit	Relative Häufigkeit
durchschnittliche Lage (außerhalb)	42186	45.16 %
gute Lage (außerhalb)	41144	44.05 %
beste Lage (außerhalb)	5859	6.27 %
zentrale durchschnittliche Lage	932	1.00 %
zentrale gute Lage	2761	2.96 %
zentrale beste Lage	527	0.56 %

Tabelle 10: Verteilung der vordefinierten Wohnlagen im finalen Datensatz

Parametrische Terme	Schätzer	Std.-Fehler	z-Wert	p-Wert
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Intercept	8.8730	1.7447	5.086	<0.001
Straßentyp: Sammelstraße	0.9863	0.2876	3.429	0.0006
Straßentyp: Wohnstraße	1.1744	0.2163	5.429	<0.001
Straßentyp: Fußgängerbereich	1.1784	0.4205	2.803	0.0051
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Intercept	1.0496	1.8394	0.571	0.5683
Straßentyp: Sammelstraße	1.5447	0.5028	3.072	0.0021
Straßentyp: Wohnstraße	1.7257	0.3870	4.460	<0.001
Straßentyp: Fußgängerbereich	0.7401	0.6229	1.188	0.2347
Geglättete Terme (Splines)	EDF	Chi-Quadrat	p-Wert	
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Distanz zum Bahnhof	7.677	105.07	<0.001	
Distanz zum Mittelzentrum	8.348	215.50	<0.001	
ÖPNV-Index	5.007	43.93	<0.001	
Distanz zum Unterzentrum	5.077	50.32	<0.001	
Hauspreisindex	3.962	138.87	<0.001	
Distanz zur U-Bahn	7.229	56.50	<0.001	
Distanz zur Bushaltestelle	7.006	78.34	<0.001	
Nahversorgungsindex	8.473	342.62	<0.001	
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Distanz zum Bahnhof	7.735	91.48	<0.001	
Distanz zum Mittelzentrum	5.937	75.15	<0.001	
ÖPNV-Index	5.757	54.40	<0.001	
Distanz zum Unterzentrum	7.660	142.13	<0.001	
Hauspreisindex	5.499	146.07	<0.001	
Distanz zur U-Bahn	4.122	29.95	<0.001	
Distanz zur Bushaltestelle	5.772	68.22	<0.001	
Nahversorgungsindex	8.223	186.64	<0.001	

Tabelle 11: Modelloutput des GAM für die zentralen Lagen. Deviance explained = 70.3 %, n = 4220

Parametrische Terme	Schätzer	Std.-Fehler	z-Wert	p-Wert
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Intercept	0.312 41	0.085 52	3.653	<0.001
Straßentyp: Sammelstraße	0.092 38	0.038 45	2.403	0.0163
Straßentyp: Wohnstraße	-0.068 46	0.034 21	-2.001	0.0454
Straßentyp: Fußgängerbereich	-0.163 41	0.057 81	-2.826	0.0047
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Intercept	-41.126 43	5.307 89	-7.748	<0.001
Straßentyp: Sammelstraße	0.300 07	0.078 95	3.801	<0.001
Straßentyp: Wohnstraße	-0.340 87	0.070 95	-4.804	<0.001
Straßentyp: Fußgängerbereich	-1.106 33	0.183 25	-6.037	<0.001
Geglättete Terme (Splines)	EDF	Chi-Quadrat	p-Wert	
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Distanz zum Bahnhof	9.689	784.4	<0.001	
Distanz zum Mittelzentrum	9.855	2189.8	<0.001	
ÖPNV-Index	8.934	1349.3	<0.001	
Distanz zum Unterzentrum	8.733	2680.5	<0.001	
Hauspreisindex	7.430	7734.3	<0.001	
Distanz zur U-Bahn	9.435	564.0	<0.001	
Distanz zur Bushaltestelle	7.999	176.1	<0.001	
Nahversorgungsindex	9.875	6053.5	<0.001	
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Distanz zum Bahnhof	9.846	1088.5	<0.001	
Distanz zum Mittelzentrum	8.281	1366.8	<0.001	
ÖPNV-Index	8.287	903.0	<0.001	
Distanz zum Unterzentrum	7.470	901.8	<0.001	
Hauspreisindex	9.035	3904.2	<0.001	
Distanz zur U-Bahn	9.812	490.3	<0.001	
Distanz zur Bushaltestelle	7.010	320.9	<0.001	
Nahversorgungsindex	7.573	663.5	<0.001	

Tabelle 12: Modelloutput des GAM für die Lagen außerhalb. Deviance explained = 43.6 %, n = 89184

Parametrische Terme	Schätzer	Std.-Fehler	z-Wert	p-Wert
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Intercept	7.2723	1.6462	4.418	<0.001
Straßentyp: Sammelstraße	0.9575	0.2700	3.546	0.0004
Straßentyp: Wohnstraße	1.2176	0.2035	5.984	<0.001
Straßentyp: Fußgängerbereich	1.7454	0.3914	4.459	<0.001
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Intercept	0.9949	1.6846	0.591	0.5548
Straßentyp: Sammelstraße	1.7556	0.4684	3.748	0.0002
Straßentyp: Wohnstraße	1.8387	0.3756	4.896	<0.001
Straßentyp: Fußgängerbereich	1.4892	0.6052	2.460	0.0139
Geglättete Terme (Splines)	EDF	Chi-Quadrat	p-Wert	
<i>Effekte für „gute Lage“ (vs. durchschnittlich)</i>				
Distanz zum Mittelzentrum	8.036	294.75	<0.001	
ÖPNV-Index	4.584	39.05	<0.001	
Distanz zum Unterzentrum	7.273	97.93	<0.001	
Hauspreisindex	4.245	142.21	<0.001	
Distanz zur U-Bahn	7.281	72.95	<0.001	
Distanz zur Bushaltestelle	6.210	63.99	<0.001	
Nahversorgungsindex	8.591	402.96	<0.001	
<i>Effekte für „beste Lage“ (vs. durchschnittlich)</i>				
Distanz zum Mittelzentrum	6.815	116.65	<0.001	
ÖPNV-Index	7.277	93.47	<0.001	
Distanz zum Unterzentrum	6.064	171.33	<0.001	
Hauspreisindex	6.194	159.78	<0.001	
Distanz zur U-Bahn	2.555	31.42	<0.001	
Distanz zur Bushaltestelle	5.629	59.02	<0.001	
Nahversorgungsindex	8.151	173.42	<0.001	

Tabelle 13: Modelloutput des angepassten GAM für die zentralen Lagen (ohne `distanz_bahnhof`). Deviance explained = 66.8 %, n = 4220

B Elektronischer Anhang

Daten, Code und Grafiken sind in elektronischer Form verfügbar.

rohdaten - Dieser Ordner enthält die rohen Datensätze.

daten - Dieser Ordner enthält folgende Dateien zur Datenverarbeitung:

- data_read.R: Code zum einlesen der Rohdaten.
- data_edit.R: Code zur Datenvorverarbeitung.
- model_data_ausserhalb_complete.RData: Aufbereiteter Datensatz der nicht-zentralen Wohnlagen.
- model_data_zentral_complete.RData: Aufbereiteter Datensatz der zentralen Wohnlagen.
- model_data_complete.RData: Aufbereiteter Datensatz mit allen Wohnlagen.

modellierung - Dieser Ordner enthält folgende Dateien zur Modellierung:

- fehlklassifikations_karten.R: Code zur Erstellung der Fehlklassifikationskarten und interaktiven Karten.
- final_gam.R: Code zur Erstellung der GAMs.
- final_model_analysis.R: Code zur Analyse der GAMs.
- model_evaluation.R: Codedatei mit Hilfsfunktionen zur Evaluierung der GAMs.
- plot_funktionen.R: Codedatei mit Hilfsfunktionen zur Erstellung der Grafiken.

modelle - Dieser Ordner enthält folgende Modelldateien:

- gam_model_ausserhalb.rds: Modell der nicht-zentralen Wohnlagen.
- gam_model_zentral.rds: Modell der zentralen Wohnlagen.
- gam_model_zentral_ohne.rds: Modell der zentralen Wohnlagen ohne `distanz_bahnhof`.

plots - Dieser Ordner enthält alle Grafiken dieser Arbeit.

interaktive_karten - Dieser Ordner enthält alle interaktiven Karten.

EDA.R - Code zur Erstellung der deskriptiven Grafiken.

paket_download.R - Code Datei zum Herunterladen der benötigten R Pakete.

Literatur

Agresti, Alan (2007). *An Introduction to Categorical Data Analysis*. Second. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471226185.

Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen (n. d.). *Mietspiegel für München*. Zuletzt abgerufen am 26.05.2025. URL: <https://www.bmwsb.bund.de/Webs/BMWSB/DE/themen/stadt-wohnen/wohnungswirtschaft/fakten-wohnungsmarkt/mietspiegel/mietspiegel-node.html>.

Clark, Michael (2022). *Generalized Additive Models. Application Using R*. Zuletzt abgerufen am 15.08.2025. URL: <https://m-clark.github.io/generalized-additive-models/application.html>.

Clark, Nicholas (2024). *How to interpret and report nonlinear effects from Generalized Additive Models*. Zuletzt abgerufen am 15.08.2025. URL: <https://ecogambler.netlify.app/blog/interpreting-gams/>.

Fahrmeier, Ludwig et al. (2013). *Regression*. Springer. ISBN: 9783642343322.

Fahrmeier, Ludwig et al. (2016). *Statistik. Der Weg zur Datenanalyse*. eighth. Springer. ISBN: 9783662503713.

James, Gareth et al. (2021). *An Introduction to Statistical Learning: with Applications in R*. Second. Springer. ISBN: 9781071614174.

Landeshauptstadt München (2023). *Mietspiegel für München 2023. Information zur ortsüblichen Miete*. Broschüre, Zuletzt abgerufen am 02.07.2025. URL: https://2023.mietspiegel-muenchen.de/broschueren/Mietspiegel_2023_Broschuere.pdf.

- (2025a). *Mietspiegel für München*. Zuletzt abgerufen am 26.05.2025. URL: <https://stadt.muenchen.de/infos/mietspiegel.html>.
- (2025b). *Mietspiegel für München 2025. Statistik, Dokumentation und Analysen*. Broschüre, Zuletzt abgerufen am 02.07.2025. URL: https://2025.mietspiegel-muenchen.de/broschueren/Dokumentation_MS25.pdf.

Landeshauptstadt München (2025). *Münchner U-Bahn: Strecken, Linien, Tickets und weitere Infos*. Zuletzt abgerufen am 10.08.2025. URL: <https://www.muenchen.de/verkehr/oeffentlicher-nahverkehr/u-bahn-muenchen-die-wichtigsten-infos>.

- Pebesma, Edzer (2018). “Simple Features for R: Standardized Support for Spatial Vector Data”. In: *The R Journal* 10.1, S. 439–446. DOI: 10.32614/RJ-2018-009. URL: <https://doi.org/10.32614/RJ-2018-009>.
- Tian, Junjiao et al. (2020). “Posterior Re-calibration for Imbalanced Datasets”. In: *CoRR* abs/2010.11820. arXiv: 2010 . 11820. URL: <https://arxiv.org/abs/2010.11820>.
- Wickham, Hadley et al. (2025). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.5.1.* Manual, Zuletzt abgerufen am 10.08.2025. URL: <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>.
- Wood, Simon N. (2017a). “Comment”. In: *Journal of the American Statistical Association* 112.517, S. 164–166. DOI: 10 . 1080 / 01621459 . 2016 . 1270050. eprint: <https://doi.org/10.1080/01621459.2016.1270050>. URL: <https://doi.org/10.1080/01621459.2016.1270050>.
- (2017b). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.9-3.* Manual, Zuletzt abgerufen am 18.08.2025. URL: <https://cran.r-project.org/web/packages/mgcv/mgcv.pdf>.
- Yee, Thomas W. (2015). *Vector Generalized Linear and Additive Models. With an Implementation in R.* Springer. ISBN: 9781493928170.
- Zumel, Nina und John Mount (2019). *vtreat: a data.frame Processor for Predictive Modeling.* arXiv: 1611 . 09477 [stat.AP]. URL: <https://arxiv.org/abs/1611.09477>.