Ludwig-Maximilians-Universität München
Institut für Statistik

Datenerhebung bei Wohnungsmarktbarometern

Bachelorarbeit

Jessica Peter

betreut von
Prof. Dr. Göran Kauermann
und Dr. Michael Windmann

21. Juni 2018

Zusammenfassung

Ziel dieser Bachelorarbeit war es herauszufinden, ob sich beim einmaligen oder wöchentlichen Abgreifen von Online-Immobilienanzeigen der Website "Immobilienscout24" eine Verzerrung in den Daten ergibt und somit zu einem verzerrten Bild der Nettomiete pro Quadratmeter für Wohnungsmarktbarometer liefert. Außerdem soll die Verzerrung gegebenenfalls mit passenden Gewichten gemindert werden. Dafür spielen vor allem die Variablen "Nettomiete pro Quadratmeter" und die "Online-Anzeigedauer" eine Rolle. Zweitere löst die hier betrachtete Verzerrung aus und trägt zu ihrer Korrektur als Gewichtung bei, denn es zeigt sich zunächst durch ein generalisiertes additives Modell, dass im Schnitt Wohnungen, die länger online sind, eine höhere Nettomiete pro Quadratmeter aufweisen. Greift man die Anzeigen einmalig ab, erhält man durchschnittlich mehr Wohnungen, die eine höhere Nettomiete pro Quadratmeter aufweisen als der Durchschnitt. Durch eine Datensimulation konnte gezeigt werden, dass der Mittelwert der Nettomiete pro Quadratmeter in einem Monat, in dem täglich Anzeigen abgegriffen wurden, im Schnitt geringer ausfällt, als beim einmaligen Abgreifen an einem beliebigen Tag im jeweiligen Monat, und somit eine Verzerrung der Daten vorliegt. Durch das Berechnen des Mittelwerts mit einer Gewichtung durch die inversen Anzeigedauer beim einmaligen Abgreifen der Anzeigen und dem Vergleich des Monatsmittels (bei täglichem Abgreifen), kann empirisch eine Entzerrung der Daten festgestellt werden, welche theoretisch begründet wird. Gleiches gilt für eine beim einmaligen Abgreifen praktisch anwendbare Gewichtung mit der inversen bisherigen Anzeigedauer, welche Werte erzeugt, die näher am Monatsmittelwert der Nettomiete pro Quadratmeter liegen als bei ersterer Gewichtung.

Inhaltsverzeichnis

1 Einleitung

Eine der wichtigsten Aufgaben eines Statistikers ist es, Daten mit adäquaten Methoden auszuwerten und so zu Informationen zu gelangen. Doch selbst die beste Auswertung kann zu Ergebnissen gelangen, die fern von der Realität sind, wenn die Daten verzerrt sind. Deswegen ist es unabdingbar, Daten möglichst sauber zu erheben. Die Datenerhebung ist weitaus weniger trivial als es auf den ersten Blick zu sein scheint, wie folgende Fälle zeigen: Möchte man in einem Ort lebende Personen befragen, reicht es nicht, irgendwann Leute an einem Platz zu befragen, denn zu unterschiedlichen Zeiten befinden sich dort andere Personengruppen und man würde überproportional viele Personen einer Gruppe befragen [1, vgl. S.5ff.]. Hat man das Ziel Informationen über ein sensibles Thema wie Einkommen von Personen zu gelangen, muss Anonymität hergestellt werden, damit die Befragten nicht in Richtung gesellschaftlicher Normen antworten [2].
In diesen einfachen Beispielen gibt es zahlreiche weitere Faktoren, die das Ergebnis verzerren können, wie z.B. das Verhalten des Interviewers.
Damit ist klar, dass auch die Datenerhebung nicht unterschätzt werden darf. Diese Arbeit beschäftigt sich mit der Datenerhebung für Wohnungsmarktbarometer, welches unter anderem den durchschnittlichen Nettomietpreis pro Quadratmeter einer Stadt angibt.
Sie hat das Ziel aufzudecken, ob sich eine Verzerrung der Daten und insbesondere der Nettomiete pro Quadratmeter ergibt, wenn die Wohnungsanzeigen einer Immobilienwebsite nur einmal oder wöchentlich abgerufen werden und falls vorhanden, diese durch geeignete Gewichte auszugleichen. Dafür wird ein Datensatz verwendet, dessen Inhalt vom sehr häufigen Abgreifen von Wohnungsanzeigen der Immobilienwebsite "Immobilienscout24" stammt.
Um eine Verzerrung aufzudecken, werden zunächst die Originaldaten, der Hintergrund für die Vermutung der Verzerrung und die Datenaufbereitung sowie die für die Analyse wichtigsten Variablen beschrieben. Anschließend wird die Theorie des generalisierten additiven Regressionsmodells sowie die Durchführung und die Ergebnisse eines solchen Modells auf Basis der aufbereiteten Daten erläutert. Danach wird eine Datensimulation vorgestellt, welche aus den aufbereiteten Daten die einmalige Informationsgewinnung sowie die tägliche Datengewinnung für einen Monat simuliert, bevor eine geeignete Gewichtung zur Reduktion der Verzerrrung vorgestellt wird. Anhand von ungewichteten und gewichteten Mittelwerten wird gezeigt, dass die Verzerrung durch Gewichtung vermindert werden kann. Zum Schluss werden die wichtigsten Ergebnisse zusammengefasst und ein Ausblick gegeben.

2 Einführung der Daten und Hintergrund der Arbeit

2.1 Datenhintergrund

Die Daten wurden durch Web Scraping von der Website "Immobilienscout24" erlangt. Web Scraping bedeutet, dass diese Informationen nicht durch einen Menschen per Hand abgelesen werden, sondern durch ein Programm, dass die Informationen automatisch von der Website extrahiert und strukturiert. Dazu werden die Daten meist anhand des HTML-Codes gewonnen. [3, vgl. S.44f] Die Website "Immobilienscout24" dient vor allem zum Finden oder Anbieten von Immobilien zum Kauf oder zur Miete. Jede Anzeige verfügt über viele Informationen. Neben Adresse, Preis und Quadratmeter finden sich unter anderem zahlreiche Angaben zur Ausstattung [4].
Somit besteht der Datensatz aus ca. 180 Variablen zur Wohnungsvermietung und besitzt Daten von den Jahren 2012 bis Anfang 2018. Der für diese Arbeit vorliegende Datensatz wurde auf private Vermietungen in München ohne Auszugsfrist reduziert und besitzt ca. 80.000 Beobachtungen.
Durch die nicht tägliche Datenerhebung wird eine Verzerrung vermutet, welche im Folgenden dargestellt wird.

2.2 Hintergrund der vermuteten Verzerrung bei wöchentlichem oder einmaligem Abgreifen der Daten

In München herrscht eine besondere Mietsituation. Die Einwohnerzahl Münchens steigt immer weiter [5], wodurch die Mieten teurer und die Wohnungen knapper werden [6]. Aufgrund dieses Mangels liegt die Vermutung nahe, dass sehr gute Angebote in kürzester Zeit nicht mehr bestehen. Dies würde für eine Immobilienwebsite bedeuten, dass die Wohnungen mit einer sehr geringen Nettomiete pro Quadratmeter binnen weniger Tage wieder von der Seite genommen werden. Geht man davon aus, dass gute Angebote weniger lang online sind als schlechtere, würde man beim wöchentlichen oder einmaligen Abgreifen durch die kürzere Onlinezeit weniger gute Angebote in den Daten finden, als anteilig vermietet werden und damit ein verzerrtes Bild der Realität erlangen.
In Abbildung 1 ist diese Theorie für eine beispielhafte Woche aufgezeigt. Hierbei steht jeder farbige Balken für eine Wohnungsanzeige, welche nach der Höhe des Nettomietpreises pro Quadratmeter eingefärbt ist. Dabei steht grün für eine verhältnismäßig niedrige, gelb für eine mittelhohe und rot für

eine hohe Nettomiete pro Quadratmeter. Nach der Theorie wurde die Anzeigedauer daran angepasst, wie günstig eine Wohnung pro Quadratmeter ist. Greift man die Daten nun einmal pro Woche ab, wie zum Beispiel am Freitag, erhält man hier aufgrund der Onlinezeit der Anzeigen alle Wohnungen mit hoher und fast alle mit mittlerer Nettomiete pro Quadratmeter, aber nur eine Wohnung mit niedrigem Nettomietpreis pro Quadratmeter und somit ein verzerrtes Bild der Wohnungsmarktsituation.

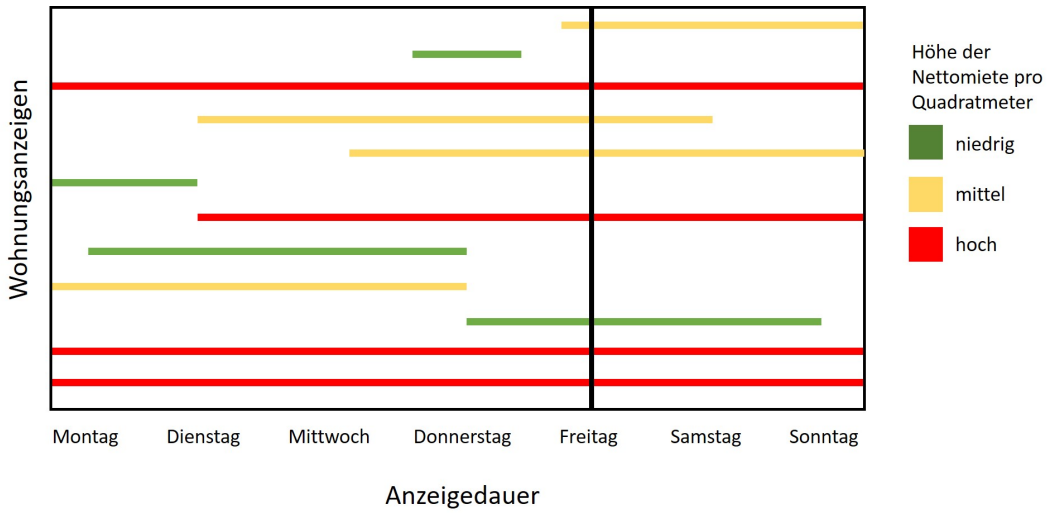

Abbildung 1:
Theorie der Verzerrung beim einmaligen oder wöchentlichem Abgreifen der Daten an einer beispielhaften Woche: Jeder Balken stellt eine Wohnungsanzeige dar. Ein grüner Balken signalisiert eine Wohnung mit niedriger, ein gelber eine mit mittlerer und ein roter eine mit hoher Nettomiete pro Quadratmeter. Die Länge der Balken gibt die Anzeigedauer an. Der vertiakle schwarze Balken signalisiert das beispielhafte Abgreifen an einem Wochentag.

3 Datenaufbereitung, Variablenerstellung und Variablenbeschreibung

3.1 Datenaufbereitung und Variablenerstellung

Im Datensatz liegen Start- und Enddatum der Anzeigen vor. Aus diesen wurde die Zeitdifferenz in Tagen berechnet, um später den Effekt der Anzeigedauer auf den Nettomietpreis pro Quadratmeter beurteilen zu können. Zwei weitere neue Variablen sind die Zentroids-Ortsdaten. Die Ortsdaten wurden durch Aufteilung der Stadt München in Polygone nach den Postleitzahlbereichen und durch Nutzung der Schwerpunkte dieser Vielecke gebildet. Die Schwerpunkte werden durch durch x- und y-Koordinaten angegeben. Die Variable, die die Anzahl von Balkonen und Terrassen angibt, wurde in eine kategoriale Variable mit den Kategorien "kein Balkon oder Terrasse", "ein Balkon oder Terrasse" und "mehr als ein Balkon oder Terrasse" umgewandelt. Zudem wurden Beobachtungen von Neuvermietungen ausgeschlossen, da diese ein spezielles, für diese Analyse irrelevantes Marktsegment darstellen. Auch wurde der Datensatz auf Beobachtungen in den mittleren 95 Prozent der Variablendaten von Nettomiete pro Quadratmeter, Fläche und Anzeigedauer in Tagen beschränkt, um extreme Werte zu entfernen. Da das 2,5 Prozent Quartil der Anzeigedauer bei dem Minimum 0 liegt, entspricht die Einschränkung für diese Variable den geringsten 97,5 Prozent der Variablendaten. Durch Ausschluss von Neuvermietungen und Ausreißern reduziert sich der Datensatz von über 80.000 Beobachtungen auf ca. 65.000.

3.2 Variablenbeschreibung und -anpassung mit deskriptiver Analyse

In diesem Abschnitt werden die für die Analyse relevanten Variablen beschrieben. Für die statistischen Kennzahlen wird der nicht reduzierte Datensatz verwendet.

Ausstattungsklassen nach Definition von Empirica

Um zu beschreiben wie gut die Ausstattung einer Wohnung ist, existiert eine Skala von einfach über normal und gut bis hochwertig. Mit 40,38 Prozent machen Wohnungen mit guter Ausstattung den höchsten Anteil aus. Nicht viel seltener gibt es normal ausgestattete Wohnungen (34,77 Prozent) und ein gutes Fünftel (21,24 Prozent) der Wohnungen sind hochwertig ausgestattet. Nur circa 3,61 Prozent der Wohnungen weisen eine einfache Ausstattung auf.

Nettomiete pro Quadratmeter in Euro

Nun wird die Variable Nettomiete pro Quadratmeter in Euro betrachtet, welche die Zielvariable des generalisierten additiven Modells darstellen wird. Die Kosten belaufen sich auf 2 bis 50 Euro pro Quadratmeter, wobei die mittleren 50 Prozent der Daten zwischen 12,50 und 16,57 Euro pro Quadratmeter liegen. Der mittlere Wert der Nettomiete pro Quadratmeter beträgt 14,32 Euro und liegt damit unter dem Mittelwert von 14,82 Euro. Somit ist die Verteilung leicht linkssteil. In Abbildung 2 ist die Verteilung der mittleren 95 Prozent der Nettomiete pro Quadratmeter zu erkennen.

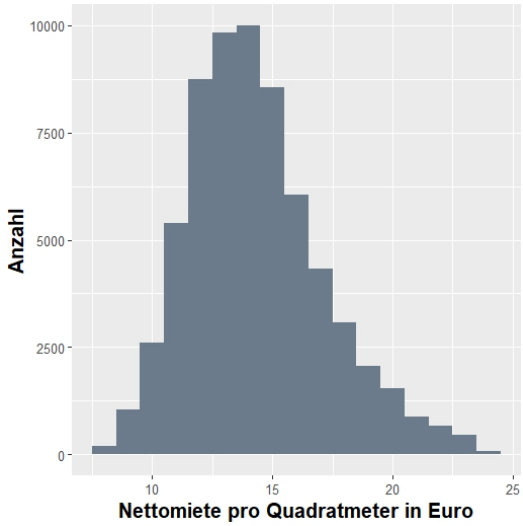

Abbildung 2:
Histogramm für die Variable Nettomiete pro Quadratmeter in Euro eingeschränkt auf die mittleren 95 Prozent der Variablendaten

Wohnfläche in Quadratmetern

Eine Einflussgröße für das generalisierte additive Modell stellt die Wohnfläche in Quadratmetern dar. Hierbei besitzt die kleinste Mietsache 8 und die größte 571,5 Quadratmeter. Die Hälfte aller Wohnungen hat eine Größe zwischen 52 und 91 Quadratmetern. Der Median beträgt 70,00 Quadratmeter und liegt 5,56 Quadratmeter unter dem Mittelwert. Somit liegt eine linkssteile Verteilung vor. Auch diese Variable wurde auf die mittleren 95 Prozent reduziert, welche von 26 bis 166 Quadratmetern reichen. In Abbildung 3 findet sich ein Histogramm der Wohnfläche eingeschränkt auf die mittleren 95 Prozent der Variablendaten.

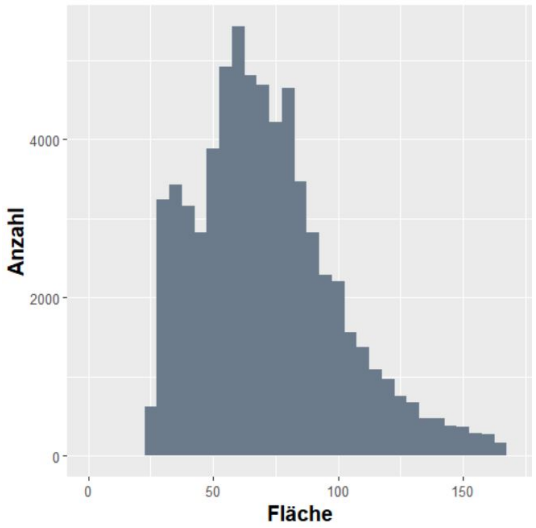

Abbildung 3:
Histogramm für die Variable Fläche in Quadratmetern eingeschränkt auf die mittleren 95 Prozent der Variablendaten

Anzeigedauer in Tagen

Eine weitere Einflussgröße für die Nettomiete ist, wie lang eine Wohnungsanzeige online war. Diese Variable berechnet sich mit Hilfe des Start- und Enddatums der Anzeige. Die Dauer wird in Tagen angegeben und das Minimum liegt bei weniger als einem Tag und das Maximum bei 1972 Tagen. Die Verteilung ist linkssteil. Dreiviertel der Anzeigen sind weniger als 37 Tage online, die Hälfte der Anzeigen weniger als 14 Tage, was auch in Abbildung 4, welche einen Ausschnitt der empirischen Verteilungsfunktion zeigt, dargestellt ist. 95 Prozent der Anzeigen sind weniger als 123 Tage online. Der Median liegt mit 13 Tagen deutlich unter dem Mittelwert von ca. 33 Tagen. Beim Betrachten von Abbildung 5, welche ein Histogramm über die auf die geringsten 97,5 Prozent der Variablendaten reduzierten Daten enthält, fällt auf, dass ein Ausreißer bei 2,5 Monaten vorliegt.

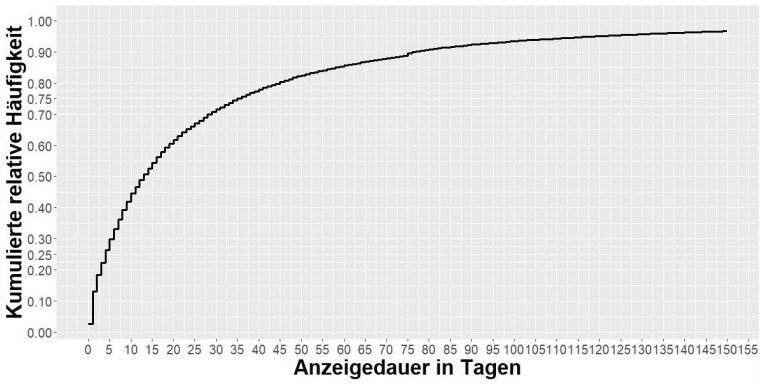

Abbildung 4:
Ausschnitt der empirischen Verteilungsfunktion der Anzeigedauer in Tagen auf 1-150 Tage

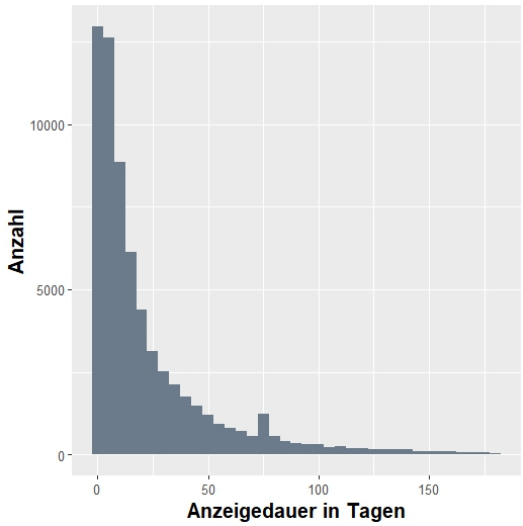

Abbildung 5:
Histogramm für die Variable Anzeigedauer in Tagen eingeschränkt auf die geringsten 97,5 Prozent der Variablendaten

Existenz einer Einbauküche

Neben den Ausstattungsklassen werden zwei weitere Ausstattungsmerkmale herangezogen. Als erstes Ausstattungsmerkmal wird betrachtet, ob sich eine Einbauküuche in der Wohnung befindet. 66,14 Prozent der Wohnungen besitzen eine Einbauküche.

Anzahl der am Objekt vorhandenen Balkone und/oder Terrassen

Ein weiteres Ausstattungsmerkmal ist die Anzahl an Balkonen und/oder Terrassen an dem Mietobjekt. Circa ein Viertel (25,64 Prozent) der Wohnungen besitzt keinen Balkon und keine Terrasse und fast zwei Drittel der Wohnungen (64,79 Prozent) besitzen einen Balkon oder eine Terrasse. Knapp zehn Prozent der Mietsachen (9,57 Prozent) haben mehr als einen Balkon und/oder Terrassen.

Wohnlage

Um den räumlichen Effekt auf die Nettomiete beurteilen zu können, dienen die Zentroid-Variablen als Einflussgröße. In Abbildung 6 sind die Schwerpunkte der Postleitzahlbereiche durch Punkte auf einer Karte von München veranschaulicht.

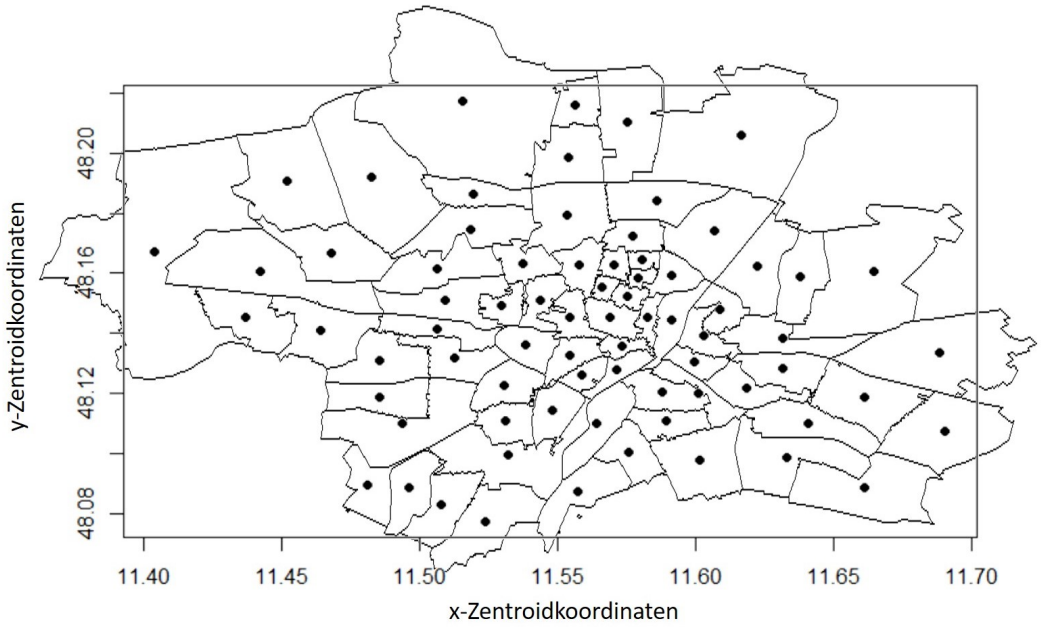

Abbildung 6:
Zentroide der Postleitzahlbereiche Münchens: Abgebildet ist München mit seinen Postleitzahlbereichen. Die Postleitzahl wurde durch den Schwerpunkt jedes Bereichs ersetzt, welcher durch x- und y-Koordinaten dargestellt wird.

4 Regressionsanalyse auf aufbereiteten Originaldaten

Um einen Überblick über Zusammenhänge zwischen Variablen und Zielgröße zu erhalten, sowie zu prüfen, ob tatsächlich bei einer höheren Anzeigedauer im Schnitt ein höherer Nettomietpreis pro Quadratmeter auftritt, wird ein generalisiertes additives Regressionsmodell mit den in Kapitel 3.2 beschriebenen Variablen durchgeführt.

4.1 Theorie des generalisierten additiven Modells

Um den Einfluss verschiedener Variablen auf die Nettomiete pro Quadratmeter zu beurteilen, wurde ein generalisiertes additives Modell (GAM) verwendet. Bei einem GAM handelt es sich um ein generalisiertes lineares Modell, dessen additiver Prädiktor η aus der Summe linearer $\beta_0+\beta_1\cdot x_{\mathrm{i1}}+\ldots+\beta_k\cdot x_{\mathrm{ik}}$ und nichtlinearer Komponenten $f_1(z_{\mathrm{i1}})+\ldots+f_{\mathrm{q}}(z_{\mathrm{iq}})$ besteht: [7, vgl. S.119] [8, vgl. S.46]

$$\eta = f_1(z_{\mathrm{i1}})+\ldots+f_{\mathrm{q}}(z_{\mathrm{iq}})+\beta_0+\beta_1\cdot x_{\mathrm{i1}}+\ldots+\beta_k\cdot x_{\mathrm{ik}} \tag{1}$$

$z_1,\ldots,z_q$ stellen metrische Kovariablen dar. $x_1,\ldots,x_q$ können metrisch, binär oder mehrkategorial sein. Zweitere wirken linear auf die Zielgröße im Gegensatz zu $z_1,\ldots,z_q$, welche einen nicht linearen Einfluss auf die Zielgröße haben. [8, vgl. S.46]
Der Erwartungswert der Zielvariable entspricht dem mit der Responsefunktion transformierten Prädiktor: $E(y)=h(\eta)$
Da lineare und nichtlineare Einflüsse betrachtet werden, wird dies als semiparametrische Regression bezeichnet. Zusätzlich zu diesem Standardterm können Interaktionen aufgenommen werden, um gemeinsame Effekte auf die Zielgröße durch zwei Variablen zu beurteilen.
Prinzipiell wird davon ausgegangen, dass ein Modell nie perfekt ist und Schätzfehler vorliegen. Jedoch gilt bei generalisierten additiven Modellen die Annahme, dass die Fehler ϵ_i zufällig sind und keine Struktur aufweisen. Ist dies gegeben, gleichen sich die Fehler aus und der Erwartungswert der Fehler ergibt null. Wegen der Zufälligkeit der Fehler werden diese zudem als unabhängig und als identisch verteilt angenommen.[8, S.19, 21, 46]
Bei einem generalisierten Modell folgt die Zielgröße, welche hier die Nettomiete pro Quadratmeter ist, einer Verteilung aus der Exponentialfamilie, wie zum Beispiel der Normalverteilung. [8, vgl. S.218]

Die Kombination aus linearen und nichtlinearen Komponenten hat den großen Vorteil, dass ein Zusammenhang zwischen der Zielgröße und einer Einflussgröße nicht nur durch eine Gerade beschrieben werden kann, sondern durch eine passende glatte Funktion. Dies ist bedeutend, da in der Praxis häufig nichtlineare Zusammenhänge auftreten. [8, vgl. S.399]
Die linearen Einflüsse des Modells werden über die Methode der kleinsten Quadrate bestimmt. Hierbei werden die Regressionskoeffizienten geschätzt, indem die Beobachtungen für eine Einflussgröße und die Zielgröße in ein Koordinatensystem eingetragen und eine Linie so durch die Punkte gelegt wird, dass die Summe von eins bis zur Stichprobengröße n der quadratischen Abstände von dem y-Wert der Beobachtungen y_i zur Regressionsgeraden $x_i' \cdot \beta$ minimiert werden.
Die Abstände werden klassischerweise quadriert, um Beobachtungen mit größeren Abständen stärker zu gewichten als Punkte mit geringen Abständen. Die unquadrierten Abstände der Beobachtungen zur Gerade werden auch als Residuen $\hat{\epsilon_i}$ bezeichnet. Diese stellen eine Schätzung für die Fehler ϵ_i des Modells dar. Durch Minimierung der KQ-Formel erhält man den β-Koeffizient für die gewählte Einflussgröße. Die KQ-Formel lautet: [8, vgl. S.63,90ff.]

$$KQ(\beta) = \sum_{i=1}^{n} (y_i - x_i' \cdot \beta)^2 = \sum_{i=1}^{n} (\epsilon_i)^2 \tag{2}$$

In Abbildung 7 sind beispielhafte Beobachtungen mit ihren Abständen zur KQ-Formel minimierenden Regressionsgeraden dargestellt. Außerdem ist für die Beobachtung x_i das Residuum $\hat{\epsilon_i}$, der y-Wert y_i und der durch die Regressionsgerade geschätzte y-Wert $\hat{y_i}$ aufgezeigt.

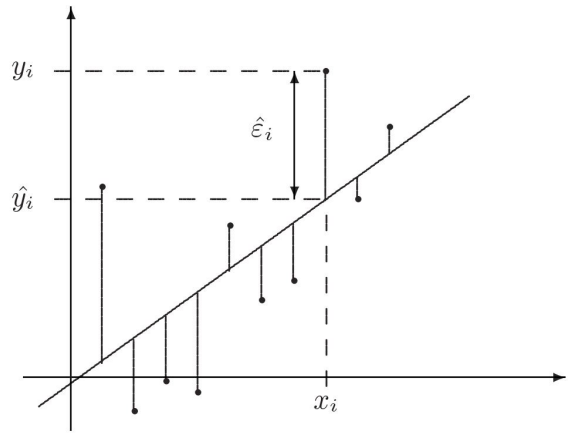

Abbildung 7:
KQ-Methode: Abstände beispielhafter Beobachtungen zur Regressionsgeraden mit Darstellung des Residuums $\hat{\epsilon}_i$, des y-Werts y_i und des durch die Regressionsgerade geschätzten y-Werts $\hat{y}_i$ einer Beobachtung x_i. [8, S.92]

Dummy-kodierte und kategoriale Variablen werden linear in das Modell aufgenommen. Metrische Variablen werden zunächst nichtparametrisch, also als glatte Funktion dargestellt. Falls sich diese als linear oder annähernd linear herausstellen, empfiehlt es sich, diese Einflussgröße linear aufzunehmen, um unnötige Komplexität zu vermeiden. Die geglätteten Funktionen können auf verschiedene Weisen erzeugt werden. Eine bewährte Möglichkeit stellen penalisierte Basic-Splines (B-Splines) dar [8, vgl. S.401]. Dies bedeutet, dass für eine Einflussgröße die Beobachtungen mit der Zielgröße in ein Koordinatensystem eingetragen werden und der Wertebereich auf der x-Achse in Abschnitte eingeteilt wird. Die Grenzpunkte der Abschnitte werden auch als Knoten bezeichnet. Zu den Knoten werden sogenannte Basisfunktionen gebildet: Bei jedem Knoten beginnt ein Polynom vom Grad l und endet so, dass sich zwischen zwei Knoten immer l+1 Teilstücke von Polynomen befinden und die einzelnen Polynome glatt in das nächste übergehen. Jedes Polynom ist dabei gleich und lediglich auf der x-Achse verschoben. In dem von k+2 Knoten aufgespannten Bereich sind die Werte positiv und sonst null. Diese B-Spline-Basen werden in Abbildung 8 beispielhaft am Grad drei mit neun äquidistanten Knoten veranschaulicht. [8, vgl. S.303ff.]

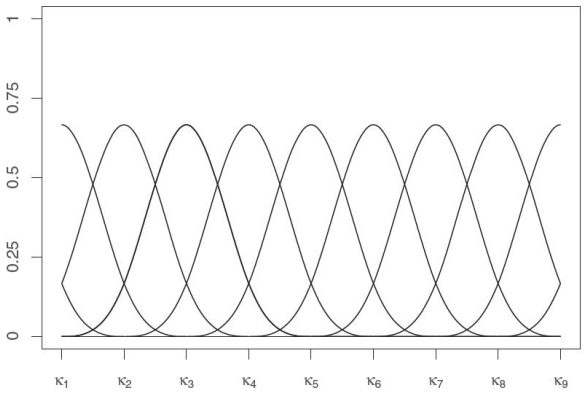

Abbildung 8:
B-Spline Basen: B-Spline-Basen mit Grad drei und neun äquidistanten Knoten [8, S.304]

Durch eine Linearkombination gewichteter Basisfunktionen $B_j(z)$ kann jegliche Spline-Funktion des Grads l zu einer gegebenen Anzahl von m Knoten und damit die gesuchte glatte Funktion durch die Beobachtungen gebildet werden. γ stellt hierbei den Parametervektor dar. [8, vgl. S.303ff.]

$$f(z) = \sum_{j=1}^{m+l-1} \gamma_j \cdot B_j(z) \tag{3}$$

Die j-te Basisfunktion zum Grad l ist dabei rekursiv zu folgendem Term definiert, wobei k_j der j-te Knoten ist: [8, vgl. S.303ff.]

$$B_j^l(z) = \frac{z - k_j}{k_{j+l} - k_j} B_j^{l-1}(z) + \frac{k_{j+l+1} - z}{k_{j+l+1} - k_{j+1}} B_{j+1}^{l-1}(z) \tag{4}$$

Für $B_j^1(z)$ gilt: [8, vgl. S.303ff.]

$$B_j^1(z) = \frac{z - k_j}{k_{j+1} - k_j} \mathbb{1}_{[k_j, k_{j+1})}(z) + \frac{k_{j+2} - z}{k_{j+2} - k_{j+1}} \mathbb{1}_{[k_{j+1}, k_{j+2})}(z) \tag{5}$$

Die Anzahl der Knoten spielt hier eine große Rolle, denn liegen zu viele Knoten vor, kommt es zu einer zu rauen und bei zu wenigen Knoten zu einer zu glatten Funktion. Um die ideale Funktion zu finden, muss entweder die Anzahl an Knoten sinnvoll gewählt werden oder eine zu raue Funktion durch

einen Penalisierungsterm bestraft werden. Zweiteres soll hier betrachtet werden. Um die Bestrafung durchzuführen, wird zunächst eine Anzahl von Knoten (z.B. 30) gewählt, die groß genug ist, damit raue Funktionen gebildet werden können. Anschließend wird die zugehörige Funktion geschätzt und der Bestrafungsterm zur KQ-Formel addiert. P-Splines sind also B-Splines mit Penalisierung. [8, vgl. S.306f.] Der Penalisierungsterm besteht aus dem Parameter λ, der die Rauheit der Funktion bestimmt und dem Integral über das Quadrat der zweiten Ableitung der im ersten Schritt erzeugten Funktion $f(z)$, da die Krümmung der Funktion als Maß für die Rauheit verwendet werden kann. Der Strafterm hat die Form: [8, vgl. S.309ff.]

$$\lambda \cdot \int (f''(z))^2 dz \tag{6}$$

Der Koeffizientenvektor wird analog zu oben mit der KQ-Formel durch Minimierung berechnet, wobei hier der Penalisierungsterm addiert wird. So ergibt sich folgende penalisierte KQ-Formel: [7, vgl. S.144]

$$PKQ(\gamma) = \sum_{i=1}^{n} (y_i - f(z))^2 + \lambda \cdot \int (f''(z))^2 dz \tag{7}$$

Im durchgeführten Modell wird für sowohl lineare als auch nicht-lineare Einflüsse eine modifizierte Form der KQ-Methode verwendet, bei welcher die Residuen auch quadriert, aber noch zusätzlich gewichtet werden (und für nicht lineare Effekte zusätzlich penalisiert werden). Das Ziel dieser Methode namens IRLS bzw. penalisierte IRLS ist es die Schätzung stabiler zu machen und die Modellannahmen besser zu erfüllen. Die Gewichte werden dabei iterativ bestimmt. [7, S76ff.,S.165f.] [9, S.9]

Zuletzt muss der Glättungsparameter λ geeignet gewählt werden, um Overfitting (zu raue Funktion) oder zu geringer Datentreue vorzubeugen. Der Einfluss von Lamda wird in Abbildung 9 an beispielhaften Daten veranschaulicht. Dabei stellt die rote Funktion jeweils die Schätzung und die schwarze das Ideal dar. Ganz links hat λ einen Wert von null und es kommt zu extremem Overfitting. Das andere Extrem, wenn Lamda gegen unendlich geht und die Funktion maximal glatt ist, wird ganz rechts veranschaulicht. In der Mitte ist ein angemesseneres Lamda dargestellt.

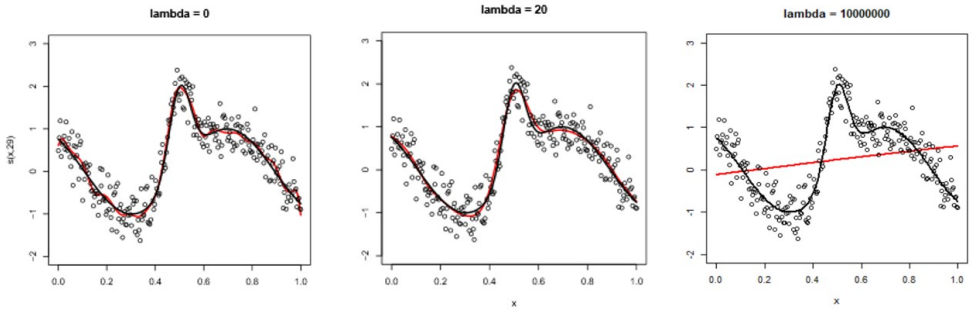

Abbildung 9:
Einfluss des Glättungsparameters λ bei P-Splines an beispielhaften Simulationsdaten: Die schwarze Kurve veranschaulicht den idealen Verlauf des Splines und die roten den geschätzten Spline. Links zeigt sich Overfitting bei zu kleinem λ von null, in der Mitte ein geeignetes Lamda von 20 und rechts eine zu starke Glättung bei Lamda gegen Unendlich mit einem Wert von 10.000.000. [10]

Um Lamda zu wählen, kann eine Kreuzvalidierung verwendet werden. Allgemein werden bei der Kreuzvalidierung eine oder mehrere Beobachtungen aus dem Datensatz entfernt und eine Schätzung auf Basis der restlichen Daten durchgeführt, um anschließend zu sehen, wie gut die Schätzung zu den gelöschten Beobachtungen passt bzw. diese voraussagt. Für Penalisierungen lässt sich zeigen, dass sich der Funktionswert für die ausgelassene Beobachtung über die Glättungsmatrix bestimmen lässt, ohne die Schätzung durchzuführen. Diese Form wird in der Praxis jedoch häufig durch einen generalisierten Kreuzvalidierungsterm approximiert, welcher die Spur $sp()$ der Glättungsmatrix S verwendet. Dieser hat den Vorteil, deutlich weniger rechenintensiver zu sein, dadurch dass sich die Spur in diesem Fall leicht berechnen lässt, da Matrizen innerhalb der Spur verschiebbar sind. Der Kreuzvalisierungsterm hat die Form: [8, vgl. S.350ff.]

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}(z_i)}{1 - sp(S)/n} \right)^2 \tag{8}$$

Bei Minimierung dieses Terms erhält man einen Glättungsparameter, der einen guten Kompromiss zwischen Glattheit und Datentreue findet. [8, vgl. S.350ff.]

Das folgende generalisierte additive Modell wurde mit dem Programm R berechnet. Hierbei wurde die gam-Funktion aus dem Paket "mgcv"verwendet, welches von Simon Wood stammt und die beschriebene Methodik nutzt. [8, vgl. S.309] [11]

Lediglich für den Effekt der Ortsvariablen werden statt P-Splines "thin-plate-splines" eingesetzt, welche gute und recheneffiziente zweidimensionale Lösungen bieten und für räumliche Effekte sehr gut geeignet sind. Diese Splines basieren auf Basisfunktionen, die aus der euklidischen Norm der Differenz von Knoten und dem Beobachtungspunkt - hier Zentroid x- und y-Variable-bestehen. Die thin-plate-splines stellen eine Erweiterung der Glättungssplines für die Schätzung von Oberflächen dar. [8, vgl. S.379]

4.2 Spezifikation und Ergebnisse des generalisierten additiven Modells

Spezifikation des GAMs

Für die Regressionsanalyse wurde aus oben genannten Gründen ein generalisiertes additives Modell gewählt. Als Verteilungsfamilie wurde die Normalverteilung gewählt, da die Nettomiete pro Quadratmeter bedingt auf den Einflussgrößen zwar nicht negativ erwartet wird, jedoch die Werte groß genug sind, um als approximativ normalverteilt angenommen zu werden. Auch der Mietspiegel Münchens, welcher als methodisch sehr gut gilt, verwendet bei gleicher Zielgröße die Normalverteilung [9, vgl. S.9]. Da als Linkfunktion die Identität gewählt wurde, entspricht $E(y)$ dem Prädiktor η. Um λ zu wählen, wird die generalisierte Kreuzvalidierung verwendet, da die wahre Varianz unbekannt ist. Für die eindimensionalen Splines wurden penalisierte B-Splines und bei den Ortsvariablen thin-plate-splines eingesetzt. Betrachtet wird das Haupteffektmodell auf den aufbereiteten Daten mit der Nettomiete pro Quadratmeter als Zielgröße. Als lineare Einflussgrößen dienen die Existenz einer Einbauküche, die Ausstattungsklassen und die Anzahl an Balkonen und/oder Terrassen. Als nicht lineare Einflussgrößen werden die Ortsvariablen, die Anzeigedauer in Tagen und die Wohnfläche verwendet. Da der Ort durch Schwerpunkte angegeben ist, welcher durch eine x- und y-Koordinate dargestellt wird, sind die x- und y-Koordinaten als Interaktion in das Modell aufgenommen. Somit hat das Modell die unten stehende Form, wobei eine vorhandene Einbauküche und eine einfache Ausstattungsklasse jeweils die Referenzkategorie zu Einbauküche ja/nein beziehungsweise zur Ausstattungsklasse ist:

$$\hat{y} = \beta_0 + \beta_1 \cdot Einbauküche_{\text{nein}} + \beta_2 \cdot Ausstattungsklasse_{\text{normal}}$$
$$+\beta_2 \cdot Ausstattungsklasse_{\text{gut}} + \beta_2 \cdot Ausstattungsklasse_{\text{hochwertig}}$$
$$+\beta_3 \cdot Balkone/Terrassen + f_1(Anzeigedauer) + f_2(Wohnfläche)$$
$$+f_3(x\text{-}Koordinate \times y\text{-}Koordinate) \quad (9)$$

Ergebnisse

Intercept

Der Intercept gibt den Mittelwert der Nettomiete pro Quadratmeter in den Referenzkategorien an. Das heißt für eine Wohnung mit einer Einbauküche, einer einfachen Ausstattung, keinem Balkon bzw. Terrasse, einem Ort mit den Koordinaten von ca. 11,51 und 48,09, einer Anzeigedauer von 10 Tagen und einer Fläche von 55 Quadratmetern kostet ein Quadratmeter durchschnittlich ca. 13,81 Euro.

Ausstattungsklassen

Verändert man zum Intercept nur die Ausstattungsklasse, erhöht sich die Nettomiete pro Quadratmeter im Vergleich zur einfachen Ausstattung für eine normale Ausstattung durchschnittlich um ca. 0,25 Euro, für eine gute Ausstattung um ca. 0,53 Euro und für eine hochwertige Ausstattung um ca. 1,30 Euro.

Wohnfläche

In Abbildung 11 ist der nicht-lineare geschätzte Effekt der Wohnfläche auf die Nettomiete pro Quadratmeter veranschaulicht. Die Nettomiete nimmt zunächst für eine zunehmende Fläche ab, bevor sich diese ab einer Fläche von ca. 70 Quadratmetern stabilisiert und ab dieser Größe sehr geringfügig steigt. Bis ca. 35 Quadratmeter Wohnfläche ist ein sehr starker Abfall des geschätzten Effekts der Nettomiete pro Quadratmeter zu verzeichnen.

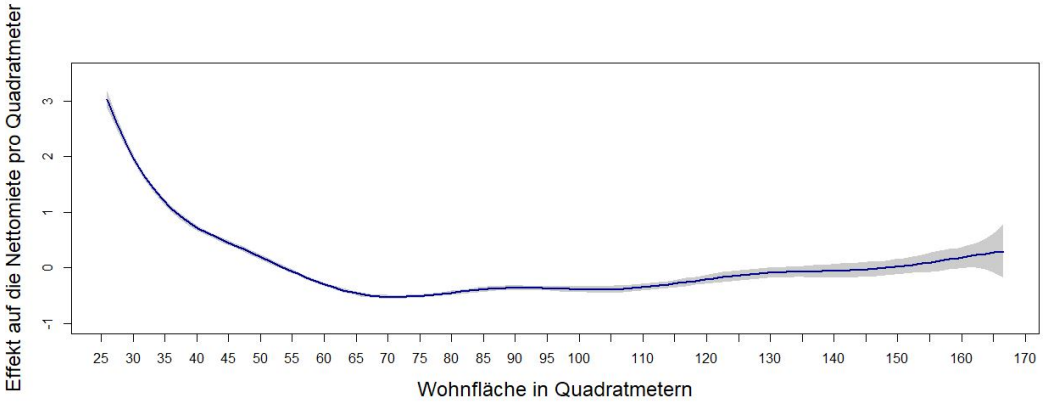

Abbildung 10:
Geschätzter Effekt der Wohnfläche auf die Nettomiete pro Quadratmeter: Die x-Achse gibt die Wohnfläche und die y-Achse den geschätzten Effekt der Wohnfläche auf die Nettomiete pro Quadratmeter an. Der graue Bereich ist der Konfidenzbereich mit jeweils der zweifachen Standartabweichung über und unter der Schätzung.

Anzeigedauer

Der Effekt der Anzeigedauer in Tagen auf die Nettomiete pro Quadratmeter geht nichtlinear in das GAM ein. Wie in Abbildung 10 veranschaulicht, steigt die Nettomiete zunächst bei wachsender Anzeigedauer bis ca. 17 Tage auf ihr globales Maximum, bevor sie langsam bis zum Ausreißer von 75 Tagen sinkt. Bei 75 Tagen wird ein lokales Minimum erreicht, welches jedoch eine höhere Nettomiete aufweist als bei einer Anzeigedauer von null bis vier Tagen. Bis zu einer Anzeigedauer von gut 100 Tagen steigt die Nettomiete erneut an. Bei einer höheren Anzeigedauer sinkt die Nettomiete geringfügig und bleibt über dem lokalen Minimum von 75 Tagen. Der geschätzte Effekt für sehr große Anzeigedauern ist mit Vorsicht zu genießen, da dort verhältnismäßig wenige Beobachtungen vorliegen. Zu Erkennen ist dies auch am breiteren Konfidenzbereich, welcher in Abbildung 10 durch den grauen Bereich dargestellt ist. Insgesamt ist zu erkennen, dass Wohnungen mit identischen Eigenschaften bei einer höheren Anzeigedauer im Schnitt einen höheren Quadratmeterpreis aufweisen. Somit ist die in Kapitel 2.2 beschriebene Theorie, dass Anzeigen mit einer längeren Onlinezeit im Schnitt eine höhere Nettomiete pro Quadratmeter haben, begründet.

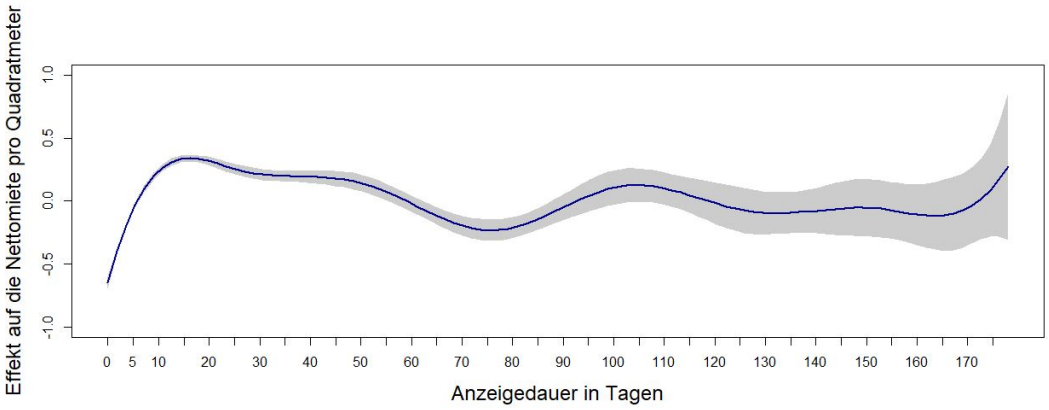

Abbildung 11:
Geschätzter Effekt der Anzeigedauer auf die Nettomiete pro Quadratmeter: Die x-Achse gibt die Anzeigedauer in Tagen und die y-Achse den geschätzten Effekt auf die Nettomiete pro Quadratmeter an. Der graue Bereich ist der Konfidenzbereich mit jeweils der zweifachen Standartabweichung über und unter der Schätzung.

Existenz einer Einbauküche

Bei einer Wohnung ohne Einbauküche und sonst einer gleichen Wohnung, wie beim Intercept, erwartet man durchschnittlich einen um ca. 0,48 Euro niedrigeren Nettoquadratmeterpreis.

Anzahl Balkone und/oder Terrassen

Hat eine Wohnung einen Balkon oder eine Terrasse, erwartet man, bei sonst festgehaltenen Variablen, einen um 0,16 Euro höheren Nettoquadratmeterpreis und bei mehr als einem Balkon oder mehr als einer Terrasse einen um ca. 0,35 Euro höheren Preis.

Wohnlage

Klar zu erkennen ist, dass die Wohnlage durchschnittlich eine große Auswirkung auf die Nettomiete pro Quadratmeter hat. Je zentraler eine Wohnung liegt, desto höher ist im Schnitt der Preis pro Quadratmeter. Im Stadtzentrum erwartet man Mieten mit 15,00 bis 16,50 Euro pro Quadratmeter, wohingegen am weit vom Zentrum entfernten Stadtrand Nettomietpreise von 12,00 bis 13,00 Euro erwartet werden. Zudem fällt auf, dass westlich und süd-süd-westlich vom Zentrum in der Nähe des Zentrums die Preise durchschnittlich höher sind als gleich weit entfernte Gebiete in anderen Richtungen. Der Zusammenhang zwischen Wohnort und Nettomiete pro Quadratmeter ist in Abbildung 12 veranschaulicht:

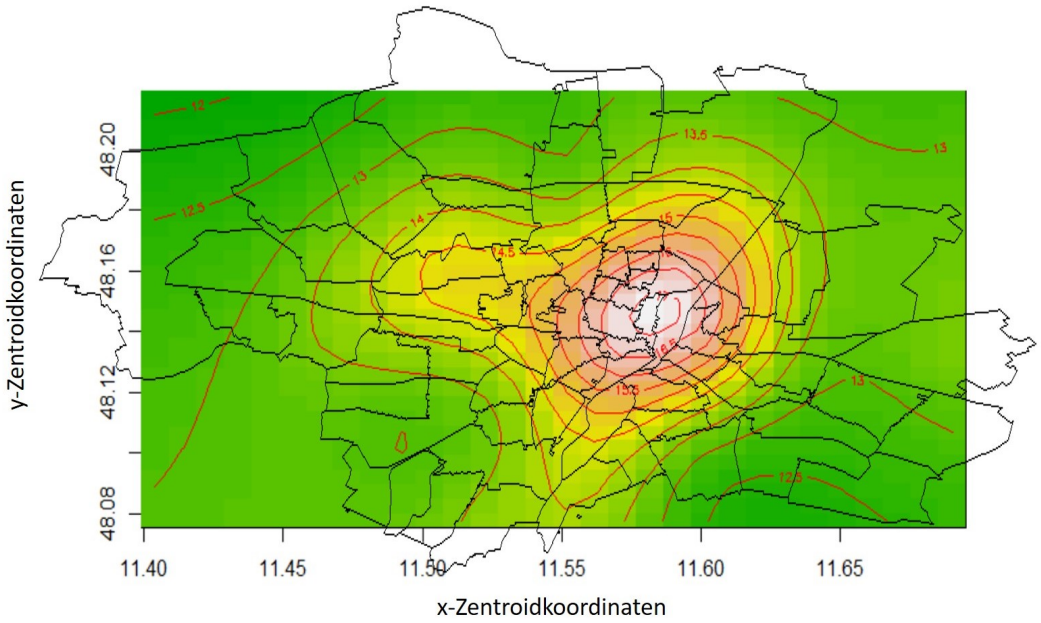

Abbildung 12:
Geschätzter Effekt der Lage der Wohnung in München auf die Nettomiete pro Quadratmeter: München ist in Form einer Karte mit Postleitzahlbereichen abgebildet. Darübergelegt ist der Effekt von den Postleitzahlbereichsschwerpunkten auf die Nettomiete pro Quadratmeter. Dabei ist weiß die teuerste, orange eine weniger teuere, gelb eine noch günstigere Wohngegend und dunkelgrün die günstigste. Die roten Linien stellen Preishöhenlinien dar. Die Zahlen auf diesen geben die durchschnittliche Nettomiete pro Quadratmeter für diesen Bereich an. Durch x- und y-Achse werden die Koordinaten der Postleitzahlbereichszentroide veranschaulicht.

Da dieses Modell zur Beschreibung von Zusammenhängen der vorliegenden Datensituation dient, wird nicht weiter auf die Modelldiagnostik eingegangen. Jedoch ist zu bemerken, dass der Normal-QQ-Plot annähernd auf der Winkelhalbierenden verläuft, was zeigt, dass die Residuen annähernd der Standartnormalverteilung folgen. Der Residuenplot, bei dem die Residuen gegen die geschätzten Werte der Zielgröße aufgetragen werden, ist annähernd strukturlos und die Werte streuen um null. Weder der Mittelwert noch die Varianz der Residuen scheinen von der Zielgröße abzuhängen. Diese Ergebnisse sprechen für ein gut gewähltes Modell. [9, vgl. S.16-20]

5 Datensimulation, Gewichtung und Ergebnisse

Da Anzeigen, die länger online sind, im Schnitt teurer sind und man beim einmaligen oder wöchentlichen Abgreifen von Wohnungen, die länger online sind mit einer höheren Wahrscheinlichkeit zieht, als solche, die kürzer online sind, erhält man somit im Schnitt beim einmaligen oder wöchentlichen Ziehen ein in Richtung einer zu hohen Nettomiete pro Quadratmeter verzerrtes Bild. Um dies zu zeigen werden Datensimulationen durchgeführt, bei denen der Mittelwert der Nettomiete pro Quadratmeter beim einmaligen Ziehen mit dem Monatsmittelwert verglichen wird. Um die Verzerrung zu minimieren, werden Gewichtungen eingesetzt, die abhängig von der Anzeigedauer sind.

5.1 Datensimulation: Einmaliges Abgreifen der Daten und tägliches Abgreifen in einem Monat

Zunächst werden Datensätze simuliert, die Daten vom einmaligen Abgreifen der Anzeigen enthalten. Dafür wird ein Tag ausgewählt und für jede Beobachtung (entspricht einer Anzeige) des aufbereitenen Originaldatensatzes überprüft, ob der gewählte Tag zwischen Start- und Enddatum der Anzeige liegt. Ist dies der Fall, wird die Anzeige zu einem Datensatz hinzugefügt, in dem die Beobachtungen gesammelt werden, die durch Web Scraping an diesem Tag entstanden wären. Da zu erwarten ist, dass die Ergebnisse von Tag zu Tag unterschiedlich sind und der Wochentag des Web Scrapings auch eine Rolle spielen könnte, wird die beschriebene Datensimulation an einem beispielhaften Monat (Mai 2017) für jeden Tag durchgeführt.
Um die Daten des täglichen Abgreifens eines Monats (hier Mai 2017) zu erhalten, wird für jede Beobachtung des aufbereiteten Originaldatensatzes geprüft, ob mindestens ein Tag des Monats zwischen Start- und Enddatum der Anzeige liegt. Falls ja, wird die Beobachtung in den Monatsdatensatz aufgenommen.

5.2 Stichprobentheorie zur Gewichtung

Um die vermutete Verzerrung auszugleichen, werden Gewichtungen eingesetzt. Bei den beim einmaligen Ziehen durch die Simulation erhaltenen Stichproben handelt es sich um designbasierte Stichproben, da eine vor der Stichprobenziehung bekannte Information (Anzeigedauer bzw. bisherige Anzeigedauer) genutzt wird, um die Elemente der Ziehung zu gewichten[12, vgl. S.93ff.].

Die Anzeigedauer wird invers als Gewicht aufgenommen, da eine natürliche Gewichtung ausgeglichen werden soll.
Wenn man das Ziel hat herauszufinden, wie hoch der Nettomietpreis pro Quadratmeter in einer Stadt durchschnittlich ist und man dazu die Daten einer Immobilienwebsite nutzt, möchte man, dass jede Anzeige gleich wahrscheinlich in die Stichprobe gelangt. Durch die unterschiedliche Onlinezeit der Anzeigen zieht man, wenn man zu einer zufälligen Zeit die Anzeigen aufruft, eine Anzeige, die zum Beispiel 7 Tage online ist, 7 mal leichter als eine Anzeige, die nur einen Tag online ist (Kapitel 2.2). Da ein Zusammenhang zwischen Onlinezeit und der Nettomiete besteht, verzerrt dies den Eindruck vom durchschnittlichen Nettomietpreis. Daher möchte man von den Anzeigen, die sieben Tage online sind, 1/7 dieser in die Stichprobe aufnehmen, um den Überschuss auszugleichen und damit für eine solche Anzeige eine Inklusionswahrscheinlichkeit von 1/7 erreichen. Da man in diesem Fall aber alle Anzeigen erhält, die zum Zugzeitpunkt online waren, muss eine Gewichtung durchgeführt werden, sodass die Beispielanzeige nur noch 1/7 von Ihrem ursprünglichen Einfluss auf den Mittelwert hat. Daraus ergibt sich der unter der Annahme $X_i \sim N(\mu, \sigma^2)$ folgender erwartungstreuer Schätzer, wobei d_i die Anzeigedauer der Anzeige i, n den Stichprobenumfang und x die Nettomiete pro Quadratmeter angibt: [12, vgl. S.238ff.]

$$\bar{x}_{gewichtet_{optimal}} = \frac{\sum_{i=1}^{n} \frac{x_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{10}$$

Beweis der Erwartungstreue:
Sei $g_i := \frac{1}{d_i}$.

$$\begin{aligned} E(\bar{X}_{gewichtet}) &= E\left(\frac{\sum_{i=1}^{n} g_i \cdot X_i}{\sum_{i=1}^{n} g_i}\right) = \frac{g_1 \cdot E(X_1) + g_2 \cdot E(X_2) + \cdots + g_n \cdot E(X_n)}{\sum_{i=1}^{n} g_i} \\ &= \frac{g_1}{\sum_{i=1}^{n} g_i} \cdot \mu + \frac{g_2}{\sum_{i=1}^{n} g_2} \cdot \mu + \cdots + \frac{g_n}{\sum_{i=1}^{n} g_n} \cdot \mu = \mu \end{aligned} \tag{11}$$

5.3 Gewichtung

Bei der Simulation für das einmalige Abgreifen sind 31 Datensätze, also für jeden Tag im Mai 2017 einer, entstanden. Für jeden dieser Datensätze werden drei Werte berechnet. Zunächst wird das arithmetische Mittel der Nettomiete pro Quadratmeter mit folgende Formel berechnet:

$$\bar{x}_{=}\frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$

Dabei gibt x_i die Nettomiete pro Quadratmeter für die Anzeige i an und n den Stichprobenumfang.

Anschließend wird der durch die inverse Anzeigedauer gewichtete Mittelwert mit dem im vorherigen Kapitel vorgestellten Schätzer berechnet:

$$\bar{x}_{gewichtet_{optimal}} = \frac{\sum_{i=1}^{n} \frac{x_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{13}$$

Der dritte Wert ist der durch die inverse bisherige Anzeigedauer gewichtete Mittelwert. Dies entspricht nicht der optimalen Gewichtung, dennoch sind diese Ergebnisse für die Praxis sehr relevant, da man beim einmaligen Web Scraping lediglich herausfinden kann, wie lange eine Anzeige bereits online war, aber nicht wie lange sie noch online sein wird. Die Formel ist analog wie die obige Formel mit der Gewichtung durch die inverse Anzeigedauer, jedoch mit der bisherigen Anzeigedauer in Tagen b statt der gesamten Anzeigedauer d:

$$\bar{x}_{gewichtet_{praxis}} = \frac{\sum_{i=1}^{n} \frac{x_i}{b_i}}{\sum_{i=1}^{n} \frac{1}{b_i}} \tag{14}$$

5.4 Ergebnisse

Für jeden Tag im Mai 2017 existiert nun ein Datensatz, wie er entstanden wäre, wenn nur an diesem Tag ein Web Scraping durchgeführt worden wäre. Für jeden Datensatz wurde das arithmetische Mittel, der nach der inversen Anzeigedauer und nach der inversen bisherigen Anzeigedauer gewichtete Mittelwert berechnet. Die Werte werden für jeden Tag in Abbildung 13 veranschaulicht. Hätte man sich zum Beispiel für den ersten Mai 2017 für das Web Scraping entschieden, hätte man für die Nettomiete pro Quadratmeter einen Durchschnittswert von ca. 16,36 Euro erhalten (Da die Simulation auf den aufbereiteten Daten basiert, wäre der am ersten Mai 2017 erhaltene Wert geringfügig anders. Auf diese Anmerkung wird in weiteren Fällen verzichtet.). Mit ca. 16,19 Euro liegt der Wert für den mit der Anzeigedauer

gewichtete Mittelwert darunter. Auch der Schätzer mit der bisherigen Anzeigedauer ergibt einen niedrigeren Wert mit ca. 16,22 Euro. Bildet man nun den Mittelwert über jeden Mittelwertstyp erhält man folgende Werte, welche auch in Abbildung 13 dargestellt sind: Für 31 einmalige Web Scrapings im Mai 2017 erhält man im Schnitt einen Nettomietpreis pro Quadratmeter von ca. 16,37 Euro. Gewichtet man die Mittelwerte der Nettomiete pro Quadratmeter jedes Web Scraping mit der Anzeigedauer, beträgt die Nettomiete pro Quadratmeter durchschnittlich 16,02 Euro und bei Gewichtung mit der bisherigen Anzeigedauer ca. 16,18 Euro.

Tag	1	2	3	4	5	6	7	8
arith.Mittel	16.36490	16.32958	16.27594	16.27996	16.24738	16.30212	16.31146	16.34947
gewichtet.Anzeigedauer	16.19262	15.95116	16.02110	15.95882	15.89258	15.89630	16.16633	16.33113
gewichtet.bisherige.A.	16.21534	16.04114	16.09140	16.04804	16.10685	16.29256	16.27436	16.36961
Tag	9	10	11	12	13	14	15	16
arith.Mittel	16.27870	16.27662	16.32830	16.32792	16.34900	16.36406	16.34276	16.36158
gewichtet.Anzeigedauer	15.93723	15.93355	16.15001	15.70712	15.89385	16.17077	16.08352	16.03154
gewichtet.bisherige.A.	15.98501	16.20754	16.37744	16.36166	16.24062	16.38075	16.27437	16.24858
Tag	17	18	19	20	21	22	23	24
arith.Mittel	16.35908	16.38977	16.36342	16.36768	16.45724	16.47162	16.44411	16.51448
gewichtet.Anzeigedauer	15.92857	15.85644	15.87042	16.01089	16.59926	16.23556	15.50599	16.02604
gewichtet.bisherige.A.	16.08722	16.15627	16.23979	16.04343	16.45915	16.24187	15.91891	16.50955
Tag	25	26	27	28	29	30	31	Durchschnitt
arith.Mittel	16.50423	16.53251	16.48280	16.44391	16.38911	16.32262	16.38249	16.37145
gewichtet.Anzeigedauer	16.53506	16.40716	16.28037	15.97072	15.60401	15.84398	15.58714	16.01868
gewichtet.bisherige.A.	16.32394	16.43930	16.27028	15.95600	15.71803	15.79311	16.05529	16.18476

Abbildung 13:
Mittelwerte der Nettomiete pro Quadratmeter beim einmaligen Web Scraping: Angegeben werden das arithmetische Mittel, das mit der inversen Anzeigedauer gewichtete Mittel und das durch die inverse Anzeigedauer bis zu diesem Tag gewichtete Mittel sowie der Durchschnitt für diese für jeden Tag im Mai 2017, wenn nur an dem jeweiligen Tag Daten abgegriffen worden wären.

Somit führt die Gewichtung zu einem niedrigeren Nettomietpreis. Dass es sich dabei tatsächlich um eine Minderung der Verzerrung handelt, kann man erkennen, wenn man die Werte mit dem Mittelwert aller Anzeigen, die man beim täglichen Web Scraping im Mai 2017 erhält, vergleicht. Der Mittelwert für den ganzen Mai beträgt ca. 16,13 Euro und liegt somit unter dem arithmetischen Durchschnittswert von den einmaligen Abgreifungen. Somit

ist zu erkennen, dass das tägliche Abgreifen zu zu hohen Werten führt. Beide gewichteten Werte liegen näher am Monatsmittelwert als das arithmetische Mittel für die einzelnen Tage und führen somit zu einer Verbesserung. Dennoch ist zu erkennen, dass auch der als optimal gewichtet angenommene Wert den Monatsmittelwert verfehlt. Da viele Anzeigen länger als einen Monat online sind, könnte auch der Monatsmittelwert verzerrt sein, weswegen ein geringerer Wert beim durch die inverse Anzeigedauer gewichteten Mittelwert plausibel ist. Der Mittelwert über die durch die bisherige Anzeigedauer gewichteten Nettomieten für das einmalige Abgreifen, liegt über dem Monatsmittelwert. Ein Erklärungsansatz hierfür ist, dass durch weniger Information bei der Gewichtung diese schwächer ausfällt. Nichts desto trotz führt diese praktisch anwendbare Gewichtung zu einer Verbesserung.
Führt man die Simulation und die Gewichtung auch für jeden anderen Monat des Jahres 2017 durch, kann man erkennen, dass bis auf einen Fall die Schätzwerte mit Gewichtungen näher am Mittelwert des Monats liegen und somit zu einer Verminderung der Verzerrung führen. Zudem ist in fast jedem Monat zu erkennen, dass der mit der inversen bisherigen Anzeigedauer gewichtete Wert näher am Monatsmittelwert liegt als der mit der inversen Anzeigedauer gewichtete Wert. Die Werte werden in Abbildung 14 dargestellt.

2017	Januar	Februar	März	April	Mai	Juni	Juli
1	15.88729	16.07934	16.06407	16.17089	16.12715	16.32420	16.30606
2	16.09063	16.28748	16.23209	16.42099	16.37145	16.49748	16.53292
3	15.79175	15.93088	15.85252	16.07228	16.01868	16.20953	16.18000
4	15.93541	16.08500	16.09832	16.18961	16.18476	16.40637	16.31143

2017	August	September	Oktober	November	Dezember
1	16.10892	16.24921	16.32260	16.40289	16.45096
2	16.39005	16.42056	16.50940	16.61637	16.50350
3	16.02202	16.11763	16.23793	16.35338	16.37598
4	16.12710	16.40214	16.36282	16.42099	16.50699

Abbildung 14:
Durchschnittliche Mittelwerte der Nettomiete pro Quadratmeter pro Monat beim einmaligen Web Scraping: Angegeben werden das arithmetische Mittel (2), das mit der inversen Anzeigedauer gewichtete Mittel (3) und das durch die inverse Anzeigedauer bis zu diesem Tag gewichtete Mittel (4) über die einmaligen Web Scrapings für jeden Tag, sowie das arithmetische Mittel bei täglichem Abgreifen im jeweiligen Monat (1).

6 Zusammenfassung

Zusammenfassend konnte die Verzerrung der Daten beim wöchentlichen oder einmaligen Abgreifen von Anzeigen von einer Immobilienwebsite bestätigt werden, da Wohnungen mit einer kürzeren Onlinezeit im Schnitt eine günstigere Nettomiete pro Quadratmeter aufweisen (Kapitel 4.2) und Wohnungen, die lange online sind, wahrscheinlicher in die Stichprobe gelangen, als Wohnungen, die kurz online sind (Kapitel 5.2). Das einmalige Abgreifen der Anzeigen wurde anhand von 365 Tagen (für alle Tage in 2017) simuliert, sowie für das tägliche Abgreifen jeden Monat in 2017. Für die Daten, die beim einmaligen Abgreifen der Anzeigen entstanden wären, wurden drei Werte berechnet: Das verzerrte arithmetische Mittel und das mit zwei verschiedenen Gewichten adjustierte Mittel. Ein Gewicht entspricht der inversen Onlinezeit in Tagen, da davon ausgegangen wird, dass eine Anzeige, die x Tage online ist, x-mal eher in die Stichprobe gelangt als eine Anzeige, die nur einen Tag online ist und durch das anschließende Gewichten mit 1/x dieser Effekt ausgeglichen wird. Da man beim einmaligen Abgreifen der Daten erfassen kann, wie viele Tage die Anzeigen bereits online sind, aber nich wie lange sie noch online sein werden, wird die inverse bisherige Anzeigedauer als zweite Gewichtung verwendet. Nachdem diese Werte für jeden Tag berechnet wurden, wurde der Monatsdurchschnitt für das arithmetische Mittel sowie für die gewichteten Mittel berechnet und mit dem jeweiligen Monatsmittelwert der Nettomiete pro Quadratmeter für das tägliche Abgreifen im dem Monat als einfache Stichprobe verglichen. Der Vergleich dieser Werte für alle Monate im Jahr 2017 zeigt, dass der Monatsmittelwert der Nettomiete pro Quadratmeter stets günstiger ist als das arithmetische Mittel vom einmaligen Abgreifen im Monatsdurchschnitt. Beide Gewichtungen führen dazu, dass der Mittelwert der einmaligen Ziehungen näher am Monatsmittelwert liegt. Somit wird die Verzerrung gemindert. Auffällig ist jedoch, dass der durch die inverse bisherige Anzeigedauer gewichtete Schätzer in der Regel näher am Monatsdurchschnitt liegt, als der durch die inverse Anzeigedauer gewichtete.
Insgesamt wird empfohlen eine Gewichtung des Mittelwerts der Nettomiete pro Quadratmeter mit der inversen bisherigen Anzeigedauer durchzuführen, wenn nur ein einmaliges oder nicht tägliches (zum Beispiel wöchentliches) Abgreifen von Wohnungen einer Onlineseite möglich ist.

7 Ausblick

Auch wenn nun eine Verzerrungsquelle gemindert werden konnte, bilden die Daten, die durch nicht tägliches Abgreifen der Immobiliendaten stammen, nicht zwingend die wahre Mietsituation Münchens ab. Weitere Verzerrungsquellen könnten zum Beispiel Preisverhandlungen sein. Denkbar wäre, dass der Mietpreis in Persona heruntergehandelt wird und die Wohnungen im Schnitt günstiger vermietet werden als durch die Onlineanzeige angegeben. Auch fasst das Abgreifen einer Website nur Wohnungen, die dort eingestellt sind. Wohnungen, die innerhalb der Familie oder unter Freunden günstiger als für den durchschnittlichen Preis vermietet werden, werden nicht erfasst. Weiter ist denkbar, dass der Wochentag des Web Scrapings einen Einfluss auf die Daten hat.
Insgesamt sollte die Datenerhebung nicht unterschätzt und möglichst sauber durchgeführt werden. Zudem sollten mögliche Verzerrungen analysiert und, wenn möglich durch die Art der Datenerhebung, ausgegelichen werden. Ist dies nicht möglich, sollten Verzerrungen durch Gewichtungen reduziert oder sogar ausgeglichen werden.

8 Literatur- und Abbildungsverzeichnis

Literatur

[1] unbekannter Verfasser. *Einführung in die Stichprobentheorie.* Projekt Neue Statistik 2003, Freie Universität Berlin, 2003.

[2] unbekannter Verfasser. *Probleme einer Datenerhebung.* Julius-Maximilians-Universität Würzburg, unbekanntes Jahr. URL:https://www2.uni-wuerzburg.de/dmuw-vhb/demo/statistik/Popups/exkurs12.html
Aufgerufen am: 18.05.2018.

[3] Vargiu Eloisa & Urru Mirko. *Exploiting web scraping in a collaborative filteringbased approach to web advertising.* Artificial Intelligence Research, 2013, Vol. 2, No. 1, 2013.

[4] *Immobilienscout 24, 2018. URL: https://www.immobilienscout24.de/*
Aufgerufen am: 20.04.2018.

[5] unbekannter Verfasser. *Münchens Einwohnerentwicklung: Steigende Bevölkerungszahlen.* Referat für Stadtplanung und Bauordnung, 2015. URL:https://www.muenchen.de/rathaus/Stadtverwaltung/Referat-fuer-Stadtplanung-und-Bauordnung/Stadtentwicklung/Grundlagen/demografie.html
Aufgerufen am: 22.05.2018.

[6] Nau Niklas & Klühspieß Anna. *Wohnungsnot in Bayern: Können wir uns Wohnen noch leisten?* BR, 2018. URL:https://www.br.de/nachrichten/wohnungsnot-wohnen-koennen-wir-uns-das-noch-leisten-100.html
Aufgerufen am: 22.05.2018.

[7] Wood Simon. *Generalized Additive Models - An Introduction with R.* Chapman & Hall/CRC, 2006.

[8] Fahrmeir Ludwig & Kneib Thomas & Lang Stefan. *Regression - Modelle, Methoden und Anwendungen.* Springer-Verlag, Berlin/Heidelberg, 2009.

[9] Singer Andreas. *Vergleich von multiplikativen und additiven Mietspiegeln.* Ludwig-Maximilians-Universtität München, Bachelorarbeit, 2015.

[10] Böhmer Valentin. *1 Semi- und Nonparametrische Regression (I).* Docplayer, 2018. URL:http://docplayer.org/70575514-1-semi-und-nonparametrische-regression-i.html
Aufgerufen am: 21.05.2018.

[11] Wood Simon. *gam: Generalized additive models with integrated smoothness estimation.* RDocumentation, 2018. URL:https://www.rdocumentation.org/packages/mgcv/versions/1.8-23/topics/gam
Aufgerufen am: 22.04.2018.

[12] Kauermann Göran & Küchenhoff Helmut. *Stichproben: Methoden und praktische Umsetzung in R.* Springer-Verlag, Berlin/Heidelberg, 2011.

Abbildungsverzeichnis

9 Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorgelegte Bachelorarbeit eigenständig und ohne fremde Hilfe verfasst und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Bachelorarbeit ist in dieser oder einer ähnlichen Form in keinem anderen Kurs und/oder Studiengang als Studien- oder Prüfungsleistung vorgelegt worden.

München, 20.06.2018

Jessica Peter