TECHNISCHE UNIVERSITÄT WIEN

Diploma Thesis

Spatial Effect in Landpricemodels

submitted in satisfaction of the requirements for the degree of
Diplom-Ingenieur
of the TU Wien, Faculty of Architecture and Planning

Diplomarbeit

Räumliche Effekte in der Bodenpreismodellierung

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Diplom-Ingenieurs
eingereicht an der Technischen Universität Wien, Fakultät für Architektur und Raumplanung

von

Moritz Starzer, BSc

Matr.Nr.: 01227314

unter der Anleitung von

ao.Univ.Prof. Dipl.-Ing. Dr. **Wolfgang Feilmayr**

Univ.Ass. Dipl.-Ing. **Robert Kalasek**

Fakultät für Architektur und Raumplanung
Fachbereich für Stadt- und Regionalforschung
Technische Universität Wien
Karlsplatz 13/E280-2, 1040 Wien

Wien, im Jänner 2019 _____

Danksagung

...

Kurzfassung

Ziel der Arbeit ist es, mithilfe eines quantitativ-statistischen Modells die Prozesse und Faktoren, die den durchschnittlichen Bodenpreis für Bauland der österreichischen Gemeinden beeinflussen, zu untersuchen. Hierfür liegen quantitative Datensätze zu 1667 Gemeinden vor.

Die Lage im Raum spielt für die Erklärung von Bodenpreisen ohne Zweifel eine wichtige Rolle. Es handelt sich somit um Daten mit starkem räumlichen Bezug. Bei der Modellierung solcher räumlichen Sachverhalte können deswegen räumliche Effekte (spatial effects) auftreten, die berücksichtigt werden müssen. Modellansätze aus der räumlichen Ökonometrie (spatial econometrics), insbesondere spatial-autoregressive-models, und Methoden aus der Geostatistik, insbesondere kriging-Methoden, eignen sich, um räumliche Effekte in der Modellentwicklung in Betracht zu ziehen.

Beim Vergleich dieser räumlichen Modellansätze mit klassischen nicht-räumlichen Modellansätzen kann gezeigt werden, dass der Erklärungsgehalt der Modelle durch die Einbeziehung räumlicher Effekte erheblich verbessert wird. Der Prozess, der die Bodenpreise in Österreich generiert, kann somit als räumlicher Prozess verstanden werden.

Abstract

folgt...

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Problemstellung

Einerseits ist der Bodenpreise eine der wichtigsten Grundlagen nach denen räumlich relevante Entscheidungen getroffen werden. Er ist maßgeblich für die Standortentscheidungen von Unternehmen beziehungsweise Haushalten und spiegelt in weiterer Folge die Standortqualität wieder. Andererseits ergibt sich dadurch indirekt eine gewaltige Steuerungsmöglichkeit für die räumliche Planung. Durch die indirekte Beeinflussung des Bodenpreises können so räumliche Entscheidungen beeinflusst werden.

Um diese Steuerungsmöglichkeit zu nutzen, müssen allerdings die Prozesse und Faktoren die den Bodenwert beeinflussen, verstanden werden. Es erscheint sinnvoll, quantitativ-statistische Modelle zu entwickeln, die diese Prozesse und Faktoren beleuchten und verständlich machen.

Die Position im Raum spielt für Bodenpreise ohne Zweifel eine wichtige Rolle. Es handelt sich somit um Daten mit stark räumlichem Bezug. Bei der Modellierung solcher Daten können deswegen räumliche Effekte (*spatial effects*) auftreten, die berücksichtigt werden müssen (vgl. Anselin 1988; LeSage und Pace 2009). In der Literatur werden verschiedene Ansätze aus verschiedenen Forschungsrichtungen unterschieden um räumliche Effekte in Betracht zuziehen (vgl. Jahanshiri et al. 2011). In dieser Arbeit sollen Ansätze aus der Geostatistik und der räumlichen Ökonometrie angewendet verglichen werden.

1.2 Forschungsfragen

Untersuchungseinheiten der Analyse sind die österreichischen Gemeinden. Für sie soll der durchschnittliche Bodenpreis für Bauland modelliert werden, um die Prozesse die den Bodenpreis generieren offen zu legen. Die zentralen Hauptfragestellungen der Arbeit lassen sich in eine inhaltliche und eine methodische Ebene gliedern.

Inhaltlich:

- Durch welche Faktoren lässt sich der Prozess, der den durchschnittlichen Bodenpreis in den österreichischen Gemeinden generiert, beschreiben?

Methodisch:

- Welche Herausforderungen ergeben sich bei der Modellierung von Sachverhalten mit räumlichen Bezug?

- Können Ansätze der räumlichen Ökonometrie und der Geostatistik eine Verbesserung des Erklärungsgehalts beziehungsweise der Vorhersagegenauigkeit von Bodenpreismodellen bewirken?

1.3 Aufbau der Arbeit

Die Arbeit gliedert sich in mehrere Abschnitte. Im nächsten Abschnitt 2 werden die theoretischen Grundlagen erarbeitet. Hier wird auf die Modellierung allgemein und auf hedonische Preismodelle im Besonderen eingegangen. Anschließend werden aus theoretischen Überlegungen sowie aus der Literatur Einflussfaktoren des Bodenpreises herausgearbeitet und Hypothesen abgeleitet. Des Weiteren wird auf die Besonderheiten von Daten mit räumlichem Bezug eingegangen. Es wird aufgezeigt, was räumliche Effekte sind und warum es wichtig ist diese in einem Modell zu berücksichtigen. Am Ende diese Abschnitts werden die, in der methodischen Literatur verbreiteten Ansätze räumliche Effekte in die Modellbildung miteinzubeziehen, beschrieben und argumentiert warum Methoden aus der räumlichen Ökonometrie und der Geostatistik bei der Bodenpreismodellierung zielführend sind.

Im Abschnitt 3 werden die zuvor argumentierten methodischen Ansätze genauer beschrieben, die in der Datenanalyse und der Modellbildung im Abschnitt 5 zur Anwendung kommen. Der Abschnitt 4 beschäftigt sich mit den verwendeten Daten. Insbesondere werden die Datenquellen und die notwendigen Schritte der Datenaufbereitung aufgezeigt. Zusätzlich werden mittels Methoden der deskriptiven Statistik die Daten beschrieben. Abschließend fasst Abschnitt 6 die Erkenntnisse der Arbeit zusammen.

Kapitel 2

Grundlagen

In diesem Abschnitt ist zunächst eine sachliche, zeitliche und räumliche Systemabgrenzung notwendig. Wie bereits erwähnt, dienen die österreichischen Gemeinden als räumliche Analyseeinheit. Zu diesen stehen einerseits aktuelle Bodenpreisdaten für Bauland und andererseits aktuelle sozioökonomische Daten zur Verfügung. Es handelt sich somit um eine Querschnittsanalyse (*cross-sectional data*). Die genaue Beschreibung der verwendeten Daten erfolgt im Kapitel 4 auf Seite 39.

Dieser Abschnitt gliedert sich in drei Kapitel. Im ersten Kapitel 2.1 wird auf die Modellierung allgemein und auf hedonische Modelle im Speziellen eingegangen. Im zweiten Kapitel 2.2 werden mögliche Einflussfaktoren (Variablen) auf den Bodenpreis aus der Literatur herausgearbeitet und Hypothesen abgeleitet. Das dritte Kapitel 2.3 beschäftigt sich mit den Besonderheiten von Daten mit räumlichem Bezug (*spatial data*), räumlichen Effekten (*spatial effects*) und deren Berücksichtigung in der Modellbildung. Unter räumlichen Effekten werden Effekte der räumlichen Abhängigkeit und der räumlichen Heterogenität zusammengefasst. Es stehen in der methodischen Literatur mehrere Ansätze zur Verfügung, räumliche Effekte zu berücksichtigen. Es wird aufgezeigt, dass im Kontext der Bodenpreismodellierung Methoden der räumlichen Ökonometrie und der Geostatistik zielführend erscheinen, um räumliche Effekte in der Modellbildung zu berücksichtigen.

2.1 Modellierung

Es sind komplexe reale Systeme und Prozesse, die den Pries des Bodens bestimmen. Modelle sind allgemein ein Versuch, diese in der Realität bestehenden Systeme abzubilden. Sie stellen somit ein Hilfsmittel dar, die Wirklichkeit zu erklären, zu verstehen und vorherzusagen (Haavelmo 1944, S.1). Um die Realität abzubilden, muss sie allerdings zweckorientiert vereinfacht (abstrahiert) werden. Diese Tatsache brachte Box und Draper (1986, S.424) zu der Aussage:

> *Essentially, all models are wrong, but some are useful.*

Ziel eines Modells muss es demnach sein, die Realität so zu vereinfachen, dass die wesentlichen Prozesse abgebildet sind. Ein in einem konkreten Fall entwickeltes Modell ist von einer Vielzahl von Faktoren abhängig. Besonders maßgebend dafür sind die zu beantwortende Fragestellung und die zur Verfügung stehenden Daten (quantitativ oder qualitativ). Für diese Arbeit stehen quantitative sozioökonomische Daten und Bodenpreisdaten der österreichischen Gemeinden zu Verfügung. Es soll in weiterer Folge der Prozess, der den Preis des Bodens bestimmt, quantitativ-statistisch modelliert werden.

Im Modellierungsprozess kann grundsätzlich zwischen theoretischen und empirischen Modellen unterschieden werden. Das theoretische Modell beschreibt den zu modellierenden Prozess inhaltlich-formal. Im Kontext dieser Arbeit geschieht dies in Kapitel 2.2. Hier werden vermutete Zusammenhänge a priori aus der Literatur herausgearbeitet und mögliche erklärende Variablen und Hypothesen definiert.

Das empirische Modell entsteht durch die zur Verfügung stehenden Daten und die angewendeten Schätzmethoden. Am weitesten verbreitet sind Regressionsmodelle. Im einfachsten Fall werden lineare Regressionsmodelle mit einer abhängigen (endogenen) und einer oder mehreren erklärenden (exogenen) Variablen geschätzt. Ziel ist es, den Einfluss von Sachverhalten (erklärende Variablen) auf eine Zielvariable zu ermitteln. Im Kontext von Bodenpreisen könnten die erklärenden Variablen beispielsweise die Erreichbarkeit einer Gemeinde abbilden und deren Erklärungsgehalt für Bodenpreise bestimmt werden. Eine zugrundeliegende Idee dieser Vorgangsweise ist die hedonische Theorie, auf die anschließend eingegangen wird. In dieser Arbeit dienen einfache lineare Regressionsmodelle als Ausgangspunkt für komplexere Modelle.

Bei der Auswahl von erklärenden Variablen kann grundsätzlich zwischen einem deduktiven und einem induktiven Ansatz unterschieden werden. Beim deduktiven Ansatz kann auf das aus der Theorie abgeleitete theoretische Modell zurückgegriffen werden. Aus diesem geht hervor, welche Variablen einen möglicherweise hohen Erklärungsgehalt haben und deshalb ins empirische Modell einfließen sollen. In Bezug auf Bodenpreise kann beispielsweise angenommen werden, dass sich eine gute Erreichbarkeit einer Gemeinde positiv auf den Bodenpreis auswirkt. Beim induktiven Ansatz wird von den Daten der einzelnen Beobachtungen ausgegangen. Es wird mithilfe statistischer Schätz- und Testtheorie versucht, Variablen mit hohem Erklärungsgehalt zu finden. In der Praxis, so wie auch in dieser Arbeit, ist es meist sinnvoll, beide Ansätze zu kombinieren. So schrieb bereits Keynes et al. (1890, S.164):

> *All induction is blind, so long as the deduction of causal connections is left out of account; and all deduction is barren so long as it does not start from observation.*

Dieses Zitat verdeutlicht einen häufigen Zielkonflikt bei der Modellierung. Auf der einen Seite soll der Modell-Fit eines Modells hoch sein (induktiver Zugang). Auf der anderen Seite sollen die Interpretierbarkeit und eine gewisse Theorienkonformität gegeben sein (deduktiver Zugang). In diesem Zusammenhang kann auch von verschiedenen Dimensionen der Modellgültigkeit gesprochen werden. Man unterscheidet theoretische, formale, empirische und pragmatische Gültigkeit. Theoretische Gültigkeit bezieht sich auf die Übereinstimmung des theoretischen Modells mit dem empirischen Modell - beispielsweise bezüglich der Beziehungen (Vorzeichen) der Variablen. Formale Gültigkeit beschreibt einerseits, ob ein Modell den statistischen Annahmen bei der Schätzung entspricht sowie die bereits angesprochene Modellperformance (Modell-Fit). Empirische Gültigkeit bezieht sich auf den Verglich zwischen dem Modell und dem abgebildeten System. Pragmatische Gültigkeit beschreibt, ob ein Modell dem Untersuchungszweck angemessen ist. Im Idealfall erfüllt ein Modell alle vier Dimensionen der Modellgültigkeit.

Literaturverweis hinzufügen

Hedonische Preistheorie

Das Konzept der hedonischen Preise ist, dass die Preise heterogener Güter durch die Summe der Eigenschaften des Gutes beschreibbar sind (Rosen 1974, S.34). Umgelegt auf Bodenpreise bedeutet das nun, dass man nicht direkt für das Gut „Boden“, sondern für dessen Eigenschaften beziehungsweise dessen Umfeld und Möglichkeiten eine gewisse Zahlungsbereitschaft hat. Anders

formuliert ergibt sich die Haupthypothese, dass der Wert des Bodens vom sozioökonomischen Umfeld, in dem er sich befindet, abhängig ist (Giuliani 2002, S.2).

Bei der Methode der hedonischen Preisen, die den *revealed-prference* Analysen (indirekte Verfahren) zugeordnet werden kann, geht es also darum, Präferenzen von Marktteilnehmern für gewisse Eigenschaften eines Gutes zu messen. Sie kann den Wirtschaftswissenschaften (Ökonometrie) zugerechnet werden. Die Ökonometrie versteht sich als Zusammenspiel zwischen Statistik, Mathematik und Ökonomischen Theorien. Ihr Ziel ist es, Zusammenhänge und Phänomene der realen Welt empirisch zu hinterlegen (Maddala und Lahiri 1992, S.3f).

Über die erstmalige Verwendung von hedonischer Regression herrscht kein Konsens. Die wahrscheinlich wichtigsten theoretischen Fundamente wurden aber von Lancaster (1966) und Rosen (1974) erarbeitet (Herath und Maier 2010, S.2). Lancaster (1966, S.133) formulierte:

> *The chief technical novelty lies in breaking away from the traditional approach that goods are the direct objectives of utility and, instead, supposing that it is the properties or characteristics of goods from which utility is derived.*

Hier wird ersichtlich, dass es um die nutzenstiftenden Eigenschaften eines heterogenen Gutes geht. Diese einzelnen nutzenstiftenden Eigenschaften sind messbar, bewertbar und determinieren den Gesamtpreis. Sie sind allerdings aufgrund ihres Wesens unteilbar und nur im Bündel greifbar (Rosen 1974, S.36). Mathematisch kann die oben definierte Überlegung wie folgt ausgedrückt werden (Wieser 2006, S.5):

$$P = P(\mathbf{z}) \tag{2.1}$$

Wobei P den Preis und $z = (z_1, z_2, ..., z_k)$ einen Vektor der verschiedenen Eigenschaften darstellt. In Bezug auf Bodenpreise von Gemeinden könnte eine dieser Eigenschaften z_i beispielsweise für die Erreichbarkeit der Gemeinde stehen. Auf diese einzelnen Eigenschaften bzw. Attribute wird im nächsten Kapitel 2.2 genauer eingegangen.

Leitet man die Preisfunktion 2.1 partiell nach den einzelnen Eigenschaften z_i ab, ergeben sich die impliziten Preisfunktionen für die jeweilige Eigenschafte eines Gutes (Wieser 2006, S.5):

$$p_{x_i}(x_i; x_{\text{-}i}) = \frac{\partial p(x)}{\partial x_i} \text{ ; (i = 1 bis k)} \tag{2.2}$$

Diese implizite Preisfunktion gibt an, um wieviel sich der Pries eines Gutes verändert, wenn sich eine Eigenschaft z_i marginal verändert und gleichzeitig alle anderen Eigenschaften $z_{\text{-}i}$ unverändert bleiben. So kann dann der Einfluss einer bestimmten Eigenschaft auf den Pries eines Gutes bestimmt werden.

Es müssen allerdings gewisse Annahmen getroffen werden, um die Methode der hedonischen Preise anzuwenden. Nach diesen Annahmen versuchen alle Marktteilnehmer ihren Nutzen zu maximieren und handeln rational. Marktteilnehmer treffen sich auf dem Markt, wo der klassischen ökonomischen Theorie zufolge Angebot und Nachfrage zusammenkommen und ein Gleichgewichtspreis entsteht. Des Weiteren wird angenommen, dass vollständige Transparenz herrscht und somit Anbieter und Nachfrager über alle marktrelevanten Informationen verfügen. Ebeso werden externe Effekte vernachlässigt (Rosen 1974, S.35).

In der Regel entsprechen diese Modellannahmen nicht der Realität. Dennoch wird die Methode der hedonischen Preise in vielen Bereichen angewandt. Beispielsweise in der Verbrauchermarktforschung, in der Kalkulation von Verbraucherpreisindices oder in der Bewertung von Autos.

Sehr häufige Anwendung findet sie allerdings in der Immobilien und Grundstücksbewertung (Herath und Maier 2010, S.1f). Es kann festgehalten werden, dass sich der Ansatz der hedonischen Preise gut für die Analyse des Immobilien- und Grundstücksmarkts eignet (Haase 2011, S.159; Giuliani 2002, S.85; Herath 2013, S.1353). Im nächsten Kapitel 2.2 wird dieser Ansatz weiter verfolgt und versucht für das theoretische Modell Variablen zu identifizieren die, die nutzenstiftenden Eigenschaften des Bodens an einem gewissen Standort und damit den Bodenpreis wiederspiegeln.

2.2 Einflussfaktoren des Bodenwerts

Boden bekommt als einer der drei klassischen Produktionsfaktoren (neben Arbeit und Kapital) seit jeher vor allem in der wirtschaftswissenschaftlichen Literatur eine große Bedeutung zugemessen. Die Betrachtung des Bodens als Produktionsfaktor, ist Ausgangspunkt für die theoretische Ableitung von Einflussfaktoren des Bodenwerts. Der klassischen ökonomischen Theorie zufolge wird der Wert eines Gutes durch Angebot und Nachfrage bestimmt. Dies gilt grundsätzlich auf für Boden. Bevor allerdings Angebot und Nachfrage des Produktionsfaktors Boden herausgearbeitet werden können, müssen zuerst dessen Besonderheiten und Begrifflichkeiten erläutert beziehungsweise geklärt werden. Diese sehr von der klassischen ökonomischen Theorie geleitete Sichtweise soll dabei helfen, zu verstehen, welche Faktoren den Bodenpreis beeinflussen.

2.2.1 Besonderheiten des Produktionsfaktors Boden

Schon Adam Smith definierte die drei Produktionsfaktoren Arbeit, Kapital und Boden die wichtig sind, für die Produktion von Gütern und Dienstleistungen. Diese sind in der klassischen Ökonomie knappe Güter und haben deshalb einen Preis. Dieser ist im Fall von Boden der Bodenpreis, Bodenwert oder die Bodenrente. Diese drei Begriffe werden meist als Synonyme verwendet. Allerdings gilt es zu beachten, dass vor allem bei dem Begriff der Bodenrente keine einheitliche Definition aus der Literatur abzuleiten ist (Giuliani 2002, S.46).

Boden hat, im Gegensatz zu den anderen klassischen Produktionsfaktoren, besondere Eigenschaften. Diese sind Unvermehrbarkeit, Unentbehrlichkeit und Immobilität (Giuliani 2002, S.25). Im Bezug auf Unvermehrbarkeit ist zu beachten, dass zwar die Gesamtmenge an Boden natürlich begrenzt wird, die Art der Nutzung aber verändert werden kann. So kann beispielsweise neues Bauland zu Ungunsten von Grünland ausgewiesen werden, um das Angebot an Bauland zu erhöhen. Unentbehrlich ist Boden, da er für die menschliche Lebensentfaltung, beispielsweise Nahrungsmittelproduktion, unabdingbar ist. Zusätzlich ist der Boden in seiner Lage fixiert. Die Position im Raum spielt also eine wichtige Rolle.

Diese Besonderheiten führen zu Nutzungskonflikten um bestimmte Standorte und zu einem begrenzten Angebot von Boden innerhalb eines Betrachtungsraumes. Man spricht von einer unelastischen Angebotskurve (Giuliani 2002, S.45f).

2.2.2 Bodennachfrage

Da das Angebot an Boden wie beschrieben gegeben und eher unelastisch ist, kann daraus allgemein abgeleitet werden, dass die Bodennachfrage maßgeblich für den Bodenpreis ist. Eine erhöhte Nachfrage an Boden bedeutet nun einen erhöhten Bodenpreis. Wie sich eine solche

Nachfrage ergibt und in weiterer Folge welche Faktoren bei der Standortwahl von Individuen und Gruppen eine Rolle spielen, sind klassische Fragestellungen in der Regionalwissenschaft beziehungsweise der Wirtschaftsgeographie als verwandte Disziplinen (Kulke 2017, S.23-58). In diesem Zusammenhang sind die sehr frühen Überlegungen von Thünen (1826), aber auch jene von Weber (1909) und das Grundrentenmodell von Alonso (1960) zu nennen. Sie alle untersuchten das Standortverhalten von Akteuren im Raum, die ihren jeweiligen Nutzen maximieren wollen und schließen dadurch auf Standortnutzungen in Abhängigkeit von Standortfaktoren. Dem Begriff der Standortfaktoren kommt also bei der Definition der Bodennachfrage eine besondere Rolle zu. Es gibt verschiedene Definitionen des Begriffs Standortfaktoren. Bökemann (1982, S.123) definiert den Begriff als *Merkmale, welche die Eignung eines Standortes für bestimmte industrielle Produktionsweisen kennzeichnen.* Behrens (1971, S.7) als *solche Merkmale eines geographischen Ortes, die ihn für die Durchführung einer industriellen Produktion attraktiv machen* (nach Kramar 2005, S.35f.). Diese Definitionen legen den Fokus auf betriebliche Nutzungen,können aber auch auf Haushalte übertragen werden (Kramar 2005, S.35f.). Dies ist vor allem im Kontext dieser Arbeit wichtig, da nicht zwischen Bauland für betriebliche Nutzung und Bauland Wohnnutzung differenziert wird.

Abschließend kann also festgehalten werden, dass sich die Bodennachfrage eines bestimmten Standortes und damit der Bodenpreis grundsätzlich erhöht ist, je attraktiver dessen Standortfaktoren sind. Attraktive Standortfaktoren können beispielsweise eine gute Erreichbarkeit oder, im Kontext von touristischer Nutzung, eine attraktive Landschaft sein. Mit welchen Variablen diese Attraktivität der Standortfaktoren allgemein gemessen und abgebildet werden kann, wird im folgenden Teil beschrieben.

2.2.3 Hypothesenbildung

Wie zuvor ausgeführt, ist es Ziel dieser Arbeit ein quantitativ-statistisches Modell für Bodenpreise in Österreich zu entwickeln. Es stellt sich nun die Frage, welche Variablen in ein solches Modell einfließen sollen. Im vorherigen Teil wurde herausgearbeitet, dass diese Variablen allgemein in der Lage sein sollen, die Attraktivität der Standortfaktoren, die der hedonischen Theorie zufolge nutzenstiftenden sind, in einer Gemeinde widerzuspiegeln.

Bodenpreismodellierung wird aufgrund ähnlicher Problemstellungen in bestehenden Arbeiten oft mit Hauspreismodellierung verbunden. Es bietet sich deshalb an, auch Literatur zu Hauspreismodellierung zu betrachten um daraus erklärende Variablen abzuleiten.

In einem ersten Schritt soll versucht werden, erklärende Variablen in Gruppen zu gliedern. Hier gibt es in der Literatur keine einheitliche Herangehensweise, jedoch sind Ähnlichkeiten zu erkennen. In der klassischen Literatur zu hedonischen Hauspreismodellen werden beispielsweise die Kategorien physische, Nachbarschafts-, Lage-, und Vertragsattribute unterschieden (Herath und Maier 2010, S.7). Wieser (2006, S.12f) unterscheidet in seiner Arbeit zu Bodenpreismodellierung in Wien zwischen Erreichbarkeitsvariablen, Nachbarschaftsmerkmalen und Umweltvariablen. In der Arbeit von Helbich et al. (2014, S.392) zur Hauspreismodellierung in Österreich wird untergliedert in physische Eigenschaften und Nachbarschaftseigenschaften. Diese Gliederung findet sich auch bei Brunauer et al. (2013, S.152) in ihrer Arbeit zu Hauspriesen in Österreich wieder. Can (1998, S.63f) betont besonders die Nachbarschaftseigenschaften und untergliedert diese weiter in Erreichbarkeitsvariablen, Variablen, die die gebaute Nachbarschaft beschreiben, Variablen zum sozioökonomischen Umfeld und Variablen zur Versorgungsqualität mit öffentlichen Einrichtungen.

Im Kontext dieser Arbeit erscheint es allerding sinnvoll, zwei Kategorien von erklärenden Variablen zu bilden. Diese sind (1) Nachbarschaftsattribute und (2) Lageattribute. Unter Nachbarschaftsattributen sollen Variablen summiert werden, die die sozioökonomische Struktur einer Gemeinde beschreiben. Lagevariablen sollen die Lage einer Gemeinde innerhalb Österreichs beschreiben. Hier liegt der Fokus auf der Erreichbarkeit.

In einem weiteren Schritt werden die möglichen erklärenden Variablen einzeln beschrieben, deren vermutete Zusammenhänge (Vorzeichen) mit dem Bodenpreis erläutert und somit Hypothesen aufgestellt (theoretisches Modell).

Nachbarschaftsattribute

Die Nachbarschaftsattribute können in Anlehnung an Giuliani (2002, S.105-111) weiter in Variablen zur Flächennutzung, Variablen zur Bevölkerungsstruktur, Variablen zur Wirtschaftsstruktur untergliedert und beschrieben werden. Tabelle 2.1 gibt einen Überblick über die vermuteten Zusammenhänge.

Die Variablen zur Flächennutzung beschrieben die derzeitige Nutzungsverhältnisse und den zur erwarteten Nutzungsdruck. So wird angenommen, dass eine hohe Bevölkerungsdichte den Nutzungsdruck auf den Boden erhöht und sich dadurch der Bodenpreis ebenfalls relativ erhöht. Ähnliches gilt für die Bautätigkeit. Ob sich der Anteil an Bauland beziehungsweise Grünland in einer Gemeinde positiv oder negativ auf den Baulandpreis auswirkt, ist in der Literatur umstritten (Giuliani 2002, S.105f). Einerseits ist ein hoher Anteil an Bauland charakteristisch für städtisch geprägte Gemeinden in denen ein höherer Nutzungsdruck herrscht und somit ein relativ höherer Bodenpreis erwartet wird. Andererseits kann argumentiert werden, dass ein hoher Baulandanteil den Nutzungsdruck durch das erhöhte Angebot an Bauland senkt. In dieser Arbeit wird der erst genannten Argumentationslinie gefolgt und ein positiver Zusammenhang zwischen hohem Baulandanteil und dem Bodenpreis, beziehungsweise ein negativer Zusammenhang zwischen hohem Grünlandanteil und dem Bodenpreis erwartet. Eine weitere sinnvolle Variable den Nutzungsdruck zu messen, ist der Baulandüberhang in einer Gemeinde. Hier wird ein positiver Zusammenhang erwartet.

Grundsätzlich wird erwartet, dass bei einer günstigen Bevölkerungs- und Wirtschaftsstruktur häufiger Bodennutzungskonflikte auftreten. Dadurch wird ein relativ erhöhter Bodenpreis erwartet (Giuliani 2002, S.107). Die Bevölkerungsstruktur einer Gemeinde lässt sich mit dem Belastungsindex beschrieben. Dieser gibt das Verhältnis von am Arbeitsmarkt nicht aktiver (0-19 bzw. über 65 Jahren) und potentiell aktiver (20-65 Jahren) Bevölkerung an. Es wird ein negativer Zusammenhang mit dem Bodenpreis erwartet, da ein hoher Anteil an potentiell aktiver Bevölkerung charakteristisch für eine günstige Wirtschaftsstruktur in einer Gemeinde ist. Zusätzlich wird davon ausgegangen, dass ein starkes Bevölkerungswachstum den Nutzungsdruck erhöht und somit der Bodenpreis steigt.

Weiters wird von einem positiven Zusammenhang zwischen dem Akademikeranteil und dem Bodenpreis und einem negativen Zusammenhang zwischen der Arbeitslosenquote und dem Bodenpreis ausgegangen. Eine günstige Wirtschaftsstruktur kann auch durch das Einkommen in einer Gemeinde beschrieben werden. Hier wird ebenfalls ein positiver Zusammenhang erwartet.

Eine weitere Möglichkeit die Wirtschaftsstruktur einer Gemeinde zu beschrieben, ist Zuteilung der Beschäftigten nach Wirtschaftssektoren. So kann davon ausgegangen werden, dass ein hoher Anteil an Beschäftigten in der Landwirtschaft eine eher schwache Wirtschaftsstruktur

widerspiegelt, wodurch ein negativer Zusammenhang mit dem Bodenpreis zu erwarten ist. Von einem positiven Zusammenhang ist hingegen bei einem hohen von Anteil der Beschäftigten im Tourismus auszugehen. In solchen Gemeinden ist von einem erhöhten Nutzungskonflikt und weiter von einem relativ höherem Bodenpreis auszugehen. Die Variable Anzahl an Übernachtungen spiegelt auch diesen Sachverhalt wieder. Für beide Variablen wird ein positiver Zusammenhang mit dem Bodenpreis erwartet.

Die Pendleraktivität kann ebenfalls dazu herangezogen werden, die Wirtschaftsstruktur einer Gemeinde zu beschreiben. So ist zu erwarten, dass in Gemeinden mit hohen Pendlersaldo er Nutzungsdruck auf Boden und somit der Bodenpreis relativ höher ist.

Tab. 2.1: Hypothesen Nachbarschaftsattribute

Variable	erwarteter Zusammenhang
Variablen zur Flächennutzung	
Bevölkerungsdichte	+
Bautätigkeit	+
Baulandanteil	+
Grünlandanteil	-
Baulandüberhang	+
Variablen zur Bevölkerungsstruktur	
Bevölkerungswachstum	+
Belastungsindex	-
Variablen zur Wirtschaftsstruktur	
Akademikerquote	+
Arbeitslosenquote	-
Einkommen	+
Beschäftigte Landwirtschaft	-
Beschäftigte Tourismus	+
Anzahl Übernachtungen	+
Pendlersaldo	+

Lageattribute

Bei den Lageattributen geht es, wie bereits erwähnt, um die Erreichbarkeit beziehungsweise die Lage von Gemeinden innerhalb Österreichs. Can (1998, S.61) betont bereits *[...]location, location, loction[...]* als die wichtigste Determinate des Preises für Immobilien. Distanzen oder Fahrzeiten zu wichtigen Infrastrukturen und Zentren dienen in der Reglen als Variablen. Herath und Maier (2013, S.166) argumentieren beispielsweise, dass die Distanz zum Stadtzentrum in jedes Hauspreismodell einfließen soll. Es stellt sich allerdings grundsätzliche die Frage, zu welchen Infrastruktureinrichtungen und Zentren die Distanz in ein Modell einfließen soll, um die Erreichbarkeit eins Standortes zu beschreiben. In Anlehnung an das Projekt *Grid-based indicators of accessibility of public utility infrastructure* der *Statistik Austria*, dass die Erreichbarkeit von öffentlicher Infrastruktur in Österreich untersucht hat, werden die Fahrzeiten zu Einzelhan-

delseinrichtungen, Bildungseinrichtungen, Gesundheitseinrichtungen, Freizeiteinrichtungen und Sicherheitseinrichtungen als Lagevariablen angenommen.

Der Zusammenhang zwischen Distanzvariablen (Fahrzeit) und dem Bodenpreis wird als neagtiv angenommen. Das heißt, mit steigender Fahrzeit ist der erwartete Bodenpreis relativ niedriger. Tabelle 2.2 gibt einen Überblick über die Lagevariablen und deren vermuteten Zusammenhänge.

Tab. 2.2: Hypothesen Lageattribute

Variable	erwarteter Zusammenhang
Variablen zur Erreichbarkeit PKW	
Fahrzeit Einzelhandelseinrichtungen	-
Fahrzeit Bildungseinrichtungen	-
Fahrzeit Gesundheitseinrichtungen	-
Fahrzeit Freizeiteinrichtungen	-
Fahrzeit Sicherheitseinrichtungen	-
Variablen zur Erreichbarkeit ÖV	
Fahrzeit Einzelhandel	-
Fahrzeit Bildungseinrichtungen	-
Fahrzeit Gesundheitseinrichtungen	-
Fahrzeit Freizeiteinrichtungen	-
Fahrzeit Sicherheitseinrichtungen	-

Grundsätzlich könnte eine Vielzahl an weiteren Variablen gefunden werden. Dies gilt sowohl für die Nachbarschaftsattribute als auch für die Lageattribute. Es stellt sich allerdings die Frage der Aussagekraft weiterer Variablen. So bilden manche Variablen einen ähnlichen oder gleichen Sachverhalt ab. Dies ist bereits bei den beschriebenen Variablen zu Wirtschaftsstruktur einer Gemeinde zu hinterfragen. Somit kann es sein, dass zwei Variablen sehr stark miteinander korrelieren (Multikollinearität). Dies gilt vor allem für Lagevariablen. Zusätzlich bestehen Probleme der Datenverfügbarkeit und Datenqualität. Die tatsächlich ins Modell einfließenden Variablen, werden weiter in Kapitel 4.2 beschrieben.

Nun, da das theoretische Modell mit den erklärenden Variablen beschrieben ist, kann im nächsten Abschnitt die mathematisch statistische Methodik für das empirische Modell erarbeitet werden. Zuvor muss allerdings noch auf Besonderheiten im Umgang mit Daten mit stark räumlichen Bezug hingewiesen werden.

2.3 Räumliche Effekte

Die Tatsache, dass Erreichbarkeit und Lage im Raum eine der bedeutendsten erklärenden Einflüsse für den Unterschied von Bodenpreisen sind, macht deutlich, dass Bodenpreise räumliche Daten sind (*spatial data*). Das heißt, sie haben einen räumlichen Bezug und es spielt eine Rolle wo sich die Beobachtungen im Raum relativ zueinander befinden. Dem Zugrunde liegt die Denkweise, dass Beobachtungen im Raum sich gegenseitig beeinflussen. So kann die Veränderung eines

Objekts im Raum Einfluss haben auf ein anderes Objekt im Raum. Logan (2012, S.507) schreibt dazu:

> *Everything happens somewhere, which means that all action is embedded in place and may be affected by its placement.*

Es ergeben sich somit räumliche Externalitäten (*spatial externalities*), insbesondere *spatial spillovers* und *spatial multipliers* (vgl. Anselin 2003). Diese werden in der Modellierung räumliche Effekte (*spatial effects*) genannt (vgl. Anselin 1988).

Diese räumlichen Effekte können auch durch Probleme in den Daten selbst, beziehungsweise durch deren Erhebung und Modellierung entstehen. So stellt es ein Problem dar, wenn ein beobachteter räumlicher Prozess sich nicht räumlich mit der Beobachtungseinheit deckt, in der, die den Prozess beschreibenden Daten, erhoben werden (*scale mismatch*) (Anselin 2001b, S.705). Auch die oft vorhandene Notwendigkeit, Daten es verschiedenen Quellen mit unterschiedlichen räumlichen Aggregationsniveaus zusammenzuführen, aber auch Messfehler, können zu solchen räumlichen Effekten führen (Haining 2009, S.7f). Des weiteren können durch das Fehlen von wichtigen erklärenden Variablen (*omitted-variable bias*) in Modellen, räumlichen Effekten auftreten. LeSage und Pace (2009, S.20) schreiben dazu:

> *Latent unobservable influences related to culture, infrastructure, recreational amenities and a host of other factors for which we have no available sample data can be accounted for by relying on neighboring values taken by the dependent variable.*

Zusammengefasst gibt es somit zwei Argumentationslinien, warum räumliche Effekte bei der Modellierung zu berücksichtigen sind. Einerseits aus theoretischen Überlegungen, da sich Beobachtungen im Raum gegenseitig beeinflussen (*spatial externaleties*) und andererseits aus Problemen, die aus den Daten selbst kommen(*scale missmatch, omitted-variable bias*) (Anselin 2001b, S.705).

Diese räumlichen Effekte können grob in zwei Gruppen eingeteilt werden. Diese sind räumliche Heterogenität (*spatial heterogeneity*) und räumliche Abhängigkeit (*spatial dependence*) (vgl. Anselin 1988). Auf diese zwei Gruppen wird im Folgenden kurz eingegangen.

2.3.1 Räumliche Abhängigkeit / Räumliche Autokorrelation

Das *First Law of Geography* von Tobler (1970, S.236) besagt:

> *Everything is related to everything else, but near things are more related than distant things*

Dieses First Law of Geography hebt nochmals die Abhängigkeit von Beobachtung im Raum hervor. Allerdings legt es den Fokus auf relative Nähe von Beobachtungen und sagt, dass solche die sich geographisch näher sind einen engeren funktionalen Zusammenhang aufweisen, als solche die weiter voneinander entfernt sind. Der Zusammenhang verringert sich also mit zunehmender Distanz zwischen zwei Beobachtungen. Man spricht von einem Distantszerfall-Effekt (*disdance decay*). Zu definieren ab welcher Distanz Beobachtungen „nahe“ oder „entfernt“ voneinander sind beziehungsweise wie stark der Zusammenhang mit zunehmender Distanz abnimmt, sind komplexe Fragestellungen die im nächsten Abschnitt 3 besprochen werden. zusätzlich kann kritisiert werden, dass das First Law of Geography lediglich eine Vereinfachung eines komplexen Sachverhalts darstellt. Jedoch stellt es in den meisten Fällen eine sinnvolle Vereinfachung dar, um die räumliche Abhängigkeit zu modellieren (Haining 2009, S.6).

2.3.2 Räumliche Heterogenität

Räumliche Heterogenität meint grundsätzlich, dass Merkmalsausprägungen von Beobachtungen im Raum variieren können. So sind die durchschnittlichen Bodenpreise in den österreichischen Gemeinden nicht überall gleich. Bildet man beispielsweise einfach den Mittelwert über alle österreichischen Gemeinden ab, wird dieser in nur wenigen Gemeinden Aussagekraft besitzen. Ebenso kann sich der Prozess, der die Bodenpreise generiert, in verschiedenen Teilräumen Österreichs unterscheiden. Dies könnte bedeuten, dass die Erklärungskraft gewisser erklärender Variablen in unterschiedlichen Teilräumen unterschiedlich hoch ist. Man spricht von räumliche Stationärität (*stationarity*) beziehungsweise räumlicher nicht-Stationärität (*non-stationarity*).

Räumliche Heterogenität beschriebt somit den Unterschied zwischen einer globalen und einer lokalen Perspektive. Aus der globalen Perspektive heraus wird räumliche Stationerität angenommen. Das heißt, der Untersuchungsraum kann als homogener Raum betrachtet werden. Es wird angenommen, dass die Einflüsse erklärender Variablen überall gleich sind. Solche Modelle verlangen grundsätzlich weniger Annahmen im statistischen Schätzprozess und sind deshalb einfacher zu schätzen und zu interpretieren. Jedoch kann die formale und die theoretische Modellgültigkeit beeinträchtigt, und die Erklärungskraft der Modelle geringer sein.

Auf der anderen Seite nehmen Modelle mit lokaler Perspektive räumliche nicht-Stationärität an. Der Untersuchgsraum wird also als heterogen angenommen. Einflüsse von Variablen und Zusammenhängen können im Raum somit variieren (A. S. Fotheringham et al. 1998, S.1905). Beim Schätzen von lokalen Modellen müssen einige Annahmen getroffen werden. Es erhöht sich somit die Komplexität und die Schwierigkeit der Interpretation der Modelle. Wenn es die Rahmenbedingungen zulassen, sind grundsätzlich einfache Modelle komplexeren Modellen vorzuziehen, da jede weitere Annahme Einschränkungen in der Interpretierbarkeit bedeutet.

2.3.3 Umgang mit räumlichen Effekten in der Modellierung

Bis jetzt ist festzuhalten, dass bei der quantitativ-statistischen Modellierung von Daten mit räumlichem Bezug sehr wahrscheinlich räumliche Effekte vorhanden sind, die in Betracht gezogen werden müssen. Es gibt eine Vielzahl an unterschiedlichen Konzepten und Zugängen, wie dies geschehen soll. Diese kommen aus unterschiedlichen Forschungsrichtungen und beachten verschiedene räumlichen Effekte. Um den Rahmen der Arbeit nicht zu sprengen, wird hier nur kurz auf die bekanntesten Ansätze verwiesen. Anschließend wir im nächsten Abschnitt 3 genauer auf die in dieser Arbeit verwendeten Modelle und Methoden eingegangen.

Jahanshiri et al. (2011) unterscheidet drei erfolgreiche Forschungsrichtungen, wie räumliche Effekte in Betracht gezogen werden können. Diese sind die räumliche Ökonometrie (*spatial econometrics*), die Geographie und die Geostatistik. All diese Forschungsrichtungen haben Überschneidungen und sind nicht immer leicht zu trennen. Die von Jahanshiri et al. (2011) getroffene Untergliederung erscheint aber trotzdem sinnvoll, um die große Bandbreite an Konzepten und Modellen zu überblicken.

Die wahrscheinlich populärste und am weitesten verbreitete Forschungsrichtung ist die räumliche Ökonometrie. Von der Ökonometrie kommend, sind meist einfache lineare Regressionsmodelle der Ausgangspunkt für komplexere Modelle. Hierzu sind beispielsweise die Arbeiten von Anselin (1988) zu zählen. Er machte unter anderem die später noch genauer betrachteten *Simultaneous Autoregressive Models (SAR)* populär (Montero et al. 2018, S.29). Hier geht es vereinfacht darum, Werte benachbarter Beobachtungen in die Regressionsgleichung mitaufzunehmen.

Aus der Geographie kommen lokale Modelle. Auch diese Modelle haben ihren Ursprung in der Ökonometrie. Zu nennen ist die stark verbreitete geographisch gewichtete Regression (*Geographically Weighted Regression*) (vgl. A. S. Fotheringham et al. 1998). Ziel dieser ist es, vereinfacht für jede Beobachtung im Raum eigene Koeffizienten für die erklärenden Variablen zu schätzen. Für die Schätzung werden nur die Beobachtungen einbezogen, die sich jeweils räumlich nahe sind.

Die dritte Forschungsrichtung ist die Geostatistik. Diese ist nicht der Ökonometrie zuzurechnen, sehr wohl aber der Statistik. Zu nennen sind vor allem *kriging-Modelle*. Regressionsmodelle stellen hier nicht den Ausgangspunkt dar (Jahanshiri et al. 2011, S.26). Der Fokus von kriging-Modellen liegt auf der Vorhersage. Im einfachsten Fall wird nur die abhängige Variable betrachtet. In dieser Arbeit ist das der Bodenpreis der österreichischen Gemeinden. Modelliert wird, wie dieser im Raum variiert. So kann kriging als eine Form der räumlichen Interpolation gesehen werden.

Eine weitere Möglichkeit Modellansätze einzuteilen, ist die Einteilung nach den räumlichen Effekten, die sie in Betracht ziehen. So sind manche Modellansätze gut geeignet, räumliche Heterogenität zu modellieren, während andere speziell zur Modellierung von räumlicher Abhängigkeit entwickelt wurden. Die klassischen Modelle aus der räumlichen Ökonometrie (SAR-Modelle) sind vor allem dazu geeignet, räumliche Abhängigkeit zu modellieren. Dies trifft auch auf kriging-Modelle zu. Bei der geographisch gewichteten Regression liegt der Fokus hingegen auf der räumlichen Heterogenität.

In der Literatur findet sich eine Vielzahl an weiteren Modellansätzen und Kombinationen verschiedener Ansätze. So vergleicht Montero et al. (2018) 13 verschiedene räumliche Modellansätze beim Versuch, Hauspreise zu modellieren. Dubin (1998a) kombiniert Regressionsmodelle mit kriging-Methoden. Auch wurden manche Modellansätze erweitert, um räumliche Effekte besser zu berücksichtigen. So entwickelte A. Fotheringham et al. (2002) die *Mixed Geographically Weighted Regression (MGWR)*. Diese unterscheidet zwischen global konstanten und lokal variierenden Koeffizienten für die erklärenden Variablen. Leroux et al. (2000) verwendet wiederum so genannte „Mixed Models“, um räumliche Effekte zu berücksichtigen.

Welcher Modellansatz in einem konkreten Fall verfolgt werden soll, hängt von der genauen Fragestellungen und den zur Verfügung stehenden Daten ab und ist im Einzelfall zu entscheiden. Allerdings erscheint es nicht immer eindeutig, welcher Modellansatz zu verwenden ist. In vielen Fällen werden deshalb mehrere Modelle geschätzt und miteinander verglichen. So auch in dieser Arbeit.

Wie zuvor bereits herausgearbeitet, ist die Lage einer der wichtigsten Faktoren, die den Wert des Bodens bestimmen. Aufgrund dessen kann davon ausgegangen werden, dass räumliche Abhängigkeit beziehungsweise räumliche Autokorrelation in den Daten gegeben ist, da nahe benachbarte Gemeinden in der Regel eine sich ähnelnde Lagequalität aufweisen. Dubin (1998b) fasst die Erkenntnis in Bezug auf Hauspreise zusammen (nach Chica-Olmo 2007, S.91):

> *Location is probably the most important variable used to explain house price. Spatial autocorrelation is present when location is very important to housing price.*

Diese Tatsache legt den Fokus auf Modellansätze, die räumliche Abhängigkeit beziehungsweise räumliche Autokorrelation in Betracht ziehen. Deshalb wird in weiterer Folge auf die Methodik von Modellansätzen aus der räumlichen Ökonometrie, insbesondere *Simultaneous Autoregressive Models (SAR)*, und aus der Geostatistik, insbesondere *kriging-Modelle*, eingegangen. Diese sind beide in der Lage, räumliche Abhängigkeit in Betracht zu ziehen.

Kapitel 3

Methodik

In diesem Kapitel wird die Methodik, von den im Abschnitt 2 (Theoretische Grundlagen) argumentierten Modellansätzen, beschrieben. Zusätzlich wird auf die Möglichkeiten, die Performance verschiedener Modelle zu vergleichen, eingegangen.

Im Bezug auf Bodenpreisdaten ist vor allem auf die räumliche Abhängigkeit zu achten. Aufgrund dessen wird argumentiert, für die Modellierung einerseits räumliche Modelle aus der Ökonometrie (Simultaneous Autoregressive Models SAR) sowie andererseits Modellansätze aus der Geostatistik (kriging-Methoden) zu verwenden. Beide Modellansätze sind geeignet, räumliche Abhängigkeit zu berücksichtigen.

Zuerst werden die Modelle aus der Ökonometrie betrachtet. Es wird ausgehend vom einfachen liniaeren Regressionsmodell aufgezeigt, welche Probleme entstehen und wie räumliche Modelle aus der Ökonometrie diese zu beheben versuchen. Anschließend werden(kriging-Methoden) betrachtet.

3.1 Ökonometrie - Regressionsmodelle

Die weiter oben angeführten Überlegungen zur hedonischen Theorie, dass der Bodenpreis einer Gemeinde durch die nutzenstiftenden Eigenschaften (sozioökonomischen Daten) der Gemeinde beschreibbar ist (siehe Formel 2.1 auf Seite 15), lassen sich grundsätzlich in eine Regressionsgleichung überführen. Die Form dieser kann unterschiedlich komplex sein. Das einfache lineare Modell stellt den Ausgangspunkt da.

3.1.1 Einfachs Lineare Modelle

Ein einfaches lineares Regressionsmodell kann in folgender bekannter Form in Matrixschreibweise geschrieben werden:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{3.1}$$

Wobei $\mathbf{y}$ einen Vektor mit Bodenpreisen je Gemeinde, $\mathbf{X}$ eine Matrix mit den Eigenschaften je Gemeinde, β einen Vektor mit Koeffizienten je Eigenschaft, und ϵ einen Vektor der Residuen (Fehlertermen) darstellt. Ziel ist es, die Koeffizienten β je Eigenschaft zu schätzen. Dafür wird meist die Methode der kleinsten Fehlerquadrate *Ordinary least Squares* (OLS) verwendet. Es wird versuch eine Regressionslinie so in die Punktwolke zu legen, dass die Summe aller quadrierten Abständen zwischen den tatsächliche erhobenen Datenpunkten und der Regressionslinie möglichst gering ist.

Es müssen jedoch einige Annahmen zutreffen, damit eine OLS-Schätzung zu validen Ergebnissen führt. Man spricht auch davon, dass der Schätzer BLUE (*Best Linear Unbiased Estimator*) sein soll. Das heißt, dass der Schätzer erwartungstreu und varianzminimal ist. Hier ist es wichtig das *Gaus-Markov-Theorem* zu erwähnen. Unter den Gaus-Markov Annahmen ist der OLS-Schätzer BLUE (Gujarati und Porter 2003, S.79). Die Gaus-Markov Annahmen spielen also eine zentrale Rolle und werden deswegen genauer betrachtet. Sie besagen (Wooldridge 2012, S.104):

- Die Beziehung zwischen der abhängigen Variable $\mathbf{y}$ und den unabhängigen Variablen $\mathbf{X}$ ist linear und durch die Koeffizienten β beschreibbar
- Falls mit einer Stichprobe gearbeitet wird (im Gegensatz zu Grundgesamtheit) , muss diese Zufällig ausgewählt werden. Zusätzlich muss die Anzahl der Beobachtungen größer sein, als die zu schätzenden Parameter.
- Zwischen den erklärenden Variablen darf keine perfekte Abhängigkeit bestehen (keine perfekte Multikollinearität)
- Der Erwartungswert des Fehlerterms ϵ ist null und somit erwartungstreu. Mathematisch ausgedrückt $E(\epsilon_i|\mathbf{X}) = 0$.
- Der Fehlerterm ϵ hat, unabhängig von dem Wertebereich der abhängigen Variable, immer die gleiche Varianz. Mathematisch ausgedrückt $Var(\epsilon_i|\mathbf{X}) = \sigma^2$. Dies bedeutet, dass die Fehlerterme Homoskedastizität aufweisen. Die letzten beiden Punkte zusammengefasst meinen, dass die Fehlerterme *i.i.d. (independent and identically distributed)* sind. In Bezug auf räumliche Daten bedeutet das auch, dass die Fehlerwerte nicht räumlich korrelieren dürfen.

Wenn all diese Annahmen erfüllt werden können, ist ein einfaches lineares Regressionsmodell mit OLS-Schätzer zielführend. Allerdings müssen diese Annahmen beim vorhanden sein von räumlichen Effekten oft verworfen werden. OLS-Modelle gehen von Unabhängigkeit der einzelnen Beobachtungen untereinander aus. Davon kann bei Bodenpreisdaten, deren größter Einfluss die Erreichbarkeit darstellt, nicht unbedingt ausgegangen werden. LeSage und Pace (2009, S.20) schreibt dazu:

> *[...]it seems unlikely that region i's network of vehicle and public transport infrastructure is independent from that of region j[...]*

Auch nicht-lineare Zusammenhänge zwischen abhängiger und erklärenden Variablen können zu systematischen Über- oder Unterschätzung in manchen Bereichen der Daten führen. Man spricht auch von der funktionalen Form der Regressionsgleichung. Einzelne Variablen in die logarithmische Form zu bringen, kann dieses Problem lösen (Wooldridge 2012, S.43). Daraus ergeben sich dann *semi-log* oder *log-log* Modellspezifikationen, wobei der semi-log Ansatz bei hedonischen Preisen am weitesten verbreitet ist (Herath und Maier 2010, S.8).

3.1.2 Räumliche Modelle

Da beim Vorhandensein von räumlichen Effekten die Gaus-Markov Annahmen nicht mehr zutreffen, können OLS-Modelle zu mangelhaften Ergebnissen führen. Räumliche Regressionsmodelle versuchen auftretende räumliche Effekte in Betracht zu ziehen. Anselin (2001a, S.311) schreibt dazu:

> *There is a growing recognition that standard econometric techniques often fail in the presence of spatial autocorrelation, which is commonplace in geographic (cross-sectional) data sets.*

Ziel ist es, räumliche Abhängigkeiten zwischen den Beobachtungen zu modellieren. Diese Abhängigkeiten lassen sich durch eine Varianz-Kovarianz-Matrix grundsätzlich darstellen. Die Varianz-Kovarianz-Matrix ist in diesem Fall eine Matrix mit allen Beobachtungen (Gemeinden) in Zeilen und Spalten und der jeweiligen Kovarianz, beziehungsweise der Varianz in der Hauptdiagonale, für alle Beobachtungspaare in den Zellen. Wenn sich in den *off-diagonalen* Elementen der Varianz-Kovarianz-Matrix nicht-null Werte befinden, ist eine Abhängigkeit zwischen den Beobachtungen gegeben. Mathematisch ausgedrückt bedeutet das (Anselin 2001a, S.312):

$$cov[y_i, y_j] = E[y_i y_j] - E[y_i] \cdot E[y_j] \neq 0, \qquad \text{for } i \neq j \tag{3.2}$$

Wobei i,j einzelne Beobachtungen und y_i,y_j die abhängige Variable, im Fall dieser Arbeit Bodenpreise, darstellt. Wenn die nicht-null Werte der Beobachtungspaare eine räumlich interpretierbare Struktur aufweisen, ist die Kovarianz (Korrelation) aus räumlicher Sicht bedeutungsvoll. Allerdings kann die Kovarianz der einzelnen Beobachtungspaare nicht einfach berechnet werden, da im Vergleich zu normalen Kovarianzberechnung nur *eine* Beobachtung (zwischen i und j) zur Verfügung steht (*incidental parameter problem*). Sie muss also modelliert und eine räumliche Struktur angenommen werden.

Es gibt in der Literatur mehrere Ansätze, eine räumliche Struktur in die Varianz-Kovarianz-Matrix zu bringen. Räumliche Struktur meint unter anderem, dass die Abhängigkeit (Kovarianz) von Beobachtungen mit zunehmender Distanz sinkt (vgl. Tobler 1970 First Law of Geographie). Am weitesten verbreitet ist der Ansatz eines *spatial stochastic process model*, und weiter insbesondere eines *simultaneous autoregressive processes model (SAR)*, dass den Wert einer Beobachtung im Raum von anderen Beobachtungen im Raum abhängig macht. Die Varianz-Kovarianz-Matrix folgt dann diesem Prozess (Anselin 2001a, S.312).

Hier ist das Konzept einer *spatially lagged variable* zu erwähnen (vgl. Anselin 1988). Diese bildet vereinfacht den gewichteten Mittelwert einer Variable in der Umgebung einer Beobachtung ab. Eine spatially lagged Variable kann dann in die Regressionsgleichung mitaufgenommen werden und so räumliche Abhängigkeit in das Modell bringen. Anders ausgedrückt, kann so die abhängige Varialbe y an einem Standort im Raum nicht nur durch die Eigenschaften des eigenen Standort, sondern auch durch dessen Nachbarschaft erklärt werden. Die Nachbarschaftsbeziehungen und somit der räumliche Prozess, werden mit einer Gewichtungsmatrix W definiert. Es gibt verschiedene Ansätze eine solche zu erstellen, auf die später noch eingegangen wird. Mit Hilfe der Gewichtungsmatrix ,kann dann die Varianz-Kovarianz-Matrix, und somit die Abhängigkeit modelliert werden.

Bei der Aufnahme einer spatially lagged Variable in die Regressionsgleichung werden zwei Hauptansätze unterschieden. Entweder wird diese in Form einer erklärenden spatially lagged Variable oder über die Fehlerterme in des Modell eingeführt. Das erstere wird *simultaneous autoregressive lag model* (SAR_{Lag}) genannt und wird angewendet wenn vermutet wird, dass die räumliche Abhängigkeit beziehungsweise Autokorrelation durch die abhängige Variable eingeführt wird. Der zweite Ansatz ist das *simultaneous autoregressive error model* (SAR_{Error}). Hier wird räumliche Abhängigkeit in den Fehlertermen vermutet (Anselin 2001a, S.316).

Das SAR_{Lag} lässt sich wie folgt darstellen:

$$\mathbf{y} = \rho \mathbf{W} y + \beta \mathbf{X} + \epsilon \tag{3.3}$$

Wobei ρ den räumlichen Autokorrelationskoeffizienten und $\mathbf{W}$ die Gewichtungsmatrix darstellt. Die Variable y auf der rechten Seite der Gleichung stellt den gewichteten Mittelwert der umliegenden Beobachtungen da (spatially lagged Variable). Der gesamte Term $\rho \mathbf{W} y$ bringt die räumliche Abhängigkeit in das Modell. Bei dem SAR_{Lag} Modell wird auch von *substantive spatial dependence* gesprochen, da direkt die räumliche Interaktion in Betracht gezogen wird (Anselin 2001a, S.316). Anders beim SAR_{Error} Modell, das wie folgt geschrieben werden kann:

$$\mathbf{y} = \mathbf{X}\beta + u \tag{3.4}$$

Wobei die räumliche Abhängigkeit über den Fehlerterm u in das Modell kommt und geschrieben werden kann als:

$$u = \lambda \mathbf{W} + \epsilon \tag{3.5}$$

Hier stellt λ den Autokorrelationskoeffizienten und $\mathbf{W}$ wiederum die Gewichtungsmatrix darstellt. Es wird davon ausgegangen, dass die Fehlerterme räumlich korrelieren, also voneinander abhängig sind. Dies kann beispielsweise durch Fehler in der Modellspezifikation oder durch das Fehlen wichtiger erklärender Variablen entstehen (*omitted-variable bias*). Welcher Ansatz besser geeignet ist, muss im Einzelfall geprüft werden. Hier sei auf das Kapitel 3.1.3 hingewiesen, in dem Spezifikationstest beschrieben werden.

Bei den räumlichen SAR_{Error} und SAR_{Lag} Modellen ist das OLS-Schätzverfahren nicht mehr zielführend (LeSage und Pace 2009, S.30). Deshalb müssen andere Verfahren angewendet werden. Am meisten verbreitet ist das *Maximum Likelihood* (ML) Schätzverfahren, das auch in dieser Arbeit verwendet wird. Die Höhe der geschätzten Koeffizienten ρ beziehungsweise λ spiegeln grundsätzlich die Stärke der Abhängigkeit wieder. Allerdings ist die Interpretation erschwert, da diese nur in Kombination mit der definierten Gewichtungsmatrix erfolgen kann.

Gewichtungsmatrix W

Die räumliche Gewichtungsmatrix definiert, wie bereits erwähnt, die Nachbarschaftsbeziehungen der Beobachtungen. Sie stellt eine Matrix mit allen Beobachtungen in Zeilen und Spalten dar. Die Elemente der Matrix definieren wie stark Beobachtungen benachbart sind (Anselin 2001a, S.212). In der einfachsten Form nehmen die Matrixelemente den Wert 1 wenn eine Nachbarschaftsbeziehung besteht und den Wert 0 an wenn dies nicht der Fall ist. Die Hauptdiagonale besteht aus 0-Werten. Damit wird vermieden, dass Beobachtungen direkt mit sich selbst benachbart sind. Meist werden alle Zeilen der Gewichtungungsmatrix standardisiert (*row-standardization*) um die Interpretation zu erleichtern (Getis und Aldstadt 2004, S.93).

Es gibt in der Literatur unterschiedliche Ansätze die Gewichtungsmatrix zu erstellen. Eine umfassende Aufzählung befindet sich in Getis und Aldstadt (2004). Grundsätzlich kann zwischen Distanz-basierten und Nachbarschafts-basierten Ansätzen unterschieden werden. Bei den Distanz-basierten Ansätzen werden solche Beobachtungen als Nachbarn definiert, die in einer gewissen Distanz zueinander liegen. Nachbarschafts-basierte Ansätze definieren solche Beobachtungen als Nachbarn, die eine gemeinsame Grenze haben. Bei beiden Ansätzen kann das Problem auftreten, dass die Anzahl der Nachbarn je Beobachtung stark variiert. Der *k-nearest neighbors* Ansatz löst

dieses Problem. Hier haben alle Beobachtungen die gleiche Anzahl an Nachbarn und zwar die, die der jeweiligen Bobachtung am nächsten liegen.

Der Gewichtungsmatrix kommt somit eine zentrale Rolle zu. Im Zuge dieser Arbeit werden verschiedene Gewichtungsmatrizen definiert, in der Modellierung verwendet und verglichen.

3.1.3 Spezifikationstests / Modelldiagnostik

Im Rahmen dieses Kapitels werden statistische Spezifikationstest vorgestellt, die hinterfragen, ob die oben beschriebenen räumlichen Regressionsmodelle überhaupt zielführend sind. Des weiteren wird die Modelldiagnostik besprochen. Es wird auf statistische Tests zur Prüfung der Modellannahmen, insbesondere auf Multikollinearität und Homoskedastizität eingegangen.

Es stellt sich nun grundsätzlich die Frage, ob räumliche Regressionsmodellspezifikationen benötigt werden oder, ob einfache liniere OLS-Modelle den Prozess, der die Bodenpreise generiert ausreichend gut beschrieben. Die Grundlage, anhand derer diese Frage beantwortet werden kann, sind die geschätzten Residuen (Fehlerterme). Der am weitesten verbreitetste Spezifiaktionstest ist die *Morans'I* Statistik (Anselin 2001a, S.323) nach Moran (1948). Diese misst die Stärke und Signifikants der räumlichen Autokorrelation und kann wie folgt geschrieben werden (Bivand et al. 2013, S.259):

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.6}$$

Wobei n die Anzahl der Beobachtungen, $\bar{x}$ den Mittelwert von x, x_i die i-te Beobachtung und w_{ij} die Gewichtung zwischen den Beobachtungen darstellt. Ist der Wert für Morans'I signifikant, von 0 verschieden besteht räumliche Autokorrelation. Es gilt zu beachten, dass der Morans'I ein so genannter *diffuser* Test ist. Dies bedeutet, dass zwar klar hervorgeht ob räumliche Autokorrelation vorliegt, nicht aber mit welcher Modellspezifikation diese in Betracht gezogen werden soll. Hierfür wurden so genannte *fokussierte* Test entwickelt. Zu nennen ist hier vor allem der *Lagrange Multiplier* (LM) Test (Anselin et al. 1996).

Der Lagrange Multiplier Test basiert auf Maximum Likelihood Prinzipien. Es werden immer zwei Modellalternativen miteinander verglichen und geprüft, ob diese sich signifikant unterscheiden. Die Ergebnisse der LM-Tests geben einen Hinweis darauf, wo sich die räumliche Autokorrelation befindet. Ob es sich also um einen SAR_{Lag} oder einen SAR_{Error} Prozess handelt (Bivand et al. 2013, S.290ff). Es wird ein eigener LM-Test für das SAR_{Lag} und für das SAR_{Error} durchgeführt. Die Modellspezifikation mit einem signifikanten Testergebnis für den jeweiligen LM-Test ist zu bevorzugen. Sind beide LM-Test signifikant, kann der robuste Lagrange Multiplier (RLM) Test angewendet werden. Hier ist dann das höhere Signifikantsniveau ausschlaggebend (Bivand et al. 2013, S.291).

3.1.3.1 Normalität

Eine weitere Annahme die Residuen betreffend ist deren Normalverteilung (Wooldridge 2012, S.118). Dies kann mit Diagnoseplots überprüft werden. Hier ist vor allem der Q-Q Plot zu nennen.

Beim Q-Q Plot werden die Residuen, geordnet nach ihrer Größe, gegen die theoretische Normalverteilung geplottet. Wenn eine Normalverteilung gegeben ist, liegen nahezu alle Residuenwerte auf einer geraden Linie.

3.1.3.2 Multikollinearität

Regressionsmodelle bilden Abhängigkeiten (Kollinearität) ab. Im Speziellen geht es um die Abhängigkeit der abhängigen Variable y von den erklärenden Variablen X. Gibt es allerdings Abhängigkeiten innerhalb der erklärenden Variablen X, spricht man von Multikollinearität. Solche Abhängigkeiten können zu Problemen bei der Schätzung und Interpretation von Regressionsmodellen führen (Wooldridge 2012, S.95). Das nicht vorliegen von Multikollinearität, wurde zuvor schon als eine der Gaus-Markov-Annahmen definiert. Liegt Multikollinearität vor ist der OLS-Schätzer zwar immer noch BLUE, es kann aber der Einfluss einzelner erklärender Variablen nicht mehr isoliert werden. Es gibt eine Vielzahl an Teststatistiken die das vorliegen von Multikollinearität testen. In dieser Arbeit wird der *Variance Inflation Factor* (VIF) wie folgt berechnet:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.7}$$

In der Literatur finden sich unterschiedliche Grenzwerte für den VIF wieder. Oft wird der Wert 10 gewählt (Wooldridge 2012, S.95).

3.1.3.3 Heteroskedastizität

Skedastizität bezieht sich auf die Varianz der Fehlerterme. Ist diese im gesamten Wertebereich der abhängigen Variablen gleich, spricht man von Homoskedastizität. Anders formuliert bedeutet das, dass beispielsweise im oberen Wertebereich die durchschnittliche Abweichung gleich hoch ist wie im unteren Wertebereich. Ist dies nicht der Fall, spricht man von Heteroskedastizität. Liegt Heteroskedastizität vor ist der OLS-Schätzer zwar erwartungstreu aber nicht mehr varianzminimal (Wooldridge 2012, S.269f).

Heteroskedastizität kann unterschiedliche Ursachen haben. Sie kann sich durch eine Fehlspezifikation des Modells ergeben. Sind Einflüsse von Variablen beispielsweise räumlich nicht-Stationär, kann dies zu Heteroskedastizität führen (Anselin 2001a, S.311). Zusätzlich kann über die räumliche Gewichtungsmatrix W Heteroskedastizität eingeführt werden. Wird die Gewichtungsmatrix beispielsweise so definiert, dass nicht alle Beobachtung gleich viele Nachbarn haben (z.B. kein k-nearest neighbors Ansatz) wird dadurch Heteroskedastizität ins Modell gebracht (Anselin 2001a, S.313).

Ob Heteroskedastizität vorliegt, kann mit dem *Breusch-Pagan* Test überprüft werden (vgl. Breusch und Pagan 1979). Ist dieser signifikant, liegt Heteroskedastizität vor. Dies deutet auch auf das vorhanden sein von räumlicher Heterogenität hin (Anselin 2001a, S.311). Zusätzlich kann dies grafisch mittels Residuenplot überprüft werden. Dafür werden die Residuen gegen die geschätzten y Werte geplottet. Zusätzlich wird eine Trendlinie in den Plot eingeführt. Ist diese annähernd horrizontal liegt kein Trend in den Residuen vor.

3.2 Geostatistik - Kriging Modelle

Der zweite Ansatz neben den Ansätzen aus der Ökonometrie die in der Arbeit verwendet wird, ist die Geostatistik. Im Gegensatz zur Ökonometrie geht es bei der Geostatistik vermehrt darum, Vorhersagen (*spatial prediction*) zu machen (Jahanshiri et al. 2011, S.26). Geostatistische Ansätze kommen ursprünglich aus der physischen Geographie und hier speziell aus dem Bergbau (vgl. Journel und Huijbregts 1978). In letzter Zeit werden die Mehtoden aber vermehrt auch für humangeographische Fragestellungen verwendet (Haining et al. 2010, S.7). Die Geostatistik beschäftigt sich allgemein mit Phänomenen die im Raum variieren. Isaaks und Srivastava (1989, S.3) schreiben dazu:

> *Geostatistics offers a way of describing the spatial continuity of natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity.*

Es geht also darum, herauszufinden, *wie* Sachverhalte im Raum variieren. Räumliche Autokorrelation beziehungsweise räumliche Abhängigkeit ist somit das zentrale Thema in der Geostatistik. Wie vorhin bereits betont, kann davon ausgegangen werden, dass bei Bodenpreisdaten räumliche Autokorrelation beziehungsweise Abhängigkeit vorhanden ist.

Die wahrscheinlich wichtigsten Beiträge im Bereich der Geostatistik kommen von Matheron (1963), Krige (1951) und Goldberger (1962). Aufeinander aufbauend, führten sie den Begriff *kriging*, für eine Methode, die es ermöglicht optimale Vorhersagen im geographischen Raum zu machen, ein und entwickelten somit einen *spatial best linear unbiased predictor* (BLUP) (Haining et al. 2010, S.10).

Dem Zugrunde liegt die Annahme, dass sich die räumliche Variation von Phänomenen aus zwei Komponenten zusammensetzt. Einerseits einer *large scale variation* und andererseits einer *small scale variation*. Dies kann wie folgt geschrieben werden (Chica-Olmo 2007, S.93):

$$Z_s = m_s + u_s \tag{3.8}$$

Z_s stellt den Bodenpreis in einer Gemeinde s dar. m_s repräsentiert die large scale Variation, die auch *trend* oder *drift* genannt wird. u_s bildet die small scale Variation ab und kann als räumlich autokorrelierende Fehlerterm verstanden werden. Die Idee ist, dass durch die Autokorrelation in den Fehlertermen zusätzliche Information vorhanden ist, die genutzt werden kann, um die Vorhersage zu verbessern.

Es gibt unterschiedliche Arten beziehungsweise Komplexitätsstufen von kriging-Modellen, deren Unterschiede sich vor allem auf das Schätzen der large scale Variation beziehen. Eine einfache Art ist *ordinary kriging*. Beim ordinary kriging wird davor ausgegangen, dass die large scale Variation überall gleich beziehungsweise konstant und somit stationär ist. in diesem Fall wird die Abweichung von einem konstanten Mittelwert mittels kriging modelliert (Chica-Olmo 2007, S.93). Betrachtet wird nur die abhängige Variable, in diesem Fall der Bodenpreis der österreichischen Gemeinden.

Komplexere kriging Arten nehmen an, dass die large scale Variation von Beobachtung zu Beobachtung variiert. Hier ist vor allem *universal* kriging zu nennen. Angenommen wird, dass die large scale Variation einem Trend folgt. Dieser Trend wird meist mit einem Regressionsmodell beschrieben. Dieses kann unterschiedlich komplex sein. Die kriging-Methode wird dann auf die räumliche autokorrilierenden Fehlerterme angewandt. Diese Fehlerterme müssen dann räumlich stationär sein. Das Regressionsmodell kann allerdings nicht mit dem OLS-Schätzer geschätzt

werden, da die Gaus-Markov-Annahmen verletzt sind (Fehlerterme sind nicht *i.i.d*). Es bieten sich der *generalized least squares estimator* (EGLS) an (Sinclair und Blackwell 2002a, S.215).

Eine weitere komplexere Art von kriging ist *Co-* kriging. Hier wird angenommen, dass es Co-Variablen gibt, die hoch mit der zu vorhersagenden Variable (im Fall dieser Arbeit Bodenpreise) korrelieren. Diese zusätzliche Information wird dann verwendet, um die Vorhersage zu verbessern (Chica-Olmo 2007, S.91).

Welche kriging art in einem Konkreten Fall angewandt werden soll, ist abhängig von den Daten und deren Eigenschaften. Speziell die Entscheidung, ob Co-kriging oder universal kriging besser geignet ist, ist oft schwer zu bestimmen. In dieser Arbeit wird geprüft, ob ein Trend in large scale Variation vorliegt. Wenn dies der Fall ist wird universal kriging angewendet.

3.2.1 Variogramm

Es stellt sich also wiederholt die Frage, wie Beobachtungen im Raum korrelieren. Beim Kriging wird diese Korrelation durch das sogenannte *Variogramm* abgebildet. Es ist das zentrale Element im kriging-Verfahren. Haining et al. (2010, S.7) schreiben über das Variogramm:

> *Geostatisticians attach much importance to estimating and modeling the variogram to explore and analyze spatial variation because of the insight it provides.*

Das Variogramm beschreibt die räumliche Abhängigkeit von Beobachtungen aufgrund ihrer Distanz zueinander (Chica-Olmo 2007, S.92). Dafür werden zuerst Distanzklassen gebildet. Anschließend wird die Varianz innerhalb dieser Distanzklassen berechnet und halbiert. Das ergibt die Semivarianz nach folgender Formel (Blöschl und Merz 2000, S.152):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i,j} (R_i - R_j)^2 \tag{3.9}$$

Wobei $\hat{\gamma}$ die Semivarianz, N die Beobachtungen je Distanzklasse, h die Distanzkalsse und R die zu vorhersagende Variable, im Fall dieser Arbeit die Residuen (Fehlerterme) der Regression darstellen. Halbiert wird die Varianz deshalb, da sonst alle Beobachtungen doppelt in die Berechnung einfließen würden (von i nach j und j nach i).

Die Semivarianz nimmt in der Regel mit größer werdender Distanz zu. Dies kommt daher, dass erwartet wird, dass sich näher liegende Beobachtungen ähnlicher sind als weiter von einander entfernte (Sinclair und Blackwell 2002b, S.194). Eine idealisiertes Beispiel eines empirischen Semivariogramm ist in Abbildung 3.1 zu sehen. Jede Distanzklasse wird durch einen Punkt repräsentiert

Dieses so erstellt Semivariogramm wird als *empirisches* beziehungsweise *experimentelles* Semivariogramm bezeichnet, da es sich aus der Zusammensetzung der Daten ergibt. Bei der Erstellung des empirischen Semivariogramms stellt sich allerdings die Frage nach der Größe der Distanzklassen. In der Regel werden unterschiedliche Distanzklassengrößen getestet. Es gilt darauf zu achten, dass in jeder Beobachtungskalsse genügend Beobachtungspaare enthalten sind. Dies gilt vor allem für kurze Distanzen, da hier meist weniger Beobachtungspaare vorhanden sind (Sinclair und Blackwell 2002b, S.196).

In Abbildung 3.1 sind auch die Kenngrößen des Semivariogramms ersichtlich. Diese sind *Sill*, *Nugget* und *Range*. Der Sill ist die Semivarianz des gesamten Prozesses. Das Nugget ist die Varianz von zwei Beobachtungen mit sehr geringem Abstand. Bei einer Distanz von 0 ist der

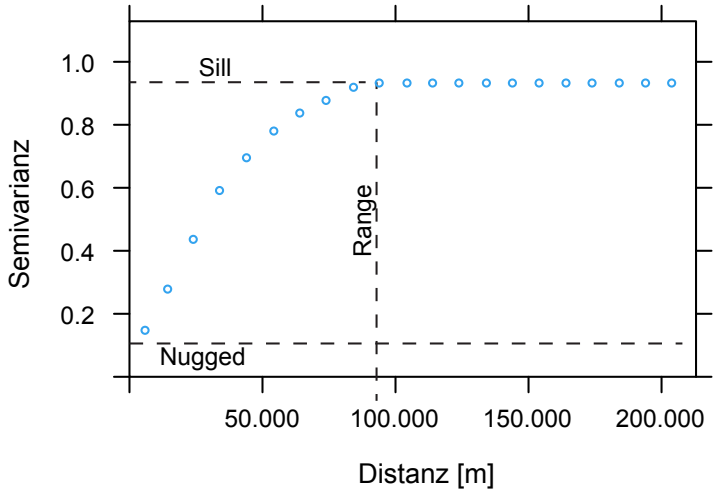

Abb. 3.1: Beispiel eines empirischen Semivariogramms

erwartete Wert für die Semivarianz ebenfalls 0. Dies ist aber aufgrund von Messfehlern meist nicht der Fall. Diesen Effekt nennt man Nugged-Effekt (Blöschl und Merz 2000, S.152). Somit ergibt sich zusätzlich die *partial Sill* als die Differenz zwischen Sill und Nugged. Die Range ist die Reichweite, an der die Semivarianz die Sill erreicht. Ab dieser Distanz ist keine signifikante Autokorrelation mehr vorhanden.

Aufbauend auf das empirische Semivariogramm muss ein theoretisches Semivariogramm geschätzt werden. Dieses wird auch Semivariogramm-Modell genannt. Es gibt eine Vielzahl an verschiedenen Ansätzen, ein Semivariogramm-Modell zu schätzen. Ziel ist es grundsätzlich immer, das empirische Semivariogramm gut zu repräsentieren. Die am meisten verbreiteten Modelle sind das *sphärische-*, das *exponential-* und das *gaussian*-Modell, wobei des spährische Modell am häufigsten angewandt wird (Sinclair und Blackwell 2002b, S.196). Diese können wie folgt geschrieben werden (Sinclair und Blackwell 2002b, S.197):

$$\text{Sphärisches Modell: } \gamma(h) = c \cdot 1,5(\frac{h}{a}) - 0,5(\frac{h}{a})^3 \quad \text{bei } h \leq a; \quad \text{sonst: } \gamma(h) = c \tag{3.10}$$

$$\text{Exponential Modell: } \gamma(h) = c \cdot (1 - exp(\frac{-3h}{a})) \tag{3.11}$$

$$\text{Gaussian Modell: } \gamma(h) = c \cdot (1 - exp(\frac{-3h^2}{a^2})) \tag{3.12}$$

Wobei c der Sill, a die Range und h die Distanz darstellt. Bei dem exponentialen Modell ergeben sich Schwierigkeiten bei der Interpretation, da die Kurve nicht eindeutig abflacht. Hier wird die *practical* Range eingeführt. Diese Beträgt 95 Prozent der asymtotischen Range. Abbildung 3.2 zeigt die unterschiedlichen Modelle:

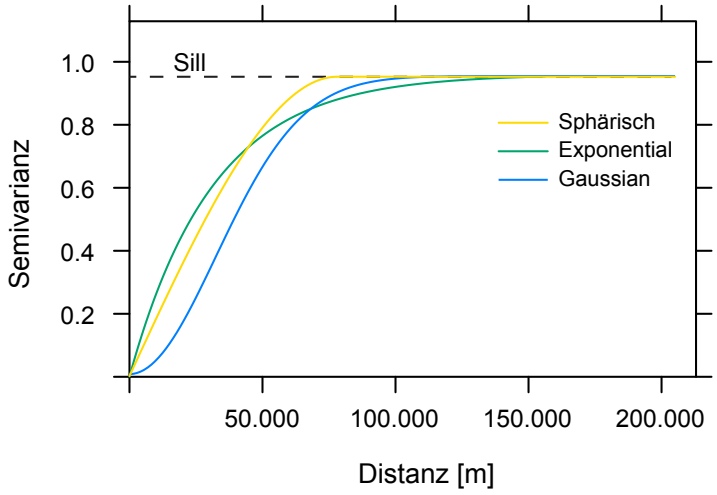

Abb. 3.2: Vergleich zwischen sphärischem und exponentialem und gaussian Semivariogramm-Modell

Es stellt sich die Frage, welches der Modelle am besten geeignet ist, um das empirische Semivariogramm und in weitere Folge die räumliche Abhängigkeit zu repräsentieren. Sinclair und Blackwell (2002b, S.198) schreiben dazu:

> *The art of fitting models to experimental semivariograms is fraught with subjective decisions.*

Dieses Zitat hebt die Schwierigkeit der Frage nach dem besten Modell hervor. In dieser Arbeit wird *Leave-one-out cross validation* (LOOCV) angewandt, um verschieden große Distanzklassen und Modellansätze zu vergleichen. Auf das LOOCV Verfahren wird im Kapitel 3.3 zum Modellvergleich genauer eingegangen.

3.2.2 Kriging Vorgangsweise

Es gilt zu beachten, dass kriging eine Methode zu Vorhersage für Beobachtungen im Raum ist, für die keine Werte vorliegen. Um die Methodik von kriging weiter zu erklären, muss deshalb angenommen werden, dass der Bodenpreis einer bestimmten Gemeinde nicht bekannt ist und vorhergesagt werden soll. Um universal kriging Anzuwenden, muss allerdings die oben beschriebene large scale Variation in dieser Gemeinde bekannt sein. In diesem Fall wird angenommen, dass dies die erklärenden Variablen X in der Regressionsgleichung sind. Dies bringt folgende Ausgangssituation: Es sind die Fehlerterme der einzelnen Gemeinden bekannt (y_i). Nun stellt sich die Frage, wie hoch dieser Fehlerterm in einer nicht beobachteten Gemeinde ist (y_0). Dies kann grafisch vereinfacht wie in Abbildung 3.3 dargestellt werden.

Kriging kann auch als räumliche Interpolation verstanden werden (Sinclair und Blackwell 2002a, S.218). Die umliegenden Beobachtungen haben nun einen unterschiedlich großen Einfluss auf die

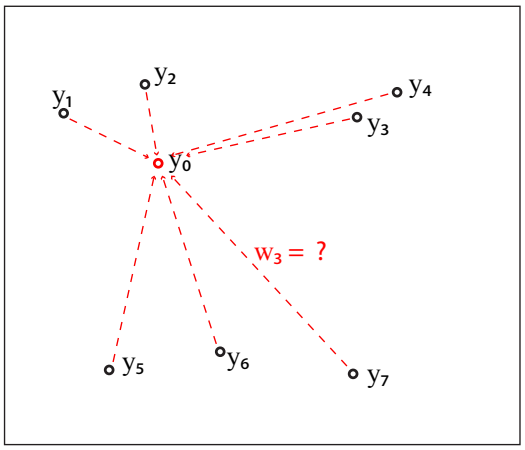

Abb. 3.3: Beispiel Ausgangssituation für kriging

zu vorhersagende Beobachtung. Sie werden also gewichtet. Daraus ergibt sich die Formel für y_0 (Sinclair und Blackwell 2002a, S.217):

$$y_0 = \sum_{i}^{n} w_i \cdot y_i \tag{3.13}$$

Wobei y_0 die zu vorhersagende Variable, y_i die bekannten Variablen und w_i die jeweilige Gewichtung des Einflusses der Variablen darstellt. Die Gewichte w_i müssen nun so berechnet werden, dass Varianz der Vorhersage so gering wie möglich ist. Die Herleitung der Formel für die Gewichtung ist sehr komplex und wird hier nicht genauer beschrieben. Für die genaue Herleitung sei auf Isaaks und Srivastava (1989, S.280-287) verwiesen. Es ergibt sich folgende Formel für die Gewichte in Matrixschreibweise (Isaaks und Srivastava 1989, S.287):

$$w = C^{-1} \cdot D \tag{3.14}$$

Wobei C^{-1} die inverse Varianz-Kovarianz-Matrix aller bekannten Beobachtungen, und D einen Vektor mit Kovarianzen zwischen allen bekannten Beobachtungen und der zu vorhersagenden Beobachtung darstellt. Die Werte für C^{-1} und D können mit dem zuvor geschätztem Semivariogramm-Modell ermittelt werden. Zusätzlich wird die Varianz der vorhergesagten Beobachtung mit folgender Formel ermittelt (Sinclair und Blackwell 2002a, S.217):

$$\sigma_0^2 = \sigma^2 - \sum_{i}^{n} w_i Cov(y_i, y_0) + \mu \tag{3.15}$$

Wobei σ^2 die Varinaz des gesamten Prozesses, $Cov(y_i, y_0)$ die Kovarianz zwischen y_i und y_0, und μ ein *Lagrange Parameter* darstellt. Der Lagrange Parameter erklärt sich aus der Herleitung der Formel 3.14 in Isaaks und Srivastava (1989, S.284). Über die Varianz der Vorhersage ist somit ein Maß verfügbar, wie genau beziehungsweise vertrauenswürdig die vorhergesagte Variable an einer bestimmten Position im Raum ist.

Beim Kriging ist die Distanz zwischen den Beobachtungen von großer Bedeutung und hat maßgeblichen Einfluss auf das Ergebnis. Da es in dieser Arbeit um den durchschnittlichen Bodenpreis für Gemeinden geht, der modelliert werden soll, stellt sich die Frage nach der Distanz zwischen den Gemeinden. Um dieses Problem zu lösen, werden die Gemeinden vereinfacht als Punkte angenommen. Hierfür wird der Schwerpunkt der bewohnten Flächen je Gemeinde errechnet. Dieser wird dann im kriging-Verfahren verwendet, um die euklidische Distanz zwischen den Gemeinden zu ermitteln.

3.2.3 Räumliche Stationarität

Die Annahme, dass räumliche Stationarität vorliegt, ist zentral für das kriging-Verfahren. Räumliche Stationarität liegt vor, wenn der Mittelwert und die Varianz der zu interpolierenden Variable über den gesamten Datensatz annähernd gleich ist (Haining et al. 2010, S.11). Im Fall von universal kriging gilt das für die Fehlerterme der Regression. Ob räumliche Staionarität vorliegt, kann grafisch mit dem empirischen Semivariogramm überprüft werden. Folgt das Semivariogramm einer annähernd idealtypischen Form, wie in Abbildung 3.1, kann von Stationarität ausgegangen werden.

3.3 Modellperformance / Modellvergleich

Bis jetzt wurde nicht auf Maßzahlen der Modellgüte (Modell-Fit) eingegangen. Dies soll in diesem Kapitel geschehen. Erst werden die verschiedenen Modellansätze einzeln betrachtet, um anschließend auf Messgrößen einzugehen, die unterschiedliche Modellansätze miteinander vergleichen können.

Für einfache lineare Regressionsmodelle (OLS) ist vor allem das Bestimmtheitsmaß R^2 zu erwähnen. Es ist das Verhältnis der SSR (*Sum of Squared Residuals*) und der TSS (*Total Sum Squared*), wobei SST die Summe aus SSR und ESS (*Explaned Sum Squared*) ist. Diese können wie folgt bestimmt werden:

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{3.16}$$

Wobei $\bar{y}$ den Mittelwert und $\hat{y}$ die gefitteten Werte darstellt. Daraus berechnet sich dann das R^2 (Wooldridge 2012, S.32):

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{3.17}$$

Es geht also darum, die Streuung von y in einen durch die erklärenden Variablen erklärbaren Teil und einen nicht erklärbaren Teil zu trennen. Der Wert von R^2 liegt somit immer zwischen 0 und 1. Je höher dieser ist, desto besser ist der Erklärungsgehalt der Modelle (Ernste 2011, S.73). Um OLS-Modelle mit einer unterschiedlichen Anzahl an erklärenden Variablen miteinander zu vergleichen, eignet sich das Bestimmtheitsmaß R^2 nur bedingt, da sich R^2 unweigerlich mit jeder

neuen erklärenden Variable x erhöht. Deswegen sollte das so genannte *adjusted* R^2 für einen Vergleich herangezogen werden (Wooldridge 2012, S.202):

$$\bar{R^2} = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} \tag{3.18}$$

Wobei n für die Anzahl der Beobachtungen und k für die Anzahl der erklärenden Variablen steht. Hier wird also um die Anzahl der Freiheitsgrade korrigiert (Ernste 2011, S.74f).

Bei räumlichen Modellen, die nicht mit OLS-Schätzern geschätzt wurden, ist das Bestimmtheitsmaß R^2 nicht anwendbar. Werden räumlichen Modelle mit dem Maximum Likelihood (ML) Schätzverfahren geschätzt, ergeben sich allerdings eine andere Maßzahlen, die den Modell-Fit beschreiben. So wird versucht, den *log-likelihood* zu maximieren. Anselin (2001a, S.320) zeigt, dass bei *simultaneous autoregressive processes Models (SAR)* die Maximierung des *log-likelihood* gleichzeitig zu einer Minimierung der SSR (*Sum of Squared Residuals*) führt.

Ein weiteres Maß, um den Modell-Fit von Modellen zu vergleichen, ist das *Akaike Information Criterion* (AIC), das von Akaike (1974) entwickelt wurde. Dieses basiert auch auf log-likelihood Funktionen, berücksichtigt aber die Anzahl der erklärenden Variablen in einem Modell. Je niedriger das AIC ist, desto besser is der Modell-Fit (Getis und Aldstadt 2004, S.98). Das AIC kann sowohl für einfache OLS-Modelle, als auch räumliche SAR-Modelle berechnet werden. Somit ist ein Vergleich dieser Modelle möglich.

In Kapitel 3.2 wurde bereits auf das *Corss-Validation* Verfahren hingewiesen. Dieses Verfahren bietet allgemein die Möglichkeit, verschiedene Modellspezifikationen miteinander zu vergleichen, um jenes mit der größten Vorhersagekraft zu finden. In dieser Arbeit wird das cross-validation Verfahren dazu verwendet, um verschiedene Semivariogramm-Modellspezifikationen zu vergleichen. Konkret wird *Leave-One-Out Cross Validation* (LOOCV) angewandt. LOOCV entfernt einen Datenpunkt aus dem Modellbildungsprozess und trifft für diesen eine Vorhersage. Dieser Prozess wird für alle Datenpunkte wiederholt (Chica-Olmo 2007, S.100). So kann anschließend der *Roor Mean Square Error* (RMSE) je Modellspezifikation errechnet werden. Die Modellspezifikation mit dem geringsten RMSE hat den besten Modell-Fit.

Berechnet man den RMSE nicht für die Beobachtungen, die zum Schätzen der Modelle genutzt wurden (in-sample-Prediction), sondern für neue Beobachtungen spricht man vom *Roor Mean Square Prediction Error* (out-of-sample-prediction). Der RMSPE eignet sich somit auch, um kriging-Modelle und Regressionsmodelle miteinander über die out-of-sample Vorhersage zu vergleichen. Der RMSPE sowie der RMSE stellt die Wurzel der mittleren quadratischen Abweichung vom tatsächlich erhobenen Wert und dem geschätzten Wert dar:

$$RMSE/RMPSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{3.19}$$

Um den RMSPE für Modelle zu berechnen, muss der Datensatz in ein Trainingsset und ein Testset aufgeteilt werden. Die Modelle werden mit dem Trainingsset geschätzt und anschließend wird versucht, die Beobachtungen des Testsets vorherzusagen. Daraus kann dann der RMSPE berechnet werden. Je kleiner dieser Wert ist, desto kleiner ist die Verzerrung, bezogen auf den erwarteten Wert (Chica-Olmo 2007, S.100). Auf die detaillierte Vorgangsweise der Modellbildung wird im Kapitel 5.1 eingegangen.

Alle Modelleschätzungen und die Datenaufbereitung in dieser Arbeit werden mit der Statistiksoftware *R* durchgefürt. Die Ermittlung des Gemeindemittelpunkts erfolgt mit der GIS-Software *Arc-Map*. Die räumliche Gewichtungsmatrizen wird der Software *GeoDa* definiert.

Kapitel 4

Daten

In diesem Abschnitt wird auf die zur Verfügung stehende und verwendete Datengrundlage eingegangen. In Kapitel 4.1 werden die Datenquellen sowie deren Aufbereitung behandelt. In Kapitel 4.2 erfolgt die deskriptive Analyse der Daten, aufgeteilt nach der abhängigen Variable und den erklärenden Variablen.

4.1 Datenherkunft und Datenaufbereitung

Die Verfügbarkeit von Daten in einer gewissen qualitativen und zeitlichen Homogenität beeinflusst das Ergebnis und die daraus gewonnen Erkenntnisse eines Modells in hohem Maße. Jedoch konnten nicht alle im Kapitel 2.2.3 Hypothesenbildung beschriebenen Variablen mit solchen Datensätzen hinterlegt werden. Die Modellbildung erfolgt somit nur mit den zur Verfügung stehenden Datensätzen

4.1.1 Abhängige Variable - Bodenpreis

Zentral für die Modellbildung sind Bodenpreisdaten zu den österreichischen Gemeinden. Ein entsprechender Datensatz steht dem Fachbereich für Stadt- und Regionalforschung zur Verfügung. Dieser besteht auf Bodenpreisdaten, die die Gemeinden selbst melden. Aufgrund dessen ist die zeitliche und qualitative Homogenität nur schwer zu gewährleisten. Für den Rahmen dieser Arbeit ist die Datenqualität allerdings ausreichend.

Der Datensatz enthält je Gemeinde eine Preisspanne für einen Quadratmeter Bauland. Für einzelne Gemeinden steht diese Preisspanne unterteilt in verschiedene Ortschaften innerhalb der Gemeinde oder für besondere Lagen wie Seegrund zusätzlich zur Verfügung. Diese besonderen Lagen (Seegrund, ...) werden vernachlässigt. Je Gemeinde wird weiters der Mittelwert des Bodenpreises über alle Ortsschaften und der Preisspanne errechnet. So entsteht ein Datensatz mit der abhängigen Variable Bodenpreis je Gemeinde, der in weiterer Folge für die Modellierung herangezogen werden kann.

4.1.2 Erklärende Variablen

Nachbarschaftsvariablen

Die Variablen zur Flächennutzung (Bevölkerungsdichte, Baulandanteil, Grünlandanteil) werden über den CORINE-Landcover Datensatz berechnet. Dieser enthält für ganz Österreich die

Landesbedeckungs- und Nutzungsklassen. Somit kann näherungsweise der Anteil an Grünland und Bauland am Dauersiedlungsraum berechnet werden. Zum Bauland werden alle Flächen der Nutzungsklasse *bebaute Flächen* und zum Grünland alle Flächen der Nutzungsklasse *landwirtschaftliche Flächen* gezählt. Der Dauersiedlungsraum ist die Summe dieser Kategorien. Die Landesbedeckungsklassen *Wälder und naturnahe Flächen*, *Feuchtflächen* sowie *Wasserflächen* werden nicht zum Dauersiedlungsraum gezählt. Die Bevölkerungsdichte der Gemeinden wird bezogen auf die Nutzungsklasse *städtisch geprägte Flächen* berechnet. Für Details zu dem CORINE-Landcover Datensatz sei auf das Umwelt Bundesamt (umwelbundesamt.at/rp_corine) verwiesen. Zu Bautätigkeit und Baulandüberhang stehen keine Daten zur Verfügung.

Die Daten zur Bevölkerung werden vom Onlineportal der Statistik Austria bezogen und die Variablen Bevölkerungswachstum und Belastungsindex berechnet. Die Variable Bevölkerungswachstum bezieht sich auf den Zeitraum der letzten zehn Jahre. Der Belastungsindex stellt das Verhältnis der nicht-aktiven (0-19 Jährige und über 65 Jährige) zur potentiell aktiven (20-65 Jährige) Bevölkerung da.

Die Daten zur Wirtschaftsstruktur haben unterschiedliche Bezugsquellen. Die abgestimmte Erwerbsstatistik der Statistik Austria wird zur Berechnung der Akademikerquote, der Arbeitslosenquote und des Pendlersaldos genutzt. Die Akademikerquote ist der Anteil der Wohnbevölkerung mit mindestens einem sekundären Bildungsabschluss. Die Arbeitslosenquote ist der Anteil an arbeitslos gemeldeten Personen bezogen auf alle Erwerbspersonen. Der Penderlsaldo ist das Verhältnis von Erwerbstätigen am Arbeitsort zu Erwerbstätigen am Wohnort in einer Gemeinde.

Die Daten zu den Beschäftigten kommen aus der Arbeitsstättenzählung der Statistik Austria. Ein entsprechender Datensatz liegt dem Fachbereich für Stadt- und Regionalforschung vor. Dieser untergliedert die Zahl der Beschäftigten je Gemeinde in ÖNACE Kategorien (A-S). Die Variable Beschäftigte in der Landwirtschaft ist der Anteil der in der ÖNACE Kategorie A (Land- und Forstwirtschaft) Beschäftigen im Verhältnis zu allen Beschäftigten in einer Gemeinde. Die Variable Beschäftigte im Tourismus ist der Anteil der in ÖNACE Kategorie I (Beherbergung und Gastronomie) Beschäftigen im Verhältnis zu allen Beschäftigten.

Ein Datensatz zum Einkommensniveau liegt dem Fachbereich in Form der Lohnsteuerstatistik der Statistik Austria ebenfalls vor. Aus diesem werden die Bruttobezüge für Arbeitnehmer und Pensionisten sowie die Anzahl der lohnsteuerpflichtigen Personen ersichtlich. Daraus kann das durchschnittliche Bruttoeinkommen je lohnsteuerpflichtiger Person errechnet werden. Dieser Wert wird in der Arbeit als Annäherung für die Variable Einkommen herangezogen.

Daten zu Übernachtungen liegen nicht vollständig für alle Gemeinden vor. Deshalb fließt die Variable Anzahl der Übernachtungen nicht in die Modellbildung ein. Allerdings ist anzunehmen, dass die Variable Beschäftigte im Tourismus eine ähnliche Information beisteuert wie die Variable Anzahl der Übernachtungen.

Lagevariablen

Aufgrund nicht zur Verfügung stehender Datensätze zu Fahrzeiten für den öffentlichen Verkehr können nur Datensätze zur Erreichbarkeit einer Gemeinde mit dem PKW weiter betrachtet werden.

Für den Datensatz zur Erreichbarkeit mit dem PKW wird auf die Ergebnisse des bereits erwähnten Projetks der Statistik Austria zurückgegriffen. Ziel des Projekts mit dem Namen *Grid-based indicators of accessibility of public utility infrastructure* war es, Erreichbarkeitsindikatoren für

öffentliche Infrastruktur auf Rasterbasis in Österreich zu modellieren (Saul 2017). Somit fließen nicht die Fahrzeiten selbst, sondern Erreichbarkeitsindikatoren in die Modellbildung ein.

Bei der Indikatorenbildung des Projekts der Statistik Austria wurde eine Vielzahl an Einrichtungen zu öffentlichen Infrastrukturen berücksichtigt. Die Daten zu öffentlichen Infrastrukturen wurden in fünf Gruppen gegliedert. Diese sind Einzelhandel, Bildungseinrichtungen, Gesundheitseinrichtungen, Freizeiteinrichtungen und Sicherheitseinrichtungen. Unter Einzelhandel wurden unter anderem Lebensmittelläden, Banken und Tankstellen subsumiert, unter Gesundheitseinrichtungen Ärzte, Krankenhäuser und Apotheken, unter Bildungseinrichtungen Kindergärten und Schulen, unter Sicherheitseinrichtungen Polizeistationen und unter Freizeiteinrichtungen Kinos, Museen und Restaurants. Für die genaue Aufzählung sei auf den Endbericht zum Projekt in Saul (2017, S.6f) verwiesen.

Anschließend wurden die Fahrzeiten von jeder bewohnten Rasterzelle (1km Raster) zu diesen Infrastruktureinrichtungen mittels Netzwerkanalyse errechnet. Für die Berechnung der Erreichbarkeitsindikatoren je Gruppe waren diese Fahrzeiten dann der Input für eine *principal components analyses* (PCA). Das Ergebnis der PCA waren die Erreichbarkeitsindikatoren je Infrastruktur-Gruppe und ein Gesamterreichbarkeitsindikator je bewohnter Rasterzelle.

Die Erreichbarkeitsindikatoren wurden skaliert auf den Wertebereich 0-100. Die Skalierung erfolgte so, dass der Rasterzelle mit der schlechtesten Erreichbarkeit (längste Fahrzeit zu den jeweiligen Infrastrukturen) der Wert 0 und der Rasterzelle mit der besten Erreichbarkeit (kürzeste Fahrzeit zu den jeweiligen Infrastrukturen) der Wert 100 zugewiesen wurde. Die anderen Rasterzellen wurden im Verhältnis zu ihren Fahrzeiten in diesem Wertebereich skaliert.

Da für den Zweck dieser Arbeit ein Wert je Gemeinde erforderlich ist, wird mittels *spatial join* der Mittelwert über alle bewohnten Zellen je Gemeinde gebildet. Somit ergibt sich ein Datensatz mit fünf Erreichbarkeitsindikatoren je Infrastruktur-Gruppe und einem Gesamterreichbarkeitsindikator für jede Gemeinde. Diese Mittelwertbildung über mehrere Rasterzellen führt auch dazu, dass die Spannweite der Erreichbarkeitsindikatoren nicht wie bei den Rasterzellen zwischen 0 und 100, sondern geringfügig darunter beziehungsweise darüber liegt.

Die benötigten Geodaten der Gemeinden werden vom Open Data Portal Österreichs (data.gv.at) bezogen. Zusätzlich liegt dem Fachbereich ein Rasterdatensatz zur österreichischen Bevölkerung vor. Dieser wird genutzt, um die Gemeindemittelpunkte gewichtet nach bewohnten Zellen zu berechnen. Diese werden benötigt, um Nachbarschaftsbeziehungen und Distanzen zwischen Gemeinden zu definieren beziehungsweise zu berechnen.

Durch die begrenzte Datenverfügbarkeit kann nicht gewährleistet werden, dass alle Variablen eine eindeutige zeitliche Homogenität aufweisen. Die Daten stammen somit aus einem Zeitraum von 2012-2018. Für die Zwecke dieser Arbeit ist diese zeitliche Homogenität allerdings ausreichend.

Alle oben beschriebenen Datensätze werden anhand der Gemeinde ID miteinander verbunden (*table join*). In die Modellbildung können nur Gemeinden einfließen, bei denen Werte zur abhängigen Variable (Bodenpreis), sowie zu allen erklärenden Variablen vorliegen. Die Bundeshauptstadt Wien wird in ihre 23 Bezirke aufgeteilt. Somit ergeben sich 2120 mögliche Beobachtungen (alle 2097 Gemeinden zuzüglich 23 Wiener Bezirke), für die Daten vorliegen könnten. Allerdings sind lediglich 1667 davon vollständig und können für die Modellbildung herangezogen werden. Da sich die fehlenden Werte gleichmäßig über den Untersuchungsraum verteilen, kann deren Fehlen im Rahmen dieser Arbeit vernachlässigt werden. Es ergibt sich somit eine Datenmatrix mit 1667 Beobachtungen und 18 Variablen.

4.2 Deskriptive Datenanalyse

Im folgenden Kapitel werden alle Variablen einzeln betrachtet und eine deskriptive Analyse durchgeführt. Ziel ist es, einen Überblick über die Datengrundlage zu schaffen.

4.2.1 Abhängige Variable - Bodenpreis

Betrachtet man die räumliche Verteilung der abhängigen Variable Bodenpreis in Abbildung 4.1, ist das zu erwartende Muster erkenntlich, dass rund um die Ballungszentren relativ höhere Bodenpreise gemeldet werden als in peripheren Teilen des Untersuchungsraums. Unter der Annahme, dass in und um Ballungszentren eine günstigere Wirtschaftsstruktur vorhanden ist, unterstützt diese Erkenntnis die in Kapitel 2.2.3 aufgestellte Hypothese, dass bei einer solchen günstigen Wirtschaftsstruktur höhere Bodenpreise zu erwarten sind.

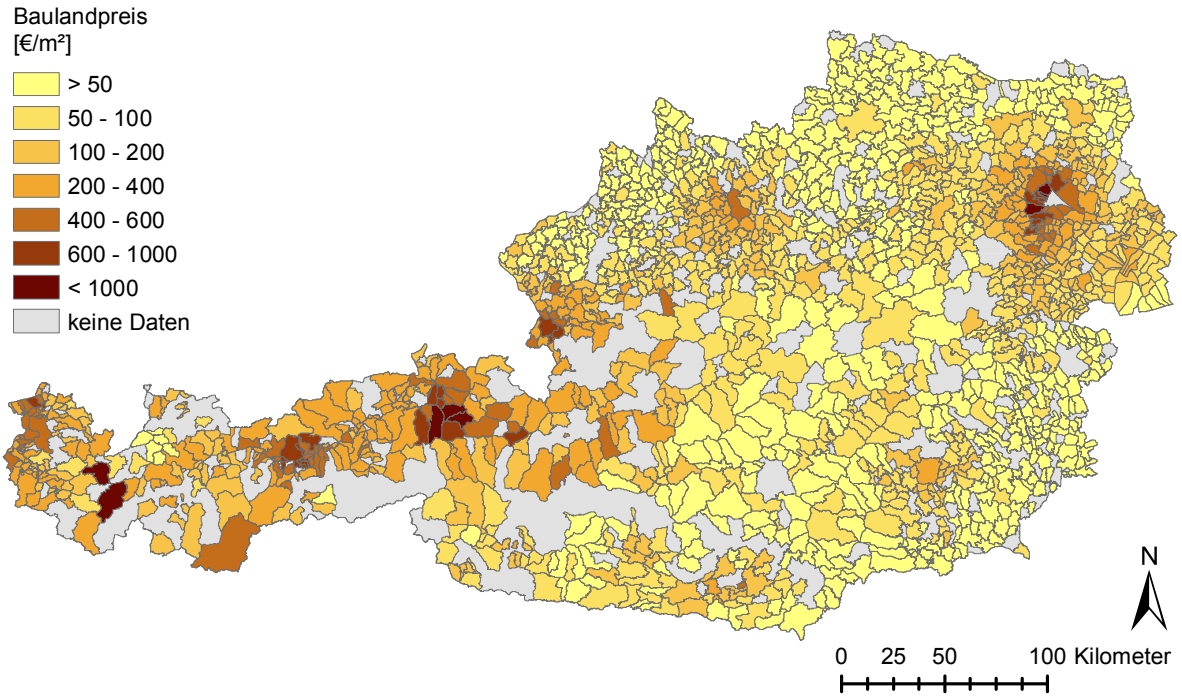

Abb. 4.1: Verortung Bodenpreis

Zusätzlich ist erkenntlich, dass der Baulandpreis im Westen tendenziell höher ist als im Osten Österreichs. Man spricht in diesem Zusammenhang von einem Trend in den Bodenpreisdaten. Dieser Trend könnte mit der höheren touristischen Nutzung in den westlichen Bundesländern zusammenhängen. Betrachtet man den Boxplot in Abbildung 4.2, bestätigt sich dieser Trend. Die drei westlichen Bundesländer Vorarlberg, Tirol und Salzburg weisen einen höheren Mittelwert auf. Wien bildet in diesem Zusammenhang eine Ausnahme, da das gesamte Bundesland als Ballungszentrum zu bewerten ist.

Um eine annähernde Normalverteilung der abhängigen Variable zu gewährleisten, wird der Bodenpreis logarithmiert und fließt so in die Modellbildung ein. In Abbildung 4.2 ist dies bereits zu erkennen.

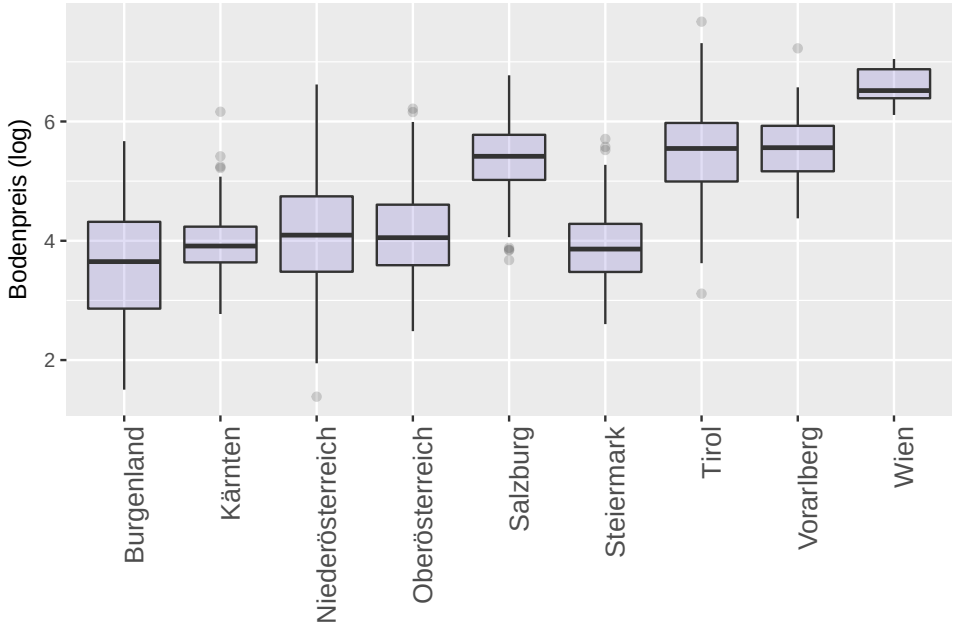

Abb. 4.2: Vergleich Bodenpreis Bundesländer

4.2.2 Erklärende Variablen

Die erklärenden Variablen werden differenziert nach Nachbarschafts- und Lagevariablen betrachtet. Es werden deren Einheiten und die statistische Verteilung beschrieben. Da ausschließlich quantitative Variablen verwendet werden, bietet es sich an, die Zusammenhänge der einzelnen Variablen mit dem Bodenpreis mittels Scatterplot darzustellen.

Nachbarschaftsvariablen

Variablen zur Flächennutzung

Betrachtet man den Zusammenhang zwischen der Bevölkerungsdichte und dem Bodenpreis in Abbildung 4.3 links, wird kein eindeutiger Zusammenhang deutlich. Dies könnte darin begründet sein, dass die Berechnung aufgrund der groben Auflösung der CORINE Landcover Daten fehleranfällig ist. Ist ein bebautes Gebiet zu klein, wird es in den CORINE Daten nicht aufgenommen. Dadurch wird des bebaute Gebiet fälschlicherweise verkleinert und die Bevölkerungsdichte steigt.

Die Anteile von Grünland und Bauland werden in Prozent des Dauersiedlungsraumes angegeben und summieren sich somit zu Eins. Es zeigt sich, dass der Grünlandanteil in den eher flachen Gemeinden im Osten mit großem Dauersiedlungsraum am höchsten ist. In Ballungszentren und eher gebirgigem Terrain im Westen ist der Baulandanteil am Dauersiedlungsraum höher. Dieser Sachverhalt ist auch bei der räumlichen Verteilung der Bodenpreisdaten zu erkennen. Es kann ein Zusammenhang in Abbildung 4.3 rechts erkannt werden. Da sich beide Variablen auf Eins ergänzen, bieten sie die gleiche Information. Somit muss nur eine der Variablen in die Modellbildung mit einfließen.

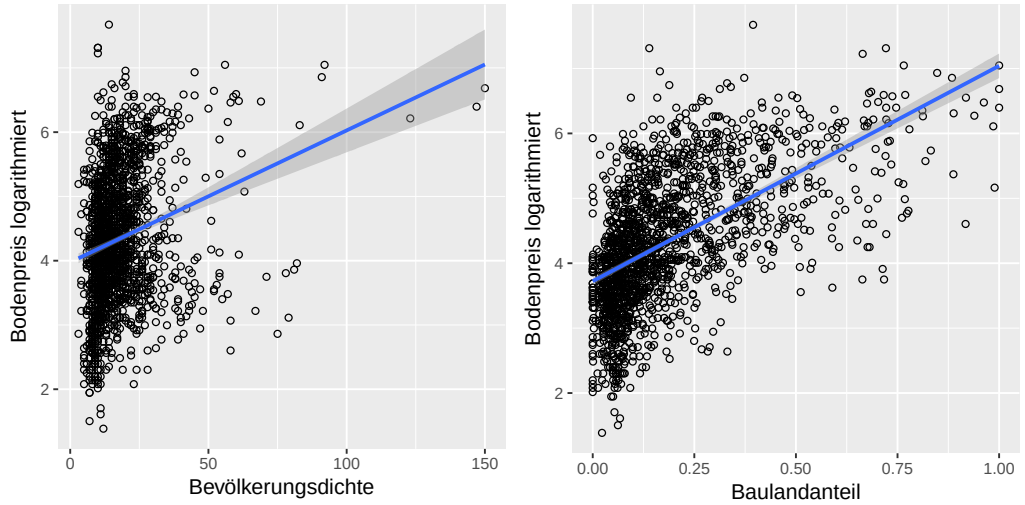

Abb. 4.3: Zusammenhang Variablen zur Flächennutzung - Bodenpreis

Variablen zur Bevölkerungsstruktur

Das Bevölkerungswachstum wird ebenfalls in Prozent angegeben. Es kann somit positiv im Falle eines Einwohnerzuwachses oder negativ im Falle eines Schrumpfungsprozesses sein. Tendenziell ist in Abbildung 4.4 links zu erkennen, dass schrumpfende Gemeinden einen niedrigeren Bodenpreis aufweisen als stark wachsende Gemeinden. Dieser Sachverhalt entspricht der zuvor aufgestellten Hypothese, dass ein Ansteigen der Bevölkerung zu höheren Bodenpreisen führt.

Der Belastungsindex hat einen Wertebereich zwischen 0,57 und 1,58. Liegt dieser über 1, so ist der Anteil der nicht-aktiven Bevölkerung größer als jener der potentiell aktiven Bevölkerung. Der erwartete negative Zusammenhang lässt sich in Abbildung 4.4 rechts erkennen.

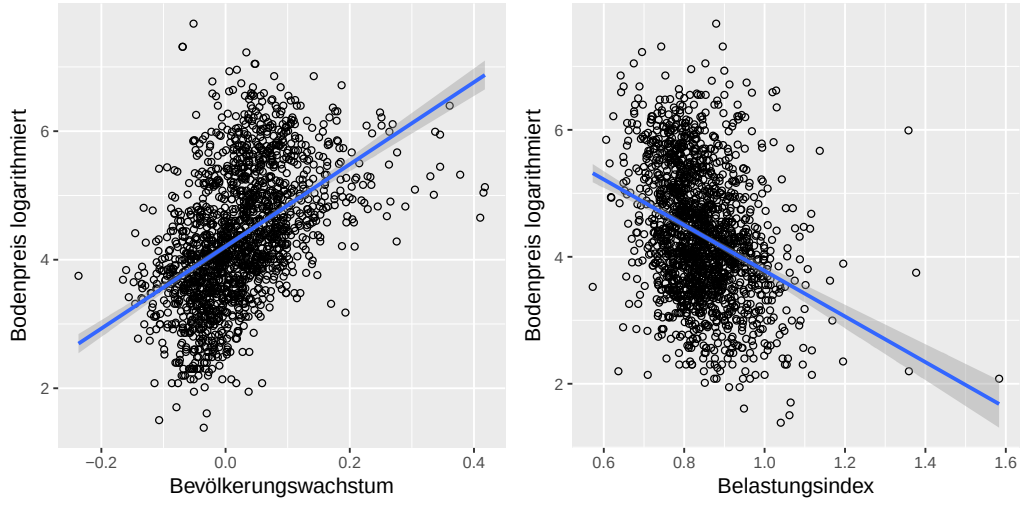

Abb. 4.4: Zusammenhang Variablen zur Bevölkerungsstruktur - Bodenpreis

Variablen zur Wirtschaftsstruktur

Die Beschäftigten in der Landwirtschaft werden in Prozent aller Erwerbstätigen einer Gemeinde angegeben. Der vermutete negative Zusammenhang ist in Abbildung 4.5 links oben geringfügig erkennbar.

Die Beschäftigten im Tourismus werden, wie die Beschäftigten in der Landwirtschaft, in Prozent aller Erwerbstätigen gemessen. Ein positiver Zusammenhang wird vermutet. Ein solcher lässt sich allerdings nur in sehr geringem Ausmaß in Abbildung 4.5 rechts oben feststellen.

Die Akademikerquote wird in Prozent aller Einwohner angegeben. Bei dieser ist ein Stadt-Land-Gefälle sichtbar. Da ein solches auch bei den Bodenpreisen selbst erkennbar ist, wird ein positiver Zusammenhang erwartet. Dies ist in geringem Maß in Abbildung 4.5 links mittig ersichtlich.

Gemessen wird die Arbeitslosenquote in Prozent der Erwerbspersonen. Zuvor wurde die Hypothese aufgestellt, dass eine hohe Arbeitslosenquote mit niedrigen Bodenpreisen einhergeht. Betrachtet man allerdings den Scatterplot in Abbildung 4.5 rechts mittig, kann kein eindeutiger Zusammenhang erkannt werden.

Der Pendlersaldo wird mittels Verhältniszahl zwischen den Erwerbstätigen am Arbeitsort und den Erwerbstätigen am Wohnort gemessen. Er liegt im Wertebereich von 0,13 bis 3,18. Ein Wert über 1 bedeutet, dass es mehr Erwerbstätige am Arbeitsort als Erwerbstätige am Wohnort in einer Gemeinde gibt. In solch einem Zusammenhang spricht man von einer Einpendlergemeinde. Es wurde zuvor die Hypothese aufgestellt, dass der Bodenpreis in Einpendlergemeinden höher ist, als in Auspendlergemeinden. Dieser Zusammenhang kann nicht eindeutig bestätigt werden, wie Abbildung 4.5 links unten zeigt.

Das Einkommen wird in Bruttobezüge pro lohnsteuerpflichtiger Person gemessen. Hier wird ein positiver Zusammenhang mit dem Bodenpreis angenommen. Dieser lässt sich in Abbildung 4.5 rechts unten auch feststellen.

Tabelle 4.1 gibt einen Überblick über alle Nachbarschaftsvariablen, deren Einheiten und Wertebereiche.

Lagevariablen

Wie bereits erwähnt, handelt es sich bei den Daten zur Erreichbarkeit nicht um Fahrzeitdaten, sondern um Erreichbarkeitsindikatoren. Diese weisen, bedingt durch den Vorgang der Datenaufbereitung, einen Wertebereich von 7,0 bis 84,3 auf, wobei eine gute Erreichbarkeit mit hohen Werten beschrieben wird. Diese Tatsache kehrt die in Kapitel 2.2.3 definierten Hypothesen um. Somit wird ein hoher Bodenpreis bei hohen Erreichbarkeitsindikatoren erwartet. Dieser erwartete positive Zusammenhang ist bei allen Erreichbarkeitsvariablen erkennbar. Allerdings ist dieser beispielsweise für die Erreichbarkeit von Sicherheitseinrichtungen weniger stark ausgeprägt.

Bei Erreichbarkeitsvariablen besteht grundsätzlich die Möglichkeit, dass in einem geschätzten Modell Multikollinearität auftritt, da die einzelnen Indikatoren in derselben Gemeinde meist in einem ähnlichen Wertebereich liegen und somit korrelieren. Auch aufgrund dessen kann das Fehlen von Erreichbarkeitsindikatoren für den öffentlichen Verkehr vernachlässigt werden.

Tabelle 4.2 fasst die Erreichbarkeitsvariablen und deren Wertebereich zusammen.

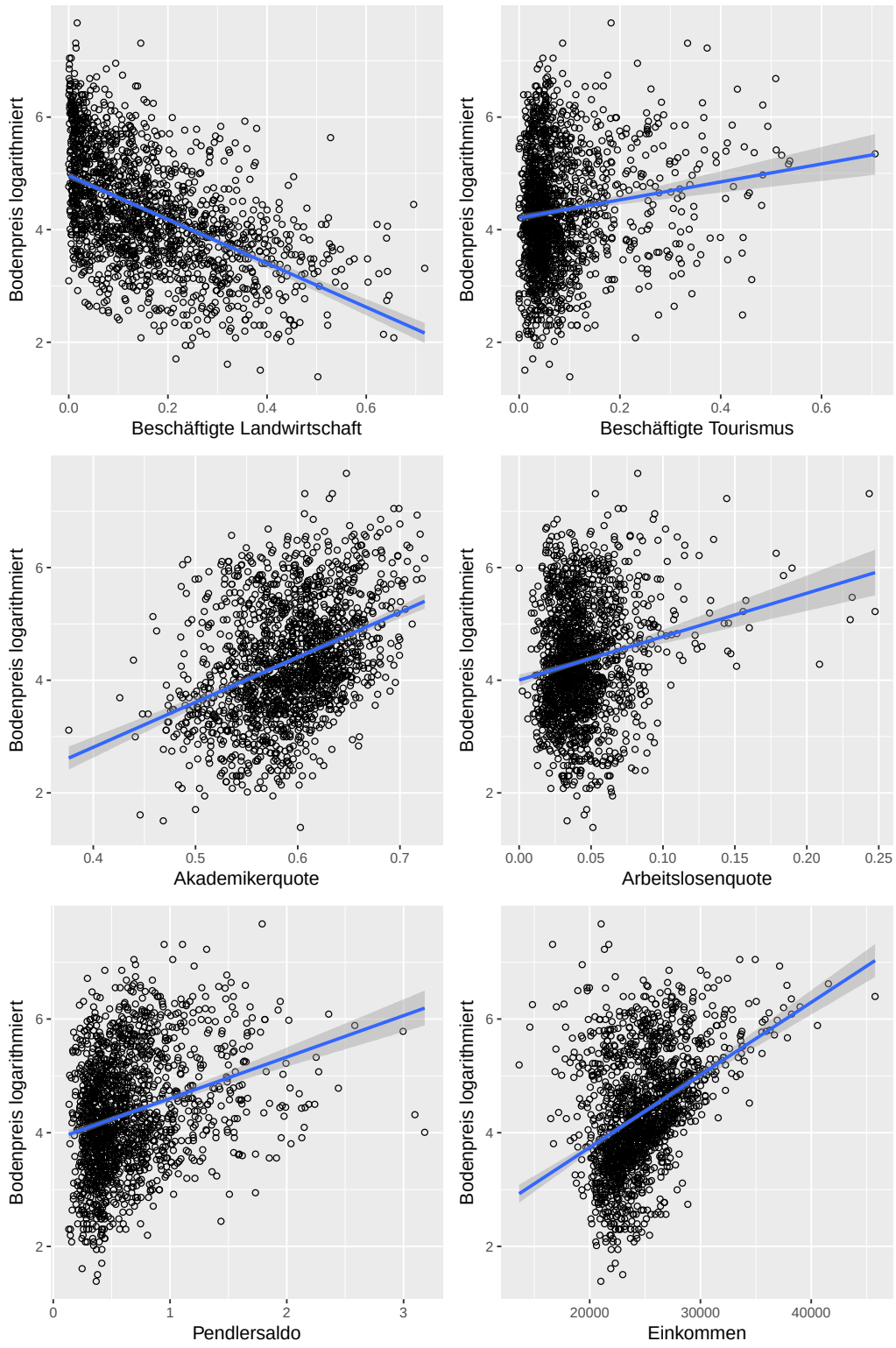

Abb. 4.5: Zusammenhang Variablen zur Wirtschaftsstruktur - Bodenpreis

Tab. 4.1: Nachabrschaftsvariablen - desktiptive Statistik

Variable	Einheit	Min	Max	Mittel	St.Abw
Variablen zur Flächennutzung					
Bevölkerungsdichte	[Ew/ha]	2,91	149,57	17,82	11,62
Baulandanteil	[% von Dsr]	0,00	1,00	0,19	0,18
Grünlandanteil	[% von Dsr]	0,00	1,00	0,81	0,18
Variablen zur Bevölkerungsstruktur					
Bevölkerungswachstum	[%] (10 Jahre)	-0,24	0,42	0,02	0,08
Belastungsindex	[(0-19)+(65+)]/[20-65]	0,57	1,58	0,84	0,09
Variablen zur Wirtschaftsstruktur					
Beschäftigte Landwirtschaft	[% von Et]	0,00	0,72	0,16	0,14
Beschäftigte Tourismus	[% von Et]	0,00	0,71	0,08	0,08
Akademikerquote	[% von Ew]	0,38	0,72	0,59	0,05
Arbeitslosenquote	[% von Ep]	0,00	0,25	0,04	0,02
Pendlersaldo	[(Et Arb/Et Woh)]	0,13	3,18	0,64	0,39
Einkommen	[€/Ep]	13.629	45.767	24.711	3.282

Ew=Einwohner; Et=Erwerbstätige; Ep=Erwerbspersonen; ha=Hekta; Dsr=Dauersiedlungsraum; Arb=Arbeitsort; Woh=Wohnort

Tab. 4.2: Lagevariablen - deskriptive Statistik

Variable	Min	Max	Mittel	St.Abw
Variablen zur Erreichbarkeit PKW				
Erreichbarkeit Einzelhandelseinrichtungen	7,60	72,61	34,15	9,61
Erreichbarkeit Bildungseinrichtungen	11,15	66,41	32,88	8,92
Erreichbarkeit Gesundheitseinrichtungen	14,40	72,26	38,10	9,37
Erreichbarkeit Freizeiteinrichtungen	10,92	73,20	34,416	9,66
Erreichbarkeit Sicherheitseinrichtungen	7,04	68,95	34,24	10,63
Erreichbarkeit gesamt	11,34	84,30	40,88	10,45

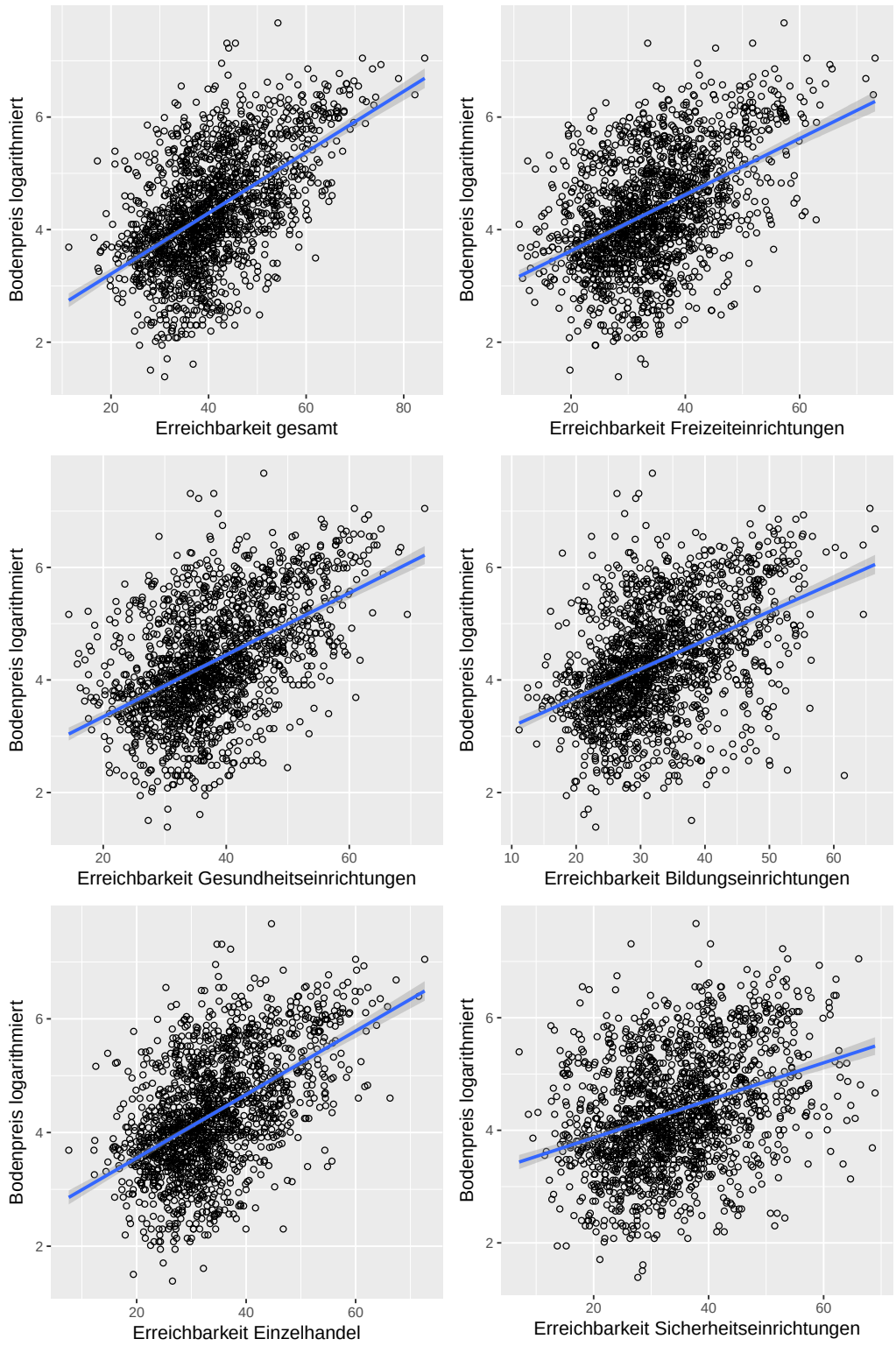

Abb. 4.6: Scatterplot Erreichbarkeitsindikatoren - Bodenpreis

Ob die optisch gefundenen Zusammenhänge ausreichen, um statistisch signifikante Ergebnisse für die Koeffizienten in den geschätzten Modellen zu erzielen, wird im nächsten Abschnitt 5 Modellbildung ersichtlich.

Kapitel 5

Modellbildung

In diesem Abschnitt werden die in Abschnitt 2 argumentierten und in Abschnitt 3 genauer beschriebenen Modellansätze geschätzt. Der verwendete Datensatz wurde in Abschnitt 4 aufbereitet.

5.1 Ablauf der Modellbildung

Der finale Datensatz besteht aus 1667 Beobachtungen (Gemeinden). In diesem sind die abhängige Variable (Bodenpreis), sowie die Nachbarschafts- und Lagevariablen der Gemeinden zusammengefasst. Um alle geschätzten Modellansätze miteinander zu vergleichen, wird der gesamte Datensatz in einen Trainingsdatensatz und einen Testdatensatz geteilt. Die Aufteilung erfolgt zufällig. 80 Prozent der Beobachtungen kommen in den Trainingsdatensatz, während 20 Prozent den Testdatensatz bilden. Zum Schätzen der Modelle wird nun ausschließlich auf den Trainingsdatensatz zurückgegriffen. Anschließend werden die geschätzten Modelle dazu verwendet, die Bobachtungen aus dem Testdatensatz vorherzusagen. So kann dann der Unterschied zwischen vorhergesagtem und tatsächlichem Bodenpreis und in weitere Folge der Root Mean Square Prediction Error (RMSPE) bestimmt werden. Je kleiner dieser Wert ist, desto kleiner ist die Verzerrung bezogen auf den tatsächlichen Wert. Des weiteren werden die Modelle aus der Ökonometrie (OLS-Modelle und SAR-Modelle) über das Akaike Information Criterion (AIC) verglichen. Auch hier gilt, je niedriger der Wert, desto besser ist der Fit des Modells. Unterschiedliche OLS-Modellspezifikationen werden zusätzlich über das adjusted R^2 verglichen, während bei unterschiedlichen SAR-Modellspezifikationen zusätzlich das log-likelihood herangezogen wird.

Bezogen auf die funktionale Form der Modelle werden semi-log Spezifikationen verwendet. So wird in allen Modellen die abhängige Variable (Bodenpreis) logarithmiert. Dadurch können nichtlineare Abhängigkeiten zwischen erklärenden Variablen und abhängiger Variable in Betracht gezogen werden. Bei semi-log Spezifikationen können die Koeffizienten β als semi-Elastizität interpretiert werden. So bedeutet die Erhöhung einer erklärenden Variable x um eine Einheit eine Zunahme der abhängigen Variablen y um $100 \cdot \beta$ Prozent.

In der Literatur wird bezüglich der Vorgangsweise bei der Suche nach der richtigen Modellspezifikation zwischen *forward stepwise strategies* und *backward stepwise strategies* unterschieden (vgl. Florax et al. 2003). Bei der forward stepwise Strategie werden einfache Modelle geschätzt und getestet, ob alle Bedingungen erfüllt sind. Diese Bedingungen betreffen in erster Linie die Residuen (beispielsweise keine Autokorrelation). Sind diese nicht erfüllt, werden komplexere Modelle geschätzt. Die Komplexität der Modellspezifikation wird so lange erhöht, bis die Bedingungen erfüllt sind. Bei der backward stepwise Strategie werden hingegen zuerst komplexe Modelle geschätzt. Anschließend werden einzelne Parameter dieser Modelle schrittweise Null

gesetzt (constraints). Nun wird getestet, ob sich die Modelle mit den Null gesetzten Parametern signifikant von den komplexen Modellen unterscheiden. Ist dies nicht der Fall, ist das einfachere Modell (mit den Null gesetzten Parametern) zielführend.

In dieser Arbeit wird die klassische forward stepwise Strategie angewandt. So werden zuerst einfache OLS-Modelle geschätzt. Bei der Auswahl der Variablen, werden die in Abschnitt 2 (Grundlagen) erwähnten induktiven und deduktiven Ansätze kombiniert. Einerseits wurden bereits Variablen aus der Theorie abgeleitet. Andererseits werden nun im Modellbildungsprozess verschiedene Variablenkombinationen getestet und verglichen. Somit kann das OLS-Modell mit dem besten Modell-Fit gefunden werden. Die Residuen der OLS-Modelle werden auf die in Abschnitt 3 (Methodik) aufgestellten Bedingungen getestet. Hier ist vor allem auf mögliche räumliche Autokorrelation der Residuen zu achten. Liegt eine solche vor, ist ein OLS-Modell nicht zielführend. Das OLS-Modell mit dem besten Modell-Fit wird anschließend als Ausgangspunkt verwendet, um die räumlichen Modelle (Modellansätze der räumlichen Ökonometrie und der Geostatistik) zu schätzen.

5.2 Geschätzte Regressionsmodelle

5.2.1 OLS Modelle

Es werden drei verschiedene OLS Modelle geschätzt. In das erste Modell (OLS_1) werden alle Nachbarschaftsvariablen aufgenommen. In das zweite Modell (OLS_3) werden alle Lagevariablen aufgenommen. Im dritten Modell (OLS_2) werden schließlich alle Variablen, die in den ersten beiden Modellen einen signifikanten Einfluss aufweisen, zusammengefasst.

Die Auswahl der Variablen im jeweiligen Modell erfolgt mittels *stepwise selection*. Als Ausgangspunkt dient ein *intercept only* Modell ohne erklärende Variablen. Anschließend wird schrittweise jene Variable hinzugefügt, die den Modell-Fit gemessen am AIC am stärksten erhöht. Dies erfolgt so lange, bis das Hinzufügen einer weiteren Variable den Modell-Fit nicht mehr verbessert. Dieser Prozess wird automatisiert mit der Statistiksoftware *R* für die Modelle OLS_1, OLS_2 und OLS_3 durchgeführt.

Modell OLS_1

In das OLS_1 werden schrittweise alle erklärenden Nachbarschaftsvariablen, mit Ausnahme der Variable Arbeitslosenquote, aufgenommen. Die Ergebnisse werden in Tabelle 5.1 dargestellt. Das Modell ist mit einer F-Statistik von 192,6 und einem p-Wert von 0,000 signifikant. Als Signifikanzniveau wird weiters ein p-Wert von 0,05 angenommen. Das adjusted R^2 liegt bei 0,564. Das bedeutet, dass rund 56 Prozent der Varianz des Bodenpreises erklärt werden können. Das Akaike Information Criterion (AIC) liegt bei 2687,477. Der Root Mean Square Prediction Error (RMSPR), also der Fehler, der sich bei der Vorhersage des Testdatensatzes ergibt, liegt bei 0,692.

Bei der Bewertung der Modellqualität gilt es, die Vorzeichen der Koeffizienten zu beachten (theoretische Modellgültigkeit). Diese sollten dem zuvor aufgestellten theoretischen Modell entsprechen. Bei den Variablen Grünlandanteil, Belastungsindex und Beschäftigte in der Landwirtschaft werden negative Zusammenhänge erwartet. Die Ergebnisse des Modells zeigen, dass dies der Fall ist. Die Höhe der Koeffizienten kann nur in Zusammenhang mit dem Skalenniveau der einzelnen

Tab. 5.1: Ergebnisse OLS_1

	Abhängige Variable = *log* Bodenpreis	N=1334
	OLS_1	
Erklärende Variablen	Koeff.	VIF
Konstante	4,137 $(0,333)^{***}$	
Nachbarschaftsvariablen		
Bevölkerungsdichte	0,013 $(0,002)^{***}$	1,10
Grünlandanteil	$-1,760$ $(0,134)^{***}$	1,65
Bevölkerungswachstum	3,295 $(0,225)^{***}$	1,60
Belastungsindex	$-1,666$ $(0,189)^{***}$	1,26
Beschäftigte Landwirtschaft	$-0,794$ $(0,189)^{***}$	2,02
Beschäftigte Tourismus	1,857 $(0,225)^{***}$	1,19
Akademikerquote	3,599 $(0,493)^{***}$	1,75
Pendlersaldo	0,178 $(0,055)^{**}$	1,48
Einkommen	0,000 $(0,000)^{*}$	2,38
Modellperformance		
R^2	0,567	
Adjusted R^2	0,564	
Akaike Inf. Crit.	2.687,477	
RMSPE	0,692	
Residuenanalyse		
Morans'I	0,574 ***	
Breusch-Pagan-Test	76,871 ***	
Anmerkung:	*p<0,1; **p<0,05; ***p<0,01	

Variablen interpretiert werden. Betrachtet man beispielsweise die Variable Bevölkerungsdichte, heißt das, dass bei einem Anstieg der Bevölkerungsdichte um eine Einheit (1 Einwohner pro Hektar städtisch geprägtem Gebiet) der Bodenpreis tendenziell um 0,013 Prozent steigt.

Weiters gilt es, das Signifikanzniveau der einzelnen Koeffizienten zu beachten. Die Standardabweichung der Koeffizienten ist in der Klammer ablesbar. Mit Ausnahmen der Koeffizienten der Variablen Einkommen und Pendlersaldo sind alle Koeffizienten auf einem Signifikanzniveau von 0,01 statistisch signifikant. Bei dem Koeffizienten zur Variable Pendlerslado ist allerdings ein Signifikanzniveau von 0,05 feststellbar. Die Variable Einkommen wird aufgrund eines zu geringen Signifikanzniveaus im weiteren Modellbildungsprozess nicht mehr berücksichtigt.

Multikollinearität ist für keine der Variablen feststellbar. Der Varianz Infation Factor (VIF) liegt unter dem definierten Grenzwert von 10.

Betrachtet man im Zuge der Residuenanalyse den Q-Q Plot (siehe Anhang), kann die Normalverteilung der Residuen angenommen werden. Es sind nur geringe Abweichungen zu erkennen, die bei ausreichend großen Datensätzen zu vernachlässigen sind (Haase 2011, S.100). Allerdings muss die Annahme von Homoskedastizität beim Betrachten des Residuenplots verworfen werden. Dies wird auch durch das Signifikanzniveau des Breusch-Pagan Tests bestätigt.

Zusätzlich muss die Annahme, dass keine räumliche Autokorrelation der Residuen vorliegt, verworfen werden. Diese ist mit einem Morans'I von 0,574 deutlich gegeben und auf einem Signifikanzniveau von 0,01 signifikant. Für die Berechnung des Morans'I ist bereits die Definition einer räumlichen Gewichtungsmatrix notwendig. Hierfür wird eine k-nearest neighbors Gewichtungsmatrix mit zehn Nachbarn definiert.

Modell OLS_2

Ein zweites OLS Modell (OLS_2) wird ausschließlich mit den Lagevariablen geschätzt. Allerdings gilt es zu beachten, dass die Variable Erreichbarkeit gesamt sich aus den anderen Variablen zur Erreichbarkeit zusammensetzt. Um die Interpretierbarkeit zu gewährleisten, kann somit entweder ausschließlich die Variable zur gesamten Erreichbarkeit oder die Variablen zu den Einzelerreichbarkeiten in das Modell einfließen. Um zu beantworten, welche der beiden Möglichkeiten zielführender ist, werden beide Möglichkeiten in eigenen OLS Modellen ($OLS_{2.1}$, $OLS_{2.2}$) geschätzt. Das Modell mit dem besseren Modell-Fit wird bei der Erstellung des OLS_3 Modells berücksichtigt.

Tabelle 5.2 zeigt die Modellergebnisse für das Modell $OLS_{2.1}$, bei dem die Erreichbarkeit gesamt als einzige erklärende Variable einfließt. Die F-Statistik des Modells ist mit einem Wert von 575,2 statistisch signifikant. Das adjusted R^2 erreicht einen Wert von 0,301. Dies bedeutet, dass rund 30 Prozent der Varianz des Bodenpreises erklärt werden können. Das Akaike Information Criterion (AIC) liegt bei 3309,071 und der Root Mean Square Prediction Error (RMSPR) bei 0.907. Das Vorzeichen des Koeffizienten stimmt mit der aufgestellten Hypothese überein. Dieser ist auf einem Signifikanzniveau von 0,01 statistisch signifikant.

Der Q-Q Plot zeigt, dass eine Normalverteilung der Residuen angenommen werden kann. Zudem liegt Homoskedastizität vor. Dies wird durch den Residuenplot einerseits und den Breuch-Pagan Test andererseits, dessen Ergebnis nicht signifikant ist, bestätigt. Allerdings muss die Nullhypothese, dass keine räumliche Autokorrelation in den Residuen vorliegt, verworfen werden. Der Morans'I ist mit einem Wert von 0,907 sehr hoch und statistisch auf einem Signifikanzniveau von 0,01 signifikant.

Tab. 5.2: Ergebnisse $OLS_{2.1}$

	Abhängige Variable = *log* Bodenpreis	N=1334
	$OLS_{2.1}$	
Erklärende Variablen	Koeff.	VIF
Konstante	2,198 $(0,092)^{***}$	
Lagevariable		
Erreichbarkeit gesamt	0,053 $(0,002)^{***}$	–
Modellperformance		
R^2	0,301	
Adjusted R^2	0,301	
Akaike Inf. Crit.	3.309,071	
RMSPE	0,907	
Residuenanalyse		
Morans'I	0,774 ***	
Breusch-Pagan-Test	0,386	
Anmerkung:	*p<0,1; **p<0,05; ***p<0,01	

Bei der Auswahl der Variablen für das Modell $OLS_{2.2}$ wird ebenfalls wie zuvor beim Modell OLS_1 mittels stepwise selection vorgegangen. Es zeigt sich, dass die Variable Erreichbarkeit von Bildungseinrichtungen die Modellgüte, gemessen am AIC, nicht erhöht und somit nicht ins Modell einfließt. Tabelle 5.3 fasst die Ergebnisse zusammen.

Das Modell ist mit einer F-Statistik von 141.0 auf einem Signifikanzniveau von 0,01 statistisch signifikant. Im Vergleich mit dem zuvor geschätzten Modell $OLS_{2.1}$, in das nur die Variable Erreichbarkeit gesamt einfloss, ist das adjusted R^2 mit einem Wert von 0,296 niedriger. Auch das Akaike Information Criterion (AIC) deutet mit einem Wert von 3322,015 auf einen geringeren Modell-Fit hin. Der Root Mean Square Prediction Error (RMSPR) ist ebenfalls mit 0,911 geringfügig größer.

Die Vorzeichen der Koeffizienten entsprechen den aufgestellten Hypothesen. Allerdings sind lediglich die Koeffizienten der Variablen Erreichbarkeit von Einzelhandel und Erreichbarkeit von Freizeiteinrichtungen auf einem Signifikanzniveau von 0,05 statistisch signifikant. Multikollinearität ist für keine der Variablen feststellbar. Der Varianz Inflation Factor (VIF) liegt unter dem definierten Grenzwert von 10.

Der Q-Q Plot zeigt, dass eine Normalverteilung der Residuen angenommen werden kann. Der Breusch-Pagan Test ist auf einem Signifikanzniveau von 0,05 statistisch nicht signifikant. Dies bedeutet, dass die Nullhypothese, dass Homoskedastizität vorliegt, angenommen werden kann. Allerdings liegt auch in dieser Modellspezifikation räumliche Autokorrelation in den Residuen vor. Der Morans'I ist bei einem Wert von 0,767 statistisch signifikant.

Aufgrund der besseren Modellgüte von Modell $OLS_{2.1}$ gegenüber Modell $OLS_{2.2}$, fließt ausschließlich die Variable Erreichbarkeit gesamt als Lagevariable in das Modell OLS_3 ein.

Tab. 5.3: Ergebnisse $OLS_{2.2}$

	Abhängige Variable = *log* Bodenpreis	N=1334
	$OLS_{2.2}$	
Erklärende Variablen	Koeff.	VIF
Konstante	2,257 $(0,099)^{***}$	
Lagevariablen		
Erreichbarkeit Einzelhandel	0,037 $(0,005)^{***}$	4,86
Erreichbarkeit Gesundheitseinrichtungen	0,012 $(0,006)^{*}$	5,17
Erreichbarkeit Freizeiteinrichtungen	0,017 $(0,004)^{***}$	2,23
Erreichbarkeit Sicherheitseinrichtungen	0,007 $(0,003)^{*}$	2,10
Modellperformance		
R^2	0,298	
Adjusted R^2	0,296	
Akaike Inf. Crit.	3.322,015	
RMSPE	0,911	
Residuenanalyse		
Morans'I	0,767 ***	
Breusch-Pagan-Test	7,867 *	
Anmerkung:	*p<0,1; **p<0,05; ***p<0,01	

Modell OLS_3

In dem Modell OLS_3 werden nun alle Nachbarschaftsvariablen und Lagevariablen mit Koeffizienten auf einem Signifikanzniveau von 0,05 zusammengefasst. Das stepwise selection Verfahren zeigt, dass alle verbleibenden Variablen den Modell-Fit gemessen am AIC erhöhen. Tabelle 5.4 fasst die Modellergebnisse zusammen. Mit einer F-Statistik von 194,3 ist das Modell auf einem Signifikanzniveau von 0,01 statistisch signifikant. Das adjusted R^2 liegt bei 0,566. Das bedeutet, dass rund 57 Prozent der Varianz des Bodenpreises erklärt werden können. Das Akaike Information Criterion (AIC) liegt bei 2680,751. Der Root Mean Square Prediction Error (RMSPR) liegt bei 0,691.

Die Koeffizienten aller erklärenden Variablen besitzen die erwarteten Vorzeichen und sind auf einem Signifikanzniveau von 0,05 statistisch signifikant. Es liegt keine Multikollinearität vor. Der Varianz Inflation Factor liegt bei jeder Variable unter dem Grenzwert von 10.

Betrachtet man im Zuge der Residuenanalyse den Q-Q Plot, kann die Normalverteilung der Residuen angenommen werden. Allerdings muss die Annahme von Homoskedastizität beim Betrachten des Residuenplots verworfen werden. Dies wird auch durch das Signifikanzniveau des Breusch-Pagan Tests bestätigt. Zusätzlich muss die Nullhypothese, dass keine räumliche Autokorrelation der Residuen vorliegt, verworfen werden.

Im Verglich aller geschätzten OLS Modelle zeigt sich erwartungsgemäß, dass das Modell OLS_3 den höchste Modell-Fit aufweist. Es hat das größte adjusted R^2, das geringste AIC und den

Tab. 5.4: Ergebnisse OLS_3

	Abhängige Variable = *log* Bodenpreis	N=1334
	OLS_3	
Erklärende Variablen	Koeff.	VIF
Konstante	3,820 $(0,352)^{***}$	
Nachbarschaftsvariablen		
Bevölkerungsdichte	0,012 $(0,002)^{***}$	1,11
Grünlandanteil	$-1,580$ $(0,147)^{***}$	2,01
Bevölkerungswachstum	3,228 $(0,289)^{***}$	1,54
Belastungsindex	$-1,627$ $(0,224)^{***}$	1,26
Beschäftigte Landwirtschaft	$-0,701$ $(0,192)^{***}$	2,08
Beschäftigte Tourismus	1,874 $(0,219)^{***}$	1,13
Akademikerquote	4,020 $(0,416)^{***}$	1,25
Pendlersaldo	0,149 $(0,051)^{**}$	1,45
Lagevariable		
Erreichbarkeit gesamt	$-0,009$ $(0,003)^{***}$	2,41
Modellperformance		
R^2	0,569	
Adjusted R^2	0,566	
Akaike Inf. Crit.	2.680,751	
RMSPE	0,691	
Residuenanalyse		
Morans'I (Residuen)	0,584 ***	
Breusch-Pagan-Test	70,038 ***	
Anmerkung:	*p<0,1; **p<0,05; ***p<0,01	

geringsten RMSPE. Deshalb wird dieses Modell als Ausgangspunkt für die nunmehr betrachteten räumlichen Modellansätze herangezogen.

5.2.2 SAR Modelle

Wie im Kapitel 3 (Methodik) herausgearbeitet, dient die Residuenanalyse dazu herauszufinden, ob räumliche Effekte vorhanden sind, die in der Modellspezifikation zu berücksichtigen sind. Es wurde bereits der Morans'I berechnet, um zu prüfen, ob räumliche Modellspezifikationen erforderlich sind. Tabelle 5.5 hebt das Ergebnis für das Modell OLS_3 nochmals hervor.

Tab. 5.5: Morans'I

Morans'I	Erwartet	Varianz	p-wert
0,5873	-0,002634488	0.000132883	0.00000***

Anmerkung: *p<0,1; **p<0,05; ***p<0,01

Die Morans'I Statistik zeigt eindeutig, dass räumliche Autokorrelation in den Residuen vorliegt. Dies deutet auf das Vorhandensein von räumlichen Effekten, insbesondere räumliche Abhängigkeit hin, die es in der Modellspezifikation zu berücksichtigen gilt. Die zwei Hauptmodellansätze aus der räumlichen Ökonometrie, diese zu berücksichtigen, sind simultaneous autoregressive error und simultaneous autoregressive lag Modellspezifikationen. Die Morans'I Statistik zeigt als diffuser Spezifikationstest lediglich auf, dass räumliche Effekte vorhanden sind, jedoch nicht, mit welcher Modellspezifikation diese Effekte am besten zu berücksichtigen sind. Es geht somit klar hervor, dass simultaneous autoregressive processes (SAR) Modelle zielführend sein können, jedoch nicht, ob es sich um einen simultaneous autoregressive lag processes (SAR_{Lag}) oder einen simultaneous autoregressive error processes (SAR_{Error}) handelt. Dies kann mittels Lagrange Multiplier Test als fokussierter Test geprüft werden. Dieser vergleicht zwei Modellalternativen miteinander und gibt einen Hinweis darauf, wo sich die räumliche Abhängigkeit befindet. Die Tabelle 5.6 fasst die Ergebnisse des Largrange Multiplier Tests zusammen.

Tab. 5.6: Lagrange Multiplier Test

	Wert	p-wert
Lagrange Multiplier (lag)	2154,8	0,00000***
Lagrange Multiplier (error)	2484,9	0,00000***
Robust LM (lag)	383,3	0,00000***
Robust LM (error)	713,3	0,00000***

Anmerkung: *p<0,1; **p<0,05; ***p<0,01

Es zeigt sich, dass die Lagrange Multiplier Tests für beide Alternativen (SAR_{Lag} und SAR_{Error}) hoch signifikant sind. Dasselbe zeigt sich bei den robusten Versionen des Lagrange Multiplier Tests. In diesem Fall kann somit nicht eindeutig abgeleitet werden, welche Modellspezifikation zielführender ist. Auf Grund dessen werden in weiterer Folge SAR_{Error} und SAR_{Lag} Modelle geschätzt und deren Modell-Fit verglichen.

Bevor die SAR-Modelle geschätzt werden können, muss die räumliche Gewichtungsmatrix, die die Nachbarschaftsbeziehungen und somit den räumlichen Prozess modelliert, definiert werden.

Gewichtungsmatrix

Der räumlichen Gewichtungsmatrix kommt eine wichtige Rolle bei der Schätzung von räumlichen Modellen zu, da diese einen großen Einfluss auf das Ergebnis hat. Die Gewichtungsmatrix sollte im Idealfall die räumliche Abhängigkeit widerspiegeln. Sarlas und Axhausen (2014, S.9) schreiben dazu:

> *The first key aspect of proceeding to the estimation of the spatial regression models is to determine the underlying spatial dependence, if any, and incorporate it accordingly in the spatial regression models, in the form of a spatial weight matrix*

Ist über den räumlichen Prozess selbst wenig bekannt, ist es sinnvoll, bei einer einfachen binären Gewichtungsmatrix zu bleiben (vgl. Bavaud 1998 nach Bivand et al. 2013, S.251). Dies bedeutet, dass die Elemente der Matrix ausschließlich die Werte 0 beziehungsweise 1 annehmen, je nachdem, ob zwei Beobachtungen als benachbart definiert werden oder nicht. In dieser Arbeit wird ein distanz-basierter Ansatz (*distanc-band*) mit einem *k-nearest neighbors* Ansatz verglichen.

Bei den distanz-basierten Ansätzen wird die Distanz, ab der Beobachtungen als benachbart gelten und damit in der Gewichtungsmatrix den Wert 1 annehmen, definiert. Diese wird schrittweise erhöht und der Modell-Fit anhand des log-likelihood verglichen. Beim k-nearest neighbors Ansatz wird die Anzahl der Nachbarn schrittweise erhöht und ebenfalls der Modell-Fit mittels log-likelihood verglichen. Als Maß für die Distanz wird die euklidische Distanz zwischen zwei Beobachtungen ermittelt. Es wäre allerdings wünschenswert, dass die tatsächlichen Fahrzeiten (Netzwerkdistanz) zwischen den Beobachtungen ausschlaggebend wären. Im Rahmen dieser Arbeit wird dies aufgrund des hohen Aufwandes und der Datenverfügbarkeit vernachlässigt und die euklidische Distanz herangezogen.

Es werden k-nearest neighbors Gewichtungsmatrizen mit 1-20 Nachbarn für das SAR_{Lag} und das SAR_{Error} Modell getestet. Dies erfolgt automatisiert mit der Statistiksoftware *R*. Es zeigt sich, dass der Modell-Fit für das SAR_{Lag} Modell mit einer Gewichtungsmatrix mit den 7 nächsten Nachbarn am höchsten ist. Beim SAR_{Error} Modell wird der höchste Modell-Fit bei einer Gewichtungsmatrix mit den 12 nächsten Nachbarn erreicht.

Beim Distanz-basierten Ansatz gilt es zu beachten, dass die Anzahl der Nachbarn je Beobachtung (Gemeinden) durch die ungleiche Verteilung im Raum nicht gleich ist. Allerdings muss gewährleistet sein, dass jede Bobachtung mindestens einen definierten Nachbarn aufweist. Damit dies sichergestellt wird, muss das Distanzband, innerhalb dessen Beobachtungen als benachbart gelten, mindestens rund 17 Kilometer betragen. Es zeigt sich, dass bei größeren Distanzbändern der Modell-Fit ohnehin sowohl für das SAR_{Lag} als auch für das SAR_{Error} Modell sinkt.

Modell SAR_{Lag}

Die Ergebnisse für die geschätzten SAR_{Lag} Modelle sind in Tabelle 5.7 ersichtlich. Darin werden die Ergebnisse der beiden räumlichen Gewichtungsmatrizen gegenübergestellt. Es gilt zu beachten, dass durch die Annahme einer spatially lagged Variable in SAR Modellen (vgl. Kapitel 3.1.2) die direkte Interpretation der Koeffizienten nicht möglich ist, da die Veränderung einer Variable für eine Beobachtung somit auch einen Einfluss auf andere Bobachtungen hat (vgl. Anselin 2002).

Tab. 5.7: Ergebnisse SAR_{Lag} Modelle

	Abh. Variable = *log* Bodenpreis	N=1334
	SAR_{Lag} W = knn 7	SAR_{Lag} W = DistBand 17km
Erklärende Variablen	Koeff.	Koeff.
Konstante	−0,190	−0,645 ***
Rho ρ	0,780 ***	0,818 ***
Nachbarschaftsvariablen		
Bevölkerungsdichte	0,003 ***	0,004 ***
Grünlandanteil	−0,232 ***	−0,297 ***
Bevölkerungswachstum	1,025 ***	0,961 ***
Belastungsindex	−0,146	−0,092
Beschäftigte Landwirtschaft	−0,476 ***	−0,325 ***
Beschäftigte Tourismus	0,859 ***	0,853 ***
Akademikerquote	1,539 ***	1,832 ***
Pendlersaldo	0,278 ***	0,307 ***
Lagevariablen		
Erreichbarkeit gesamt	0,007 ***	0,008 ***
Modellperformance		
log-Likelihood	−507,303	−499,759
Akaike Inf. Crit.	1038,600	1021,520
Breusch-Pagan-Test	44,332 ***	69,417 **
RMSPE	0,331	0,326

Anmerkung: *p<0,1; **p<0,05; ***p<0,01

Hervorzuheben ist der Autokorrelationskoeffizient *Rho*. Dieser ist mit einem Wert von 0,780 auf einem Signifikanzniveau von 0,01 statistisch signifikant. Die spatially lagged Variable hat somit einen signifikanten Einfluss. Es zeigt sich, dass das Modell mit der distanz-basierten räumlichen Gewichtungsmatrix einen besseren Modell-Fit aufweist. Der log-Likelihood liegt bei -499,759 im Verglich zu -507,303 bei dem Modell, in dem die k-nearest neighbors Gewichtungsmatrix verwendet wurde. Auch das Akaike Information Criterion (AIC) weist auf diesen Sachverhalt hin.

Der Breuch-Pagan Test gegen Heteroskedastizität ist in beiden SAR_{Lag} Modellen statistisch signifikant. Die Nullhypothese, dass Homoskedastizität vorliegt, muss also verworfen werden.

Modell SAR_{Error}

Die Modellergebnisse in Tabelle 5.8 zeigen, dass der Autokorrealtionskoeffizient *Lambda* in beiden geschätzten SAR_{Error} Modellen auf einem Signifikanzniveau von 0,01 statistisch signifikant ist. Die Modellperformance der beiden SAR_{Error} Modelle mit unterschiedlicher Gewichtungsmatrix ist sehr ähnlich. Das Modell mit der distanz-basierten Gewichtungsmatrix weist einen geringfügig besseren Modell-Fit, gemessen sowohl am log-likelihood als auch am Akaike Information Criterion (AIC), auf.

Auch bei den SAR_{Error} Modellen ist der Breuch-Pagan Test gegen Heteroskedastizität statistisch signifikant. Die Nullhypothese, dass Homoskedastizität vorliegt, muss also verworfen werden. Dies könnte mit vorhandener räumlicher Heterogenität zusammenhängen, die auch möglicherweise durch die räumliche Gewichtungsmatrix mit distanz-basiertem Ansatz in das Modell eingeführt werden kann.

Vergleicht man die SAR_{Lag} und SAR_{Error} Modelle untereinander, kann festgehalten werden, dass die SAR_{Error} Modellspezifikationen einen besseren Modell-Fit aufweisen. Die räumliche Abhängigkeit folgt somit eher einem simultaneous autoregressive error Prozess. Zu betonen ist allerdings, dass in beiden Modellspezifikationen die Modellperformance gegenüber den nicht-räumlichen OLS Modellen deutlich verbessert werden kann. Das Akaike Information Criterion (AIC) kann von 2680,751 für das OLS_3 Modell auf 1021,520 für das bessere SAR_{Lag} Modell, beziehungsweise 931,891 für das bessere SAR_{Error} Modell reduziert werden. Bei der Betrachtung des Root Mean Prediction Error (RMSPR) ist Folgendes zu beachten: Die Vorhersage (out-of-sample Prediction) in SAR Modellen, insbesondere in SAR_{Error} Modellen, zur Berechnung des RMSPE ist komplex und mit vielen Fragestellungen verbunden (vgl. Kelejian und Prucha 2007). Im Rahmen dieser Arbeit wird auf diese Thematik nicht genauer eingegangen. Für die Vorhersage wird die Funktion *predict.sarlm* auf dem Paket *spded* für die Statistiksoftware *R* verwendet (vgl. Bivand et al. 2013). Dieses bietet für die Vorhersage verschiedene Methoden an. Für deren genaue Beschreibung sei auf Goulard et al. (2017) verwiesen. Grundsätzlich stehen für die Vorhersage in SAR_{Error} Modellen weniger Methoden zu Verfügung. In dieser Arbeit wird vereinfacht die Methode ausgewählt, die den geringsten RMSPE errechnet. Trotzdem ist der berechnete RMSPE für die SAR_{Error} Modelle unerwartet hoch. Für die SAR_{Lag} Modelle entspricht dieser allerdings den Erwartungen und ist im Verglich zum OLS_3 Modell, das einen Wert von 0,691 erreicht, mit 0,326 für das bessere SAR_{Lag} Model deutlich geringer.

Tab. 5.8: Ergebnisse SAR_{Error} Modelle

	Abh. Variable = *log* Bodenpreis	N=1334
	SAR_{Error} W = knn 12	SAR_{Error} W = DistBand 17km
Erklärende Variablen	Koeff.	Koeff.
Konstante	1,085	1,016 ***
Lambda λ	0,941 ***	0,946 ***
Nachbarschaftsvariablen		
Bevölkerungsdichte	0,004 ***	0,005 ***
Grünlandanteil	−0,161 ***	−0,198 **
Bevölkerungswachstum	0,995 ***	1,021 ***
Belastungsindex	0,1536	0,170
Beschäftigte Landwirtschaft	−0,348 ***	−0,278 ***
Beschäftigte Tourismus	0,717 ***	0,693 ***
Akademikerquote	3,687 ***	3,835 ***
Pendlersaldo	0,229 ***	0,241 ***
Lagevariablen		
Erreichbarkeit gesamt	0,017 ***	0,016 ***
Modellperformance		
log-Likelihood	−454,200	−453,946
Akaike Inf. Crit.	932,400	931,891
Breusch-Pagan-Test	47,267 ***	87,547 ***
RMSPE	0,863	0,861

Anmerkung: *p<0,1; **p<0,05; ***p<0,01

5.3 Geschätzte Kriging-Modelle

Der zweite im Kapitel 3 (Methodik) beschriebene Ansatz, räumliche Effekte, insbesondere räumliche Abhängigkeit, bei der Modellierung in Betracht zu ziehen, stammt aus der Geostatistik. In dieser Arbeit werden kriging-Methoden angewandt. Beim Kriging wird die räumliche Abhängigkeit mittels empirischem Semivariogramm dargestellt und anschließend in einem Semivariogramm-Modell modelliert. Eine wichtige Voraussetzung, damit kriging Methoden zu validen Ergebnissen führen, ist räumliche Stationarität. Betrachtet man die Mittelwerte des Bodenpreises für die einzelnen Bundesländer in Abbildung 4.2 auf Seite 43, wird ersichtlich, dass ein Trend in den Bodenpreisen vorliegt und diese somit nicht räumlich stationär sind. Es wird versucht, den Trend durch das bereits zuvor geschätzte OLS_3 Modell abzubilden. Allerdings wird dieses mit dem generalized least squares Schätzer erneut geschätzt. Anschließend wird universal kriging auf die räumlich autokorrelierenden Residuen angewandt.

Bei der Erstellung des empirischen Semivariogramms stellt sich die Frage nach den richtigen Distanzklassen. In dieser Arbeit werden Distanzklassen von 5, 10, 15 und 20 Kilometern getestet. Bei kleineren Distanzklassen würde die Möglichkeit bestehen, dass nicht genügend Beobachtungspaare je Distanzklasse vorhanden wären, die nötig sind, um ein valides empirisches Semivariogramm zu erstellen. Bei größeren Distanzklassen besteht die Gefahr, dass sich das im Anschluss zu schätzende Semivariogramm-Modell an einer zu geringen Anzahl an Punkten orientiert.

Aufbauend auf das empirische Semivariogramm wird das Semivariogramm-Modell geschätzt. Getestet werden das sphärische, das exponentielle und das gausian Modell. Somit werden zwölf verschiedene Kombinationen aus Distanzklassen und Modellansätzen mittels Leave-one-out cross validation (LOOCV) getestet. Dies wird automatisiert in der Statistiksoftware *R* durchgeführt. Die Kombination mit dem geringsten Root Mean Square Error (RMSE) wird bei einer Distanzklasse von 10 Kilometern in Kombination mit einem exponentiellem Modellansatz erreicht. Dies wird in Abbildung 5.1 dargestellt.

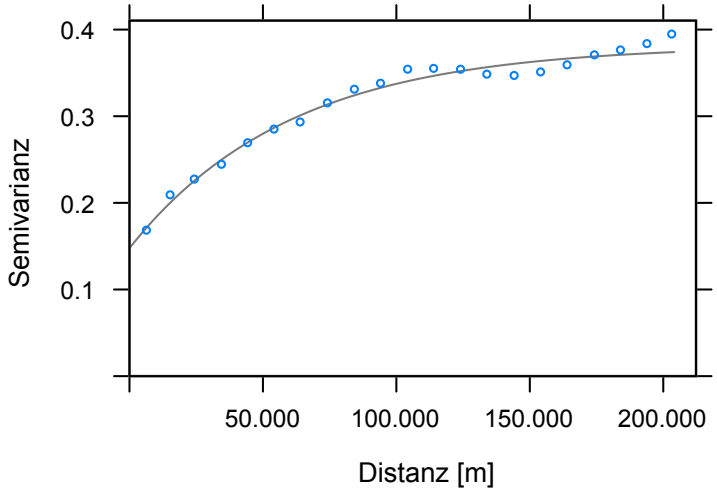

Abb. 5.1: Empirisches Semivariogramm (Distanzklasse 10km) und Semivariogramm-Modell (exponentiell)

Das empirische Semivariogramm folgt grundsätzlich einer idealtypischen Form, weshalb räumliche Stationarität in den Fehlertermen angenommen wird. Mit dem gefitteten Semivariogramm-Modell kann die universal kriging Methode angewandt werden. Es gilt zu beachten, dass kriging eine Methode zur Vorhersage für Beobachtungen im Raum ist, für die keine Werte vorliegen. Für die Ermittlung des empirischen Semivariogramms und dem Trend wurde wie bei allen anderen Modellansätzen lediglich der Trainingsdatensatz verwendet. Durch die universal kriging Methode können nun die Beobachtungen (Gemeinden) aus dem Testdatensatz vorhergesagt werden und der Root Mean Square Prediction Error errechnet werden. Dieser beträgt 0,328.

5.4 Modellgegenüberstellung

Tabelle 5.9 stellt alle oben geschätzten Modellansätze und deren Modell-Fit Kennzahlen nochmals gegenüber. Grundsätzlich ist zu erkennen, dass Modellansätze, die räumliche Effekte in Betracht ziehen, einen höheren Modell-Fit aufweisen. Sie beschrieben somit den unbekannten Prozess, der die Bodenpreise generiert, besser. Das Akaike Information Criterion (AIC) kann erheblich verringert werden. Es liegt somit nahe, dass räumliche Effekte, insbesondere räumliche Abhängigkeit, den Prozess, der die Bodenpreise generiert, beeinflussen. Die Modellergebnisse legen weiters nahe, dass räumliche Heterogenität ebenfalls explizit zu beachten ist. Dies ist unter anderem am Signifikanzniveau der Breusch-Pagan Tests zu erkennen.

Tab. 5.9: Modellgegenüberstellung

Modellspezifikation	adj. R^2	AIC	log-Likelihood	RMSPE
Nicht-räumliche Modellansätze				
OLS_1	0,564	2.687		0,692
$OLS_{2.1}$	0,301	3.309		0,907
$OLS_{2.2}$	0,296	3.322		0,911
OLS_3	0,566	2.681		0,691
Räumliche Modellansätze - räumliche Ökonometrie				
SAR_{Lag} (Knn7)		1.039	-507,3	0,331
SAR_{Lag} (Dist. 17km)		1.021	-499,8	0,326
SAR_{Error} (Knn12)		932	-454,2	0,863
SAR_{Error} (Dist. 17km)		932	-454,2	0,861
Räumliche Modellansätze - Geostatistik				
Universal kriging				0,328

Beim Verglich zwischen Modellansätzen aus der räumlichen Ökonometrie und der Geostatistik ist der unterschiedliche Fokus der beiden Forschungsrichtungen zu beachten. So geht es in der Geostatistik meist darum, valide räumliche Vorhersagen zu machen. In der räumlichen Ökonometrie ist räumliche Vorhersage nur ein Anwendungsfeld. Hier steht vermehrt das Verstehen von räumlichen Prozessen im Vordergrund. Somit kann ein Vergleich dieser Ansätze sinnvollerweise lediglich über die Vorhersagegenauigkeit (out-of-sample prediction), gemessen am RMSPE, geschehen. Vergleicht man den PMSPE der SAR_{Lag} Modellspezifikationen mit dem der universal

kriging Methode, fällt deren Ähnlichkeit auf. Im Vergleich zu den nicht-räumlichen Modellansätzen kann die Vorhersage erheblich verbessert werden. Der errechnete RMSPE für die SAR_{Error} Modellspezifikationen ist aufgrund der Komplexität der Methodik nur bedingt aussagekräftig und bedarf einer weiteren Vertiefung.

Kapitel 6

Abschließende Betrachtung

Ziel dieser Arbeit war es, die Prozesse, die den Bodenpreis in den österreichischen Gemeinden generieren, mittels eines quantitativen statistischen Modells zu beschreiben, um diesen zu verstehen. Durch das Verstehen der Prozesse hat die räumliche Planung die Möglichkeit, diese Prozesse zu beeinflussen und somit die räumliche Entwicklung zu steuern.

In Bezug auf die erste Forschungsfrage,

> **Durch welche Faktoren lässt sich der Prozess, der den durchschnittlichen Bodenpreis in den österreichischen Gemeinden generiert, beschreiben?**

wurde festgestellt, dass es die nutzenstiftenden Eigenschaften der Gemeinden sind, die helfen können, diese Prozesse zu beschreiben. Zu den über die nutzenstiftenden Eigenschaften aufgestellten Hypothesen (Zusammenhänge) ist Folgendes festzuhalten:

Es konnte ein positiver Zusammenhang zwischen der Bevölkerungsdichte in einer Gemeinde und dem Bodenpreis festgestellt werden. Dieser ist in allen geschätzten Modellen signifikant. Ein solcher positiver Zusammenhang ist auch beim Baulandanteil zu beobachten. Dies bestätigt die Hypothese, dass ein hoher Nutzungsdruck durch eine hohe Bevölkerungsdichte zu erhöhten Bodenpreisen führt. Weiters konnte die Hypothese bestätigt werden, dass ein erhöhtes Bevölkerungswachstum ebenfalls den Nutzungsdruck und damit die Bodenpreise erhöht. Die Hypothese, nach der eine günstige Wirtschaftsstruktur in einer Gemeinde zu erhöhten Bodenpreisen führt, kann teilweise bestätigt werden. So konnten keine signifikanten Zusammenhänge zwischen dem Einkommen beziehungsweise der Arbeitslosenquote und dem Bodenpreis einer Gemeinde gefunden werden. Allerdings wurden signifikante und positive Zusammenhänge zwischen der Akademikerquote beziehungsweise dem Pendlersaldo und dem Bodenpreis gefunden. Diese beiden Sachverhalte, sowie der negative Zusammenhang zwischen den Beschäftigten in der Landwirtschaft und dem Bodenpreis, deuten auf eine Bestätigung der genannten Hypothese bezüglich einer günstigen Wirtschaftsstruktur hin. Die Hypothese, dass sich ein hoher Anteil an Beschäftigten im Tourismus positiv auf die Bodenpreise auswirkt, kann auch bestätigt werden. Bezüglich der Hypothese zum Belastungsindex ist anzumerken, dass in den OLS-Modellspezifikationen ein signifikanter erwarteter negativer Zusammenhang feststellbar war, dieser jedoch in den SAR-Modellspezifikationen nicht mehr nachweisbar war. Auch der erwartete positive Zusammenhang zwischen guter Erreichbarkeit und hohem Bodenpreis kann bestätigt werden.

Durch die Tatsache, dass einige wünschenswerte Variablen nicht verfügbar waren, sowie, dass die Erreichbarkeit und damit die Lage im Raum stark determinierend für den Bodenpreis ist, ergaben sich Herausforderungen in der Modellierung. Im Bezug auf die zweite Forschungsfrage,

> **Welche Herausforderungen ergeben sich bei der Modellierung von Sachverhalten mit räumlichem Bezug?**

ist deshalb Folgendes festzuhalten: Bei Sachverhalten, die einen starken räumlichen Bezug aufweisen, ist es sehr wahrscheinlich, dass räumliche Effekte, insbesondere räumliche Abhängigkeit vorhanden ist, auf die es im Modellierungsprozess zu achten gilt. Es wurde aufgezeigt, dass räumliche Effekte einerseits aus theoretischen Überlegungen abgeleitet werden können (spatial externaleties) oder andererseits durch die verwendeten Daten selbst entstehen können (omitted-variable bias).

Es wurden Methoden und Modellansätze aus den Forschungsrichtungen der räumlichen Ökonometrie und der Geostatistik, die beide räumliche Effekte in Betracht ziehen, vorgestellt und angewendet. Im Bezug auf die dritte Forschungsfrage,

> **Können Ansätze der räumlichen Ökonometrie und der Geostatistik eine Verbesserung des Erklärungsgehalts beziehungsweise der Vorhersagegenauigkeit von Bodenpreismodellen bewirken?**

ist anzumerken, dass die vorgestellten räumlichen Modellspezifikationen und Methoden den Erklärungsgehalt und die Vorhersagegenauigkeit der Bodenpreismodelle eindeutig verbessern. So kann festgehalten werden, dass der Prozess, der Bodenpreise in den österreichischen Gemeinden bestimmt, ein räumlicher ist. Das bedeutet, dass der Prozess nicht auf eine Gemeinde beschränkbar ist, sondern Einfluss auf umliegende Gemeinden hat.

Zu betonen ist, dass die Verwendung einer bestimmten räumlichen Modellspezifikation nicht prädeterminiert ist. Hier ist es wichtig, anzumerken, dass die vorgestellten Methoden und Modellansätze nur einen Teil der in der Literatur vorhanden Methoden und Ansätze darstellen. So sind die in dieser Arbeit geschätzten Modelle beispielsweise globale Modellansätze. Dies bedeutet, dass versucht wird, einen einheitlichen Prozess für den gesamten Untersuchungsraum zu finden. Somit können räumlich heterogene Effekte nur teilweise berücksichtigt werden. Es gibt eine Vielzahl an weiteren Modellspezifikationen und Methoden, die unterschiedliche Effekte berücksichtigen. Diese Arbeit bietet in diesem Zusammenhang lediglich einen Einstieg in die Thematik.

Um die Steuerungsmöglichkeit für die räumliche Planung bezüglich Bodenpreise vollständig zu entfalten, ist es, wie bereits erwähnt, essenziell, den Prozess, der den Bodenpreis generiert, zu verstehen. Für ein noch besseres Verständnis sollten bei weiteren Forschungen noch mehr verschiedene Modellansätze und Methoden getestet und aus deren Ergebnissen Erkenntnisse gewonnen werden. Festzuhalten ist jedoch, dass diese Methoden in der Lage sein sollten, räumliche Effekte zu berücksichtigen.

Literaturverzeichnis

Akaike, H. (1974). „A new look at the statistical model identification“. In: *IEEE transactions on automatic control* 19.6, S. 716–723.

Alonso, W. (1960). „A theory of the urban land market“. In: *Papers in Regional Science* 6.1, S. 149–157.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (2001a). „A Companion to Theoretical Econometrics“. In: Hrsg. von B. H. Baltagi. Wiley-Blackwell. Kap. Spatial Econometrics, S. 310–330.

– (2001b). „Spatial effects in econometric practice in environmental and resource economics“. In: *American Journal of Agricultural Economics* 83.3, S. 705–710.

– (2002). „Under the hood: Issues in the specification and interpretation of spatial regression models“. In: *Agricultural economics* 27.3, S. 247–267.

– (2003). „Spatial externalities, spatial multipliers, and spatial econometrics“. In: *International regional science review* 26.2, S. 153–166.

Anselin, L., A. K. Bera, R. Florax und M. J. Yoon (1996). „Simple diagnostic tests for spatial dependence“. In: *Regional science and urban economics* 26.1, S. 77–104.

Bavaud, F. (1998). „Models for spatial weights: a systematic look“. In: *Geographical analysis* 30.2, S. 153–171.

Behrens, K. C. (1971). „Allgemeine Standortbestimmungslehre. UTB 27. Opladen, ßevölkerungs- und Sozialgeographie 1972: Deutscher Geographentag Erlangen 1971. Münchner Studien zur Soz.-u. Wirtsch“. In: *Geogr* 8, S. 1–2.

Bivand, R. S., E. Pebesma und V. Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. URL: `http://www.asdar-book.org/`.

Blöschl, G. und R. Merz (2000). „Methoden Der Hydrologischen Regionalisierung im Zusammenhang mit der Niederschlag-Abflussmodellierung“. In: *Wiener Mitteilungen: Niederschlag-Abfluss Modellierung - Simulation und Prognose* 164, S. 149–178.

Bökemann, D. (1982). „Theorie der Raumplanung: Regionalwissenschaftliche Grundlagen für die Stadt“. In: *Regional-und Landesplanung, München*.

Box, G. E. P. und N. R. Draper (1986). *Empirical Model-building and Response Surface*. New York, NY, USA: John Wiley & Sons, Inc. ISBN: 0-471-81033-9.

Breusch, T. S. und A. R. Pagan (1979). „A simple test for heteroscedasticity and random coefficient variation“. In: *Econometrica: Journal of the Econometric Society*, S. 1287–1294.

Brunauer, W. A., S. Lang und W. Feilmayr (2013). „Hybrid multilevel STAR models for hedonic house prices“. In: *Jahrbuch für Regionalwissenschaft* 33.2, S. 151–172.

Can, A. (1998). „GIS and spatial analysis of housing and mortgage markets“. In: *Journal of Housing Research* 9.1, S. 61–86.

Chica-Olmo, J. (2007). „Prediction of housing location price by a multivariate spatial method: cokriging“. In: *Journal of Real Estate Research* 29.1, S. 91–114.

Dubin, R. A. (1998a). „Predicting house prices using multiple listings data“. In: *The Journal of Real Estate Finance and Economics* 17.1, S. 35–59.

– (1998b). „Spatial autocorrelation: a primer“. In: *Journal of housing economics* 7.4, S. 304–327.

Ernste, H. (2011). *Angewandte Statistik in Geografie und Umweltwissenschaften*. vdf Hochschulverlag AG.

Florax, R. J., H. Folmer und S. J. Rey (2003). „Specification searches in spatial econometrics: the relevance of Hendry's methodology". In: *Regional Science and Urban Economics* 33.5, S. 557–579.

Fotheringham, A. S., M. E. Charlton und C. Brunsdon (1998). „Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis". In: *Environment and planning A* 30.11, S. 1905–1927.

Fotheringham, A., C. Brunsdon und M. Charlton (Jan. 2002). „Geographically Weighted Regression: The Analysis of Spatially Varying Relationships". In: *John Wiley and Sons Ltd, West Sussex, 1 Auflage* 13.

Getis, A. und J. Aldstadt (2004). „Constructing the spatial weights matrix using a local statistic". In: *Geographical analysis* 36.2, S. 90–104.

Giuliani, G. (2002). „Landwirtschaftlicher Bodenmarkt und landwirtschaftliche Bodenpolitik in der Schweiz". Diss. ETH Zurich.

Goldberger, A. S. (1962). „Best linear unbiased prediction in the generalized linear regression model". In: *Journal of the American Statistical Association* 57.298, S. 369–375.

Goulard, M., T. Laurent und C. Thomas-Agnan (2017). „About predictions in spatial autoregressive models: Optimal and almost optimal strategies". In: *Spatial Economic Analysis* 12.2-3, S. 304–325.

Gujarati, D. N. und D. C. Porter (2003). *Basic Econometrics. 4th*. New York: McGraw-Hill.

Haase, R. (2011). „Ertragspotenziale: Hedonische Mietpreismodellierungen am Beispiel von Büroimmobilien". Diss. ETH Zurich.

Haavelmo, T. (1944). „The probability approach in econometrics". In: *Econometrica: Journal of the Econometric Society*, S. iii–115.

Haining, R. (2009). „The special nature of spatial data". In: *The Sage Handbook of Spatial Analysis, Sage, Thousand Oaks*, S. 5–24.

Haining, R., R. Kerry und M. A. Oliver (2010). „Geography, Spatial Data Analysis, and Geostatistics: An Overview." In: *Geographical Analysis* 42.1, S. 7–31.

Helbich, M., W. Brunauer, E. Vaz und P. Nijkamp (2014). „Spatial heterogeneity in hedonic house price models: the case of Austria". In: *Urban Studies* 51.2, S. 390–411.

Herath, S. (2013). „Discovering the urban structure using a spatial hedonic house price model". In: *ICCREM 2013: Construction and Operation in the Context of Sustainability*, S. 1352–1372.

Herath, S. und G. Maier (2010). „The hedonic price method in real estate and housing market research: a review of the literature". In: *SRE-Discussion 2010/03*.

– (Aug. 2013). „Local particularities or distance gradient: What matters most in the case of the Viennese apartment market?" In: *Journal of European Real Estate Research* 6.2, S. 163–185.

Isaaks, E. H. und R. M. Srivastava (1989). *An introduction to applied geostatistics*. Oxford University Press, New York.

Jahanshiri, E., T. Buyong und A. R. M. Shariff (2011). „A review of property mass valuation models". In: *Pertanika Journal of Science & Technology* 19.1, S. 23–30.

Journel, A. G. und C. J. Huijbregts (1978). *Mining geostatistics*. Academic press.

Kelejian, H. H. und I. R. Prucha (2007). „The relative efficiencies of various predictors in spatial econometric models containing spatial lags". In: *Regional Science and Urban Economics* 37.3, S. 363–374.

Keynes, J. N. et al. (1890). „The scope and method of political economy". In: *History of Economic Thought Books*.

Kramar, H. (2005). *Innovation durch Agglomeration: zu den Standortfaktoren der Wissensproduktion*. Fachbereich Stadt-u. Regionalforschung, Department f. Raumentwicklung, Infrastruktur-u. Umweltplanung, Techn. Univ. Wien.

Krige, D. G. (1951). „A statistical approach to some basic mine valuation problems on the Witwatersrand". In: *Journal of the Southern African Institute of Mining and Metallurgy* 52.6, S. 119–139.

Kulke, E. (2017). *Wirtschaftsgeographie*. Bd. 2434. 6. Auflage. UTB. ISBN: 978-3-8252-4709-6.

Lancaster, K. J. (1966). „A new approach to consumer theory". In: *Journal of political economy* 74.2, S. 132–157.

Leroux, B. G., X. Lei und N. Breslow (2000). „Estimation of disease rates in small areas: a new mixed model for spatial dependence". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, S. 179–191.

LeSage, J. und R. K. Pace (2009). *Introduction to spatial econometrics*. Chapman und Hall/CRC.

Logan, J. R. (2012). „Making a place for space: Spatial thinking in social science". In: *Annual review of sociology* 38, S. 507–524.

Maddala, G. S. und K. Lahiri (1992). *Introduction to econometrics*. Bd. 2. Macmillan New York.

Matheron, G. (1963). „Principles of geostatistics". In: *Economic geology* 58.8, S. 1246–1266.

Montero, J.-M., R. Minguez und G. Fernandez-Aviles (2018). „Housing price prediction: parametric versus semi-parametric spatial hedonic models". In: *Journal of Geographical Systems* 20.1, S. 27–55.

Moran, P. A. (1948). „The interpretation of statistical maps". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10.2, S. 243–251.

Rosen, S. (1974). „Hedonic prices and implicit markets: product differentiation in pure competition". In: *Journal of political economy* 82.1, S. 34–55.

Sarlas, G. und K. W. Axhausen (2014). „Towards a direct demand modeling approach". In: *14th Swiss Transport Research Conference (STRC 2014)*. Swiss Transport Research Conference (STRC).

Saul, S. (2017). *Grid-based indicators of accessibility of public utility infrastructure*. Forschungsber. Statistics Austria.

Sinclair, A. J. und G. H. Blackwell (2002a). „KRIGING". In: *Applied Mineral Inventory Estimation*. Cambridge University Press, S. 215–241. DOI: 10.1017/CBO9780511545993.011.

– (2002b). „SPATIAL (STRUCTURAL) ANALYSIS: AN INTRODUCTION TO SEMIVARIOGRAMS". In: *Applied Mineral Inventory Estimation*. Cambridge University Press, S. 192–214. DOI: 10.1017/CBO9780511545993.010.

Thünen, J. v. (1826). „Der isolierte Staat". In: *Beziehung auf Landwirtschaft und Nationalökonomie*.

Tobler, W. R. (1970). „A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46.sup1, S. 234–240.

Weber, A. (1909). *Über den standort der industrien, 1. teil: Reine theorie des standortes. English translation: On the location of industries*.

Wieser, R. (2006). „Wirkungen der U-Bahn auf den Bodenmarkt in Wien". In: *Fachbereich Finanzwissenschaft und Inf*.

Wooldridge, J. M. (2012). *Introductory Econometrics - A Modern Approach*. Hrsg. von M. Worls. South-Western Cengage Learning.

Anhang

OLS Diagnoseplots

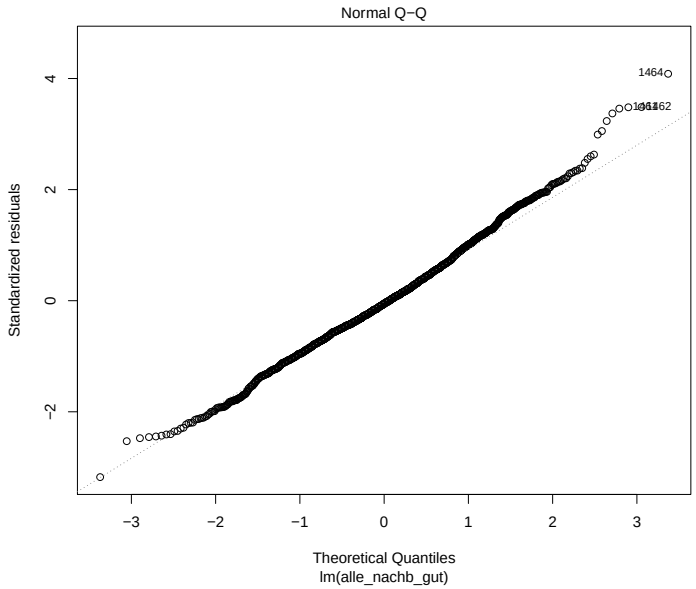

Abb. .1: OLS_1 Qunatil-Quantil Plot

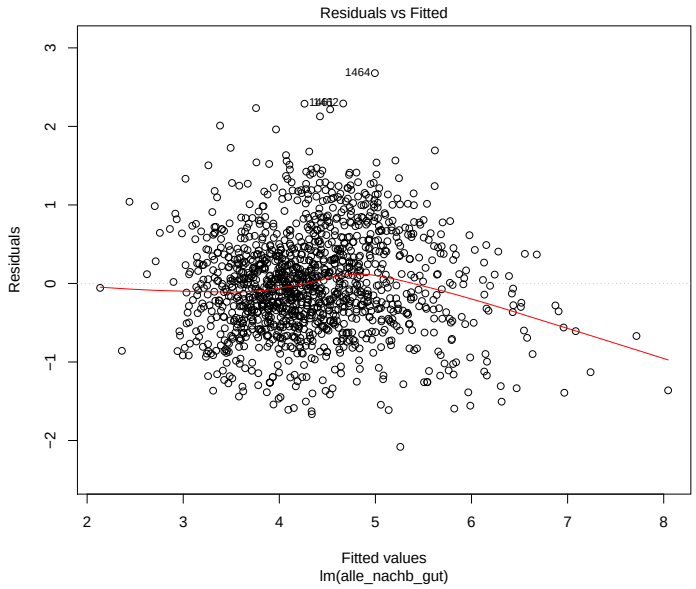

Abb. .2: OLS_1 Residuenplot

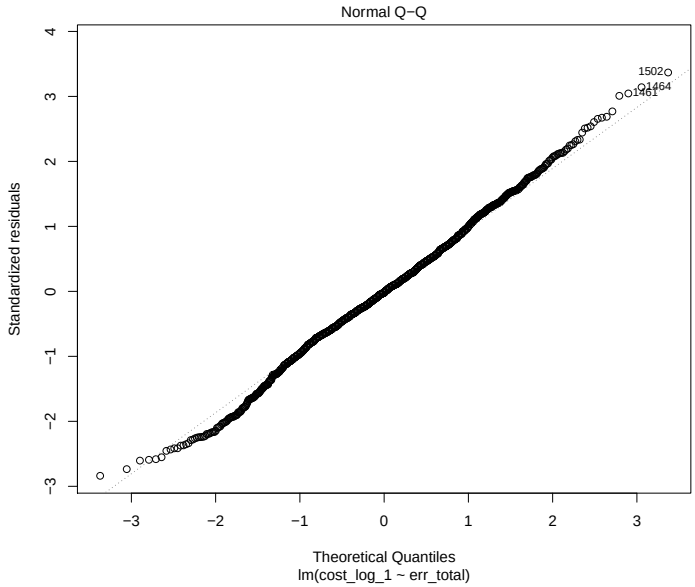

Abb. .3: $OLS_{2.1}$ Qunatil-Quantil Plot

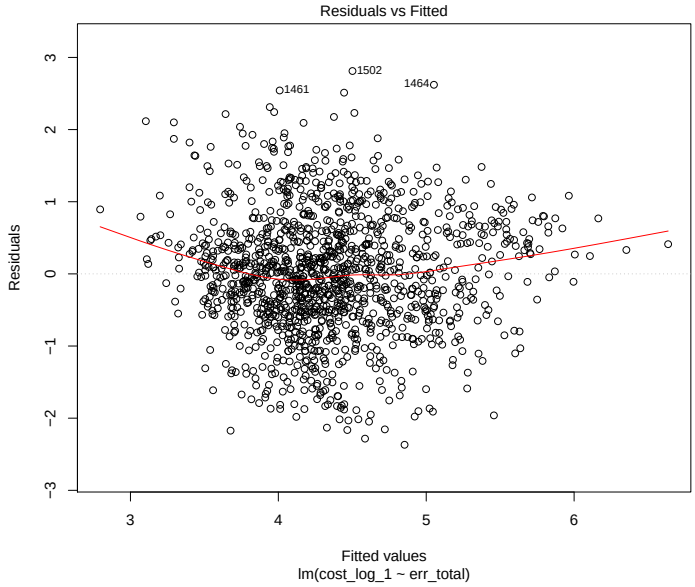

Abb. .4: $OLS_{2.1}$ Residuenplot

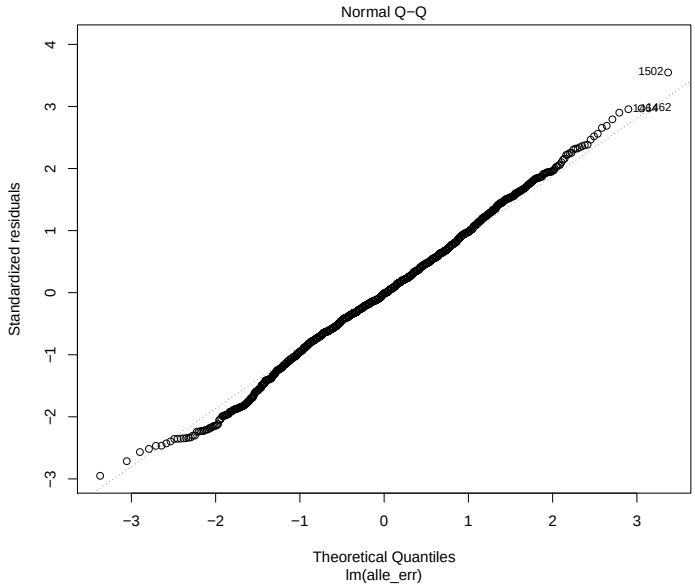

Abb. .5: $OLS_{2.2}$ Qunatil-Quantil Plot

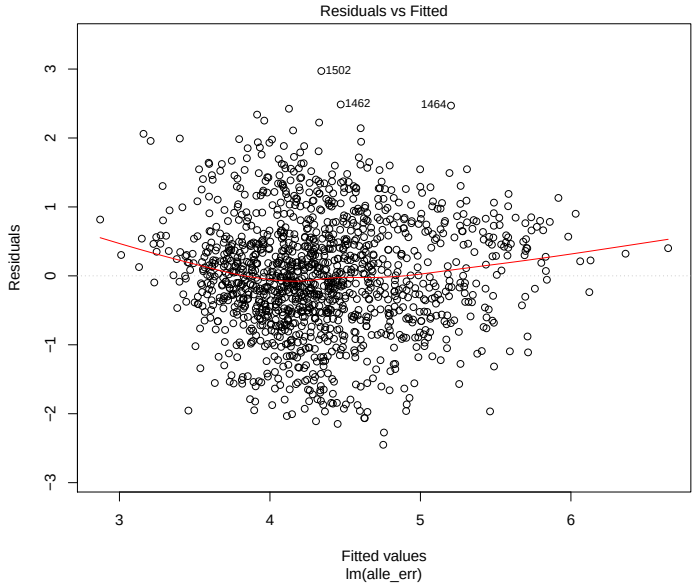

Abb. .6: $OLS_{2.2}$ Residuenplot

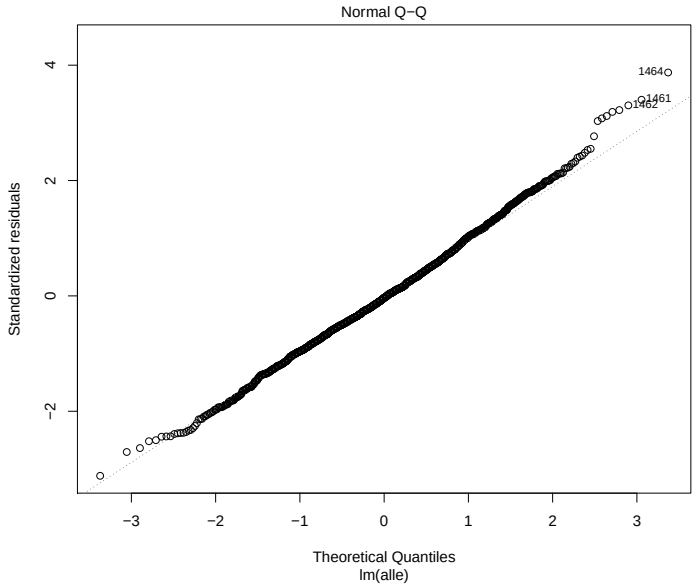

Abb. .7: OLS_3 Qunatil-Quantil Plot

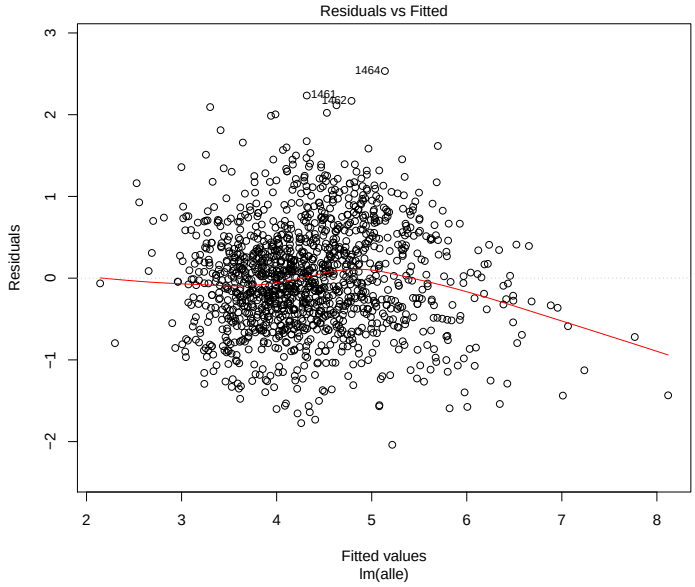

Abb. .8: OLS_3 Residuenplot