

Ramstein, Axel

**Article**

## Das Analysepotenzial georeferenzierter Daten

WISTA - Wirtschaft und Statistik

**Provided in Cooperation with:**

Statistisches Bundesamt (Destatis), Wiesbaden

*Suggested Citation:* Ramstein, Axel (2022) : Das Analysepotenzial georeferenzierter Daten, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 74, Iss. 3, pp. 39-52

This Version is available at:

<https://hdl.handle.net/10419/260631>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

---

# DAS ANALYSEPOTENZIAL GEOREFERENZIERTER DATEN

---

Axel Ramstein

---

↳ **Schlüsselwörter:** Georeferenzierung – raumbezogene Analyse – Mietkosten – Mikrozensus – Wohnlage

## ZUSAMMENFASSUNG

Seit 2013 ist es gesetzlich möglich, Erhebungsmerkmale der amtlichen Statistik dauerhaft zu geografischen quadratischen Gitterzellen von mindestens 100 Metern Kantenlänge zu verorten. Dadurch ergeben sich Analysemöglichkeiten in großer räumlicher Tiefe. Der Artikel untersucht auf Basis von Daten des Mikrozensus 2018 sowie der Zusatzerhebung „Wohnen“, ob sich auf Haushaltsebene raumbezogene Charakteristika finden lassen, die einen Teil der Varianz der Nettokaltmiete erklären. Anhand der identifizierten räumlichen Variablen wird anschließend ein Indikator abgeleitet, der als ein Ansatz zur qualitativen Abbildung des Einflusses des Merkmals „Lage“ auf Wohnkosten vorgeschlagen wird. Allerdings bedarf dieser Indikator weiterer Untersuchungen und ist daher als experimentelle Statistik zu betrachten.

↳ **Keywords:** georeferencing – geospatial analysis – rental costs – German microcensus – location

## ABSTRACT

*Since 2013, legislation has permitted the geocoding of survey variables in official statistics to geographic quadratic grid cells with a minimum edge length of 100 metres on a permanent basis. This allows for analyses at a high level of spatial detail. Based on data from the microcensus 2018 and its housing module, the article analyses whether geospatial characteristics can be found at household level that explain part of the variance of net rent exclusive of heating expenses. The spatial variables identified are then used to derive an indicator and it is suggested that this indicator could be an approach to describing the influence of location on rental costs in qualitative terms. The indicator needs to be analysed further, however, and is therefore to be considered as an experimental statistic.*

### Axel Ramstein

studierte Public Economics in Berlin und ist seit 2019 wissenschaftlicher Mitarbeiter im Referat „Forschungsdatenzentrum, Methoden der Datenanalyse“ des Statistischen Bundesamtes. Schwerpunkte seiner Arbeit sind Projekte zur Nutzung raumbezogener Daten für die amtliche Statistik.

## 1

### Einleitung

---

Seit der Änderung des § 10 Bundesstatistikgesetz zum 1. August 2013 ist die regionale Zuordnung von Erhebungsmerkmalen der amtlichen Statistik zu (quadratischen) geografischen Gitterzellen (oft als Geokodierung bezeichnet) gesetzlich zulässig. Bei einer minimalen Kantenlänge von 100 Metern ist eine kleinräumige Speicherung von Daten unterhalb der Gemeindeebene erlaubt. Die räumlich höhere Auflösung und zeitliche Stabilität des Rasters bietet die Möglichkeit zu flexiblen Analysen in einem feingranularen räumlichen Kontext, gleichzeitig ist Vergleichbarkeit im Längsschnitt gegeben. Zum Ende des Jahres 2019 waren alle amtlichen Statistiken geokodiert, die geokodierbar sind; mit Stand November 2021 traf dies auf 175 Statistiken zu.

Kann der Merkmalskranz einer amtlichen Statistik durch Verschneidung mit weiteren rasterbasierten Daten ohne Belastung der Auskunftspflichtigen erweitert werden, um damit neue statistische Erkenntnisse zu gewinnen? Dies wurde in einem in diesem Artikel vorgestellten Projekt am Beispiel des geokodierten Mikrozensus 2018 mit der Zusatzerhebung „Wohnen“ untersucht.

Ziel des Projekts war, mittels einer multiplen linearen hedonischen Regression (MLR) räumliche Charakteristika zu identifizieren, mit denen sich Kosten für gemieteten Wohnraum (im Folgenden kurz: Wohnkosten) erklären lassen. Anhand der ausgewählten Charakteristika wurde anschließend ein experimenteller Indikator abgeleitet, der als ein erster Ansatz zur qualitativen Abbildung des Merkmals „Lage“ vorgeschlagen wird. Damit ist allerdings noch nicht der Qualitätsstandard einer amtlichen Statistik erreicht, da Validität und Aussagekraft des experimentellen Indikators noch weiterer Untersuchung bedarf.

Das folgende Kapitel 2 stellt zunächst dar, welche Daten neben den Ergebnissen der Zusatzerhebung „Wohnen“ zum Mikrozensus 2018 für das Projekt genutzt wurden. Den Modellansatz und die Modellreduktion beschreibt Kapitel 3. Die Ergebnisse der räumlichen Analyse der geokodierten Mikrozensusdaten ohne und mit externer Kovariate sowie den abgeleiteten Indikator für Wohnkosten enthält Kapitel 4. Der Artikel schließt mit einer Zusammenfassung und einem Ausblick in Kapitel 5.

## 2

### Datengrundlage

---

Der Mikrozensus 2018 stellt mit der vierjährigen Zusatzerhebung Daten zur Wohnsituation der Auskunftspflichtigen bereit (Statistisches Bundesamt, 2019). Neben der Wohnfläche stehen weitere Merkmale zur Verfügung, wie der Wohngebäudetyp, das Jahr des Einzugs, das Baujahr des Gebäudes als kategoriale Variable und Informationen über die Eigentumsverhältnisse am Haus oder an der Wohnung. Als abhängige Variable wird die Nettokaltmiete herangezogen, die aus den Angaben zu den gesamten Zahlungen an die Vermieterin oder den Vermieter und der Höhe der Nebenkosten abgeleitet ist. Zur Ausstattung der Wohnung stehen keine weiteren Daten zur Verfügung, sodass eine Kontrolle auf individuelle Qualitätsunterschiede in der Regression nicht möglich ist. Hilfsgrößen zur Beschreibung der Ausstattungsqualität, wie die Energieart zur Beheizung der Wohnräume oder Aspekte der Barrierefreiheit, liefern nur einen sehr geringen Erklärungsbeitrag bezogen auf das statistische Gütemaß  $R^2$ . Sie wurden daher nicht berücksichtigt. Die Daten werden auf Haushaltsebene ausgewertet und sind durch Geokodierung jeweils einer quadratischen Gitterzelle mit 100 Metern Kantenlänge zugeordnet. Deren eindeutiger Identifikator dient als Schlüssel für die Verschneidung mit weiteren Daten.

Die Analyse erfolgte lediglich exemplarisch für die Bundesländer Nordrhein-Westfalen und Hessen. Gründe dafür waren zum einen die im Folgenden beschriebene notwendige weitere Datengewinnung und zum anderen die Begrenzung der Anzahl der Beobachtungen auf ein handhabbares Maß. Hessen wurde ausgewählt, da es vollständig innerhalb des Staatsgebiets der Bundesrepublik Deutschland liegt und somit Verzerrungen in der Nähe der Staatsgrenze durch fehlende Daten aus den Nachbarstaaten gegebenenfalls minimiert werden können.

Bei der Charakterisierung des räumlichen Kontexts unterscheidet man in der Literatur – basierend auf der betrachteten Ausdehnung des Umfelds – zwischen Makrolage, der regionalen überörtlichen Situation, und Mikrolage, der kleinräumigen Nachbarschaft (Voß/Bakker, 2017). Hierfür wurde eine Reihe unterschiedlicher Quellen herangezogen, die sich in zwei Kategorien dif-

ferenzieren lassen: Zum einen sind dies Daten der amtlichen Statistik (beispielsweise aus dem Zensus 2011 zur Beschreibung der Demografie und aus dem statistischen Unternehmensregister zur Ableitung von Indikatoren der Wirtschaftskraft). Zum anderen handelt es sich um externe Daten zur weiteren Beschreibung der Lage im Hinblick auf die Bebauung und Umwelt und die örtliche Nähe beziehungsweise Erreichbarkeit sogenannter Points-of-Interest (POI). Dazu zählen unter anderem Verkehrsinfrastrukturpunkte, Institutionen des öffentlichen Dienstes oder private Einrichtungen, beispielsweise zur Freizeitgestaltung oder zur Deckung des täglichen Bedarfs. Derzeit gibt es keine umfassende und regelmäßig aktualisierte Quelle für derartige Rasterdaten, auf die die amtliche Statistik für raumbezogene Analysen zurückgreifen könnte.<sup>1</sup> Deshalb wurden diese extern bezogen oder projektspezifisch generiert.

Um die Oberflächenbeschaffenheit zu charakterisieren wurde auf das Landbedeckungsmodell des Bundesamtes für Kartographie und Geodäsie (BKG, 2018a) zurückgegriffen und daraus Informationen zur Flächenversiegelung und Begrünung des Umfelds einer Gitterzelle gewonnen. Außerdem hat das Bundesinstitut für Bau-, Stadt- und Raumentwicklung (BBSR) Rasterdaten zur Pkw-Erreichbarkeit von Flughäfen, Autobahnanschlüssen, Bahnhöfen des Fernverkehrs, Oberzentren und Mittel- oder Oberzentren zur Verfügung gestellt. Weiterhin hat das BBSR Daten zu täglichen Abfahrtszahlen des öffentlichen Verkehrs sämtlicher Haltestellen des Bundesgebiets bereitgestellt, unterschieden nach den Transportmodi Bus, Bahn und Tram für jeweils einen Referenzsamstag, -sonntag und -werktagsdurchschnitt. Daten zur Pkw-Erreichbarkeit von Bildungseinrichtungen (Universitäten, Fachhochschulen, Grund-, weiterführende und Berufsschulen) wurden vom BKG bezogen. Bei den herangezogenen Quellen für die Erreichbarkeitsinformationen unterscheiden sich nicht nur die eingesetzte Routing-Engine und das zugrunde liegende Straßennetz, sondern auch die Berechnungsansätze. Diese Diskrepanz ließ sich kurzfristig nicht beseitigen und wurde für das Projekt zur Begrenzung des Aufwands hingenommen. Ob und wie

stark sich diese unterschiedlichen Berechnungsansätze auf die Ergebnisse auswirken, wurde aus dem gleichen Grund nicht näher analysiert.

Weitere Kategorien von Points-of-Interest wurden zum einen dem Datenangebot „POI-Bund“ des Bundesamtes für Kartographie und Geodäsie (BKG, 2018b) entnommen (beispielsweise Justizvollzugsanstalten und Kindertagesstätten [Kitas]<sup>2</sup>) und zum anderen über einen automatisierten API<sup>3</sup>-Zugriff (Olbricht, 2018) vom Datenbestand des Projekts „OpenStreetMap“ (OSM) extrahiert. Neben Kultur- und Freizeiteinrichtungen wie Kinos und Theater wurden auf diese Weise auch soziale Einrichtungen, Gaststätten, Supermärkte und Einrichtungen des Freizeitsports (Sportzentren, Schwimmhallen und -bäder, Fitnesscenter, Eishallen, Golfplätze) gewonnen. Zur Generierung der gerouteten Erreichbarkeitsinformationen dieser zusätzlichen Points-of-Interest wurde die API des vom Bundesamt für Kartographie und Geodäsie für Bundesbehörden angebotenen Dienstes RoutingPlus<sup>4</sup> (BKG, 2021) herangezogen. Mit dieser wurden für jede Punktcoordinate einer POI-Kategorie sogenannte Isochronen<sup>5</sup> für das Transportmittel Pkw und „Isodistanzen“<sup>6</sup> für die Fortbewegung zu Fuß zu jeweils vordefinierten Fahrzeiten (5 Minuten, 10 Minuten, 15 Minuten und so weiter) beziehungsweise Wegstrecken (200 Meter, 500 Meter, 1 000 Meter, ...) berechnet. Diese wurden anschließend auf das geografische Gitter verschnitten. Im Ergebnis liegen je Thema auf Gitterzellenebene die (maximale) Fahrzeit zum nächstgelegenen Point-of-Interest als kategoriale Variable und die Anzahl der erreichbaren Points-of-Interest zu den gegebenen Fahrzeiten vor. Man spricht hier vom Isochronenansatz für die Erreichbarkeitsberechnung. Zusätzlich wurden auch Daten nach dem sogenannten Rasteransatz gewonnen. Bei diesem ist die Bewegungsrichtung entgegengesetzt, das heißt es wird ausgehend von jedem Gitterzellenmittelpunkt auf den Point-of-Interest als Ziel geroutet. Hierdurch ergibt sich für jede Gitterzelle die Wegstrecke beziehungsweise Fahrzeit als kontinuier-

1 Das Statistische Bundesamt entwickelt zurzeit gemeinsam mit den Projektpartnern Bundesamt für Kartographie und Geodäsie und Johann Heinrich von Thünen-Institut eine solche Datenbank auf Gitterzellenebene im Rahmen eines von der Europäischen Union geförderten Projekts zum Thema „Integration of geospatial data and statistics“.

2 Für Kitas wurden aufgrund der sehr hohen Dichte keine gerouteten Erreichbarkeitsdaten berechnet, sondern der euklidische Abstand verwendet.

3 API steht für application programming interface und bedeutet Programmierschnittstelle.

4 Der Dienst basiert auf dem vom Heidelberg Institute for Geoinformation Technology (HeiGIT) entwickelten Dienst OpenRouteService und nutzt das OSM-Straßennetz.

5 Isochronen bezeichnen Polygone gleicher Reisezeit.

6 Isodistanzen bezeichnen Polygone gleicher Wegstrecke.

liche Variable. Ein Problem stellt hierbei die Abwägung zwischen Komplexitätsreduktion und Genauigkeit dar: Die Berechnung sämtlicher Routen von jedem Gitterzellenmittelpunkt zu jedem Point-of-Interest mit anschließender Wahl des jeweiligen Minimums ist in der Praxis nur in Ausnahmefällen operabel. Beim Verzicht auf das Routing, um den Aufwand zu verringern, und bei Betrachtung des einfachen euklidischen Abstands (Luftlinie) können sich wiederum größere Diskrepanzen ergeben, wenn beispielsweise natürliche Barrieren wie Wasserflächen (Flüsse, Seen) geschnitten werden. Für das Projekt wurden in einer Näherung je Gitterzelle die Routen zu den fünf euklidisch nächstgelegenen Points-of-Interest berechnet und dabei angenommen, dass darunter auch der nach Reisezeit nächstgelegene POI fällt. Aus diesen Routen wurde dann das Minimum gewählt. Die Wahl der geeigneten Methode, um Erreichbarkeitsinformationen zu berechnen, hängt grundsätzlich vom Anwendungsfall, der zur Verfügung stehenden IT-Infrastruktur, der POI-Kategorie und der Methodik zur Lagebewertung ab (siehe zum Beispiel WalkScore®; Lerner, 2014).

Auf potenziell relevante Einflussgrößen – beispielsweise um Lärmimmissionen abzubilden – wurde aufgrund fehlender Daten mit flächiger Abdeckung der Bundesländer verzichtet. Darüber hinaus wurden weitere Variablen der Mikrolage, wie Höhenlage, Hangneigung und Exposition (Promann, 2012), nicht berücksichtigt, da eine derart feingranulare Betrachtung nur bei punktscharfen Untersuchungen sinnvoll ist. Für die vorliegende Analyse auf Gitterzellenebene kommt sie nicht in Betracht.

### 3

## Modellansatz und -reduktion

Der multiplen linearen Regression liegt die Annahme eines hedonischen Modells zugrunde, das auf der von Lancaster (1966) vertretenen Hypothese fußt: Ein Gut stiftet nicht selbst Nutzen, sondern dessen inhärente spezifische Eigenschaften. Diese hat Rosen (1974) für seine Theorie impliziter Preise aufgegriffen, die gemeinhin als Grundlage hedonischer Preismodelle gilt. Die Gleichung des Grundmodells lautet damit:

$$(1) \ln(nkm) = \beta_0 + \beta_1 \mathbf{et} + \beta_2 \mathbf{bj} + \beta_3 \mathbf{efh} + \beta_4 \mathbf{wohnd} + \beta_5 \ln(\mathbf{wfl}) + \varepsilon,$$

$$\text{mit Var}(\varepsilon) = \begin{bmatrix} \Omega_1 & 0 & 0 & 0 \\ 0 & \Omega_2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \Omega_C \end{bmatrix}, \text{ wobei}$$

**nkm** die Nettokaltmiete bezeichnet,

**et** und **efh** als kategoriale beziehungsweise als Dummy-Variable die Informationen zum Eigentümer und Wohng Gebäudetyp verdichten,

**bj** kategoriale den Bauzeitraum abbildet und

**wohnd** und **wfl** die Wohndauer (Vertragslaufzeit) und -fläche beschreiben.

$\varepsilon$  bezeichnet einen Vektor mittelwertfreier Fehlerterme.

Üblicherweise werden die Fehler  $\varepsilon$  als unabhängig und identisch verteilt angenommen. Im vorliegenden Fall ist jedoch zu berücksichtigen, dass es sich beim Mikrozensus um eine Flächenstichprobe (Klumpenstichprobe) handelt (Bihler/Zimmermann, 2016). Deren Auswahl-einheiten sind sogenannte Auswahlbezirke (AWB), die jeweils in vier fachlichen Schichten der Anschriften (Anschriftengrößenklassen, GKL1 bis GKL4) gebildet werden. Die Zugehörigkeit einer Anschrift zu einer fachlichen Schicht bestimmt sich aus der Anzahl der Wohnungen, die an dieser vorhanden sind; dabei deckt die GKL4 Anschriften ohne Wohnungen und Sonderbereiche ab. Die GKL1 fasst Anschriften mit bis zu vier Wohnungen zu einem Auswahlbezirk mit zehn bis dreizehn Wohnungen zusammen. Die GKL2 umfasst Anschriften mit fünf bis zehn Wohnungen und in GKL3 werden Anschriften mit elf oder mehr Wohnungen künstlich in Auswahlbezirke mit einer Richtgröße von sechs Wohnungen aufgeteilt. Es ist davon auszugehen, dass Beobachtungen innerhalb eines Auswahlbezirks der GKL3 eine stärkere räumliche Nähe und damit Ähnlichkeit der Lage (und gegebenenfalls auch der Ausstattung, sofern sie im selben Wohnhaus gelegen sind) aufweisen als beispielsweise Beobachtungen eines Auswahlbezirks der GKL1. Das stützt die Annahme, dass die Residuen von Beobachtungen eines Auswahlbezirks korreliert sind. Für das Modell werden deshalb für  $\varepsilon$  nach Auswahlbezirken geclusterte Fehlerterme angenommen, das heißt eine blockdiagonale Kovarianzmatrix, wobei  $C$  in obiger Gleichung die Anzahl der Cluster – hier Auswahlbezirke – beschreibt.

Nach dem Ersten Gesetz der Geographie von Tobler (1970) hängt alles mit allem zusammen, aber näher beieinander Liegendes stärker als weiter entfernt Liegendes.<sup>17</sup> Diesem Gesetz, dessen Gültigkeit für Wohnkosten intuitiv nachvollziehbar ist, wird hier ausschließlich durch die Regressoren der räumlichen Charakteristika Rechnung getragen. In der Literatur werden komplexere Modelle untersucht, die räumliche Abhängigkeit und mögliche nicht lineare Zusammenhänge strukturell im Modellansatz berücksichtigen (Montero und andere, 2018; Veie/Panduro, 2013). Für Modelle mit räumlicher Autoregressivität der abhängigen Variablen ist die Kleinste-Quadrate-Schätzung (OLS-Ansatz) allerdings verzerrt und inkonsistent (Anselin, 1988), weshalb die Schätzung dieser sogenannten Spatial-Lag-Modelle mittels Maximum Likelihood erfolgt. Mit steigender Anzahl von Beobachtungen wachsen die Anforderungen an die Rechenleistung mit dem verwendeten R-Paket „spatialreg“ (Bivand und andere, 2021) jedoch stark, sodass die praktische Berechnung schnell an Grenzen stößt.

Die Identifikation der räumlichen Größen mit dem größten Erklärungsgehalt für Wohnkosten birgt mehrere Herausforderungen. Zum einen ist dies die Selektion der sachlichen Kategorie, zum anderen die Wahl des betrachteten Einflussradius der jeweiligen Variablen, das heißt ob sie eher in der Makro- oder Mikrolage ausschlaggebend ist. Beide Aspekte sind nicht von vornherein festgelegt. Für das Projekt wurde die Auswahl mittels Modellreduktionsverfahren über statistische Gütemaße getroffen. Als besondere Hürde erweist sich hierbei die starke Korrelation der Regressoren. Zum einen gibt es dafür sachliche Gründe – beispielsweise korreliert die Bevölkerungsdichte stark mit der Dichte von Kindertagesstätten. Zum anderen wird sie verursacht durch die mehrfache Einbindung derselben Variable mit unterschiedlichem Einflussradius (zum Beispiel die Anzahl der erreichbaren Supermärkte in 1000 Metern, 2000 Metern, 5000 Metern), um die Wirkung in der Makro- oder Mikrolage zu identifizieren. Grundsätzlich, jedoch über dieses Projekt hinausgehend, lässt sich eine Priorisierung beziehungsweise Vorauswahl und Einordnung von Lagefaktoren über Befragungen realisieren (Prioritätenauswahl; siehe Voß/Bakker, 2017).

<sup>17</sup> “Everything is related to everything else, but near things are more related than distant things.”

## 4

### Ergebnisse

Eine einfache Regression der abhängigen Variable auf die Kovariate des Mikrozensus der vollständigen Beobachtungen ohne räumliche Charakteristika resultiert in vergleichsweise niedrigem adjustiertem  $R^2$  – für Nordrhein-Westfalen beispielsweise 0,34. Zurückführen lässt sich dies auf zwei Ursachen: Zum einen stehen, wie bereits beschrieben, keine aussagekräftigen Ausstattungsmerkmale der Wohnungen zur Verfügung, die die Varianz der Wohnkosten besser erklären könnten. Zum anderen sind Beobachtungen mit sehr niedrigen Nettokaltmieten von unter 2 Euro je Quadratmeter enthalten. Diese sind vermutlich besonderen Umständen der Mietverhältnisse geschuldet und nicht als marktbestimmte Mieten mit Einflüssen räumlicher Gegebenheiten abgeleitet. Entfernt man das unterste Perzentil der Daten als potenzielle Sonderfälle, steigt das adjustierte  $R^2$  beispielsweise für Nordrhein-Westfalen auf rund 0,42.

↘ Tabelle 1

**Tabelle 1**

Ergebnisse der multiplen linearen Regression des Grundmodells

	Nordrhein-Westfalen	Hessen
N0	39 641	11 461
Adjustiertes $R^2$	0,34	0,22
Räumliche Residuenkorrelation (I) <sup>1</sup>	0,18	0,26
Erstes Perzentil Nettokaltmiete je Quadratmeter [EUR]	1,81	1,25
Bei abgeschnittenem ersten Perzentil		
N	39 244	11 347
Adjustiertes $R^2$	0,42	0,30
Räumliche Residuenkorrelation (I)	0,23	0,33
...und mit Regressoren ROR und KT		
Adjustiertes $R^2$	0,50	0,48
Räumliche Residuenkorrelation (I)	0,12	0,14

<sup>1</sup> Gauß-Distanzfunktion, adaptive Bandbreite.  
ROR: Raumordnungsregion; KT: siedlungsstruktureller Kreistyp.

Als Maß für die räumliche Korrelation der Beobachtungen dient Moran's I (Moran, 1950), ein Index, der sich wie folgt berechnet:

$$(2) \quad I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \text{ wobei}$$

$n$  die Anzahl der Beobachtungen bezeichnet,  
 $y_i$  die untersuchte Variable mit dem Mittelwert  $\bar{y}$  und  
 $w_{ij}$  die auf der räumlichen Nachbarschaftsbeziehung zwischen Beobachtung  $i$  und  $j$  (zum Beispiel Distanz) beruhenden Gewichte<sup>8</sup>.

Der Erwartungswert für eine vollkommen zufällige Verteilung der Werte im Raum beträgt dabei  $E(I) = \frac{-1}{(n-1)}$ .

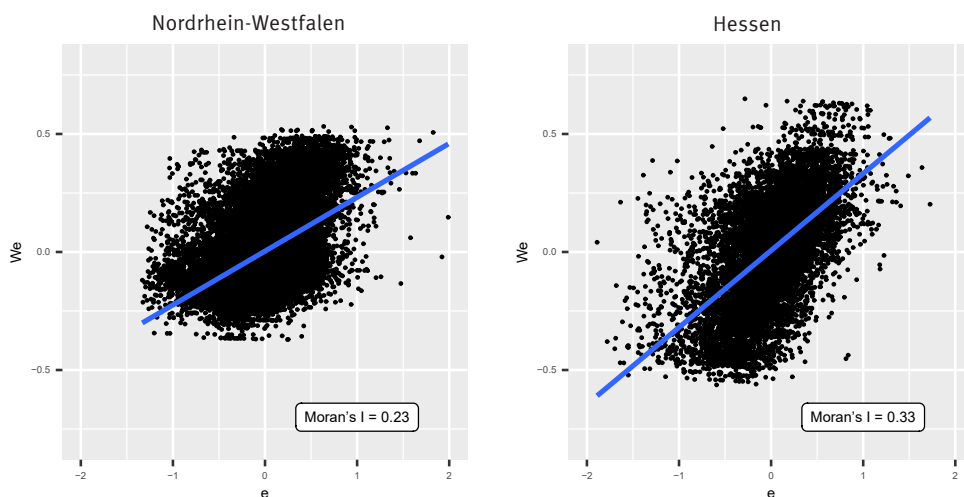
Für zeilennormierte Gewichte ( $\sum_j w_{ij} = 1$ ) liegt der Wertebereich von  $I$  zwischen  $-1$  (überdurchschnittliche Werte sind umgeben von unterdurchschnittlichen Werten, ähnlich einem Schachbrettmuster) und  $1$  (über- und unterdurchschnittliche Werte bilden überschneidungsfreie Klumpungen). Die Residuen des Grundmodells ohne zusätzliche räumliche Kovariate weisen für Hessen und Nordrhein-Westfalen eine deutliche Korrelation auf. [↘ Grafik 1](#) Die Häufigkeit und Größe urbaner Zentren mit hoher Siedlungsdichte und entsprechend höherem Anteil an Auswahlbezirken der

GKL3 (und damit der Anzahl von Beobachtungen, da in jeder Schicht unabhängig von den anderen ein Auswahlatz von 1 % realisiert wurde) der beiden Bundesländer kommen als eine mögliche Ursache infrage. Mit einer Erweiterung des Modells mit in der Praxis gängigen fixen Effekten für die Raumordnungsregion und den siedlungsstrukturellen Kreistyp (BBSR, 2021) ist es möglich, räumlicher Differenzierung Rechnung zu tragen. Damit sinkt die räumliche Korrelation der Regressionsresiduen in beiden betrachteten Fällen deutlich (siehe Tabelle 1). Das bedeutet, dass die Berücksichtigung dieser räumlichen Strukturen einen wesentlichen Anteil der räumlichen Korrelation erklären kann. Die Strukturen basieren jedoch auf einer Vielzahl von Variablen und erlauben keine Betrachtung in höherer räumlicher Tiefe.

<sup>8</sup> Wobei  $w_{ii} = 0$ , da eine Beobachtung nicht sich selbst Nachbar ist.

**Grafik 1**

Moran-Plots der Residuen des Grundmodells



2022 - 0145

## Räumliche Analyse der geokodierten Mikrozensusdaten ohne externe Kovariate

Geografisch gewichtete Regression (Geographically Weighted Regression [GWR]; Brunsdon und andere, 1996) ermöglicht über die Berechnung individueller Regressionskoeffizienten für eine beliebige Punktkoordinate die Analyse, ob ein Koeffizient über den Raum variiert oder konstant ist. Die Regressionsgewichte beruhen dabei analog zu Moran's I auf dem euklidischen Abstand und einer gewählten Distanzfunktion (zum Beispiel inverse, biquadratische oder Gauß-Funktion). Die Grafiken 2 und 3 bilden im 100-Meter-Raster die so gewonnenen Punktschätzungen für die Elastizität der Nettokaltmiete bezüglich der Wohnfläche und die Semielastizität bezüglich der Wohndauer ab.

Es zeigt sich, dass in ländlichen Gebieten Wohneinheiten mit größerer Wohnfläche ceteris paribus günstiger sind als solche mit kleinerer Wohnfläche. Dagegen wird

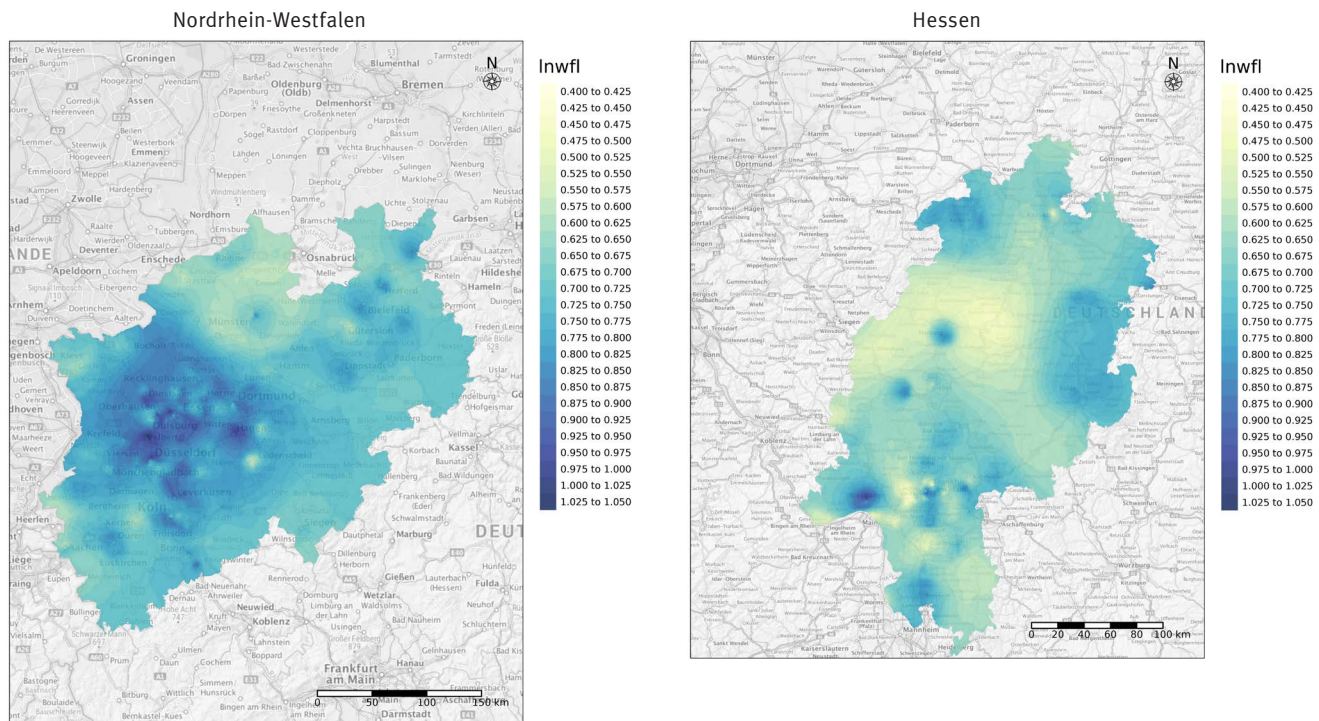
dieser regressive Zusammenhang in urbanen Zentren schwächer und kehrt sich vereinzelt lokal sogar um, das heißt größere Wohnungen sind dort vergleichsweise teurer als kleinere Wohnungen. Weiterhin ist der Kostenvorteil von Bestandsmieten zu erkennen, insbesondere in größeren Städten.

## Räumliche Analyse der geokodierten Mikrozensusdaten mit externen Kovariaten

Aus der Datenaufbereitung liegt wie bereits beschrieben eine Vielzahl stark fachlich und räumlich korrelierter Regressoren vor. Die Variablen mit mehreren räumlichen Einzugsbereichen wurden dazu zunächst auf jeweils zwei Regressoren reduziert. Dazu wurden mittels Partial Least Squares (PLS) des um die jeweilige Variablenkategorie erweiterten Grundmodells von den (zuvor normalisierten) Variablen diejenigen mit den betragsmäßig

### Grafik 2

Punktschätzung des Koeffizienten des Regressors Wohnfläche (logarithmiert) im Grundmodell mittels geografisch gewichteter Regression (GWR) 100-m-Raster



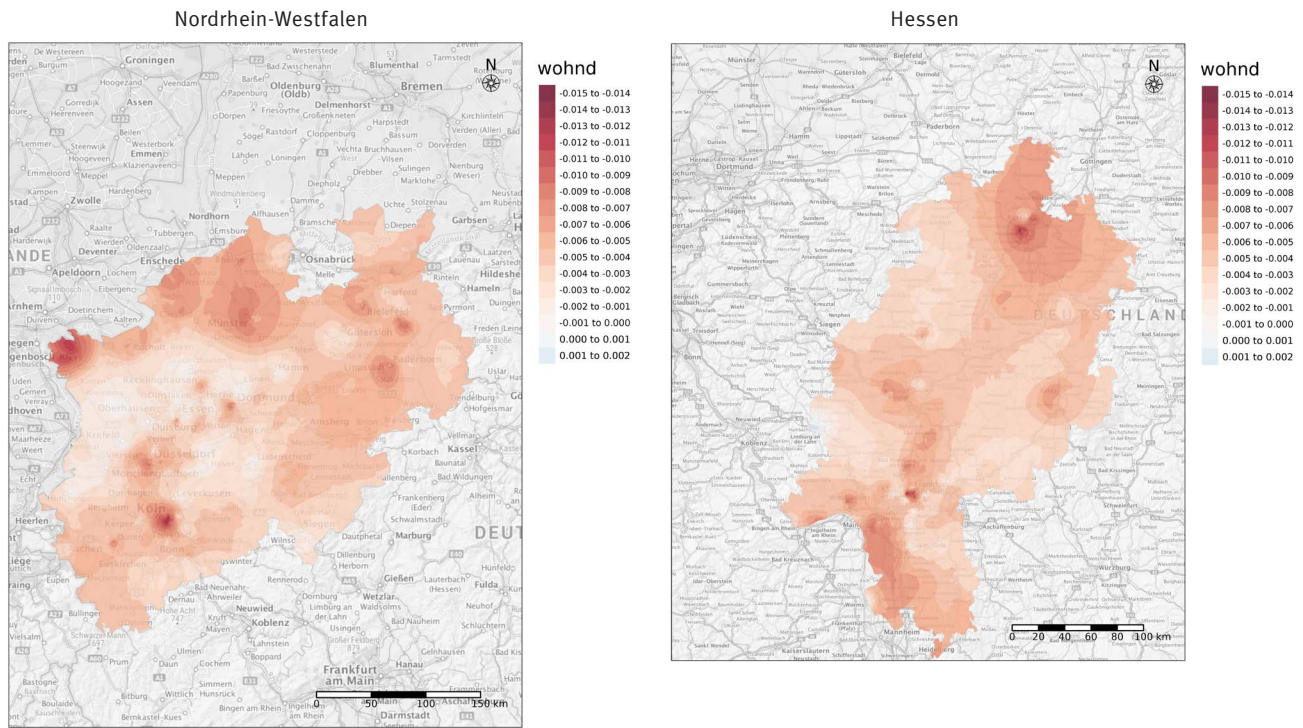
Hintergrundkarte: ©Bundesamt für Kartographie und Geodäsie (2020), TopPlusOpen

2022 - 0146



**Grafik 3**

**Punktschätzung des Koeffizienten des Regressors Wohndauer im Grundmodell mittels geografisch gewichteter Regression (GWR)  
100-m-Raster**



Hintergrundkarte: ©Bundesamt für Kartographie und Geodäsie (2020), TopPlusOpen

2022 - 0147

höchsten Ladungswerten der ersten und zweiten Komponente ausgewählt. Es zeigt sich, dass in der ersten Komponente tendenziell eine Variable mit größerem Deckungsbereich aus der Makrolage dominiert und für die zweite orthogonale Komponente ein Regressor aus der Mikrolage. Die Modelloptimierung insgesamt wurde anschließend mit einem iterativen Forward-/Backward-selectionansatz<sup>9</sup> vorgenommen. Dabei wurden zwei unterschiedliche Gütemaße betrachtet: zum einen Akaikes Information Criterion (AIC) und zum anderen Moran's I. Die Resultate waren jedoch weitgehend ähnlich.

Zum Vergleich der auf diese Weise optimierten linearen Regressionen wurden für die beiden ausgewählten Datensätze jeweils Modelle berechnet, die räumliche Abhängigkeiten strukturell berücksichtigen. In der Literatur (zum Beispiel LeSage/Pace, 2009) wird grundsätzlich zwischen Modellen mit einer Korrelation der abhängigen Variable (Spatial Autoregression Model – SAR), des

Fehlerterms (Spatial Error Model – SEM) oder einer Kombination aus beiden mit einer (optionalen) Berücksichtigung von Effekten der unabhängigen Variablen benachbarter Beobachtungen (Manski-Modell) unterschieden. Es gibt allerdings noch Varianten mit unterschiedlichen Gewichtsmatrizen für jeweils den autoregressiven und den Fehlerterm (Spatial Autoregression Confused – SAC). Für die Auswahl des passenden Modells werden in der Literatur verschiedene Herangehensweisen vorgeschlagen (INSEE, 2018): Bottom-up-Ansätze gehen vom einfachen multiplen linearen Regressionsmodell aus, ein Top-down-Ansatz startet mit einem komplexeren räumlichen Modell. Die verschiedenen Herangehensweisen differenzieren die jeweiligen Ausgangsmodelle schrittweise mittels statistischer Tests aus. Im Grundmodell spiegeln sich die Einflüsse der Lage auf die Wohnkosten (neben den nicht berücksichtigten individuellen Ausstattungsmerkmalen der Wohnungen) in den Fehlertermen wider. Dass Wohnungseigenschaften wie Wohnfläche, Dauer des Mietverhältnisses oder Eigentümerstruktur unmittelbaren Einfluss auf die Wohnkosten benachbarter

<sup>9</sup> Eine Übersicht über Selektions- und Regularisierungsmethoden findet sich beispielsweise in Hastie und andere (2009).

Wohnräume haben, ist wenig plausibel. Auf den ersten Blick erscheint daher für die vorgesehene Analyse das SEM sachlich am geeignetsten. Tatsächlich resultiert der in der oben genannten Literatur einfache Bottom-up-Ansatz, bei dem die Modellselektion zwischen einfachem MLR, SEM und SAR anhand (robuster) LM-Tests<sup>10</sup> mit jeweils der Nullhypothese „nicht SAR“ und „nicht SEM“ vorgenommen wird, in der Wahl eines SEM. Dieses Modell stellt das räumliche Analogon zur Autokorrelation bei Zeitreihen dar und lautet in allgemeiner Matrixschreibweise

$$(3) \quad \mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u} \text{ mit } \mathbf{u} = \lambda \cdot \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon},$$

wobei  $\mathbf{Y}$  die abhängige und  $\mathbf{X}$  die unabhängigen Variablen,  $\boldsymbol{\beta}$  einen Parametervektor,  $\mathbf{u}$  den Residuenvektor,  $\lambda$  den räumlichen Korrelationseffekt beschreibt.  $\mathbf{W}$  bezeichnet die als exogen angenommene Gewichtsmatrix und  $\boldsymbol{\varepsilon}$  einen Vektor mit identisch und unabhängig verteilten Fehlern, der für eine Maximum-Likelihood-Schätzung als homoskedastisch normalverteilt angenommen wird ( $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \cdot \mathbf{I}_n)$ ). Die Gewichtsmatrix hängt von der Wahl der Distanzfunktion und gegebenenfalls einer Bandbreite ab, wodurch Annahmen hinsichtlich der Struktur der wechselseitigen Beeinflussung benachbarter Beobachtungen getroffen werden. In welchem Umfang sich diese Wahl auf die Ergebnisse auswirkt, ist in der Literatur umstritten. LeSage und Pace (2014) argumentieren, dass bei unterschiedlicher Spezifikation von  $\mathbf{W}$  abweichende Ergebnisse auf ein unpassend gewähltes Modell hindeuten und vermeintliche Diskrepanzen auf einer falschen Interpretation der Ergebnisse basieren. [↘ Tabelle 2](#) stellt die Punktschätzwerte der einzelnen Modelle für die Koeffizienten der Wohndauer und des Logarithmus der Wohnfläche einander gegenüber. Es zeigt sich, dass der Logarithmus der Wohnfläche im Grundmodell für Nordrhein-Westfalen und Hessen eine statistisch signifikante Verzerrung aufweist. Das SEM weist mit einem vergleichsweise hohen Lambda eine scheinbar starke räumliche Korrelation der Fehlerterme aus, allerdings ist die Beziehung zwischen den Indikatoren Lambda und Moran's I nichtlinear. Dies bedeutet, dass bereits eine betragsmäßig niedrige Korrelation gemessen mit Moran's I zu einem betragsmäßig hohen Wert von Lambda führt, das für stärkere Korrelation gegen  $\pm 1$  in Sättigung läuft. Insgesamt liefern die mit

10 Ein Lagrange-Multiplier-(LM)-Test ist ein Hypothesentest, der mittels statistischer Analyse Aussagen über die Gültigkeit einer (Null-)Hypothese im Vergleich zur alternativen Hypothese erlaubt.

**Tabelle 2**

**Modellvergleich – Punktschätzungen der Koeffizienten für die Wohndauer und den Logarithmus der Wohnfläche**

	Nordrhein-Westfalen	Hessen
Grundmodell		
Wohndauer	-0,0036 (0.0001)	-0,0040 (0.0003)
Wohnfläche (logarithmiert)	0,7494 (0.0091)	0,6297 (0.0176)
Akaikes Information Criterion (AIC)	22 715	11 256
Optimiertes Modell = Grundmodell + räumliche Kovariate		
Wohndauer	-0,0039 (0.0001)	-0,0051 (0.0003)
Wohnfläche (logarithmiert)	0,7929 (0.0074)	0,6847 (0.0142)
Akaikes Information Criterion (AIC)	14 543	6 805
Spatial Error Model (SEM)		
Wohndauer	-0,0037 (0.0001)	-0,0049 (0.0003)
Wohnfläche (logarithmiert)	0,7924 (0.0044)	0,6877 (0.0086)
Lambda	0,912 (0.0069)	0,894 (0.0083)
Akaikes Information Criterion (AIC)	12 788	6 707
p-Wert Hausman-Test	< 0,001	0,003
Spatial Durbin Model		
Wohndauer	-0,0037 (0.0001)	-0,0049 (0.0003)
Wohnfläche (logarithmiert)	0,7930 (0.0044)	0,6894 (0.0086)
Lag. Wohndauer	0,0053 (0.0011)	0,0069 (0.0015)
Lag. Wohnfläche (logarithmiert)	-0,7519 (0.0244)	-0,6089 (0.0284)
Rho	0,879 (0.0087)	0,839 (0.0113)
Akaikes Information Criterion (AIC)	12716	6626
p-Wert Lagrange-Multiplier-Test <sup>11</sup> auf räumliche Residuenkorrelation	0,129	0,266

1 Getestet wird in einem SAC-beziehungswise Manski-Modell auf  $\lambda = 0$  (Anselin, 1988b). In runden Klammern unter den Werten sind die (für Grund- und optimiertes Modell geclusterten) Standardfehler angegeben.

räumlichen Kovariaten erweiterten Grundmodelle ähnliche Modellgüten (gemessen mit AIC) und vergleichbare Punktschätzwerte für die betrachteten Variablen wie die Modelle räumlich mit korreliertem Fehlerterm. Die Hypothese, dass ein SEM den zugrunde liegenden datengenerierenden Prozess vollständig abbildet, wird jedoch durch einen Hausman-Test<sup>11</sup> für Nordrhein-Westfalen und Hessen verworfen. Dies lässt sich durch im Modell

11 LeSage/Pace (2009), hier: Abschnitt 3.3.1.

fehlende, mit den Regressoren korrelierende Variablen erklären, was auf eine Modellgleichung mit autoregressivem Term der abhängigen und einem sogenannten Spatial Lag der Regressoren führt (Spatial Durbin Model;  $\rho$  bezeichnet den Koeffizienten der Autoregression und  $\theta$  den Parametervektor der räumlich korrelierten Regressoren):

$$(4) \quad Y = \rho \cdot WY + X \cdot \beta + WX\theta + \varepsilon$$

Sowohl für Nordrhein-Westfalen als auch für Hessen zeigt die Verbesserung der Modellgüte gemessen anhand des AIC an, dass das Spatial Durbin Model die Zusammenhänge treffender abbildet als ein SEM. Die Interpretation der geschätzten Regressorkoeffizienten ist für die Modelle mit autoregressivem Term der abhängigen Variablen jedoch deutlich schwieriger als für MLR oder SEM. Grund dafür ist, dass sie nicht wie bei Letzteren den marginalen Effekt einer Variablen auf den Regressanden darstellen. Hier kommen zu direkten Effekten einer Änderung einer Variablen  $x_i$  in Beobachtung  $i$  auf die abhängige Variable  $y_i$  noch indirekte Effekte auf die übrigen Beobachtungen aufgrund des autoregressiven Terms hinzu.

Die erweiterten multiplen Regressionsmodelle bieten insofern gegenüber strukturell komplexeren Modellen den Vorteil, dass sie die räumlichen Korrelationen transparenter abbilden. Dabei hängen jedoch die durch Optimierung selektierten und als die Korrelation vermittelnd angenommenen Kovariate und letztlich die Modellgüte offenkundig von der Grundmenge an verfügbaren räumlichen Charakteristika ab. [Tabelle 3](#) zeigt exemplarisch für Hessen die gewählten Regressoren und stellt die beiden Ansätze zur Erreichbarkeitsberechnung (Isochronen- und Rasteransatz) einander gegenüber. Die Variablenselektion für Nordrhein-Westfalen ist deutlich heterogener, mit wesentlichen Unterschieden zwischen beiden Ansätzen. Hinsichtlich der erreichten Modellgüte sind beide Ansätze für die jeweiligen Datensätze hingegen vergleichbar. Das optimierte Modell mit Variablen des Rasteransatzes verwendet tendenziell eine kleinere Anzahl an Regressoren bei leichten Abstrichen hinsichtlich des adjustierten  $R^2$  und der resultierenden räumlichen Residuenkorrelation. Für die folgende Indikatorenberechnung wird jeweils auf das etwas einfachere Modell basierend auf dem Rasteransatz zurückgegriffen.

**Tabelle 3**

Identifizierte räumliche Charakteristika mit jeweiligem Einzugsgebiet für Hessen differenziert nach Berechnungstyp der Erreichbarkeitsvariablen

Modell 1: Isochronenansatz	Modell 2: Rasteransatz
Umsatz (Radius 15 km, 50 km)	Umsatz (Radius 15 km, 50 km)
	Fahrzeit zum/zur nächstgelegenen > Mittel- oder Oberzentrum, > Flughafen, > Fernverkehrsbahnhof, > Hochschule, > Gymnasium, > sozialen Einrichtung
Bevölkerungsanteil Nichtdeutscher (250 m), Bevölkerungsanzahl (3 000 m)	Bevölkerungsanteil Nichtdeutscher (250 m), Bevölkerungsanzahl (3 000 m)
Flächenanteil Grünbedeckung (500 m), Flächenversiegelung (700 m)	
Anzahl der erreichbaren Grundschulen (60 min), Supermärkte (30 min), Hochschulen (30 min), sozialen Einrichtungen (9 min), Freizeitsporteinrichtungen (9 min)	
Durchschnittliche Gebäudehöhe (100 m)	Durchschnittliche Gebäudehöhe (100 m)
Summe aller ÖPNV-Abfahrten wochentags (750 m)	
Adjustiertes $R^2$	
0,53	0,53
Räumliche Residuenkorrelation (I)	
0,06	0,06

Iterative Forward-/Backwardselection, Kriterium: Moran's I

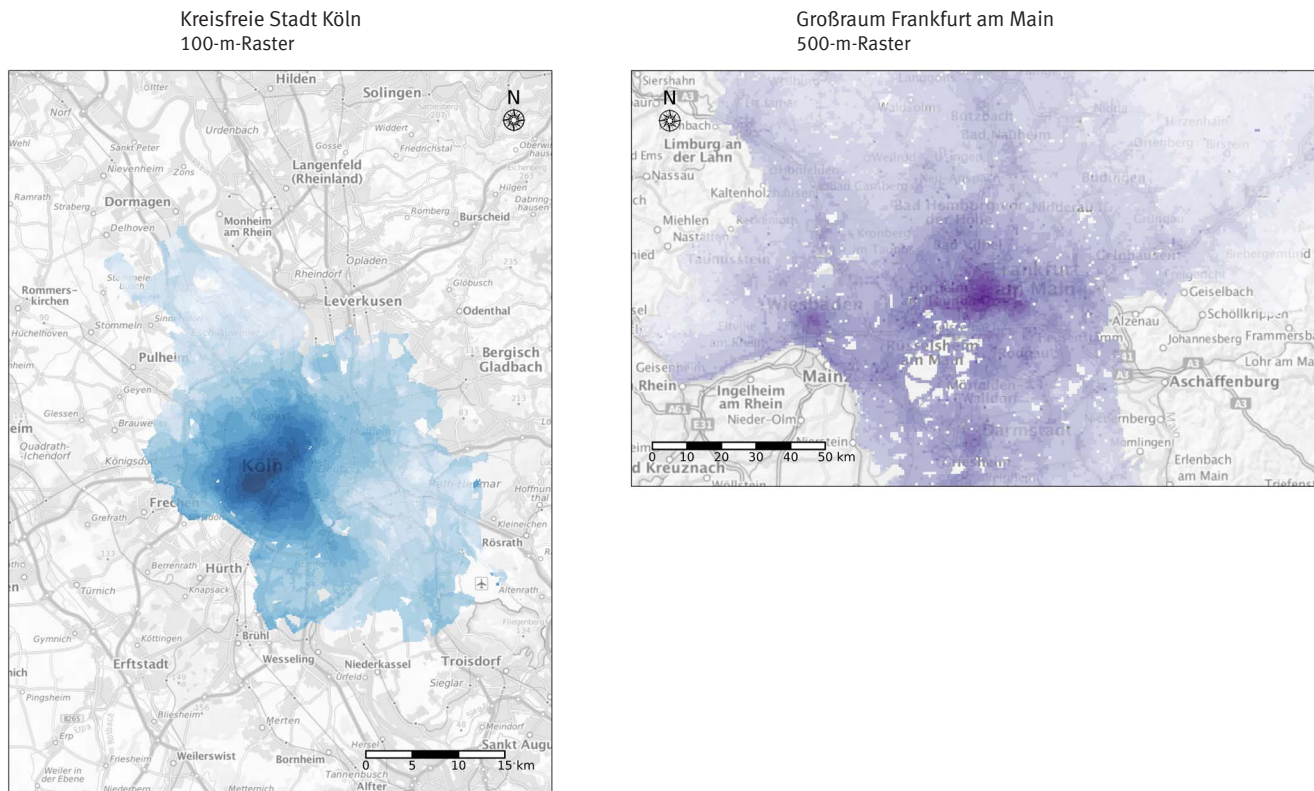
## Ableitung eines qualitativen Indikators für Wohnkosten basierend auf den identifizierten räumlichen Charakteristika

Mittels Partial Least Squares (PLS) wurden schließlich aus den räumlichen Regressoren der optimierten Modelle orthogonale Komponenten mit maximaler Varianz bei gleichzeitiger Maximierung der Korrelation mit der abhängigen Variablen abgeleitet. Anhand der so ermittelten Regressorgewichte wurde jeweils für die erste Komponente ein experimenteller Indikator berechnet, der als ein erster Ansatz zur qualitativen Abbildung des Merkmals „Lage“ vorgeschlagen wird. Diese Interpretation steht unter der Prämisse, dass auch für diejenigen Gitterzellen, für die keine Mikrozensus-Beobachtungen vorliegen, dieselben Gewichte wie für die vorliegende Stichprobe gelten. Beispiele hierfür sind in [Grafik 4](#) dargestellt. Zur Reduktion des Auf-

wands der Datenaufbereitung wurde für den Großraum Frankfurt/Main der Indikator jeweils im 500-Meter-Raster berechnet, das heißt, für die konzentrische 100-Meter-Gitterzelle wurde jeweils das Ergebnis für die gesamte 500-Meter-Gitterzelle ausgewiesen. Der Indikator für die kreisfreie Stadt Köln ist im 100-Meter-Raster berechnet und dargestellt. Je stärker die Farbsättigung einer Gitterzelle ausfällt, desto höher sind tendenziell die lagebedingten Wohnkosten. Umgekehrt weist eine schwache Einfärbung auf eine Lage mit vergleichsweise niedrigen Kosten hin.

**Grafik 4**

Experimentelle qualitative Indikatoren zur Lageabhängigkeit der Wohnkosten basierend auf räumlichen Charakteristika




Für die Darstellung wurden die separat berechneten Indikatoren jeweils für das betrachtete Gebiet normiert und die Farbsättigung maximal gespreizt. Hintergrundkarte: ©Bundesamt für Kartographie und Geodäsie (2020), TopPlusOpen

2022 · 0148

## 5

## Zusammenfassung und Ausblick

Das in diesem Aufsatz beschriebene Projekt liefert einen Einblick in die Möglichkeiten, die Analysen auf der Basis integrierter statistischer und geografischer Informationen bieten. Durch Verschneidung des geokodierten Mikrozensus 2018 und der Zusatzerhebung „Wohnen“ mit externen Geodaten lassen sich in einem multiplen linearen Regressionsmodell räumliche Charakteristika identifizieren, die Wohnkosten erklären und gleichzeitig die räumliche Korrelation der Residuen senken. Es zeigt sich, dass ausgehend von einem einfachen Grundmodell mit wenigen die Wohnungscharakteristika beschreibenden Variablen durch das Hinzufügen räumlicher Regressoren die Varianz der abhängigen Variable besser erklärt werden kann. Aus den identifizierbaren räumlichen Variablen wurde ein experimenteller qualitativer Indikator für die Lageabhängigkeit der Kosten gemieteten Wohnraums abgeleitet. Unter der Annahme, dass der Erklärungsgehalt und die Gewichte der identifizierbaren räumlichen Merkmale im gesamten betrachteten Gebiet gleich sind, wurde diese auch für Gitterzellen berechnet, die nicht oder nur in geringem Maße von der (Klumpen-) Stichprobe des Mikrozensus abgedeckt sind. Die Gültigkeit der beschriebenen Annahme ist allerdings noch zu prüfen, weswegen die Ergebnisse als experimentelle Statistik zu verstehen sind. Außerdem ist bei ihrer Interpretation zu berücksichtigen, dass ein Erklärungsbeitrag einer Variablen nicht zwangsläufig einen kausalen Zusammenhang vermittelt. Vielmehr sind Wohnkosten verknüpft mit der Dynamik der Demografie und Einkommensstruktur der Mieterinnen und Mieter. Diese wiederum sind interdependent mit der Entwicklung von Infrastruktur wie Bildungseinrichtungen, Einrichtungen der Gesundheitsvorsorge und der Freizeitgestaltung oder Geschäften zur Deckung des täglichen Bedarfs. Das Ergebnis dieses rein statistisch-technischen Ansatzes verdeutlicht jedoch, dass eine breite und diverse Palette raumbezogener Daten für umfassende Analysen notwendig ist. Hierfür ist der Aufbau einer entsprechenden IT-Infrastruktur mit einer Datenbank erforderlich („Gitterzellendatenbank“, siehe Fußnote 1), die die einfache Generierung, Verwaltung und den Zugriff auf räumliche Charakteristika auf Gitterzellenebene ermöglicht.

Die starke Korrelation einzelner räumlicher Charakteristika stellt für die Identifikation geeigneter Korrelate eine Herausforderung dar. Dieser kann beispielsweise mit sinnvoller Vorauswahl durch Fachleute begegnet werden, um die Robustheit der Resultate zu erhöhen und die Interpretation der Ergebnisse zu erleichtern. Liegt der Schwerpunkt eher auf der Prädiktion der abhängigen Variablen und weniger auf der Identifikation räumlicher Korrelation<sup>12</sup>, bietet sich der Einsatz komplexerer Modelle oder von Machine Learning an. Komplexe Modelle berücksichtigen räumliche Abhängigkeiten strukturell, in Prozesse des Maschinellen Lernens können mehrere hundert Merkmale einfließen (Rattay/Schäffner, 2019). Die Bewertung dieser Modelle bei Verwendung hoch aufgelöster räumlicher Charakteristika bietet ein weites Feld für künftige wissenschaftliche Arbeiten. Der experimentelle Indikator zur Beschreibung des Einflusses des Merkmals „Lage“ auf Wohnkosten soll als ein möglicher Ansatz und Anregung für weitere Untersuchungen dienen. Insbesondere die Gültigkeit der zugrunde liegenden Annahme, dass der Erklärungsgehalt und die Gewichte der identifizierten räumlichen Merkmale im gesamten betrachteten Raumbereich identisch sind, sowie die Aussagekraft und Validität des Indikators in hoher räumlicher Tiefe bedürfen zusätzlicher Analyse. 

12 Beispielsweise zur Prädiktion von Immobilienpreisen (Horvath und andere, 2019).

### LITERATURVERZEICHNIS

---

Anselin, Luc. *Spatial Econometrics: Methods and Models*. 1988a.

Anselin, Luc. *Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity*. In: *Geographical Analysis*. Jahrgang 20. Ausgabe 1/1988, Seite 1 ff. 1988b.

BBSR (Bundesinstitut für Bau-, Stadt- und Raumforschung). *Raumabgrenzungen*. 2021. [Zugriff am 2. Mai 2022]. Verfügbar unter: [www.bbsr.bund.de](http://www.bbsr.bund.de)

Bihler, Wolf/Zimmermann, Daniel. *Die neue Mikrozensusstichprobe ab 2016*. In: *WISTA Wirtschaft und Statistik*. Ausgabe 6/2016, Seite 20 ff.

Bivand, Roger/Millo, Giovanni/Piras, Gianfranco. *A Review of Software for Spatial Econometrics in R*. In: *Mathematics*. Jahrgang 9. Ausgabe 11/2021, Seite 1276 ff. [Zugriff am 2. Mai 2022]. Verfügbar unter: [dx.doi.org/10.3390/math9111276](https://doi.org/10.3390/math9111276)

BKG (Bundesamt für Kartographie und Geodäsie). *Digitales Landbedeckungsmodell für Deutschland, Stand 2018 (LBM-DE2018)*. 2018a. [Zugriff am 2. Mai 2022]. Verfügbar unter: [gdz.bkg.bund.de](http://gdz.bkg.bund.de)

BKG (Bundesamt für Kartographie und Geodäsie). *Points of Interest Bund (POI-Bund)*. 2018b. [Zugriff am 2. Mai 2022]. Verfügbar unter: [gdz.bkg.bund.de](http://gdz.bkg.bund.de)

BKG (Bundesamt für Kartographie und Geodäsie). *RoutingPlus-Dienst (web\_ors)*. 2021. [Zugriff am 2. Mai 2022]. Verfügbar unter: [gdz.bkg.bund.de](http://gdz.bkg.bund.de)

Brunsdon, Chris A./Fotheringham, Stewart/Charlton, Martin E. *Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity*. In: *Geographical Analysis*. Jahrgang 28. Ausgabe 4/1996, Seite 281 ff.

Hastie, Trevor/Tibshirani, Robert/Friedman, Jerome. *The Elements of Statistical Learning*. New York 2009.

Horvath, Sabine/Soot, Matthias/Weitkamp, Alexandra/Neuner, Hans. *Künstliche neuronale Netze in der Immobilienwertermittlung*. In: *avn. allgemeine vermessungsnachrichten*. Ausgabe 8-9/2019, Seite 199 ff.

INSEE (Institut national de la statistique et des études économiques). *Handbook of Spatial Analysis*. 2018. [Zugriff am 2. Mai 2022]. Verfügbar unter: [ec.europa.eu](http://ec.europa.eu)

Lancaster, Kelvin J. *A New Approach to Consumer Theory*. In: *The Journal of Political Economy*. Jahrgang 74. Ausgabe 2/1966, Seite 132 ff.

Lerner, Matthew R./Kocher, Jesse G./Matieu, Michael C. *Systems, Techniques, And Methods for providing Location Assessments*. United States Patent No. US8738422B2, Patenthalter: Walk Score Management, LLC, Seattle 2014.

LeSage, James/Pace, Robert Kelley. *Introduction to Spatial Econometrics*. New York 2009.

LeSage, James/Pace, Robert Kelley. *The Biggest Myth in Spatial Econometrics*. In: *Econometrics*. Jahrgang 2. Ausgabe 4/2014, Seite 217 ff.

## LITERATURVERZEICHNIS

---

Montero, José María/Minguez, Román/Fernández-Avilés, Gema. *Erratum to Housing Price Prediction: parametric versus semiparametric spatial hedonic models*. In: Journal of Geographical Systems. Jahrgang 20. Ausgabe 1/2018, Seite 107 ff.

Moran, P. A. P. *Notes on continuous stochastic phenomena*. In: Biometrika. Jahrgang 37. Ausgabe 1/2, Seite 17 ff. Oxford 1950.

Olbricht, Roland. *Overpass-API*. 2018. [Zugriff am 4. Mai 2022]. Verfügbar unter: <https://github.com/drolbr/Overpass-API>, Instanz-URL: [overpass-api.de](https://overpass-api.de)

Promann, Johannes. *Die Berücksichtigung des Wohnwertmerkmals Lage in den Mietspiegeln der deutschen Großstädte*. Band 8 der Schriftenreihe Industrieökonomik. 2012.

Rattay, M./Schäffner, L. *Mit Algorithmen zur Lagebewertung 2.0*. gis.Business, Ausgabe 06/2019.

Rosen, Sherwin. *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*. In: The Journal of Political Economy. Jahrgang 82. Ausgabe 1/1974, Seite 34 ff.

Statistisches Bundesamt. *Qualitätsbericht Mikrozensus 2018*. [Zugriff am 2. Mai 2022]. Verfügbar unter: [www.destatis.de](http://www.destatis.de)

Tobler, Waldo R. *A Computer Movie Simulating Urban Growth in the Detroit Region*. In: Economic Geography. Jahrgang 46. 1970. Seite 234 ff.

Veie, Kathrine Lausted/Panduro, Toke Emil. *An alternative to the standard spatial econometric approaches in hedonic house price models*. IFRO working paper 2013/18. Kopenhagen 2013.

Voß, Winrich/Bakker, Keno. *Einsatz von Geodaten für die Immobilienwertermittlung*. In: Meinel, Gotthard/Schumacher, Ulrich/Schwarz, Steffen/Richter, Benjamin (Herausgeber). *Flächennutzungsmonitoring IX. Nachhaltigkeit der Siedlungs- und Verkehrsentwicklung?* Leibniz-Institut für ökologische Raumentwicklung. IÖR Schriften. Band 73. 2017. Seite 113 ff.

## RECHTSGRUNDLAGEN

---

Artikel 13 des Gesetzes zur Förderung der elektronischen Verwaltung sowie zur Änderung weiterer Vorschriften vom 25. Juli 2013 (BGBl. I Seite 2749): Änderung des Bundesstatistikgesetzes (BGBl. I Seite 2756 f.).

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) in der Fassung der Bekanntmachung vom 20. Oktober 2016 (BGBl. I Seite 2394), das zuletzt durch Artikel 2 des Gesetzes vom 14. Juni 2021 (BGBl. I Seite 1751) geändert worden ist.

**Herausgeber**  
Statistisches Bundesamt (Destatis), Wiesbaden

---

**Schriftleitung**  
Dr. Daniel Vorgrimler  
Redaktion: Ellen Römer

---

**Ihr Kontakt zu uns**  
[www.destatis.de/kontakt](http://www.destatis.de/kontakt)

---

**Erscheinungsfolge**  
zweimonatlich, erschienen im Juni 2022  
Ältere Ausgaben finden Sie unter [www.destatis.de](http://www.destatis.de) sowie in der [Statistischen Bibliothek](#).

---

Artikelnummer: 1010200-22003-4, ISSN 1619-2907

---

© Statistisches Bundesamt (Destatis), 2022  
Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.