Inaugural-Dissertation zur Erlangung des akademischen Grades

**Dr. rer. pol.**

zum Thema

# Einsatzmöglichkeiten von Mobilfunkdaten in der amtlichen Statistik

an der

Professur für Angewandte Statistik
Fachbereich Wirtschaftswissenschaft
Freie Universität Berlin

vorgelegt von

**Sandra Hadam, M. Sc.**
aus Berlin

Berlin, 2022

Sandra Hadam, *Einsatzmöglichkeiten von Mobilfunkdaten in der amtlichen Statistik,*
August 2022

Erstgutachter: Prof. Dr. Timo Schmid
Zweitgutachter: Prof. Dr. Markus Zwick

Ort:
Berlin

Tag der Disputation:
13. Februar 2023

## Danksagung

Zuallererst möchte ich meinem Doktorvater Prof. Dr. Timo Schmid (Otto-Friedrich-Universität Bamberg, Deutschland) meinen großen Dank aussprechen. Seine Unterstützung und beständige Ermutigung waren von unschätzbarem Wert für den Erfolg dieses Vorhabens und ohne dessen Zuspruch ich diese Arbeit nicht hätte beenden können.

Sehr dankbar bin ich auch Prof. Dr. Markus Zwick (Goethe-Universität Frankfurt am Main, Deutschland) für die zahlreichen Diskussionen sowie Anregungen in den letzten Jahren zu diesem neuen Forschungsthema in der amtlichen Statistik und gleichfalls für seine Unterstützung, diese Arbeit im Statistischen Bundesamt zu finalisieren.

Ein großer Dank gebührt meinen Kolleginnen und Kollegen am Institut für Forschung und Entwicklung in der Bundesstatistik im Statistischen Bundesamt sowie meinen Ko-Autorinnen, insbesondere am Lehrstuhl für Angewandte Statistik der Freien Universität Berlin, die mich auf dem Weg zu dieser Arbeit begleitet und vor allem in schwierigen Momenten immer wieder aufgebaut haben.

Meine Arbeit selbst widme ich meiner Familie – meinen Eltern, meinem Bruder und seiner Familie sowie meinem Partner, meinem größten Glück, dessen Lektorat und Korrekturlesen dieser Arbeit den Feinschliff gab. Danke für eure bedingungslose Unterstützung und Zuversichtlichkeit, die ich selbst nicht immer hatte, eure Geduld und Liebe. Ohne euch hätte ich es nicht geschafft. Euch gehört mein größter Dank.

## Liste der Veröffentlichungen

Die unten aufgeführten Veröffentlichungen sind das Ergebnis der im Rahmen dieser Dissertation durchgeführten Forschungsarbeiten unter dem Titel "Einsatzmöglichkeiten von Mobilfunkdaten in der amtlichen Statistik". Die Kapitel 1 und 4 wurden entsprechend der aufgeführten Veröffentlichungen mit Koautoren erstellt.

1. Hadam, S., Schmid, T., und Simm, J. (2020). **Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland**. In: Klumpe, B., Schröder, J., und Zwick, M. (Hrsg.), Qualität bei zusammengeführten Daten, S. 27–44. Wiesbaden: Springer VS. Doi: `https://doi.org/10.1007/978-3-658-31009-7_3`.

2. Hadam, S. (2023). **Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten**. AStA Wirtschafts- und Sozialstatistisches Archiv. Doi: `https://doi.org/10.1007/s11943-023-00320-2`.

3. Hadam, S. (2021). **Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik**. AStA Wirtschafts- und Sozialstatistisches Archiv, 15: 197–235. Doi: `https://doi.org/10.1007/s11943-021-00294-z`.

4. Hadam, S., Würz, N., Kreutzmann, A.-K., und Schmid, T. (2023). **Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany**. *Working paper*.

# Inhaltsverzeichnis

# Einleitung

Im Rahmen der allgemein fortschreitenden Digitalisierung ist die amtliche Statistik gefordert, neue Datenquellen zu erforschen und zu integrieren. In der Nutzung neuer digitaler Daten wird Potenzial für eine möglicherweise schnellere und präzisere amtliche Statistikproduktion gesehen.

Für eine faktengestützte Politikgestaltung müssen valide, aktuelle und kleinräumige Daten vorliegen. Beispielsweise wird die geografische Verteilung der aktuellen Bevölkerung, der Erwerbslosenquote oder des Pendlerverhaltens der Bevölkerung dazu verwendet, Entscheidungen über die Verteilung von Ressourcen zu treffen. Die amtliche Statistik übernimmt als Statistikproduzentin eine wichtige Funktion in diesem Entscheidungsprozess. Aufgrund des Stichprobencharakters vieler Erhebungen und der geringen Stichprobenanzahl in den meisten Statistiken können Aussagen zu kleinräumigen Regionen oder unterschiedlichen Bevölkerungsgruppen nicht getätigt werden. Die kontinuierliche und zeitnahe Erhebung ist mit traditionellen Erhebungsmethoden nicht gewährleistet. Hierzu wird eine Kombination aus alternativen Datenquellen sowie administrativen Daten und Befragungsdaten erforderlich.

Dies muss grundsätzlich unter Beachtung des Adäquationsproblems erfolgen. Unter Adäquation wird dabei die Übertragung idealtypischer theoretischer Probleme in statistische Begriffe durch dafür erhobene Daten verstanden, die dem Zweck empirischer Untersuchungen dienen (Grohmann, 1985). Dies trifft jedoch nur auf Primärerhebungen in der amtlichen Statistik zu, wohingegen neue digitale Daten wie z.B. Mobilfunkdaten nicht für statistische Zwecke erfasst werden und daher nicht dem Prinzip der Adäquation folgen. Damit wird das Adäquationsproblem erst nach dem Datenerhalt aufgegriffen, da hierbei initial geprüft werden muss, ob und inwieweit die neue Datenquelle notwendige Informationen zur Beschreibung der interessierenden Realität besitzt (Rendtel et al., 2022). Um dies beantworten zu können, werden die hier betrachteten amtlichen Statistiken als Maßstab für die Untersuchung der zur Verfügung stehenden Mobilfunkdaten verwendet, um den zu interessierenden Teil der Daten für die Untersuchungen herausfiltern zu können und einen ersten Ansatz für die jeweilige Fragestellung und das Unterstützungspotenzial der amtlichen Statistik herauszuarbeiten.

Angesichts der hohen Penetrationsrate mobiler Endgeräte in der Bevölkerung nach dem Statistischen Bundesamt (2021a) kann ein Teil der genannten Herausforderungen durch die Verwendung von Mobilfunkdaten ggf. bewerkstelligt werden. Hierbei muss beachtet werden, dass Mobilfunkdaten durch die hohe Penetrationsrate in Deutschland die realen Aufenthaltsorte der Bevölkerung ggf. adäquater abbilden können, als es die amtliche Bevölkerungsstatistik aufgrund einer idealisierten Datenerhebung ermöglicht. Die Herausforderung in dieser Arbeit besteht daher mitunter darin, die Mobilfunkdaten derart zu definieren und aufzubereiten, um einerseits die amtliche Statistik zu unterstützen und andererseits den Vorteil dieser Datenquelle hinsichtlich realer Bevölkerungsverhältnisse durch die Hilfestellung des Benchmarkings bzw. der Maßstabssetzung anhand der amtlichen Statistik nicht zu verlieren. Durch die Kombination von Mobilfunkdaten und Daten der amtlichen Statistik können sodann Schätzungen mittels verschiedener prädiktiver Verfahren für tiefer gegliederte Gebietseinheiten ausgegeben werden.

Um die Potenziale dieser Datenquelle einschätzen und benennen zu können, werden in dieser Arbeit konkrete Anwendungsfelder von Mobilfunkdaten in der amtlichen Statistik ermittelt und ihre Einsatzmöglichkeiten in der amtlichen Statistikproduktion beurteilt. Da sich Mobilfunkdaten durch ihre hohe räumliche und zeitliche Auflösung auszeichnen, liegt der Fokus hierbei auf der Einbindung der statischen und der dynamischen Bevölkerung aus Mobilfunkdaten deutscher Mobilfunkanbieter.

Aufgrund datenschutzrechtlicher Regelungen erhält das Statistische Bundesamt in Zusammenarbeit mit verschiedenen Mobilfunkdatenanbietern am deutschen Mobilfunkmarkt nur anonymisierte, aggregierte Mobilfunkdaten, welche anhand der Forschungsfragen entsprechend aufbereitet sind. Ein Vergleich unterschiedlich aufbereiteter Mobilfunkdaten soll bewusst zeigen, dass insbesondere verschiedene Datenaufbereitungsarten die Einsatzmöglichkeiten in der amtlichen Statistik bestmöglich gewährleisten. Das beigefügte Glossar (siehe S. 131) bietet weiterhin die Möglichkeit, die wichtigsten Fachbegriffe aus dem Bereich der Mobilfunkdaten, der verwendeten Fachstatistiken sowie aus der Statistik und Geowissenschaft nachzulesen.

Konkret werden in Teil I Anwendungsfelder im Zusammenhang mit Bevölkerungsdarstellungen zu bestimmten Zeitpunkten analysiert, indem grundlegend die Unterstützung der Bevölkerungsstatistik mit Hilfe von Mobilfunkdaten analysiert wird. Mit anonymisierten aggregierten statischen Mobilfunkdaten kann die Tages- und Wohnbevölkerung in Deutschland zeitnah und kleinräumig abgebildet werden, wie in Kapitel 1 dargelegt wird. Unter Verwendung einer nicht-parametrischen Kerndichteschätzung wird zudem die Möglichkeit geboten, die Mobilfunkdaten auf weitere interessierende Regionen – unabhängig von der zugrundeliegenden räumlichen Einheit – umzuverteilen und die Bevölkerungsverteilung valide wiederzugeben. Hierbei wird ein starker und positiver Zusammenhang der kleinräumigen statischen Mobilfunkdaten mit den georeferenzierten Bevölkerungszahlen des Zensus 2011 nachgewiesen. Diese

Erkenntnis wird genutzt, um darauf aufbauend in Kapitel 2 die aktuelle Bevölkerungszahl basierend auf der Bevölkerungsfortschreibung mit Hilfe von Mobilfunkdaten kleinräumig auf einem INSPIRE-konformen Raster umzuverteilen und als experimentelles georeferenziertes Produkt zu nutzen, bis die erste amtliche georeferenzierte Bevölkerungszahl des Zensus 2022 vorliegt. Anhand eines einfach umzusetzenden Verteilungsverfahrens durch die gruppenspezifischen Ziehungswahrscheinlichkeiten innerhalb der Mobilfunkdaten sowie einem speziellen Rundungsverfahren ist es möglich die aktuelle Bevölkerungszahl der Bevölkerungsfortschreibung kleinräumig valide zu verteilen und gleichzeitig die Eckwerte der Bevölkerungsfortschreibung auf einer höher aggregierten Ebene zu garantieren. Durch die Verwendung zusätzlicher Geodaten des Bundesamtes für Kartographie und Geodäsie kann die experimentelle georeferenzierte Bevölkerungszahl je Gitterzelle auf Plausibilität geprüft werden.

Dynamische Mobilfunkdaten dagegen geben die Bewegung eines Mobilfunkgerätes im Raum in einem 24-Stunden-Zeitintervall durch sogenannte Quelle-Ziel-Matrizen wieder, wodurch sich Aussagen zum Mobilitätsverhalten der Mobilfunknutzenden tätigen lassen. In Teil II werden folglich Nutzungsmöglichkeiten von Mobilfunkdaten in Zusammenhang mit der Bevölkerungsmobilität untersucht. Hierbei werden Quelle-Ziel-Matrizen verwendet, um in Kapitel 3 die Unterstützungsmöglichkeiten von Mobilfunkdaten in der amtlichen Pendlerrechnung zu ermitteln. Durch die Generierung und Filterung bestimmter Bewegungsströme aus den Mobilfunkdaten werden Ansätze hergeleitet, um das Bewegungsverhalten von Pendlern abzubilden. In Kapitel 4 werden Mobilfunkdaten weiterführend als Hilfsinformationen zusätzlich zu den bereits erhobenen (Befragungs-)Daten genutzt, um die Erwerbslosenquote der Arbeitskräfteerhebung, auch bekannt als Labour Force Survey, im Rahmen eines Small-Area-Verfahrens kleinräumiger zu schätzen. Die Arbeitskräfteerhebung ist darauf ausgerichtet, verlässliche Schätzungen zu Indikatoren bezüglich der zivilen Erwerbsbevölkerung auf einer groben räumlichen Auflösung in Deutschland wiederzugeben. Um politische Handlungsempfehlungen beispielsweise für Kommunen formulieren zu können, müssen die Schätzungen der Indikatoren für räumlich disaggregierte Ebenen abgeleitet werden. Unter Verwendung eines transformierten Fay-Herriot-Modells mit Bias-Korrektur wird die Erwerbslosenquote auf kleinräumiger Ebene der städtischen Gebiete geschätzt. Gleichzeitig sind Rückschlüsse auf Pendlerbewegungen in die Schätzung eingeflossen, um einerseits die pendelnde Bevölkerung in die Schätzung und andererseits das Phänomen der Stadtzuwanderung in Zusammenhang mit höheren Erwerbslosenquoten in städtischen Gebieten als im Umland einzubeziehen. Mobilfunkdaten greifen hierbei erstmals aktiv in die Schätzung eines amtlichen Indikators ein, bei der zum einen kleinräumig verlässliche Schätzer resultieren sowie im Rahmen einer alternativen Erwerbslosenquote den Aspekt der dynamischen Bevölkerung bzw. genauer der Pendler einfließen lässt.

Zusammenfassend stellen Kapitel 1 und 3 damit praktische Anwendungsfälle für eine pri-

märe Nutzung von Mobilfunkdaten dar. Darunter werden Aussagen bzw. Ergebnisse zu den entsprechenden Fragestellungen verstanden, die ausschließlich auf Basis der ausgewerteten Mobilfunkdaten getätigt werden. Kapitel 2 und 4 bauen auf die primären Nutzungsmöglichkeiten von Mobilfunkdaten aus Kapitel 1 und auch 3 auf, wobei Mobilfunkdaten weiterführend als Hilfsinformationen genutzt werden, um aktiv eine bestehende amtliche Statistik bzw. die Ergebnisse kleinräumiger valide zu schätzen und als experimentelles Produkt zu veröffentlichen. Diese Arbeit legt insgesamt dar, wie bestehende amtliche Statistiken in Kombination mit Mobilfunkdaten sowie verschiedenen Herangehensweisen erweitert bzw. unterstützt werden können.

# Teil I

# Nutzungsmöglichkeiten von Mobilfunkdaten zur Bevölkerungsdarstellung

# Kapitel 1

# Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland

## 1.1 Einleitung

Der Zensus gehört zu den elementaren Konzepten der Bevölkerungsstatistik. Das wichtigste Ziel der Volkszählungen ist dabei die Ermittlung der aktuellen Einwohnerzahl in Deutschland zu einem bestimmten Stichtag (Statistisches Bundesamt, 2016). Die Relevanz des Zensus ist dabei für politische und gesellschaftliche Entscheidungen unumstritten. Zuverlässige Kenntnisse über die Verteilung der Bevölkerung und die Einwohnerzahl eines Landes auf kleinstmöglicher geografischer Ebene sind für eine solide evidenzbasierte Politikgestaltung unerlässlich. Diese sind beispielsweise für die öffentliche Verwaltung relevant auf deren Grundlage über die lokale Infrastruktur, Schul- und Verkehrsplanung, Förderung der Bildung und Kultur oder Sozialleistungen entschieden wird.

Ehemals fand die Bevölkerungsabbildung ausschließlich mittels traditioneller Daten statt, die auf Primärerhebungen basierten, also einer Totalerhebung. Nach dem Volkszählungsboykott im Jahr 1982 begann in Deutschland ein schrittweises Umdenken von einer Totalerhebung hinzu einem registerbasierten Zensus (Grohmann, 2011; Heinzel, 2006). Der Zensus wird derzeit alle zehn Jahre durch die amtliche Statistik durchgeführt und ermittelt die Wohnbevölkerung der Bundesrepublik, welche für eine faktengestützte Politikgestaltung jährlich fortgeschrieben werden muss. Als Grundlage dieser jährlichen Fortschreibungen dient die letzte Volkszählung und wird mithilfe von verschiedenen administrativen Daten der Meldebehörden und Standesämter zum Stichtag (31.12.) aktualisiert. Die Qualität der jährlichen Fortschreibungen ist von der

Vollständigkeit und Genauigkeit der durch die Ämter und Behörden gelieferten Daten abhängig. Mit steigendem Abstand zur letzten Volkszählung werden die Ergebnisse der Bevölkerungsfortschreibungen ungenauer (Statistisches Bundesamt, 2019b).

Eine bestehende Herausforderung ist das Umdenken von einer statischen hin zu einer dynamischen (zeitlich aktuellen) Bevölkerung. Zeitnahe und detaillierte Informationen darüber, wo sich die Individuen im Tagesverlauf aufhalten sind nicht nur im Fall von Katastrophen, Epidemien oder Konflikten entscheidend (Deville et al., 2014), sondern spielen bspw. auch bei der Regional- und Verkehrsplanung eine entscheidende Rolle. Diese Dynamik kann mit traditionellen Daten nicht erfasst werden. Durch den Einsatz administrativer Daten zur Ermittlung und Lokalisierung der Einwohnerzahl im registerbasierten Zensus konnten die Auskunftsgebenden zwar teils entlastet werden, allerdings blieben die Herausforderungen in der Abbildung einer dynamischen Bevölkerung und der Bevölkerungsfortschreibung weiter bestehen.

Neue digitale Daten haben das Potenzial diese Herausforderungen zu lösen. Sie sind eine Folge der digitalen Revolution und die daraus entstehenden enormen Datenmengen das Resultat der verschiedenen Kommunikationsarten und werden daher über die drei V definiert: Volume, Velocity und Variety (Daas et al., 2013; Zwick, 2016). Im Rahmen der allgemein fortschreitenden Digitalisierung ist auch die amtliche Statistik gefordert, diese neuen Datenquellen zu erforschen und einzusetzen und ihre Prozesse und Verfahrensweisen entsprechend auszurichten. Durch die Nutzung solcher Daten wird Potenzial für eine möglicherweise schnellere, präzisere und kostengünstigere amtliche Statistikproduktion sowie eine eventuelle Entlastung der Auskunftsgebenden gesehen (Wiengarten und Zwick, 2017). Darüber hinaus können neue digitale Daten völlig neue Erkenntnisse liefern, was mit traditionellen und administrativen Daten derzeit nicht möglich ist. Diese sollen bzw. können die traditionellen Daten jedoch nicht vollständig ersetzen, sondern sollen zukünftig als sogenannte „blended data“ mit Befragungs- und administrativen Daten verknüpft werden (Wiengarten und Zwick, 2017).

Beispielsweise können Mobilfunkdaten zu einer dynamischen und zeitnäheren Schätzung der Bevölkerung beitragen. Das Statistische Bundesamt führt daher Machbarkeitsstudien zur Integration von Mobilfunkdaten in die amtliche Statistik durch. Hierbei werden Analysen zur Abbildung der Wohn- und Tagesbevölkerung mittels Mobilfunkdaten durchgeführt. Zu den Nutzungsmöglichkeiten von Mobilfunkdaten für statistische Zwecke existieren bereits diverse wissenschaftliche Studien. De Meersman et al. (2016) beurteilten beispielsweise die Qualität von Mobilfunkdaten als Quelle für die amtliche Statistik. Deville et al. (2014) nutzten Mobilfunkdaten, um eine dynamische Bevölkerung darstellen zu können. Makita et al. (2013) ermittelten, ob Mobilfunkdaten genutzt werden können, um die Bevölkerung in kleinräumigen Gebieten (Small Area) zu schätzen. Schmid et al. (2017) bestimmten soziodemografische Indikatoren (etwa Alphabetisierung im Senegal) basierend auf Umfragedaten in Kombination mit Mobilfunkdaten

als Hilfsinformation, um die Indikatoren ebenfalls mittels Small-Area-Verfahren (vgl. Tzavidis et al., 2018) auf kleinräumigen Ebenen darstellen zu können. All diese Studien basieren auf sogenannten Call Detail Records (CDRs). Dies sind Einzeldaten zur Art und Weise einer mobilen Kommunikation bzw. Aktivität. Sie entstehen bei jeder vom Mobilfunknutzer getätigten aktiven Kommunikation, wie beispielsweise durchs Telefonieren oder mobile Datenverbindungen, und enthalten Informationen über Ort, Dauer, Art der Aktivität, sowie die ID der SIM-Karte des Mobilfunknutzers und seines Gesprächspartners.[1]

Um diese Informationen nutzbar zu machen, wurden verschiedene Methoden zur geografischen Lokalisierung und Verteilung der einzelnen mobilen Aktivitäten verwendet und analysiert. Bedingt wird die Nutzung verschiedener Methoden durch Zusatzinformationen über den Standort der Antenne, deren Frequenz, Höhe, Leistung und Strahlungsrichtung. Je mehr Informationen vorliegen, desto genauer kann die Position der einzelnen mobilen Aktivität geschätzt werden. Aus datenschutzrechtlichen Gründen erhält das Statistische Bundesamt (Destatis) keine Einzeldaten aus dem deutschen Mobilfunknetz, und daher ist eine Verwendung dieser Verfahren in Deutschland nicht direkt möglich. Es werden ausschließlich Aggregatdaten vom Mobilfunkunternehmen zur Verfügung gestellt, welche Informationen über die Anzahl mobiler Aktivitäten für eine bestimmte geografische Einheit und Zeitintervall beinhalten. Um diese Aggregate umfangreich nutzen zu können, wird im Folgenden eine Umverteilungsmethode basierend auf einer nicht-parametrischen Kerndichteschätzung unter Messfehlern (Groß et al., 2017) verwendet. Somit können die Destatis zur Verfügung stehenden Mobilfunkdaten mit den offiziellen Bevölkerungszahlen des Zensus 2011 auf verschiedenen geografischen Ebenen verglichen werden.

Der Aufbau des Artikels stellt sich wie folgt dar. In Abschn. 1.2 werden zunächst die Mobilfunkdaten und erste deskriptive Analysen vorgestellt. Der Kerndichteschätzer unter Messfehlern zur Umverteilung der Bevölkerungszahlen wird in Abschn. 1.3 „Methodik“ erläutert. Anschließend wird der vorgestellte Algorithmus zur Bestimmung der kleinräumigen Bevölkerungszahlen (Kreise und Gemeinden) basierend auf Mobilfunkdaten in Abschn. 1.4 angewendet. Abschn. 1.5 schließt mit einer Zusammenfassung und Ausblick.

## 1.2 Beschreibung der Mobilfunkdaten

Zur Erforschung des Themas „Mobilfunkdaten“ für die amtliche Statistik ist Destatis im September 2017 eine Kooperation mit T-Systems International GmbH und Motionlogic GmbH (beide

[1] In diesem Kapitel sei nachträglich darauf hingewiesen, dass aus Gründen der besseren Lesbarkeit im Text verallgemeinernd das generische Maskulinum für Begriffe wie „Nutzerinnen und Nutzer“ oder „Kundin und Kunde“ verwendet wird.

100% Tochterunternehmen der Deutschen Telekom AG) eingegangen. Die Deutsche Telekom teilt sich den deutschen Mobilfunkmarkt mit Vodafone und Telefónica mit jeweils einem Drittel Marktanteil. Die Konzepte für die geplanten Machbarkeitsstudien wurden gemeinsam mit der Bundesnetzagentur, der Bundesbeauftragten für den Datenschutz und die Informationsfreiheit und in Kooperation mit T-Systems abgestimmt. Mittel- bzw. langfristiges Ziel ist es, die Tages- und Wohnbevölkerung mit Hilfe der Mobilfunkdaten bundesweit valide abbilden und schätzen zu können. Daher befassen sich die ersten Analysen mit der Frage, ob und inwieweit Mobilfunkdaten dazu genutzt werden können, die Bevölkerung valide abzubilden. Zur Überprüfung der Repräsentativität dieser Daten werden die Bevölkerungszahlen des Zensus 2011 als Vergleichsmaßstab herangezogen. Aufgrund datenschutzrechtlicher Regelungen erhält Destatis nur anonymisierte aggregierte Mobilfunkaktivitäten von T-Systems.

Der zur Verfügung stehende Datensatz enthält mobile Aktivitäten von Telekom-Kunden für Nordrhein-Westfalen (NRW) für eine statistische Woche aus ausgewählten Tagen aus den Monaten April, Mai und September des Jahres 2017 in einem 24-Stunden-Zeitraum.[2] Unter einer mobilen Aktivität wird ein Signal mit vordefinierter Aufenthaltsdauer an einem Ort ohne Bewegung verstanden, wobei alle Signaldaten ausgewertet werden, d.h. Telefonate, SMS und Datenverbindungen. Im Gegensatz zu CDRs entstehen Signaldaten automatisch, in regelmäßigen Abständen und registrieren lediglich die Ortsangabe des Funkmastes, mit dem ein mobiles Endgerät zu einem bestimmten Zeitpunkt verbunden ist.

Bei jeder Aktivität eines mobilen Endgerätes, wie beispielsweise das Verfassen einer SMS, verbindet sich dieses über Funkwellen mit der nächstgelegenen Basisstation, welche nur ein bestimmtes Gebiet versorgen, worunter eine Funkzelle verstanden wird. Die Funkzellen bilden zusammen ein mobiles Netzwerk, wobei die Größe der Funkzellen von der erwarteten Anzahl der Nutzer abhängig ist. Somit werden ländliche Gebiete von wenigen großen Funkzellen abgedeckt und städtische Gebiete von mehreren kleinen. Für die Übertragung der Daten zwischen Basisstation und dem mobilen Endgerät werden verschiedene Frequenzen verwendet. Diese finden sich in den Mobilfunkstandards wieder, da jedem Standard (2G bis künftig 5G) ein bestimmter Frequenzbereich zugewiesen wird und sich ihre Nutzungsmöglichkeiten wiederum durch unterschiedliche Reichweiten charakterisieren (Krzossa, 2019). In ländlichen Gebieten werden bspw. eher niedrige Frequenzbereiche (wie zum Beispiel 2G) genutzt, da sie aufgrund der größeren Reichweiten größere Gebiete versorgen können. Mit steigender Frequenz sinkt jedoch die Reichweite, weswegen höhere Frequenzen in den städtischen Gebieten ihren Einsatz finden (Krzossa, 2019).

Die Mobilfunkaktivitäten, welche Destatis zur Verfügung stehen, enthalten die durchschnittlichen Aktivitäten ausgewählter Wochentage und liegen dabei in fünf Tagestypen vor, wobei die

[2] Ausgeschlossen werden hierbei Ferien- und Feiertage.

Tage von Dienstag bis Donnerstag zusammengefasst werden. Des Weiteren enthalten die Mobilfunkdaten unter anderem Informationen über die soziodemografischen Charakteristiken der Mobilfunknutzer, wie die Altersgruppe, das Geschlecht und die Nationalität der SIM-Karte.[3] Aufgrund datenschutzrechtlicher Regelungen wurden die Mobilfunkaktivitäten anonymisiert[4] und aggregiert, wobei erst Wertangaben ab einer Mindestzahl von 30 Aktivitäten pro Gitterzelle[5], im Folgenden auch Grid genannt, an Destatis übermittelt wurden. Die Gitterzellen sind INSPIRE-konform und sind deckungsgleich zu den Zensus-Gitterzellen des Zensusatlas 2011.[6]

Die Anzahl der Mobilfunkaktivitäten hängt von der Lage und Anzahl der Funkmasten in den verschiedenen Gitterzellen ab. Wie bereits erläutert unterscheiden sich je nach Lage der Funkmasten (ländlich oder städtisch) ihre Frequenzen und führen mitunter zu ihrer ungleichmäßigen Verteilung in den verschiedenen Regionen. Demzufolge können in einer vorliegenden Geometrie 5 bis 20 Funkmasten enthalten sein. Infolgedessen werden einige Geometrien zusammenfasst, um die Mindestzahl von 30 Aktivitäten pro Gitterzelle zu gewährleisten. Da die Anzahl mobiler Aktivitäten durch die Verweildauer mobiler Endgeräte bedingt wird, werden je nach Länge der Verweildauer entsprechend lange mobile Aktivitäten gezählt und in den Datensatz einbezogen und kurze mobile Aktivitäten demzufolge außer Acht gelassen. Unter einer Verweildauer wird dabei eine Aufenthaltsdauer eines mobilen Endgerätes an einem Ort bzw. in einer Gitterzelle ohne Bewegung verstanden. Im vorliegenden Datensatz beträgt die Verweildauer zwei Stunden, um kurze mobile Aktivitäten, unter anderem hervorgerufen durch schnelle Wechsel zwischen den Gitterzellen, herauszufiltern und so eine möglichst unverfälschte Darstellung der Wohnbevölkerung mittels Mobilfunkdaten zu gewährleisten.

Ziel der ersten Analysen ist es, die Wohn-, Tages- bzw. Arbeitsbevölkerung valide abzubilden. Hierzu werden im Folgenden die mobilen Aktivitäten aller deutschen Mobilfunknutzer der Deutschen Telekom in NRW für eine statistische Woche betrachtet. Abb. 1.1 stellt die Anzahl der Aktivitäten nach den zur Verfügung stehenden Wochentagen und dem 24-Stunden-Auswertungszeitraum dar. Die Grafik visualisiert die Veränderungen der Aktivitäten im Tagesverlauf und lässt eine Unterscheidung der Aktivitäten durch eine mögliche Tages- und Wohnbevölkerung zu. Es ist ersichtlich, dass weniger mobile Aktivitäten in der Tagesmitte registriert werden und deutlich mehr Aktivitäten in den Morgen- und Abendstunden zu verzeichnen sind.

[3] Hier sind sowohl Vertrags-, Prepaid-, Congstar- wie auch Businesskunden enthalten. Allerdings liegen nur für die Vertragskunden die charakteristischen Merkmale vor.

[4] Die Telekom AG anonymisiert die Daten in einem mit der Bundesbeauftragten für den Datenschutz und die Informationsfreiheit (BfDI) abgestimmten Verfahren.

[5] Unter einer Gitterzelle versteht man eine geografische Einheit in Form eines Quadrats mit variierenden oder einheitlichen Gitterweiten mit Zell- und Raumbezug. Diese sind unabhängig von nationalen Verwaltungsgrenzen und bilden eine sachbezogene Gebietsabgrenzung. Mehrere Gitterzellen ergeben zusammen ein Raster, worunter ein flächendeckendes Bezugssystem verstanden wird.

[6] Näheres zum Zensusatlas vgl. hierzu: `https://atlas.zensus2011.de/`.

In den Abendstunden beträgt die Summe der Aktivitäten in NRW ca. 10,5 Millionen bei einer Einwohnerzahl von ca. 17,5 Millionen. Dies lässt auf Veränderungen der Aktivitäten durch die Arbeitsbevölkerung schließen. Alle Wochentage weisen zudem einen ähnlichen Kurvenverlauf auf und lassen in diesem Sinne keine weiteren Aussagen zu möglichen Unterschieden zwischen den Werktagen und des Wochenendes durch eine dynamische Bevölkerung zu.

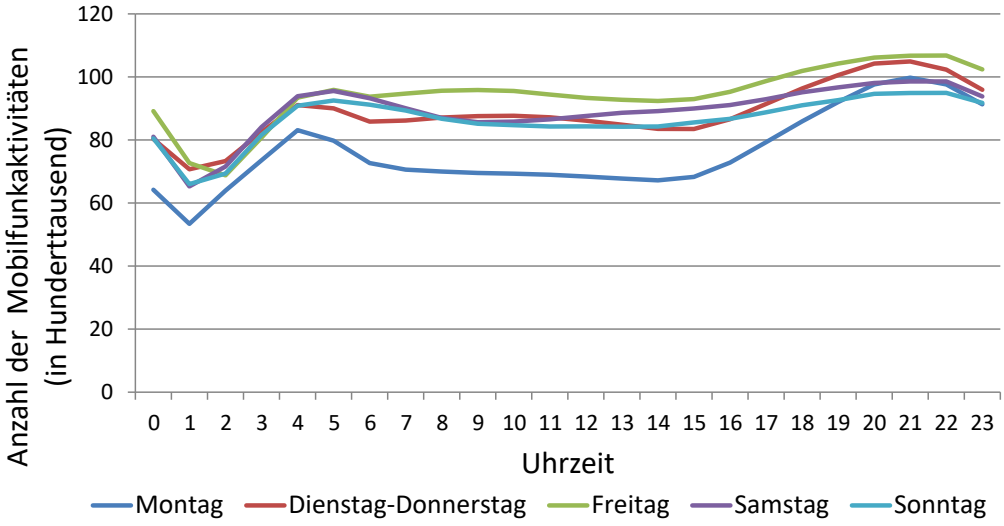



Abbildung 1.1: Anzahl mobiler Aktivitäten im Tages- und Wochenverlauf, eigene Darstellung.

Für einen ersten Zusammenhang mit dem Zensus 2011 wurde weiterhin die Korrelation zwischen den Mobilfunkaktivitäten und den Bevölkerungszahlen des Zensus 2011 auf Basis der Gitterzellen nach Tagestypen und Uhrzeit für NRW ermittelt, wie nachfolgend in Abb. 1.2 dargestellt. Die Koeffizienten weisen insgesamt eine hohe Korrelation von 0.8 zwischen mobilen Aktivitäten und Bevölkerungszahlen für Samstag und Sonntag im kompletten Tagesverlauf auf. Wochentags bzw. Werktags sinkt die Korrelation in einem Zeitraum von 5 Uhr früh bis 16 Uhr auf unter 0.7, was auf stärkere Unterschiede zwischen Wohnbevölkerung basierend auf dem Zensus 2011 und dem Standort der Mobilfunkaktivitäten zum angegebenen Zeitraum hindeutet. Daraus kann geschlossen werden, dass sich die Mobilfunknutzer in diesem Zeitraum mit höherer Wahrscheinlichkeit nicht an ihrem Wohn- sondern an ihrem Arbeitsort o.ä. aufhalten. Die Korrelationen zeigen, dass mittels Mobilfunkdaten eine Unterscheidung zwischen Wohn- und Tagesbevölkerung möglich ist. Die hohe Korrelation in den Abendstunden deutet auf eine mögliche Nutzung von Mobilfunkaktivitäten für die Darstellung der Wohnbevölkerung hin. Die niedrige Korrelation in der Tagesmitte kann hingegen durch die Veränderung durch die Tages- bzw. Arbeitsbevölkerung in den Mobilfunkdaten erklärt werden.

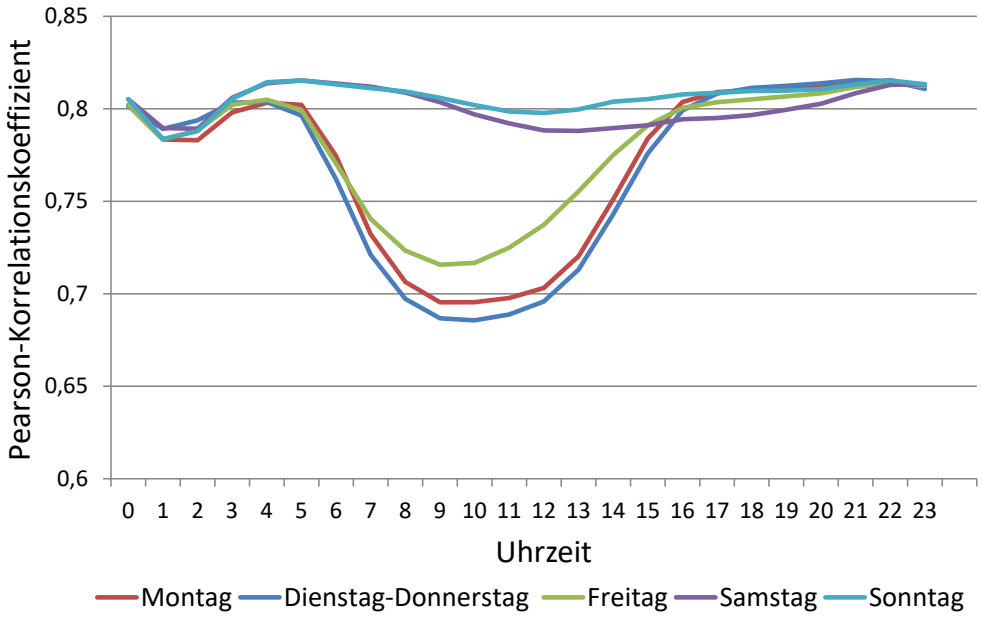



Abbildung 1.2: Pearson-Korrelationskoeffizienten bezüglich der Zensuswerte und der Mobilfunkaktivitäten im Tages- und Wochenverlauf, eigene Berechnung.

Somit können die Wohn- bzw. Arbeitsstandorte der Bevölkerung annähernd lokalisiert werden, welche durch Gebiete mit über- und unterdurchschnittlich aktiven SIM-Karten zu bestimmten Zeitpunkten im Tagesverlauf gekennzeichnet sind. Zur Darstellung der Wohnbevölkerung wurde aufgrund der hohen Korrelation und Plausibilität ein Zeitraum von 20 bis 23 Uhr ausgewählt. Hierbei wird angenommen, dass sich die Mobilfunknutzer in ihrem Wohnort befinden und ihre mobilen Endgeräte gleichzeitig noch mit höherer Wahrscheinlichkeit aktiv sind. In der gängigen Literatur wird häufig ein Zeitraum von 18 Uhr abends bis 6 Uhr früh gewählt. Die Analysen haben aber keine signifikanten Unterschiede zu dem hier gewählten Zeitraum gezeigt, weshalb der vier-Stunden-Zeitraum präferiert wird. Für die weiterführenden Analysen zur Darstellung der Wohnbevölkerung mittels Mobilfunkdaten wird weiterhin eine Umrechnung und Kalibrierung der Mobilfunkaktivitäten durchgeführt, welche im Folgenden näher erläutert wird. Dies ermöglicht zudem einen direkten Vergleich mit den Bevölkerungszahlen des Zensus 2011.

## 1.3 Methodik: Eine nicht-parametrische Kerndichteschätzung

In diesem Abschnitt werden kurz die theoretischen Grundlagen einer nicht-parametrischen Kerndichteschätzung beschrieben. Ein weit verbreitetes und simples Verfahren zur Bestimmung

einer Dichte stellt das Histogramm dar. Statistisch etwas fortgeschrittener ist die Anwendung von Kerndichteschätzern. Dabei handelt es sich um eine Annäherung der Dichtefunktion $f(x)$ einer stetigen mehrdimensionalen Zufallsvariablen $X$ aus den Beobachtungen der Stichprobe $X_i$ mit $i = 1, \ldots, n$. Die Methodik gehört – wie das Histogramm – zu den nicht-parametrischen Verfahren, d.h. es gibt keine direkten Annahmen über die Gestalt der Dichtefunktion und es handelt sich um ein datengetriebenes Verfahren.

Im Folgenden betrachten wir nur den zwei-dimensionalen Fall mit $X_i = (X_{i1}, X_{i2})$, wobei $X_{i1}$ und $X_{i2}$ Längen- und Breitengrade der Koordinaten der Beobachtungen darstellen. Ein bivariater Kerndichteschätzer an einer Stelle $x = (x_1, x_2)$ ist gegeben durch

$$\hat{f}_h(x) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} K(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2})$$

mit vorab definierten Bandbreiten $h_1$ und $h_2$ für die beiden Dimensionen und einer Kernfunktion $K$, etwa einem bivariaten Gaußkern. Für weitere Details bezüglich der Wahl der Bandbreite und der Kernfunktion wird auf Wand und Jones (1994) verwiesen.

Wie bereits erwähnt stehen die Mobilfunkaktivitäten nur in aggregierter Form für unterschiedlich große Gitterzellen zur Verfügung. So können den einzelnen Aktivitäten keine exakten Längen- und Breitengrade (Koordinaten) zugewiesen werden. Bei einer Kerndichteschätzung müssen die aggregierten Aktivitäten zunächst auf einen Punkt innerhalb der Gitterzelle – etwa dem Mittelpunkt – projiziert werden. Dieser Prozess kann als eine Art zweidimensionales Runden (Messfehler) aufgefasst werden. Bei einer Verwendung eines klassischen Kerndichteschätzers würden an den Mittelpunkten Häufungen entstehen, die durch den Rundungsprozess bzw. durch die fehlenden geografischen Koordinaten der Mobilfunkaktivitäten zu begründen sind. Verzerrte Kerndichteschätzungen sind die direkte Folge. Daher wird eine nicht-parametrische Kerndichteschätzung nach Groß et al. (2017) durchgeführt. Bei der Methode handelt es sich um einen Stochastic-Expectation-Maximization Algorithmus (SEM; Celeux und Diebolt, 1985). In der Anwendung des Algorithmus auf die aggregierten Mobilfunkaktivitäten wird durch das wiederholte Anwenden der Kerndichteschätzung und anschließender Stichprobenziehung ein Wert der Dichte ermittelt – eine Art Simulation von Geokoordinaten. Da bei diesem iterativen Ansatz die Modalwerte der Dichte nicht so stark an die Häufungspunkte/Mittelwerte gekoppelt sind, kann die Methodik als eine Art Kerndichteschätzung unter Einbeziehung von Messfehlern/Rundungen betrachtet werden. Der Algorithmus lässt sich wie folgt beschreiben (Groß et al., 2017):

1. Berechnung einer naiven Kerndichteschätzung $\hat{f}_h(x)$ von $f(x)$ basierend auf den aggregierten Mobilfunkdaten aus den Gitterzellen.

2. Ziehung von exakten Koordinaten der Mobilfunkaktivitäten aus der geschätzten Dichte

$\hat{f}_h(x)$, wobei die Anzahl der Mobilfunkaktivitäten je Gitterzelle erhalten bleibt – d.h. Mobilfunkaktivitäten können nicht in andere Gitterzellen „wechseln“.

3. Berechnung der optimalen Bandbreiten $h_1$ und $h_2$ nach Wand und Jones (1994) basierend auf den Mobilfunkaktivitäten mit exakten Koordinaten.

4. Berechnung der naiven Kerndichteschätzung $\hat{f}_h(x)$ von $f(x)$ basierend auf den Mobilfunkaktivitäten mit exakten Koordinaten und den optimalen Bandbreiten $h_1$ und $h_2$.

5. Wiederholung der Schritte 2-4 insgesamt $B$ (Burn-in) plus $N$ mal.

6. Berechnung einer finalen Dichteschätzung durch Mittelung der $N$ geschätzten Kerndichteschätzungen $\hat{f}_h(x)$.

Für die Berechnung des Algorithmus wurde das R-Paket **Kernelheaping** (Groß, 2018) verwendet. Die Software ermöglicht mit dem Befehl `dshapebivr` eine Schätzung der Kerndichte der mobilen Aktivitäten und erstellt eine kontinuierliche Karte, anhand der die Dichte (der Mobilfunkaktivitäten) der Regionen in NRW identifiziert werden können. Im zweiten Schritt erfolgt mithilfe des Befehls `toOtherShape` eine Umrechnung der resultierenden Dichten auf die gewünschte Geometrie (etwa Kreise oder Gemeinden).

## 1.4 Anwendung: Bestimmung von Bevölkerungszahlen in NRW

In diesem Abschnitt werden die aggregierten Mobilfunkaktivitäten mit Hilfe der nichtparametrischen Kerndichteschätzung aus Abschn. 1.3 in die interessierenden Geometrien umgewandelt, um Vergleiche mit den Bevölkerungszahlen in NRW zu ermöglichen.

Ein erster Eindruck der Verteilung der Mobilfunkaktivitäten auf Ebene der Gitterzellen ist in Abb. 1.3 (links) zu sehen. Da die Gitterzellen stark unterschiedliche Größen aufweisen, sind visuelle Vergleiche mit anderen Datenquellen, etwa mit den Bevölkerungszahlen des Zensus 2011, nur schwer durchzuführen. Wie in Abb. 1.3 (links) deutlich zu erkennen ist, weisen die städtischen Gebiete teilweise sehr kleine Gitterzellen auf, so dass die Anzahl mobiler Aktivitäten nicht mehr eindeutig identifiziert werden kann. Durch die Anwendung der in Abschn. 1.3 vorgestellten Kerndichteschätzung können die mobilen Aktivitäten von der geografischen Rasterstruktur (Gitterzellen) gelöst werden und die visuellen Unschärfe aufgehoben werden. Mithilfe dieses Verfahrens wird zunächst die Kerndichte der mobilen Aktivitäten basierend auf den zugrunde liegenden Gitterzellen geschätzt und anschließend eine einheitliche Karte mit der Dichteverteilung der mobilen Aktivitäten, wie in Abb. 1.3 (rechts) dargestellt, erstellt. Anhand dessen kann die Dichteverteilung der vorliegenden mobilen Aktivitäten räumlich hervorgehoben und somit

auch die städtischen Gebiete mit einer hohen Dichte sichtbar gemacht werden. Im vorliegenden Fall wird die Wahrscheinlichkeitsverteilung mobiler Aktivitäten für das Bundesland NRW geschätzt.

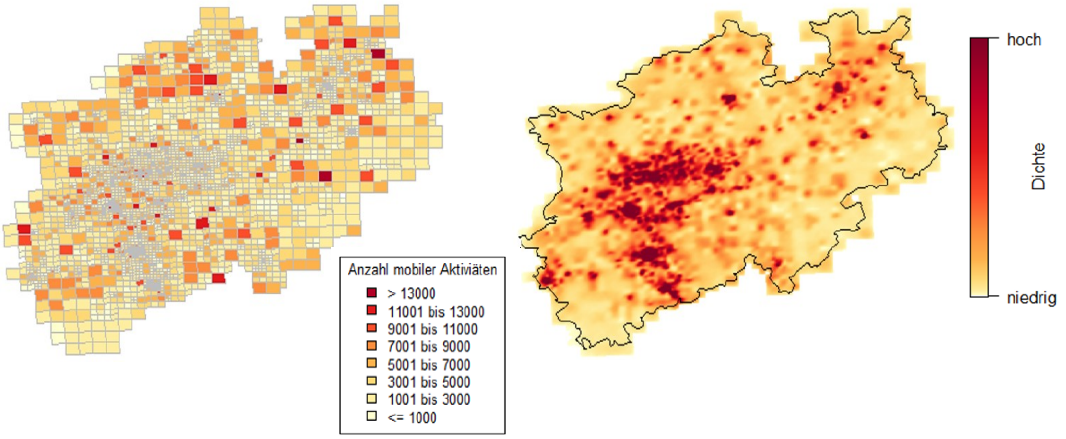



Abbildung 1.3: Mobilfunkaktivitäten: Gitterzellen (**links**) und Kerndichteschätzung (**rechts**), eigene Berechnung.

Da die Wohnbevölkerung mittels Mobilfunkdaten abgebildet werden soll, beziehen sich alle bisherigen und folgenden Ergebnisse auf die Mobilfunkaktivitäten an einem statistischen Sonntagabend. Dies wurde auf Basis der Korrelation (Abb. 1.1) ausgewählt. Hierzu werden die Schätzergebnisse der Kerndichteschätzung des gewählten Zeitraumes anhand der geografischen Koordinaten auf Kreis- und Gemeindeebene umgerechnet. Da in Deutschland eine hierarchische Struktur von Bundesland, Regierungsbezirk, Kreis bis hin zur Gemeinde vorliegt, wurden die Schätzungen auf diese Ebenen für NRW umgerechnet.[7] Genauer unterteilen sich diese in 5 Regierungsbezirke, 22 kreisfreie Städte, 31 Kreise und darunter 374 Gemeinden, wobei eine Umrechnung der Kerndichten auf Regierungsbezirksebene aufgrund ihrer Größe keinen direkten Mehrwert liefert.

[7] Die Gitterzellen lassen sich nicht eindeutig in die hierarchische Struktur einordnen.

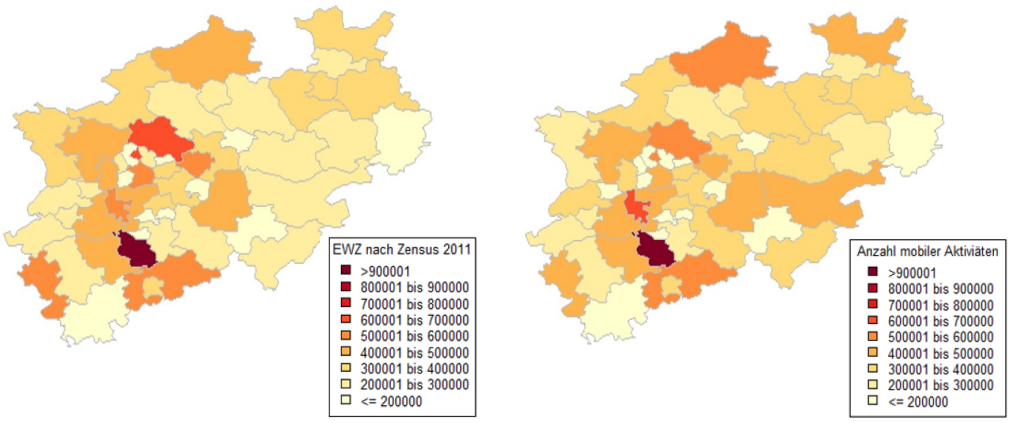



Abbildung 1.4: Zensus 2011 (**links**) und basierend auf Mobilfunkaktivitäten (**rechts**) auf Kreisebene, eigene Berechnung.

Tabelle 1.1: Absoluter und relativer Bias zwischen Mobilfunkaktivitäten und Zensus 2011 auf Kreisebene, eigene Berechnung.

| Bias | Min. | 1st Qu. | Median | 3rd Qu. | Max |
|---|---|---|---|---|---|
| Absoluter | -102250 | -35757 | -4928 | 29370 | 163033 |
| Relativer [%] | -36.3 | -10.3 | -2.2 | 9.8 | 61.2 |

Da die Anzahl mobiler Aktivitäten deutlich niedriger ist als die der Einwohnerzahl vom Zensus 2011, werden die mobilen Aktivitäten anhand der Einwohnerzahlen in NRW kalibriert. Dies ist notwendig um absolute Vergleiche mit dem Zensus vornehmen zu können. Die Kalibrierung setzt sich aus einem Faktor basierend auf dem Verhältnis der Gesamtanzahl der Bevölkerung vom Zensus zur Gesamtanzahl der Mobilfunkaktivitäten zusammen. Ein direkter Vergleich ohne weitere Hochrechnung würde lediglich zum Ergebnis führen, dass die mobilen Aktivitäten niedriger sind als die offiziellen Bevölkerungszahlen. Durch den angewandten Korrekturfaktor werden die Aktivitäten von ca. 9 Mio. auf die Bevölkerung von NRW mit ca. 17,5 Mio. Einwohner hochgerechnet, wie in den folgenden Abb. 1.4 und 1.5 dargestellt.

Zunächst ist in beiden Abbildungen kein offensichtlicher visueller Unterschied zwischen den Mobilfunkaktivitäten und den Bevölkerungszahlen des Zensus 2011 sichtbar. Dies deutet daraufhin, dass die Verteilung der mobilen Aktivitäten denjenigen des Zensus 2011 ähnelt. Die Ermittlung der Differenzen bzw. des Bias in Tab. 1.1 und 1.2 liefert zusätzlich quantitative Informationen darüber, ob und inwieweit die Destatis zur Verfügung stehenden mobilen Aktivitäten mit den Bevölkerungszahlen aus dem Jahr 2011 übereinstimmen. Ein positiver Bias bzw. ei-

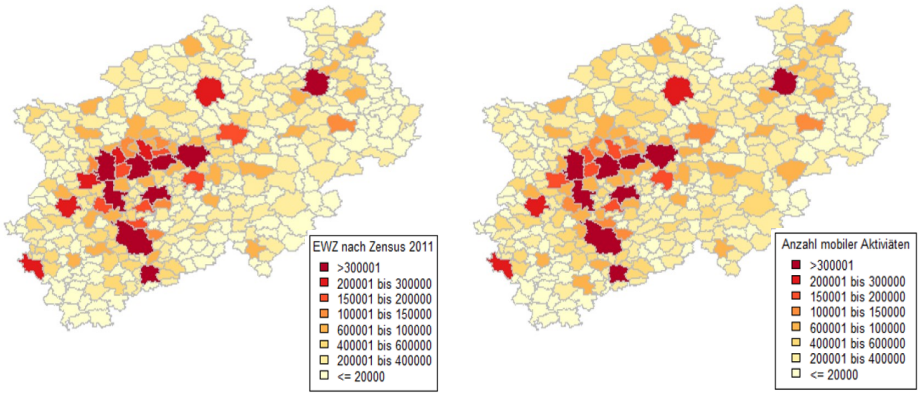



Abbildung 1.5: Zensus 2011 (**links**) und basierend auf Mobilfunkaktivitäten (**rechts**) auf Gemeindeebene, eigene Berechnung.

Tabelle 1.2: Absoluter und relativer Bias zwischen Mobilfunkaktivitäten und Zensus 2011 auf Gemeindeebene, eigene Berechnung.

| Bias | Min. | 1st Qu. | Median | 3rd Qu. | Max |
|---|---|---|---|---|---|
| Absoluter | -107052 | -2182.5 | 1840 | 5514.5 | 60079 |
| Relativer [%] | -73.6 | -7.1 | 9.8 | 31.4 | 157.4 |

ne positive Verzerrung deutet auf eine Überschätzung der Bevölkerung mittels Mobilfunkdaten hin und eine negative Verzerrung bedeutet eine Unterschätzung der Bevölkerung mittels Mobilfunkdaten. Im Durchschnitt werden die Bevölkerungszahlen auf Kreisebene (Tab. 1.1) deutlich besser geschätzt als auf Gemeindeebene (Tab. 1.2).

Als eine erste gute Schätzung der Bevölkerung mittels Mobilfunkdaten können Schätzergebnisse mit einem Bias von +-10 Prozent gesehen werden. Alle anderen Ergebnisse sind zu stark verzerrt und unter- oder überschätzen die tatsächliche Anzahl der Bevölkerung deutlich. Im vorliegenden Beispiel werden in 57 Prozent der Kreise die Bevölkerung mit Mobilfunkaktivitäten am Sonntagabend in einem Zeitraum von 20 bis 23 Uhr zufriedenstellend geschätzt und auf Gemeindeebene dagegen nur 33 Prozent. Zudem sind die Über- und Unterschätzungen auf Kreisebene relativ ausgeglichen. Dagegen überwiegt auf Gemeindeebene die Überschätzung der Bevölkerung mittels Mobilfunkdaten.

Die Deutsche Telekom ist vor allem in den ländlichen Gebieten stark vertreten und besitzt in den städtischen Gebieten vergleichsweise weniger Marktanteile, woraufhin auch Abb. 1.6 hindeutet. Sie zeigt die geografische Differenz bzw. den absoluten Bias aus Tab. 1.1 und 1.2

zwischen den Mobilfunkaktivitäten und den Einwohnerzahlen vom Zensus 2011 auf Kreis- und Gemeindeebene. Sind die Gebiete rot hinterlegt bedeutet dies, dass dort mehr (kalibrierte) mobile Aktivitäten registriert wurden als Einwohner gemeldet sind. Blaue Flächen deuten hingegen auf Regionen hin, in denen weniger (kalibrierte) mobile Aktivitäten registriert wurden als Einwohner gemeldet sind.

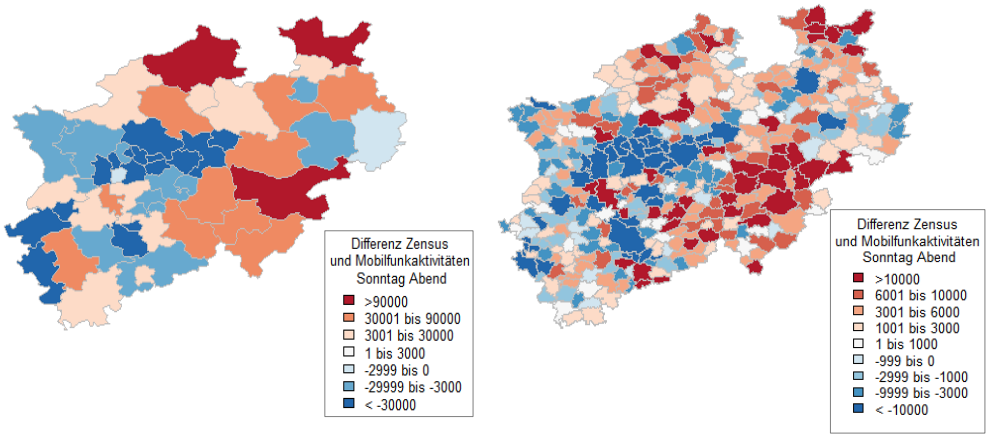



Abbildung 1.6: Differenz zwischen Mobilfunkaktivitäten an einem statistischen Sonntag in einem Zeitraum von 20 bis 23 Uhr und den Einwohnerzahlen nach dem Zensus 2011 auf Kreisebene (**links**) und Gemeindeebene (**rechts**), eigene Berechnung.

Im Vergleich zum Zensus 2011 sind die mobilen Aktivitäten in Abb. 1.6 auf Kreis- und Gemeindeebene in den ländlicheren Gebieten deutlich höher. Dagegen sind die mobilen Aktivitäten in den urbanen Gebieten deutlich niedriger als die angegebenen Bevölkerungszahlen vom Zensus 2011, was unter anderem durch die Marktanteile des Mobilfunkanbieters verursacht werden könnte.

## 1.5 Zusammenfassung und Ausblick

Die Ergebnisse zeigen vom Grundsatz her, dass die Bevölkerung mit den vorliegenden Mobilfunkdaten teilweise gut abgebildet werden könnte. Beobachtbare Unterschiede in der Bevölkerungsdarstellung mittels Mobilfunkdaten und den Zensuswerten können teilweise durch die zeitliche Differenz zwischen den Mobilfunkdaten aus dem Jahr 2017 und den Zensusdaten aus dem Jahr 2011, aber auch durch das seitens des Datenproviders angewandte Hochrechnungsverfahren hervorgerufen werden. Die Hochrechnung basiert auf den regionalen Marktanteilen der Deutschen Telekom am gesamtdeutschen Mobilfunkmarkt. Die Gewichtung der Mobilfunkak-

tivitäten basiert auf dem Standort bzw. der Postleitzahl des allerersten Signals eines mobilen Endgerätes zu Beginn seiner Aktivitätenkette. Das bedeutet, dass die Hochrechnung aller gezählten Aktivitäten eines mobilen Endgerätes im Tagesverlauf nur durch den Marktanteil der Postleitzahl der ersten gezählten Aktivität bedingt wird. Zudem erfolgt die Hochrechnung nur auf die Gesamtanzahl aller Mobilfunknutzer. Aktuell besitzen rund 80 Prozent der deutschen Bevölkerung ein Mobiltelefon.[8] Entsprechend werden 20 Prozent der Bevölkerung nicht in der Hochrechnung berücksichtigt.

Durch den Einsatz einer Kerndichteschätzung kann zudem unabhängig von der zugrundeliegenden geografischen Einheit eine Umverteilung auf weitere zu interessierende Geometrien erfolgen. Durch dieses Verfahren entsteht allerdings eine zusätzliche Unsicherheit in den Mobilfunkaktivitäten, die sich vor allem dann stärker auswirkt, je kleiner die Geometrien werden. Zudem wurden nur Mobilfunkaktivitäten eines Anbieters in Deutschland analysiert. Die dadurch entstehenden Verzerrungen werden durch die jeweiligen Marktanteile bedingt und sind in der räumlichen Verteilung der Mobilfunkaktivitäten sichtbar. Auch in den soziodemografischen Merkmalen spiegelt sich die Kundenstruktur des Mobilfunkanbieters wieder, welche aufgrund der starken Selektivität nicht näher vorgestellt wurden. Da die bundesweite Repräsentativität der Daten essenziell ist, müssen weitere Schritte unternommen werden, um künftig möglichst Daten aller Mobilfunkanbieter in Deutschland zu erhalten.

Neben der Bevölkerungsdarstellung mittels Mobilfunkdaten sollte sich eine weitere Forschungsfrage damit auseinandersetzen, inwieweit eine Modellierung von zeitlichen Bevölkerungsfortschreibungen mit Hilfe routinemäßig erhobener Mobilfunkdaten erfolgen kann bzw. inwieweit Mobilfunkdaten als zeitliche Fortschreibung für intra-zensus Perioden genutzt werden können. Dies kann durch eine mögliche Kalibration der Mobilfunkdaten auf den künftigen Zensus 2021 erfolgen. Anhand der damit errechneten Korrekturfaktoren kann anschließend mittels aktuellster Mobilfunkdaten ein regelmäßiges Update der Bevölkerung auch nach soziodemografischen Merkmalen erfolgen und so als eine Ergänzung der jährlichen Bevölkerungsfortschreibung gesehen werden.

Neben der hier beschriebenen direkten Nutzung von Mobilfunkdaten können diese auch als Zusatz- oder Hilfsinformationen für andere Statistiken oder Indikatoren genutzt werden. Beispielsweise wurden im Rahmen des ESSnet Projektes „City Data from LFS and Big Data“ Indikatoren der Arbeitskräfteerhebung mittels Small-Area-Verfahren unter Verwendung von Mobilfunkdaten auf kleinräumige Ebenen geschätzt (European Commission, 2019). Durch dieses Verfahren konnten für Gebiete ohne Beobachtungen verlässliche Schätzer ausgegeben werden. Gleichzeitig konnte hierdurch die Unsicherheit bei der Schätzung der Erwerbslosenquote auf

[8] Vgl. `https://de.statista.com/statistik/daten/studie/585883/umfrage/anteil-der-smartphone-nutzer-in-deutschland/`, Zugriff am 08.06.2019.

Ebene von sogenannten „funktionalen städtischen Gebieten“ – welche auf ausgewählten Gemeinden und Kreisen basieren – verringert werden.

Um diese Verfahren allerdings dauerhaft umsetzen zu können, wird ein uneingeschränkter Datenzugang benötigt. Neue digitale Daten werden vorwiegend in privaten Unternehmen generiert und gehalten. Um sie langfristig in die amtliche Statistikproduktion integrieren zu können, müssen neue Rechtsgrundlagen geschaffen werden, um den Zugang zu privat gehaltenen Daten dauerhaft zu sichern.

# Kapitel 2

# Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten

## 2.1 Motivation

Aktuelle und valide Einwohnerzahlen sind für politische Entscheidungsfindungen unerlässlich und bspw. für den Finanzausgleich zwischen Bund und Ländern, für die Einteilung des Bundesgebietes in Wahlkreise und zur Bestimmung der Größe der Wahlbezirke oder für allgemeine Planungsaufgaben von Bedeutung (Statistisches Bundesamt, 2021b). Auch bei Fragestellungen zur Verkehrsnutzung oder zur räumlichen Gesundheitsversorgung können diese von hoher Relevanz sein.

Der Bedarf an möglichst kleinräumigen Bevölkerungszahlen wird aktuell vor allem im Rahmen von georeferenzierten Auswertungen im Bereich des Gesundheitssystems, zum Beispiel hinsichtlich des Zugangs zu Gesundheitseinrichtungen, gesehen. Bspw. untersuchte Information und Technik Nordrhein-Westfalen (IT.NRW, 2018) die Verteilung von Arztpraxen in Nordrhein-Westfalen anhand von georeferenzierten Einwohnerzahlen aus dem Zensus 2011. Eine weitere Verwendungsmöglichkeit wird in diesem Zusammenhang im Einbezug kleinräumiger Bevölkerungszahlen beim Krankenhaus-Atlas[1] gesehen, der deutschlandweit die Krankenhäuser unter Angabe ihrer jeweiligen Fachabteilungen, ihrer Erreichbarkeit sowie in Abhängigkeit von der

[1] Siehe hierzu den Krankenhausatlas der Statistischen Ämter des Bundes und der Länder: `https://krankenhausatlas.statistikportal.de/`.

kleinräumigen Bevölkerungsdichte basierend auf dem Zensus 2011 interaktiv darstellt. Hierzu ist anzumerken, dass die zeitliche Diskrepanz in der Bevölkerungsdichte zwischen dem Zensus 2011 sowie den zu einem späteren Zeitpunkt hieran anknüpfenden Forschungsfragen eine mögliche Unsicherheit bei der Interpretation der Ergebnisse darstellt, da keine aktuelleren Ergebnisse der jährlichen Bevölkerungsfortschreibung auf kleinräumiger Ebene vorlagen.

Dieses Problem findet sich auch bei Arbeiten zur Raumforschung und bei der Bewertung von Raumentwicklungen. Bei Fina et al. (2019) werden bspw. anhand kleinräumiger Bevölkerungszahlen aus dem Zensus 2011 Analysen hinsichtlich der innerstädtischen Mobilität durchgeführt, um Zusammenhänge zwischen bspw. Mobilitätsarmut und der soziodemografischen Struktur auf stadtregionaler Ebene herzustellen. Sie betonen, dass eine Fortschreibung differenzierter kleinräumiger Bevölkerungsdaten für derartige Analysen aufgrund abnehmender Aktualität der Zensus-Ergebnisse erforderlich sei.

Die Bevölkerungsfortschreibung ermöglicht die Angabe aktueller Einwohnerzahlen auf geografischer Ebene der administrativen Einheiten. Die Einwohnerzahl wird hierbei auf Basis des Zensus 2011 anhand von Angaben der Statistiken zu Geburten und Sterbefällen sowie der Wanderungsstatistik laufend bzw. jährlich fortgeschrieben. Die kleinste administrative Ebene, auf der hierbei veröffentlicht werden kann, ist die Gemeindeebene. Entsprechend ist es nicht möglich Bevölkerungsdaten für nicht-administrative Einheiten zu ermitteln, die sich nicht aus Gemeinden zusammensetzen (Statistisches Bundesamt, 2021b). Hierunter fallen unter anderem georeferenzierte Daten in Form von INSPIRE[2]-konformen Gitterzellen.

Gitterzellen sind gleich große Quadrate, die bei einer flächendeckenden, gleichmäßigen Anordnung bzw. Verteilung, ein Raster bilden. Die sogenannten INSPIRE-konformen Gitterzellen stellen ein europaweit einheitliches geografisches Gitter dar und sind dadurch unabhängig von administrativen Einheiten, wodurch sie zugleich unabhängig von Gebietsstandsänderungen sind (BKG, 2020b). Demzufolge sind INSPIRE-konforme Gitterzellen zeitlich stabil und bleiben dauerhaft vergleichbar, selbst bei Zusammenfassung der Gitterzellen auf einer höheren Ebene. Sie erlauben somit flexible Auswertungen nicht nur für administrative oder statistische Gebiete, sondern auch für frei definierbare Gebiete wie innerstädtische Strukturen oder Stadt-Umland-Übergänge (Wonka et al., 2009; BBSR, 2021; Kirchner et al., 2014). Ein weiterer Vorteil neben der beliebigen räumlichen Zusammenfassung besteht in der vereinfachten Verschneidung mit weiteren Datenquellen auf dieser räumlich neutralen und – in Europa – länderübergreifend standardisierten Struktur. Flächendeckende Bevölkerungsdaten auf Rasterebene ermöglichen demnach eine differenziertere Betrachtung soziodemografischer Entwicklungen in Gemeinden, Ortsteilen etc. Da der Bedarf an kleinräumigeren Bevölkerungszahlen aktuell nicht von der Bevölkerungsfortschreibung gedeckt werden kann, wird ein neuer experimenteller Lösungsansatz ver-

[2]INSPIRE = INfrastructure for SPatial InfoRmation in Europe.

folgt.

Demzufolge gibt es bereits diverse Herangehensweisen bzw. Ansätze, um kleinräumige und aktuelle Bevölkerungszahlen zu erzeugen. Neben mittlerweile gängigen Methoden wie der sogenannten Small Area Estimation, eine kleinräumige Schätzmethode zur Schätzung von kleinräumigen Bevölkerungszahlen wie bereits in Simpson et al. (1996) diskutiert, werden zusätzlich insbesondere Fernerkundungsdaten – vorzugsweise Satellitendaten – zur Herleitung kleinräumiger Bevölkerungsverteilungen verwendet. Hierbei werden Bevölkerungszahlen, insbesondere Bevölkerungsdichten, vorrangig durch Kartierungsmethoden oder kleinräumige Schätzverfahren anhand von Fernerkundungsdaten auf Basis der letzten Volkserhebung auf räumlich feine Einheiten prognostiziert (Stevens et al., 2015).[3] Lloyd et al. (2017) nutzen hierbei im sogenannten WorldPop Programm[4] basierend auf diversen Geodaten einen gewichteten dasymetrischen Ansatz, worunter ein Prozess der räumlichen Umverteilung von interessierenden Größen durch eine flächenhafte Interpolation zu verstehen ist, bei dem anhand eines Random Forest Modells Bevölkerungszahlen kleinräumig geschätzt werden.[5] Schug et al. (2021) kartieren die Bevölkerung in Deutschland unter Verwendung von Gewichtungsschichten, die bspw. von der Gebäudedichte, Gebäudehöhe und den Gebäudetypen aus Satellitendaten (Copernicus Sentinel-1 und Sentinel-2 Daten) hergeleitet werden. Neuere Ansätze wie in Koebe et al. (2022) kombinieren Satellitendaten und das Small Area-Schätzverfahren SPREE[6], eine Methode der strukturerhaltenden Schätzung, die insbesondere für regional und demografisch differenzierte Bevölkerungsfortschreibungen zwischen den Zensen auf kleinräumiger Ebene verwendet wird. Anhand der Hilfsinformationen aus den Satellitendaten werden folglich kleinräumige Bevölkerungszahlen für den Senegal selbstständig fortgeschrieben.

Um die Qualitätsaspekte der amtlichen Statistik möglichst nicht zu tangieren, werden in dieser Arbeit die amtlich fortgeschriebenen Bevölkerungszahlen anhand neuer digitaler Daten kleinräumig umverteilt. Die Qualität der fortgeschriebenen Bevölkerungszahl ist ab der Gemeindeebene aufwärts unangetastet, da sie den Bevölkerungszahlen der Bevölkerungsfortschreibung entsprechen. Mit diesem Verfahren wird die amtliche Statistik unterstützt – jedoch nicht ersetzt – da keine Bevölkerungszahlen fortgeschrieben werden, sondern diese nur anhand einer zusätz-

[3] Ferner vergleichen Leyk et al. (2019) anhand verschiedener dasymetrischer Kartierungsmethoden erzeugte und zugängliche Rasterdatensätze zu Bevölkerungszahlen und -dichten mit den Genauigkeiten und Qualitäten der Ergebnisse und geben Hilfestellung zum beabsichtigten Verwendungszweck.

[4] Siehe hierzu auch: `https://www.worldpop.org/methods/populations`.

[5] Kartografisch dargestellt wird dabei die Bevölkerungsdichte anhand der geschätzten Anzahl der Wohnbevölkerung pro Gitterzelle, wobei die hier verwendeten Gitterzellen nicht INSPIRE-konform sind. Die geschätzte Gesamtbevölkerung der Länder wird so angepasst, dass sie mit den entsprechenden offiziellen Bevölkerungsschätzungen der Vereinten Nationen und nicht mit den amtlich veröffentlichten Einwohnerzahlen – sofern vorhanden – der einzelnen Länder übereinstimmen.

[6] SPREE = Structure preserving estimation.

lichen externen Datenquelle – genauer Mobilfunkdaten – kleinräumig unterhalb der Gemeindeebene verteilt werden.

Dass die Verteilung der Bevölkerung mit den vorliegenden Mobilfunkdaten grundsätzlich gut und zeitnah abgebildet werden kann, zeigen Hadam et al. (2020) bereits in den bisherigen Analysen zur Bevölkerungsdarstellung mit Mobilfunkdaten. Der ausschlaggebende Vorteil der Mobilfunkdaten im Vergleich zu anderen Datenquellen oder Hilfsinformationen besteht hierbei in den starken Zusammenhängen der Mobilfunkdaten mit der Bevölkerung sowie in ihrer zeitlich und räumlich hohen Auflösung. Zudem sind Mobilfunkdaten robust gegenüber administrativen Gebietsstrukturänderungen und können für jede gewünschte räumliche Einheit aufbereitet werden und sind dadurch auch im Zeitverlauf vergleichbar. Im Gegensatz zu anderen Datenquellen können Mobilfunkdaten die tatsächlichen Aufenthaltsorte der Bevölkerung somit valide und zeitnah darstellen.[7]

Douglass et al. (2015) haben den Nutzen von Mobiltelefondaten zur Darstellung hochauflösender Bevölkerungsschätzer bereits erkannt und fokussieren sich darauf, die Bevölkerung in Mailand durch ein Random Forest Modell für den Zeitraum zwischen den Zensen zu schätzen, das auf den bekannten Zensusdaten trainiert wird. Sie verwenden hierzu sogenannte individuelle Mobiltelefondaten oder auch Call Detail Records (CDRs). Deville et al. (2014) zeigen ferner, wie Mobiltelefondaten bzw. CDRs die gängigen Ergebnisse der Volkszählung durch kleinräumige Schätzungen oder auch bei der Messung der Bevölkerungsdynamik ergänzen können. Zudem vergleichen sie die geschätzte Bevölkerungsdichte, die auf Basis von CDRs sowie durch Fernerkundungsdaten hergeleitet wird, anhand der amtlichen Bevölkerungszahlen in Portugal und schlussfolgern, dass die Kombination beider Datenquellen und Methoden eine Verbesserung der räumlichen und zeitlichen Auflösung verspricht.[8]

Der Vorteil dieser CDRs liegt mitunter in der sehr individuellen Angabe von Informationen zu Mobiltelefonnutzenden auf einer hohen räumlichen Auflösung, die im Gegensatz zu den Signaldaten – im Folgenden nur noch als Mobilfunkdaten bezeichnet – jedoch ereignisbasiert sind. Bei den Mobilfunkdaten werden alle erzeugten Signale im entsprechenden Mobilfunknetz vom Netzbetreiber erfasst (Hadam, 2021). Die CDRs sind daher nur verfügbar, wenn der Telefonnutzende bspw. aktiv einen Anruf tätigt oder eine SMS bzw. mobile Daten sendet. Zudem liegen CDRs nur von Vertragskundinnen und -kunden vor, die im Rechnungssystem des Mobilfunkanbieters hinterlegt sind. Um Aussagen über die Bevölkerungszahlen anhand von Daten mobiler

[7] Daher bieten sich Mobilfunkdaten auch als potenzielle Informationsquelle für den Katastrophenschutz an, insbesondere um zu ermitteln, wo sich die Tagesbevölkerung im Zeitverlauf (Tages- und Wochenverlauf) befindet. Der Informationsgehalt ist jedoch abhängig von der erforderlichen räumlichen und zeitlichen Auflösung.

[8] Die bessere räumliche Auflösung findet sich bei Deville et al. (2014) in den Fernerkundungsdaten, deren Aufnahmen jedoch abhängig von den Wetterverhältnissen sind. Die zeitliche Genauigkeit findet sich hierbei in den Mobiltelefondaten.

Endgeräte zu tätigen, bieten sich CDRs aufgrund der offensichtlichen Selektivitäten (Vertrag- vs. Prepaid-Kundin/-Kunde) daher nicht an.

Im Projekt *Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten* wird darauf aufbauend erforscht, ob und inwieweit mit Mobilfunkdaten die vorhandene Bevölkerungsfortschreibung unter Verwendung eines Verteilungsverfahrens kleinräumig, von der Gemeindeebene bundesweit auf INSPIRE-konforme 1x1 km Gitterzellen, verteilt und abgebildet werden kann. Bis die erste amtliche georeferenzierte Bevölkerungszahl auf Basis des Zensus 2022 vorliegt, kann die zeitliche Lücke behelfsweise durch die Nutzung von Mobilfunkdaten geschlossen und als experimentelles Ergebnis genutzt werden. Zudem werden die erstellten kleinräumigen Ergebnisse anhand von Geodaten der deutschen Landesvermessung des Bundesamtes für Kartographie und Geodäsie (BKG) validiert und die experimentelle Bevölkerungsverteilung auf Plausibilität geprüft. Fehlzuweisungen, die unplausible Ergebnisse hervorrufen, werden weiterhin durch eine nachfolgende Modifizierung der Mobilfunkdaten bestmöglich korrigiert. Insgesamt stellt der Artikel damit den aktuellen Stand der Arbeit zur experimentellen Georeferenzierung der Bevölkerungszahl mittels Mobilfunkdaten dar.

Dieser Artikel ist wie folgt gegliedert: Im nachfolgenden Abschnitt werden die Datengrundlagen beschrieben, die sich in die amtliche Bevölkerungsfortschreibung sowie die verwendeten Mobilfunkdaten aufteilen. Hierbei wird insbesondere auf die Datenstrategie und -aufbereitung der Mobilfunkdaten eingegangen. In Abschn. 2.3 wird die Methode – genauer das Verteilungsverfahren – zur Umverteilung der Bevölkerungszahlen anhand der zuvor beschriebenen Daten erläutert. Die resultierenden Ergebnisse werden in Abschn. 2.4 diskutiert und auf Plausibilität geprüft. Im letzten Abschnitt wird ein Fazit zur hier beschriebenen Erstellung kleinräumiger Bevölkerungszahlen unter Verwendung von Mobilfunkdaten gezogen und es werden weitere Schritte sowie Schlussfolgerungen zur Diskussion gestellt.

## 2.2 Datengrundlage: Bevölkerungsfortschreibung und Mobilfunkdaten

Die Bevölkerungsfortschreibung ermöglicht die Angabe aktueller Einwohnerzahlen bis auf Ebene der Gemeinden und gibt die Bevölkerungszahl und die Zusammensetzung der Bevölkerung untergliedert nach Geschlecht, Alter, Familienstand und Staatsangehörigkeit wieder (Statistisches Bundesamt, 2021b).

Die Einwohnerzahl wird nach § 5 BevStatG auf Basis der letzten Volkszählung (gegenwärtig Zensus 2011) anhand von Angaben der Statistiken zu Geburten und Sterbefällen, zu Staats-

angehörigkeitswechseln und Lösungen von Ehen und Lebenspartnerschaften sowie der Wanderungsstatistik laufend fortgeschrieben (Statistisches Bundesamt, 2021b). Die Fortschreibung der Bevölkerungszahlen insgesamt sowie untergliedert nach Alter und Geschlecht resultiert aus den statistischen Ergebnissen der Bevölkerungsbewegungen, worunter Wanderungen, Geburten, Sterbefälle und Eheschließungen zu verstehen sind. Untergliedert werden diese nach den natürlichen Bevölkerungsbewegungen, hierunter fallen Geburten sowie Sterbefälle, und nach den räumlichen Bevölkerungsbewegungen, den Zu- und Abwanderungen über Gemeindegrenzen hinweg, die aus entsprechenden Verwaltungsdaten wie Standesämter und Meldebehörden gezogen werden (Statistisches Bundesamt, 2021b).

Die demografischen Merkmale der Zusammensetzung aus der Bevölkerungsfortschreibung liegen zudem in unterschiedlicher regionaler Gliederungstiefe vor, wobei die Merkmale Geschlecht, Alter und Staatsangehörigkeit (deutsch/nicht-deutsch) bis auf Gemeindeebene und der Familienstand nur auf der Kreisebene sowie einzelne Staatsangehörigkeiten auf der Landesebene vorliegen (Statistisches Bundesamt, 2021b). Insgesamt werden die Ergebnisse auf Ebene der Gemeinden, Kreise, Bundesländer und das Bundesgebiet nach dem Gemeindeverzeichnis[9] des Statistischen Bundesamtes ausgewiesen.

Neben der Bevölkerungsfortschreibung stellen Mobilfunkdaten die zweite elementare Datengrundlage in diesem Artikel dar. Aufgrund des Potenzials, die Verteilung der Tages- und Wohnbevölkerung gut und zeitnah abzubilden (Hadam et al., 2020), stellen sie – besonders durch die starken Zusammenhänge mit der Wohnbevölkerung – eine geeignete Grundlage dar, um die Ergebnisse der Bevölkerungsfortschreibung kleinräumig zu verteilen.[10] Seit dem Jahr 2019 besitzen über 97% der privaten Haushalte in Deutschland ein mobiles Endgerät (Statistisches Bundesamt, 2021a), weshalb die gezählten Mobilfunkaktivitäten bundesweit flächendeckend zu einer realitätsnahen Darstellung der Tages- und Wohnbevölkerung in Deutschland beitragen können. Im Vergleich zu anderen Datenquellen, insbesondere traditionellen Erhebungsdaten, liegen Mobilfunkdaten damit zeitnah, hochaktuell und kleinräumig zur Verfügung und sind grundsätzlich nicht von äußeren Einflüssen, wie Wetterbedingungen, beeinflussbar. Zudem ist der Aufwand der Datenerfassung und -aufarbeitung bei Mobilfunkdaten tendenziell geringer, weshalb die zeitliche Aktualität bei anderen Datenquellen, so wie bspw. traditionellen Erhebungsdaten,

[9] Siehe hierzu: `https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html`.

[10] Hierbei wird vereinfachend von Wohnbevölkerung – genauer der potenziellen Wohnbevölkerung – in den Mobilfunkdaten gesprochen, die jedoch gleichbedeutend mit der Aufenthaltsbevölkerung in den Mobilfunkdaten ist. Dies ist darin begründet, dass die Mobilfunkdatenaufbereitung zwangsweise anderen Definitionen zur Ermittlung der Wohnbevölkerung unterliegt als die amtliche Statistik. Jedoch wurden die Mobilfunkdaten mit dem Ziel aufbereitet, ausschließlich bzw. bestmöglich die potenzielle Wohnbevölkerung abzubilden. Weitere Informationen und Analysen zur Tagesbevölkerung oder auch Aufenthaltsbevölkerung sind in Statistisches Bundesamt (2019a), Hadam et al. (2020) und Hadam (2021) aufgeführt.

tendenziell abnimmt.

Ziel der hier aufbereiteten Mobilfunkdaten ist eine möglichst perfekte Abbildung der potenziellen Wohnbevölkerung in den Mobilfunkdaten, um anhand dessen die Bevölkerungszahlen der Bevölkerungsfortschreibung kleinräumig zu verteilen. Zu diesem Zweck werden Mobilfunkdaten aus dem Netz der Telefónica Deutschland verwendet, die vom Datenanbieter Teralytics GmbH aufbereitet und zur Verfügung gestellt werden. Die Mobilfunkdaten liegen aus 8 ausgewählten Wochen aus dem Jahr 2019, exkl. Ferien und Feiertage, auf einem flächendeckenden INSPIRE-konformen 1x1 km Raster[11] vor, um Verzerrungen durch touristische und freizeitliche Aktivitäten zu vermeiden.[12] Da es sich hierbei um Signaldaten handelt, werden alle Signale im entsprechenden Mobilfunknetz vom Netzbetreiber automatisch erfasst, sofern das mobile Endgerät nicht ausgeschaltet ist oder sich im Flugmodus befindet. Dabei wird lediglich die Ortsangabe des Funkmastes registriert, mit dem das mobile Endgerät zu einem bestimmten Zeitpunkt verbunden ist.

Abb. 2.1 stellt beifolgend die Pearson-Korrelationskoeffizienten zwischen den Bevölkerungszahlen des Zensus 2011 und den aufbereiteten Mobilfunkdaten aus dem Netz der Telefónica Deutschland basierend auf einem Mischraster, wie in Statistisches Bundesamt (2019a), nach Wochentag und Uhrzeit erstmalig für ganz Deutschland in einem Liniendiagramm dar. Eine hohe Korrelation in Abb. 2.1 lässt schlussfolgern, dass zu den entsprechenden Zeitpunkten die mobilen Aktivitäten am Wohnort getätigt wurden, da der Zusammenhang zwischen den Bevölkerungszahlen des Zensus 2011 und den Mobilfunkdaten entsprechend stark positiv ist.

Ferner stehen zwei Strategien zur Zählung mobiler Aktivitäten zur Abbildung der potenziellen Wohnbevölkerung zur Verfügung. Die beiden möglichen Datenstrategien sind visuell in den rötlich hinterlegten Rechtecken in Abb. 2.1 hervorgehoben.

Die erste bereits bekannte Option nach Hadam et al. (2020) und dem Statistischen Bundesamt (2019a) bildet einen Datensatz für einen statistischen Sonntagabend, wie in Abb. 2.1 dargestellt, welcher den Durchschnittswert aller Mobilfunkaktivitäten von 20 bis 23 Uhr an den ausgewählten Sonntagen mit einer zweistündigen Verweildauer im Untersuchungsgebiet enthält. Aufgrund der in Abb. 2.1 sichtbaren höchsten Korrelation zwischen den Mobilfunkdaten am statistischen Sonntagabend und den Bevölkerungszahlen des Zensus 2011 wird angenommen, dass dieser Zeitraum einen guten Indikator für die Darstellung der Bevölkerungsverteilung liefert.

[11] Entsprechend der INSPIRE-Richtlinie wurde dazu die Lambert Azimuthal Equal Area Projektion verwendet (ETRS89-LAEA Europe - EPSG:3035).

[12] Eine weitere Unterscheidung der potenziellen Wohnbevölkerung in den Mobilfunkaktivitäten zwischen Erst- und Zweitwohnsitz ist zudem nicht möglich, beides ist in den Mobilfunkdaten vorhanden. Da die Mobilfunkdaten jedoch derart aufbereitet sind, dass der hauptsächliche Aufenthaltsort der Aktivitäten im Jahresdurchschnitt ermittelt wird, ist diese Unterscheidung besonders für die weitere Ermittlung der Tagesbevölkerung nicht ausschlaggebend.

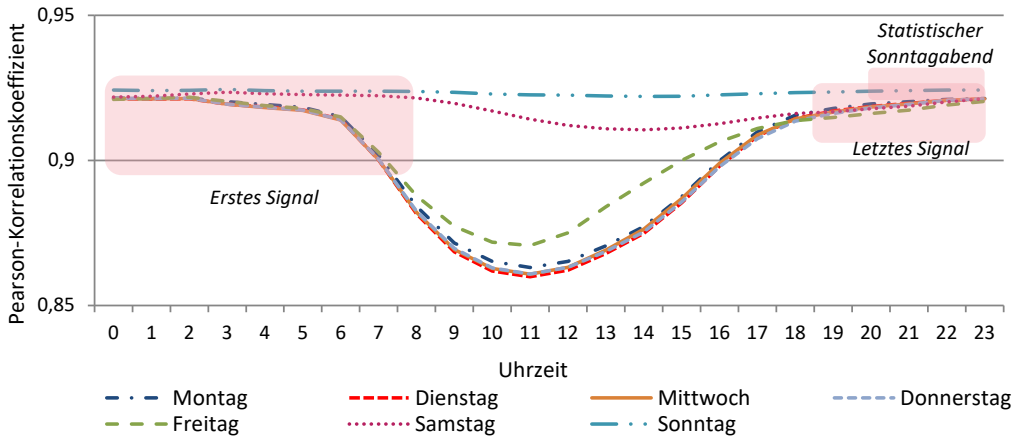



Abbildung 2.1: Pearson-Korrelationskoeffizienten zwischen den Bevölkerungszahlen des Zensus 2011 und den Mobilfunkdaten basierend auf einem bundesweiten Mischraster sowie Darstellung/Hervorhebung der beiden möglichen Mobilfunkdatenstrategien.

Die zweite und damit alternative Datenstrategie zur Ermittlung der Wohnbevölkerung wird als Heimatort-Strategie bezeichnet und ermittelt den Herkunftsort aller erfassten Mobilfunksignale anhand des ersten und des letzten Signals innerhalb von 24 Stunden. Hierbei gibt es verschiedene Möglichkeiten, die die Erfassungen der Signale betreffen, die in Abb. 2.1 hypothetisch im Korrelationsdiagramm in den rötlichen Rechtecken (*Erstes Signal, Letztes Signal*) hervorgehoben sind.[13]

Die eingängigste Definition stellt das räumlich identische erste und letzte Signal dar. Eine räumliche Einheit wird als Herkunftsort bzw. potenzieller Wohnort aus den Mobilfunkdaten bestimmt, wenn das erste und letzte Signal des mobilen Endgerätes innerhalb von 24 Stunden in derselben räumlichen Einheit erfasst wurde. Alternativ hierzu können das erste Signal oder auch das letzte Signal innerhalb von 24 Stunden separat verwendet werden, sofern diese nicht räumlich übereinstimmen. Dabei wird der potenzielle Wohnort auf dem ersten Signal innerhalb von 24 Stunden bestimmt, falls das Signal vor 8 Uhr erfasst wurde und das erste und letzte Event nicht übereinstimmen und vice versa bei der Bestimmung nach dem letzten Signal innerhalb von 24 Stunden.

Weiterhin wird bei der Heimatort-Strategie im Gegensatz zum statistischen Sonntagabend ein Werktagsdurchschnitt gebildet, ohne das Wochenende sowie den Freitag einzubinden, was am veränderten (Bewegungs-)Verhalten der Bevölkerung bzw. der Mobilfunknutzenden in die-

[13] Um welche Uhrzeit das erste und letzte Signal innerhalb von 24 Stunden getätigt wird, ist nicht definiert. Daher werden die Hervorhebungen in Abb. 2.1 zur vereinfachten Veranschaulichung der Datenstrategie angenommen.

sem Zeitraum liegt (siehe auch Hadam, 2021). Zudem muss beachtet werden, dass die Angabe des Herkunftsortes nur über das separate erste oder letzte Signal verzerrt ist, wie Abb. 2.2 veranschaulicht.

Werden die räumlich identischen ersten und letzten Signale ins Verhältnis zu allen verfügbaren (ersten u./o. letzten) Signalen über die Werktage Montag bis Donnerstag gesetzt, sticht das Autobahnnetz in Deutschland insbesondere im östlichen Teil Deutschlands in Abb. 2.2 sichtbar hervor. Der Anteil des separaten ersten oder letzten Signals in diesen Bereichen erscheint gering im Vergleich zu den anderen Regionen, jedoch bilden sie an den Bundesgrenzen sowie auf dem Autobahnnetz die Hauptaktivitäten in den verfügbaren Mobilfunkdaten. Insgesamt macht das erste Signal nur 2,2% aller Mobilfunkaktivitäten aus, das letzte Signal nur 3,9% aller Aktivitäten und damit fallen 93,8% aller Aktivitäten der ausgewählten Werktage auf die räumlich identischen ersten und letzten Signale, weshalb durch diese Anforderung kein Informationsverlust in den Mobilfunkdaten entsteht. Schlussendlich stellen diese Herkunftsorte nicht die potenzielle Wohnbevölkerung in Deutschland dar und werden schließlich in den nachfolgenden Analysen nicht weiter einbezogen.

Da nur Mobilfunkdaten eines von insgesamt drei Mobilfunkanbieter auf dem deutschen Markt zur Verfügung stehen, wurden die Mobilfunkdaten vom Datenanbieter extrapoliert, wobei ein konstanter Extrapolationsfaktor auf Landkreisebene basierend auf Einwohnerzahlen der Bevölkerungsfortschreibung berechnet wurde. Dabei wurde die Extrapolation nur für Mobilfunkaktivitäten deutscher SIM-Karten durchgeführt, um Verzerrungen durch ausländische oder touristische Aktivitäten zu vermeiden. Dies erfolgte durch eine sogenannte Roamerkorrektur, wobei die Roamer (nicht-deutsche SIM-Karten) bei der Berechnung der Extrapolationsfaktoren herausgerechnet wurden.

Um flächendeckende Mobilfunkdaten für das 1x1 km Raster zu erhalten, mussten die Mobilfunkaktivitäten durch den Datenanbieter in einem letzten Schritt – sofern notwendig – anhand von weiteren Bevölkerungszahlen[14] modelliert bzw. räumlich verteilt werden. Das bedeutet, dass die Anzahl der mobilen Aktivitäten nicht in jedem Fall eindeutig einer einzelnen Gitterzelle zugewiesen werden kann.

Abb. 2.3 visualisiert hierbei vereinfachend das Modellierungsprinzip. Im Idealfall wird jede Gitterzelle von mindestens einer Mobilfunkzelle abgedeckt, auf deren Basis die mobilen Aktivitäten initial erfasst werden (siehe Abb. 2.3 **a**). Bei einem bundesweiten 1x1 km Raster wird diese Bedingung nur in dicht besiedelten Regionen erfüllt. Sofern die Mobilfunkzelle mehr als eine Gitterzelle, wie in Abb. 2.3 **b** und **c**, abdeckt, werden die Mobilfunkaktivitäten vom Datenanbieter Teralytics anhand der ihnen zur Verfügung stehenden Bevölkerungszahlen probabilistisch in die Gitterzellen verteilt. Dies ist vor allem in ländlichen oder weniger dicht besiedelten Regio-

[14] Hierbei handelt es sich um Daten eines Customer Intelligence Unternehmens auf Ebene von Wohnbezirken.

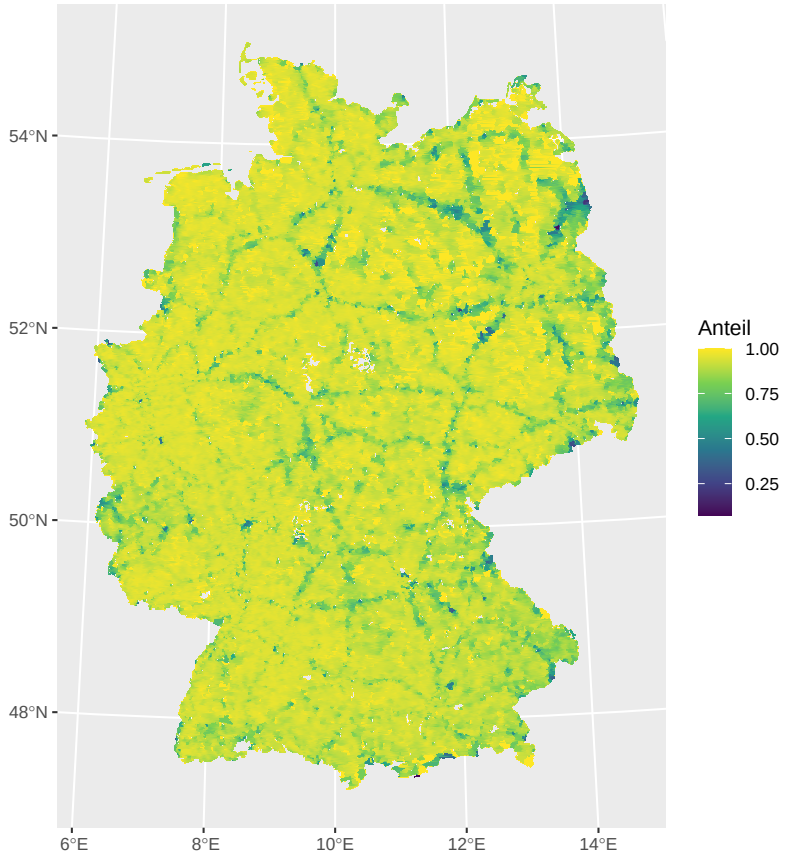



Abbildung 2.2: Anteil der identischen ersten und letzten Signale an allen verfügbaren Signalen (erstes u./o. letztes) über die Werktage (Mo.-Do.) mit sichtbar höherem Anteil des separaten ersten oder letzten Signals auf dem deutschen Autobahnnetz.

nen der Fall. Der Modellierungsgrad des Datenanbieters zeigt auf, dass 5,6% der Mobilfunkdaten den 1x1 km Gitterzellen eindeutig zugewiesen werden konnten (vergleichbar mit Abb. 2.3 **a**). 37,5% wurden kleinräumig modelliert (vergleichbar mit Abb. 2.3 **b**), was bedeutet, dass die Mobilfunkzelle zwischen zwei und neun Gitterzellen abdeckt. Deckt die Mobilfunkzelle mehr als neun Gitterzellen ab, werden die Mobilfunkaktivitäten großräumig modelliert bzw. mit einfachen Annahmen räumlich verteilt und die Genauigkeit der Zuweisungen mobiler Aktivitäten lässt deutlich nach (vergleichbar mit Abb. 2.3 **c**). Dies ist bei 56,9% der Gitterzellen der Fall und der größte Treiber möglicher Unsicherheiten in den resultierenden Ergebnissen.

Zu guter Letzt liegen ebenfalls die soziodemografischen Merkmale Altersgruppe und Geschlecht ausschließlich der Vertragskundinnen und -kunden vor (Datenstand: 2021).

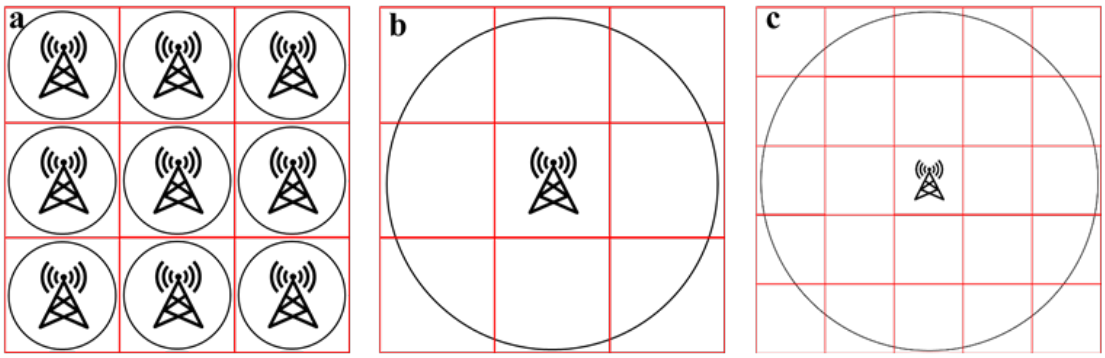



Abbildung 2.3: Räumliche Verteilung/Modellierung der Mobilfunkaktivitäten durch den Datenanbieter nach Modellierungsgrad (**a**) keine, (**b**) kleinräumig und (**c**) großflächig modelliert.

## 2.3 Methodik

### 2.3.1 Räumliche Zuordnung der Gitterzellen und Mobilfunkaktivitäten

Für eine präzise Umverteilung der Ergebnisse der Bevölkerungsfortschreibung von der Gemeindeebene auf Ebene der 1x1 km Gitterzellen bedarf es einer akkuraten Zuordnung der Gitterzellen zu der überdeckenden Gemeinde. Andernfalls besteht die Gefahr einer fehlerhaften kleinräumigen Verteilung von Bevölkerungszahlen innerhalb ihrer Gemeinde durch die aufbereiteten Mobilfunkdaten.

Die einfachste Methode bietet eine geografische Mittelpunktzuordnung, bei der der Mittelpunkt einer Gitterzelle anhand ihrer räumlichen xy-Koordinaten in einer eindeutig überdeckenden Gemeinde verortet und dieser zugeordnet wird. D.h., die Gitterzelle wird derjenigen Gemeinde zugewiesen, deren Mittelpunkt sie überdeckt. Der Vorteil dieser Methode ist die schnelle und einfache Umsetzung in gängigen Geoinformationssystem-Softwareprodukten. Der Nachteil liegt in der teilweise fehlerhaften oder auch nicht möglichen Zuordnung von Gitterzellen zu Gemeinden. Dies ist vorwiegend in Bundesländern mit flächenmäßig kleinen Gemeinden der Fall, bei der eine Zuordnung nicht oder nur stark verzerrt möglich ist. Im vorliegenden Fall resultieren bei einer Mittelpunktzuordnung 46 Gemeinden in Deutschland, die nicht von Gitterzellen überdeckt und zugeordnet werden können (siehe Tab. 2.1).

Eine zweite und zuverlässigere Möglichkeit bietet die Zuordnung der Gitterzelle zu einer überdeckenden Gemeinde ausschließlich anhand ihrer Fläche, mit der die Gitterzelle die Gemeinde überdeckt. Dies wurde auch in dieser Arbeit umgesetzt. Das Ziel hierbei ist es, die Gitterzelle derjenigen Gemeinde zuzuordnen, die den größten Flächenanteil an einer Gemeinde besitzt, unabhängig von den zugrundeliegenden Bevölkerungsdichten.

In einem ersten Schritt werden die Flächenanteile jeder Gitterzelle zur überdeckenden Ge-

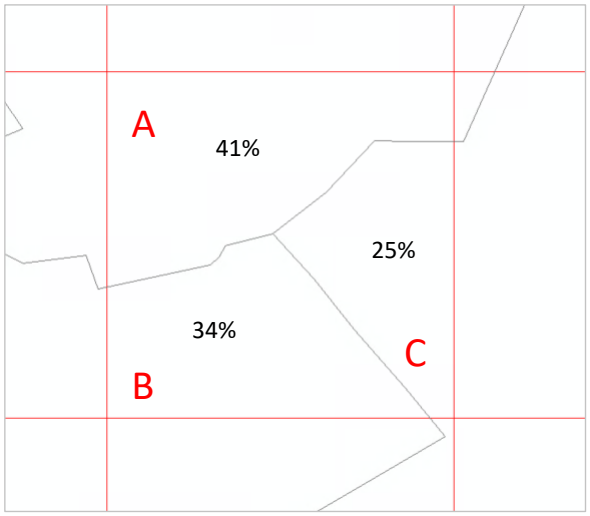



Abbildung 2.4: Flächenzuordnung der Gitterzelle zu den überdeckenden Gemeinden anhand der berechneten Flächenanteile.

meinde berechnet. Abb. 2.4 zeigt beispielhaft eine Gitterzelle, die drei Gemeinden überdeckt. Die Berechnung der Fläche ergibt, dass Gemeinde A den größten Flächenanteil mit 41% an der Gitterzelle besitzt. In einem zweiten Schritt werden die Mobilfunkaktivitäten aus dieser Gitterzelle anhand der jeweiligen Flächenanteile anteilsmäßig auf die drei Gemeinden A, B und C verteilt und damit nicht zu 100% einer Gemeinde zugeteilt. Dadurch erfolgt in diesem Beispiel eine implizite Dreiteilung der Gitterzelle bei der Berechnung der gruppenspezifischen Ziehungswahrscheinlichkeit in Abschn. 2.3.2. Hiermit werden weniger Verzerrungen bzw. Unsicherheiten in den Ergebnissen durch grenzüberlappende Gitterzellen verursacht, weil die Mobilfunkdaten anteilsmäßig den Gemeinden zugeordnet werden, in denen diese flächenmäßig liegen. Im letzten Schritt wird jede Gitterzelle sowie die berechnete experimentelle georeferenzierte Bevölkerungszahl eindeutig der Gemeinde mit dem höchsten Flächenanteil zugeordnet (hier Gemeinde A).

In den hier aufgezeigten Zuordnungsstrategien und im nachfolgend beschriebenen Verteilungsverfahren wird damit implizit eine Strukturgleichheit auf Gemeindeebene sowie dem 1x1 km Raster angenommen. Diese wird gleichsam in sowie zwischen den Mobilfunkdaten und den zu verteilenden Bevölkerungszahlen der Bevölkerungsfortschreibung übernommen und geht aus der nachweisbaren Korrelation beider Datenquellen hervor. Letzteres resultiert zudem in der Annahme, dass sich die vorliegende (kleinräumige) Verteilung der Mobilfunkdaten proportional zur Verteilung der amtlichen Bevölkerungszahlen verhält, weshalb diese als Verteilungsgrundlage in dieser Arbeit Bestand haben. Entsprechend ist auch eine triviale Aufteilung der Mobilfunkakti-

vitäten von der Gitterzelle auf mehrere Gemeinden anhand der Flächenanteile umsetzbar.

In Bezug zu den beiden Mobilfunkdatenstrategien aus Abschn. 2.2 werden beim letzten Verfahren damit alle Gemeinden – mit Ausnahme von zwei Gemeinden, in denen keine Mobilfunkdaten unter anderem durch fehlende Mobilfunkmasten vorliegen – mit Gitterzellen abgedeckt, wie in Tab. 2.1 gelistet. Durch die Abhängigkeit der Datenverfügbarkeit aufgrund der ausgewählten Mobilfunkdatenstrategie bzw. Datenwahl kann auch die präferierte Flächenzuordnung zu fehlenden Ergebnissen einzelner Gemeinden führen, wie es im Fall des statistischen Sonntagabends zur Abbildung der potenziellen Wohnbevölkerung vorliegt (vgl. Tab. 2.1). Wegen fehlender Mobilfunkaktivitäten beim statistischen Sonntagabend können zwei weitere Gemeinden nicht kleinräumig durch Gitterzellen dargestellt werden. Dies trägt maßgeblich zur Wahl der Heimatort-Strategie für die Umsetzung in Abschn. 2.4.1 bei.

Tabelle 2.1: Anzahl fehlender Gemeinden nach Zuordnungsstrategie der Gitterzellen sowie Mobilfunkdatenwahl.

| | Mittelpunktzuordnung (Sonntagabend) | Flächenzuordnung (Sonntagabend) | Flächenzuordnung (Heimatort-Werktagsdurchschnitt) |
|---|---|---|---|
| Anzahl fehlender Gemeinden | 46 | 4 | 2 |

### 2.3.2 Verteilungs- und Rundungsverfahren

Im Rahmen eines Verteilungsverfahrens werden die Ergebnisse der Bevölkerungsfortschreibung von der Gemeindeebene auf eine kleinräumigere Ebene umverteilt. Für die kleinräumige Verteilung der Bevölkerungszahlen aus der Bevölkerungsfortschreibung werden die den Gemeinden räumlich zugeordneten Gitterzellen benötigt, um aus den zugrundeliegenden Mobilfunkaktivitäten den Verteilungsvorgang herzuleiten. Hierfür werden gruppenspezifische Ziehungswahrscheinlichkeiten ($P_{ID}$) für jede Gitterzelle aus den Mobilfunkdaten in Abhängigkeit von der zugrundeliegenden Gemeinde errechnet:

$$P_{ID} = \frac{n_{ID}}{n_i}, \tag{2.1}$$

wobei $n_{ID}$ die Anzahl der Mobilfunkaktivitäten pro Gitterzelle $ID$ und $n_i$ die Summe aller Mobilfunkaktivitäten in der zugeordneten Gemeinde $i$ ist. Vereinfachend ausgedrückt, wird der Anteil der Mobilfunkaktivitäten in Gitterzelle $ID$ im Verhältnis zur Gesamtanzahl aller Aktivitäten in der zugeordneten Gemeinde $i$ berechnet, sodass $P_{ID}$ innerhalb der Gemeinde

variiert.

Anhand der Ziehungswahrscheinlichkeiten $P_{ID}$ wird die amtliche Bevölkerungszahl im nächsten Schritt kleinräumig verteilt. Die experimentelle georeferenzierte Bevölkerungszahl ($ExpGeoBFS_{ID}$) pro Gitterzelle $ID$ ergibt sich dabei aus der Multiplikation der Bevölkerungszahl der Bevölkerungsfortschreibung in Gemeinde $i$ mit der gruppenspezifischen Ziehungswahrscheinlichkeit $P_{ID}$ aus Gleichung 2.1:

$$ExpGeoBFS_{ID} = BFS_i * P_{ID}, \tag{2.2}$$

wobei $BFS_i$ die Einwohnerzahl der Bevölkerungsfortschreibung in Gemeinde $i$ darstellt.

Wie in Abb. 2.5 dargestellt, ermöglicht Gleichung 2.2 eine kleinräumige Verteilung der Bevölkerungszahl von jeder möglichen administrativen Einheit – wie im vorliegenden Fall die Gemeindeebene – auf jede mögliche kleinräumigere Ebene. Notwendig hierfür sind gleichgroße kleinräumige Strukturen, bei denen die gruppenspezifischen Ziehungswahrscheinlichkeiten errechnet werden können. Burgdorf (2010) und Steinnocher et al. (2005) führen im Vergleich hierzu eine räumliche Disaggregation von Bevölkerungsdaten mittels Bebauungsinformation bzw. Bebauungsdichten aus dem amtlichen Digitalen Basis-Landschaftsmodell (ATKIS-Basis-DLM) durch. Burgdorf (2010) nimmt dabei für bestimmte Objektarten unterschiedliche Bevölkerungsdichten an, die zur Vergabe von Gewichten für die Umverteilung der Bevölkerungszahl verwendet werden. Bei Steinnocher et al. (2005) erfolgt die räumliche Aufteilung über eine gewichtete Summenfunktion. Sie leiten bspw. für die untersuchte Region einen spezifischen Faktor ab, der innerhalb der Region konstant ist und vom Verhältnis der Gesamtbevölkerung zur Summe des Flächenanteils und der Bebauungsdichte der entsprechenden Bebauungsklasse abhängt.

Hierbei wird zudem die Relevanz einer bestmöglichen Zuordnung von Gitterzelle zu Gemeinde aus Abschn. 2.3.1 deutlich. Wird die Gitterzellenzuordnung nicht akkurat durchgeführt, hat dies andere gruppenspezifische Ziehungswahrscheinlichkeiten ($P_{ID}$) aus Gleichung 2.1 zur Folge. Daraus ergeben sich zwangsläufig veränderte experimentelle georeferenzierte Bevölkerungszahlen ($ExpGeoBFS_{ID}$) in Gleichung 2.2.

Um nun die amtliche Bevölkerungszahl je Gemeinde aus der Bevölkerungsfortschreibung zu erhalten, werden die experimentellen kleinräumigen Bevölkerungszahlen in einem entsprechenden Verfahren gerundet. Norman et al. (2008) oder Rees et al. (2003) verwenden dafür die sogenannte iterative proportionale Anpassung (Iterative Proportional Fitting (IPF)), um Bevölkerungsgruppen kleinräumig zu disaggregieren und gleichzeitig die Randwerte zu erhalten, sodass Zeilen- und Spaltensummen immer der Gesamtzahl der Bevölkerungsgruppe entsprechen. Im Vergleich dazu werden hier vereinfacht die aus Gleichung 2.2 resultierenden

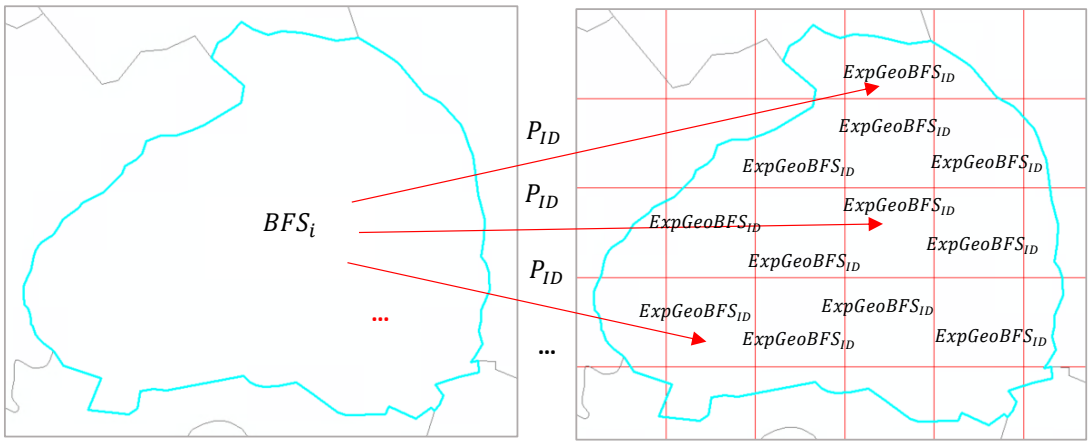



Abbildung 2.5: Visuelle Darstellung des Verteilungsverfahrens.

$ExpGeoBFS_{ID}$ auf Gemeinde $i$ aufsummiert und anschließend anhand der jeweiligen Bevölkerungszahl pro Gemeinde ($BFS_i$) gerundet, wobei die Summe der $ExpGeoBFS_{ID}$ pro Gemeinde $i$ ($\sum_i ExpGeoBFS_{ID}$) dem Wert der Einwohnerzahl der Bevölkerungsfortschreibung ($BFS_i$) entsprechen muss:

$$\sum_i ExpGeoBFS_{ID} = BFS_i. \tag{2.3}$$

Daraus ergeben sich experimentelle georeferenzierte Bevölkerungszahlen, deren Eckwerte denen der amtlichen Bevölkerungsfortschreibung entsprechen.

Gleichung 2.3 wird dabei wie folgt ausgeführt: Im ersten Schritt werden alle kleinräumig verteilten Bevölkerungszahlen anhand der angegebenen Dezimalstellen abgerundet, entgegen der klassischen Rundungsregel. Basierend darauf wird die Differenz zwischen der amtlichen Bevölkerungszahl pro Gemeinde und der aufsummierten kleinräumig verteilten experimentellen Bevölkerungszahl für die zugrundeliegende Gemeinde gebildet.

Als Beispiel sei der folgende Zahlenvektor für die experimentellen georeferenzierten Bevölkerungszahlen ($ExpGeoBFS_{ID}$) aufgeführt: (27,27273; 34,34343; 38,38384). Nach der klassischen Rundungsregel, oder auch Fünferrundung genannt, würde eine Summe von 99 resultieren.[15] Der wahre Wert beträgt jedoch 100, welcher durch das implizite Rundungsverfahren in Gleichung 2.3 hervorgeht. Weiter ergibt sich nun eine Differenz zwischen dem gerundeten und wahren Wert von 1. Daraufhin werden die zuvor abgerundeten $ExpGeoBFS_{ID}$ nach ihren Dezimalstellen geordnet. Dies ist notwendig, um die abgerundeten Dezimalstellen von denjenigen $ExpGeoBFS_{ID}$ um den Faktor 1 zu erhöhen, die letztlich die größten Nachkom-

[15] Klassisch gerundet auf: (27, 34, 38).

mastellen aufweisen. Die eingangs ermittelte Differenz, hier als Beispiel in Höhe von 1, zwischen $BFS_i$ und der aufsummierten kleinräumig verteilten experimentellen Bevölkerungszahl ($\sum_i ExpGeoBFS_{ID}$) bestimmt hierbei die Anzahl der $ExpGeoBFS_{ID}$ pro Gemeinde $i$, deren Werte erhöht bzw. aufgerundet werden müssen, um Gleichung 2.3 zu erfüllen. Im angegebenen Beispiel wird von den drei Werten im Zahlenvektor durch die Differenz von 1 nur ein Wert mit der größten Dezimalstelle aufgerundet, hier 38,38384 auf 39. In Summe resultiert nach Gleichung 2.3:

$$27 + 34 + 39 = 100.$$

Das hier beschriebene Verfahren lässt sich grundsätzlich auch auf weitere Angaben zur Bevölkerung, wie bspw. soziodemografische Merkmale umsetzen, sofern geeignete Mobilfunkdaten oder andere räumlich passende Datenquellen vorliegen, die eine kleinräumige Umverteilung anhand gruppenspezifischer Ziehungswahrscheinlichkeiten zulassen.

## 2.4 Diskussion der resultierenden experimentellen georeferenzierten Bevölkerungszahlen

### 2.4.1 Die experimentelle georeferenzierte Bevölkerungszahl

Wie in Abschn. 2.2 bereits beschrieben, stehen zwei mögliche Mobilfunkdatenstrategien für die Bestimmung experimenteller georeferenzierter Bevölkerungszahlen nach Abschn. 2.3.2 zur Verfügung. Die Wahl der Datenstrategie hängt von zwei grundlegenden Aspekten ab: bundesweit flächendeckende Abdeckung bzw. Datenverfügbarkeit sowie bestmögliche Wiedergabe der Bevölkerungsverteilung.

Aus Tab. 2.1 wurde bereits sichtbar, dass die Mobilfunkdaten nach dem statistischen Sonntagabend nicht flächendeckend vorliegen bzw. die Datenverfügbarkeit durch die restriktive Annahme der ausschließlich gezählten sonntagabendlichen Mobilfunkaktivitäten abnimmt. Daher findet hier die Heimatort-Strategie Verwendung. Wie bereits Abb 2.2 veranschaulichte, werden hierbei zudem nur die räumlich identischen ersten und letzten Signale mobiler Aktivitäten einbezogen. Da das hier beschriebene Verfahren für die folgenden Berichtsjahre 2020 und 2021 umgesetzt wird, ist die flächendeckende Datenverfügbarkeit für die Folgejahre der ausschlaggebende Entscheidungsfaktor. Weitere Informationen zur Mobilfunkdatenwahl und die Auswirkungen der gewählten Mobilfunkdatenstrategie auf die Berechnung der experimentellen georeferenzierten Bevölkerungszahlen sind im Appendix 2.6.1 dargestellt.

Die experimentelle georeferenzierte Bevölkerungszahl nach Abschn. 2.3.2 wird anhand der

Heimatort-Strategie, basierend auf dem Werktagsdurchschnitt von Montag bis Donnerstag, und dem räumlich identischen ersten und letzten Mobilfunksignal der Mobilfunknutzenden innerhalb von 24 Stunden berechnet.[16] Die Ergebnisse sind weiterhin in Abb. 2.6 in einer statischen Karte dargestellt.

Abb. 2.6 stellt die räumliche Verteilung der experimentellen georeferenzierten Bevölkerungszahl anhand einer klassierten Skala dar, wobei die Grenzen der Bundesländer für die Einordnung der Werte hervorgehoben sind. Experimentelle georeferenzierte Bevölkerungszahlen zwischen 0 und 3 werden durch die Angabe eines Intervalls geheim gehalten und in der Karte als solches sowie farblich hell hinterlegt. Hohe Werte der experimentellen georeferenzierten Bevölkerungszahl werden dunkel schattiert (farblich rot) hervorgehoben und niedrige Werte hell schattiert (gelb/orange). Auf den ersten Blick erscheinen die berechneten Werte und die Verteilung plausibel, da besonders dicht besiedelte Regionen bzw. Städte wie Berlin, Hamburg, München, Köln, Bonn oder das Ruhrgebiet entsprechend stark in der Karte hervorstechen. Der eher weniger dicht besiedelte oder auch der ländliche Raum sind entsprechend gelb/orange in Abb. 2.6 hinterlegt. Der weniger dicht besiedelte Raum stellt visuell den Großteil der Fläche in Deutschland dar. Um die Ergebnisse kleinräumiger und individueller betrachten zu können, wurde zusätzlich eine interaktive Rasterkarte erstellt und auf der Seite *Statistik visualisiert* des Statistischen Bundesamtes veröffentlicht (weitere Informationen siehe hierzu den Appendix 2.6.2).[17]

Die Eckwerte der experimentellen georeferenzierten Bevölkerungszahlen können in einigen Gebietsstrukturen oberhalb der Gemeindeebene (Kreis, Bundesland) leicht von den Ergebnissen der amtlichen Bevölkerungsfortschreibung[18] abweichen. Grund hierfür sind nicht verfügbare Mobilfunkdaten in einigen Gitterzellen, die zu fehlenden experimentellen georeferenzierten Bevölkerungszahlen in den Ergebnissen der zugrundeliegenden Gemeinde führen können. Im vorliegenden Fall können zwei Gemeinden in Schleswig-Holstein (Helgoland, Nieby) wegen nicht vorhandener Mobilfunkaktivitäten unter anderem durch fehlende Mobilfunkmasten mit experimentellen georeferenzierten Bevölkerungszahlen ausgewiesen werden (vgl. auch Tab. 2.1). Insgesamt handelt es sich hier um rund 0,05% der Gesamtbevölkerung in Schleswig-Holstein, die folglich nicht mit einer experimentellen georeferenzierten Bevölkerungszahl abgebildet werden können. Eine Aggregation der experimentellen georeferenzierten Bevölkerungszahlen von Gemeinde- auf bspw. Kreisebene kann demnach zu einer geringeren Einwohnerzahl führen als amtlich angegeben.

[16] Tab. 2.6 im Appendix hebt die entsprechenden (finalen) Ergebnisse aus der Summenstatistik kräftig hervor.

[17] Siehe hierzu die am 14. Feb. 2022 erstmals veröffentlichte Anwendung unter: `https://www.destatis.de/DE/Service/Statistik-Visualisiert/Bevoelkerung-Geo/Bevoelkerung_Karten.html`.

[18] Siehe hierzu: `https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html`.

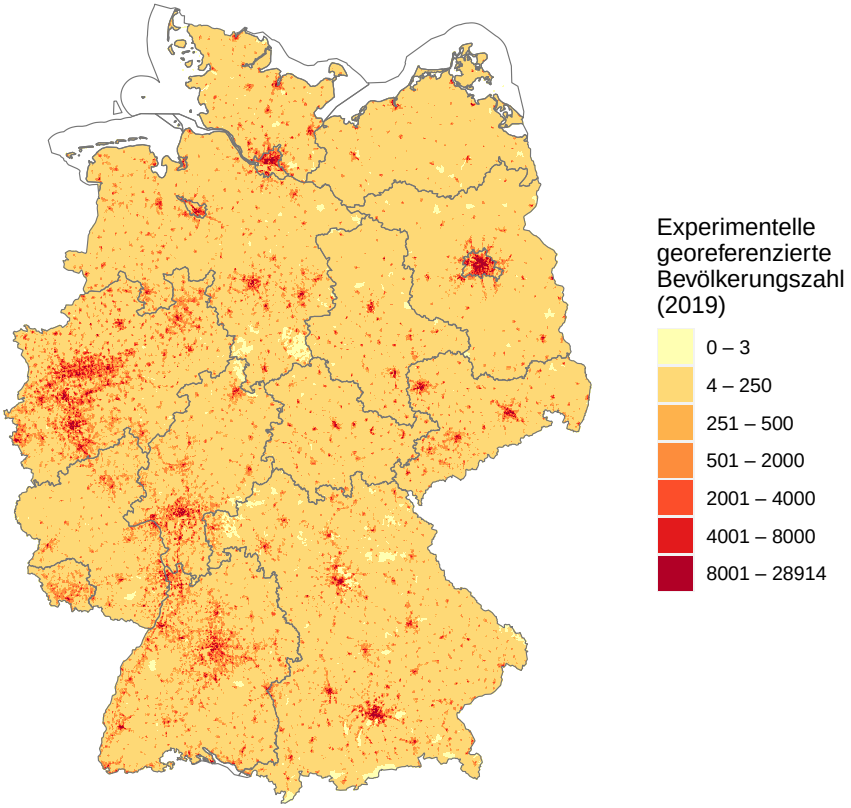



Abbildung 2.6: Bundesweit visualisierte experimentelle georeferenzierte Bevölkerungszahl auf Ebene der 1x1 km Gitterzellen.

Weiterhin wurde untersucht, inwieweit sich das in Abschn. 2.3.2 beschriebene Verteilungsverfahren auf die soziodemografischen Angaben, wie Altersgruppe und Geschlecht, anwenden lässt. Wie durch das Statistische Bundesamt (2021e) bereits ausführlich dargestellt wurde, unterliegen die soziodemografischen Angaben der Mobilfunkanbieter starken Verzerrungen, die sich auch in dem hier beschriebenen Verfahren wiederfinden lassen. Um dieser Verzerrung entgegenzuwirken, wurden Anpassungsfaktoren auf nationaler Ebene durch den Datenanbieter berechnet und die Verteilungen entsprechend angepasst. Besonders ein Fehlen der Nicht-Vertragsmündigen sowie der Prepaid-Kundinnen und -Kunden verhindert eine plausible Darstellung der experimentellen georeferenzierten Bevölkerungszahl differenziert nach Altersgruppen sowie Geschlecht.

Ferner ergeben sich Herausforderungen bei der Nutzung der soziodemografischen Merkmale für das hier beschriebene Verfahren in der resultierenden Datenverfügbarkeit durch die Kreuzkombinationen von Mobilfunkaktivität pro Gitterzelle und soziodemografischem Merkmal beim

Datenanbieter sowie dem umgesetzten Anonymisierungsverfahren. Durch die Unterteilung der Merkmale in mehrere Untergruppen bzw. Klassen erhöht sich das Risiko, dass viele 1x1 km Gitterzellen ohne entsprechenden Wert ausgegeben werden und dadurch keine flächendeckenden Ergebnisse berechnet und dargestellt werden können. Zusätzliche Ausführungen zur Verwendung soziodemografischer Merkmale für die experimentelle georeferenzierte Bevölkerungsfortschreibung sind im Appendix 2.6.3 ausgeführt.

Eine Qualitätseinschätzung der Mobilfunkdaten sowie der Ergebnisse (der experimentellen georeferenzierten Bevölkerungszahl) ist nur anhand weiterer vergleichbarer Datengrundlagen gegeben, welche in den nachfolgenden Abschnitten besprochen werden.[19]

### 2.4.2 Validierung der Ergebnisse – Erste Plausibilitätsprüfung anhand des Zensus 2011

In Abschn. 2.4.1 wurden die ermittelten experimentellen georeferenzierten Bevölkerungszahlen visualisiert und beschrieben. Hierbei wurden erwartbare regionale Differenzen der experimentellen Bevölkerungsdichte sichtbar, die insbesondere urbane Gebiete im Vergleich zum ländlichen Raum stark hervorheben. Aufgrund der räumlich genaueren Aufbereitung der Mobilfunkaktivitäten im urbanen Raum (vgl. Abschn. 2.2, Abb. 2.3) war dies zu erwarten und unterstützt die Annahme, dass die Ergebnisse – insbesondere die Verteilung der experimentellen georeferenzierten Bevölkerungszahlen – in urbanen Regionen plausibel sind. Umgekehrt wird angenommen, dass Unsicherheiten insbesondere in ländlichen Gebieten auftreten und die Ergebnisse dort weniger plausibel erscheinen.

Da es sich bei den Ergebnissen um keine Schätzung im eigentlichen Sinne handelt, ist eine gängige Bestimmung von Unsicherheitsmaßen, wie dem Mean Squared Error (MSE) o.ä., nicht gegeben. Auch eine Prüfung der absoluten Werte der experimentellen georeferenzierten Bevölkerungszahl ist aktuell nicht möglich, da es hierfür keine vergleichbare Datengrundlage gibt. Dies wird erst mit den Ergebnissen des Zensus 2022 möglich sein.

Stattdessen wird eine zweistufige Plausibilisierung durchgeführt. Als erster Schritt zur Validierung der Ergebnisse aus Abschn. 2.4.1 werden diese den georeferenzierten Bevölkerungszahlen des Zensus 2011 gegenübergestellt, um eine erste Einschätzung der Ergebnisse zu erhalten. Im zweiten Plausibilisierungsschritt wird dann in Abschn. 2.4.3 die räumliche Verteilung der experimentellen georeferenzierten Bevölkerungszahl anhand aktueller amtlicher Geodaten auf

[19] Anhand eines vorliegenden Pearson-Korrelationskoeffizienten zwischen den hier verwendeten Mobilfunkdaten und den Einwohnerzahlen aus der Bevölkerungsfortschreibung 2019 auf Ebene der Gemeinden in Höhe von 0,999 kann bereits festgehalten werden, dass die Mobilfunkdaten gut auf höherer administrativer Ebene aufbereitet wurden und eine geeignete Datengrundlage für die kleinräumige Verteilung der amtlichen Bevölkerungszahlen aus der Bevölkerungsfortschreibung darstellen.

Plausibilität geprüft. Grundsätzlich soll anhand dessen im ersten Plausibilisierungsschritt ermittelt werden, inwieweit sich die hier ermittelten experimentellen Bevölkerungszahlen von denen aus dem Zensus 2011 unterscheiden und worin die Unterschiede bzw. mögliche Fehlerquellen bestehen. Aufgrund der zeitlichen Differenz beider Datenquellen stellt dies nur eine grobe Annäherung dar und ermöglicht keine absoluten Aussagen. Entsprechend müssen die Befunde mit Vorsicht interpretiert werden.

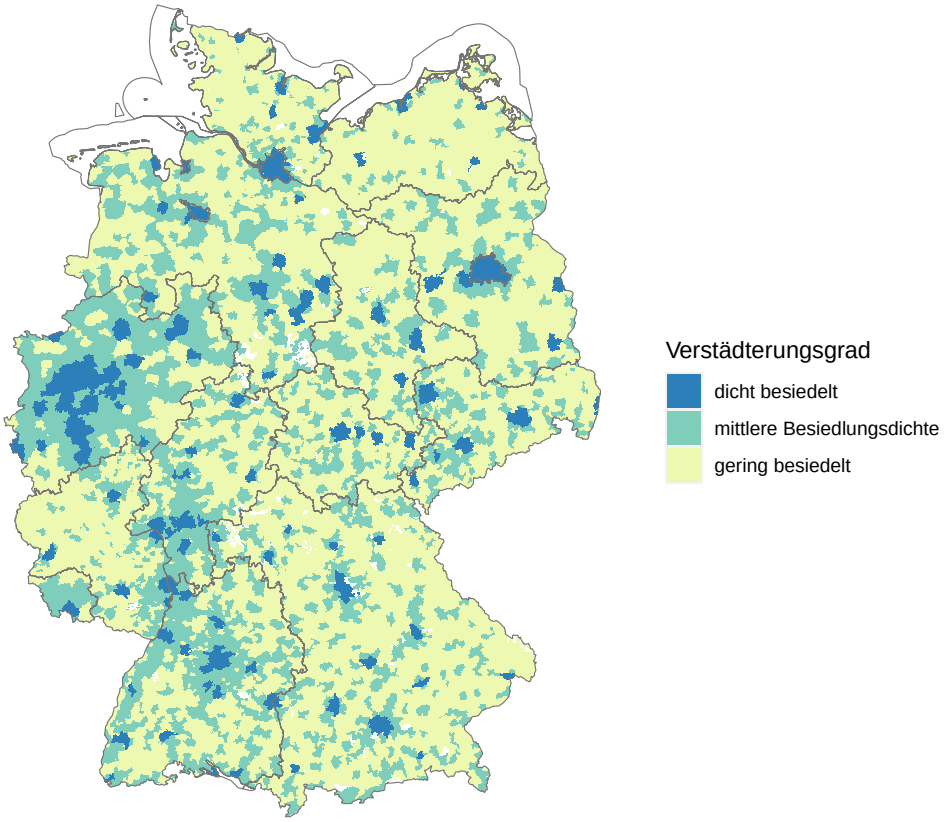



Abbildung 2.7: Verstädterungsgrad in Deutschland.

Um regionale Unterschiede bei der Gegenüberstellung beider Bevölkerungszahlen feststellen zu können, werden diese anhand des Verstädterungsgrades in Deutschland aufgegliedert. Der Verstädterungsgrad wird nach der Definition von der Europäischen Kommission und Eurostat (2021) in drei Kategorien unterteilt, die anhand des Anteils der regionalen Bevölkerung ermittelt werden. Es wird nach Städten oder dicht besiedelten Gebieten unterschieden, die mindestens 50% ihrer Bevölkerung in städtischen Zentren nachweisen. Hinzu kommen kleinere Städte und Vororte bzw. Gebiete mit mittlerer Bevölkerungsdichte, die weniger als 50% ihrer Bevölkerung in städtischen Zentren und nicht mehr als 50% ihrer Bevölkerung in ländlichen Gebieten

vorweisen sowie ländliche Gebiete oder dünn besiedelte Gebiete, die mehr als 50% ihrer Bevölkerung in ländlichen Gebieten aufweisen. Abb. 2.7 stellt den Verstädterungsgrad anhand der drei beschriebenen Kategorien kartografisch dar. Hierbei ist weiterhin entscheidend, wie sich der Flächenanteil sowie die deutsche Wohnbevölkerung auf die drei Raumtypen verteilen (siehe Tab. 2.2).

Aus Abb. 2.7 wird visuell ersichtlich, dass der urbane Raum den geringsten Flächenanteil in Deutschland ausmacht und der ländliche Raum dagegen den größten. In Deutschland gelten nach Tab. 2.2 rund ca. 2% der Fläche als dicht besiedeltes Gebiet, ca. 24,5% als Gebiet mit mittlerer Besiedlungsdichte und ca. 73,5% als ländliches und demnach gering besiedeltes Gebiet. Jedoch wird in Tab. 2.2 gleichzeitig sichtbar, dass knapp 80% der deutschen Bevölkerung auf nur rund 27% der Fläche mit dichter sowie mittlerer Besiedlungsdichte angesiedelt sind. D.h. im Umkehrschluss, dass ein Großteil der Fläche in Deutschland gering besiedelt ist oder nicht bewohnt wird. Weiterhin wird beim Vergleich der beiden Abb. 2.6 und 2.7 deutlich, dass der Verstädterungsgrad sowie die räumliche Verteilung bzw. Dichte der experimentellen georeferenzierten Bevölkerungszahl im Bundesgebiet deckungsgleich sind.

Tabelle 2.2: Flächen- und Bevölkerungsanteile nach Verstädterungsgrad in Deutschland.

| Grad der Verstädterung | Flächenanteil in % | Anteil der ansässigen Bevölkerung je Kategorie in % |
|---|---|---|
| dicht besiedelt | 2,0 | 39,5 |
| mittlere Besiedlungsdichte | 24,5 | 40,3 |
| gering besiedelt | 73,5 | 20,2 |

Für die Feststellung möglicher regionaler Unterschiede bei der Gegenüberstellung beider Bevölkerungszahlen, werden in Abb. 2.8 nun die Pearson-Korrelationskoeffizienten der Gegenüberstellung bzw. die Zusammenhänge der experimentellen georeferenzierten Bevölkerungszahl 2019 und der georeferenzierten Einwohnerzahl basierend auf dem Zensus 2011 und den 1x1 km Gitterzellen differenziert nach dem Verstädterungsgrad betrachtet. Auf der x-Achse ist die absolute Einwohnerzahl aus dem Zensus 2011 hinterlegt und auf der y-Achse die experimentelle georeferenzierte Bevölkerungszahl 2019.

Zunächst geht aus den Korrelationskoeffizienten je Verstädterungsgrad in Abb. 2.8 insgesamt hervor, dass die Zusammenhänge beider Bevölkerungszahlen trotz zeitlicher Differenz von 8 Jahren stark positiv sind und mit einem Koeffizienten von maximal 0,94 in dicht besiedelten Gebieten einhergehen. Weiterhin wird ersichtlich, dass die Korrelationen mit abnehmender Bevölkerungszahl – in Form des Verstädterungsgrades und damit der Bevölkerungsdichte – stetig abnehmen. In ländlichen bzw. dünn besiedelten Gebieten fällt die Korrelation am geringsten aus, womit auch der positive Zusammenhang mit dem Zensus 2011 geringer wird. In den

Gitterzellen, die den dünn besiedelten Gemeinden zugeordnet werden, wird außerdem dazu tendiert, die experimentelle georeferenzierte Bevölkerungszahl im Vergleich zum Zensus 2011 zu überschätzen, was durch die sichtbare Streuung oberhalb der Diagonalen in Abb. 2.8 (unterstes Streudiagramm) sichtbar wird. Hier besteht die Möglichkeit, dass durch die Überschätzung im ländlichen Raum eine Unterschätzung im städtischen Raum durch die anzunehmende fehlerhafte Verteilung in einigen Gitterzellen vorliegen könnte.

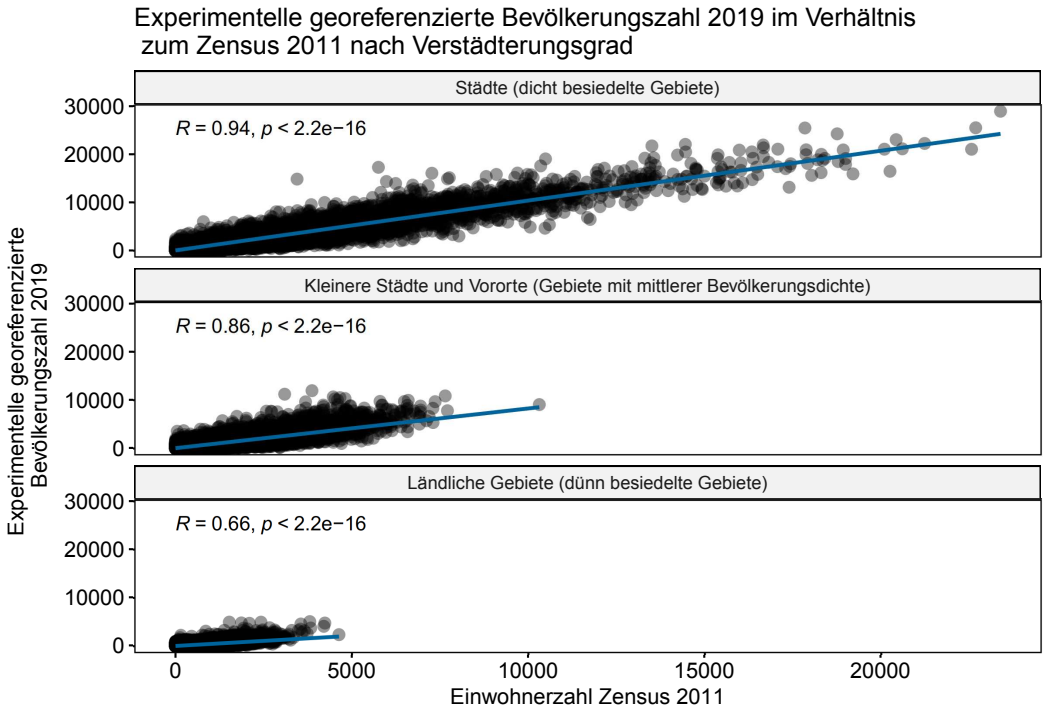



Abbildung 2.8: Korrelationsdiagramm der experimentellen georeferenzierten Bevölkerungszahl 2019 und der Einwohnerzahl basierend auf dem Zensus 2011 nach Verstädterungsgrad.

Tab. 2.3 schlüsselt weiterhin mögliche Fehlschätzungen in den experimentellen georeferenzierten Bevölkerungszahlen auf, die aus dem Vergleich mit den Bevölkerungszahlen aus dem Zensus 2011 auf Ebene der 1x1 km Gitterzellen resultieren. Hierbei fällt besonders auf, dass in 39% aller Gitterzellen mit dem hier verwendeten Verteilungsverfahren Bevölkerung kleinräumig verteilt werden, in denen laut dem Zensus 2011 keine Einwohner gemeldet waren. Dies ist vorwiegend in dünn besiedelten Gemeinden mit knapp 70% der betroffenen Gitterzellen der Fall sowie zu einem merklichen Anteil von 27,3% in kleineren Städten und Vororten. In 0,2% der Gitterzellen gibt die experimentelle georeferenzierte Bevölkerungszahl keine Bevölkerung oder einen geheim zuhaltenden Wert aus, in denen dies laut dem Zensus 2011 nicht der Fall ist. Offensichtlich wird aus Tab. 2.3 und Abb. 2.8, dass Unsicherheiten und Fehlschätzungen bei den

kleinräumigen Zuordnungen der Bevölkerungszahl im ländlichen oder weniger dicht besiedelten Raum in den (vorwiegend modellierten) Gitterzellen angenommen werden müssen.

Tabelle 2.3: Mögliche Fehlschätzungen in der experimentellen georeferenzierten Bevölkerungszahl im Vergleich zum Zensus 2011 auf Ebene der 1x1 km Gitterzellen.

| Mögliche Fehlschätzungen im Verteilungsverfahren (laut Zensus 2011) | Fehlschätzungen pro Gitterzelle in % | Verstädterungsgrad in % | | |
|---|---|---|---|---|
| | | dicht besiedelt | mittlere Besiedlungsdichte | gering besiedelt |
| Keine Einwohner | 39,0 | 3,0 | 27,3 | 69,7 |
| Keine Geheimhaltung/ bewohnt | 0,2 | 4,0 | 27,2 | 68,8 |

Die Erkenntnisse aus Tab. 2.2 relativieren hierbei die weniger guten Ergebnisse in ländlichen Gebieten in Abb. 2.8 sowie Tab. 2.3. Zwar wird ein flächenmäßig großer Anteil in Deutschland mit dem hier beschriebenen Verfahren mit tendenziell nicht plausiblen experimentellen georeferenzierten Bevölkerungszahlen ausgegeben. Jedoch wird demgegenüber aufgrund der Verteilung der Bevölkerung auf die drei kategorialen Gebiete angenommen, dass der Großteil der Bevölkerung besonders im urbanen Raum nachvollziehbar und plausibel kleinräumig verteilt wird. Diese Annahmen werden im folgenden zweiten Plausibilitätsschritt eingehender geprüft.

### 2.4.3 Zweite Plausibilitätsprüfung anhand amtlicher Geodaten

In einem zweiten Plausibilisierungsschritt werden die Erkenntnisse aus der Gegenüberstellung mit dem georeferenzierten Zensus 2011 aus Abschn. 2.4.2 aufgegriffen und die Ergebnisse anhand weiterer Datenquellen, genauer Geodaten aus amtlichen Vermessungsdaten, untersucht, die Informationen zu räumlichen Gegebenheiten in Bezug auf Landnutzung und Wohnflächen liefern. Sie werden verwendet, um die räumliche Verteilung der Ergebnisse in den Gitterzellen auf Plausibilität zu prüfen.

Der Vorteil bei der Nutzung von Vermessungsdaten bzw. Geodaten im Vergleich zu originären Fernerkundungsdaten liegt in den dort bereits aufbereiteten Geometrien. Sie enthalten belastbare Informationen zur Landnutzung und sind vergleichsweise einfach zu verarbeiten, können jedoch aufgrund der aufwändigen Datenaufbereitung seitens des BKG und der Vermessungsämter der Länder meist nur in einem Turnus von mehreren Jahren aktualisiert werden. Somit ist eine gewisse zeitliche Differenz zwischen den zugrundeliegenden Daten vorhanden.

Wie von Deville et al. (2014) bereits vorgeschlagen, wird eine Kombination aus Mobilfunkdaten sowie Geodaten, welche teilweise aus Fernerkundungs- oder genauer Satellitendaten hergeleitet oder mit diesen aktualisiert werden, umgesetzt. Jedoch wird anhand dessen keine zusätzliche Schätzung der experimentellen georeferenzierten Bevölkerungszahl wie in Schug et al.

(2021) erfolgen, stattdessen wird anhand dieser Geodaten eine zusätzliche Plausibilitätsprüfung durchgeführt und bei Bedarf werden Verbesserungsmaßnahmen in der Aufbereitung der Mobilfunkdaten definiert.

Zu diesem Zweck werden aktuelle Geodaten der deutschen Landesvermessung des BKG verwendet, um die Verteilung der experimentellen georeferenzierten Bevölkerungszahl auf Plausibilität zu prüfen. Hierfür werden die *amtlichen Hausumringe Deutschland* (HU-DE) sowie die Geodaten aus dem Datensatz *Haushalte Einwohner Bund* (HH-EW-Bund) verwendet. Dabei wird analysiert, welche Gitterzellen Wohnflächen bzw. eine Wohnnutzung aufweisen und ob folglich eine Wohnbevölkerung in dieser zu erwarten ist oder ausgeschlossen werden kann.

Allein die Fläche zu betrachten, ist für die Plausibilisierung der Ergebnisse nicht ausreichend, da bspw. in ländlichen Gebieten Einfamilienhäuser eine relativ große Fläche für vergleichsweise wenige Einwohnerinnen und Einwohner beanspruchen können. Durch die Kombination der HU-DE mit den Daten des HH-EW-Bund lassen sich genauere Aussagen zur absoluten Verteilung der experimentellen georeferenzierten Bevölkerungszahl treffen. Hierzu werden die georeferenzierten Umringpolygone der Gebäudegrundrisse in Deutschland nach Wohngebäuden anhand der Gebäudefunktionskennung gefiltert und die Anzahl der Hausumrisse in der entsprechenden Gitterzelle bestimmt (BKG, 2021b). Anhand dessen wird primär die räumliche Verteilung der experimentellen georeferenzierten Bevölkerungszahl auf Plausibilität geprüft. Zusätzlich werden die georeferenzierten Daten des HH-EW-Bund verwendet, um die Wohngebiete des HU-DE mit georeferenzierten Adressdaten (GA)[20] sowie die Anzahl der Haushalte pro Adresse und Gitterzelle zu ergänzen. Letztere werden dabei von der infas 360 GmbH anhand diverser Quellen sowie einer stichprobenhaften Erhebung hergeleitet (BKG, 2021a). Der Stand des HH-EW-Bund Datensatzes ist auf Januar 2019 datiert und der des HU-DE auf das Jahr 2021, weshalb die Kombination beider Geodatenquellen aufgrund der zeitlichen Differenz in einigen Fällen zu nicht konsistenten Angaben führen kann.[21]

Eine grundlegende Annahme, die bei der Plausibilitätsprüfung mit diesen Geodaten getroffen wird, ist, dass die Höhe der experimentellen georeferenzierten Bevölkerungszahl mit der Wohnfläche bzw. der Wohnbebauung, den Adresspunkten sowie der damit verbundenen Anzahl an Haushalten einhergeht. Konkret bedeutet dies, dass je mehr Adressdaten pro Gitterzelle vorliegen und je höher der Anteil der Wohnfläche bzw. der Hausumringe sowie die Anzahl der Haushalte pro Gitterzelle ist, desto höher sollte die experimentelle georeferenzierte Bevölkerungszahl tendenziell liegen. Im gleichen Maße geht damit einher, dass keine experimentelle georeferenzierte Bevölkerungszahl in einer Gitterzelle vorliegen darf, in der keine Wohnbebau-

[20] Bei den verwendeten Adressen handelt es sich um die Haus-/Postanschrift aus den Quelldaten (BKG, 2021a).

[21] Grund für diese Differenz ist, dass die genutzte Gebäudefunktionskennung erst im HU-DE Datensatz von 2021 vorliegt.

ung bzw. -fläche, Adressdaten etc. vorliegen.

Daraus ergeben sich folgende Kennzahlen im Sinne eines hier definierten Ampel-Konzepts, das die experimentelle georeferenzierte Bevölkerungszahl in drei Plausibilisierungskategorien einteilt: plausibel (grün), teils plausibel (gelb) und unplausibel (rot).

Als *unplausibel* gelten hierbei alle Gitterzellen mit experimentellen georeferenzierten Bevölkerungszahlen, die einem unbewohnten Gebiet bzw. den dazugehörigen Gitterzellen (keine Wohnfläche bzw. Wohngebäude, Adressen oder Haushalte) zugeordnet wurden oder keine in einer bewohnten Gitterzelle. Als *plausibel* gelten alle Bevölkerungszahlen, die in Gitterzellen verteilt wurden, die mit einem entsprechend hohen Wohnflächenanteil, einer hohen Anzahl an Adresspunkten und an Haushalten einhergehen sowie Gitterzellen, die eine experimentelle Bevölkerungszahl von Null angeben, in der auch keine Wohnflächen etc. ausgewiesen werden. Als *teils plausibel* werden Gitterzellen bewertet, deren zugeordnete experimentelle georeferenzierte Bevölkerungszahl zu hoch oder zu niedrig in Zusammenhang mit der vorhandenen Wohnfläche bzw. der Anzahl an Hausumringen, den Adresspunkten und den geführten Haushalten erscheinen. Hierbei wird zudem als Schwellenwert zur Bewertung der Höhe der experimentellen georeferenzierten Bevölkerungszahl die Anzahl der Haushalte pro Gitterzelle sowie ihre durchschnittliche Anzahl an Personen in Höhe von zwei angeführt (Statistisches Bundesamt, 2020a).[22]

Die in Abb. 2.9 **a** visualisierten Ergebnisse dieser zweiten Plausibilitätsprüfung unterstützen die ersten Annahmen aus Abschn. 2.4.2, dass insbesondere in urbanen Räumen plausible experimentelle georeferenzierte Bevölkerungszahlen ermittelt bzw. zugewiesen werden. Umgekehrt bestätigt sich, dass ländliche Räume tendenziell häufiger mit unplausiblen Ergebnissen, aufgrund der hier verwendeten Mobilfunkdaten und des verwendeten Verteilungsverfahrens, einhergehen. Insgesamt geben die Kennzahlen rund 27,5% der Gitterzellen und die ihnen zugeordneten experimentellen Bevölkerungszahlen als plausibel, 37,2% als teils plausibel und rund 35,3% als unplausibel an (vgl. Tab. 2.5). Hierbei wird erneut in Abb. 2.9 **a** visuell deutlich, dass die weniger dicht besiedelten bzw. ländlichen Gebiete (vgl. mit Abb. 2.7) mit überdurchschnittlich vielen unplausiblen Werten einhergehen. Der Unterschied wird bspw. im Raum Berlin-Brandenburg ersichtlich.

Um weitere Rückschlüsse der Plausibilität der Ergebnisse durch regionale Unterschiede herzuleiten, werden die Ergebnisse aus Abb. 2.9 **a** zusätzlich nach dem Verstädterungsgrad (vgl. Abschn. 2.4.2, Abb. 2.7) aufgeschlüsselt. Tab. 2.4 listet den gerundeten prozentualen Anteil der

[22] Als Beispiel für ein teils plausibles Ergebnis wird eine Gitterzelle mit einer experimentellen georeferenzierten Bevölkerungszahl von 20 angenommen, in der jedoch über 200 Gebäude sowie Adresspunkte und doppelt so viele Haushalte verortet werden. Es kann hierbei davon ausgegangen werden, dass die tatsächliche Bevölkerungszahl aufgrund der hergeleiteten Wohnkapazitäten deutlich höher sein wird, als mit der experimentellen georeferenzierten Bevölkerungszahl berechnet und daher nur als teilweise plausibel eingestuft werden kann.

Kennzahlen nach Plausibilität – sowie farblich nach Abb. 2.9 **a** – und nach dem Grad der Verstädterung auf.

Tabelle 2.4: Auflistung der Ergebnisse der berechneten Kennzahlen (Ampel-Konzept) nach dem Verstädterungsgrad.

| Grad der Verstädterung | Plausibilität der Ergebnisse in % | | |
|---|---|---|---|
| | plausibel (grün) | teils plausibel (gelb) | unplausibel (rot) |
| dicht besiedelt | 53,0 | 31,4 | 15,6 |
| mittlere Besiedlungsdichte | 30,6 | 39,1 | 30,3 |
| gering besiedelt | 23,2 | 36,9 | 39,9 |

Offensichtliche Unsicherheiten ergeben sich bei Zuordnungen der Bevölkerungszahlen in ländlichen, weniger dicht besiedelten Gebieten vorrangig durch die kleinräumige Aufbereitung der mobilen Aktivitäten beim Datenanbieter.[23] 53,0% der Werte je Gitterzelle werden in dicht besiedelten Regionen als plausibel gekennzeichnet, während es in gering besiedelten Gebieten nur 23,2% sind (vgl. Tab. 2.4). Dagegen sind nur 15,6% der Gitterzellen in Städten mit nicht plausiblen experimentellen Bevölkerungszahlen versehen, während es 39,9% der Gitterzellen im ländlichen Gebiet sind. Die teils plausiblen Ergebnisse verteilen sich zu relativ gleichen Teilen auf alle Raumeinheiten bzw. etwas verstärkter in Gebieten mit mittlerer Besiedlungsdichte. Die teils plausiblen Werte sollten zudem mit Vorsicht interpretiert werden. Da es sich hierbei – aufgrund der grundsätzlich nachvollziehbaren räumlichen Verteilung aber der gleichzeitig nicht validierbaren Höhe der zugewiesenen Bevölkerung – um schwer einzuschätzende experimentelle georeferenzierte Bevölkerungszahlen handelt, wird von einer weiteren Analyse der absoluten Werte abgesehen. Insgesamt muss hierbei daher beachtet werden, dass die absoluten Werte der einzelnen experimentellen georeferenzierten Bevölkerungszahlen mit den Geodaten nicht hinreichend bzw. nicht absolut verglichen werden können. Eine Einschätzung hinsichtlich der Plausibilität der Verteilung und der Werte kann mit dem hier beschriebenen Verfahren aber grundsätzlich abgegeben werden.

### 2.4.4 Schlussfolgerungen der Plausibilitätsprüfung – Räumliche Korrektur der Mobilfunkdaten

Insgesamt bleibt damit die Schlussfolgerung bestehen, dass das hier durchgeführte Verfahren insbesondere im urbanen Raum zu nachvollziehbaren, plausiblen experimentellen georeferen-

[23] Die plausiblen Werte verteilen sich weiterhin zu 65,6% auf die nicht modellierten Mobilfunkaktivitäten bzw. die zugrundeliegenden Gitterzellen, zu 29,5% auf die kleinräumig modellierten und nur zu 20,9% auf die modellierten Gitterzellen. Dagegen sind über 43,5% der unplausiblen Ergebnisse in den modellierten, 28,2% in kleinräumig modellierten und nur 8,0% in nicht modellierten Gitterzellen zu finden. Dies bestärkt weiterhin die Problematik der Mobilfunkdatenaufbereitung aus Abschn. 2.2.

zierten Bevölkerungszahlen führt. Ungefähr ein Drittel der Gitterzellen in gering besiedelten Gebieten und Gebieten mit mittlerer Besiedlungsdichte werden dagegen mit unplausiblen experimentellen georeferenzierten Bevölkerungszahlen ausgewiesen. Hierbei handelt es sich um unbewohnte Gitterzellen, denen durch die hier verwendeten Mobilfunkdaten und dem Verteilungsverfahren fälschlicherweise Bevölkerungszahlen zugewiesen werden, was auch in Abschn. 2.4.2 durch den Vergleich mit dem Zensus 2011 bereits angedeutet wurde. Diese offensichtlichen Fehlzuweisungen können allerdings nicht im Nachgang korrigiert werden. Hierfür müssen die Mobilfunkdaten, also die Datengrundlage anhand derer die Bevölkerungszahlen der Bevölkerungsfortschreibung kleinräumig verteilt werden, entsprechend modifiziert werden.

Gründe für die unplausible Verteilung speziell in gering besiedelten Gebieten sind vor allem in der Aufbereitung der Mobilfunkaktivitäten zu finden, die aus der Modellierung der Gitterzellen herrühren. Durch die probabilistische Verteilung der Aktivitäten durch den Datenanbieter ohne Einbezug weiterer Datenquellen als Hilfsinformationen (siehe Abschn. 2.2, Abb. 2.3) werden nicht bewohnte Regionen wie Naturschutzgebiete, Waldgebiete oder Industriegebiete bei der Verteilung der Aktivitäten stets zu gleichen Anteilen mitberücksichtigt.

Liegt nun wie in Deutschland der Fall vor, dass sich die Dichte des Mobilfunknetzes an der regionalen Bevölkerungsdichte orientiert, wird die Netzabdeckung in urbanen Räumen flächendeckender und engmaschiger und in ländlichen Räumen grobmaschiger und möglicherweise lückenhafter.[24] Dadurch sind Stadtzentren bzw. Innenstädte entsprechend mit kleinen Mobilfunkzellen und städtische Randbezirke und generell weniger dicht besiedelte Gebiete mit größeren ausgelegt. Daraus resultieren in urbanen Räumen kleinräumige und räumlich genauere Verortungen von gezählten Mobilfunkaktivitäten als im weniger dicht besiedelten Raum.

Im Rahmen der Mobilfunkdatenaufbereitung konnte diese Netzabdeckung bei der Verteilung der Mobilfunkaktivitäten auf das hier verwendete 1x1 km Raster nicht einbezogen werden. Stattdessen wurden die Aktivitäten in weniger dicht besiedelten Gebieten mit einfachen Annahmen räumlich modelliert bzw. gleichmäßig verteilt. Folglich resultieren unplausible Werte aus den räumlich ungenauen Verteilungen und damit ergibt sich eine Unsicherheit in den Ausgangsdaten sowie den resultierenden Ergebnissen.

Die hieraus resultierende Konsequenz besteht in einer Verbesserung der Mobilfunkdatenaufbereitung beim Datenanbieter für das Berichtsjahr 2020, um anhand einer Modifizierung die Bevölkerungszahlen der Bevölkerungsfortschreibung auf kleinräumiger Ebene genauer zu verteilen. Dies wurde in Form einer Optimierung der Modellierung, genauer einer Steigerung der räumlichen Genauigkeit, von Mobilfunkaktivitäten auf das 1x1 km Raster anhand zusätzlicher

[24] Siehe hierzu die Netzabdeckung in Deutschland von der Bundesnetzagentur im Mobilfunk-Monitoring nach Mobilfunknetzbetreiber und Technologie: `https://www.breitband-monitor.de/mobilfunkmonitoring/karte`.

realitätsbasierter Annahmen erzielt und durch Verwendung der Geodaten des *Landbedeckungsmodells für Deutschland* (LBM-DE) umgesetzt.[25]

Eine präferierte Lösung, um die unbebaute Fläche strikt aus dem Modellierungsprozess der Mobilfunkdaten zu entfernen, gleichzeitig jedoch nicht Wohnflächen herauszufiltern bzw. auszuschließen, bietet eine Filterung der Gesamtfläche des Bundesgebietes. Hierfür wurden räumliche Gebiete nach Landbedeckungs- bzw. Landnutzungskategorien gefiltert, die nicht bebaut sind oder der Landnutzung *Wohnen*[26] nicht zugehörig sind, wie bspw. Industrieanlagen oder bebaute Flächen für den Verkehr, und wo prinzipiell keine Wohnbevölkerung verortet werden darf (siehe hierzu BKG, 2020a). Somit werden alle nicht bebauten Flächen sowie Flächen ohne Siedlungsfunktion exkludiert, was in Deutschland ca. 90% der Gesamtfläche ausmacht, die daraufhin für Verteilungsmaßnahmen nicht mehr beachtet werden.[27]

Die Auswirkungen dieser räumlichen Korrektur auf die Ergebnisse werden in Abb. 2.9 präsentiert. Sie stellt die resultierenden Plausibilitätsprüfungen für das Berichtsjahr 2019 dem Berichtsjahr 2020 gegenüber.[28] Letzteres basiert dabei auf den experimentellen georeferenzierten Bevölkerungszahlen unter Verwendung der räumlich korrigierten Mobilfunkdaten. Hierbei wird bereits visuell deutlich, dass im Ergebnis deutlich plausiblere Ergebnisse durch die Verwendung von Landnutzungsinformationen im Rahmen der Mobilfunkdatenaufbereitung erzeugt werden (vgl. Abb. 2.9 **b**). Insgesamt nehmen damit die als plausibel eingestuften zugeordneten experimentellen Bevölkerungszahlen in Tab. 2.5 um über 40 Prozentpunkte zu, 22,1% gelten als teils plausibel und nur noch 10,1% als unplausibel.

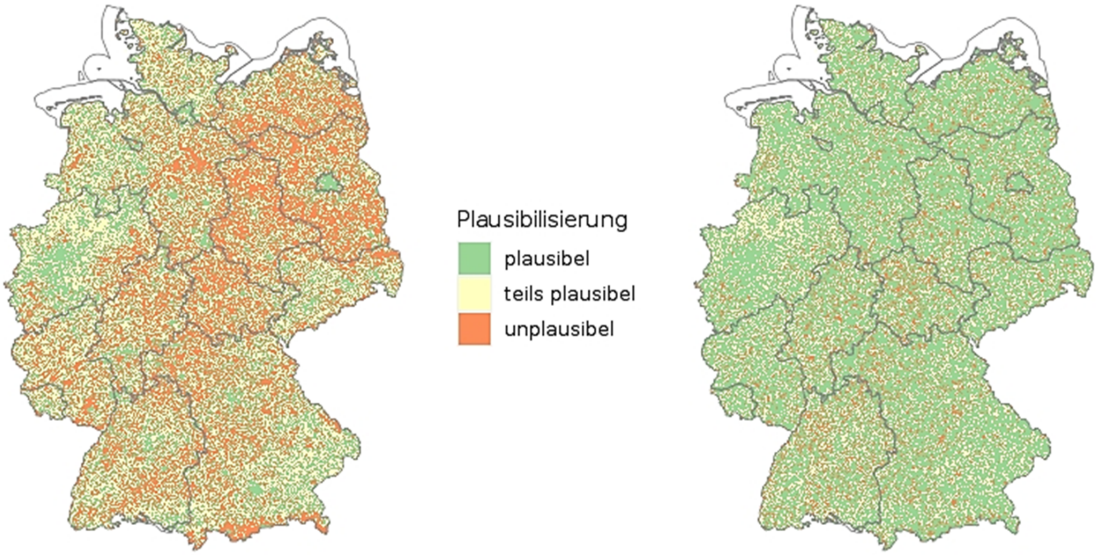
Ferner werden nun 69,5% der Werte je Gitterzelle in dicht besiedelten Regionen als plausibel gekennzeichnet sowie 69,4% in gering besiedelten Gebieten. Dagegen sind nur noch 10,3% der Gitterzellen im ländlichen Gebiet mit nicht plausiblen experimentellen Bevölkerungszahlen versehen. Die teils plausiblen Ergebnisse verteilen sich weiterhin verstärkter in Gebieten mit

[25] Der aktuelle Datenstand des LBM-DE bezieht sich auf das Referenzjahr 2018 und wird in einem Dreijahreszyklus aktualisiert. Dabei werden die amtlichen Vermessungsdaten des ATKIS-Basis-DLM für die Landnutzung sowie Bilddaten der RapidEye und Sentinel-2 Satelliten als auch digitale Orthophotos aus Überfliegungen der Landesvermessungsämter für die Landbedeckung genutzt (BKG, 2020a). Insofern basiert die Abgrenzung von bspw. Wohngebieten oder Produktionsstandorten auf den Daten der Vermessungsbehörden der Länder und weisen eine hohe Genauigkeit auf.

[26] Darunter fallen auch Gebiete wie *Fußgängerzonen* oder *Grasland mit Bäumen zur Wohnnutzung gehörig* (BKG, 2020a).

[27] Die Bodenfläche nach Nutzungsarten in Deutschland zum Stichtag 31.12.2020 kann auch nachvollzogen werden unter: `https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Landwirtschaft-Forstwirtschaft-Fischerei/Flaechennutzung/Tabellen/bodenflaeche-insgesamt.html`.

[28] Siehe hierzu auch die Vorveröffentlichung der Ergebnisse für das Berichtsjahr 2020 vom 6. Juli 2022 des Statistischen Bundesamtes (2022b).




(a) 2019 (b) 2020

Abbildung 2.9: Kartografische Visualisierung der Kennzahlen (Ampel-Konzept) der Plausibilitätsprüfung anhand der regionalen Ergebnisse für die Berichtsjahre 2019 (**a**) und 2020 (**b**) (mit räumlicher Korrektur der Mobilfunkdaten).

Tabelle 2.5: Ergebnisse der räumlichen Anpassung für die experimentelle georeferenzierte Bevölkerungszahl.

| Plausibilität | Plausibilität der Ergebnisse in % | |
|---|---|---|
| | 2019 | 2020 |
| plausibel (grün) | 27,5 | 67,8 |
| teils plausibel (gelb) | 37,2 | 22,1 |
| unplausibel (rot) | 35,3 | 10,1 |

mittlerer Besiedlungsdichte.[29]

Die zusätzliche Einbindung von Landnutzungsdaten bzw. Bebauungsinformation bei der Datenaufbereitung des Mobilfunkdatenanbieters bewirkt eine plausiblere Umverteilung der Mobilfunkdaten mit deutlich weniger Unsicherheit und eine höhere Qualität der Daten sowie der resultierenden Ergebnisse. Dadurch werden insgesamt nicht die Intensitäten der Mobilfunkaktivitäten beeinflusst, aber die Verteilungen deutlich verbessert, weshalb auch bereits plausibel geschätzte experimentelle georeferenzierte Gitterzellen Änderungen an ihren absoluten Werten erfahren haben. Ein weiterer positiver Nebeneffekt dieses Vorgehens liegt in einer leicht umsetzbaren und

[29] Die unplausiblen Ergebnisse verteilen sich zudem nur noch zu 14,8% in den modellierten, 13,0% in kleinräumig modellierten und nur zu 5,0% in nicht modellierten Gitterzellen.

nachvollziehbaren Änderung der Methodik zur Datenaufbereitung beim Datenanbieter sowie dem damit einhergehenden Einfluss und der Mitgestaltung bei der Mobilfunkdatenaufbereitung.

## 2.5 Fazit und Schlussfolgerungen

Im Projekt *Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten* werden im Rahmen eines Verteilungsverfahrens die Ergebnisse der Bevölkerungsfortschreibung von der Gemeindeebene anhand von Mobilfunkdaten bundesweit auf INSPIRE-konforme 1x1 km Gitterzellen kleinräumig umverteilt. Durch den starken nachweisbaren Zusammenhang zwischen Mobilfunkdaten und amtlichen Bevölkerungszahlen bieten Mobilfunkdaten eine geeignete Datenquelle, um die Wohnbevölkerung in Deutschland kleinräumig zu verteilen. Mit diesem Vorgehen wird die vorhandene amtliche Bevölkerungsfortschreibung um ein experimentelles kleinräumiges Ergebnis ergänzt und kann damit zur Schließung der Lücke einer fehlenden aktuellen und georeferenzierten Bevölkerungszahl beitragen, bis die ersten georeferenzierten Bevölkerungszahlen auf Basis des Zensus 2022 vorliegen. Durch die Erstellung und Veröffentlichung eines frei zugänglichen experimentellen Produktes stehen die Ergebnisse zudem uneingeschränkt zur Verfügung. Weiterhin ist das Verfahren zur Erstellung experimenteller kleinräumiger Bevölkerungszahlen grundsätzlich bei allen weiteren statistischen Ämtern umsetzbar, die eine laufende Bevölkerungsfortschreibung erstellen sowie Zugang zu anonymisierten und aggregierten Mobilfunkdaten haben.

Neben den üblichen Nutzungszwecken der amtlichen Bevölkerungsfortschreibung für politische Entscheidungsfindungen können weitere Anwendungsfälle wie bei Fina et al. (2019) oder im Zusammenhang mit dem Krankenhaus-Atlas der Statistischen Ämter des Bundes und der Länder aufgegriffen werden. Grundsätzlich werden die experimentellen georeferenzierten Bevölkerungszahlen sämtlichen Forschungsvorhaben zugutekommen, die aktuelle und kleinräumige Bevölkerungszahlen benötigen und für die bislang nur die georeferenzierten Bevölkerungszahlen des Zensus 2011 zur Verfügung standen. Zwar erfüllen die Ergebnisse nicht den Qualitätsanspruch der amtlichen Statistik, jedoch wird deren Qualität mit den Erkenntnissen aus der Plausibilitätsprüfung verbessert.

Durch das hier verwendete Verteilungsverfahren resultiert ein regional differenziertes Bild der Bevölkerung, das im Schnitt ein plausibles Ergebnis insbesondere im urbanen Raum darstellt. Fehlzuweisungen kleinräumiger Bevölkerungszahlen werden durch die Unsicherheiten im ländlichen Raum durch die Modellierung der dort befindlichen Gitterzellen und ihrer Mobilfunkaktivitäten hervorgerufen, die aufgrund der probabilistischen Zuordnung beim Datenanbieter zu unplausiblen Ergebnissen führen kann. Dies ist verstärkt in den Ergebnissen des Berichtsjahres 2019 der Fall. Die Verwendung von amtlichen Geodaten des HU-DE und des

HH-EW-Bund ermöglichen in der Plausibilisierungsprüfung, diese Bereiche zu identifizieren und anhand des Verstädterungsgrades zu kategorisieren. Bei diesem Verfahren ist eine genaue, uneingeschränkte Prüfung der absoluten Zahlen der experimentellen georeferenzierten Bevölkerungszahl allerdings nicht gegeben. Jedoch können anhand der verwendeten Geodaten Tendenzen zur Höhe der Werte angegeben werden. Die ultimative Validierungsgrundlage stellen daher die georeferenzierten Ergebnisse des Zensus 2022 dar, welche voraussichtlich im Jahr 2023 zur Verfügung stehen.

Aufgrund der Resultate der hier durchgeführten Plausibilitätsprüfung in den zu ermittelnden experimentellen georeferenzierten Bevölkerungszahlen für das Berichtsjahr 2019 kam es folglich zu einer Modifizierung der Mobilfunkdatenaufbereitung anhand des LBM-DE, indem die unbewohnten Flächen in der kleinräumigen Verteilung der mobilen Aktivitäten vom Datenanbieter nicht wieder einbezogen wurden und damit zu einer sichtlichen Verbesserung der Ergebnisse für das Berichtsjahr 2020 führte. Wie Deville et al. (2014) bereits vorgeschlagen haben oder auch Schug et al. (2021) in anderer Form umsetzen, kommt es dadurch für die nachfolgenden Berichtsjahre von einer indirekten (in Form der Plausibilitätsprüfung) zu einer direkten Kombination (in Form der Modifizierung der Mobilfunkdatenaufbereitung) aus Geodaten aus dem LBM-DE sowie Mobilfunkdaten, was zur Qualitätssteigerung der Ergebnisse führt.

Dennoch wird durch die weiterhin bestehende Abhängigkeit vom Datenanbieter hinsichtlich der Mobilfunkdatenaufbereitung die Qualität der Ergebnisse tangiert bzw. eine Aussage hierzu deutlich erschwert, so dass diese nicht die Qualität der amtlichen Statistik vorweisen können. Durch die Verwendung von Mobilfunkaktivitäten nur eines Netzanbieters in Deutschland werden dadurch entstehende Abweichungen und Unsicherheiten in den Ergebnissen sowie in den soziodemografischen Merkmalen durch die jeweiligen regionalen Marktanteile und die verwendete und nicht im Detail offengelegte Methodik des Datenanbieters bei der Datenaufbereitung bedingt (siehe hierzu auch Hadam, 2021).Verzerrungen in unterschiedlichen Regionen können durch ein besseres Extrapolationsverfahren bzw. Gewichtungsverfahren entgegengewirkt werden. Hierzu müssen Stärke und Lage der Verzerrung im Raum bekannt sein, wie es ansatzweise in Statistisches Bundesamt (2019a, 2021e) aufgeführt wird. Für ein zielorientierteres Gewichtungsverfahren sind weitere Informationen zur Mobilfunknutzung in der Bevölkerung erforderlich. Diese könnten bei zukünftigen Haushaltserhebungen, wie dem Mikrozensus, durch zusätzliche Fragen, zum genutzten Mobilfunkanbieter sowie zur Anzahl und Nutzung von Mobilfunkgeräten ermittelt werden, ohne die Auskunftsgebenden hierbei zu stark zu belasten.

Neben der schwierigen Ableitung weiterer Merkmale bleibt grundsätzlich auch die Positionsschätzung der mobilen Aktivitäten, genauer der aktiven SIM-Karten, unterschiedlich präzise, wie in Saidani et al. (2022) aufgeführt. Sie zeigen auf, dass die Ableitung von Positionen der SIM-Karten aufgrund des ungleichförmigen Mobilfunknetzes bspw. anhand von Voronoi-

Polygonen keine genaue Schätzung der mobilen Aktivitäten ermöglicht. Fehlzuweisungen in dem hier beschriebenen Verteilungsverfahren resultieren daher aus Beeinträchtigungen in der Datenverfügbarkeit durch das Mobilfunknetz des Anbieters, sofern Regionen nicht durch das entsprechende Mobilfunknetz abgedeckt werden oder eine genaue Verortung der Mobilfunkaktivitäten aufgrund zu großer Mobilfunkzellen nicht gegeben ist. Ausfälle von Mobilfunkmasten spielen hierbei keine ausschlaggebende Rolle, da für die Erstellung der experimentellen georeferenzierten Bevölkerungszahl ein Jahresdurchschnitt der mobilen Aktivitäten gebildet wird. Auch mögliche Doppelzählungen von SIM-Karten in den Mobilfunkdaten, wie bspw. durch Zweitverträge oder SIM-Karten aus anderen nicht personenbezogenen Geräten provoziert, die nicht durch eine Deduplizierung ausgeschlossen werden konnten, können ebenfalls Einschränkungen in den Ergebnissen hervorrufen. Da weiterhin keine modellbasierte Schätzung vorliegt, ist die Umsetzung einer klassischen Varianzschätzung nicht gegeben, die besonders vor dem Hintergrund der Qualitätsstandards der amtlichen Statistik notwendig ist. Hierbei eignen sich insbesondere Small-Area-Methoden, um die Genauigkeit von kleinräumig geschätzten Bevölkerungsschätzungen zu evaluieren oder auch eine Varianzreduzierung zu bewirken (Simpson et al., 1996; Rao und Molina, 2015). Insgesamt bleiben daher Einschränkungen in der Qualitätseinschätzung der Ergebnisse bestehen.

Die grundsätzlich angestrebte Nutzung von Mobilfunkdaten für die Produktion amtlicher Statistiken kann nur anhand von Daten aller Mobilfunkanbieter in Deutschland erfolgen, um die bundesweite Repräsentativität und Qualität der Daten einschätzen und verbessern zu können. Hierfür bedarf es der Schaffung einer Rechtsgrundlage, um den Zugang zu privat gehaltenen Daten zu ermöglichen und dauerhaft zu sichern und diese langfristig in die amtliche Statistikproduktion integrieren zu können. Erst wenn diese Datenquelle für die amtliche Statistik dauerhaft und vollständig zugänglich ist, kann diese dem Qualitätsanspruch der amtlichen Statistik entsprechend aufbereitet und dauerhaft genutzt werden.

## 2.6 Appendix

### 2.6.1 Zusätzliche Information zur Mobilfunkdatenwahl für die kleinräumige Verteilung der Bevölkerungszahlen

Die Wahl der Mobilfunkdatenstrategie hat Auswirkungen auf die Berechnung der experimentellen georeferenzierten Bevölkerungszahl, auf welche nachfolgend eingegangen wird.

Tab. 2.6 listet die Ergebnisse der nach Abschn. 2.3.2 berechneten experimentellen georeferenzierten Bevölkerungszahlen nach den drei möglichen Optionen der Datenwahl (statistischer Sonntagabend, Heimatort Werktagsdurchschnitt ‚erstes u./o. letztes Signal' und ‚nur identisches erstes u. letztes Signal') anhand einer Summenstatistik. Wie bereits genannt, wird bei der Heimatort-Strategie aufgrund der Art der Mobilfunkdatenaufbereitung durch das erste und letzte Signal innerhalb von 24 Stunden jeweils ein Werktagsdurchschnitt von Montag bis Donnerstag gebildet, um veränderte Gewohnheiten an Wochenenden durch die Mobilfunknutzenden zu exkludieren. Offensichtlich wird, dass sich die Ergebnisse im Durchschnitt nicht voneinander unterscheiden und sich ausschließlich Unterschiede ab dem 3. Quantil zeigen. Im Maximum findet sich bei allen drei Optionen in einer 1x1 km Gitterzelle eine experimentelle georeferenzierte Bevölkerungszahl von fast 30.000. Die final ausgewählten und diskutierten Ergebnisse sind in Tab. 2.6 kräftig hervorgehoben.

Tabelle 2.6: Summenstatistik der berechneten experimentellen georeferenzierten Bevölkerungszahlen nach Mobilfunkdatenwahl.

| Mobilfunkdatenwahl/-Strategie | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Statistischer Sonntagabend | 0 | 26 | 55 | 230,2 | 128 | 29.008 |
| Heimatort Werktagsdurchschnitt (erstes u./o. letztes Signal) | 0 | 26 | 56 | 230,2 | 130 | 28.846 |
| **Heimatort Werktagsdurchschnitt (nur identisches erstes u. letztes Signal)** | **0** | **26** | **55** | **230,2** | **128** | **28.914** |

Abb. 2.10 visualisiert die Verteilung der berechneten experimentellen georeferenzierten Bevölkerungszahlen nach der jeweiligen Mobilfunkdaten-Strategie aus Tab. 2.6 weiterhin in einem Streudiagramm. Auf der x-Achse sind die Ergebnisse des räumlich identischen ersten und letzten Signals laut der Heimatort-Strategie abgebildet. Auf der y-Achse werden die Ergebnisse nach den alternativen Optionen abgetragen – der statistische Sonntagabend sowie der Werktagsdurchschnitt ‚erstes u./o. letztes Signal'. Hier wird noch einmal sichtbar, dass sich die Ergebnisse nicht signifikant voneinander unterscheiden. Jedoch fällt in Abb. 2.10 eine merkliche Streuung um die rote Diagonale (entspricht einer identischen Verteilung der Ergebnisse) auf, die insbesondere durch die unterschiedlichen Ergebnisse der Sonntagabend-Strategie entsteht (blaue Punkte).

Diese schätzen die experimentelle georeferenzierte Bevölkerungszahl in Gitterzellen ab einer geschätzten Bevölkerungszahl von 10.000 tendenziell höher ein als die Heimatort-Strategie ‚nur identisches erstes u. letztes Signal'.

Weiterhin gibt es wenig sichtbare Unterschiede bei den beiden Heimatort-Strategien. Da anhand der Abb. 2.2 jedoch bereits der negative Einfluss der separaten ersten und letzten Signale zum Vorschein kam, werden im Artikel (vgl. Abschn. 2.4.1) die experimentellen georeferenzierten Bevölkerungszahlen anhand der räumlich identischen ersten und letzten Mobilfunksignale innerhalb von 24 Stunden erstellt. Der statistische Sonntagabend wurde vor allem durch die schlechtere Flächenabdeckung und Datenverfügbarkeit (vgl. Tab. 2.1) nicht ausgewählt.

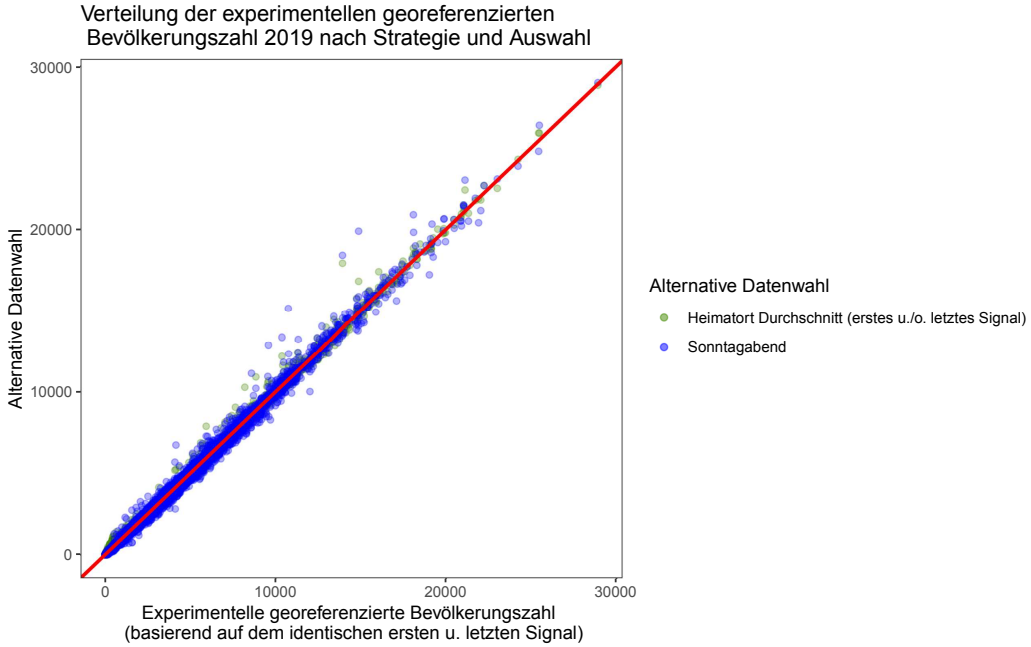



Abbildung 2.10: Vergleich der Verteilungen der experimentellen georeferenzierten Bevölkerungszahlen nach Mobilfunkdaten-Strategie.

### 2.6.2 Die experimentelle georeferenzierte Bevölkerungszahl als interaktive Karte

Die Ergebnisse der experimentellen georeferenzierten Bevölkerungszahl können anhand einer ArcGIS Online-Anwendung auf der Seite *Statistik visualisiert* des Statistischen Bundesamtes individuell und nutzerfreundlich gefiltert, heruntergeladen und für weitere Forschungsfragen ausgewertet werden. Die erstmals am 14. Feb. 2022 veröffentlichte Anwendung ist unter `https://www.destatis.de/DE/Service/Statistik-Visualisiert/Bevoelkerung-Geo/Bevoelkerung_Karten.html` zu finden.

Der Mehrwert der experimentellen kleinräumigen Bevölkerungszahlen wird insbesondere auf der 1x1 km Ebene unter Verwendung zusätzlicher Geodaten-Layer sichtbar. Die experimentelle georeferenzierte Bevölkerungszahl liegt auf Basis von 1x1 km sowie zusätzlich von 10x10 km Gitterzellen vor. Dabei wird für jede Gitterzelle die ermittelte experimentelle georeferenzierte Bevölkerungszahl ausgewiesen, sofern die Zellen zum aktuellen Zeitpunkt mit mobilen Aktivitäten gefüllt sind und zusätzlich nicht der Anonymisierung sowie der anschließenden Geheimhaltung unterliegen.

Abb. 2.11 stellt einen Ausschnitt der interaktiven Anwendung dar. Beispielhaft werden die Ergebnisse für den Raum Berlin dargestellt. Abgebildet werden die experimentellen georeferenzierten Bevölkerungszahlen mit ansteigender Anzahl: In hell eingefärbten Zellen fällt die experimentelle georeferenzierte Bevölkerungszahl gering aus, in dunklen Zellen ist sie höher. Dies ermöglicht zusätzlich einen regionalen Vergleich der aktuellen Bevölkerungsverteilung. Erwartungsgemäß zeigen sich deutliche Unterschiede in der regionalen Verteilung der experimentellen georeferenzierten Bevölkerungszahl zwischen städtischen und ländlichen Gebieten.

### 2.6.3 Verwendung soziodemografischer Merkmale für die experimentelle georeferenzierte Bevölkerungsfortschreibung

Wie bereits in Abschn. 2.4.1 erläutert, wurde die Anwendung des Verteilungsverfahrens auf die soziodemografischen Angaben, wie Altersgruppe und Geschlecht, geprüft und aufgrund der starken Verzerrungen in den soziodemografischen Angaben der Mobilfunkanbieter als nicht umsetzbar erachtet.

Abb. 2.12 stellt zur Veranschaulichung der vorliegenden Verzerrung beispielhaft die Verteilungen der Altersklassen nach der Bevölkerungsfortschreibung 2019 sowie nach den vorliegenden Mobilfunkdaten aus dem Netz der Telefónica Deutschland in einem Balkendiagramm dar. Es werden die jeweiligen Anteile nach Altersgruppe angegeben, die in den Daten vorliegen. Abb. 2.12 unterscheidet zusätzlich die Verteilung der Altersklassen in der Bevölkerungsfortschreibung mit und ohne die minderjährige Bevölkerung, wobei Letzteres durch die Verteilung der Altersgruppe in den Mobilfunkdaten bedingt wird. Hierbei wird vereinfachend mit der Altersgruppe der unter 20-Jährigen der Anteil der Minderjährigen abgedeckt. Die Altersgruppen 18 und 19 Jahre werden für die Vergleichbarkeit in die Altersgruppe 20–30 gezählt, da dies in dieser Form einfachheitshalber durch den Datenanbieter umgesetzt wurde.

Durch die ausschließliche Nutzung der Merkmale der Vertragskundinnen und -kunden wird bereits aus Abb. 2.12 ersichtlich, dass anhand dieser Datenquelle keine Angaben zu den Minderjährigen gemacht werden können und diese ausgehend von der Bevölkerungsfortschreibung die

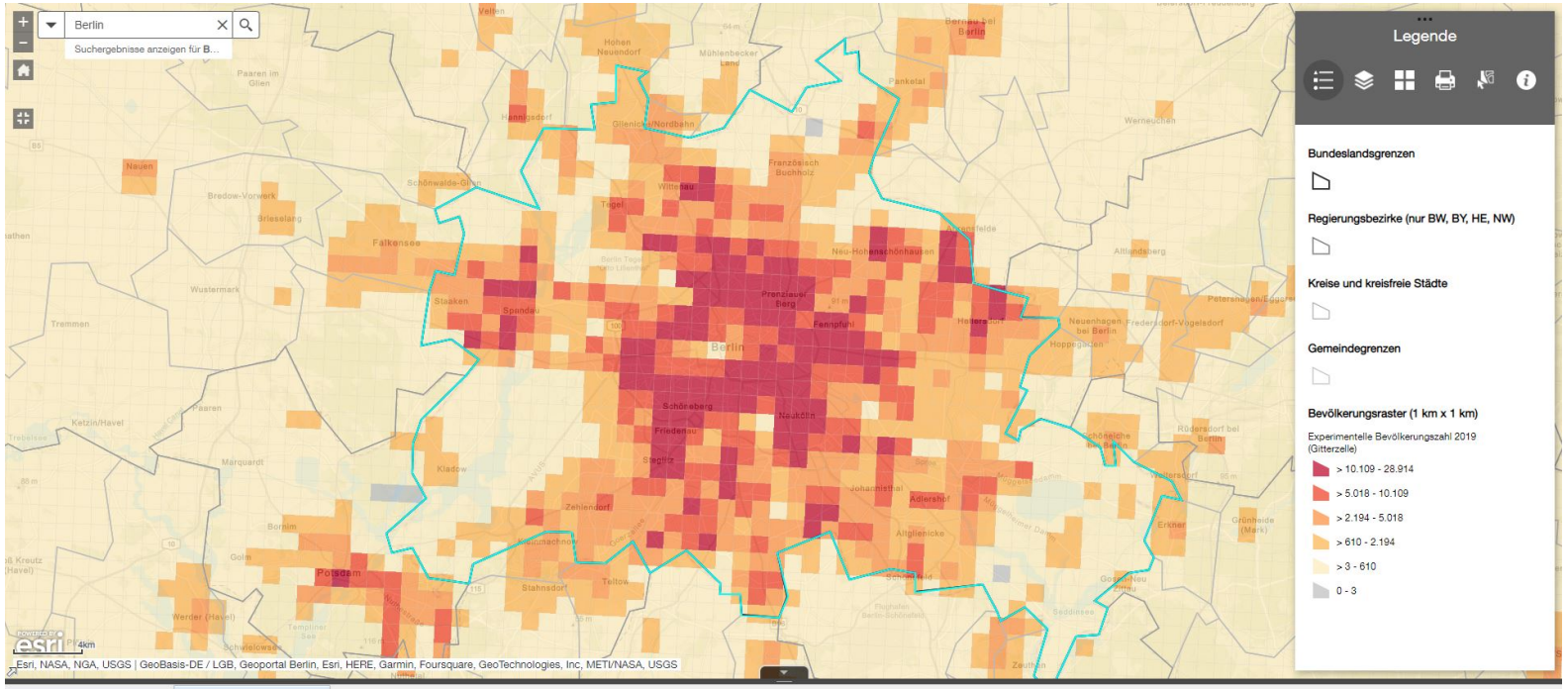



Abbildung 2.11: Ausschnitt der interaktiven Karte im Internetauftritt des Statistischen Bundesamtes (`https://www.destatis.de/DE/Service/Statistik-Visualisiert/Bevoelkerung-Geo/Bevoelkerung_Karten.html`) zur interaktiven Darstellung der experimentellen georeferenzierten Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten.

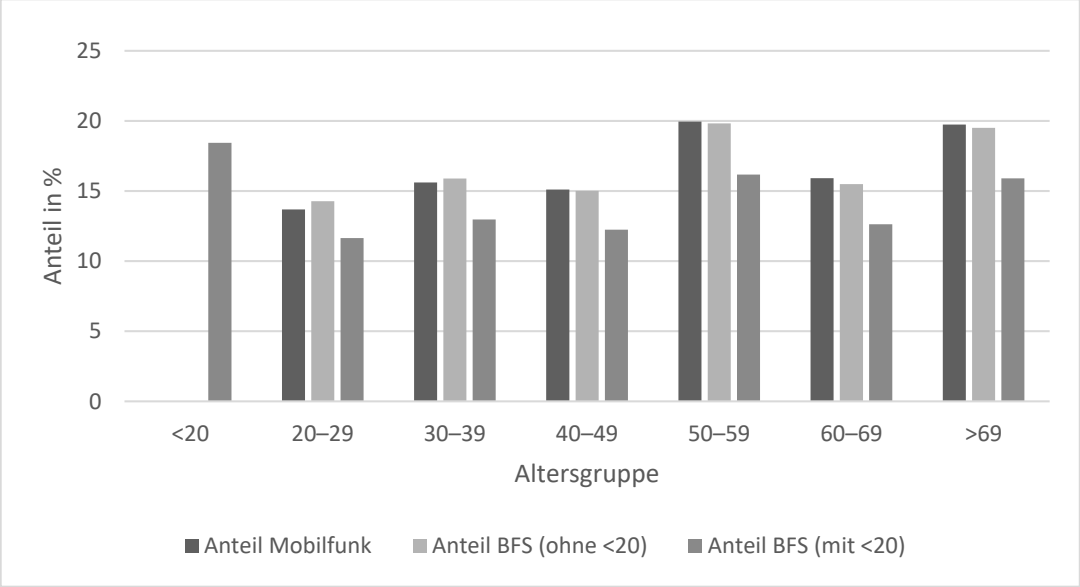



Abbildung 2.12: Gegenüberstellung der Verteilungen der Altersgruppen nach der Bevölkerungsfortschreibung (BFS) 2019 sowie nach den vorliegenden Mobilfunkdaten aus dem Netz der Telefónica Deutschland.

größte Gruppe aller dargestellten Altersgruppen ausmachen. Entsprechend ist keine plausible Darstellung der Bevölkerungsverteilung nach Altersgruppen möglich. Demzufolge ist auch eine Berechnung der Anteile der verschiedenen Altersgruppen für die 1x1 km Gitterzellen nicht gegeben, da bereits die Grundgesamtheit fehlerhaft ist. Bei der Nutzung des Merkmals Geschlecht aus den Mobilfunkdaten sind die Folgen weniger drastisch als bei den Altersgruppen, jedoch sind die ungleichen Verteilungen und sichtbaren Verzerrungen zwischen Mobilfunkdaten und Bevölkerungsfortschreibung wie in vorherigen Untersuchungen des Statistischen Bundesamtes (2021e) auch hier weiterhin der Fall, zumal ebenfalls beim Merkmal Geschlecht alle Minderjährigen sowie Prepaid-Kundinnen und -Kunden nicht einbezogen werden.

# Teil II

# Nutzungsmöglichkeiten von Mobilfunkdaten in Zusammenhang mit dem Mobilitätsverhalten der Bevölkerung

# Kapitel 3

# Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik

## Können Mobilfunkdaten die amtliche Pendlerrechnung unterstützen?

### 3.1 Einleitung

Für diverse politische Entscheidungsfindungen ist es von hoher Bedeutung, Informationen über die Aufenthaltsorte der Bevölkerung zu haben. Durch die regionale Verteilung von Arbeitsplätzen und Wohnorten der Arbeitsbevölkerung entwickeln sich regionale Strukturen, die wiederum durch räumliche Pendlerverflechtungen gekennzeichnet sind (Pütz, 2015). Unter Pendlerverflechtungen werden durch die Gesamtheit der zielgerichteten Pendlerbewegungen entstandene Bewegungsmuster verstanden. Das aus Standortentscheidungen resultierende Pendlerverhalten wird weiter durch die Verfügbarkeit und Qualität der vorhandenen Verkehrsinfrastruktur mitbestimmt, wobei die Erreichbarkeit der Standorte ein wichtiges Merkmal der Qualität der Infrastruktur darstellt. Faktoren wie der Zielort, die benötigte Fahrzeit und -entfernung sowie die Erreichbarkeit sind nach Pütz (2015) besonders für Verkehrsplaner von großer Bedeutung, um adäquat auf signifikante Veränderungen in der Verkehrsnutzung und -auslastung reagieren zu können. Eine laufende Verbesserung der Verkehrsinfrastruktur trägt dazu bei, die Erreichbarkeit der Zielorte zu verbessern und den Zeitaufwand des Pendelns zu reduzieren. Weiterhin liefern die Pendlerverflechtungen wichtige Informationen zur Bestimmung von funktionalen Räumen wie Arbeitsmarktregionen oder Stadt-Land-Regionen.

Der Berufsverkehr ist nach Pütz (2015) eine bestimmende Größe in der Inanspruchnahme der Verkehrsinfrastruktur neben anderen Verwendungszwecken wie für Freizeit-, Erholungs- oder Einkaufsaktivitäten. Da das Pendeln zwischen Wohn- und Arbeitsort nach Haas und Hamann (2008) als „flexible Form der Arbeitskräftemobilität“ an Bedeutung zunimmt, werden die räumlichen Pendlerverflechtungen besonders durch Arbeitsplatzkonzentrationen geprägt (Pütz, 2015). Eine grundsätzlich verbesserte räumliche Erreichbarkeit führt folglich zur Bildung oder Verlagerung von Arbeitsplätzen in Regionen, die bis dato an verkehrsungünstig gelegenen Standorten lagen, woraus sich wiederum neue Pendlerverflechtungen ergeben. Um diese Entwicklungen aufzeigen und nachvollziehen zu können, bedarf es zeitlich hochaktueller und räumlich genauer Informationen zu den aktuellen Pendlerverflechtungen.

Sowohl die amtliche Statistik wie auch andere Institutionen decken diesen Bedarf anhand der sogenannten Pendlerstatistik oder auch Pendlerrechnung ab. Die Pendlerrechnung ist eine Sekundärstatistik, die die benötigten Angaben zum Arbeits- und Wohnort der Erwerbsbevölkerung sowie die Merkmale der Pendler aus unterschiedlichen Statistiken bzw. Datenquellen heranzieht. Beispielsweise veröffentlichen der Landesbetrieb Information und Technik Nordrhein-Westfalen (IT.NRW) oder auch die Statistik der Bundesagentur für Arbeit (BA) jährlich die aktuellen Pendlerverflechtungen in Form eines interaktiven Pendleratlanten. Durch die unterschiedliche Datenlage in beiden Angeboten besteht jedoch eine zeitliche Verzögerung bis zur Veröffentlichung der Ergebnisse. Die BA verwendet beispielsweise ihre eigene Beschäftigungsstatistik mit mehr als 30 Mio. Datenpunkten basierend auf aktuellen Daten aus dem Meldeverfahren zur Sozialversicherung und ermöglicht dadurch eine Veröffentlichung der Ergebnisse mit einem ca. einjährigen Zeitverzug. Allerdings beinhaltet die Pendlerstatistik der BA nur die sozialversicherungspflichtig Beschäftigten (Bundesagentur für Arbeit, 2020). Darauf aufbauend enthält auch die Pendlerrechnung von IT.NRW diese Daten, welche jedoch mit weiteren Informationen aus dem Mikrozensus sowie der Personalstandstatistik zu weiteren Erwerbstätigen unterfüttert werden, wodurch der Informationsgehalt gesteigert wird. Durch die Kombination mehrerer Datenquellen resultiert in den Ergebnissen bei der Veröffentlichung jedoch ein zweijähriger Zeitverzug. Weiterhin findet die räumliche Auflösung nur bis auf die Kreis- oder Gemeindeebene statt. Insgesamt offenbaren diese beiden Beispiele zeitliche und räumliche Verbesserungsmöglichkeiten in der Wiedergabe und Darstellung der aktuellen Pendlerverflechtungen. Besonders eine Verbesserung bzw. Weiterentwicklung der Pendlerrechnung der amtlichen Statistik – hierbei ist exemplarisch die Pendlerrechnung von IT.NRW zu nennen – ist für die Statistischen Ämter des Bundes und der Länder von Bedeutung.

Diverse Studien und Forschungsarbeiten haben sich daher mit der Frage beschäftigt, inwieweit alternative Datenquellen zeitlich hochaktuelle und räumlich genaue Mobilitätsanalysen der Bevölkerung zulassen. Mobilfunkdaten, sogenannte Signaldaten, bieten eine vielversprechende

Datenquelle. Hierbei werden alle Signale im entsprechenden Mobilfunknetz vom Netzbetreiber erfasst. Sie werden automatisch erzeugt, sofern das mobile Endgerät nicht ausgeschaltet ist oder sich im Flugmodus befindet und registrieren lediglich die Ortsangabe des Funkmastes, mit dem das mobile Endgerät zu einem bestimmten Zeitpunkt verbunden ist. Aufgrund der hohen Penetrationsrate von Nutzerinnen und Nutzern mobiler Endgeräte in der Bevölkerung (Statistisches Bundesamt, 2020b) und der zeitlich und räumlich genauen Verortung dieser Geräte lassen sich vielversprechende Fragestellungen zur Mobilität der Bevölkerung betrachten.

Im Zuge erster Machbarkeitsstudien hat das Statistische Bundesamt bereits anhand von Korrelationsmodellen den Zusammenhang zwischen statischen Mobilfunkdaten und amtlichen Bevölkerungszahlen für das Bundesland Nordrhein-Westfalen (NRW) auf kleinräumiger Ebene der Gitterzellen herstellen können (Hadam et al., 2020). Die zeitliche Aktualität und Kleinräumigkeit spiegelt sich dabei auch in der Dynamik der Mobilfunkdaten wider und in der Fragestellung, wohin sich die Mobilfunkaktivitäten im Zeitverlauf bewegen. Wird diese Fragestellung auf die amtliche Statistik projiziert, so kann anhand dieser Daten erforscht werden, ob das Pendlerverhalten und die Pendlerverflechtungen der erwerbstätigen Bevölkerung anhand von Mobilfunkdaten abgebildet werden kann. Deville et al. (2014) zeigen bereits anhand von Mobiltelefondaten, sogenannten Call Detail Records (CDRs), auf, wie dynamisch die Verteilung einer Bevölkerung im Zeitverlauf ist. CDRs liefern individuelle Informationen zu Mobiltelefonnutzenden auf einer hohen räumlichen Auflösung, welche im Gegensatz zu den Signaldaten jedoch ereignisbasiert sind. Die CDRs sind daher nur verfügbar, wenn der Telefonnutzende bspw. aktiv einen Anruf tätigt oder eine SMS bzw. mobile Daten sendet (Jacques, 2018). Am Beispiel Frankreichs und Portugals stellen Deville et al. (2014) mit diesen Daten die Verlagerung der jeweiligen Bevölkerung an die Küstenregionen in den Sommermonaten präzise dar. Diese Informationen sind aktuell auch im Fall von Katastrophen oder Epidemien entscheidend. Ein Beispiel stellen die Auswirkungen der Covid-19 Pandemie auf die Bewegungsfreiheit und damit die Mobilität der Bevölkerung dar. Anhand von Mobilfunkdaten kann das Mobilitätsverhalten der Bevölkerung tagesaktuell ausgewertet und Beschränkungsmaßnahmen können auf ihre Wirkung geprüft werden (Statistisches Bundesamt, 2021c). Unabhängig davon betrachten De Jonge et al. (2012), inwieweit sich Bewegungsmuster in den Niederlanden aus CDRs für die amtliche Statistik ableiten lassen und vergleichen diese mit der niederländischen Mobilitätserhebung, genauer dem Verkehrsaufwand nach Verkehrsmittel in den Niederlanden.[1] Sie berechnen dabei die Reisedistanzen der Mobiltelefonnutzenden und verwenden diesen als Näherungswert für die Mobilitätsstatistik. Aufgrund der Limitation der enthaltenen Informationen, die mit den CDR Daten

[1] Siehe hierzu die niederländische Mobilitätserhebung *Onderzoek Verplaatsingen in Nederland (OViN)* unter `https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingen-in-nederland--ovin--`.

einhergehen, werden die Anzahl der Bewegungen sowie die gesamten Mobilitätsverflechtungen aus den CDRs mitunter unterschätzt oder nicht flächendeckend abgebildet. Im Zusammenhang mit den CDR Daten wird oftmals auch verstärkt die wirtschaftliche Aktivität für räumliche Regionen – wie bspw. Arbeitsmarktregionen – abgeleitet, indem die Intensität und damit das Volumen von aktiven Mobiltelefonen als Indikator für wirtschaftliche Aktivität interpretiert wird (Arhipova et al., 2020). Novak et al. (2013) wiederum betrachten Pendlerströme, worunter aggregierte zielgerichtete Pendlerbewegungen zu verstehen sind, anhand von Standortdaten der Mobiltelefone in Estland. Die Standortdaten bilden in ihrer Studie die wichtigsten Pendlerströme ab und identifizieren dabei wichtige regionale Zentren. Inwieweit kleinräumige und zeitlich aktuelle Pendlerdaten in Verknüpfung mit anderen Datenquellen weiterführend genutzt werden können, zeigen Hadam et al. (2020). Sie nutzen die Intensitäten und die Verteilung der Mobilfunkdaten, um eine alternative Erwerbslosenquote für funktionale städtische Gebiete anhand des Arbeitsortes und des Pendlerverhaltens der Erwerbspersonen zu berechnen.

Das hier beschriebene Projekt *Pendler Mobil* wird in Kooperation mit IT.NRW durchgeführt und sieht Analysen zur Bevölkerungsmobilität in NRW auf Basis von Mobilfunkdaten aus dem Netz der Deutschen Telekom vor. Fortführend wird das Ziel verfolgt, Bereiche anhand des Beispiels von NRW zu identifizieren, in denen Mobilfunkdaten zu einer bundesweiten Ergänzung der bisherigen Pendlerstatistik bzw. -rechnung beitragen können. Anhand von Quelle-Ziel-Matrizen wird untersucht, ob Daten aus dem Mobilfunknetz genutzt werden können, um Pendlerströme im Tagesverlauf abzubilden und inwieweit die aktuelle Pendlerrechnung für NRW mit diesen Daten ergänzt werden kann. Ein Vorteil der Nutzung von Mobilfunkdaten zur Unterstützung der Pendlerrechnung liegt mitunter darin, dass Mobilfunkdaten robust gegenüber Unschärfen bei der Betriebsstättenzuordnung sind. Der Zielort in den Mobilfunkdaten entspricht dem realen potenziellen Arbeitsplatz und damit einhergehenden existenten Pendlerverflechtungen. Letztere sind dadurch nicht abhängig vom eingetragenen Hauptsitz des Unternehmens, welcher bei der amtlichen Pendlerrechnung durch die entsprechenden Datenquellen zwar hinterlegt ist, aber zu dem de facto nicht gependelt wird. Eines der wesentlichsten Vorhaben besteht folglich darin, die arbeitende Bevölkerung und daraus die Pendler aus den Mobilfunkdaten abzuleiten, um diese unter anderem mit der amtlichen Pendlerrechnung von IT.NRW vergleichen zu können. Da die Pendlerrechnung von IT.NRW, wie auch generell, nur auf Gemeindeebene zur Verfügung gestellt wird und aufgrund der dafür notwendigen Daten jährlich mit einem Zeitverzug der Ergebnisse von zwei Jahren aktualisiert werden kann, wird ergänzend geprüft, ob Mobilfunkdaten zu einer kleinräumigeren und schnelleren Abbildung von Pendlerströmen in Form einer *experimentellen Pendlerrechnung* beitragen können. Die Pendlerrechnung dient hierbei als Vergleichsmaßstab, um die Plausibilität der aufbereiteten Mobilfunkdaten anhand dieser zu überprüfen.

Der Artikel ist wie folgt gegliedert: Im nachfolgenden Abschnitt wird die Datengrundlage

beschrieben und auf die amtliche Pendlerrechnung sowie mögliche Ergänzungen bzw. Erweiterungsmöglichkeiten dieser eingegangen. Zudem werden die hierfür benötigten Mobilfunkdaten sowie ihre Aufbereitung anhand der Pendlerrechnung vorgestellt, um die Daten anschließend in Abschn. 3.3 mit den amtlich ermittelten Pendlerströmen zu validieren. Weiterführend werden kleinräumige Pendlerbewegungen für eine erweiterte Zielorts-Bestimmung in Städten hergeleitet und Zusammenhänge zwischen Berufspendlern nach Beschäftigungsumfang und den Mobilfunkdaten geknüpft. In Abschn. 3.4 diskutieren wir die zuvor ermittelten Ergebnisse und gehen auf mögliche Einflüsse auf die Pendlerbewegungen in den Mobilfunkdaten ein. Weiterhin werden Modifizierungsansätze in der Definition und Erstellung der Mobilfunkdaten diskutiert. Im letzten Abschnitt wird ein Fazit des Erweiterungspotenzials der amtlichen Pendlerrechnung durch Mobilfunkdaten gezogen und weitere Schlussfolgerungen zur Diskussion gestellt.

## 3.2 Datengrundlage

### 3.2.1 Die amtliche Pendlerrechnung und ihre Erweiterungsmöglichkeiten

Die amtliche Pendlerrechnung dient zur Ermittlung der täglich zur Arbeit pendelnden Personen, um damit flächendeckende Angaben zur pendelnden Bevölkerung zu ermöglichen. Als Pendler[2] werden erwerbstätige Personen verstanden, die einen Arbeitsweg vom Wohn- zum Arbeitsort zurücklegen. Die Bedingung hierbei ist, dass sich der Arbeitsort vom Wohnort unterscheiden muss, wobei sich der Arbeits- und Wohnort für gewöhnlich auf die räumliche Gebietsstruktur der Gemeinden bezieht (IT.NRW, 2020; Bundesagentur für Arbeit, 2021). Als erwerbstätig werden im Folgenden, entsprechend der Definition der internationalen Arbeitsorganisation (ILO), alle Personen im arbeitsfähigen Alter verstanden, die eine oder mehrere gegen Entgelt ausgerichtete Tätigkeiten ausüben, unabhängig von der Dauer oder dem Umfang der tatsächlich geleisteten Tätigkeit (ILO, 2013).[3]

In der Pendlerrechnung der amtlichen Statistik werden zu den Erwerbstätigen alle sozialversicherungspflichtig Beschäftigten gezählt, deren Grundlage die Beschäftigungsstatistik der BA bildet (IT.NRW, 2020; Bundesagentur für Arbeit, 2021).[4] Die Pendlerstatistik der BA verwendet bspw. nur ihre eigene Beschäftigungsstatistik als Datengrundlage. Der daraus resultierende

[2]Unter dem Fachbegriff „Pendler“ werden sowohl Pendlerinnen als auch Pendler verstanden. Aus Gründen der besseren Lesbarkeit wird im Text verallgemeinernd das generische Maskulinum verwendet.

[3]Neben der ILO-Definition finden sich weitere Definitionen für den Begriff der Erwerbstätigkeit, auf die hier jedoch nicht weiter eingegangen werden. Interessierte Leserinnen und Leser sind beispielsweise auf das U.S. Bureau of Labor Statistics (`https://www.bls.gov/cps/definitions.htm#employed`) verwiesen.

[4]Insgesamt fließen rund 33 Mio. sozialversicherungspflichtig Beschäftigte in die Beschäftigungsstatistik der BA ein, deren Pendlerverflechtungen anhand der Betriebsnummer des Arbeitgebers und der Anschrift des Versicherten einem Arbeitsort und einem Wohnort bis auf Ebene der Gemeinden zugeordnet werden können (Pütz, 2015; Bundesagentur für Arbeit, 2020).

interaktive Pendleratlas gibt somit Informationen aller sozialversicherungspflichtig pendelnden Beschäftigten für jeden Kreis in Deutschland des Vorjahres zum Stichtag 30. Juni wieder, wobei die zehn größten Pendlerströme aus den nahegelegensten Kreisen mit einer Entfernung von ca. 150 km ausgewiesen werden.[5]

Aufgrund einer speziellen Landesstatistik in den Bundesländern Baden-Württemberg, Hessen und NRW wird eine Pendlerrechnung entsprechend nur durch diese drei Statistischen Landesämter anhand einer abgestimmten Methodik der Statistischen Ämter der Länder erstellt und veröffentlicht. Daher liegt bislang keine bundesweite Pendlerrechnung der Statistischen Ämter des Bundes und der Länder vor. In den Pendlerrechnungen der genannten Statistischen Landesämter werden neben den sozialversicherungspflichtig Beschäftigten und geringfügig Beschäftigten weiterhin Beamtinnen und Beamte, Richterinnen und Richter aus der Personalstandstatistik sowie die Selbstständigen und mithelfenden Familienangehörigen aus dem Mikrozensus hinzugezählt (IT.NRW, 2020; Dettmer und Emmel, 2018; Statistisches Landesamt Baden-Württemberg, 2019b). Damit bauen die Pendlerrechnungen der Statistischen Landesämter auf identischen Datengrundlagen auf. IT.NRW veröffentlicht bspw. jährlich die Pendlerbewegungen der Erwerbstätigen auf Gemeinde- und Kreisebene mit einem zweijährigen Zeitverzug und stellt diese für alle Erwerbstätigen in NRW in einem interaktiven Pendleratlas dar (IT.NRW, 2020).[6] Neben den interaktiven Pendleratlanten von IT.NRW und der Statistik der BA wurden die Pendlerrechnungen – jedoch nicht in Form einer interaktiven, d.h. jährlich aktualisierten Karte – vom Statistischen Landesamt Baden-Württemberg und vom Hessischen Statistischen Landesamt in entsprechenden Veröffentlichungen für die jeweiligen Bundesländer publiziert (siehe hierzu Statistisches Landesamt Baden-Württemberg, 2019a; Dettmer und Emmel, 2018).

Deutschlandweit wird die Pendlerstatistik nur auf Basis des Mikrozensus im Rahmen eines vierjährigen Zusatzprogramms ermittelt und vom Statistischen Bundesamt veröffentlicht, wobei die Beantwortung der Fragen zum Pendlerverhalten freiwillig ist. Die Auswertungen geben den Bundesdurchschnitt der pendelnden Erwerbstätigen, nachfolgend Berufspendler genannt, nach der Stellung im Beruf, der Entfernung, dem Zeitaufwand und dem benutzten Verkehrsmittel für den Hinweg zur Arbeitsstätte wieder (Statistisches Bundesamt, 2017a,c). Aktuell liegen die

[5] Siehe hierzu den interaktiven Pendleratlas der BA: `https://statistik.arbeitsagentur.de/DE/Navigation/Statistiken/Interaktive-Angebote/Pendleratlas/Pendleratlas-Nav.html`. Die Anzahl der sozialversicherungspflichtig Beschäftigten am Arbeits- und Wohnort sowie nach Ein- und Auspendler über Gemeindegrenzen oder auf Kreisebene kann im Statistikportal der Statistischen Ämter des Bundes und der Länder heruntergeladen werden (siehe hierzu auch: `https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Publikationen/Downloads/regiostatkatalog-2019.pdf?__blob=publicationFile`).

[6] Siehe hierzu den interaktiven Pendleratlas von IT.NRW: `https://www.pendleratlas.nrw.de/`.

damit ermittelten Angaben der Berufspendler aus dem Mikrozensus für das Jahr 2016 vor.[7]

Berufspendler werden als übergemeindliche Pendler definiert, sofern ihr Arbeitsort nicht in derselben Gemeinde wie ihr Wohnort liegt. Andernfalls spricht man von innergemeindlichen Pendlern. Übergemeindliche Pendler kategorisiert man weiterhin nach Ein- und Auspendlern. Einpendler sind erwerbstätige Personen, die nicht in ihrer Arbeitsgemeinde wohnen, wohingegen Auspendler nicht in der Gemeinde arbeiten, in der sie wohnen (IT.NRW, 2020; Bundesagentur für Arbeit, 2021). Zum besseren Verständnis sei als Beispiel Person $i$ wohnhaft in Gemeinde A und berufstätig in Gemeinde B angeführt. D.h., Person $i$ pendelt von Gemeinde A nach Gemeinde B und umgekehrt. Aus Sicht von Gemeinde A ist Person $i$ demnach ein Auspendler, da Person $i$ außerhalb ihres Wohnortes arbeitet und sich damit aus ihrem Wohnort rausbewegt. Aus Sicht von Gemeinde B ist Person $i$ hingegen ein Einpendler, da diese in einer Gemeinde arbeitet, in der sie nicht wohnt und damit in Gemeinde B einpendelt bzw. sich reinbewegt.

Zusätzliche Differenzierungen der Berufspendler werden unter anderem nach den Merkmalen Alter, Geschlecht, Beschäftigungsumfang oder dem Wirtschaftsbereich vorgenommen. Dabei wird die Strecke des Pendelweges bzw. die Distanz zwischen dem Wohn- und Arbeitsort für die Plausibilisierung der Pendlerströme verwendet. Die Distanz wird durch die Luftlinienentfernung zwischen den geografischen Mittelpunkten des Arbeits- und Wohnortes berechnet. Ein Pendelweg wird hierbei als plausibel bewertet, sofern die berechnete Distanz zwischen Wohn- und Arbeitsort die Grenze von 80 Kilometer nicht überschreitet und diese damit noch als täglich zu bewerkstelligen eingestuft wird (IT.NRW, 2020).

Insgesamt bildet die Pendlerrechnung von IT.NRW das Pendlerverhalten von ungefähr 91% der Erwerbstätigen in NRW ab.[8] Da die Qualität der Pendlerrechnung von IT.NRW insgesamt als sehr gut zu bewerten ist (IT.NRW, 2020), wird im Folgenden nicht die Qualität der veröffentlichten Pendlerströme, sondern potenzielle Erweiterungsmöglichkeiten der amtlichen Pendlerrechnung in NRW für das aktuell veröffentlichte Jahr 2019 an diesem Fallbeispiel betrachtet. Bundesweite Ergänzungen werden zum einen in der Darstellung der Pendlerverflechtungen in räumlichen Gebieten unterhalb der Gemeindeebene gesehen. Besonders kleinräumige Pendlerverflechtungen in großen Städten liefern wertvolle Informationen zu stark frequentierten städtischen Bereichen oder bis dahin nicht aufgezeigten Arbeitsplatzkonzentrationen. Auch im Hinblick auf die Infrastruktur des öffentlichen Nahverkehrs bieten kleinräumige Informationen eine größere und vor allem genauere Planungsmöglichkeit bei der effizienten Gestaltung. Zudem lassen sich ausschließlich anhand kleinräumiger Angaben zum Wohn- und Arbeitsort innergemeindliche Verflechtungen sichtbar machen.

[7] Zudem gibt es weitere alternative Pendleratlanten kommerzieller bzw. privater Anbieter wie den Pendleratlas der Mitfahrzentrale für Pendler, die Daten zu Pendlerbewegungen in Deutschland kombiniert (siehe `https://www.pendleratlas.de/`).

[8] Grund hierfür ist die Vollerhebung in der Beschäftigungs- und Personalstandstatistik.

Da auch die Aktualität der von IT.NRW veröffentlichten Pendlerrechnung des Jahres 2019 aufgrund der verschiedenen Datenquellen Verbesserungspotenzial aufzeigt, wird auch eine mögliche Unterstützung in der zeitlichen Komponente durch die nachfolgend beschriebenen Mobilfunkdaten diskutiert. Zum anderen besteht eine weitere Ergänzungsmöglichkeit in der Abbildung von Bildungspendlern, worunter Schülerinnen und Schüler sowie Studierende zu verstehen sind,[9] da diese aufgrund von Dateninkonsistenzen bei den Wohnortangaben in der Pendlerrechnung nicht berücksichtigt werden (IT.NRW, 2020).[10] Wegen der aktuell begrenzten Möglichkeiten ausschließlich Bildungspendler aus den Mobilfunkdaten abzuleiten, fokussieren wir uns in diesem Artikel primär auf die Darstellung der Berufspendler.[11] Dennoch besteht der Bedarf, regional tiefere Verflechtungen der Bildungspendler zu ermitteln, wie die Zahlen des Statistisches Bundesamt (2017b) zu den Bildungspendlern nach Entfernung, Zeitaufwand und benutztem Verkehrsmittel für den Hinweg zur Schule oder Hochschule für das Jahr 2016 andeuten. Insbesondere zeigt der hohe Anteil von ca. 47% der Bildungspendler, die regelmäßig den öffentlichen Personenverkehr nutzen und auch beanspruchen, wie relevant kleinräumige und aktuelle Angaben zum Pendlerverhalten dieser Personengruppe für eine Optimierung sowie Quantifizierung der Auslastung der Verkehrsinfrastruktur ist.

### 3.2.2 Mobilfunkdaten: Datendefinition und -aufbereitung

Mobilfunkdaten können je nach Spezifikation zahlreiche Strukturen annehmen und damit diverse Fragestellungen beantworten. Daher ist es essenziell die Daten entsprechend dem Forschungsziel aufzubereiten. Ziel dieser Arbeit muss es sein, die Berufspendler bestmöglich aus den Mobilfunkdaten abzuleiten, um diese mit der amtlichen Pendlerrechnung 2019 vergleichen zu können.

Das Statistische Bundesamt und IT.NRW nutzen Quelle-Ziel-Matrizen aus dem Netz der Deutschen Telekom, um Mobilitätsanalysen der Bevölkerung mit besonderem Fokus auf die Pendlerströme in NRW durchzuführen. Quelle-Ziel-Matrizen stellen dabei anonymisierte und aggregierte Bewegungen von Signalen bzw. mobilen Aktivitäten im Mobilfunknetz vom Start- zum Zielort (daher Quelle-Ziel-Matrix), im Folgenden auch als Bewegungsverflechtungen bezeichnet, dar. Unter einer mobilen Aktivität wird ein Signal an einem Funkmast verstanden,

[9] Ausgenommen sind Auszubildende mit Ausbildungsvergütung.

[10] Angaben zu Bildungspendlern werden vom Statistischen Bundesamt nur als bundesweiter Durchschnitt aus dem Mikrozensus 2016 (Statistisches Bundesamt, 2017b) oder wie in Dettmer und Wolf (2018) als Durchschnittswert der Ausbildungspendler für das Bundesland Hessen sowie nach Gemeindegrößenklassen der Wohnsitzgemeinde als mögliche Ergänzung zur Pendlerrechnung veröffentlicht.

[11] Die in Abschn. 3.2 beschriebenen Mobilfunkdaten haben diverse Limitationen, wie die repräsentative Angabe der Altersgruppen (siehe bspw. Statistisches Bundesamt, 2021e), die es nicht ermöglichen plausible Pendlerverflechtungen von jungen Personengruppen aufzuzeigen (siehe hierzu auch Abschn. 3.4.2).

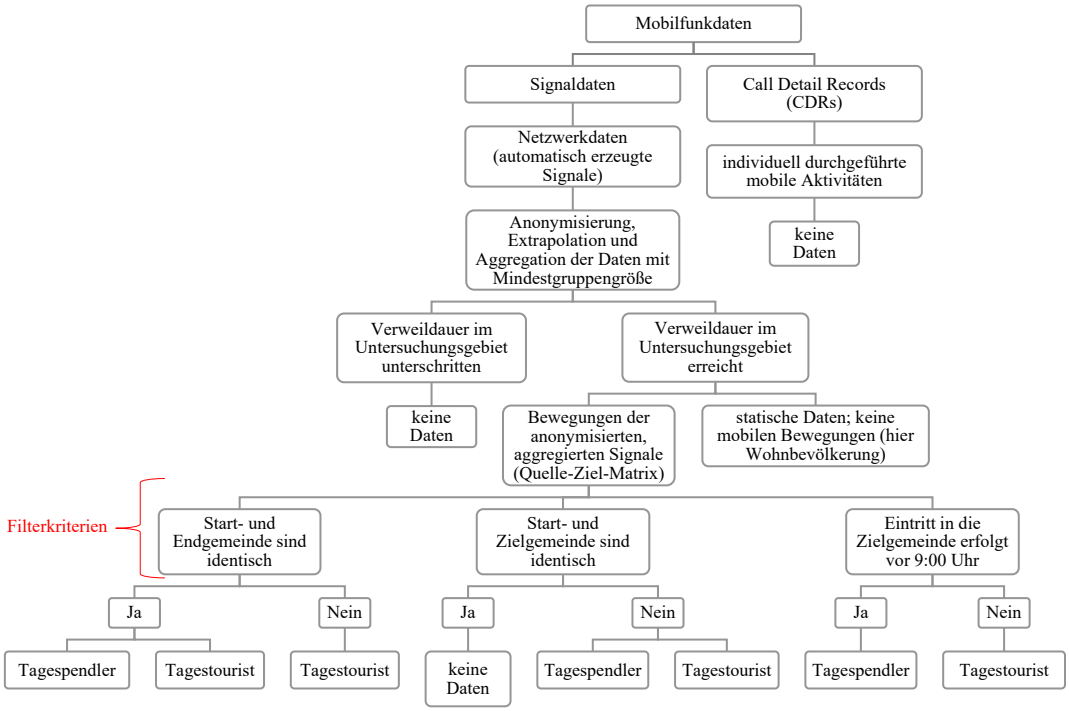



Abbildung 3.1: Schematische Darstellung der verwendeten Mobilfunkdaten sowie der umgesetzten Filterkriterien zur Eingrenzung der mobilen Bewegungsverflechtungen.

welches durch eine Mindestverweildauer des mobilen Endgerätes in einem Untersuchungsgebiet bedingt wird. Das mobile Endgerät darf dabei nicht ausgeschaltet oder im Flugmodus sein. Im Gegensatz zu CDR Daten sind bei den hier vorliegenden Signaldaten keine Informationen zur durchgeführten Aktivität, wie getätigte Anrufe, SMS, mobile Datennutzung o.ä., enthalten.

Die Daten enthalten weiterhin tägliche und extrapolierte Bewegungsverflechtungen für die Monate August, September und Oktober aus dem Jahr 2019 mit einer Verweildauer von zwei Stunden am Zielort, im Folgenden auch Untersuchungsgebiet genannt. Die Verweildauer gibt an, wie lange sich das mobile Endgerät durchschnittlich in einem Untersuchungsgebiet aufhalten muss, um als Aktivität gezählt zu werden. D.h., alle mobilen Aktivitäten, die weniger als zwei Stunden im Untersuchungsgebiet aktiv waren, sind in den vorliegenden Mobilfunkdaten nicht enthalten. Alle Aktivitäten, die länger als zwei Stunden im Untersuchungsgebiet aktiv waren, sind auch in den Daten erhalten. Abb. 3.1 stellt die beschriebene Datenlage schematisch dar.

Das Untersuchungsgebiet stellt in diesem Artikel generell eine Gemeinde[12] dar. Hierbei

[12] Städte sind inkludiert und fallen hier vereinfachend unter die Gebietsstruktur der Gemeinden.

werden Städte mit mehr als 100.000 Einwohnern[13] zusätzlich in Gitterzellen mit Gitterweiten von 250x250 m bis zu 4x4 km, abhängig vom Mobilfunknetz des Anbieters, unterteilt. Die aggregierten Mobilfunkdaten wurden weiterhin – aufgrund datenschutzrechtlicher Regelungen – erst ab einer Mindestgruppengröße von fünf mobilen Aktivitäten an das Statistische Bundesamt und IT.NRW übermittelt. Durch das Anonymisierungsverfahren des Datenanbieters können die Bewegungsverflechtungen darüber hinaus lediglich tageweise (innerhalb von 24 Stunden) nachvollzogen werden (Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, 2017), weshalb auch nachfolgend der Begriff „Tagespendler“ in den Mobilfunkdaten Verwendung findet. Eine Validierung der Bewegungsverflechtungen durch die Betrachtung längerer, ununterbrochener Zeitperioden ist daher nicht möglich. D.h., es kann nicht nachvollzogen werden, ob die identifizierten Bewegungen täglich bzw. regelmäßig stattfinden oder diese ggf. nur Ausnahmeerscheinungen sind. Auch erlaubt die Auswertung der anonymisierten mobilen Aktivitäten keinen direkten Rückschluss auf die Ursache der ermittelten Bewegung.

Um sicherzustellen, dass vor allem die Berufspendler in den Mobilfunkdaten enthalten sind, wurden weiterhin Bedingungen bzw. Filterkriterien formuliert, die sich an der zuvor erläuterten Definition eines Berufspendlers orientieren. Die mobilen Aktivitäten in den Mobilfunkdaten unterliegen somit folgenden Bedingungen (vgl. Abb. 3.1):

a) Die Start- und Endgemeinde sind identisch. Berufspendler starten an ihrem Wohnort und kehren üblicherweise an denselben zurück, weshalb das erste und das letzte Signal des Tages eines mobilen Endgerätes in der gleichen Gemeinde und damit dem potenziellen Wohnort verzeichnet werden müssen.

b) Die Start- und Zielgemeinde sind nicht identisch. Damit wird die Bedingung gestellt, dass die mobilen Aktivitäten in der Zielgemeinde bzw. dem Untersuchungsgebiet – demnach am potenziellen Arbeitsort – außerhalb der potenziellen Wohnsitzgemeinde stattfinden und sich folglich aus der Wohnsitzgemeinde entfernen müssen. Dadurch kann gewährleistet werden, dass nur übergemeindliche Bewegungen in den Mobilfunkdaten enthalten sind. Weiterhin lässt sich daraus interpretieren, dass eine Person mit einem mobilen Endgerät, die ihr erstes und letztes Mobilfunksignal des Tages außerhalb des Untersuchungsgebietes tätigt, im Untersuchungsgebiet arbeitet und nach der Pendlerdefinition im Untersuchungsgebiet als Einpendler angesehen wird.

c) Eintritt in die Zielgemeinde erfolgt vor 9:00 Uhr. Für eine trennschärfere Unterscheidung zwischen Tagespendlern und Bewegungen zu anderen Zwecken, wie bspw. Freizeit- oder

[13] Unter dem Begriff „Einwohner“ werden im Folgenden sowohl Einwohnerinnen als auch Einwohner verstanden. Aus Gründen der besseren Lesbarkeit wird im Text verallgemeinernd das generische Maskulinum verwendet.

touristische Aktivitäten, wird weiterhin angenommen, dass Berufspendler früher am Tag ihre Wohngemeinde verlassen als Nicht-Berufspendler.

Zusätzlich wird zur Bestimmung des Wohnortes behelfsmäßig die Vertragspostleitzahl neben dem ersten und letzten Signal des Tages durch den Datenanbieter verwendet, sofern diese zur Verfügung steht. Ebenfalls werden nur die nationalen SIM-Karten in den Daten beibehalten, um Verzerrungen durch touristische Aktivitäten von ausländischen Tagestouristen zu minimieren.

Insgesamt stellt die letzte Bedingung (c) der Filterkriterien jedoch eine besondere Schwachstelle in den Daten bzw. der schwierigen Datenspezifikation dar, da dadurch zwangsweise Schichtarbeitende, Teilzeitbeschäftigte oder auch mobile Arbeitskräfte, nicht oder nur marginal einbezogen werden.

Aufgrund der schwierigen Eingrenzung der Pendlerbewegungen in den vorliegenden Mobilfunkdaten finden sich daher neben den Tagespendlern folglich zwei weitere Personenkategorien, die Tagestouristen[14] und die Wohnbevölkerung, die im Wesentlichen die Bewegungsverflechtungen der Nicht-Berufspendler abdecken. Die sogenannten Tagestouristen weisen hierbei ein ähnliches Verhalten wie die Tagespendler in den Mobilfunkdaten auf und sind daher schwierig voneinander abzugrenzen, wie in Abb. 3.1 anhand der umgesetzten Filterkriterien offensichtlich wird. Die grundlegende Unterscheidung beider Personenkategorien ist das Ziel bzw. die Motivation der Bewegung, welche beim Tagestouristen auf Freizeit- oder touristische Aktivitäten zurückzuführen ist. Da der Zielort der mobilen Bewegungen allein jedoch nicht zu belastbaren Aussagen führt, werden weitere Unterscheidungsmerkmale der Personenkategorien eingeführt. Anders als der Tagespendler muss nach der Definition des Datenanbieters ein Tagestourist nicht wieder an seinen Wohnort zurückkehren (vgl. Abb. 3.1). Zudem tritt der Tagestourist Annahme bedingt erst nach 9:00 Uhr Lokalzeit in die Zielgemeinde ein. Dies trifft sicherlich auch auf Berufspendler zu, allerdings stellt diese Bedingung, wie in der schematischen Darstellung in Abb. 3.1 klargestellt wird, das einzige Filterkriterium dar, das konsequent zwischen Tagespendler und Tagestourist unterscheidet.

Weiterhin wird die Wohnbevölkerung in den Daten ausgewiesen, die jedoch keine Bewegung aufzeigt, da grundsätzlich nur der Wohnort als Untersuchungsgebiet betrachtet wird und folglich kein Wechsel der Gemeinde erfolgen darf. Dementsprechend werden sie in der folgenden Arbeit nicht weiter berücksichtigt.

Um anhand der gefilterten Mobilfunkdaten im nachfolgenden Abschnitt Vergleiche mit der Pendlerrechnung von IT.NRW des Jahres 2019 durchführen zu können, müssen die Daten zeitlich und räumlich weiter aufbereitet werden. Hierzu werden zum einen nur ferienfreie Werktage

[14] Unter dem Begriff „(Tages-)Tourist“ werden im Folgenden sowohl (Tages-)Touristinnen als auch (Tages-)Touristen verstanden. Aus Gründen der besseren Lesbarkeit wird im Text verallgemeinernd das generische Maskulinum verwendet.

aus den Monaten August, September und Oktober 2019 aus den Mobilfunkdaten verwendet, da Ferienzeiten und Feiertage die alltäglichen Muster des Pendlerverhaltens im Regelfall nicht adäquat wiedergeben. Zum anderen müssen die regionalen Gebiete in den Mobilfunkdaten denen der Pendlerrechnung angepasst werden, um unter anderem die Distanzen zwischen Start- und Zielgemeinde entsprechend der Pendlerrechnung zu berechnen. Das bedeutet, dass die Startorte der mobilen Bewegungen, welche in den Mobilfunkdaten auf Ebene der 5-stelligen Postleitzahlen vorliegen, den überdeckenden Gemeinden zugeordnet werden müssen.

Des Weiteren werden die täglichen Bewegungsverflechtungen zu einem Durchschnittswert zusammengefasst. So kann annäherungsweise gewährleistet werden, dass die regulären Pendlerbewegungen in den Mobilfunkdaten stärker einbezogen werden. Dafür werden zunächst alle Pendlerströme mit Zielorten auf Ebene der Gitterzelle – sofern vorhanden – pro Tag auf die zugehörige Gemeinde aufaggregiert und daraus ein Durchschnittswert der Bewegungsverflechtungen aller verbliebenen Auswertungstage für alle enthaltenen Kombinationen von Start- und Zielgemeinde gebildet.[15] Für die Angabe der Distanzen zwischen diesen Start- und Zielgemeinden in den Mobilfunkdaten wird die Entfernung der geografischen Koordinaten (ETRS89) der Gemeindemittelpunkte unter Verwendung des Satzes des Pythagoras in Kilometern berechnet.[16]

Für die Zusammenführung der aufbereiteten Mobilfunkdaten mit der Pendlerrechnung werden anschließend die einzelnen Start-Ziel-Verbindungen in beiden Datenquellen mit einer eindeutigen selbsterstellten Kennung versehen, um diese im nachfolgenden Abschn. 3.3.1 miteinander verknüpfen und vergleichen zu können.

## 3.3 Mobilfunkdaten in der amtlichen Pendlerrechnung

### 3.3.1 Vergleich der Pendlerbewegungen auf Basis von Mobilfunkdaten mit der amtlichen Pendlerrechnung

Bevor Möglichkeiten zur Anreicherung der amtlichen Pendlerrechnung mit den vorliegenden Mobilfunkdaten diskutiert werden können, müssen wir zunächst ermitteln, inwieweit die Quelle-Ziel-Matrizen der Mobilfunkdaten 2019 mit denen der Pendlerrechnung 2019 übereinstimmen.

[15] Eine nachträgliche Aufsummierung der Mobilfunkaktivitäten kann zu Mehrfachzählungen führen, da eine Deduplizierung der Mobilfunkaktivitäten nur auf Basis der ursprünglichen geografischen Ebene durchgeführt wurde. Für einen Abgleich mit der Pendlerrechnung ist es jedoch zwingend erforderlich, die Aktivitäten von der Ebene der Gitterzellen auf die Gemeindeebene zu aggregieren. Verzerrungen oder Unschärfen werden in den Bewertungen der Ergebnisse berücksichtigt.

[16] Die Distanz ergibt sich dabei aus: $Distanz_i = \sqrt{(dx)^2 + (dy)^2}$ wobei $dx = 111,3 * cos(lat) * (lon1 - lon2)$, $lat = \frac{(lat1+lat2)}{2*0,01745}$ und $dy = 111,3 + (lat1 - lat2)$. $Lat1$, $lat2$, $lon1$, $lon2$ entsprechen Breiten- und Längengrade jeder Start- und Zielgemeinde. Die Konstante 0,01745 ergibt sich aus der Umrechnung von Grad in Radiant (Bogenmaß) durch: $1^\circ = \frac{\pi}{180 rad}$. Der Abstand zwischen zwei Breitenkreisen in km wird durch die Konstante 111,3 und der variable Abstand zwischen zwei Längenkreisen durch $111,3 * cos(lat)$ angegeben (Kompf, 2020).

Durch eine räumliche Zuordnung aller Pendlerverflechtungen in beiden Datenquellen anhand der eindeutigen Kennung ist es möglich, all jene Pendlerbewegungen aus den Mobilfunkdaten zu extrahieren, die in der Pendlerrechnung von IT.NRW ausgewiesen werden und sie dementsprechend zu validieren.

Hierfür bedarf es einer geeigneten Mobilfunkdatenbasis. Da eine trennscharfe Abgrenzung der einzelnen Personengruppen und damit der Pendlerverflechtungen durch die zuvor in Abschn. 3.2.2 beschriebenen Filterkriterien jedoch nur approximativ möglich ist, stehen zur Abbildung der Pendlerverflechtungen zwei Optionen zur Verfügung. Entweder es werden nur die Tagespendler einbezogen, womit eine belastbare Datenbasis für die Darstellung der Pendlerverflechtungen der potenziellen Berufspendler aus den Mobilfunkdaten gebildet wird. Dadurch wird allerdings ein Informationsverlust an Bewegungsverflechtungen entstehen, da diese Daten durch die in Abschn. 3.2.2 diskutierten Filterkriterien sehr stark selektiert sind. Alternativ können alle Personenkategorien verwendet werden, das schließt die Tagespendler, die Tagestouristen sowie die Wohnbevölkerung ein, wodurch deutlich mehr Bewegungsverflechtungen bzw. mobile Aktivitäten in den Daten enthalten sind. In diesem Fall geht hingegen die Möglichkeit verloren, charakteristischere Aussagen zu den Berufspendlern zu treffen. Tab. 3.1 illustriert diesen Zielkonflikt anhand des Pearson-Korrelationskoeffizienten für die mit der Pendlerrechnung 2019 übereinstimmenden mobilen Pendlerströme auf Basis der Tagespendler sowie aller drei Personenkategorien. Der Pearson-Korrelationskoeffizient ermittelt den linearen Zusammenhang bzw. die Stärke des Zusammenhangs zwischen den gemeinsamen Pendlerströmen in den Mobilfunkdaten und der Pendlerrechnung.

Tabelle 3.1: Pearson-Korrelationskoeffizient (r) für die mit der Pendlerrechnung übereinstimmenden Pendlerströme unter Einbeziehung der Tagespendler sowie aller Personenkategorien in den Mobilfunkdaten wie auch die Summe ($\sum$) aller – von der Pendlerrechnung unabhängigen – Mobilfunkbewegungen.

| Übereinstimmende Pendlerströme | r | t-Wert | df | $Pr(>\|t\|)$ | $\sum$Mobilfunkbewegungen |
|---|---|---|---|---|---|
| Tagespendler | 0,846 | 233,6 | 21.698 | < 2.2e-16 | 28.509 |
| Alle Personenkategorien | 0,906 | 368,73 | 29.839 | < 2.2e-16 | 89.093 |

Im Ergebnis liegt die Korrelation der übereinstimmenden Pendlerströme beider Datenquellen, unter ausschließlichem Einbezug der Tagespendler, bei 0,85 (vgl. Tab. 3.1) und deutet damit einen sehr starken positiven linearen Zusammenhang zwischen den Pendlerströmen beider Datenquellen an. Weiterhin wird ersichtlich, dass die Einbeziehung aller Personenkategorien zu einer moderaten Zunahme der Korrelation führt und damit der zuvor diskutierte und erwartete höhere Informationsgehalt zum Tragen kommt. Dies wird neben dem Korrelationskoeffizienten

in Höhe von 0,91 auch bei der gestiegenen Anzahl an Freiheitsgraden (df) sichtbar.

Betrachten wir weiterhin die gesamten Mobilfunkbewegungen ($\sum$Mobilfunkbewegungen) in Tab. 3.1, ohne Verbindung zur Pendlerrechnung, stehen uns unter Einbezug aller Personenkategorien rund 89.000 Mobilfunkbewegungen zur Verfügung und mit den Tagespendlern nur rund 28.500 Mobilfunkbewegungen. Der Einbezug aller Personenkategorien beinhaltet offensichtlich Mobilitätsverflechtungen, die überwiegend nicht den Pendlerverflechtungen und stattdessen eindeutig touristischen Aktivitäten zuzuordnen sind, wie der Besuch von Naherholungsorten, Einkaufstandorten oder Veranstaltungen. Bei dieser Datenauswahl würde außerdem die Möglichkeit verloren gehen, weitere Berufspendlerströme aus Mobilfunkdaten herzuleiten, welche möglicherweise noch nicht in der Pendlerrechnung erfasst wurden, da Aussagen hierzu aufgrund der Vermischung mit anderweitigen Mobilitätsverflechtungen nicht mehr getroffen werden können. Daher werden beim Einbezug aller Personenkategorien keine gerechtfertigten oder belastbaren Erweiterungsmöglichkeiten der Pendlerrechnung gesehen, wie eine charakteristische kleinräumige Zielorts-Bestimmung der Berufspendler (siehe Abschn. 3.3.2).

Da das Ziel dieser Arbeit die fachliche Unterstützung der amtlichen Pendlerrechnung bzw. eine Prüfung der Möglichkeiten hierfür ist, sollen möglichst plausible Pendlerwege in den Daten ermittelt werden. Daher werden für die nachfolgenden Analysen ausschließlich die Tagespendler aus den Mobilfunkdaten verwendet.[17] Insgesamt erhalten wir somit 21.700 übereinstimmende Pendlerströme aus beiden Datenquellen durch die eindeutige Verknüpfung der jeweils gelisteten Start-Ziel-Verbindung. Zur Einordnung der Größenordnung liegen ca. 44.600[18] amtlich erfasste Pendlerbewegungen und insgesamt 28.509 Mobilfunkbewegungen (siehe $\sum$Mobilfunkbewegungen in Tab. 3.1) in der Analyse vor. Ungefähr 22.900 Pendlerverflechtungen aus der amtlichen Pendlerrechnung stimmen damit nicht mit den Mobilfunkdaten überein oder sind nicht in den Mobilfunkdaten enthalten.[19] Mit der ausschließlichen Nutzung der Tagespendler liegen in den Mobilfunkdaten andererseits, durch die Differenz aller Mobilfunkbewegungen der Tagespendler und der mit der Pendlerrechnung übereinstimmenden Bewegungen (vgl. Tab. 3.1), 6.809 potenzielle Pendlerverflechtungen vor, die bislang noch nicht

[17] Natürlich beinhalten die Quelle-Ziel-Matrizen der Mobilfunkdaten deutlich mehr Wege und damit Verflechtungen, die durch den Nicht-Einbezug weiterer Personenkategorien in diesem Artikel jedoch nicht weiter berücksichtigt werden können.

[18] Hierbei handelt es sich vereinfacht um insgesamt 44.559 Auspendlerströme sowie 44.659 Einpendlerströme aus der Pendlerrechnung von IT.NRW.

[19] Würde die Korrelation aller Pendlerwege aus der amtlichen Pendlerrechnung mit den Mobilfunkdaten berechnet werden, so würde die Korrelation von 0,85 aller übereinstimmenden Pendlerverflechtungen auf nur noch 0,76 bei den Auspendlerströmen und 0,648 bei den Einpendlerströmen sinken. Grund dafür ist, dass knapp 50% der amtlich ermittelten Pendlerverflechtungen nicht in den Mobilfunkdaten enthalten sind.

durch die amtliche Pendlerrechnung abgedeckt werden.[20]

Um ergänzend zu Tab. 3.1 unterschiedlich strukturierte Regionen und damit mögliche regionale Unterschiede beim Vergleich der Pendlerrechnung und der Mobilfunkdaten berücksichtigen zu können, werden in Abb. 3.2 **a**, **b** die Pearson-Korrelationskoeffizienten für die übereinstimmenden Ein- und Auspendlerströme, differenziert nach fünf Einwohnergrößenklassen, betrachtet. Die Einwohnergrößenklassen geben die Anzahl der Einwohner je Gemeinde in gleichgroßen Klassen wieder, beginnend bei Gemeinden mit weniger als 50.000 Einwohnern bis hin zu 500.000 Einwohnern aufwärts. Weiterhin wird die absolute Anzahl der Ein- und Auspendlerströme zwischen den Mobilfunkdaten 2019 und der Pendlerrechnung 2019 anhand der Einwohnergrößenklasse in einem Streudiagramm gegenübergestellt. Auf der x-Achse sind die absoluten Zahlen der Aus- und Einpendlerströme der Pendlerrechnung hinterlegt und auf der y-Achse die der Mobilfunkbewegungen der Tagespendler. Ebenfalls finden sich für jede Gegenüberstellung die Korrelationskoeffizienten je Einwohnergrößenklasse.

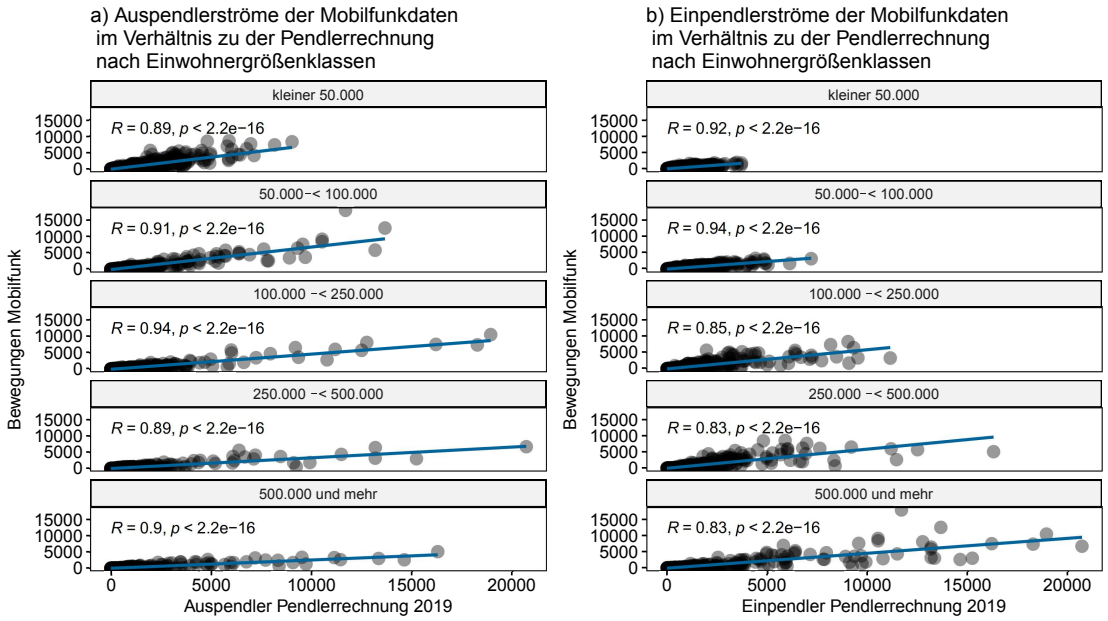



Abbildung 3.2: Korrelationsdiagramm der Bewegungsverflechtungen aus den Mobilfunkdaten 2019 und der Pendlerrechnung 2019 nach Auspendlerströmen (**a**) und Einpendlerströmen (**b**) unterteilt in fünf Einwohnergrößenklassen.

[20] Von diesen zusätzlichen Pendelbewegungen gelten 2.241 jedoch – gemäß der Pendlerrechnung – als nicht plausibel, da Entfernungen von über 80 km zurückgelegt wurden. Damit verbleiben 4.568 potenziell neue Pendlerwege aus den Mobilfunkdaten. Diese beinhalten vorwiegend zusätzliche und von der Anzahl geringere Pendlerverflechtungen (im Durchschnitt 22 Pendler pro Weg) von Städten und Gemeinden mit einer Einwohnerzahl unter 100.000 Einwohnern.

Abb. 3.2 **a**, **b** zeigen, dass die Korrelationen über alle Einwohnergrößenklassen hinweg einen sehr starken positiven linearen Zusammenhang sowohl bei den Ein- wie auch Auspendlerströmen aufweisen und daher auf ähnliche Intensitäten der Pendlerströme hindeuten. Trotz des insgesamt sehr starken positiven linearen Zusammenhangs beider Datenquellen in allen Einwohnergrößenklassen werden die absoluten Pendlerströme in den Mobilfunkdaten bei den Ein- und Auspendlerströmen im Vergleich zur Pendlerrechnung deutlich unterschätzt. Dies wird bereits an den unterschiedlichen x- und y-Achsenlängen ersichtlich.[21] Größere Gemeinden oder Städte mit einer Einwohnerzahl ab 250.000 Einwohnern aufwärts deuten bspw. den schwächsten linearen Zusammenhang mit Korrelationskoeffizienten unter 0,9 an. In diesen Fällen ist davon auszugehen, dass die Pendlerströme in größeren bzw. einwohnerreichen Gemeinden mit den Mobilfunkdaten tendenziell unterschätzt werden. Als Beispiel für die systematische Unterschätzung wird in Tab. 3.2 eine Gegenüberstellung der 10 größten (absoluten) Ein- und Auspendlerströme der Stadt Köln anhand der Anzahl von Ein- und Auspendlern nach der Pendlerrechnung 2019 und den Mobilfunkdaten 2019 betrachtet sowie die prozentuale Abweichung der gelisteten Mobilfunkdatenströme zur Pendlerrechnung. Die prozentuale Abweichung der Mobilfunkdatenströme richtet sich dabei nach den identischen Pendlerströmen der Pendlerrechnung. Letztere sind daher nicht zwangsläufig bzw. als Pendant in Tab. 3.2 aufgeführt.

Dabei werden in Tab. 3.2 exemplarisch zwei Kernergebnisse dieses Artikels verdeutlicht, die auch in anderen Pendlerverflechtungen NRWs auf Basis der Mobilfunkdaten vorzufinden sind. Zum einen wird in den Mobilfunkdaten eine andere Reihenfolge der größten Ein- und Auspendlerströme nach und aus Köln wiedergegeben als in der amtlichen Pendlerrechnung. Zum anderen werden alle Pendlerverflechtungen mit den Mobilfunkdaten – unter anderem mit Ausnahme der Stadt Pulheim – deutlich unterschätzt, wie auch die prozentuale Abweichung zur Pendlerrechnung in Tab. 3.2 verdeutlicht. Pulheim ist eine an Köln angrenzende Stadt mit weniger als 50.000 Einwohnern und der ersten Einwohnergrößenklasse zugeordnet, deren Einpendlerströme um mehr als 6.000 Einpendler überschätzt werden. Insgesamt wird jedoch die Anzahl der gesamten Pendlerströme in NRW mit den Mobilfunkdaten um ca. -49,79% unterschätzt, wie die absoluten Werte der Ein- und Auspendler in den Mobilfunkdaten in Tab. 3.2 zeigen.

Tab. 3.3 listet abschließend den Anteil der Gemeinden auf, deren Pendlerverflechtungen nach der Pendlerrechnung 2019 mit den vorliegenden Mobilfunkdaten über- oder unterschätzt werden. Dies wird, wie in Abb. 3.2, differenziert nach Einwohnergrößenklassen sowie durch eine Unterscheidung nach Ein- und Auspendlerströmen betrachtet. Grundsätzlich wird nochmals

[21] Die x-Achsenlänge beträgt maximal 20.000 gezählte Pendlerbewegungen und die y-Achsenlänge maximal 15.000 Bewegungen. Beide Achsenlängen resultieren aus den vorliegenden Pendlerströmen in den Mobilfunkdaten und der Pendlerrechnung.

Tabelle 3.2: Die 10 größten Ein- und Auspendlerströme in Köln nach der Pendlerrechnung 2019 und den Mobilfunkdaten 2019 sowie die prozentuale Abweichung der ermittelten Mobilfunkdatenströme zur Pendlerrechnung.

| **Pendlerrechnung 2019** | | **Mobilfunkdaten 2019** | | |
|---|---|---|---|---|
| **Auspendlerströme (Wohnort Köln)** | | | | |
| Arbeitsort | Anzahl Auspendler | Arbeitsort | Anzahl Auspendler | Abweichung zur Pendlerrechnung in % |
| Bonn, Stadt | 16.275 | Bonn, Stadt | 5.103 | -68,65 |
| Düsseldorf, Stadt | 14.608 | Leverkusen, Stadt | 3.292 | -70,41 |
| Leverkusen, Stadt | 11.125 | Düsseldorf, Stadt | 2.603 | -82,18 |
| Bergisch Gladbach, Stadt | 7.705 | Bergisch Gladbach, Stadt | 2.513 | -67,38 |
| Hürth, Stadt | 6.112 | Troisdorf, Stadt | 1.869 | -56,26 |
| Frechen, Stadt | 5.036 | Wesseling, Stadt | 1.805 | -23,32 |
| Troisdorf, Stadt | 4.273 | Hürth, Stadt | 1.726 | -71,76 |
| Pulheim, Stadt | 3.926 | Brühl, Stadt | 1.544 | -46,41 |
| Brühl, Stadt | 2.881 | Frechen, Stadt | 1.325 | -73,69 |
| Kerpen, Stadt | 2.818 | Dormagen, Stadt | 1.091 | -57,93 |
| **Einpendlerströme (Arbeitsort Köln)** | | | | |
| Wohnort | Anzahl Einpendler | Wohnort | Anzahl Einpendler | Abweichung zur Pendlerrechnung in % |
| Bergisch Gladbach, Stadt | 18.245 | Pulheim, Stadt | 17.952 | 53,66 |
| Leverkusen, Stadt | 16.181 | Hürth, Stadt | 12.557 | -7,98 |
| Bonn, Stadt | 15.215 | Frechen, Stadt | 9.145 | -13,05 |
| Hürth, Stadt | 13.646 | Leverkusen, Stadt | 7.511 | -53,58 |
| Pulheim, Stadt | 11.683 | Bergisch Gladbach, Stadt | 7.365 | -59,63 |
| Frechen, Stadt | 10.517 | Niederkassel, Stadt | 6.968 | 20,22 |
| Düsseldorf, Stadt | 9.717 | Erftstadt, Stadt | 4.807 | -24,77 |
| Kerpen, Stadt | 9.681 | Brühl, Stadt | 4.477 | -29,18 |
| Troisdorf, Stadt | 8.890 | Bornheim, Stadt | 4.321 | -12,69 |
| Bergheim, Stadt | 7.783 | Dormagen, Stadt | 4.057 | -29,04 |

deutlich, dass mit den Mobilfunkdaten besonders einwohnerarme Gemeinden überschätzt werden (vgl. Tab. 3.3). Bei ca. 41-42% der enthaltenen Gemeinden mit weniger als 50.000 Einwohnern werden die Ein- und Auspendlerströme durch die verfügbaren Mobilfunkdaten überschätzt. D.h., die absoluten Mobilfunkdatenströme sind größer als die der Pendlerrechnung 2019, wie am Beispiel der Stadt Pulheim bereits exemplarisch abgeleitet wurde (vgl. Tab. 3.2). Je größer die Einwohnergrößenklasse jedoch wird, desto weniger werden die Ein- und Auspendlerströme mit den Mobilfunkdaten überschätzt und tendieren eher dazu diese zu unterschätzen. Insbesondere bei ca. 97% der Gemeinden mit 500.000 Einwohnern und mehr werden die Einpendlerströme

unterschätzt sowie bei 93% dieser Gemeinden hinsichtlich der Auspendlerströme. Der Anteil an Gemeinden, deren Pendlerströme mit den Mobilfunkdaten identisch geschätzt werden können, ist dagegen vernachlässigbar gering. Zusammenfassend kann somit ein Stadt-Land- oder auch Stadt-Umland-Gefälle in den Mobilfunkdaten impliziert werden, bei dem die vorliegenden Mobilfunkdaten die Bewegungen in urbanen oder einwohnerreichen Gebieten deutlich unterschätzen und jene in ländlicheren oder einwohnerärmeren Gebieten tendenziell überschätzen.[22] Von einer zeitlich schnelleren Abbildung von Pendlerströmen mittels Mobilfunkdaten als experimentelle Pendlerrechnung muss aufgrund der deutlichen Unter- bzw. Überschätzung der Pendlerverflechtungen daher abgesehen werden.

Tabelle 3.3: Anteil der Gemeinden, deren Pendlerströme mit den Mobilfunkdaten über-, unter- oder gleichgeschätzt werden, aufgeteilt nach Einwohnergrößenklassen und Ein- und Auspendlerströmen im Vergleich zur Pendlerrechnung 2019.

| | **Schätzung der Auspendlerströme pro Gemeinde in %** | | | **Schätzung der Einpendlerströme pro Gemeinde in %** | | |
|---|---|---|---|---|---|---|
| Einwohnergrößenklasse | überschätzt | unterschätzt | identisch | überschätzt | unterschätzt | identisch |
| kleiner 50.000 | 41,30 | 56,22 | 2,47 | 41,81 | 55,54 | 2,65 |
| 50.000 $\leq$ 100.00 | 29,87 | 68,39 | 1,74 | 38,29 | 59,91 | 1,81 |
| 100.000 $\leq$ 250.000 | 20,38 | 78,03 | 1,58 | 14,74 | 84,10 | 1,12 |
| 250.000 $\leq$ 500.00 | 11,82 | 87,53 | 0,65 | 7,09 | 92,37 | 0,05 |
| 500.000 und mehr | 5,83 | 93,20 | 0,97 | 2,23 | 97,35 | 0,04 |

### 3.3.2 Kleinräumige Pendlerbewegungen in Städten – eine erweiterte Zielorts-Bestimmung

Die Abgleiche beider Datenquellen machen deutlich, dass die vorliegenden Mobilfunkdaten grundsätzlich das Potenzial haben, die Pendlerverflechtungen der amtlichen Pendlerrechnung abzubilden und zu unterstützen. Eine Erweiterungsmöglichkeit der Pendlerrechnung findet sich daher in der kleinräumigen Darstellung der Pendlerverflechtungen unterhalb der Gemeindeebene, auf der die amtliche Pendlerrechnung die Ergebnisse aktuell maximal publizieren kann. Bislang liegen keine Informationen zu innerstädtischen Verflechtungen oder besonders stark frequentierten kleinräumigen Zielorten vor. Solche räumlich differenzierten Betrachtungsmöglichkeiten liefern Informationen zur Lokalisierung oder Abgrenzung von Arbeits- und Wohngebieten und produzieren eine wertvolle Grundlage zur Infrastrukturnutzung sowie deren Planung und Ausbau. Darüber hinaus können die Pendlerverflechtungen geografisch feiner betrachtet werden.

[22] Hierbei wird die Anzahl der Mobilfunkdatenströme, die kleiner als die der Pendlerrechnung sind, um ca. -56% unterschätzt. Dagegen wird die Anzahl der Mobilfunkdatenströme, die größer als die der Pendlerrechnung sind, um ca. 42% überschätzt.

Die zuvor durchgeführten Korrelationsanalysen unterstützen dabei eine für die kleinräumige Darstellung notwendige Bedingung, bei der sich die zielgerichteten Bewegungen in den Mobilfunkdaten sowie der Pendlerrechnung sehr stark ähneln müssen. Daher kann weiterhin angenommen werden, dass die Mobilfunkdaten, die sich auf regional größeren Gebieten mit der Pendlerrechnung stark gleichen, ebenfalls aufschlussreiche Informationen zu kleinräumigen Gebieten liefern und damit eine potenziell wertvolle Erweiterung der amtlichen Pendlerrechnung bieten können. Insbesondere soll daher analysiert werden, wohin die Berufspendler pendeln und wie stark die Zielorte frequentiert werden, wenn sich diverse Arbeitgeber in der Zielgemeinde befinden.

Da die Mobilfunkdaten den Zielort bzw. den potenziellen Arbeitsort für Städte mit mehr als 100.000 Einwohnern auf Ebene der Gitterzellen mit Gitterweiten ab 250x250 m bis hin zu 4x4 km beinhalten, wird nachfolgend die Anzahl der potenziellen Einpendler je Zielort bzw. vermeintlichen Arbeitsort der Berufspendler aufsummiert und zusammengefasst dargestellt. Da nur die Zielorte der Bewegungsströme in den Mobilfunkdaten auf Ebene der Gitterzellen zur Verfügung stehen und nicht der Wohnort bzw. die Startgemeinde, werden im Folgenden nur die Einzugsgebiete der Einpendler in den Mobilfunkdaten betrachtet. Wir betrachten somit in der kleinräumigen Zielorts-Bestimmung die frequentierten Zielorte aller Einpendler und lassen dabei die Herkunft oder auch Quelle der Bewegungen außer Acht. Hierbei fließen dementsprechend nur die mit der Pendlerrechnung übereinstimmenden Mobilfunkbewegungen ein.

Weiterhin wurde im vorherigen Abschnitt bereits aufgezeigt, dass die Mobilfunkdaten die Pendlerströme in einwohnerreicheren Gemeinden deutlich unterschätzen. Daher fallen auch die absoluten Zahlen der Pendlerbewegungen in den kleinräumigen Zielorten tendenziell zu gering aus, weshalb der Fokus vorrangig auf die Lokalisierung der Zielorte und die regional bzw. urban stärksten Einzugsgebiete der Berufspendler gerichtet wird.

Zur Veranschaulichung der Kernergebnisse greifen wir hierzu auf Beispiele in Abb. 3.3 zurück, die aus der zugehörigen interaktiven Karte (Onlinematerial 1) entnommen wurden.[23] Die Darstellung visualisiert die Einzugsgebiete der ermittelten Einpendler auf Ebene der unterschiedlich großen Gitterzellen. Da sich die Größe der Gitterzellen an dem Mobilfunknetz des Anbieters orientiert und damit auch an der Dichte der Mobilfunknutzenden im Untersuchungsgebiet, sind Stadtzentren bzw. Innenstädte entsprechend mit kleinen Gitterzellen ausgelegt und weniger dicht besiedelte städtische Randbezirke mit größeren Gitterweiten. Damit wird es bereits möglich die Stadtzentren anhand der Gitterweiten zu lokalisieren. Um jedoch auf die durch die Berufspendler stärker frequentierten Einzugs- bzw. Arbeitsgebiete schließen zu können, wird

[23] Interessierte Leserinnen und Leser sind für eine vollumfassende visuelle sowie individuelle Betrachtung der Zielorte auf die zum Download verfügbare interaktive Karte „Onlinematerial1.html“ verwiesen. Bei der Karte ist zu beachten, dass hier ausschließlich die Einzugsgebiete auf Ebene der Gitterzellen visualisiert werden.

die Anzahl der Einpendler je Gitterzelle farblich hervorgehoben. Je dunkler die Gitterzelle schattiert ist,[24] desto höher ist die Anzahl der einpendelnden Berufspendler aus den Mobilfunkdaten in der entsprechenden Gitterzelle. Für die Verortung der frequentierten Gitterzellen und Interpretation derselben sind zusätzliche Geodaten von OpenStreetMap[25] in den Karten hinterlegt.

Die Ermittlung und Visualisierung der kleinräumigen Mobilfunkbewegungen der potenziellen Berufseinpendler offenbaren vier charakteristische Zielorte, die sich teilweise klar den Berufspendlern zuordnen lassen sowie andere, bei denen es sich vermutlich um keine Arbeitsstätte handelt. Die Beispiele sind in Abb. 3.3 hinterlegt.

Einen großen Einfluss auf die Zielorte der kleinräumigen Bewegungsverflechtungen in den Mobilfunkdaten haben Gewerbe- und Industriestandorte. Unter anderem bieten Chemieparks[26], bspw. in Leverkusen oder Krefeld, zahlreiche Arbeitsplatzangebote in den Regionen, die auch in den Zielorten der Mobilfunkdaten stark frequentiert werden (siehe Abb. 3.3 **c**). Zudem bilden Gewerbegebiete, wie die Duisburg-Ruhrorter Häfen, Kernpfeiler der Arbeitsmarktregion in Städten wie Duisburg (siehe Abb. 3.3 **b**). Der Mediapark in Köln, ein für Medienunternehmen konzipierter Gewerbepark (siehe Abb. 3.3 **a**), oder das Vallourec Deutschland, ein Walzwerk in Düsseldorf-Rath, sind genauso stark frequentierte Einzugsgebiete der Berufspendler, um nur einige Beispiele zu nennen.

Neben den offensichtlichen Produktions- oder Gewerbestätten fallen Stadtzentren, Altstädte oder Innenstädte sowie im Regelfall der dazugehörige Hauptbahnhof, wie exemplarisch in Abb. 3.3 **a**, **b** dargestellt, besonders stark auf. Sie stechen in jeder Stadt in NRW mit einer hohen Zahl an Einpendlern hervor. Stadtzentren bieten zahlreiche Arbeitsmöglichkeiten in diversen Dienstleistungen, in administrativen Tätigkeiten in der kommunalen Verwaltung o.ä., die mitunter historisch bedingt aus der Altstadt bzw. Innenstadt größerer Städte entstanden sind. Natürlich ist auch hier hervorzuheben, dass nicht ausgeschlossen werden kann, dass sich neben den gewünschten Berufspendlern auch andere freizeitbedingte mobile Bewegungen, wie bspw. die der Einkaufstouristen, in den ausgewählten Daten befinden. Durch die Filtersetzung, dass ein Eintritt der mobilen Aktivitäten in den Zielort vor 9 Uhr erfolgen muss, kann jedoch angenommen werden, dass ein Großteil dieser unerwünschten Bewegungsverflechtungen rausgefiltert wurde.

Darüber hinaus bilden Messegelände einen weiteren Zielort potenzieller Berufspendler in den Mobilfunkdaten ab. Exemplarisch hierzu treten die Gitterzellen in Abb. 3.3 **a** besonders stark hervor, die das Messegelände bei Köln/Messe Deutz oder die Messe Düsseldorf beinhalten. Auch beim Messebesuch muss wieder zwischen Arbeitskräften und Messebesuchern ohne Arbeitsauftrag unterschieden werden.

[24] In der Farbabbildung weist die rote Schattierung auf eine höhere Anzahl an Einpendlern hin.

[25] ©OpenStreetMap-Mitwirkende: Basiskarte und Daten von OpenStreetMap und OpenStreetMap Foundation in Abb. 3.3 und 3.4 und Onlinematerial 1.

[26] Auch als Chempark, früher als Bayerwerk, bekannt.

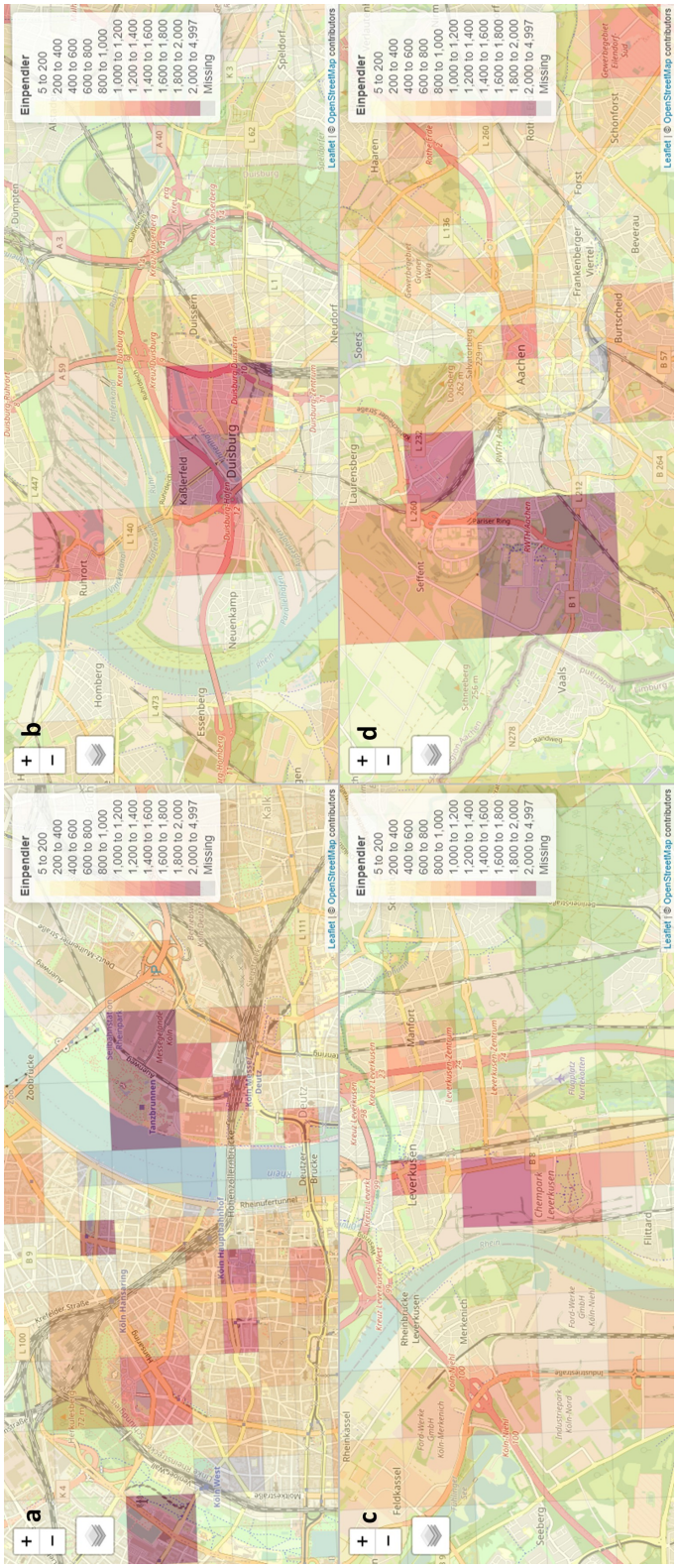



Abbildung 3.3: Frequentierte kleinräumige Zielorte der Mobilfunkverflechtungen der ermittelten potenziellen Berufseinpendler. (**a**) Köln – Innenstadt/Hauptbahnhof, Mediapark und Messegelände. (**b**) Duisburg – Duisburg-Ruhrorter Häfen. (**c**) Leverkusen – Stadtzentrum und Chempark. (**d**) Aachen – Innenstadt und RWTH Aachen.

Den dritten charakteristischen Zielort stellen Kliniken, unter anderem auch Universitätskliniken, dar, die eine große Anzahl von Arbeitsplätzen stellen. Bei den Kliniken treten gleichsam dieselben Herausforderungen in den Mobilfunkdaten auf wie in der Messewirtschaft oder in den Innenstädten. Zunächst können wir nicht sagen, ob es sich ausschließlich um Arbeitskräfte oder auch um Besuchende in den Daten handelt, die vor 9 Uhr eine Einrichtung bzw. ein Gelände aufsuchen. Zudem stellen Kliniken klassische Schichtbetriebe dar. Das bedeutet, dass mit dem 9 Uhr Filter in den Mobilfunkdaten im Sinne der Schichtarbeit nur die Arbeitskräfte der Frühschicht erfasst werden können. Die Spät- und Nachtschicht entfällt hiermit komplett, so dass davon auszugehen ist, dass die absolute Anzahl der Berufseinpendler in diesen Zellen unterschätzt wird. Zudem stellen unter anderem auch Universitätskliniken durch die Studierenden vor Ort eine weitere exemplarische Herausforderung in der zielorientierten Aufbereitung der Mobilfunkdaten dar, welche nachfolgend anhand des letzten charakteristischen Zielortes erörtert wird.

Bildungseinrichtungen wie allgemeinbildende Schulen oder Universitäten sind in den Mobilfunkdaten besonders stark frequentiert, wie bspw. die Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen (siehe Abb. 3.3 **d**) oder die Ruhr-Universität Bochum. Diese treten unter anderem stärker hervor, da ihre Fakultäten durch den Status einer Campus-Universität zentral an einem Ort vorzufinden sind und nicht – wie teilweise bei anderen Hochschulen – diffuser in der Stadt verteilt sind. Bei den Bildungseinrichtungen ergibt sich nun das Problem, dass offensichtlich Studierende sowie Schülerinnen und Schüler in den Daten enthalten sind. Laut Definition gehören sie nicht den Berufspendlern an, sondern werden separat als Bildungspendler ausgewiesen. Durch den Eintritt der mobilen Aktivitäten in den Zielort vor 9 Uhr und bspw. dem Beginn der ersten Unterrichtsstunde um 8 Uhr provozieren wir einen deutlich stärkeren Einbezug von Bildungspendlern mit den hier verwendeten Filterkriterien in den Mobilfunkdaten. Auch bei den Universitätskliniken ist es problematisch die Bildungspendler der angegliederten medizinischen Fakultät der Universität von den Berufspendlern des Klinikums zu unterscheiden. Das hängt mitunter auch mit den unterschiedlichen Gitterweiten der Gitterzellen zusammen, die keine räumlich scharfe Trennung ermöglichen. Insgesamt bedeutet das, dass die gefilterten Pendlerverflechtungen in den Mobilfunkdaten neben den potenziellen Berufseinpendlern auch die Bildungspendler beinhalten. Da die amtliche Pendlerrechnung keine Bildungspendler aufgrund von Dateninkonsistenzen ausgibt, wäre die Ermittlung der Bewegungsverflechtungen der Bildungspendler eine zusätzliche Erweiterungsmöglichkeit der Pendlerrechnung, die in weiterer Forschungsarbeit mit den Mobilfunkdaten aufgegriffen werden könnte.

Was alle charakteristischen Zielorte letztlich gemeinsam haben, ist die sehr gute Erreichbarkeit der kleinräumigen Zielorte durch das Straßen- und Schienennetz in NRW. Jede dunkel schattierte Gitterzelle enthält eine Anbindung an eine Autobahn, Bundesstraße oder einen Bahnhof

oder liegt bereits in der sehr gut angebundenen Innenstadt.[27] Eine gute räumliche Erreichbarkeit von Arbeitsplätzen bedingt somit maßgeblich die Bildung oder Verlagerung von Arbeitsmärkten. Weiterhin wird an den Beispielen in Abb. 3.3 deutlich, dass innerhalb der Städte durchaus sehr große innerstädtische Unterschiede hinsichtlich ihrer Frequentierung bestehen können und damit auch die Infrastruktur einer Stadt unterschiedlich stark belastet wird.

Als letztes Beispiel wird eine mögliche Konsequenz betrachtet, die entsteht, sofern die Datengrundlage zu viele nicht definierbare Bewegungsverflechtungen enthält. Damit greifen wir noch einmal die in Abschn. 3.3.1 geführte Diskussion hinsichtlich der für diesen Artikel passenden Personenkategorien auf.

Abb. 3.4 stellt das bekannte Fußballstadion in Dortmund, den Signal Iduna Park, sowie die daran angrenzenden Westfalenhallen dar. Der Signal Iduna Park bietet ca. 81.000 Plätze für Zuschauende von Fußballspielen. Die benachbarten Westfallenhallen sind ein Messe-, Kongress- und Veranstaltungszentrum und haben eine Kapazität von 15.400 Plätzen. Beides zusammengenommen stellen sie daher besonders Zielorte für touristische bzw. freizeitliche Aktivitäten und weniger für Berufspendler dar, wie auch Abb. 3.4 verdeutlicht. In Abb. 3.4 **a** sind die Summen der Einpendler unter Einbezug der Tagespendler abgebildet und in Abb. 3.4 **b** die Summen der Einpendler unter Einbezug aller Personenkategorien, einschließlich der Tagestouristen. Den Spitzenwert der Skala mit einem Besucherniveau von durchschnittlich 9.798[28] Einpendlern ist am Signal Iduna Park durch den Einbezug aller in den Mobilfunkdaten befindlichen Bewegungsverflechtungen erreicht, während es unter Einbezug der Tagespendler nur 978 gezählte Einpendler sind. Findet ein Fußballspiel oder bspw. ein Konzert an einem in den Daten enthaltenen Werktag statt, wird durch die sehr hohe Besucheranzahl das durchschnittliche Besucherniveau in den jeweiligen Gitterzellen, wie in Abb. 3.4 **b**, entsprechend angehoben.

Dieses Phänomen wird bei allen Bundesligastadien in NRW sichtbar und liefert insofern gegensätzliche Aussagen zu der Zielsetzung dieser Arbeit. Damit ist ein Einbezug aller Daten für die Erweiterung der amtlichen Pendlerrechnung bewiesenermaßen nicht geeignet. Dieses Beispiel zeigt eindrücklich, wie essenziell eine nicht wohldurchdachte Datengrundlage die Ergebnisse hinsichtlich der Messung von Berufspendlern beeinflussen oder verzerren kann. Ähnliche aber weit weniger drastische Effekte werden bei Naherholungsgebieten wie Naturparks, Halden oder großen Einkaufszentren, wie bspw. das CentrO in Oberhausen, sichtbar.

[27] Siehe hierzu auch die Beispiele in Abb. 3.3 oder die zum Download verfügbare interaktive Karte „Onlinematerial1.html“.

[28] Es wird ein Durchschnittswert aller Aktivitäten über die verbliebenen Auswertungstage aus drei Monaten gebildet, wobei Wochenenden und Feiertage nicht enthalten sind.


Abbildung 3.4: Anzahl der Einpendler aus den Mobilfunkdaten im Dortmunder Fußballstadion Signal Iduna Park und den angrenzenden Westfallenhallen unter Einbezug der Tagespendler (**a**) und aller Personenkategorien (**b**).

### 3.3.3 Zusammenhänge zwischen Berufspendlern nach Beschäftigungsumfang und Mobilfunkdaten

Die amtliche Pendlerrechnung gibt neben der Pendlerart auch Informationen zum Beschäftigungsumfang der Berufspendler wieder. Der Beschäftigungsumfang ermöglicht Angaben zum

Umfang der erbrachten Arbeitsleistung der Beschäftigten. Diese werden nachfolgend in Teilzeit- und Vollzeitbeschäftigung sowie die zusammengefasste Beschäftigung unterschieden. Unter Teilzeitbeschäftigung werden nach § 2 Teilzeit- und Befristungsgesetz (TzBfG) Arbeitnehmende verstanden, deren regelmäßige Wochenarbeitszeit kürzer ist als die der vergleichbaren vollzeitbeschäftigten Arbeitnehmenden. Die in Teilzeit zu leistende Arbeitszeit kann jedoch sehr unterschiedlich gestaltet sein und sehr flexibel eingeteilt werden und zählt theoretisch bereits bei weniger als 40 Wochenarbeitsstunden als solche. Daher ergibt sich die Fragestellung, inwieweit die vorliegenden Mobilfunkdaten diese beiden allgemeinen Arbeitszeitmodelle beinhalten und abbilden können. Insbesondere die eingesetzte zweistündige Verweildauer der mobilen Aktivitäten in einem Untersuchungsgebiet könnte zu kurz erscheinen, um nur Berufspendler aus den Mobilfunkdaten zu extrahieren. Daher betrachten wir im Folgenden weitere Zusammenhänge der Mobilfunkdaten mit der Pendlerrechnung, ohne dabei auf die konkreten Pendlerverflechtungen einzugehen. Stattdessen wird die Verweildauer in den Mobilfunkdaten mit dem Beschäftigungsumfang der Berufspendler aus der Pendlerrechnung in Zusammenhang gebracht. Anhand dessen soll überprüft werden, inwieweit die gegebene Verweildauer geeignet ist, um die Berufspendler insgesamt sowie unterteilt nach Beschäftigungsumfang darzustellen.

Im Vergleich zu den Pendlerverflechtungen werden die Berufspendler nach Beschäftigungsumfang in der amtlichen Pendlerrechnung nicht in Bewegungen bzw. Verflechtungen ausgedrückt, sondern als Jahresdurchschnitt pro Gemeinde angegeben. Selbige Auflösung findet sich auch in den Mobilfunkdaten wieder. Es werden die Tagessummen der mobilen Aktivitäten im Untersuchungsgebiet und damit dem potenziellen Arbeitsort wiedergegeben. Zusätzlich liegen in den Mobilfunkdaten die durchschnittlichen Verweilzeiten aller gruppierten mobilen Aktivitäten in einem Halbstundentakt, von einer halben Stunde bis hin zu 23 Stunden Verweilzeit, im Untersuchungsgebiet vor. Diese Information bietet uns die Möglichkeit, die Verweildauer in den Mobilfunkdaten in Zusammenhang mit dem Beschäftigungsumfang der Berufspendler zu setzen und mögliche Gemeinsamkeiten beider Datenquellen sowie Alternativen aufzuzeigen.

Für die Ermittlung möglicher Zusammenhänge werden erneut die Korrelationen zwischen dem Beschäftigungsumfang und der Pendlerart aus der amtlichen Pendlerrechnung zu den Tagessummen in den Mobilfunkdaten in Abhängigkeit von der halbstündigen Verweildauer innerhalb der Mobilfunkdaten berechnet. Damit soll zum einen analysiert werden, inwieweit die zweistündige Verweildauer in den Daten gerechtfertigt ist oder ob eine andere Verweilzeit in den Mobilfunkdaten den Beschäftigungsumfang besser abbilden kann.

Abb. 3.5 stellt die Korrelationen zwischen Beschäftigungsumfang und Pendlerart aus der amtlichen Pendlerrechnung zu den Tagessummen aus den Mobilfunkdaten in einem Netzdiagramm dar. Die Besonderheit in den berechneten Korrelationen liegt darin, dass diese in

Abhängigkeit von der halbstündigen Verweildauer in den Mobilfunkdaten gebildet wurden und damit pro Verweildauer ein Korrelationswert ausgegeben wird. Die gepunkteten oder durchzogenen Datenreihen geben die Pearson-Korrelationskoeffizienten der Aus- und Einpendler nach Vollzeitbeschäftigung, Teilzeitbeschäftigung sowie insgesamt wieder. Das Hauptgitternetz beinhaltet die Korrelationskoeffizienten in Bezug zu den halbstündigen Verweilzeiten als Rubrikenachsenbeschriftung, beginnend bei 0,5-1 Stunden bis hin zu 22,5-23 Stunden Verweildauer. Hierbei gilt, dass je weiter außen am Rand des Hauptgitternetzes die Linien der sechs Datenreihen verlaufen, desto höher ist die Korrelation zwischen den eingeordneten Pendlern der Pendlerrechnung und den Mobilfunkdaten. Anhand der Kreisverläufe können sodann Rückschlüsse auf die Informationsgüte der einzelnen Verweilzeiten geschlossen werden.

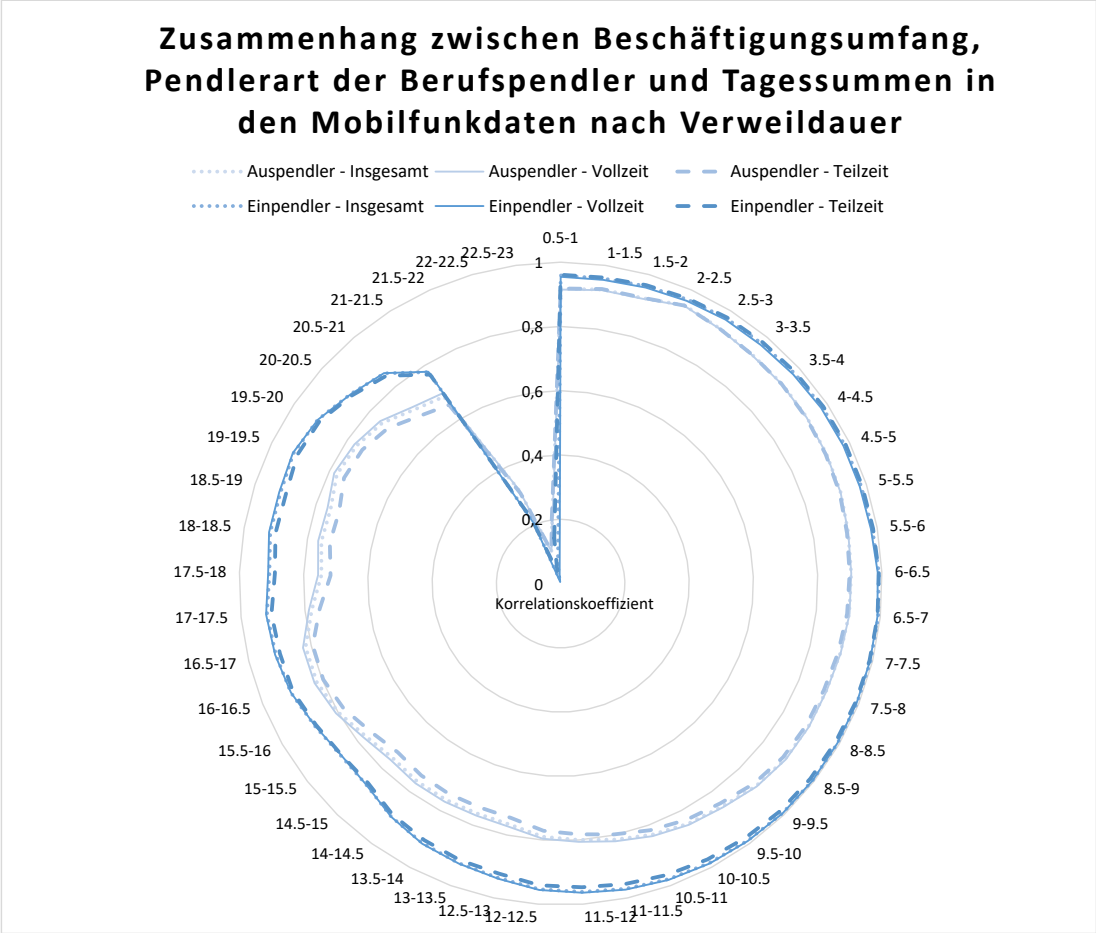



Abbildung 3.5: Korrelationsdiagramm des Beschäftigungsumfangs und der Pendlerart der Berufspendler aus der amtlichen Pendlerrechnung und den Tagessummen in den Mobilfunkdaten in Abhängigkeit von der halbstündigen Verweildauer innerhalb der Mobilfunkdaten.

Zunächst kann anhand von Abb. 3.5 festgestellt werden, dass grundsätzlich zwischen der Anzahl der Pendler nach Beschäftigungsumfang sowie der Pendlerart aus der Pendlerrechnung und den Tagessummen der Mobilfunkdaten eine sehr hohe Korrelation besteht. Insgesamt werden jedoch Unterschiede bei der Differenzierung der Korrelationen nach der Pendlerart offensichtlich, genauer zwischen Ein- und Auspendlern.

Die vorliegenden Mobilfunkdaten sind mit einer Verweildauer von zwei Stunden aufbereitet worden. Aus Abb. 3.5 wird nun ersichtlich, dass sich diese Verweildauer für die ausschließliche Abbildung der vollzeit- oder teilzeitbeschäftigten Einpendler weniger eignet. Stattdessen offenbaren die Korrelationskoeffizienten marginale Unterschiede in der Verweildauer nach Teilzeit- und Vollzeitbeschäftigung, die insbesondere bei längeren Verweildauern sichtbar werden. Bei den Einpendlern würde die höchste Korrelation (zwischen Einpendlern nach Beschäftigungsumfang aus der Pendlerrechnung und den Tagessummen der Mobilfunkdaten) und damit die optimale Verweilzeit zur Abbildung der Teilzeitbeschäftigten (*Einpendler - Teilzeit*) in den Mobilfunkdaten bei einer Verweildauer von 6,5-7 Stunden liegen und bei den Vollzeitbeschäftigten (*Einpendler - Vollzeit*) bei 8,5-9 Stunden (siehe Rubrikenachsenbeschriftung in Abb. 3.5).

Im Gegensatz dazu weisen die Auspendler einen anderen Trend auf. Die Korrelationen der Auspendler nach dem Beschäftigungsumfang sind insgesamt niedriger als die der Einpendler, auch wenn sie für sich genommen mit Korrelationskoeffizienten um 0,8 ebenfalls sehr hoch sind (siehe Abb. 3.5). Hier zeigt sich jedoch, dass eine niedrige Verweildauer von 2-2,5 Stunden die höchste Korrelation zwischen Beschäftigungsumfang der Auspendler zu den Mobilfunkdaten verursacht. Das ist dahingehend nicht verwunderlich, da die Tagessummen in den Mobilfunkdaten die Hauptaktivitäten am Tag je Untersuchungsgebiet und damit die durchschnittlichen Aktivitäten am potenziellen Arbeitsort wiedergeben. Damit werden nicht die durchschnittlichen Aktivitäten am potenziellen Wohnort angesprochen, was wiederum für die Ermittlung der Auspendler erforderlich wäre. D.h., die Tagessummen in den Mobilfunkdaten sind theoretisch nicht geeignet die Auspendler für die Gemeinden wiederzugeben. Insgesamt lässt sich die Fragestellung in Bezug auf die Zusammenhänge zwischen Beschäftigungsumgang und Pendlerart für die Einpendler beantworten, wobei die Ergebnisse für die Auspendler mit Vorsicht betrachtet werden müssen.

## 3.4 Diskussion einer alternativen Mobilfunkdatenaufbereitung

### 3.4.1 Einflüsse auf die Pendlerbewegungen in den Mobilfunkdaten

Die vorangegangenen Betrachtungen aus Abschn. 3.3 zu den Vergleichen mit der Pendlerrechnung, zur kleinräumigen Zielorts-Bestimmung sowie den Zusammenhängen mit dem Beschäf-

tigungsumfang geben Grund zur Annahme, dass die Mobilfunkdaten diversen Einflüssen unterliegen, welche die Ergebnisse maßgeblich beeinflussen. Neben offensichtlichen methodischen Aspekten, wie die Verteilung und Extrapolation der mobilen Aktivitäten durch den Datenanbieter, auf Letzteres wird in Abschn. 3.4.2 kurz eingegangen,[29] betrachten wir nachfolgend Einflüsse weiterer Variablen auf die Mobilfunkdaten.

Wie zuletzt bereits diskutiert, hat die in den Mobilfunkdaten eingebrachte Verweildauer einen beträchtlichen Einfluss auf die Verortung und die Quantität der mobilen (stationären) Aktivitäten in Form von Tagessummen. Im Gegensatz zu Abschn. 3.3.3 werden folgend erneut die halbstündigen Verweildauern, diesmal jedoch in Abhängigkeit von den Pendlerverflechtungen, betrachtet. Die mit der Pendlerrechnung übereinstimmenden Pendlerverflechtungen in den Mobilfunkdaten werden dabei anhand der Verweildauer korreliert. Dasselbe Verfahren wird erstmals für die durchschnittliche Distanz bzw. Entfernung zwischen potenziellem Wohn- und Arbeitsort durchgeführt.

Die durchschnittliche Distanz dient als Maßstab, um unplausible Pendlerwege herauszufiltern. Daher wird davon ausgegangen, dass die zurückgelegte Distanz ebenfalls von der Verweildauer abhängen könnte, da annahmegemäß Berufspendler länger an ihrem Zielort verweilen. Möglicherweise hat auch eine größere Entfernung zwischen Wohn- und Arbeitsort einen längeren Aufenthalt am Zielort zur Folge. Abb. 3.6 stellt daher die Zusammenhänge zwischen den übereinstimmenden Pendlerverflechtungen beider Datenquellen sowie die Zusammenhänge zwischen Pendlerverflechtungen aus den Mobilfunkdaten und der durchschnittlich hinterlegten Distanz vom Wohn- zum Arbeitsort, jeweils in Abhängigkeit von der halbstündigen Verweildauer, in einem Liniendiagramm dar.

Zunächst kann festgestellt werden, dass die übereinstimmenden Pendlerverflechtungen aus beiden Datenquellen grundsätzlich mittel bis stark positiv mit der Verweildauer korrelieren. Der Trend in Abb. 3.6 deutet jedoch darauf hin, dass die Korrelation sinkt, je höher die Verweildauer wird. Das bedeutet, je länger die Verweildauer am Zielort andauert, desto weniger Einfluss hat diese auf die potenziellen Pendlerverflechtungen in den Mobilfunkdaten und desto schlechter wird die amtliche Pendlerrechnung mit den vorliegenden Mobilfunkbewegungen abgebildet.

Weiterhin kann aufgedeckt werden, dass die durchschnittliche Distanz eher eine negativ schwache bis mittlere Korrelation mit der zielgerichteten Anzahl der Mobilfunkbewegungen in Abhängigkeit von der Verweildauer aufweist. Das bedeutet, dass theoretisch die durchschnittlich zurücklegte Distanz bei steigenden Pendlerverflechtungen in den Mobilfunkdaten sinkt. Ein klarer Trend ergibt sich, wie bei der Korrelation zwischen der übereinstimmenden Pendlerverflechtung und Verweildauer, jedoch nicht, da die Korrelationskoeffizienten bei steigender Ver-

[29] Weitere Ausführungen zu den methodischen Aspekten bzgl. der Mobilfunkdaten sind in Hadam et al. (2020) und Statistisches Bundesamt (2019a, 2021e) zu finden.

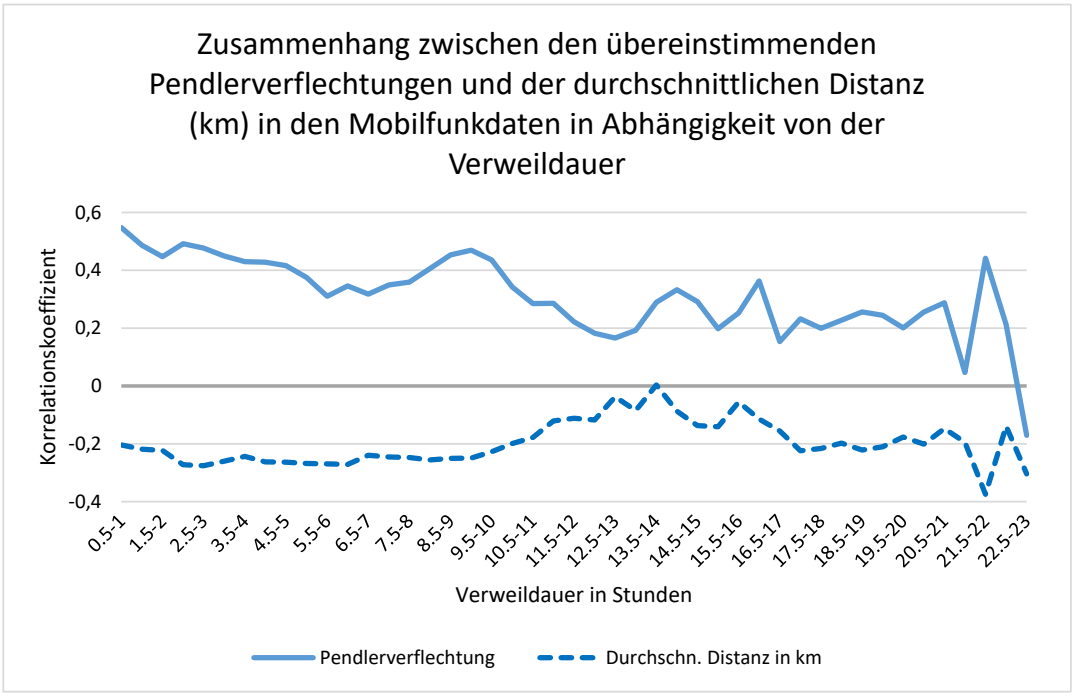



Abbildung 3.6: Korrelationsdiagramm der mit der Pendlerrechnung übereinstimmenden Pendlerverflechtungen sowie der durchschnittlichen Distanz zwischen potenziellen Wohn- und Arbeitsort in km zu den Bewegungsverflechtungen in den Mobilfunkdaten in Abhängigkeit von der durchschnittlichen Verweildauer.

weildauer von bis zu 14 Stunden sogar gar keine Zusammenhänge zwischen Distanz und Mobilfunkbewegungen mehr aufweisen. Der Einfluss der Verweildauer erscheint in diesem Fall eher diffus zu sein.

Dennoch ist anhand von Abb. 3.6 ein Einfluss der Distanz auf die Pendlerverflechtungen in den Mobilfunkdaten erkennbar, wenn auch nicht oder nur mäßig in Abhängigkeit von der Verweildauer. Daher fokussieren wir uns in einem zweiten Schritt auf das reine Verhältnis von Distanz zu den Pendlerverflechtungen der Mobilfunkdaten.

Für eine übersichtliche Darstellung werden fünf Distanzklassen beginnend mit weniger als 10 km Luftlinienentfernung bis hin zu 40 km und mehr gebildet, wobei die maximale Entfernung vom Wohn- zum Arbeitsort, entsprechend der Methodik der Pendlerrechnung, maximal 80 km betragen darf. Die in die Distanzklassen unterteilten Zusammenhänge zwischen den übereinstimmenden Pendlerverflechtungen aus den Mobilfunkdaten und der Pendlerrechnung werden in Abb. 3.7 wiedergegeben. Wie auch in Abb. 3.2 sind die absoluten Zahlen der Pendlerströme der Mobilfunkdaten 2019 auf der y-Achse und die der Pendlerrechnung 2019 auf der x-Achse abgetragen. Ebenfalls finden sich für jede Gegenüberstellung die Korrelationskoeffizienten je

Distanzklasse in der Darstellung.

Ergänzend zur ersten Annäherung des Einflusses der Distanz auf die Mobilfunkdaten (vgl. Abb. 3.6) ergibt sich aus Abb. 3.7 weiterhin, dass zwischen der Pendlerrechnung und den Mobilfunkdaten in den ersten vier Distanzklassen bis zu 40 km, hier definiert und nachfolgend bezeichnet als Kurzdistanzwege, ein sehr starker positiver Zusammenhang besteht. Dieser ist in der kleinsten Distanzklasse am größten mit einem Pearson-Korrelationskoeffizient von 0,9 und bei steigender Distanz tendenziell abnehmend und folglich in der letzten bzw. größten Distanzklasse (40 km und mehr) mit einem Korrelationskoeffizienten von nur noch 0,55 am geringsten.

Da die Zahlen der Pendlerrechnung von den Filterkriterien oder Distanzmaßen der Mobilfunkdaten nicht beeinflusst werden, ergeben sich die Effekte nur aus den Mobilfunkdaten. Daraus lässt sich ableiten, dass mit steigender Distanz die Bewegungsverflechtungen in den Mobilfunkdaten mit denen der Pendlerrechnung weniger übereinstimmen und die Distanz somit, wie anhand der Abhängigkeit von der Verweildauer bereits angedeutet, einen negativen Einfluss auf die Mobilfunkdaten zu haben scheint.

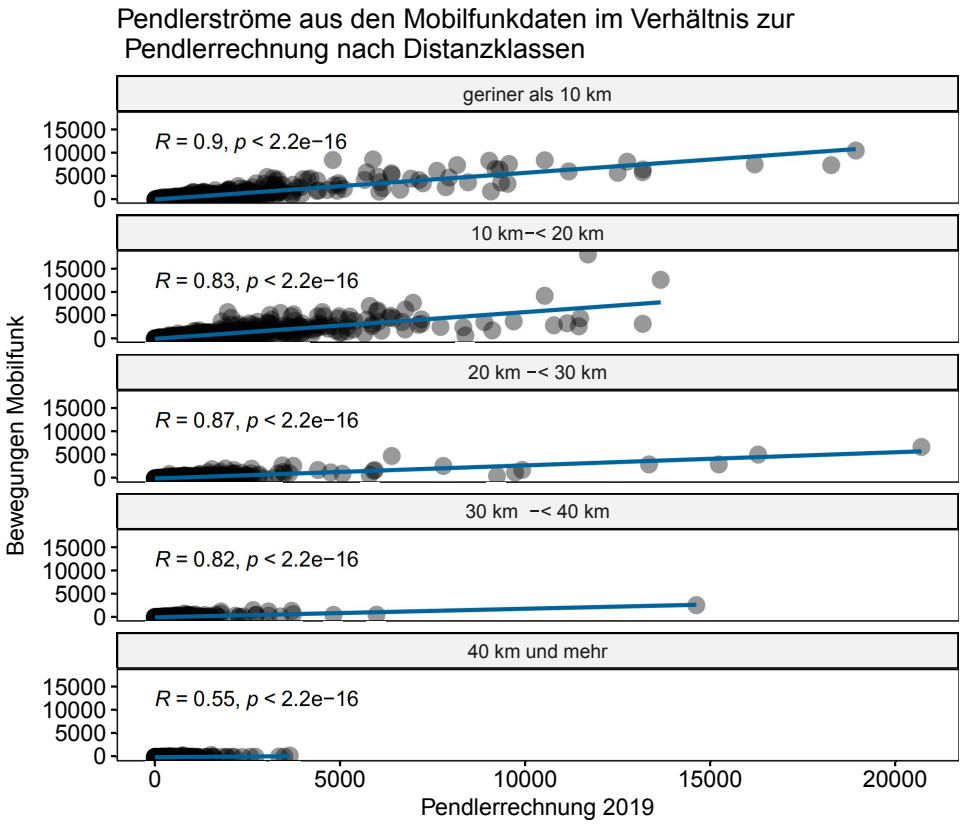



Abbildung 3.7: Korrelationsdiagramm der übereinstimmenden Pendlerverflechtungen aus den Mobilfunkdaten 2019 und der Pendlerrechnung 2019 unterteilt in fünf Distanzklassen.

Für das Erweiterungspotenzial der Pendlerrechnung durch Mobilfunkdaten können wir daraus zudem schlussfolgern, dass eine Erweiterung mit Bewegungsverflechtungen und einer zurückgelegten Distanz von bis zu 40 km (Kurzdistanzwege) aus den Mobilfunkdaten am geeignetsten sind, da diese die Pendlerverflechtungen aus der Pendlerrechnung am wahrscheinlichsten wiedergeben können. Bei solchen Kurzdistanzwegen kann auf Pendlerverflechtungen geschlossen werden, die von angrenzenden oder räumlich nahe gelegenen Gemeinden zur Zielgemeinde führen. Besonders interessant wird diese Erkenntnis für eine Erweiterung der Pendlerrechnung mit der kleinräumigen Zielorts-Bestimmung, kleinräumigen Pendlerverflechtungen oder auch die Angabe innergemeindlicher Berufspendler in Form einer *experimentellen kleinräumigen Pendlerrechnung*. Nehmen wir die Stadt Köln als Beispiel, so würden wir es schaffen mit den Kurzdistanzwegen allein in der Stadt zu verleiben und damit räumlich tiefere Betrachtungen der innergemeindlichen Pendler durchführen zu können, sofern geeignete Daten vorlägen. Dennoch bleibt auch in diesem Fall das Problem bestehen, dass mit den Mobilfunkdaten die Anzahl der mit der Pendlerrechnung übereinstimmenden Pendlerströme in allen Distanzklassen deutlich unterschätzt wird (siehe x-y-Achse in Abb. 3.7).

Durch die Aufbereitungsart der Mobilfunkdaten, wie bspw. die Anwendung diverser Zeit- und Personenfilter, ist ein spürbarer Anteil an Bewegungsverflechtungen aus den Daten verloren gegangen. Weiterhin wurde bereits beleuchtet, dass die Verweildauer oder die Distanz ebenfalls einen Einfluss auf die in den Mobilfunkdaten befindlichen Bewegungsverflechtungen haben. In einem letzten Schritt wollen wir diese sowie weitere mögliche Einflüsse auf die reinen gezählten Mobilfunkaktivitäten bzw. -verflechtungen in einem Regressionsmodell quantifizieren.

Für die vorliegenden Mobilfunkdaten eignen sich sogenannte Zählmodelle zur Modellierung von Zähldaten, wobei die Mobilfunkdaten die Anzahl der täglichen, zielgerichteten Pendlerverflechtungen der einbezogenen Werktage eines dreimonatigen Zeitraums in 2019 wiedergeben. Für die Modellierung von Zähldaten wird in der Regel die Poisson Regression herangezogen, welche jedoch anfällig ist bei Verletzung der Annahme, dass der Erwartungswert und die Varianz der Verteilung gleich $\lambda$ sind. Dabei beschreibt $\lambda$ die mittlere Anzahl der zu erwartenden Ereignisse. Verletzen wir diese Annahme, kann dies, wie im vorliegenden Fall, zu einer Überdispersion führen. Eine Überdispersion liegt vor, wenn die Varianz größer als der Erwartungswert ist.[30] Sie kann zur Über- oder Unterschätzung der Standardfehler der Regressionskoeffizienten führen und dadurch zu einer fehlerhaften Einschätzung der Signifikanz derselben (Hilbe, 2011, Kap. 7). Um der vorliegenden Überdispersion entgegenzuwirken,[31] findet das dafür empfohlene

[30] Eine Überdispersion tritt unter anderem auf, wenn die Verteilungsannahme der Daten durch eine unbeobachtete Heterogenität in den Daten und dadurch die stochastische Unabhängigkeit der Ereignisse verletzt wird. Dies tritt bspw. bei geclusterten Daten auf (Hilbe, 2011, Kap. 7).

[31] Der Dispersionsparameter beträgt im vorliegenden Fall 9,563.

negative Binomiale Regressionsmodell Verwendung.[32] Dieses nimmt die unbeobachtete Heterogenität, durch welche die Überdispersion entsteht, durch einen zusätzlichen Parameter auf (für Details siehe Hilbe, 2011, Kap. 8).

In der durchgeführten negativen Binomialen Regression stellen die gezählten Mobilfunkbewegungen in Abhängigkeit aller möglichen enthaltenen Einflussfaktoren in den Mobilfunkdaten die zu interessierende Größe dar. Als zu erklärende Variablen fließen die bereits diskutierte Verweildauer (`verweildauer`), die Distanz vom Wohn- zum Arbeitsort (`dist_mean`) sowie die bekannten Einwohnergrößenklassen als kategoriale Variable (`class2019...`) im Vergleich zum Referenzwert der ersten Einwohnergrößenklasse mit weniger als 50.000 Einwohnern ein. Des Weiteren sind die möglichen Auswirkungen des Zielortes auf die Anzahl der Mobilfunkdaten von Interesse, insbesondere ob Gitterzellen in der Gemeinde bzw. Stadt enthalten sind (`TypStadt mit MTC`) oder ob eine Stadt ohne Gitterzellen vorliegt (`TypStadt`) im Vergleich zum Referenzwert der Gemeinde mit unter 100.000 Einwohnern. Damit soll ermittelt werden, ob einwohnerreichere Gemeinden oder Städte mit mehr Bewegungsverflechtungen einhergehen.

Trotz der geringen Anzahl an Kovariaten wurde eine Modellselektion mittels einer schrittweisen Variablenselektion unter Verwendung des Akaike-Informationskriteriums durchgeführt, wobei die Auswahl des Modells unberührt blieb. Weiterhin deutet die Devianz des Modells auf einen guten Modellfit im Vergleich zu einem alternativen Poisson Modell hin.[33]

Die resultierenden Koeffizientenschätzer des negativen Binomialen Modells sind in Tab. 3.4 dargestellt. Für die Interpretation der Einflüsse bzw. der geschätzten Effekte der Koeffizientenschätzer ($\hat{\beta}$) auf die gezählten Mobilfunkbewegungen werden ergänzend die Exponenten der Schätzer berechnet ($exp(\hat{\beta})$). Negative oder positive Auswirkungen gehen aus den Vorzeichen der Koeffizientenschätzer hervor. Die numerischen Effekte entsprechen den berechneten Exponenten unter der Voraussetzung, dass alle anderen Koeffizienten konstant gehalten werden.

Die Koeffizientenschätzer in Tab. 3.4 haben insgesamt einen signifikanten Einfluss auf die gezählten Mobilfunkdaten (siehe hierzu $Pr(> |z|)$) und demnach einen Einfluss auf die Anzahl der mobilen Aktivitäten. Betrachten wir die Koeffizientenschätzer einzeln, bestätigen sie mitunter vorherige Erkenntnisse wie die negativen Einflüsse der Verweildauer und der Distanz auf die Anzahl an mobilen Bewegungen. Die Anzahl der Mobilfunkverflechtungen sinkt erwartungsgemäß um den Faktor 0,988, wenn die Verweildauer um eine halbe Stunde steigt und die übrigen Variablen im Modell konstant gehalten werden. Steigt die durchschnittlich zurückgelegte Distanz um einen Kilometer, so sinkt die Anzahl der Mobilfunkverflechtungen um den Faktor 0,976. Diese beiden negativen Effekte haben, im Vergleich zu den anderen Koeffizientenschät-

[32] Vgl. für die praktische Umsetzung in R Venables und Ripley (2002, Kap. 7.4) und Zeileis et al. (2008).

[33] Zudem konnte die Überdispersion von 9,563 auf 4,953 reduziert werden.

zer, den geringsten zu erwartenden Einfluss auf die Mobilfunkdaten.

Anders verhält es sich beim Zielort und der Einwohnergrößenklasse. Da diese beiden kategorialen Variablen mehrere Klassen oder Kategorien enthalten, wird im Folgenden jeweils die erste Kategorie beider Variablen als Referenzwert genommen, woran sich die Effekte der aufgelisteten Kategorien in Tab. 3.4 messen. Dabei stellen wir fest, dass Untersuchungsgebiete bzw. Zielorte mit weiterer Unterteilung in Gitterzellen erwartungsgemäß knapp doppelt so viele Mobilfunkverflechtungen aufweisen wie Gemeinden mit unter 100.000 Einwohnern, wobei die übrigen Variablen konstant gehalten werden (siehe Tab. 3.4 `TypStadt mit MTC` mit Faktor 1,9). Damit üben die Untersuchungsgebiete den größten Einfluss auf die gezählten mobilen Aktivitäten aus. Diese Erkenntnisse werden auch in den Effekten der vier gelisteten Einwohnergrößenklassen gefunden. Je größer die Einwohnergrößenklasse im Vergleich zu Gemeinden mit weniger als 50.000 Einwohnern ist, desto größer wird der Faktor der jeweiligen Klasse, um den die Anzahl der Mobilfunkverflechtungen erwartungsgemäß steigt. Auch hier gilt die Schlussfolgerung nur, wenn alle anderen Variablen im Modell konstant gehalten werden.

Tabelle 3.4: Geschätzte Koeffizienten des negativen Binomialen Modells zur Bestimmung von Einflüssen auf die gezählten Bewegungsverflechtungen in den Mobilfunkdaten.

| Variable | $\hat{\beta}$ | $exp(\hat{\beta})$ | Std. Fehler | z-Wert | $Pr(>\|z\|)$ |
|---|---|---|---|---|---|
| `Intercept` | 2,824 | 16,846 | 0,006 | 511,76 | **0,000** |
| `verweildauer` | -0,012 | 0,988 | 0,0002 | -60,37 | **0,000** |
| `TypStadt` | 0,368 | 1,445 | 0,005 | 76,48 | **0,000** |
| `TypStadt mit MTC` | 0,640 | 1,897 | 0,007 | 98,11 | **0,000** |
| `class201950.000-<100.000` | 0,136 | 1,146 | 0,005 | 27,24 | **0,000** |
| `class2019100.000-<250.000` | 0,123 | 1,130 | 0,008 | 16,00 | **0,000** |
| `class2019250.000-<500.000` | 0,176 | 1,192 | 0,010 | 18,41 | **0,000** |
| `class2019500.000 und mehr` | 0,191 | 1,210 | 0,012 | 15,84 | **0,000** |
| `dist_mean` | -0,024 | 0,976 | 0,0002 | -125,20 | **0,000** |

Zusammenfassend kann daraus abgeleitet werden, dass mehr mobile Bewegungsverflechtungen in einwohnerreichen Regionen provoziert werden und diese weiterhin mit einer kurzen zurückgelegten Distanz sowie kurzer Verweildauer einhergehen. Die letzte Aussage wirkt möglicherweise widersprüchlich zur Schlussfolgerung, dass längere Verweildauern geeigneter sind, um Berufspendler abzubilden. Hier muss daher berücksichtigt werden, dass in Tab. 3.4 nur die Einflüsse auf die Anzahl der Mobilfunkverflechtungen der potenziellen Berufspendler und keine Zusammenhänge mit der amtlichen Pendlerrechnung betrachtet werden.

### 3.4.2 Diskussion möglicher Modifizierungsansätze der Mobilfunkdatenaufbereitung

Die hier umgesetzten Datenaufbereitungsschritte und Definitionen sowie die daraus resultierenden Variablen in den Mobilfunkdaten haben maßgebliche Einflüsse auf die Anzahl der mobilen Aktivitäten sowie auf die Zusammenhänge mit der amtlichen Pendlerrechnung. Die hohen Korrelationen der übereinstimmenden Pendlerströme zwischen der Pendlerrechnung und den Mobilfunkdaten unterstützen grundsätzlich die Annahme, dass Mobilfunkbewegungen die amtlich ermittelten Pendlerverflechtungen teilweise gut abbilden und theoretisch auch unterstützen könnten. Dennoch wird in allen Analysen und resultierenden Ergebnissen in Abschn. 3.3 erkenntlich, dass die absoluten Pendlerverflechtungen auf Basis der Mobilfunkdaten systematisch fehlgeschätzt werden. Letzteres wird vor allem durch den Zielkonflikt zwischen ausschließlicher Extrahierung der Berufspendler und dem Informationsgehalt aller verfügbaren mobilen Bewegungsverflechtungen bedingt. Nachfolgend sollen daher Modifizierungsansätze diskutiert werden, wie sowohl eine bessere Darstellung der Berufspendler als auch ein höherer Informationsgehalt in den Mobilfunkdaten aufgrund einer alternativen Datenaufbereitung erreicht werden könnten.

Aufgrund der Abhängigkeit vom Datenanbieter im Rahmen der individuellen Mobilfunkdatenaufbereitung bleiben nur wenige Spielräume, um die Datenaufbereitung optimal an die Zielsetzung anzupassen. Dennoch können anhand der ermittelten Einflüsse auf die Mobilfunkdaten einige Treiber der Falscheinschätzung der potenziellen Pendlerverflechtungen in den Mobilfunkdaten benannt werden.

Ein erstes zu modifizierendes Filterkriterium kann vor allem in der Erhöhung der Verweildauer mobiler Aktivitäten am Zielort gesehen werden. Die hier eingeführte zweistündige Verweildauer führt zu nicht gewünschten Bewegungsverflechtungen, die nicht denen der Berufspendler entsprechen (siehe hierzu Abschn. 3.3). Die Wahl der Verweildauer ist jedoch abhängig von der Fragestellung. Das bedeutet, dass Teilzeitbeschäftigte eine kürzere Zeitdauer an ihrem potenziellen Arbeitsort verweilen als Vollzeitbeschäftigte (siehe hierzu Abb. 3.5).[34] Bei der Zielsetzung dieser Arbeit wäre eine Verweildauer von sechs Stunden aufwärts wahrscheinlich zielführender. Bei der Wahl einer höheren Verweildauer muss allerdings darauf geachtet werden, dass die Anzahl mobiler Aktivitäten tendenziell abnehmen und damit die Anzahl an potenziellen Berufspendlern geringer ausfallen wird als diejenigen der Pendlerrechnung und folglich zum oben beschriebenen Zielkonflikt zwischen Genauigkeit und Informationsverlust führt (vgl. Tab. 3.4; Abb. 3.6). Im vorliegenden Datensatz würde eine Erhöhung der Verweildauer zu einem extremen Informationsverlust führen, welcher jedoch teilweise abgefangen werden könnte, wenn

[34] Je nach Fragestellung kann es daher notwendig sein, mehrere Mobilfunkdatensätze zu verwenden.

der Eintritt der gruppierten Mobilfunkaktivitäten in die Zielgemeinde vor 9 Uhr aufgehoben wird.

Für die Unterscheidung zwischen Bewegungsverflechtungen durch Freizeitaktivitäten oder durch Berufstätigkeit ist vor allem der Eintrittszeitpunkt in einen Zielort in den Daten zuständig. Problematisch ist der daraus resultierende Ausschluss von Beschäftigten im Schichtdienst o.ä. durch die bedingte Eintrittszeit. Wird die Verweildauer entsprechend erhöht und gleichzeitig der Eintrittszeitpunkt in eine Zielgemeinde entfernt, wird potenziell der Informationsgehalt in den Mobilfunkdaten gesteigert und vorwiegend mehr Berufspendler in den Daten wiedergegeben. Hierbei sollte beachtet werden, dass das erste und letzte Signal des Tages eines mobilen Endgerätes wieder zwingend in derselben Gemeinde erfasst werden muss. Dies ist bei allen Personenkategorien in den vorliegenden Mobilfunkdaten, mit Ausnahme der Tagespendler, nicht zwangsläufig der Fall.

Dennoch bleibt es aufgrund des 24-Stunden-Auswertungszeitraums weiterhin schwierig bspw. Beschäftigte im Nachtdienst zu filtern, solange ihr letztes Signal des Tages nicht am Wohnort registriert wird. Letzten Endes ist eine Abwägung zwischen der detailgetreuen Abbildung der Berufspendler oder anderer Zielgruppen und der Anzahl mobiler Aktivitäten nicht zu vermeiden.

Weiterhin resultieren die Über- und Unterschätzungen der Pendlerströme in den Mobilfunkdaten aus der angewandten Extrapolationsmethodik des Datenanbieters. Eine Extrapolation ist grundsätzlich erforderlich, da der Mobilfunkanbieter Deutsche Telekom ca. ein Drittel des deutschen Mobilfunkmarktes umfasst und damit auch nur ca. 30% aller deutschen Mobilfunkkundinnen und -kunden abbilden kann.[35] Für repräsentative Aussagen basierend auf allen Mobilfunkkundinnen und -kunden müssen die resultierenden Mobilfunkdaten entsprechend korrigiert werden. Hierbei erfolgt die Extrapolation auf die Gesamtzahl aller aktiven Mobilfunkgeräte sowie einer anschließenden Hochrechnung auf die Gesamtbevölkerung und ignoriert dabei eine darauffolgende Einteilung der Mobilfunkdaten in verschiedene Sub- bzw. Zielgruppen, wie bspw. die Berufspendler. Folglich kommt es zur Unter- oder auch Überschätzung dieser Zielgruppen, da die Extrapolation die Mobilfunkdaten auf die Grundgesamtheit zwar angleicht bzw. korrigiert, dabei aber eine Korrektur der Daten nach der jeweiligen Zielgruppe außer Acht lässt, wie die Anzahl der Pendlerverflechtungen der Berufspendler in den vorliegenden Mobilfunkdaten demonstrieren. Demzufolge würden nach Zielgruppen extrapolierte Werte zu weniger starken Verzerrungen führen, da diese durch die Kalibrierung auf Ebene der Subgruppen zielorientierter korrigiert werden würden, weshalb eine Extrapolation nach den

[35] Siehe hierzu Bundesnetzagentur: `https://www.bundesnetzagentur.de/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Marktbeobachtung/Deutschland/Mobilfunkteilnehmer/Mobilfunkteilnehmer_node.html`.

gewünschten Subgruppen vorzuziehen ist.

Für die effektivere Filterung der Zielgruppen ist weiterhin eine geografische Selektion der Mobilfunkbewegungen anhand von bspw. Gewerbeparks, Bildungseinrichtungen etc. zu prüfen. Dadurch bestünde die Möglichkeit, Bewegungsverflechtungen konkret den Zielorten zuzuordnen und wirksamer von anderen Bewegungsverflechtungen abzugrenzen. Mit dem verfügbaren Mischraster ist dies nicht zu bewerkstelligen, da die Gitterzellen aufgrund ihrer unterschiedlichen Größe diverse Gebäude und damit Personengruppen in den Mobilfunkdaten einfangen. Hierfür bedarf es zudem geeigneter georeferenzierter Daten zu allen Arbeitgebern, Bildungseinrichtungen etc.

Ebenfalls wurden Überlegungen angestellt, um anhand soziodemografischer Merkmale der Mobilfunkkundinnen und -kunden bestimmte Bevölkerungsgruppen zu extrahieren. Beispielsweise könnten anhand der Altersgruppe Bildungspendler von Berufspendlern an Schulen, Universitäten oder anderen Bildungseinrichtungen separiert werden. Da die soziodemografischen Merkmale der Mobilfunkanbieter jedoch nur von Vertragskundinnen und -kunden stammen und diese – je nach Anbieter – stark unterschiedliche Kundenstrukturen aufweisen, ist von einer Nutzung dieser zur Filterung bestimmter Bevölkerungsgruppen abzusehen. Resultierende Ergebnisse würden aufgrund der nicht zu korrigierenden Verzerrungen in den Merkmalsausprägungen zu verfälschten Aussagen führen (siehe hierzu Statistisches Bundesamt, 2021e).

Grundsätzlich ist es zudem ratsam einen längeren Auswertungszeitraum zu betrachten als – wie im vorliegenden Fall – drei Monate im Spätsommer/Herbst 2019. Der Wochentag und die Ferienzeiten sind weitere Kriterien, die das Bewegungsverhalten und damit die Anzahl der Bewegungsverflechtungen in den Mobilfunkdaten und die Übereinstimmung mit der Pendlerrechnung beeinflussen. Durch die Exklusion der Schulferien und Feiertage blieben von über 90 Auswertungstagen nur 35 Werktage erhalten. Dies könnte zur Identifizierung von regelmäßigen Pendlerverflechtungen zu wenig sein.

Generell ist eine Nachverfolgung von Bewegungsverflechtungen über 24 Stunden hinaus zu bevorzugen, die jedoch aufgrund von Datenschutzbestimmungen bezüglich des Anonymisierungsverfahrens des Mobilfunkanbieters aktuell nicht verfügbar ist (siehe hierzu Bundesbeauftragte für den Datenschutz und die Informationsfreiheit, 2017). Die meisten Modifizierungsansätze zur Korrektur von Unschärfen[36] bleiben jedoch methodischer Art, auf die nur der Datenanbieter Einfluss hat.

[36] Als Unschärfen werden hier Umgangsformen des Datenanbieters mit ungeeigneten Gegebenheiten in den Mobilfunkdaten verstanden. Darunter fallen Situationen wie das Vorhandensein gemeindeübergreifender Mobilfunkzellen und die Zuordnung mobiler Aktivitäten in eine Gemeinde oder der Umgang mit fehlenden Signalen, wenn bspw. das Mobiltelefon bereits im Untersuchungsgebiet ausgeschaltet wurde und kein letztes Signal in der Endgemeinde mehr erzeugt bzw. registriert wird.

## 3.5 Fazit und Schlussfolgerung

Im Rahmen des Projekts *Pendler Mobil* wurden in Kooperation mit IT.NRW Analysen zur Bevölkerungsmobilität in NRW auf Basis von Mobilfunkdaten aus dem Netz der Deutschen Telekom durchgeführt. Das Ziel des Projekts war es, Bereiche zu identifizieren, in denen Mobilfunkdaten zu einer Ergänzung der bisherigen Pendlerrechnung beitragen können. Die starken Zusammenhänge zwischen den Mobilfunkdaten und der amtlichen Pendlerrechnung am Beispiel von NRW machen deutlich, dass Mobilfunkdaten im Allgemeinen grundsätzlich das Potenzial haben die amtliche Pendlerrechnung zu unterstützen bzw. zu ergänzen. Besonders eine mögliche Erweiterung durch die kleinräumige Zielorts-Bestimmung, d.h. eine Lokalisierung der am stärksten frequentierten Arbeitsgebiete unter Verwendung der Mobilfunkdaten, bringt einen praktischen Mehrwert für die Optimierung der Infrastruktur oder die Bestimmung von funktionalen Räumen und daraus resultierende Standortentscheidungen von Arbeitgebern. Insgesamt offenbaren sich jedoch noch deutliche Unterschiede zwischen der Anzahl der Pendlerströme aus den vorliegenden Mobilfunkdaten sowie einer flächendeckenden Wiedergabe dieser und der Pendlerrechnung von IT.NRW, die sich besonders auf regionaler Ebene stark auswirken.

Mögliche Gründe für die Über- oder Unterschätzungen der Pendlerverflechtungen liegen unter anderem in den Filterkriterien und Aufbereitungsprozessen der Mobilfunkdaten. Besonders die zweistündige Verweildauer erscheint zu kurz, um nur Berufspendler aus den Mobilfunkdaten zu extrahieren. Hierbei wird ein Zielkonflikt zwischen der Genauigkeit der Abgrenzung von Bewegungsverflechtungen bestimmter Zielgruppen und dem Informationsgehalt in den Mobilfunkdaten provoziert. Die Anzahl mobiler Aktivitäten nach Zielgruppen ist hierbei ein entscheidender Faktor für die Bestimmung der Pendlerverflechtungen. Durch die umgesetzte Extrapolation der Mobilfunkdatenströme auf die Gesamtbevölkerung durch den Datenanbieter entsteht eine Unter- oder auch Überschätzung der Berufspendlerströme im Vergleich zur Pendlerrechnung, da eine Kalibrierung der Verzerrungen nach Subgruppen nicht stattfindet. Dies schlägt sich in der Summe aller Mobilfunkbewegungen im Vergleich zur Pendlerrechnung nieder und folglich in einer nicht plausiblen Abbildung der Pendlerverflechtungen. Von einer zeitlich schnelleren Abbildung von Pendlerströmen oder einer zeitnahen fachlichen Unterstützung der Pendlerrechnung mittels Mobilfunkdaten in Form einer experimentellen Pendlerrechnung muss aufgrund der noch deutlichen Änderungs- bzw. Verbesserungsbedarfe der vorliegenden Mobilfunkdaten aktuell abgesehen werden.

Mögliche Lösungsansätze für eine Verbesserung der Ergebnisse oder auch zielführendere Resultate finden sich einerseits in einer Neuformulierung der Filterkriterien oder auch in der Einbindung einer geografischen Selektion zur Extraktion bestimmter Zielgruppen. Weiterhin bedarf es einer methodischen Änderung bzw. Weiterentwicklung der Extrapolation nach Subgruppen,

welche entsprechend eine Kalibrierung der Werte in den Mobilfunkdaten nach ausgewählten Zielgruppen durchführt. Hierdurch würden die Fehlschätzungen der Pendlerströme deutlich minimiert werden. Jedoch sind diese Änderungsbedarfe weiterhin als schwierig umsetzbar anzusehen, solange die Datenaufbereitungsprozesse allein vom Datenanbieter durchgeführt werden und wichtige Methoden und Prozessierungsschritte aufgrund von Geschäftsgeheimnissen bzw. Datenschutzgründen nicht offengelegt werden können. Durch die externe Datenaufbereitung bleiben außerdem die Transparenz sowie die Qualitätseinschätzung der Mobilfunkdaten wie auch der Ergebnisse nicht unberührt.

Dennoch ergeben sich aus den kleinräumigen Zielorts-Bestimmungen vielversprechende Anknüpfungsmöglichkeiten für eine experimentelle kleinräumige Pendlerrechnung bspw. in Form eines Indikators, welcher auf kleinräumige Arbeitsmarktregionen hinweist. Ebenso lassen sich aus den obigen Analysen interessante Anknüpfungspunkte zu einer weiterführenden Untersuchung einer möglichen Darstellung der Bildungspendler und der innergemeindlichen Pendler auf Ebene der Gitterzellen finden, welche in weiteren Machbarkeitsstudien näher betrachtet werden könnten.

Letztlich bezieht sich dieser Artikel nur auf die klassische Form des Pendlerverhaltens, das tägliche Pendeln vom Arbeits- zum Wohnort. Weitere Formen des Pendlerverhaltens sollten in weiteren Forschungsarbeiten fokussierter betrachtet und miteinbezogen werden, wie bspw. die zunehmende Bedeutung der Fern- oder Wochenendpendler (Pütz, 2015), die bislang in der amtlichen Pendlerrechnung weniger berücksichtigt werden. Diese können anhand geeigneter Mobilfunkdaten mit modifizierten Definitionen und Datenaufbereitungsmechanismen ggf. herausgearbeitet und für zusätzliche Erweiterungsmöglichkeiten in der amtlichen Pendlerrechnung genutzt werden.

Weitere fachübergreifende Forschungsarbeiten könnten, neben den zuvor beschriebenen methodischen Erweiterungen, eine Verknüpfung von mobilen Bewegungsverflechtungen der potenziellen Berufspendler mit weiteren sozioökonomischen Aspekten zum Gegenstand haben. Bspw. könnten durch eine Betrachtung des Einkommens oder der beruflichen Qualifikation weitere Fragestellungen zum Pendlerverhalten und ihren Ursachen untersucht werden. Zusätzlich ist eine Erweiterung der kleinräumigen Zielorts-Bestimmung zu einem Indikator für die wirtschaftliche Aktivität der identifizierten kleinräumigen Arbeitsregionen, wie in Arhipova et al. (2020), mitzudenken. Für die Erstellung eines solchen Indikators bedarf es weiterer Informationen zum touristischen Verhalten oder der Wohnbevölkerung, die mitunter auch aus geeigneten Mobilfunkdatensätzen gewonnen werden können.

Zuletzt muss betont werden, dass diese Auswertungen nur anhand eines Bundeslandes getätigt wurden. Bundesweite Aussagen zu Pendlerverflechtungen eröffnen möglicherweise weitere bis dahin nicht bekannte Beurteilungen der Mobilfunkdaten. Als langfristiges Ziel wird daher

eine Prüfung bundesweiter Erweiterungsmöglichkeiten der amtlichen Pendlerrechnung mit neu aufbereiteten Mobilfunkdaten auf Basis der hier angeführten Modifizierungsansätze angeregt. Für eine bundesweite Repräsentativität und demnach Vervollständigung der Daten werden möglichst Daten aller drei Mobilfunkanbieter in Deutschland benötigt, um so auch Fehleinschätzungen durch die Extrapolationsverfahren zu vermeiden, die durch die Nutzung von Mobilfunkdaten nur eines Netzanbieters entstehen. Die hier vorgestellten Analysen sind demnach ein weiterer Schritt zur Integrierung alternativer Datenquellen in die amtliche Statistik, jedoch bedarf es weiterführender Arbeiten, um letztendlich den Übergang von einer experimentellen hin zu einer amtlichen Statistik zu erreichen. Hierfür ist ein nachhaltiger Zugang zu Mobilfunkdaten unerlässlich. Aufgrund einer fehlenden gesetzlichen Grundlage betreffend den Zugang zu Mobilfunkdaten für Eignungsprüfungen der amtlichen Statistik ist dies jedoch aktuell nicht umsetzbar.

# Kapitel 4

# Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany

## 4.1 Introduction

Since jobs are predominantly located in cities, more people move to the cities. For example, the continuous growth of cities is creating shortages on the German housing and real estate markets (Möbert, 2018). Most large cities have higher population growth rates than the national average (see e.g., an interactive map of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR, 2017)). Due to urban labour migration, the number of people living in cities is steadily increasing nationwide. As Buch et al. (2014) showed, smaller cities recorded less net immigration than large cities, which is caused by the attractiveness of larger cities and the advantages of living in them. These are better infrastructure, more education and job opportunities, an extensive cultural infrastructure, and other location-specific amenities (Buch et al., 2014; Gans, 2017).

In contrast to this trend, unemployment rates in Germany are higher in the cities compared to its surroundings. The unemployment rate is one of the most important economic indicators. Unemployment has far-reaching indirect effects on the respective region: It favours the decline of wage levels, educational activities within companies, population mobility, life and health satisfaction, intelligence, and school performance, as well as rising right-wing extremism (Grözinger, 2009). The persistence of spatial disparities in unemployment in an economy is also shown by Elhorst (2003). He points out that regional unemployment is influenced by labour supply (affected by changes in the labour force, such as migration and commuting), labour demand,

and wage-setting factors. Kosfeld und Dreger (2006) conduct a spatial analysis of the German regional labour market, showing that strong spatial dependencies can distort the relationship between employment and unemployment. Also Patuelli et al. (2011) include spatial linkages to effectively predict regional economic variables and to uncover spatial patterns. Particularly, there are strong relationships of dependence between cities and their surrounding areas. Identifying the cities as job magnets and finding high unemployment rates at the same time seems contradictory. According to Grözinger (2018), this phenomenon is a 'false' effect and can be explained by the common definition of unemployment. Traditional unemployment rates are defined by the International Labour Organization (ILO) as the number of unemployed persons counted at their place of residence divided by the total number of persons in the labour force who are resident in the target area. This definition includes only the place of residence as a focal point for calculating these rates. In contrast to traditional unemployment rates, an alternative definition using the workplace as a focal point enables other insightful interpretation possibilities. Following Grözinger (2018), this alternative definition puts the resident unemployed of an area in relation to the labour force of the same area counted at the workplace. The alternative unemployment rate include commuters at their workplace and thus reflect the difference in the supply of jobs. This definition provides valuable information on missing workplaces in regional areas and support policy decisions in urban planning. Thereby, policymakers can identify regions where it might be useful to promote the settlement of companies to lower their unemployment rate and shorten commuter movements. For cities, lower alternative unemployment rates are assumed compared to the traditional definition. Low alternative unemployment rates contribute to the attractiveness of cities and the moving and commuting behaviour towards urban areas. Grözinger (2018) investigates this difference, among others, for regional areas in the German federal states Bavaria and Schleswig-Holstein. Furthermore, the comparison of both rates also provides valuable information on commuting behaviour in regional areas.

For analysing unemployment rates in the context of commuter behaviour, we look at the regional level of Functional Urban Areas (FUAs). For member countries of the Organisation for Economic Co-operation and Development (OECD), FUAs have been created as harmonised geometries describing urban areas (Dijkstra und Poelman, 2011). These regional areas are composed of city cores and their commuting zones. In this application, we use the FUAs in particular to include commuters and commuter areas to a greater extent. Hence, we are interested in considering only the city core and commuter zone separately, which is a spatial level underneath the FUA. We refer to our regional target level in the following as the FUA sublevel. This spatial level is particularly suitable for comparing the two unemployment rates described above, which differ in the spatial reference of the working population. Since this regional level is available for all OECD countries, our comparison of unemployment rates is transferable to other OECD

countries and does not represent a purely German phenomenon. Furthermore, due to data availability, we only consider Germany and particularly the federal state of North Rhine-Westphalia (NRW) which is the federal state with the highest number of commuters in Germany (Bundesagentur für Arbeit, 2022b).

To estimate unemployment rates, our primary data source is the European Union Labour Force Survey (LFS). The LFS enables the estimation of both unemployment rates. The survey is designed on the governmental regions level, which is a higher regional level than the FUA sublevel (Eurostat, 2019b). According to the Nomenclature of Territorial Units for Statistics (NUTS) of the European Union, the German governmental region correspond to the NUTS 2-level and the districts to the NUTS 3-level. The FUA sublevel can be composed from the NUTS 3-level. Estimates on the spatial fine FUA sublevel that are only based on survey data (direct estimates) are likely to have large variances due to relatively small sample sizes. To increase the accuracy of the direct estimates on lower spatial levels, small area estimation (SAE) methods can be used (see e.g., Rao und Molina, 2015; Tzavidis et al., 2018). SAE methods generally combine survey data with other data sources. For example, Costa et al. (2006), Pereira et al. (2011), and Martini und Loriga (2017) estimate unemployment rates using SAE methods by using administrative data as auxiliary information. Molina und Strzalkowska-Kominiak (2020) discuss different types of SAE estimators to calculate the percentage of people in the labour force for Swiss communes out of the LFS. They use administrative data that are provided at unit-level as auxiliary information. Similarly, Marino et al. (2019) propose semi-parametric empirical best prediction for unemployment rates that requires unit-level information. For many research questions, appropriate register or administrative data is not available. In particular, unit-level data is strictly protected. Furthermore, aggregated data is often not available at spatial finer resolutions, so that information at the target level is missing. One possibility is to use alternative data sources as covariates. Toole et al. (2015) and Steele et al. (2017) propose the usage of passively collected mobile phone data as auxiliary information, as they have a finer spatial resolution, high timeliness, and are available in real time. Basically, mobile network data can serve as a basis for producing statistics with a very high level of spatial, temporal and population coverage. For example, Steele et al. (2017) use Call Detail Records (CDRs) from the mobile network and remote sensing data for estimating poverty indices in developing countries. Toole et al. (2015) estimate changes in unemployment rates after shocks in the economy in case of mass layoffs at a plant by using mobile phone data. Moreover, Marchetti et al. (2015) have investigated solutions for a broad range of applications in using new digital data. They suggest three ways to use new digital data together with SAE techniques and show the potential of these data sources to mirror aspects of well-being and other socio-economic phenomena.

Our analyses are based on dynamic mobile network data, which is more widely available and

has more information content than mobile phone data. This data source validly reflects actual commuting behaviour as well as daytime and residential population, which is important for providing auxiliary information. Since commuters and daytime population affect unemployment rates, the usefulness of these covariates for estimating the traditional and alternative unemployment rates becomes apparent. Our application combines mobile network data with data from the LFS to improve the estimation of both unemployment rates on the FUA sublevel. The aim is to compare both definitions of unemployment rates at the level of interest, thus highlighting the influence of commuters. As sample sizes are small at the FUA sublevel SAE methods are needed. From a methodological perspective, we consider the Fay-Herriot (FH) model (Fay und Herriot, 1979) using mobile network data as auxiliary information. The inverse sine transformation of the dependent variable is used frequently in literature to estimate proportions when applying the FH model (Casas-Cordero et al., 2016; Burgard et al., 2016; Schmid et al., 2017). The transformation offers the advantage of stabilization of the sampling variances and helps to approximate better the normality assumptions of the model. Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) apply a naive back-transformation to obtain FH estimates and their confidence intervals on the original scale. In contrast, we use a bias corrected back-transformation following Sugasawa und Kubokawa (2017) while using as well the inverse sine transformation. To measure the uncertainty of these specific FH estimates, we propose a parametric bootstrap procedure orientated on González-Manteiga et al. (2008) to receive not only confidence intervals but also estimates for the mean squared error (MSE). The methodology is validated with official rates based on the Urban Audit. In a model-based simulation study, we show the benefit of a bias corrected back-transformation compared to a naive one.

The paper is structured as follows: Section 4.2 defines both types of unemployment rates and explains how they deal differently with commuters. Subsequently, this section introduces the data sources for constructing these indicators. Section 4.3 describes the statistical methodology. The SAE methods and the corresponding MSE estimation is applied in Section 4.4 to estimate both unemployment rates for the German federal state NRW on FUA sublevel. Section 4.5 investigates the methodology on German data for estimating the traditional unemployment rates and compares the results with official data. Furthermore, in Section 4.6, we conduct a model-based simulation study to assess the quality of the proposed estimator. Section 4.7 discusses further research potential.

## 4.2 Data sources and definitions for regional unemployment rates

In this section, we first introduce the two definitions of unemployment rates each dealing differently with commuters as well as our regional target level: the FUA sublevel (Section 4.2.1).

Subsequently, our two data sources are described: the LFS data (Section 4.2.2) and mobile network data (Section 4.2.3).

### 4.2.1 Traditional and alternative definition of unemployment rates

The unemployment rate according to the definition of the ILO provides an international comparable indicator (ILO, 2018). Following the ILO-definition, the traditional unemployment rate $\theta_{\text{UR}_1,i}$ for regional area $i$ is given by

$$\theta_{\text{UR}_1,i} = \frac{\text{N}_{i,\text{unempl. (residence)}}}{\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (residence)}}}. \tag{4.1}$$

This unemployment rate is defined by the number of unemployed persons living in area $i$ ($\text{N}_{i,\text{unempl. (residence)}}$) divided by the labour force of area $i$. The labour force is composed of the number of unemployed and employed persons living in area $i$ ($\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (residence)}}$). For traditional unemployment rates, the focal point for counting employed and unemployed persons is their place of residence, where persons aged $15$ to $74$ are considered in the ILO-definition (ILO, 2018; Eurostat, 2018a). Please note that for reasons of comparability with German official statistics, we use the age range of $15$-$64$ years throughout the analysis. In contrast to the traditional definition, the second definition proposed by Grözinger (2018) uses the workplace as a focal point and thus counts employed persons at the area $i$ where their workplace is located. Since unemployed persons have no place of work, they count at area $i$ where they live. The definition changes to

$$\theta_{\text{UR}_2,i} = \frac{\text{N}_{i,\text{unempl. (residence)}}}{\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (workplace)}}}. \tag{4.2}$$

We refer to $\theta_{\text{UR}_2,i}$ as alternative unemployment rate for area $i$. It is composed by the number of unemployed persons ($\text{N}_{i,\text{unempl. (residence)}}$) divided by the labour force aged 15 to 64 ($\text{N}_{i,\text{unempl. (residence)}} + \text{N}_{i,\text{empl. (workplace)}}$). Comparing alternative unemployment rates to traditional ones, the denominator changes as employed persons count in the area $i$ where they work. Overall, both unemployment rates treat commuters differently. If commuting is not exactly balanced, the two unemployment rates differ, and this difference reveals the influence of commuters. If the traditional unemployment rate in area $i$ ($\theta_{\text{UR}_1,i}$) is higher than the alternative one ($\theta_{\text{UR}_2,i}$), there is a stronger commuter movement from other areas to area $i$ than the other way around which is assumed for larger cities.

We focus on the alternative definition of unemployment rates as defined in Equation 4.2. However, there are other alternative definitions such as those of the Federal Labour Office

in Germany (Bundesagentur für Arbeit, 2022a) or the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics, 2021), which take into account a more socio-political perspective and the relative underutilisation of the labour supply. In contrast to the alternative definition according to Grözinger (2018) used here, the labour force remains the same as in the traditional unemployment rate, while the numerator changes.

In this study, the geographical target level for investigating unemployment rates is the FUA sublevel which is particularly suitable to illustrate the difference in both definitions of unemployment rates caused by commuter flows. To the best of our knowledge, the FUA sublevel is the only OECD harmonised geometry that allows a distinction between city cores and their commuter zones. City cores are urban centres with at least 50 000 inhabitants. The commuting zone contains the surrounding travel-to-work areas of the city core where at least 15% of their employed residents are working in the respective city core (Eurostat, 2018b). Please note that the FUA sublevel as well as the FUA do not cover the whole territory of a country. Germany has in total 208 units, which are relevant for determining FUAs. These are composed of 125 city cores and 83 commuting zones. Since some commuting zones can be assigned to several city cores, there are fewer commuting zones than city cores.

### 4.2.2 Labour Force Survey

The LFS (Eurostat, 2019b) enables the estimation of the traditional and alternative unemployment rates introduced in Section 4.2.1. It is a household survey conducted in 35 countries including all 27 EU member states and the United Kingdom, which provides information about the labour market participation. In Germany, the LFS is part of the German Microcensus, which is a one-percent sample of the population and collected annually. All inhabitants who have their main or secondary residence in Germany and live in private or collective households are included. The sampling design corresponds to a stratified single-stage cluster sample, where neighbouring buildings are sampled and all households and persons within this cluster are surveyed. The sample districts are stratified according to region and size of the buildings (Eurostat, 2019c). In the used LFS data, regional disaggregation is carried out using the EU-harmonised NUTS classification (Eurostat, 2018c). In Germany, the NUTS 1-level corresponds to the 16 federal states, the NUTS 2-level to 38 governmental regions, and the NUTS 3-level to the 401 administrative districts (European Parliament and Council, 2003). Traditional unemployment rates using LFS data are published on the 38 governmental regions level (NUTS 2-level). However, our target level is the smaller FUA sublevel which can be composed from the NUTS 3-level in Germany. As all LFS observations contain information about the NUTS 3-level and even finer, we can use the individual information of the LFS participants to match a) the place of residence

and b) the place of work to the corresponding FUA sublevel.

In addition to the FUA sublevel, there are other possible spatial levels that are suitable to examine unemployment rates. The so-called Labour Market Areas (LMAs) are functional spatial areas that capture regional labour market structures based on commuting flows (Franconi et al., 2017). In principle, both territorial structures pursue the same goal. Nevertheless, there are practical reasons and advantages to prefer the FUA sublevel in the context of this work: First, the LMAs are compiled from commuter statistics using a specially developed algorithm. In contrast, FUAs are based on territorial structure, are already harmonised, and comparable across countries. Second, the separation of the city cores and commuter zones is an advantage of FUAs versus LMAs, which is fundamental for our analysis. Third, Germany provides indicators for the Urban Audit, which is an official statistic and publishes labour market indicators, including traditional unemployment rates, at the level of the entire FUA (one level above our target level) which we can use for external validation. All in all, the FUAs are more suitable for our analyses than the LMAs, since they fit better to the research question and are easier to handle.

In this work, we consider the year 2016 with an overall sample size of 369 986 observations in the LFS. Since the FUA sublevel does not cover the whole territory, the sample size decreases to 271 587 observations. Due to known gender differences in employment, the following analyses are carried out separately by sex. Men work more often full-time, while the proportion of women employed part-time has increased in recent years, so that overall fewer women than men are unemployed (Klammer und Menke, 2020; Statistisches Bundesamt, 2021d). Due to the still existing classical gender role model, women commute fewer and shorter distances than men. These differences in behaviour justify why it is meaningful to examine unemployment separately by sex (Augustijn, 2018). Table 4.1 represents the sample sizes in the LFS based on the published NUTS 2-level and on the FUA sublevel by sex. It can be seen that the sample sizes are smaller in case of the FUA sublevel. On average, the sample sizes decrease by a factor of 7.3. Since the LFS was designed to produce reliable estimates on NUTS 2-level, the challenge of this work is to estimate reliable unemployment rates on the smaller FUA sublevel. Even if the sample sizes for FUA sublevel appear to be rather high, with a median of 368 and 421 for women and men, respectively, results in Section 4.4 show that the coefficient of variation (CV) of the direct estimates often exceeds the threshold of 20% which specifies reliable estimates at Eurostat (Eurostat, 2019a). SAE methods are discussed to obtain more reliable model-based estimates on the FUA sublevel. Since SAE methods take advantage of auxiliary variables from other data sources the auxiliary information used here is described in more detail in the next Section 4.2.3.

Table 4.1: Distribution of sample sizes in the LFS on NUTS 2-level and FUA sublevel in Germany by sex.

| | Sex | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|
| NUTS 2-level | Female | 1 162 | 2 916 | 4 104 | 4 623 | 5 521 | 10 684 |
| | Male | 1 318 | 3 304 | 4 565 | 5 114 | 6 108 | 11 675 |
| FUA sublevel | Female | 100 | 216 | 368 | 635 | 646 | 7 973 |
| | Male | 97 | 244 | 421 | 702 | 749 | 8 559 |

### 4.2.3 Mobile network data

To estimate unemployment rates on FUA sublevel using SAE methods, we take advantage of suitable auxiliary information. Many SAE applications are based on register data as a second data source. These data sources are not timely or are aggregated to higher (regional) levels. Alternative data sources have the potential to overcome these disadvantages. For example, Toole et al. (2015) or Steele et al. (2017) have used mobile phone data for SAE. Mobile network data are explored to estimate daytime population, commuter patterns or tourism behaviour (see e.g., De Meersman et al., 2016; Galiana et al., 2018). Mobile network data represent mobile activities or signals from the mobile network of the respective mobile network operator. A mobile activity is defined as an event caused by a length of stay in a specific geometry without movement (also known as dwell time). Signalling data are produced automatically, regularly and only register the location of the cell tower to which a mobile device is connected at a specific time. Therefore, they are collected as a by-product and tend to be less costly compared to official survey data. The major advantage of these data sources are their real-time availability, high temporal actuality, nationwide availability, and their finer spatial resolution. Mobile activities can be obtained at the spatial resolution of cities, communities or grid cells, so that a simple assignment to other spatial levels such as the FUA sublevel is possible. This spatial flexibility and high resolution are not feasible with register or administrative data. In many countries, like Germany, register data are strictly protected and thus not available at high resolution or on specific regional levels. In addition, mobile network data are dynamic, so that the movement of activities can be observed over the course of the day as well as daily, during a week or a month. Previous analyses in Germany have shown that mobile network data correlate strong with register-based census data like population figures and with the population mobility, more precisely commuter movements (Hadam, 2018, 2021). This is, among other things, due to the high penetration rate of mobile devices in the German population (Statistisches Bundesamt, 2022a). Accordingly, mobile network data provide a reliable picture of the real physical locations of the German daytime and nighttime population or with other words the resident and working population compared to official

statistics with a fixed reporting date. Since our aim is to estimate an alternative unemployment rate accounting for commuters mobile network data reflecting resident and working population are especially suitable auxiliary information (cf. Hadam (2021)).

In Germany, there are three mobile network operators: Deutsche Telekom, Vodafone, and Telefónica Deutschland, with a respective market share of one-third each. The data records available to Destatis and used for this work contain mobile activities of Deutsche Telekom customers. In compliance with data protection rules, the mobile activities are anonymised and aggregated. Regionally fluctuating market shares of each mobile network operator are adjusted regionally as part of the extrapolation procedure at the respective operator. Thus, the estimated local market shares of each operator are used as weights to adjust the mobile network data. The data records include contract, prepaid, and further customers. In addition, mobile network data contain information on socio-demographic characteristics of mobile device users, such as age group, sex, and nationality of the SIM card owner. However, the characteristics are only available for contract customers. Furthermore, the following assumptions were made in the data provider's data generation process: Since the number of mobile activities depends on the dwell time of mobile devices, long mobile device activities are counted and included in the data record according to the length of the dwell time, while short mobile activities are not considered. The dwell time in the data record available is two hours to filter out short mobile device activities (for example, quick movements between the grid cells). Finally, only values based on a minimum number of 30 activities per geometry were provided due to data protection reasons.

Our aim is to analyse the effect of commuters on the two proposed unemployment rates. Since we use a model-based method, suitable covariates are crucial. We only use mobile network data for this purpose and no further covariates. As we will show in Section 4.4.1, our models with only mobile network covariates lead to high coefficients of determination, so mobile network data are sufficient as SAE covariates in our case.

We define from the mobile network data 27 auxiliary variables. Between 7 to 16 auxiliary variables are chosen by model selection procedure (cf. Section 4.4.1). The data contains mobile activities for a statistical week that consists of 24-hour days. These were selected from the months April, May, and September in 2017 without school or public holidays to avoid distortions in the representation of commuters. The mobile activities comprise the average activities on the selected weekdays. The weekdays are categorised according to five types of days, with the days from Tuesday to Thursday being grouped together. Since the counted activities of mobile devices alone are not meaningful enough, further covariates are constructed from the available mobile network data at the FUA sublevel. The aim in creating the covariate is to highlight the differences between the daily and resident population and thus the commuters themselves. This is particularly reflected in the changes in the intensities of mobile activities. Based on this, co-

variates are calculated in the form of ratios, shares, and change values which reflect exactly these differences. Since it is assumed that the unemployed persons are more likely to stay at home during the day and the employed are more likely to stay at the place of work, the rate and change of activities in the morning and evening hours are calculated. This means, that the change from place of work to the place of residence and vice versa is modelled. This includes the change in mobile activities of working hours and hours spent at home as well as the change in activities of potential commuters. In addition, the change in activities during the day is calculated and the differences in core times or peaks in mobile activities are determined. The core times are based on the usual working times in Germany (7 am to 4 pm). Furthermore, differences in mobile network activities among to socio-demographic characteristics such as age, nationalities (summarised by continent), and sex can also be considered. These characteristics also have an influence on commuting behaviour. An overview of the selected mobile network covariates can be found in the supplementary material in Table 4.6.

## 4.3 Small area method

In this section, the statistical methodology for estimating unemployment rates on FUA sublevel is described. As the LFS is designed for higher regional levels, a model-based approach enriched by auxiliary variables from mobile network data is used. We use the FH model (Fay und Herriot, 1979), an area-level model that links direct estimates to area level covariates. The FH model is especially useful in countries with strict data protection requirements like Germany, as the auxiliary variables and the direct estimates only need to be available on an aggregated level. As in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017) we use the inverse sine transformation on the dependent variable to estimate proportions using area-level models. Following Sugasawa und Kubokawa (2017), we derive the inverse sine transformed FH model including a bias correction for the back-transformation. A parametric bootstrap, which incorporates the bias correction, is proposed.

### 4.3.1 Fay-Herriot estimates

In the following, we assume a finite population of size $N$, which is divided into $d$ areas. The present sample consists of areas with different sample sizes $n_1, ..., n_d$ drawn by a complex design from the population. To refer to the actual area, we use the subscript $i$. The population size and sample size of this area is indicated with $N_i$ and $n_i$, respectively. The FH model is a special case of a linear mixed model. The FH model is composed of two levels. The model for the frist level

is the sampling model

$$\hat{\theta}_i^{\text{direct}} = \theta_i + e_i \quad \text{with } i \in 1, ..., d.$$

$\hat{\theta}_i^{\text{direct}}$ is an unbiased direct estimator for a population indicator of interest $\theta_i$. The sampling errors $e_i \sim N(0, \sigma_{e_i}^2)$ are independent and normal distributed and their variance $\sigma_{e_i}^2$ is assumed to be known. In applications, the sampling variance is typically supplied by the data provider or estimated from unit-level sample data. We use the **survey** package from R (Lumley, 2004; R Core Team, 2019) and consider the sampling design of the LFS and the survey weights to estimate $\hat{\theta}_i^{\text{direct}}$ and the respective sampling variances.

The second stage of the FH model links a vector with $p$ area-specific covariates $\mathbf{x}_i$ (aggregates, e.g. area-level means) to the direct estimate ($\hat{\theta}_i^{\text{direct}}$) using an area-specific random effect $u_i$ for each area $i \in 1, ..., d$:

$$\hat{\theta}_i^{\text{direct}} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2),$$

where $\beta$ is a vector of unknown fixed-effects parameters.

Combining both levels results in the FH model:

$$\hat{\theta}_i^{\text{direct}} = \mathbf{x}_i^T \boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{e_i}^2).$$

The model assumes that the random effects $u_i$ are identically independently normally distributed and the sampling errors $e_i$ are independently normally distributed. $\hat{\theta}_i^{\text{direct}}$ is the direct estimate for the unemployment rates for a certain area $i$ and $\hat{\sigma}_{e_i}^2$ its variance estimate that are estimated with the **survey** package from R (Lumley, 2004; R Core Team, 2019) considering the sampling design of the LFS and the survey weights. The regression parameters $\hat{\boldsymbol{\beta}}$ can be estimated as best linear unbiased estimator of $\boldsymbol{\beta}$ and the random effect $\hat{u}_i$ as empirical best linear unbiased predictor of $u_i$ (Rao und Molina, 2015). For the estimation of the variance of the random effects $\sigma_u^2$, several approaches are available: The FH method of moments, the maximum likelihood method (ML), and the restricted maximum likelihood method (REML) among others (Rao und Molina, 2015). For our analysis, we use the REML method.

Through this combination, we obtain the resulting FH estimator, which is an empirical best linear unbiased predictor of $\theta_i$. It is as a weighted combination of the direct estimator $\hat{\theta}_i^{\text{direct}}$ and the synthetic estimator $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ for each area $i$:

$$\begin{aligned} \hat{\theta}_i^{\text{FH}} &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \hat{\gamma}_i \hat{\theta}_i^{\text{direct}} + (1 - \hat{\gamma}_i) \mathbf{x}_i^T \hat{\boldsymbol{\beta}}, \end{aligned} \tag{4.3}$$

where the shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ defines the weight on both parts for each area $i$. Whenever the variance of the sampling errors is relatively small for a specific area $i$, more weight is assigned on its direct estimator.

### 4.3.2 Back-transformed Fay-Herriot estimates

As unemployment rates are a percentage, we transform the dependent variable to profit from the variance stabilization of the sampling variance. Thus, we use the inverse sine transformation $h(x) = \sin^{-1}(\sqrt{x})$ as in Casas-Cordero et al. (2016), Burgard et al. (2016), and Schmid et al. (2017). Note that Schmid et al. (2017) compared in a design-based simulation study the inverse sine transformation with alternative modelling options, for instance an estimator based on a normal-logistic distribution. Both estimators lead to very similar results regarding MSE and bias. Raghunathan et al. (2007) defends the choice of the inverse sine transformation for estimating cancer risk factors rates against generalized linear models with their higher complex design features and computational tasks. While they all use a naive back-transformation $h^{-1}(x) = \sin^2(x)$, we transform the FH estimator back to the original level with consideration to the back-transformation bias. Burgard et al. (2016) mentioned the methodology for a bias corrected back-transformation. We derive the back-transformation following Sugasawa und Kubokawa (2017), who introduce the FH model for general transformations on the dependent variable. Following Jiang et al. (2001), we approximate the sampling variances of the transformed direct estimates by $\tilde{\sigma}_{e_i}^2 = 1/4\tilde{n}_i$, where $\tilde{n}_i$ denotes the effective sample size. The design effects and thus the effective sample size can also be estimated with the **survey** package (Lumley, 2004; R Core Team, 2019). For the model on the transformed scale, we consider the assumptions of the FH model

$$\sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + e_i, \quad u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \tilde{\sigma}_{e_i}^2). \tag{4.4}$$

Out of the FH model on transformed scale in Equation 4.4, $\hat{\boldsymbol{\beta}}$ and $\hat{u}_i$ can be estimated, as described in the previous Section 4.3.1. Replacing the model parameters with their estimates leads to the FH estimator on the transformed level:

$$\hat{\theta}_i^{\text{FH}^*} = \hat{\gamma}_i \sin^{-1}\left(\sqrt{\hat{\theta}_i^{\text{direct}}}\right) + (1 - \hat{\gamma}_i)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}.$$

However, the goal is to get the FH estimator on the original scale $\left(\hat{\theta}_i^{\text{FH, trans}}\right)$. For this reason, $\hat{\theta}_i^{\text{FH}^*}$ must be back-transformed. According to the Jensen-inequality (Jensen et al., 1906), a naive back-transformation $\left(\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right)$ leads to biased results due to the non-linearity of the

transformation. To avoid this bias, the following formula using the known distribution of the FH estimator on the transformed level $\hat{\theta}_i^{\text{FH}^*} \sim \mathcal{N}\left(\hat{\theta}_i^{\text{FH}^*}, \frac{\hat{\sigma}_u^2 \tilde{\sigma}_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}\right)$ is used

$$\begin{aligned}
\hat{\theta}_i^{\text{FH, trans}} &= E\left\{\sin^2\left(\hat{\theta}_i^{\text{FH}^*}\right)\right\} \\
&= \int_{-\infty}^{\infty} \sin^2(t) f_{\hat{\theta}_i^{\text{FH}^*}}(t) dt \\
&= \int_{-\infty}^{\infty} \sin^2(t) \frac{1}{2\pi \frac{\hat{\sigma}_u^2 \tilde{\sigma}_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}} \exp\left(-\frac{\left(t - \hat{\theta}_i^{\text{FH}^*}\right)^2}{2\frac{\hat{\sigma}_u^2 \sigma_{e_i}^2}{\hat{\sigma}_u^2 + \tilde{\sigma}_{e_i}^2}}\right) dt, \qquad (4.5)
\end{aligned}$$

where $\hat{\theta}_i^{\text{FH, trans}}$ denotes the transformed FH estimator. To solve this integral, numerical integration techniques are applied. In Section 4.6, the proposed bias corrected FH estimator ($\hat{\theta}_i^{\text{FH, trans}}$) is evaluated in a close to reality model-based simulation study.

### 4.3.3 Uncertainty estimation

As a measurement of uncertainty for $\hat{\theta}_i^{\text{FH, trans}}$, a parametric bootstrap MSE as well as parametric bootstrap confidence intervals are constructed. When using a FH model without transformations or with a log transformation, analytical solutions to estimate the MSE are known (Prasad und Rao, 1990; Datta und Lahiri, 2000; Slud und Maiti, 2006). Up to our knowledge, no analytical solution is available in the case of the inverse sine transformation. Bootstrap methods are very promising to estimate the MSE. Casas-Cordero et al. (2016) construct confidence intervals using a parametric bootstrap procedure, in which confidence interval limits are built on the transformed scale with subsequent naive back-transformation for each bootstrap replication. In contrast to this methodology, our goal is to construct confidence intervals and a MSE for FH estimates from a model using the inverse sine transformation. Another difference is that, instead of the naive back-transformed FH estimates, the bias corrected back-transformed FH estimates are included within the bootstrap procedure. Our parametric bootstrap is orientated on the bootstrap procedure of González-Manteiga et al. (2008). In the following, the steps of the used bootstrap method to construct both measurements of uncertainty are shown:

- From the model on the transformed scale (Equation 4.4), take $\tilde{\sigma}_{e_i}^2$ and estimate $\hat{\sigma}_u^2$ and $\hat{\boldsymbol{\beta}}$ using the sample data.
- For $b = 1, ..., B$
  - Generate area specific random effects $u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2)$ and sampling errors $e_i^* \sim \mathcal{N}(0, \tilde{\sigma}_{e_i}^2)$.

  - Bootstrap samples:
    * Use $u_i^*$ and $e_i^*$ to construct the bootstrap sample on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}_{i,(b)}^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i^* + e_i^*.$$

    * Use the bootstrap sample to estimate the FH estimator on the transformed scale $\left(\hat{\theta}_{i,(b)}^{\text{FH}^*}\right)$ as described in Section 4.3.2.
    * Determine the FH estimates on the original scale $\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}}\right)$ using (Equation 4.5) to account for the bias correction.
  - Bootstrap population:
    * Use $u_i^*$ to construct the bootstrap population on the transformed scale

$$\sin^{-1}\left(\sqrt{\hat{\theta}_{i,(b)}^{\text{direct}}}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + u_i^*.$$

    * For each bootstrap population, calculate the population mean on the original scale

$$\theta_{i,(b)}^{\text{trans}} = \sin^2\left(\mathbf{x}_i^T\boldsymbol{\beta} + u_i^*\right).$$

- Predict the MSE and the 95% confidence intervals

$$\text{MSE}(\hat{\theta}_i^{\text{FH, trans}}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)^2 \tag{4.6}$$

$$\begin{aligned}\text{CI}(\hat{\theta}_i^{\text{FH, trans}}) = \Big[&\hat{\theta}_i^{\text{FH, trans}} + q_{0.025}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right);\\ &\hat{\theta}_i^{\text{FH, trans}} + q_{0.975}\left(\hat{\theta}_{i,(b)}^{\text{FH, trans}} - \theta_{i,(b)}^{\text{trans}}\right)\Big],\end{aligned} \tag{4.7}$$

  where $q_{0.025}$ is the 2.5% quantile over the bootstrap replications and $q_{0.975}$ respectively the 97.5 % quantile.

The methodology presented above for constructing uncertainty measurements for the back-transformed FH estimates is also evaluated within a simulation study (cf. Section 4.6).

## 4.4 Alternative unemployment rates including commuters in North Rhine-Westphalia

In this section, we determine and discuss traditional and alternative unemployment rates that deal differently with commuters. For this purpose, we use the LFS data from Section 4.2.2 and

the mobile network data from Section 4.2.3. Traditional and alternative unemployment rates have been introduced in Section 4.2.1. The members of the labour force are counted for the two rates at different reference points: At the place of residence (traditional unemployment rates) or at the place of work (alternative unemployment rates). In particular, they assign commuters to different small areas. When using traditional unemployment rates, the contradiction of high unemployment rates in the city cores results from the exclusion of commuting. Alternative unemployment rates are expected to exceed traditional ones in commuter zones and to be lower in city cores. We confirm this empirically. The rates are estimated separately by sex and at the target level of the FUA sublevel.

### 4.4.1 Model selection and validation

Four models need to be created and validated. Following Schmid et al. (2017), the Bayesian information criterion for a simple linear regression model is used for the model selection. As dependent variable, we use the inverse sine transformed direct estimates from LFS and the auxiliary information is mobile network data (cf. Section 4.2.3). In total, 6 to 16 of 27 potential mobile network covariates are selected depending on the model. The covariates of all four models are listed in Table 4.6 within the Appendix 4.8.1. Since the models are built on the transformed scale, the coefficients have no natural interpretation in terms of expected values at the original level, but their direction is directly interpretable. The chosen covariates reflect most likely relationships between working and non-working hours and the changes in mobile activities due to commuting during the day and evening. The latter is represented less strongly in the females model, which is in line with lower commuting patterns of women. An increase of covariates that proxy possible commuter movements generally leads to a decrease of alternative unemployment rates ($UR_2$). The reverse is the case for traditional ones ($UR_1$). All models include changes from night to day activities of other nationalities, most likely tourists, which have a positive impact on regional employment. As expected, negative values have been observed for these coefficients.

To investigate the explanatory power of the models, we use the modified $R^2$ from Lahiri und Suntornchost (2015) and obtain values of at least 57% as shown in Table 4.2. Furthermore, we check whether meaningful results are obtained for estimating the variance of the random effects using REML estimation. As Table 4.2 shows, positive values were estimated in all cases. Thus, the potential problem of negatively estimated variances does not occur. For each FH model on the transformed scale, the assumptions on the error terms (level 1 and 2) are checked. The normality assumptions of the random effects (level 2) as well as of the residuals (level 1) – obtained from fitting the model (Equation 4.4) – are tested. The p-values of the Shapiro-Wilk test in Table 4.2 confirm that in all cases the normality assumption for both error terms cannot be rejected.

Overall, all four models could be validated and are suitable for subsequent analyses.

Table 4.2: Measurements to validate the FH models for traditional ($UR_1$) and alternative unemployment rates ($UR_2$) separated by sex: This table shows the estimated variance of the random effects ($\hat{\sigma}_u^2$), the Shapiro-Wilks (S.-W.) p-value for level 1 and level 2 error terms as well as the modified $R^2$.

| | Men | | Women | |
|---|---|---|---|---|
| | $UR_1$ | $UR_2$ | $UR_1$ | $UR_2$ |
| $\hat{\sigma}_u^2$ | 0.000320 | 0.000361 | 0.000716 | 0.000880 |
| S.-W. p-value: level 1 | 0.308668 | 0.495064 | 0.809323 | 0.866098 |
| S.-W. p-value: level 2 | 0.695112 | 0.549476 | 0.861257 | 0.901708 |
| modified $R^2$ | 0.772521 | 0.908642 | 0.632059 | 0.575550 |

### 4.4.2 Gain in accuracy

To assess the gain in the reliability of the estimators, we compare the CVs. Figure 4.1 visualises this measurement for the different methods and definitions of unemployment rates. Eurostat considers estimators with a CV below 20% to be reliable (Eurostat, 2019a). If we use direct estimation 53.7% (men; $UR_1$), 29.3% (women; $UR_1$), 53.7% (men; $UR_2$), and 31.7% (women; $UR_2$) of the CVs are below 20%. The use of the transformed FH model achieves a distinct increase of CVs below this threshold. As a result, 85.4% (men; $UR_1$), 73.2% (women; $UR_1$), 82.9% (men; $UR_2$), and 78.0% (women; $UR_2$) of the CVs are below 20%. This illustrates that the use of dynamic mobile network data in combination with SAE methodology is a powerful tool to increase the precision of both estimated unemployment rates for NRW on FUA sublevel. If we compare the direct estimates to the estimates from the proposed transformed FH model, both are often close to each other. For regions with smaller samples sizes like Witten and Paderborn, these values can deviate clearly from each other. Due to the higher uncertainty of the direct estimates for regions with lower sample sizes, the synthetic part within Equation 4.3 is weighted higher and bigger differences to the direct estimates appear.

### 4.4.3 Discussion of the estimated unemployment rates for NRW

Figure 4.2 illustrates the differences between the alternative and traditional unemployment rates. If the traditional unemployment rates are the same as alternative one, the commuter behaviour is balanced and the calculated difference would be zero. Please note, that the FUA sublevels do not cover the entire federal territory in NRW, these areas are white in Figure 4.2. The bluish colors indicate areas where the alternative unemployment rate is higher than the traditional one.

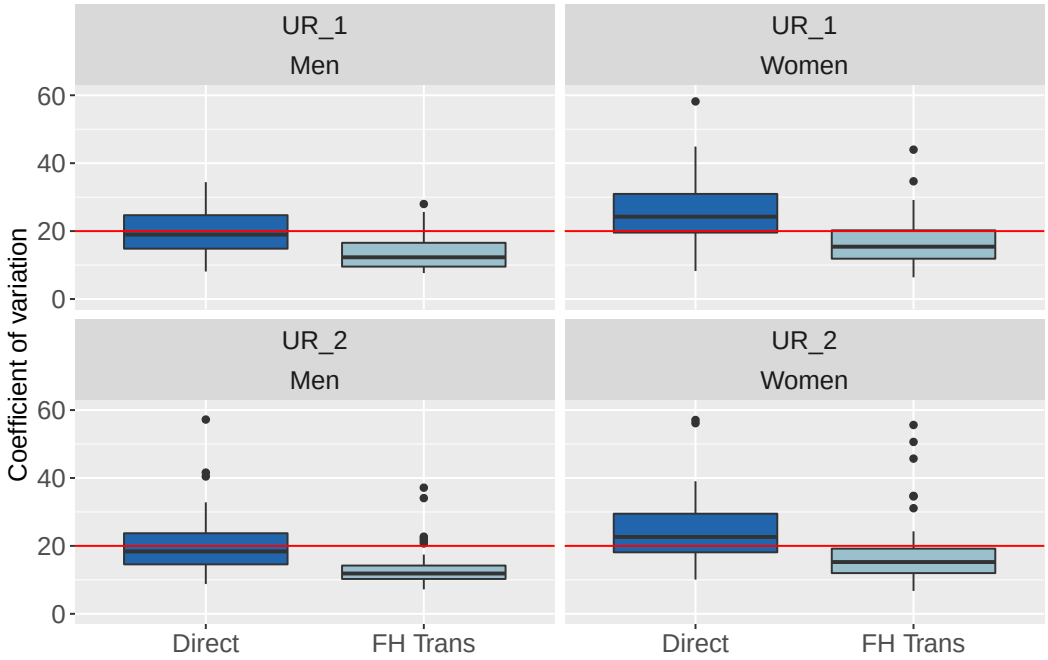



Figure 4.1: Reduction of the coefficient of variation by using the transformed FH model instead of direct estimation for estimating unemployment rates in NRW.

Those are mainly the commuter zones in both models, i.e., the commuter flow is directed out of this area. With one exception in the female model, all commuting zones are coloured blue. This means that these areas are the place of residence of many employed people who commute from those areas to their workplace. The reddish areas, however, imply that the alternative unemployment rate is lower than the traditional unemployment rate. This is mainly the case for the city cores of the FUAs. This observation is consistent with Grözinger (2018) motivation for creating an alternative unemployment rate. Nevertheless, a negative value (blue colouring) was detected for a few city cores. This is the case for nine city cores simultaneously in both models. These are the city cores Recklinghausen, Bottrop, Moers, Oberhausen, Duisburg, and Mühlheim an der Ruhr. These six are located in the Ruhr region, which includes the large city cores Essen and Dortmund, to which many people commute from the Ruhr region. Furthermore, this trend was found for the two small city cores (Solingen and Sankt Augustin) and Aachen, which is located directly on the Belgian border. Since most city cores are job engines, many employed people living in the surrounding travel-to-work areas, which is their place of residence, commute into the city cores to work. In the males model, the differences are higher than in the females model, which leads to the conclusion that women are not commuting as often or as far as men

(IT.NRW, 2019). Possible reasons for this could be the conservative role model of women, the spatial closeness to the family that is guaranteed by the woman (to the school/kindergarten of the children, etc.) or, for example, a work in small, nearby companies/enterprises (Bauer-Hailer, 2019).

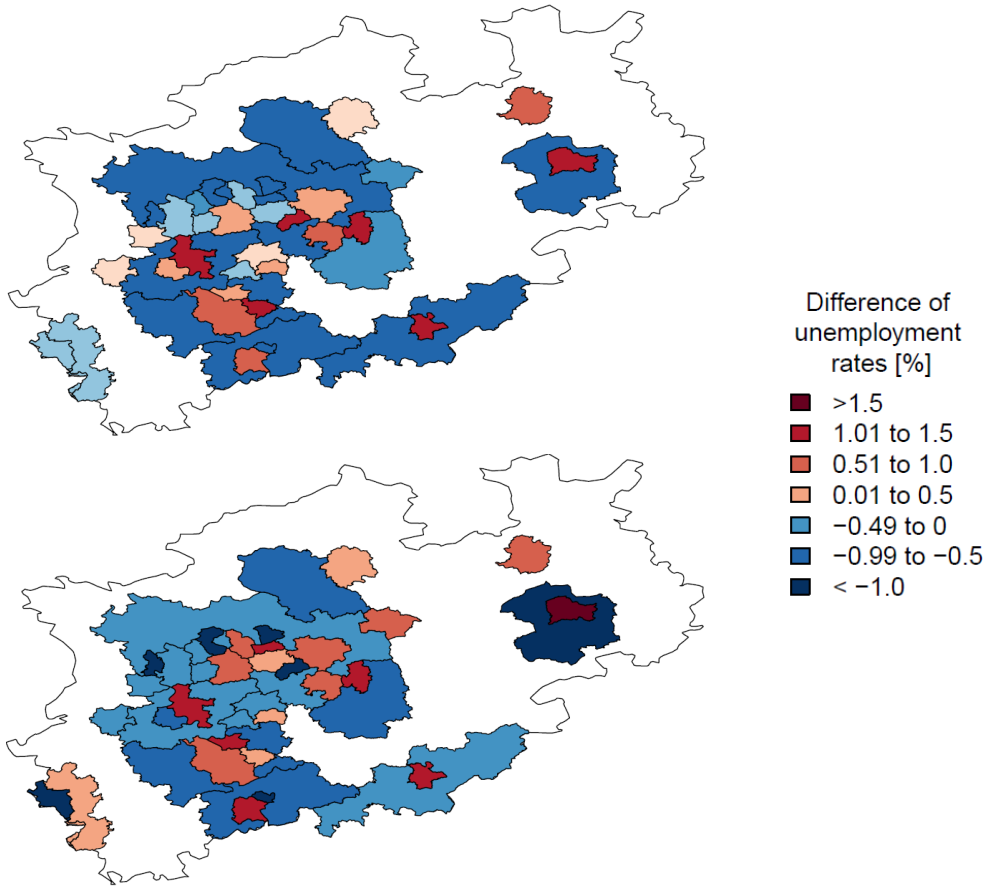



Figure 4.2: Difference of unemployment rates due to including commuters for men (above) and women (below). The spatial assignment of city names to the FUA sublevels is shown in the Appendix 4.8.2.

## 4.5 Validity of the proposed method

In the following, we evaluate the methodology used in Section 4.4 to estimate unemployment rates at the FUA sublevel through official data. For Germany, the database Urban Audit provided by Eurostat in cooperation with Destatis and Kommunales Statistisches Informationssystem (KOSIS) is the only source for German unemployment rates at the FUA level (KOSIS-

Gemeinschaft Urban Audit, 2013; Eurostat, 2017, 2019d). This official data source provides traditional unemployment rates, but no alternative unemployment rates for all German FUAs. Thus, the Urban Audit enables a comparison of traditional unemployment rates estimated by using the transformed FH estimator (Equation 4.4) with mobile network data as auxiliary information with the officially published values. As mentioned in Section 4.2.1, we have used the 15 to 64 age range for the definitions of unemployment rates to ensure comparability with the Urban Audit. Please note, the comparison in this section is made on the entire FUA level and not on the FUA sublevel as in the application in Section 4.4.

For the German federal state NRW, we have an extensive mobile network data record available as auxiliary information. However, we have only limited access to mobile network data and accordingly a data set with less information for the rest of the country. Thus, less covariates are available for the validation. In contrast to Section 4.4, where we use dynamic signalling data, we only have static mobile network activities of a typical Sunday evening for the whole of Germany. We focus on the time period from 8 to 11 pm of the average of eight Sundays of the months April, June, and July in 2018 without school or public holidays. For Sunday evenings, a high correlation has been identified between population figures from the 2011 census and the mobile network activities on the weekend and especially on Sunday evening (Hadam, 2018). As traditional unemployment rates are based on the place of residence, it is reasonable to assume that mobile network data of a Sunday evening is suitable as auxiliary variables. In the following, we validate the proposed transformed FH model by comparing the FH estimates with official unemployment rates of the Urban Audit. We use the SAE method and model selection as applied in Section 4.4 with the difference that a) the regional focus is now FUAs across Germany and b) we can only use mobile network data from Sunday evening. In the males model, the selected mobile network covariates explain around 47% of the variance in terms of the modified $R^2$ following Lahiri und Suntornchost (2015) and in the females model around 37%.

For the validation of the proposed method, Figure 4.3 shows the estimated unemployment rates using mobile network covariates (FH Trans), the direct, and the published official estimates from Urban Audit by sex. First, it can be seen that we get similar rates compared to the Urban Audit by using the transformed FH model. Comparing the direct estimator from the LFS with the FH Trans estimator, the FH Trans estimator corrects the direct estimator in such a way that the resulting value is closer to the Urban Audit. This trend is quantified in Table 4.3. It reports the distribution of the absolute difference of the females and males unemployment rates obtained by the two estimation methods for all FUAs in Germany compared to the Urban Audit. For almost all distribution values, we get a higher absolute difference for the direct estimates compared to the FH Trans estimates. Only in the males model the 25% quantile for the absolute difference is slightly higher for the FH Trans estimates. As expected, it can be noted that for FUAs with

sample size under 600 estimated unemployment rates of both estimation methods show higher values for the absolute difference.

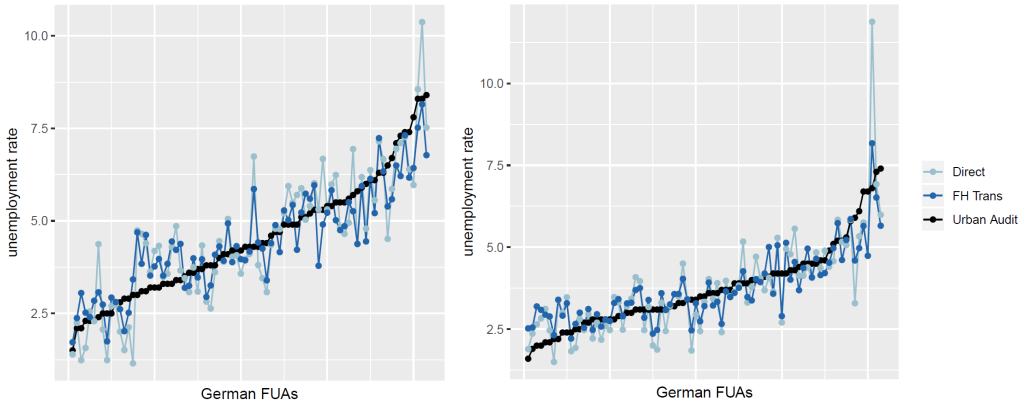



Figure 4.3: Comparison of traditional unemployment rates ($UR_1$) published in Urban Audit (black), estimated with the transformed FH model (dark blue) and the direct estimates from the LFS (light blue) for men (left) and women (right) for all German FUAs.

Table 4.3: Distribution of the absolute difference to the Urban Audit estimates of the females and males traditional unemployment rates over all German FUAs and in particular over FUAs with small sample sizes below 600.

| Areas | Sex | Estimator | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|---|
| All | Female | Direct | 0.017 | 0.246 | 0.459 | 0.638 | 0.800 | 5.078 |
| | | FH Trans | 0.005 | 0.173 | 0.415 | 0.512 | 0.748 | 1.959 |
| | Male | Direct | 0.009 | 0.202 | 0.625 | 0.713 | 0.998 | 2.440 |
| | | FH Trans | 0.008 | 0.221 | 0.428 | 0.573 | 0.824 | 1.690 |
| Sample size <600 | Female | Direct | 0.030 | 0.416 | 0.628 | 0.930 | 1.120 | 5.078 |
| | | FH Trans | 0.015 | 0.281 | 0.516 | 0.627 | 0.896 | 1.959 |
| | Male | Direct | 0.068 | 0.697 | 1.095 | 1.129 | 1.764 | 2.073 |
| | | FH Trans | 0.038 | 0.373 | 0.676 | 0.704 | 1.027 | 1.690 |

## 4.6 Model-based simulation

In the previous two sections, we use the proposed transformed FH model to estimate alternative unemployment rates and subsequently evaluate the suggested methodology with official statistics obtained from Urban Audit. This model-based simulation study is used to investigate how much we benefit from the more complicated transformed FH model with a bias corrected back-

transformation compared to the naive back-transformation. According to the Jensen-inequality (cf. Section 4.3.2), the naive back-transformation is biased under the inverse sine transformation. Furthermore, we want to show, that the proposed MSE and confidence intervals lead to reasonable results. We investigate these aims in a close to reality environment. The input values of the model-based setting are based on the real data.

The simulation study is implemented with $R = 1\ 000$ Monte Carlo replications. Within each replication, we generate the covariates ($\mathbf{x}_i$) initially from a lognormal distribution with parameters (-0.5, 0.04). The number of areas is fixed to the number of the FUA sublevels in Germany ($d = 208$). We draw the random effect and the sampling errors from normal distributions: $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and $e_i \sim \mathcal{N}(0, \sigma_{e_i}^2)$. According to the males model for Germany on FUA sublevel, $\sigma_u \approx 0.029$ is defined analogously. In addition, we adopt the variation of the sampling errors $\sigma_{e_i}$ and keep them constant over the replications. The regression coefficients are set to $\beta_0 = 0.01$ and $\beta_1 = 0.35$. As data generating process, we consider $\hat{\theta}_i^{\text{direct}} = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i + e_i\right)$ to get synthetic direct estimates. The true small area means are $\bar{y}_i = \sin^2\left(\beta_0 + \mathbf{x}_i^T\beta_1 + u_i\right)$. Table 4.4 shows the distribution of the variation of the sampling errors and the resulting shrinkage factor as well as the distribution of the direct estimates for the simulation (over all replications) and the actual direct estimated unemployment rates for males in Germany. The distributions are close to each other.

Table 4.4: Distribution of important parameters in the simulation setting: The sampling error variation $\sigma_{e_i}$ and the resulting shrinkage factor $\gamma_i$ coincide with the male model for Germany on FUA sublevel. The direct estimates $\left(\hat{\theta}_i^{\text{direct}}\right)$ of the simulation study are close to the values for the FUA sublevel.

| | | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|
| $\sigma_{e_i}$ | | 0.0063 | 0.0202 | 0.0275 | 0.0288 | 0.0366 | 0.0785 |
| $\gamma_i$ | | 0.1199 | 0.3848 | 0.5265 | 0.5355 | 0.6730 | 0.9548 |
| $\hat{\theta}_i^{\text{direct}}$ | sim. | 0.0000 | 0.0340 | 0.0495 | 0.0538 | 0.0688 | 0.2826 |
| | FUA sublevel | 0.0054 | 0.0328 | 0.0484 | 0.0508 | 0.0647 | 0.1134 |

For each replication, we estimate small area means from the transformed FH model: With respect to the back-transformation bias $\left(\hat{\theta}_i^{\text{FH,trans}}\right.$, cf. Equation 4.5$\left.\right)$ and with naive back-transformation $\left(\hat{\theta}_i^{\text{FH,naive}}\right)$. To assess the quality of the estimates, we obtain for $R = 1\ 000$ Monte Carlo replications the absolute Bias (aB) and the root mean squared error (RMSE) of the esti-

mates, defined as

$$\text{aB}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)} \right) \right| * 100$$

$$\text{and RMSE}_i = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left( \hat{\theta}_i^{\text{FH},(r)} - \bar{y}_i^{(r)} \right)^2} * 100,$$

where $\hat{\theta}_i^{\text{FH},(r)}$ is the estimated respective FH value and $\bar{y}_i^{(r)}$ the true value within replication $r$. Figure 4.4 shows the reduction of aB. For instance, the median of the aB using a naive back-transformation is 1.86 times higher than with a bias corrected back-transformation. At the same time, we observe nearly the same RMSE (cf. Figure 4.4) when we use a bias corrected back-transformation instead of a naive back-transformation. In summary, there is a clear reduction in bias at the cost of a slightly higher RMSE.

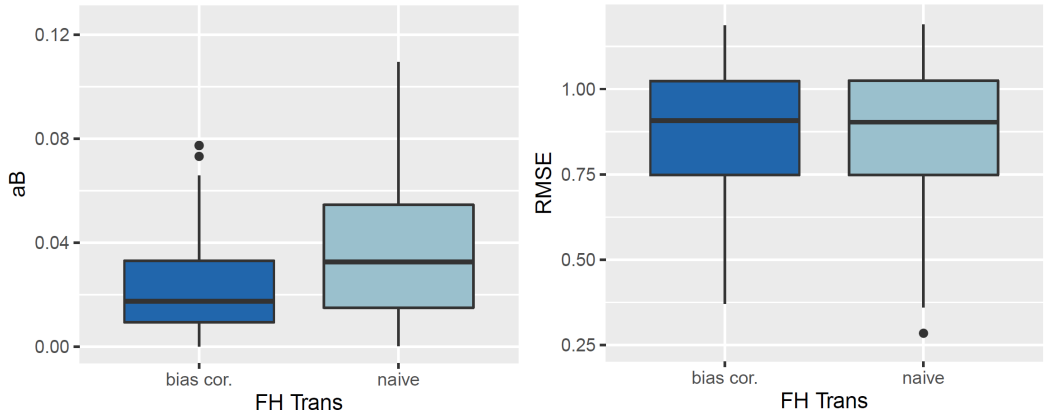



Figure 4.4: Distribution of the aB and the RMSE for the transformed FH estimator with bias corrected and naive back-transformation.

We next investigate the properties of the proposed MSE and the confidence intervals. Please note that we compare the bootstrap estimated RMSE (Equation 4.6) to the empirical RMSE, which we treat as the true one. For calculating these uncertainty measurements, we use 1 000 bootstrap replications within each Monte Carlo run. As quality measurements, we calculate the relative bias of the uncertainty estimation (rB RMSE) and the relative RMSE of the uncertainty

estimation (rRMSE RMSE). They are defined as

$$\text{rB RMSE}_i = \left( \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \text{MSE}_{\text{est},i}^{(r)}} - \text{RMSE}_{\text{true},i}}{\text{RMSE}_{\text{true},i}} \right) * 100$$

$$\text{and rRMSE RMSE}_i = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R} \left( \text{RMSE}_{\text{est},i}^{(r)} - \text{RMSE}_{\text{true},i} \right)^2}}{\text{RMSE}_{\text{true},i}} * 100,$$

where $\text{RMSE}_{\text{est},i}^{(r)}$ is the estimated RMSE out of the bootstrap procedure (cf. Section 4.3.3) for each Monte Carlo replication $r$ and $\text{RMSE}_{\text{true},i}$ is the empirical RMSE over the Monte Carlo replications. The relative bias is close to zero as Table 4.5 shows. On average, we get an underestimation of 0.55% over all areas. The interquartile range goes from -2.12% to 1.11%. In addition, the relative RMSE of the estimated RMSE is important to assess its quality. We get a mean relative RMSE of 18.74% for the estimated RMSE. The low bias and the RMSE show that the proposed MSE estimator yields good results. In addition to the MSE, we can also get bootstrap confidence intervals (cf. Section 4.3.3). The coverage is defined as the proportion of the time that the estimated confidence interval contains the true value. For the proposed confidence intervals (Equation 4.7), we get in mean a coverage of 94.34%. We can recognize a slight underestimation of the coverage, but the values are close to the target value of 95%. These three measures show that the proposed bootstrap-estimated MSE works.

Table 4.5: Distribution of the quality measurements for the estimated RMSE and the corresponding confidence intervals using the bootstrap procedure as described in Section 4.3.3.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| rB RMSE | -9.34 | -2.12 | -0.62 | -0.55 | 1.11 | 7.13 |
| rRMSE RMSE | 17.04 | 18.03 | 18.53 | 18.74 | 18.99 | 44.97 |
| Coverage | 86.70 | 93.90 | 94.40 | 94.34 | 94.90 | 96.00 |

Overall, our close to reality simulation study shows the reduction of bias while using the transformed FH estimator with bias corrected back-transformation instead of a naive back-transformation. Furthermore it demonstrate the good performance of the newly proposed MSE estimator and confidence intervals for the transformed FH estimator with bias corrected back-transformation.

## 4.7 Concluding remarks

The traditional unemployment rate is based on the place of residence of the labour force. Due to the high level of commuting, this may give a distorted impression of regional labour markets. For Germany, traditional unemployment rates show higher rates in city cores compared to its surroundings. For analysing unemployment rates in the context of commuter behaviour, the regional target area are city cores and their commuting zones, which can be extracted from FUAs. In this work, we estimate an alternative unemployment rate, where the focal point of the labour force is their workplace. It adjusts the traditional definition by including commuters.

Since the LFS is not designed to produce indicators on smaller areas than NUTS 2-level, a FH approach is used to estimate alternative and traditional unemployment rates on the FUA sublevel. From a methodological point of view we use a bias corrected back-transformed FH estimator and propose a MSE estimator to measure its uncertainty. As the FH approach relies on a model-based method, suitable covariates are required. We select covariates constructed from dynamic mobile network data and validate the selected models. The benefit of dynamic mobile network data is that they can represent the changes of the counted aggregated mobile devices during the day and in space. This information can be used to derive the commuting behaviour of the population.

The resulting differences between the traditional and the alternative unemployment rates show that the rates in city cores are mainly lower than officially indicated. The assumption that unemployment rates in city cores are lower can be confirmed and thus contributes to the explanation why so many people move to city cores due to more job opportunities. Furthermore, the alternative definition of the unemployment rate removes the static picture of the population, especially of the labour force. The labour force does not necessarily live in the same place where they work. This dynamic cannot be achieved with traditional survey methods and with traditional data. However, exactly this knowledge is necessary to make better decisions regarding urban planning. Moreover, these alternative rates provide potential employers with additional information about the current regional labour market and on missing workplaces. This will help to identify regions for which it might be useful to promote business settlement in order to reduce unemployment rates and shorten commuting distances, as new details of potentially available local workforce are available. The increasing number of commuters should be taken into account in official statistics in the future. Although the application in this paper refers to NRW, the model is also applicable to countries that perform the LFS and have implemented an FUA structure. Thus, this analysis is transferable to at least all European countries.

In Germany, we are facing some limitations in mobile network data. We do not have access to individual signalling data or CDRs. No individual activity movements or changes in individual

social behaviour can be used for the estimation. For instance, Toole et al. (2015) have shown that unemployed persons have different mobile phone usage profiles than employed ones. This information may increase the explanatory power in estimating unemployment rates compared to the used distribution of mobile activities over time.

From a methodological point of view, we leave the uncertainty of the difference between the two unemployment rates as further research. So far, we propose an MSE and confidence intervals for each unemployment rate separately. To obtain these two measures for the difference, it is necessary to calculate the covariance between both unemployment rates. For the special case of the difference between a design-based estimator and a FH estimator from the same repeated survey at different points in time, van den Brakel et al. (2016) derives the covariance. It is assumed that the design-based estimator is unbiased and that the covariates for the FH estimator come from the same survey as the design-based estimator. Since these assumptions are not applicable to our case, further research is needed to apply these results to the present case.

In addition, the following research opportunities remain open from an applied perspective. Steele et al. (2017) uses a combination of satellite and mobile phone data to gain more explanatory power in the estimation of poverty indicators. Satellite data include valuable information on a small regional level of building intensities and heights of buildings to differentiate between socially impoverished people, who live in socially weak urban districts, and wealthy people, who are living more likely in less densely populated areas, which could also be suitable for our question. Furthermore, it is of interest to which extend the same differences in unemployment rates also apply to other countries or whether it is a national phenomenon.

## Acknowledgments

Würz gratefully acknowledges support by a scholarship of Studienstiftung des deutschen Volkes.

## 4.8 Appendix

### 4.8.1 Mobile network covariates

Table 4.6: Mobile network covariates: The last four columns refer to the four different models on unemployment rates at the FUA sublevel. The covariates are based on mobile network data of Deutsche Telekom for the years 2017 and 2018 and represent a statistical week. For each selected variable, the regression coefficient is shown.

| Definition of variables | | $UR_1$ male | $UR_2$ male | $UR_1$ female | $UR_2$ female |
|---|---|---|---|---|---|
| Intercept | | 118.5287 | 14.8136 | 0.5674 | -14.3924 |
| *Proportion of mobile activities of specific subgroup at defined time* | | | | | |
| Central European | 7 am to 4 pm | -2.6305 | -2.7364 | -2.2433 | |
| Central European | 5 pm to 11 pm | 3.1693 | 4.0103 | | |
| *Proportion of mobile activities of specific subgroups at defined time on Sunday* | | | | | |
| under 50s | Sunday 8 pm to 11 pm | | -0.1478 | -0.6384 | |
| 20 to 30 year olds | Sunday 8 pm to 11 pm | | | 0.8168 | |
| *Change of mobile activities by nationality from night-time (5 pm to 11 pm) to day-time (7 am to 4 pm)* | | | | | |
| African | | 0.0012 | 0.0013 | -0.0034 | -0.0024 |
| Australia Oceania | | -0.0001 | -0.0001 | | |
| Eastern Europe | | -0.0680 | -0.0836 | | |
| North American | | -0.0245 | -0.0695 | -0.0273 | -0.0446 |
| Northern Europe | | -0.0176 | -0.0449 | | |
| Southeast Europe | | -0.1070 | -0.1654 | -0.1145 | -0.1165 |
| Southern Europe | | | 0.1158 | 0.1021 | |
| Asia | | | | | 0.0135 |
| Central Europe | | | | | -5.2348 |
| *Relative change of mobile activities between two specific times: (time point 1−time point 2)/time point 2* | | | | | |
| 10 am | 9 pm | -3.2224 | 4.7858 | | 2.4082 |
| 8 pm to 10 pm | 9 am to 11 am | 3.2963 | -3.7265 | | |
| 4 pm | 10 am | | -1.2954 | -1.1901 | |
| 9 am to 11 am | 3 am to 5 am | | | 2.4185 | |
| *Ratio of mobile activities between two specific times: time point 1/time point 2* | | | | | |
| 7 am to 4 pm | 5 pm to 11 pm | 3.0494 | | 5.2589 | |
| 5 pm to 5 am | whole day | -119.1781 | -28.0838 | | |
| 9 am to 11 am | 8 pm to 10 pm | -3.9171 | -3.5458 | -5.5749 | |
| 6 am to 4 pm | whole day | -111.7206 | | | 30.2304 |
| 12 pm to 6 am | 7 am to 4 pm | | 2.7691 | | |
| 3 am to 5 am | 9 am to 11 am | | | 2.0348 | |

### 4.8.2 Map of FUA city cores and commuter zones in NRW

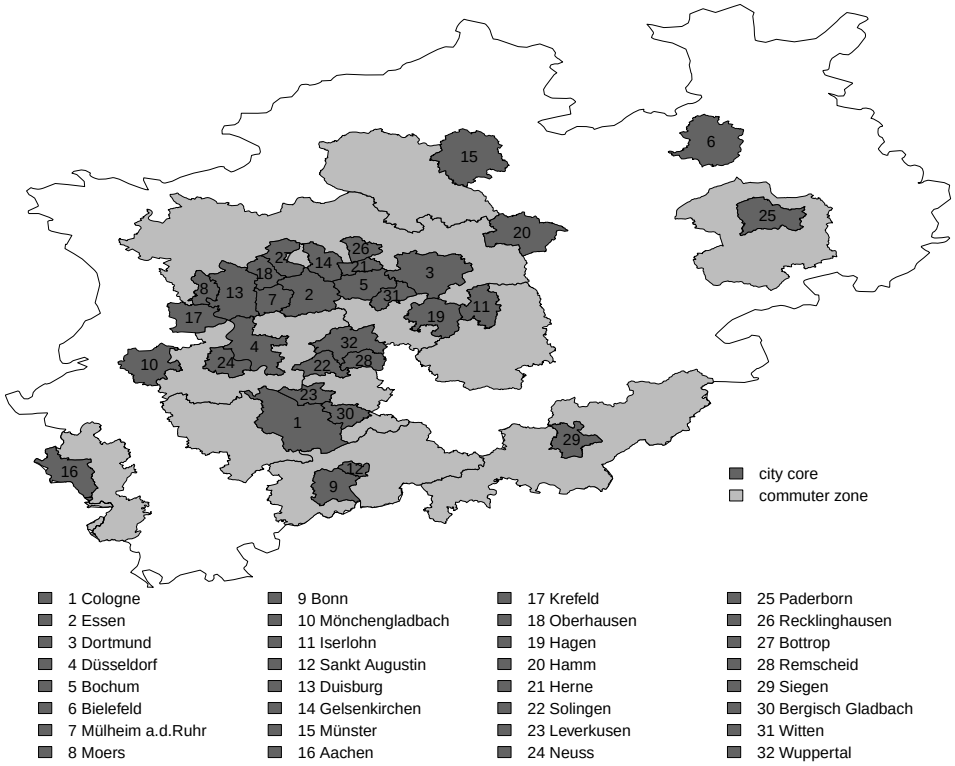



Figure 4.5: Assignment of city names to FUA city cores and geographical location of the commuter zones for NRW.

# Glossar

**Auspendler** Erwerbstätige Personen, die nicht in der Gemeinde arbeiten, in der sie wohnen.

**Beschäftigungsumfang** Umfang der erbrachten Arbeitsleistung der Beschäftigten; hier vereinfachend in Teilzeit- und Vollzeitbeschäftigung unterschieden.

**Bevölkerungsfortschreibung** Stellt die Zahl und die Zusammensetzung der Bevölkerung nach demografischen Merkmalen und ihre Veränderung fest. Die Bevölkerungszahlen ergeben sich durch Fortschreibung der Ergebnisse der jeweils letzten Volkszählung.

**Bewegungsverflechtungen** Anonymisierte und aggregierte Bewegungen von Signalen bzw. mobilen Aktivitäten im Mobilfunknetz vom Start- zum Zielort, hier auch als Mobilfunkverflechtungen bezeichnet; siehe auch Quelle-Ziel-Matrix.

**Bias-Korrektur** Empirisch-statistisches Fehlerkorrekturverfahren bzw. Korrektur einer systematischen Abweichung, auch Bias genannt (engl.: bias correction).

**Bildungspendler** Pendelnde Schülerinnen und Schüler sowie Studierende zum Ausbildungsort (bspw. Schule oder Hochschule).

**Call Detail Records (CDRs)** Individuelle Informationen zu Mobiltelefonnutzenden sowie ihrer mobilen Aktivitäten, welche im Gegensatz zu den Signaldaten ereignisbasiert sind.

**Dynamische Mobilfunkdaten** Gezählte (dynamische) mobile Aktivitäten vom Start- zum Zielort durch Bewegung, auch als Quelle-Ziel-Matrix bezeichnet.

**Einpendler** Erwerbstätige Personen, die nicht in ihrer Arbeitsgemeinde wohnen.

**Erwerbslosenquote** Anteil der Erwerbslosen an den Erwerbspersonen (bestehend aus Erwerbstätigen und Erwerbslosen) (engl.: unemployment rate). Als erwerbslos gilt jede Person im erwerbsfähigen Alter, die in einem betrachteten Zeitraum nicht erwerbstätig ist, aber aktiv nach einer Tätigkeit sucht.

**Funkmast** Antenne zum Empfang und Senden von Signalen mobiler Endgeräte für die Kommunikation im Mobilfunknetz.

**Funktionale städtische Gebiete** Zusammensetzung aus einer Stadt und ihrem Pendlereinzugsgebiet (engl.: Functional Urban Areas).

**Geodaten** Digitale Informationen über Objekte und Landbedeckung der Erdoberfläche zur räumlichen Lage und weiterer Eigenschaften in einem geografischen Koordinatensystem.

**Georeferenzierung** In Bezug setzen eines Datensatzes in einen räumlichen Kontext.

**Gitterzelle** Geografische Einheit in Form eines Quadrats mit variierenden oder einheitlichen Gitterweiten mit Zell- und Raumbezug; sogenannte INSPIRE-konforme Gitterzellen stellen ein europaweit einheitliches geografisches Gitter bzw. Raster dar (engl.: grid cells).

**Kerndichteschätzung** Statistisches Verfahren zur Schätzung der Wahrscheinlichkeitsverteilung einer Zufallsvariablen.

**Konfidenzintervall** Bereich, in dem ein Parameter (z.B. der Mittelwert) mit einer gewissen Wahrscheinlichkeit liegt (engl.: confidence interval).

**Korrelation** Misst die Stärke einer statistischen Beziehung zweier Variablen zueinander; hier Verwendung des Pearson-Korrelationskoeffizienten, der den linearen Zusammenhang bzw. die Stärke des Zusammenhangs zweier Variablen ermittelt.

**Kurzdistanzwege** Bewegungsverflechtungen bzw. Pendlerwege mit einer zurückgelegten Distanz von bis zu 40 km.

**Mittlere quadratische Abweichung** Gibt in der Schätztheorie an, wie sehr ein Punktschätzer um den zu schätzenden Wert streut (engl.: mean squared error).

**Mobile Aktivitäten** Signal an einem Funkmast, welches durch eine Mindestverweildauer des mobilen Endgerätes in einem Untersuchungsgebiet bedingt wird.

**Mobilfunkdaten** Automatisch erfasste Daten/Aktivitäten mobiler Endgeräte aus dem Funknetz des Mobilfunkanbieters; weitere Unterscheidung in statische und dynamische Mobilfunkdaten sowie zwischen CDRs und Signaldaten.

**Mobilfunkdatenströme** Aggregierte zielgerichtete Mobilfunkbewegungen, werden weiterhin nach Ein- und Auspendlerströmen unterschieden; Pendant zu Pendlerströmen.

**Mobilfunknetz** Technische Infrastruktur eines Mobilfunkanbieters, die für die Kommunikation zwischen mobilen Endgeräten erforderlich ist und in einzelne Funkzellen unterteilt ist. Die Netzdichte orientiert sich grundsätzlich am verwendeten Mobilfunkstandard (GSM (2G), UMTS (3G), LTE (4G), 5G) sowie der regionalen Bevölkerungsdichte.

**Pendler** Erwerbstätige Personen, die einen Arbeitsweg vom Wohn- zum Arbeitsort zurücklegen, wobei sich der Arbeitsort vom Wohnort unterscheiden muss; auch bezeichnet als Berufspendler. Weitere Unterscheidung in übergemeindliche Pendler, sofern der Arbeitsort nicht in derselben Gemeinde wie der Wohnort liegt, andernfalls spricht man von innergemeindlichen Pendlern.

**Pendlerart** Unterscheidung der Berufspendler nach Ein- und Auspendlern.

**Pendlerbewegung** Zielgerichteter zurückgelegter Weg von und zur Arbeitsstelle, auch als Pendlerweg bezeichnet.

**Pendlerrechnung** Sekundärstatistik, die die benötigten Angaben zum Arbeits- und Wohnort der Erwerbsbevölkerung sowie die Merkmale der Pendler aus unterschiedlichen Statistiken bzw. Datenquellen heranzieht.

**Pendlerströme** Aggregierte zielgerichtete Pendlerbewegungen; werden weiterhin nach Ein- und Auspendlerströmen unterschieden.

**Pendlerverflechtungen** Durch die Gesamtheit der zielgerichteten Pendlerbewegungen entstandene Bewegungsmuster.

**Pendlerverhalten** Individuelles Verhalten der Berufspendler hinsichtlich der Verkehrsnutzung, wie bspw. Wahl des Verkehrsmittels.

**Quelle-Ziel-Matrix** Anonymisierte und aggregierte Bewegungen von Signalen bzw. mobilen Aktivitäten im Mobilfunknetz vom Start- zum Zielort und entsprechen dynamischen Mobilfunkdaten, auch als Mobilfunkbewegungen bezeichnet; siehe auch Bewegungsverflechtungen.

**Raster** Mehrere zusammenhängende Gitterzellen mit einer flächendeckenden, gleichmäßigen Anordnung bzw. Verteilung.

**Signaldaten** Mobile Aktivitäten bzw. Signale im Netz des Mobilfunkanbieters ohne Rückschluss auf die Art der getätigten Aktivitäten. Sie werden automatisch erzeugt, sofern das mobile Endgerät nicht ausgeschaltet ist oder sich im Flugmodus befindet.

**Simulation** Erzeugung von Zufallsvariablen mit Hilfe einer mathematischen Technik (bspw. Monte-Carlo-Simulation) zur Bestimmung der Unsicherheit, auch als künstliche Durchführung von Zufallsexperimenten verstanden.

**Small-Area-Verfahren** Methode zur Schätzung von Indikatoren oder Merkmalen auf einer tiefer gegliederten Ebene (engl.: small area estimation).

**Statische Mobilfunkdaten** Gezählte (statische) mobile Aktivitäten pro Ort und ausgewähltem Zeitraum ohne Bewegung, auch als Frequenzdaten bezeichnet.

**Tagespendler** Bewegungsverflechtungen potenzieller Berufspendler aus den Mobilfunkdaten, die lediglich tageweise (innerhalb von 24 Stunden) nachvollzogen werden.

**Tagessummen** Durchschnittliche Anzahl der mobilen Aktivitäten im Untersuchungsgebiet bzw. am potenziellen Arbeitsort; Pendant zur Angabe der Berufspendler nach Beschäftigungsumfang in der amtlichen Pendlerrechnung als Jahresdurchschnitt pro Gemeinde.

**Transformation** Konvertierung bzw. Umwandlung von Daten, um die Voraussetzungen statistischer Verfahren an die Verteilung der Daten bzw. an die Verteilung der Residuen zu erfüllen.

**Untersuchungsgebiet** Interessierende räumliche Einheit, auf deren räumlicher Ebene die Mobilfunkdaten aufbereitet werden; im Zusammenhang mit Pendler-/Mobilitätsanalysen auch als Zielort der Bewegungen verstanden.

**Verweildauer** Definierte Aufenthaltsdauer eines mobilen Endgerätes an einem Ort bzw. in einer Gitterzelle ohne Bewegung.

**Zensus** Volkszählung in Deutschland, genauer eine Bevölkerungs-, Gebäude- und Wohnungszählung, die die statistischen Ämter des Bundes und der Länder durchführen.

# Literaturverzeichnis


Arhipova, I., G. Berzins, E. Brekis, J. Binde, M. Opmanis, et al. (2020). Mobile phone data statistics as a dynamic proxy indicator in assessing regional economic activity and human commuting patterns. Expert Systems 37(5), e12530.

Augustijn, L. (2018). Berufsbedingte Pendelmobilität, Geschlecht und Stress. https://www.uni-due.de/imperia/md/content/soziologie/dbsf-2018-02.pdf. [Zugegriffen: 15.01.2021].

Bauer-Hailer, U. (2019). Berufspendler im Bundesländervergleich. https://www.statistik-bw.de/Service/Veroeff/Monatshefte/PDF/Beitrag19_02_02.pdf. [Zugegriffen: 07.12.2019].

BBSR (2017). Wachsen und Schrumpfen von Städten und Gemeinden. https://gis.uba.de/maps/resources/apps/bbsr/index.html?lang=de. [Zugegriffen: 11.06.2019].

BBSR (2021). Der demografische Wandel: ein wichtiger Faktor für die Entwicklung regionaler Teilmärkte. Dezembertagung des DGD-Arbeitskreises "Städte und Regionen" in Kooperation mit dem BBSR Bonn am 5. und 6. Dezember 2019 in Berlin. BBSR-Online-Publikation (01), 4–14.

BKG (2020a). Dokumentation: Digitales Landbedeckungsmodell für Deutschland – LBM-DE2018. https://sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/lbm-de2018.pdf. [Zugegriffen: 26.02.2022].

BKG (2020b). Dokumentation: Geographische Gitter für Deutschland – GeoGitter. https://sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/geogitter.pdf. [Zugegriffen: 20.01.2022].

BKG (2021a). Dokumentation: Haushalte Einwohner Bund – HH-EW-Bund. https:

//sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/hh-ew-bund.pdf. [Zugegriffen: 26.02.2022].

BKG (2021b). Dokumentation: Hausumringe Deutschland – HU-DE. https://sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/hu-de.pdf. [Zugegriffen: 10.03.2022].

Buch, T., S. Hamann, A. Niebuhr, und A. Rossen (2014). What makes cities attractive? The determinants of urban labour migration in Germany. Urban Studies 51(9), 1960–1978.

Bundesagentur für Arbeit (2020). Grundlagen: Qualitätsbericht – Statistik der sozialversicherungspflichtigen und geringfügigen Beschäftigung. https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Qualitaetsberichte/Generische-Publikationen/Qualitaetsbericht-Statistik-Beschaeftigung.pdf?__blob=publicationFile&v=8. [Zugegriffen: 07.10.2021].

Bundesagentur für Arbeit (2021). Glossar der Statistik der Bundesagentur für Arbeit (BA) – Grundlagen: Definitionen. https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Definitionen/Glossare/Generische-Publikationen/Gesamtglossar.pdf?__blob=publicationFile&v=9. [Zugegriffen: 01.03.2021].

Bundesagentur für Arbeit (2022a). Arbeitslosenquote und Unterbeschäftigungsquote. https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Definitionen/Berechnung-der-Arbeitslosenquote/Berechnung-der-Arbeitslosenquote-Nav.html. [Zugegriffen: 16.10.2022].

Bundesagentur für Arbeit (2022b). Pendlerverflechtungen der sozialversicherungspflichtig Beschäftigten nach Ländern - Stichtag: 30.06.2016. https://statistik.arbeitsagentur.de/SiteGlobals/Forms/Suche/Einzelheftsuche_Formular.html?nn=20934&topic_f=beschaeftigung-pendler-blxbl&dateOfRevision=201606-202106. [Zugegriffen: 29.10.2022].

Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (2017). 26. Tätigkeitsbericht 2015–2016. https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Taetigkeitsberichte/26TB_15_16.pdf?__blob=publicationFile&v=9. [Zugegriffen: 25.05.2021].

Burgard, J. P., R. Münnich, und T. Zimmermann (2016). Impact of sampling designs in small area estimation with applications to poverty measurement. In M. Pratesi (Hrsg.), Analysis of Poverty Data by Small Area Estimation, S. 85–108. Hoboken: John Wiley & Sons.

Burgdorf, M. (2010). Disaggregation von Bevölkerungsdaten mittels ATKIS Basis DLM. In J. Strobl et al. (Hrsg.), Angewandte Geoinformatik 2010 - 22. AGIT-Symposium, S. 474–483. Heidelberg: Wichmann.

Casas-Cordero, C., J. Encina, und P. Lahiri (2016). Poverty mapping for the Chilean comunas. In M. Pratesi (Hrsg.), Analysis of Poverty Data by Small Area Estimation, S. 379–403. Hoboken: John Wiley & Sons.

Celeux, G. und J. Diebolt (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly (2), 73–82.

Costa, A., A. Satorra, und E. Ventura (2006). Improving small area estimation by combining surveys: New perspectives in regional statistics. Statistics and Operations Research Transactions 30(1), 101–121.

Daas, P., M. Puts, B. Buelensand, und P. van den Hurk (2013). Big data and official statistics. In New Techniques and Technologies for Statistics (NTTS) Conference 2013. Brüssel, Belgien.

Datta, G. S. und P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10, 613–627.

De Jonge, E., M. van Pelt, und M. Roos (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. Discussion paper 201214. Statistics Netherlands.

De Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, et al. (2016). Assessing the Quality of Mobile Phone Data as a Source of Statistics. In European Conference on Quality in Official Statistics (Q2016). Madrid, Spanien.

Dettmer, B. und W. Emmel (2018). Pendlerrechnung Hessen – Methodenbericht. StaWi – Staat und Wirtschaft in Hessen (2), 29–36.

Dettmer, B. und I. Wolf (2018). Mobilität der hessischen Bevölkerung. StaWi – Staat und Wirtschaft in Hessen (2), 3–11.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. R. Stevens, et al. (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences 111(45), 15888–15893.

Dijkstra, L. und H. Poelman (2011). Archive:European cities – the EU-OECD functional urban area definition. https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:European_cities_%E2%80%93_the_EU-OECD_functional_urban_area_definition#A_harmonised_definition. [Zugegriffen: 11.06.2019].

Douglass, R. W., D. Meyer, M. Ram, D. Rideout, und D. Song (2015). High resolution population estimates from telecommunications data. EPJ Data Science 4(4), 1–13.

Elhorst, P. J. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanationa. Journal of Economic Surveys 17(5), 709–748.

European Commission (2019). Study: City data from LFS and Big Data (Regional Policy: Newsroom). https://ec.europa.eu/regional_policy/en/newsroom/news/2019/06/26-06-2019-study-city-data-from-lfs-and-big-data. [Zugegriffen: 11.07.2019].

European Parliament and Council (2003). Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS). Official Journal of the European Union 154(1), 1–41.

Europäische Kommission und Eurostat (2021). Applying the Degree of Urbanisation — A methodological manual to define cities, towns and rural areas for international comparisons — 2021 edition. Publications Office of the European Union.

Eurostat (2017). City statistics (urb): National Reference Metadata in Euro SDMX Metadata Structure (ESMS). https://ec.europa.eu/eurostat/cache/metadata/EN/urb_esms_de.htm. [Zugegriffen: 09.11.2018].

Eurostat (2018a). Dataset details: Harmonised unemployment rate by sex. https://ec.europa.eu/eurostat/web/products-datasets/-/teilm020. [Zugegriffen: 29.10.2018].

Eurostat (2018b). Glossary: Functional urban area. https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area. [Zugegriffen: 13.05.2018].

Eurostat (2018c). NUTS - Systematik der Gebietseinheiten für die Statistik: Hintergrund. https://ec.europa.eu/eurostat/web/nuts/background. [Zugegriffen: 21.08.2019].

Eurostat (2019a). DataCollection: Precision level DCF. https://datacollection.jrc.ec.europa.eu/wordef/precision-level-dcf. [Zugegriffen: 06.06.2019].

Eurostat (2019b). EU Labour Force Survey Database User Guide. https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf. [Zugegriffen: 22.08.2019].

Eurostat (2019c). Labour Force Survey in the EU, candidate and EFTA countries: Main characterstics of national surveys, 2018. https://ec.europa.eu/eurostat/de/web/products-statistical-reports/-/KS-FT-19-008?inheritRedirect=true. [Zugegriffen: 14.01.2021].

Eurostat (2019d). Städte (Urban Audit): Datenbank. https://ec.europa.eu/eurostat/de/web/cities/data/database. [Zugegriffen: 21.08.2019].

Fay, R. E. und R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74(366), 269–277.

Fina, S., C. Gerten, K. Gehring-Fitting, und J. Rönsch (2019). Geomonitoring und die große Transformation – Methoden zur kritischen Bewertung nachhaltiger Raumentwicklung. ILS-TRENDS [extra]. https://www.ils-forschung.de/files_publikationen/pdfs/trends-dez-19.pdf. [Zugegriffen: 07.11.2021].

Franconi, L., D. Ichim, M. D'Alò, und S. Cruciani (2017). Guidelines for Labour Market Area delineation process: From definition to dissemination. https://ec.europa.eu/eurostat/cros/system/files/guidelines_for_lmas_production08082017_rev300817.pdf. [Zugegriffen: 15.01.2021].

Galiana, L., B. Sakarovitch, und Z. Smoreda (2018). Understanding socio-spatial segregation in French cities with mobile phone data. https://dgins2018.statisticsevents.ro/wp-content/uploads/2018/10/08-FR-dgins_segregation_1800905.pdf. [Zugegriffen: 25.06.2022].

Gans, P. (2017). Urban population development in Germany (2000-2014): The contribution of migration by age and citizenship to reurbanisation. Comparative Population Studies 42, 319–352.

González-Manteiga, W., M. J. Lombardía, I. Molina, D. Morales, und L. Santamaría (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics & Data Analysis 52(12), 5242–5252.

Groß, M. (2018). Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data. R package version 2.0.0.

Groß, M., U. Rendtel, T. Schmid, S. Schmon, und N. Tzavidis (2017). Estimating the density of ethnic minorities and aged people in berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. Journal of the Royal Statistical Society, Series A 180(1), 161–183.

Grohmann, H. (1985). Vom theoretischen Konstrukt zum statistischen Begriff – das Adäquationsproblem. Allgemeines statistisches Archiv 69, 1–15.

Grohmann, H. (2011). Volkszählung und Mikrozensus. In H. Grohmann, W. Krämer, und A. Steger (Hrsg.), Statistik in Deutschland, S. 207–221. Berlin/Heidelberg: Springer.

Grözinger, G. (2009). Achtung Lebensgefahr! Indirekte Effekte regionaler Arbeitslosigkeit auf Lebensweise und -qualität. European Journal of Economics and Economic Policies: Intervention 6(1), 12–24.

Grözinger, G. (2018). Regionale Arbeitslosigkeit: Falsche Eindrücke von Stadt-Land-Differenzen. Wirtschaftsdienst 98(1), 68–70.

Haas, A. und S. Hamann (2008). Pendeln – ein zunehmender Trend, vor allem bei Hochqualifizierten. Ost-West-Vergleich. IAB-Kurzbericht 6/2008, Nürnberg.

Hadam, S. (2018). Use of mobile phone data for official statistics. METHODS - APPROACHES - DEVELOPMENTS: Information of the German Federal Statistical Office 1(2), 6–9.

Hadam, S. (2021). Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik. AStA Wirtschafts- und Sozialstatistisches Archiv 15, 197–235.

Hadam, S., T. Schmid, und J. Simm (2020). Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland. In B. Klumpe, J. Schröder, und M. Zwick (Hrsg.), Qualität bei zusammengeführten Daten, S. 27–44. Wiesbaden: Springer VS.

Hadam, S., N. Würz, und A.-K. Kreutzmann (2020). Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany. Refubium - Freie Universität Berlin Repository.

Heinzel, A. (2006). Volkszählung 2011: Deutschland bereitet sich auf den registergestützten Zensus vor. Berliner Statistik 7, 321–328.

Hilbe, J. M. (2011). Negative Binomial Regression (2. Aufl.). New York: Cambridge University Press.

ILO (2013). Resolution concerning statistics of work, employment and labour underutilization. http://www.ilo.ch/wcmsp5/groups/public/---dgreports/---stat/documents/normativeinstrument/wcms_230304.pdf. [Zugegriffen: 15.06.2021].

ILO (2018). Unemployment rate. https://ilostat.ilo.org/resources/concepts-and-definitions/description-unemployment-rate/. [Zugegriffen: 26.06.2022].

IT.NRW (2018). Raum neu gefasst: Wie verteilen sich Arztpraxen in Nordrhein-Westfalen? Potenziale georeferenzierter Auswertungen des statistischen Unternehmensregisters. Statistik kompakt 09/2018. https://webshop.it.nrw.de/gratis/Z259%20201859.pdf. [Zugegriffen: 07.11.2021].

IT.NRW (2019). Berufspendler 2011 – 2018 nach Pendlerart, Beschäftigungsumfang und Geschlecht. https://www.it.nrw/statistik/eckdaten/berufspendler-2011-2018-nach-pendlerart-beschaeftigungsumfang-und-geschlecht. [Zugegriffen: 07.12.2019].

IT.NRW (2020). Pendlerrechnung Nordrhein-Westfalen – Methodenbeschreibung. https://www.pendleratlas.nrw.de/pdf/Pendlerrechnung_Methodenbeschreibung_lang.pdf. [Zugegriffen: 01.03.2021].

Jacques, D. C. (2018). Mobile Phone Metadata for Development. Technical Report. arXiv preprint arXiv:1806.03086.

Jensen, J. L. W. V. et al. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica 30, 175–193.

Jiang, J., P. Lahiri, S.-M. Wan, und C.-H. Wu (2001). Jackknifing in the Fay-Herriot model with an example. Technical Report, Department of Statistics, University of Nebraska, Lincoln.

Kirchner, T., F. Pflanz, A. Techen, und L. Wagenknecht (2014). Kleinräumige Gliederung, Georeferenzierung und Rasterdarstellung im Zensus. Zeitschrift für amtliche Statistik Berlin-Brandenburg (3), 28–32.

Klammer, U. und K. Menke (2020). Gender-Datenreport. https://www.bpb.de/izpb/307413/geschlechterdemokratie. [Zugegriffen: 15.01.2021].

Koebe, T., A. Arias-Salazar, N. Rojas-Perilla, und T. Schmid (2022). Intercensal updating using structure-preserving methods and satellite imagery. Journal of the Royal Statistical Society, Series A 185(Suppl. 2), S170–S196.

Kompf, M. (2020). Entfernungsberechnung. https://www.kompf.de/gps/distcalc.html. [Zugegriffen: 12.10.2020].

Kosfeld, R. und C. Dreger (2006). Thresholds for employment and unemployment. A spatial analysis of German regional labour markets 1999–2000. Papers in Regional Science 85, 523–542.

KOSIS-Gemeinschaft Urban Audit (2013). Das deutsche Urban Audit - Städtevergleich im Europäischen Statistischen System. https://ec.europa.eu/eurostat/cache/metadata/Annexes/urb_esms_de_an4.pdf. [Zugegriffen: 25.06.2022].

Krzossa, T. (2019). Nachgefragt: Was genau kann welche Frequenz? (Vodafone Newsroom). https://www.vodafone.de/medien/netz/5g-auktion-welche-frequenz-eignet-sich-wofuer. [Zugegriffen: 05.04.2019].

Lahiri, P. und J. Suntornchost (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. The Indian Journal of Statistics 77(2), 312–320.

Leyk, S., A. E. Gaughan, S. B. Adamo, A. de Sherbinin, D. Balk, et al. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. Earth System Science Data 11(3), 1385–1409.

Lloyd, C. T., A. Sorichetta, und A. J. Tatem (2017). High resolution global gridded data for use in population studies. Scientific Data 4, 170001.

Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software 9(1), 1–19.

Makita, N., M. Kimura, M. Terada, M. Kobayashi, und Y. Oyabu (2013). Can mobile phone network data be used to estimate small area population? A comparison from Japan. Statistical Journal of the IAOS 29, 223–232.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, et al. (2015). Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263–281.

Marino, M. F., M. G. Ranalli, N. Salvati, und M. Alfo (2019). Semiparametric empirical best prediction for small area estimation of unemployment indicators. Annals of Applied Statistics 13(2), 1166–1197.

Martini, A. und S. Loriga (2017). Small Area Estimation of employment and unemployment for Local Labour Market Areas in Italy. https://www.dst.dk/ext/4299453410/0/formid/4-2-Small-Area-Estimation-of-employment-and-unemployment-for-Local_Labour-Market-Areas-in-Italy--pdf. [Zugegriffen: 10.11.2018].

Möbert, J. (2018). The German housing market in 2018. https://www.dbresearch.com/PROD/RPS_EN-PROD/PROD0000000000460528/The_German_housing_market_in_2018.pdf. [Zugegriffen: 25.06.2022].

Molina, I. und E. Strzalkowska-Kominiak (2020). Estimation of proportions in small areas: application to the labour force using the Swiss Census Structural Survey. Journal of the Royal Statistical Society, Series A 183(1), 281–310.

Norman, P., L. Simpson, und A. Sabater (2008). 'Estimating with Confidence' and hindsight: New UK small area population estimates for 1991. Population, Space and Place 14(5), 449–472.

Novak, J., R. Ahas, A. Aasa, und S. Silm (2013). Application of mobile phone location data in mapping of commuting patterns and functional regionalization: a pilot study of Estonia. Journal of Maps 9(1), 10–15.

Patuelli, R., D. A. Griffitha, M. Tiefelsdorf, und P. Nijkamp (2011). Spatial filtering and eigenvector stability: Space-time models for German unemployment data. International Regional Science Review 34(2), 253–280.

Pereira, L. N., J. Mendes, und P. Coelho (2011). Estimation of unemployment rates in small areas of Portugal: A best linear unbiased prediction approach versus a hierarchical bayes approach. 17th European Young Statisticians Meeting, Faculdade de CIâncias e Tecnologia - Universidade Nova de Lisboa.

Prasad, N. G. N. und J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association 85(409), 163–171.

R Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Pütz, T. (2015). Verkehrsbild Deutschland. BBSR-Analysen kompakt 15/2015. Bonn: BBSR (Bundesinstitut für Bau-, Stadt- und Raumforschung im Bundesamt für Bauwesen und Raumordnung).

Raghunathan, T. E., D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, et al. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association 102(478), 474–486.

Rao, J. N. K. und I. Molina (2015). Small area estimation (2. Aufl.). Hoboken: John Wiley & Sons.

Rees, P., D. Brown, P. Norman, und D. Dorling (2003). Are socioeconomic inequalities in mortality decreasing or increasing within some British regions? An observational study, 1990-1998. Journal of Public Health Medicine 25(3), 208–214.

Rendtel, U., W. Seidel, C. Müller, F. Meinfelder, J. Wagner, et al. (2022). Statistik zwischen Data Science, Artificial Intelligence und Big Data: Beiträge aus dem Kolloquium „Make Statistics great again“. AStA Wirtschafts- und Sozialstatistisches Archiv 16, 97–147.

Saidani, Y., S. Bohnensteffen, und S. Hadam (2022). Qualität von Mobilfunkdaten – Projekterfahrungen und Anwendungsfälle aus der amtlichen Statistik. WISTA - Wirtschaft und Statistik 74(5), 55–297.

Schmid, T., F. Bruckschen, N. Salvati, und T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society, Series A 180(4), 1163–1190.

Schug, F., D. Frantz, S. van der Linden, und P. Hostert (2021). Gridded population mapping for germany based on building density, height and type from earth observation data using census disaggregation and bottom-up estimates. PLoS ONE 16(3), e0249044.

Simpson, S., I. Diamond, P. Tonkin, und R. Tye (1996). Updating Small Area Population Estimates in England and Wales. Journal of the Royal Statistical Society, Series A 159(2), 235–247.

Slud, E. V. und T. Maiti (2006). Mean-squared error estimation in transformed Fay-Herriot models. Journal of the Royal Statistical Society, Series B 68(2), 239–257.

Statistisches Bundesamt (2016). Zensus 2011 (Qualitätsbericht). `https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/zensus-2011.pdf`;

jsessionid=51691D6AC6AAFF54705C71CDD61EFBE5.live741?__blob=publicationFile. [Zugegriffen: 11.07.2019].

Statistisches Bundesamt (2017a). Erwerbstätigkeit – Berufspendler. https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/pendler1.html;jsessionid=0AD869EFDA68AE904675EB7106D3B7F9.live721#fussnote-1-103722. [Zugegriffen: 18.05.2021].

Statistisches Bundesamt (2017b). Erwerbstätigkeit – Bildungspendler. https://www.destatis.de/DE/Themen/Arbeit/Arbeitsmarkt/Erwerbstaetigkeit/Tabellen/pendler2.html. [Zugegriffen: 15.05.2021].

Statistisches Bundesamt (2017c). Qualitätsbericht Mikrozensus 2016. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2016.pdf;jsessionid=9F2E5A1DF05CAF3E49C54CDC914A59BF.live721?__blob=publicationFile. [Zugegriffen: 18.05.2021].

Statistisches Bundesamt (2019a). Bevölkerungsdarstellung mit Mobilfunkdaten. https://www.destatis.de/DE/Service/EXDAT/Datensaetze/mobilfunkdaten.html. [Zugegriffen: 18.11.2021].

Statistisches Bundesamt (2019b). Fortschreibung des Bevölkerungsstandes (Bevölkerungsfortschreibung 2017). Qualitätsbericht. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/bevoelkerungsfortschreibung-2017.pdf?__blob=publicationFile. [Zugegriffen: 11.07.2019].

Statistisches Bundesamt (2020a). 2040 wird voraussichtlich jeder vierte Mensch in Deutschland alleine wohnen – Pressemitteilung Nr. 069 vom 2. März 2020. https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/03/PD20_069_122.html. [Zugegriffen: 24.11.2021].

Statistisches Bundesamt (2020b). Ausstattung mit Gebrauchsgütern – Daten aus den Laufenden Wirtschaftsrechnungen (LWR) zur Ausstattung privater Haushalte mit Informationstechnik. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsgueter/Tabellen/a-infotechnik-d-lwr.html. [Zugegriffen: 17.03.2021].

Statistisches Bundesamt (2021a). Ausstattung mit Gebrauchsgütern – Daten aus den Laufenden Wirtschaftsrechnungen (LWR) zur Ausstattung privater Haushalte mit Informationstechnik. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsgueter/Tabellen/a-infotechnik-d-lwr.html. [Zugegriffen: 19.02.2022].

Statistisches Bundesamt (2021b). Fortschreibung des Bevölkerungsstandes (Bevölkerungsfortschreibung 2020). Qualitätsbericht. https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/bevoelkerungsfortschreibung-2020.pdf?__blob=publicationFile. [Zugegriffen: 04.11.2021].

Statistisches Bundesamt (2021c). Mobilitätsindikatoren auf Basis von Mobilfunkdaten. https://www.destatis.de/DE/Service/EXDAT/Datensaetze/mobilitaetsindikatoren-mobilfunkdaten.html. [Zugegriffen: 17.03.2021].

Statistisches Bundesamt (2021d). Registered unemployed, unemployment rate by sex. https://www.destatis.de/EN/Themes/Labour/Labour-Market/Unemployment/Tables/lrarb002.html. [Zugegriffen: 15.01.2021].

Statistisches Bundesamt (2021e). Strukturvergleich von Mobilfunkdaten zweier Mobilfunkanbieter. https://www.destatis.de/DE/Service/EXDAT/Datensaetze/mobilfunkanbieter-strukturvergleich.html. [Zugegriffen: 21.11.2021].

Statistisches Bundesamt (2022a). Equipment of households with information and communication technology (Germany). https://www.destatis.de/EN/Themes/Society-Environment/Income-Consumption-Living-Conditions/Equipment-Consumer-Durables/Tables/liste-equipment-households-information-communication-technology-germany.html#55714. [Zugegriffen: 27.10.2022].

Statistisches Bundesamt (2022b). Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten. https://www.destatis.de/DE/Service/EXDAT/Datensaetze/bevoelkerung-geo-mobilfunkdaten.html. [Zugegriffen: 02.11.2022].

Statistisches Landesamt Baden-Württemberg (2019a). Berufspendlerrechnung für Baden-Württemberg – Ergebnisse. https://www.statistik-bw.de/Pendler/Ergebnisse/. [Zugegriffen: 18.05.2021].

Statistisches Landesamt Baden-Württemberg (2019b). Methode der Berufspendlerrechnung. https://www.statistik-bw.de/Pendler/Methode/. [Zugegriffen: 18.05.2021].

Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, et al. (2017). Mapping poverty using mobile phone and satellite data. Journal of the Royal Society Interface 14(127), 20160690.

Steinnocher, K., F. Petrini, T. Tötzer, und J. Weichselbaum (2005). Räumliche Disaggregation von sozio-ökonomischen Daten. In J. Strobl et al. (Hrsg.), Angewandte Geoinformatik 2005 - 17. AGIT-Symposium, S. 702–707. Heidelberg: Wichmann.

Stevens, F. R., A. E. Gaughan, C. Linard, und A. J. Tatem (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLoS ONE 10(2), e0107042.

Sugasawa, S. und T. Kubokawa (2017). Transforming response values in small area prediction. Computational Statistics & Data Analysis 114, 47–60.

Toole, J. L., Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González, und D. Lazer (2015). Tracking employment shocks using mobile phone data. Journal of the Royal Society Interface 12(107), 20150185.

Tzavidis, N., L. C. Zhang, A. Luna, T. Schmid, und N. Rojas-Perilla (2018). From start to finish: A framework for the production of small area official statistics. Journal of the Royal Statistical Society, Series A 181(4), 927–979.

U.S. Bureau of Labor Statistics (2021). Labor force statistics from the current population survey: Concepts and definitions. https://www.bls.gov/cps/definitions.htm#:~:text=The%20unemployment%20rate%20represents%20the,%C3%B7%20Labor%20Force)%20x%20100. [Zugegriffen: 17.10.2022].

van den Brakel, J. A., B. Buelens, und H.-J. Boonstra (2016). Small area estimation to quantify discontinuities in repeated sample surveys. Journal of the Royal Statistical Society, Series A 179, 229–250.

Venables, W. N. und B. D. Ripley (2002). Modern Applied Statistics With S (4. Aufl.). New York: Springer.

Wand, M. und M. Jones (1994). Multivariate plug-in bandwidth selection. Computational Statistics 9(2), 97–116.

Wiengarten, L. und M. Zwick (2017). Neue digitale Daten in der amtlichen Statistik. WISTA - Wirtschaft und Statistik (5/2017), 19–30.

Wonka, E., I. Kaminger, und G. Katzlberger (2009). Regionalstatistische Auswertungen mit geographischen Rastern in der Raumplanung. Informationen zur Raumentwicklung (10/11), 661–675.

Zeileis, A., C. Kleiber, und S. Jackman (2008). Regression Models for Count Data in R. Journal of Statistical Software 27(8), 1–25.

Zwick, M. (2016). Big Data und amtliche Statistik. In B. Keller, H.-W. Klein, und S. Tuschl (Hrsg.), Marktforschung der Zukunft - Mensch oder Maschine? Bewährte Kompetenzen in neuem Kontext, S. 157–172. Wiesbaden: Springer Gabler.

# Zusammenfassungen

## Kurzzusammenfassungen auf Deutsch

### Zusammenfassung: Kleinräumige Prädiktion von Bevölkerungszahlen basierend auf Mobilfunkdaten aus Deutschland

Im Rahmen der allgemein fortschreitenden Digitalisierung ist die amtliche Statistik gefordert, neue Datenquellen zu erforschen und einzusetzen. Hierbei sind zuverlässige Kenntnisse über die Verteilung der Bevölkerung und die Einwohnerzahl eines Landes auf kleinstmöglicher geografischer Ebene für eine solide evidenzbasierte Politikgestaltung unerlässlich. Mobilfunkdaten haben das Potenzial diese Herausforderungen zu lösen und können zu einer dynamischen und zeitnahen Schätzung der Bevölkerung beitragen.

Dieser Artikel befasst sich folglich mit der Fragestellung, inwieweit Mobilfunkdaten geeignet sind, die Bevölkerung valide und kleinräumig darzustellen. Hierzu werden die in INSPIRE-konformen Gitterzellen regional tief gegliederten Mobilfunkdaten mit den analogen Zellen des Zensus 2011 abgeglichen. Die Ergebnisse zeigen vom Grundsatz, dass die Bevölkerung mit den vorliegenden Mobilfunkdaten teilweise gut abgebildet werden kann. Durch Verwendung einer nicht-parametrischen Kerndichteschätzung können zudem unabhängig von der räumlichen Aufbereitung und den zugrundeliegenden Geometrien, die Aufenthaltsorte der deutschen Bevölkerung wiedergegeben werden. Beobachtbare Unterschiede in der Bevölkerungsdarstellung mittels Mobilfunkdaten und den Zensuswerten können teilweise durch die zeitliche Differenz zwischen den Mobilfunkdaten einer statistischen Woche aus verschiedenen Monaten des Jahres 2017 und den Zensusdaten aus dem Jahr 2011 erklärt werden. Aber auch das durch den Mobilfunkdatenanbieter angewandte Hochrechnungsverfahren könnte hierfür ursächlich sein.

**Schlüsselwörter**: Mobilfunkdaten, Mobilfunkaktivitäten, Bevölkerungsdarstellung, Kerndichteschätzung, Zensus 2011

## Zusammenfassung: Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten

Kleinräumige und aktuelle Bevölkerungszahlen sind für politische Entscheidungsfindungen unerlässlich. Die Bevölkerungsfortschreibung ermöglicht die Angabe aktueller Einwohnerzahlen auf geografischer Ebene der Gemeinden. Die Einwohnerzahl wird hierbei auf Basis des Zensus 2011 anhand von Angaben der Statistiken zu Geburten und Sterbefällen sowie der Wanderungsstatistik laufend fortgeschrieben. Um den wachsenden Bedarf an kleinräumigeren Bevölkerungszahlen kurzfristig zu decken, wird die Bevölkerungsfortschreibung mit einem neuen experimentellen Lösungsansatz ergänzt.

Im Projekt *Experimentelle georeferenzierte Bevölkerungszahl auf Basis der Bevölkerungsfortschreibung und Mobilfunkdaten* werden im Rahmen eines Verteilungsverfahrens die Ergebnisse der Bevölkerungsfortschreibung von der Gemeindeebene anhand von Mobilfunkdaten bundesweit auf INSPIRE-konforme 1x1 km Gitterzellen umverteilt und mittels einer interaktiven Karte frei nutzbar zur Verfügung gestellt. Mobilfunkdaten bieten aufgrund ihres starken Zusammenhangs mit der Bevölkerungsverteilung sowie ihrer hohen zeitlichen und räumlichen Auflösung eine geeignete Datengrundlage für die kleinräumige Verteilung der amtlichen Einwohnerzahl. Auf Basis zusätzlicher Geodaten der deutschen Landesvermessung, wie bspw. dem Landbedeckungsmodell für Deutschland, werden die resultierenden experimentellen georeferenzierten Bevölkerungszahlen auf Plausibilität geprüft und Verbesserungen in der Mobilfunkdatenaufbereitung hergeleitet und umgesetzt. Im Ergebnis resultieren bundesweit experimentelle georeferenzierte Bevölkerungszahlen, die in urbanen, dicht besiedelten Regionen plausibel erscheinen, jedoch im ländlichen, weniger dicht besiedelten Raum zu systematischen Fehlschätzungen neigen.

**Schlüsselwörter:** Mobilfunkdaten, Bevölkerungsfortschreibung, Georeferenzierung, Experimentell, INSPIRE-konform, Bevölkerungszahl

## Zusammenfassung: Pendler Mobil: Die Verwendung von Mobilfunkdaten zur Unterstützung der amtlichen Pendlerstatistik

Die Verfügbarkeit von kleinräumigen und aktuellen Pendlerverflechtungen sind für politische wie auch kommunale Entscheidungsfindungen von hoher Bedeutung. Aus dem Pendlerverhalten lassen sich Rückschlüsse auf Arbeitsmarktregionen und die Verteilung der Wohnbevölkerung ziehen, was unter anderem zu einer laufenden Verbesserung der Verkehrsinfrastruktur beiträgt. Die dafür notwendigen Daten veröffentlicht die amtliche Pendlerrechnung. Jedoch weist sie Verbesserungspotenzial im Hinblick auf die zeitliche und räumliche Darstellung der Pendlerverflechtungen von Erwerbstätigen sowie eine fachliche Erweiterung hinsichtlich der Bildungs-

pendler auf.

Dieser Artikel beschreibt die mit dem Projekt *Pendler Mobil* geprüften Erweiterungsmöglichkeiten der amtlichen Pendlerrechnung auf Basis von Quelle-Ziel-Matrizen aus Mobilfunkdaten. Mobilfunkdaten stellen aufgrund ihrer zeitlichen Aktualität und räumlich feinen Auflösung eine robuste Datengrundlage zur flexiblen Abbildung von potenziellen und regelmäßigen Pendlerbewegungen dar. Die potenzielle Leistungsfähigkeit der Mobilfunkdaten ermöglicht damit eine externe Validierung bestehender Pendlerrechnungen oder Pendlerstatistiken sowie eine beiderseitige Ergänzung zur Ermittlung und Darstellungen weiterer Formen des Pendelns der Erwerbsbevölkerung.

Am Fallbeispiel des Bundeslandes Nordrhein-Westfalen werden im Folgenden Gemeinsamkeiten und Unterschiede der übereinstimmenden Pendlerverflechtungen auf Basis von Mobilfunkdaten und der amtlichen Pendlerrechnung erörtert. Dabei gehen wir auf die Herausforderungen der Aufbereitung und Definition geeigneter Mobilfunkdaten durch den Datenanbieter sowie weitere Einflüsse auf die Mobilfunkdaten, wie bspw. durch die zurückgelegte Distanz oder die Verweilzeiten mobiler Aktivitäten, ein. Besonders die Unterschätzung der mobilen Pendlerströme im Vergleich zur amtlichen Pendlerrechnung legt nahe, Modifizierungsansätze der Mobilfunkdaten zu diskutieren. Im Ergebnis können die vorliegenden Mobilfunkdaten potenziell die amtliche Pendlerrechnung durch kleinräumige Pendlerbewegungen in Städten in Form einer erweiterten Zielorts-Bestimmung unterstützen und die Identifizierung von stark frequentierten Arbeitsorten in Städten ermöglichen.

**Schlüsselwörter:** Mobilfunkdaten, Pendlerrechnung, Quelle-Ziel-Matrix, Bewegung, Mobilität, Berufspendler

## Zusammenfassung: Schätzung der regionalen Erwerbslosigkeit mit Mobilfunkdaten für funktionale städtische Gebiete in Deutschland

In mehreren Ländern führt das anhaltende Wachstum der Städte mit ihren besseren Beschäftigungsmöglichkeiten zu verstärkten arbeitsbedingten Pendlerströmen. Obwohl immer mehr Menschen in die Städte pendeln und ziehen, weist der Arbeitsmarkt in städtischen Gebieten höhere Erwerbslosenquoten auf als das Umland. Dieses Phänomen wird auf regionaler Ebene anhand einer alternativen Definition der Erwerbslosenquote untersucht, in der das Pendlerverhalten einbezogen wird. Für das Bundesland Nordrhein-Westfalen in Deutschland werden Daten aus der Arbeitskräfteerhebung mit dynamischen Mobilfunkdaten unter Verwendung von Small-Area-Modellen kombiniert. Aus methodischer Sicht wird ein transformiertes Fay-Herriot-Modell mit Bias-Korrektur für die Schätzung der Erwerbslosenquoten angewandt. Unter Einbezug der Bias-Korrektur wird die mittlere quadratische Abweichung mit einem parame-

trischen Bootstrap-Verfahren geschätzt. Die Leistungsfähigkeit der vorgeschlagenen Methode wird in einer Fallstudie unter Verwendung von amtlichen Daten sowie in modellbasierten Simulationen bewertet. Die Ergebnisse der Anwendung zeigen, dass die (um Pendler bereinigten) Erwerbslosenquoten in deutschen Städten niedriger sind, als die traditionell ermittelten amtlichen Erwerbslosenquoten indizieren.

**Schlüsselwörter**: Bias-Korrektur, Fay-Herriot-Modell, Mittlere quadratische Abweichung, Small-Area-Schätzung, Erwerbslosenquote

## Abstracts in English

### Abstract: Small-scale prediction of population figures based on mobile network data from Germany

In the ongoing process of overall digitization, official statistics are challenged to explore and use new data sources. Within this context, reliable knowledge of a country's population distribution and population size at the smallest possible geographic scale is essential for sound evidence-based policy making. Mobile network data have the potential to address these challenges and can contribute to dynamic and timely population estimation.

This article consequently addresses the question to what extent mobile network data are suitable to represent the population in a valid and small-scale manner. For this purpose, the mobile network data, which are regionally deeply subdivided into INSPIRE-compliant grid cells, are compared with the analogous cells of the 2011 census. The results show in principle that the population can be partially well represented with the available mobile network data. Moreover, by using a non-parametric kernel density estimation, the locations of the German population can be reproduced independent of the spatial processing and underlying geometries. Observable differences in population representation using mobile network data and census values can be partly explained by the temporal difference between mobile network data of a statistical week from different months of 2017 and census data from 2011. However, the extrapolation procedure used by the mobile network data provider could also be responsible for this.

**Keywords**: Mobile network data, Mobile network activity, Population representation, Kernel density estimation, Census 2011

## Abstract: Experimental georeferenced population figure based on intercensal population updates and mobile network data

Small-area and up-to-date population figures are essential for policy decision-making. The intercensal population update makes it possible to provide current population figures at the geographic level of municipalities. The number of inhabitants is continuously updated on the basis of the 2011 census using data from statistics on births and deaths and migration statistics. In order to satisfy the demand for small-area population figures in the short term, the population update is supplemented with a new experimental approach.

In the project *Experimental georeferenced population figure based on intercensal population updates and mobile network data*, the results of the intercensal population update are redistributed nationwide from the municipality level to INSPIRE-compliant 1x1 km grid cells using mobile network data and made available for free public use on an interactive map. Mobile network data offer a suitable data basis for the small-area distribution of the official population due to their strong correlation with the distribution of the population as well as their high temporal and spatial resolution. Based on additional geodata from the German Land Survey, such as the land cover model for Germany, the resulting experimental georeferenced population figures are tested for plausibility and improvements in the mobile network data processing are derived and implemented. As result, nationwide experimental georeferenced population figures are obtained, that appear plausible in urban, densely populated regions, but tend to be systematically misallocated in rural, less densely populated areas.

**Keywords**: Mobil network data, Intercensal population update, Georeferencing, Experimental, INSPIRE-compliant, Population figure

## Abstract: Pendler Mobil: The use of mobile network data to support official commuter statistics

The availability of small-scale and real-time commuting patterns is of great importance for political as well as communal decision-making processes. From commuting behaviour, conclusions can be drawn about labour market regions and the distribution of the resident population, which contributes, among other things, to the ongoing improvement of the public transport infrastructure. The data required for this purpose is published by the official commuter statistics. However, it shows potential for improvement with respect to the temporal and spatial representation of commuting patterns of employed persons as well as a technical extension with respect to educational commuters.

This article describes the possibilities for extending the official commuter statistics based on origin-destination matrices from mobile network data, which were investigated by the project

*Pendler Mobil*. Due to their temporal timeliness and spatially fine resolution, mobile network data provide a robust data basis for the flexible mapping of potential and regular commuter movements. The potential performance of mobile network data thus enables an external validation of existing commuter calculations or commuter statistics, as well as a two-way complement to identify and represent other forms of commuting by the employed population.

Using the state of North Rhine-Westphalia as a case study, we discuss similarities and differences between corresponding commuting patterns based on mobile network data and the official commuter statistics. In this context, we address the challenges of processing and defining suitable mobile network data by the data provider as well as other influences, such as the distance covered or dwell times of mobile activities on them. Especially the underestimation of mobile commuter flows compared to the official commuter statistics suggests to discuss modification approaches of the mobile network data. As a result, the available mobile network data can potentially support the official commuter statistics by providing small-scale commuter flows in cities as an extended destination determination and enables the identification of highly frequented potential work locations in cities.

**Keywords**: Mobile network data, Commuter statistics, Origin-destination matrices, Movement, Mobility, Commuter

## Abstract: Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany

The ongoing growth of cities due to better job opportunities is leading to increased labour-related commuter flows in several countries. On the one hand, an increasing number of people commute and move to the cities, but on the other hand, the labour market indicates higher unemployment rates in urban areas than in the surrounding areas. We investigate this phenomenon on regional level by an alternative definition of unemployment rates in which commuting behaviour is integrated. We combine data from the Labour Force Survey with dynamic mobile network data by small area models for the federal state North Rhine-Westphalia in Germany. From a methodical perspective, we use a transformed Fay-Herriot model with bias correction for the estimation of unemployment rates and propose a parametric bootstrap for the mean squared error estimation that includes the bias correction. The performance of the proposed methodology is evaluated in a case study based on official data and in model-based simulations. The results in the application show that unemployment rates (adjusted by commuters) in German cities are lower than traditional official unemployment rates indicate.

**Keywords**: Bias correction, Fay-Herriot model, Mean squared error, Small area estimation, Unemployment rates

# Eidesstattliche Erklärung

## Erklärung gem. § 4 Abs. 2 Promotionsordnung zum Dr. rer. pol. des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin vom 13. Februar 2013

Hiermit erkläre ich, dass ich mich noch keinem Promotionsverfahren unterzogen oder um Zulassung zu einem solchen beworben habe, und die Dissertation in der gleichen oder einer anderen Fassung bzw. Überarbeitung einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule nicht bereits zur Überprüfung vorgelegen hat.

*Wiesbaden, 25. August 2022*

Sandra Hadam

## Erklärung gem. § 4 Abs. 2 der oben genannten Promotionsordnung

Hiermit erkläre ich, dass ich für die Dissertation folgende Hilfsmittel und Hilfen verwendet habe: Open Source Statistikprogramm R, Geoinformationssystem-Softwareprodukt ArcGIS, Schriftsatz Softwarepaket LaTeX, Office-Software Microsoft Excel und Microsoft Word, PDF-Programm Adobe Acrobat Pro DC, genannte Datenquellen/-sätze, genannte Koautoren und die im Literaturverzeichnis angegebene Literatur. Auf dieser Grundlage habe ich die Arbeit selbstständig verfasst.

*Wiesbaden, 25. August 2022*

Sandra Hadam