

SCHÄTZUNG VON BEVÖLKERUNGEN MIT HOHER RÄUMLICHER AUFLÖSUNG ANHAND VON MOBILFUNKNETZDATEN

Daniel Vester

Zusammenfassung: Die präzise Erfassung von Bevölkerungsdaten ist grundlegend für Entscheidungsprozesse in vielfältigen Anwendungsbereichen. Traditionelle Volkszählungen veralten schnell, weshalb alternative Schätzmethoden notwendig sind. Mobiltelefone bieten aufgrund ihrer weiten Verbreitung und genauen Standortbestimmung eine vielversprechende Grundlage dafür. Dieser Beitrag untersucht, ob geoinformatische Methoden zur Schätzung der statischen Bevölkerung eingesetzt werden können. Hierbei werden datenschutzkonforme Mobilfunknetzdaten mit CORINE-Landbedeckungs- und Gebäudedaten kombiniert, um ein Modell zu erstellen, welches die Bevölkerungsverteilung in München und umliegenden Kreisen auf einer Auflösung von 100 × 100 Metern schätzt. Die Ergebnisse zeigen, dass die räumliche Verteilung gut abgebildet wird, jedoch eine Tendenz zur Unterschätzung besteht, wobei die initiale Datenverteilung entscheidend für die Modellqualität ist.

Schlüsselwörter: Bevölkerungsverteilung, Mobilfunkdaten, CORINE-Landnutzung, Geoinformatik, Dasymetrische Modellierung

ESTIMATING POPULATION ON A HIGH SPATIAL RESOLUTION USING MOBILE NETWORK DATA

Abstract: The precise estimation of population data is essential for decision-making processes in a variety of application areas. Traditional censuses become outdated quickly, making alternative estimation methods necessary. Mobile phones, due to their widespread use and precise location data, provide a promising foundation for this. This study investigates whether geoinformatics methods can be used to estimate the static population. It combines privacy-compliant mobile network data with CORINE land cover and building data to create a model that estimates the population distribution in Munich and surrounding areas at a resolution of 100 × 100 meters. The results show that the spatial distribution is well captured. However, there is a tendency to underestimate, with the quality of the initial data distribution being critical for the model's accuracy.

Keywords: Static population density, mobile network data, CORINE land cover, geoinformatics, dasymetric modelling

1 EINLEITUNG

Die räumliche Verteilung der Bevölkerung festzustellen, wird schwieriger, je höher die Auflösung ist. Gleichzeitig erlaubt eine genauere Abschätzung unabhängig von administrativen Grenzen effektivere Planungen im Gesundheitswesen, Katastrophenschutz, Bildungsbereich oder in der Infrastruktur. Klassische Zensusmethoden sind allerdings aufwendig, teuer und schon nach kurzer Zeit

veraltet. Des Weiteren erfassen diese die amtliche Einwohnerzahl, also alle Personen jeglicher Nationalität, die in einem bestimmten Gebiet gemeldet sind (Statistisches Bundesamt 2023). Nicht oder falsch gemeldete Personen werden nicht beachtet. Durch die Verbreitung mobiler Technologien ermöglichen Mobilfunkdaten eine präzisere und aktuellere Schätzung der Bevölkerung. Da mobile Geräte inzwischen von einem Großteil der Bevölkerung täglich genutzt werden, liefern sie wertvolle Einblicke in Routinen, Interaktionen und Mobilitätsmuster und können als Datenpunkte in einem dezentralen globalen Sensornetz gesehen werden.

Ein praktisches Anwendungsbeispiel für die Nützlichkeit dieser Daten war die Analyse der Mobilitätsreduktion während der Covid-19-Pandemie in Deutschland. Das Statistische Bundesamt konnte damit die Effektivität von Lock-down-Maßnahmen bewerten (Bohnensteffen et al. 2021). Darüber hinaus könnten Mobilfunkbetreiber präzisere und aktuellere Bevölkerungsdaten nutzen, um etwa die Auswirkungen von Netzausbau oder Netzänderungen

Autor

M. Sc. Daniel Vester
Lauterlech 40
D-86152 Augsburg
E: daniel@vesteronline.de

besser abzuschätzen oder Potenziale für eine gezieltere Vermarktung zu identifizieren.

Die Methodik, mit der die Bevölkerungsverteilung bestmöglich geschätzt werden kann, ist ein aktives Feld der Forschung. Offene Fragen sind neben der effektivsten Methodik die mögliche räumliche Auflösung sowie die Frage, ob Modelle generisch sind oder auf spezifische Gebiete trainiert werden müssen.

Vor diesem Hintergrund konzentriert sich dieser Beitrag auf die zentrale Forschungsfrage: Kann die Bevölkerungsverteilung mithilfe von Mobilfunkdaten mit Methoden aus der Geoinformatik geschätzt werden?

Weitere Fragen sind, wie gut bestehende Methoden auf höhere räumliche Auflösungen und ein größeres Untersuchungsgebiet mit einem höheren Anteil ländlicher Gebiete angewendet werden können.

Die hier vorgestellte Fallstudie wendet eine Methodik nach Khodabandelou et al. (2019) auf Funkzellen-Performance-Rohdaten ohne Personenbezug aus dem deutschen LTE-Mobilfunknetz der Telefónica Germany GmbH & OHG an und untersucht die nächtliche Wohnbevölkerung in der Stadt München und den umliegenden, teils ländlichen Kreisen auf einem Raster mit einer hohen Auflösung von 100 m × 100 m. Der Zugang zu deutschen Mobilfunkdaten dieser Art und Granularität ist in der Forschung unüblich und eröffnet neue Möglichkeiten für Analysen, die unabhängig von administrativen Grenzen und grober räumlicher Aggregation auskommen. Im Rahmen der Studie wird ein univariates Modell zur Schätzung der Nachtbevölkerung erstellt und verglichen, welche Mobilfunkdaten am besten geeignet sind.

Abschließend diskutiert der Beitrag die Ergebnisse und Limitationen sowie die Anwendbarkeit auf größere und ländlichere Gebiete, um zukünftige Forschung zu unterstützen und praxisnahe Lösungen zu fördern.

Dieser Beitrag ist eine gekürzte Version der Masterarbeit „Estimating dynamic population using cellular network data“ des Autors. Neben der hier vorgestellten Fallstudie enthält die Arbeit eine weitere Methodik zur zeitabhängigen Schätzung der Tagbevölkerung (Vester 2023).

Der Autor dieses Beitrags ist Angestellter der Telefónica Germany GmbH & Co. OHG, welche die für diese Studie genutzten Mobilfunk- sowie Gebäudedaten bereitstellte. Telefónica hatte jedoch keinen Einfluss auf die Analyse, Interpretation oder Veröffentlichung der Ergebnisse.

2 STAND DER FORSCHUNG

Die Nutzung von Mobilfunkdaten zur Bevölkerungsschätzung ist ein aktives Forschungsfeld. Deville et al. (2014) zeigten, dass sich Mobilfunkdaten effizient für diesen Zweck einsetzen lassen. Sie verteilten die Daten räumlich mithilfe von Voronoi-Polygonen und verwendeten einen RANSAC-Regressor zur Modellierung. Ihre Ergebnisse belegen, dass Mobilfunkdaten eine präzise und effiziente Schätzung der Bevölkerung ermöglichen.

Einen großen Sprung in diesem Forschungsfeld gab es, als die italienische Firma TIM S.p.A 2012 im Rahmen einer Big-Data-Challenge einen großen Datensatz an Mobilfunkdaten für die Städte Mailand und Trient veröffentlichte. Douglass et al. (2015) nutzten diesen Datensatz, um die Bevölkerung mit einer Auflösung von 235 m × 235 m zu schätzen. Zur Schätzung der Nachtbe-

völkerung wurde wieder die lineare Regression genutzt, für die dynamische Bevölkerung ein Random-Forest-Verfahren. Sie kamen zu dem Schluss, dass die Genauigkeit höher ist, wenn die Bevölkerung innerhalb einer räumlichen Einheit über einem bestimmten Schwellenwert liegt.

Khodabandelou et al. (2019) nutzten denselben Datensatz von TIM S.p.A und den gleichen RANSAC-Regressor für die Nachtbevölkerung wie Deville et al. (2014). Die räumliche Aufteilung war ein irreguläres Gitternetz mit Zellgrößen zwischen 255 m × 325 m und 2 km × 2,5 km. Sie teilten das Untersuchungsgebiet in unterschiedliche Landnutzungsklassen auf, die mithilfe der Median-Week-Signature-Methode von Furno et al. (2015) generiert wurden. Pro Landnutzungsklasse wurde ein eigenes Regressionsmodell trainiert. Für die dynamische Bevölkerung verfolgten sie einen neuen Ansatz, bei dem die durchschnittliche Aktivität pro User zum Trainieren eines multivariaten Regressionsmodells genutzt werden. Ihr Modell zeigte eine hohe Genauigkeit bei der Vorhersage typischer Tagestrends und erwies sich auch bei Großveranstaltungen wie Fußballspielen als zuverlässig. Dieser Artikel bildet die Grundlage für die in diesem Beitrag verwendete Methodologie.

Ein wichtiger Aspekt der Bevölkerungsschätzung ist die initiale räumliche Verteilung der Mobilfunkdaten auf Gebietseinheiten. Tennekes & Gootzen (2022) verglichen verschiedene bayesische Methoden für diese Verteilung und stellten fest, dass eine Kombination mehrerer Methoden zusammen mit einem signalstärkebasierten Ausbreitungsmodell die besten Ergebnisse liefert. Aasa et al. (2021) nutzten anonymisierte Call-Data-Records, um fünf verschiedene Methoden zur räumlichen Verteilung zu testen: Punkt-in-Polygon, flächengewichtete Interpolation, gebäudegewichtete Interpolation, adaptive Morton-Gitter-Interpolation und eine Random-Forest-Regression. Die Analyse zeigte, dass die ersten beiden Methoden ungeeignet sind, während die übrigen vergleichbar gute Ergebnisse liefern. Dieses Ergebnis wurde von Ricciato & Coluccia (2023) bestätigt. Die gebäudegewichtete Interpolation wird in diesem Beitrag zusammen mit Ausbreitungsdaten für die initiale Verteilung der Daten genutzt. Ein neuerer Ansatz in der räumlichen Verteilung und Vorhersage der Bevölkerung sind Deep-Learning-Modelle. Zhu et al. (2020) demonstrieren, dass diese Modelle besser zur räumlichen Interpolation von Mobilfunkdaten geeignet sind als traditionelle Verfahren wie inverse Distanzgewichtung. Dazu zeigen Cecaş et al. (2020), dass Deep Learning für die Vorhersage der Bevölkerung auf Basis von Mobilfunkdaten mit zellbasierten Zeitreihen vergleichbare Ergebnisse wie statistische Ansätze liefert. Beide Ansätze verdeutlichen das Potenzial von Deep Learning zur Optimierung der räumlichen Verteilung mobiler Daten, was durch Zong et al. (2019) unterstützt wird.

3 FALLSTUDIE – BEVÖLKERUNG IN MÜNCHEN

Die Fallstudie untersucht das Potenzial von Mobilfunkdaten zur Vorhersage der Bevölkerung in München und den umliegenden Landkreisen. Dabei wird auf der Methode von Khodabandelou et al. (2019) aufgebaut, jedoch mit wesentlichen Abweichungen: Wie Schöpflin et al. (2024) feststellen, sind personenbezogene Mobilfunkdaten aus Datenschutzgründen oft nur mit beschränkten raumzeitlichen Auflösungen verfügbar. Daher werden in diesem Beitrag statt Call-Data-Records (CDR) aggregierte Mobilfunk-Nut-

zungsdaten auf Zellenebene verwendet, die keinerlei Personenbezug, auch nicht in anonymisierter Form, aufweisen. Dies ermöglicht ihre datenschutzkonforme Nutzung.

Zur Bestimmung der Netzabdeckung werden keine approximativen Voronoi-Polygone verwendet. Stattdessen kommen Bestserver-Daten zum Einsatz, die auf einem signalstärkebasierten Ausbreitungsmodell beruhen. Diese Daten geben für jede Rasterzelle an, welche Mobilfunkzelle pro Frequenzband die höchste Signalstärke liefert. Ferner erfolgt die räumliche Aufteilung der Untersuchungsregion nicht über eine vorherige dynamische Clusterbildung auf Basis von Wochentags-Mobilfunknutzungsmustern, sondern mittels modifizierter CORINE-Landbedeckungsdaten. Außerdem wird die Bevölkerung nicht für administrative Aufteilungen geschätzt, sondern auf Basis eines 100 m × 100 m INSPIRE-konformen Rasters. Des Weiteren umfasst die Fallstudie nicht nur ein Stadtgebiet, sondern auch umliegende ländliche Gebiete.

3.1 VERWENDETE DATEN UND UNTERSUCHUNGSGEBIET

Für die Erstellung eines präzisen Modells werden in diesem Beitrag mehrere Datensätze aus unterschiedlichen Quellen und Jahren herangezogen. Die Mobilfunk-, Bestserver- und Gebäude-Daten wurden von der Telefónica Germany GmbH & Co. OHG bereitgestellt und sind nicht öffentlich zugänglich. Die Volkszählungsdaten von 2011 dienen dazu, das Bevölkerungsmodell zu kalibrieren, während die übrigen Volkszählungsdaten zur Validierung der Modellergebnisse verwendet werden. Die verwendeten Datensätze umfassen:

- ▶ Vier verschiedene LTE-Mobilfunkmetriken auf Funkzellenebene, stündlich erfasst im Zeitraum von April bis Juni 2023;
- ▶ Bestserver-Daten für jedes Frequenzband von April bis Juni 2023;
- ▶ Gebäude-Polygon-Daten, Stand 01.01.2023 (Level of Detail 1);
- ▶ Volkszählungsdaten zum 09.05.2011 in einem Raster von 100 m × 100 m (Statistisches Bundesamt 2014);
- ▶ Bayerische Volkszählungsdaten zum 31.12.2021 auf Kreisebene (Bayerisches Landesamt für Statistik 2021);
- ▶ Münchner Volkszählungsdaten auf Stadtteileebene zum 31.12.2022 (Statistisches Amt München 2022);
- ▶ CORINE-Landnutzungsdaten für 2017 bis 2018 (European Union's Copernicus Land Monitoring Service Information 2020);
- ▶ Kreisgrenzen, Stand 01.01.2023 (BKG, 2023).

Bei den Mobilfunkmetriken handelt es sich um Kennzahlen aller aktiven LTE-Netzzellen, die im Untersuchungsgebiet liegen oder in es hinein strahlen. Die Rohdaten sind in die sechs verschiedenen Frequenzbänder unterteilt, die vom Telefónica-Netz für LTE verwendet werden. Die vier Metriken umfassen „Aktive Nutzer“, „Anruferdauer“, „Anruferanzahl“ und „Datennutzung“. Für diese Daten wurden Werte aus vier Wochen ohne Schulferien und Feiertage

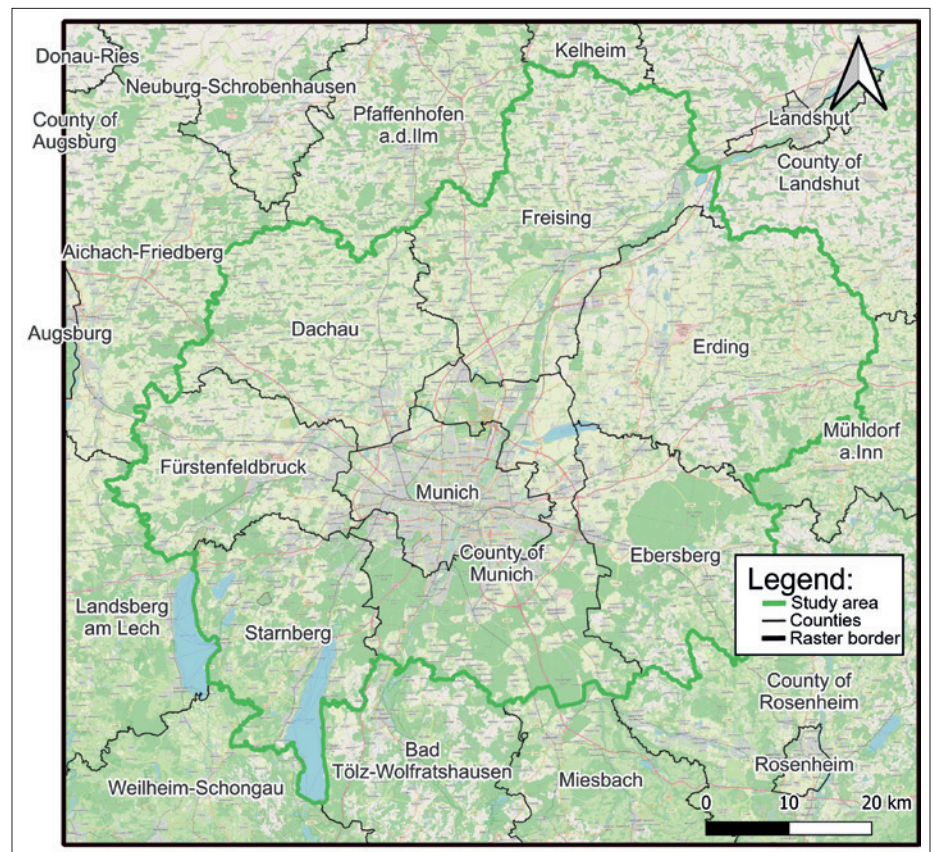


Abbildung 1: Karte des Studiengebiets

in den Monaten April bis Juni 2023 genutzt, um eine Durchschnittswoche zu bilden.

Die amtlichen Volkszählungsdaten aus dem Jahr 2011 wurden gewählt, da die Zensusdaten aus dem Jahr 2021 zum Zeitpunkt der Studie noch nicht vorlagen.

Die Landnutzungsclassen der CORINE-Daten werden zu acht relevanten Klassen zusammengefasst. Diese sind: Durchgängig städtisch, diskontinuierlich städtisch, Wald-/Wasserflächen, Landwirtschaft, Infrastruktur, urbane Grünflächen, Industrie/Gewerbe und karges Land.

Alle Daten wurden für das gesamte Untersuchungsgebiet (siehe Abbildung 1) erfasst. Es umfasst die bayerischen Kreise München (Stadt), Landkreis München, Fürstenfeldbruck, Dachau, Freising, Erding, Ebersberg und Starnberg. Abbildung 1 zeigt das Gebiet grün umrandet. Im Jahr 2021 hatte das Untersuchungsgebiet mit einer Fläche von 4696 km² eine Gesamtbevölkerung von etwa drei Millionen Einwohnern. Der Großteil der Bevölkerung, ca. 1,5 Millionen Menschen, lebt in der Stadt München.

3.2 METHODIK

Die in diesem Beitrag verwendete Methodik gliedert sich in zwei Hauptteile (siehe Abbildung 2), wobei der zweite Teil von Khodabandelou et al. (2019) übernommen wurde. Der erste Schritt, das Vorbereiten mittels Gebäudegewichtung, basiert auf Aasa et al. (2021).

Die grundlegende Annahme hinter der Nutzung einer Gebäudegewichtung ist, dass ein Großteil der Bevölkerung in Gebäuden wohnt und sie damit ein zuverlässiger Indikator für die Bevölkerungsverteilung sind (Aasa et al. 2021). Das Gewichtungsraster gibt für jede 100 m × 100 m große Zelle an, wie hoch der Anteil

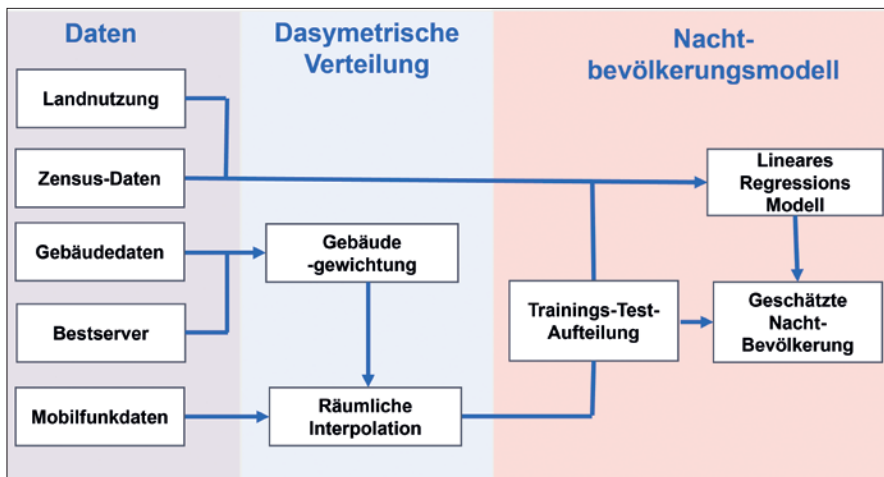


Abbildung 2: Übersicht über die Methodik zur Schätzung der Nachtbevölkerung

der Gebäudegrundfläche dieser Rasterzelle an der Gesamtgebäudegrundfläche der jeweiligen Bestserver-Fläche ist. Einer Rasterzelle mit 10 % der Gebäudegrundfläche wird auch 10 % der gesammelten Metriken zugewiesen. Dieser Gewichtungssatz wird zusammen mit den Performance-Daten jeder Frequenz genutzt, um für jede Stunde des Untersuchungszeitraums ein Raster mit den über alle Frequenzen aggregierten Metriken zu erzeugen.

Daraufhin wurde untersucht, welche der vier Metriken und welche Uhrzeit die höchste Korrelation mit der Nachtbevölkerung hat. Die Ergebnisse dieses Vergleichs sind in Tabelle 1 dargestellt. Da die Anrufdauer am Dienstag um 22 Uhr die höchste Korrelation hat, wurde dieser Datensatz zur Berechnung des Bevölkerungsmodells genutzt.

Im zweiten Schritt der Methodik wird für jede der acht Landnutzungsklassen ein separates lineares Regressionsmodell mit dem RANSAC-Regressor erstellt. Der ursprüngliche Ansatz von Khodabandelou et al. (2019), die Daten dynamisch in Cluster mit unterschiedlicher Mobilfunknutzung zu klassifizieren, wurde nicht übernommen, sondern stattdessen wurde auf CORINE-Landnutzungsdaten zurückgegriffen.

In Anlehnung an Bettencourt (2013), Deville et al. (2014) und Douglass et al. (2015) wird die Annahme getroffen, dass sich der Zusammenhang zwischen Bevölkerung p_i und Mobilfunkaktivi-

tät σ_i in Rasterzelle i durch $p_i = \alpha \sigma_i^\beta$ beschreiben lässt. Hierbei sind α und β die Modellparameter. Durch Log-Transformation $\log(p_i) = \log(\alpha) + \beta \log(\sigma_i)$ wird die Beziehung linear und kann mittels der gewöhnlichen Methode der kleinsten Quadrate geschätzt werden.

Zur Parameterschätzung werden nur Rasterzellen i gewählt, in denen die Zensusbevölkerung größer als null ist (Zensuswerte kleiner 2 sind aus Datenschutzgründen ebenfalls auf null gesetzt). Dieser Schritt vereinfacht die Log-Transformation, könnte aber die Modellqualität in dünn besiedelten Gebieten verringern. Der verwendete Regressor erkennt Ausreißer und schließt sie aus. Ergebnis der Modellanpassung sind acht $\hat{\alpha}$ - und $\hat{\beta}$ -Werte, mit

denen für jede Landnutzungsklasse und Rasterzelle die geschätzte Bevölkerung \hat{p}_i wie folgt berechnet werden kann: $\hat{p}_i = \hat{\alpha} \sigma_i^{\hat{\beta}}$.

Zur Abschätzung der Modellgüte und zum Vergleich, welche der vier Mobilfunkmetriken zu welcher Uhrzeit die besten Ergebnisse liefern, wurde für jede Metrik-Zeit-Kombination ein eigenes Modell trainiert. Anschließend erfolgte eine Leave-One-Out-Kreuzvalidierung mit zehn Durchläufen, anhand derer NRMSE und R^2 berechnet wurden. Zudem kam eine gewichtete Zufallsstichprobe zum Einsatz, um sicherzustellen, dass alle Landnutzungsklassen in jedem Modell vertreten sind.

4 ERGEBNISSE DER FALLSTUDIE

Dieser Abschnitt präsentiert die Ergebnisse des in Abschnitt 3 beschriebenen Modells. Zunächst werden Beispielkarten gezeigt, um die räumliche Verteilung der Ergebnisse aufzuzeigen. Da es zum Zeitpunkt der Fallstudie noch keine neueren Zensusdaten mit einer räumlichen Auflösung von 100 m x 100 m gab, ist die genaue Auswertung der räumlichen Verteilung nicht möglich.

Zur Überprüfung der Übertragbarkeit wird ein Modell, das ausschließlich mit Daten der Stadt München trainiert wurde, dem Modell, das auf dem gesamten Untersuchungsgebiet basiert, gegenübergestellt. Zuletzt wird untersucht, inwiefern der Wochentag und die Uhrzeit der zugrunde liegenden Daten das Modellergebnis beeinflusst.

4.1 BEISPIELERGEBNISSE DER NACHTBEVÖLKERUNGSSCHÄTZUNG

Als Beispiel für die räumliche Verteilung der Ergebnisse zeigt Abbildung 3 zwei Ausschnitte. Links wird die prognostizierte Bevölkerungsverteilung in fünf Bezirken im Münchner Stadtzentrum dargestellt. Die farbliche Darstellung reicht von Weiß, Grau und Hellblau (niedrige Werte) bis hin zu Dunkelrot und Braun (hohe Werte), wobei zahlreiche 100 m x 100 m große Zellen Werte von 250 oder mehr aufweisen, deutlich über dem Münchner Durchschnitt von 48 pro Zelle. Gleichzeitig werden unbebaute Flächen, wie die Theresienwiese und der Englische Garten, durch die Gebäudegewichtung als bevölkerungsfrei dargestellt. Bebaute Gebiete ohne Wohnbevölkerung werden allerdings mit einer geringen Bevölkerung von bis zu 10 angezeigt.

Metrik	Wochentag	Uhrzeit	Korrelationskoeffizient
Aktive Nutzer	Montag	23 Uhr	0,622
Datennutzung	Montag	23 Uhr	0,577
Anrufdauer	Dienstag	22 Uhr	0,655
Anrufanzahl	Montag	21 Uhr	0,621

Tabelle 1: Wochentage und Uhrzeiten mit den höchsten Korrelationskoeffizienten nach Bravais-Pearson für jede der vier Metriken mit den Zensusdaten von 2011

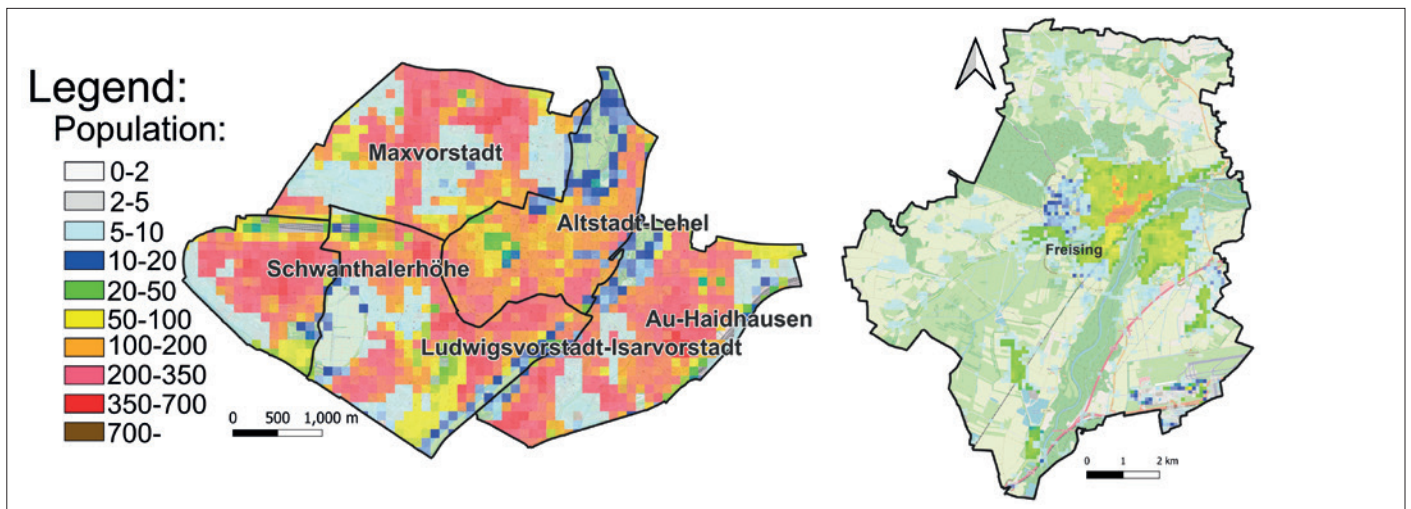


Abbildung 3: Geschätzte Nachtbevölkerung: links in der Münchner Innenstadt, rechts in der Gemeinde Freising

Die rechte Karte zeigt Freising und Umgebung einschließlich eines kleinen Teils des Münchner Flughafens in der südöstlichen Ecke. Das Stadtzentrum weist eine mittlere Dichte (orange) auf, während abgelegene Zellen meist Werte unter zehn haben. Unbebaute Bereiche erscheinen ohne prognostizierte Population und lediglich die Orte Pulling und Achering im nördlichen Bereich zeigen jeweils zwei kleine Zentren mit moderater Dichte. Allerdings wird für die Gebäude auf dem Flughafengelände ebenfalls eine Nachtbevölkerung bis zu 50 geschätzt. Dies ist eine ähnliche Problematik wie die Industriegebiete in der Münchner Innenstadt.

Dies könnte mithilfe einer Anreicherung der Gebäudedaten mit Informationen zur Gebäudefunktion verhindert werden, wie von Järv et al. (2017) vorgeschlagen. Zusammengefasst zeigt das Modell eine grundsätzlich plausible räumliche Verteilung der Nachtbevölkerung, weist jedoch in Bereichen ohne tatsächliche Bevölkerung, aber mit vorhandenen Gebäuden fälschlicherweise eine geringe Bevölkerungsdichte auf.

4.2 VERGLEICH VON TRAININGSDATEN MIT UNTERSCHIEDLICHER RÄUMLICHER AUSDEHNUNG

Um den Einfluss der Auswahl der Trainingsdaten auf die Modellgüte zu testen, wurde ein Modell ausschließlich mit Daten aus dem Stadtgebiet München trainiert und ein weiteres mit Daten aus dem gesamten Stadtgebiet.

Areal	Trainingsdaten	NRMSE	R ²
Stadtteile München	Stadt München	0,17	0,78
	Gesamtes Gebiet	0,5	-0,9
Kreise	Stadt München	0,157	0,82
	Gesamtes Gebiet	0,155	0,82

Tabelle 2: NRMSE- und R²-Werte für ein Modell trainiert mit Daten aus dem gesamten Studiengebiet und einem Modell trainiert mit Daten der Stadt München. Kennzahlen werden einmal auf Stadtteilebene von München und auf Landkreisebene verglichen.

Zur Bestimmung der Güte der beiden Nachtbevölkerungsmodelle wurden mithilfe einer leave-One-Out-Kreuzvalidierung der NRMSE und der R² berechnet. Tabelle 2 zeigt die Ergebnisse. Die Ergebnisse für beide Modelle wurden einmal mit Bevölkerungsdaten der Stadtbezirke Münchens von 2022 verglichen, und einmal mit Bevölkerungsdaten der restlichen Kreise aus dem Jahr 2021.

Für den NRMSE zeigt das auf München trainierte Modell durchweg niedrige Werte, was auf einen guten Fit hindeutet. Das Modell, das für das gesamte Studiengebiet trainiert wurde, erzielt hingegen nur für die Landkreise niedrige NRMSE-Werte, während für München mittlere Werte vorliegen. Die R²-Werte bestätigen, dass das Münchner Modell mit 0,78 bzw. 0,823 einen hohen Anteil der Variabilität erklärt, während das Studiengebietmodell für München einen ungewöhnlich negativen R² von -0,9 und für die Landkreise einen hohen Wert von 0,827 erreicht. Der negative Wert bedeutet, dass das Modell schlechter zu Bevölkerungsvorhersage geeignet ist als der durchschnittliche Bevölkerungswert. Daraus folgt, dass das auf München trainierte Modell deutlich besser für die Schätzung der Bevölkerung in München geeignet ist und ähnlich gut für die Schätzung in den anderen Kreisen. Das Studiengebietmodell sollte wegen des niedrigen R² nicht verwendet werden. Etwas anders sieht es aus, wenn die absoluten Zahlen verglichen werden. In Tabelle 3 werden die aggregierten Ergebnisse der beiden Modelle mit den Zensusdaten von 2021 auf Landkreisebene verglichen.

Der Vergleich zeigt, dass für die Kreise Fürstenfeldbruck, Ebersberg, Dachau, Erding, Starnberg und Stadt München das auf München trainierte Modell eine geringere Abweichung hat. Für den Landkreis München und Freising schneidet hingegen das auf das gesamte Studiengebiet trainierte Modell besser ab. Der sehr niedrige R² aus Tabelle 2 spiegelt sich auch hier wider, da das auf das gesamte Gebiet trainierte Modell die Bevölkerung in der Stadt München um 61 % unterschätzt.

Laut Ahas et al. (2015) und Douglass et al. (2015) unterscheidet sich die Aktivität sowohl zwischen Städten als auch innerhalb derselben, wobei manche Städte frühere Aktivitäten und ausgeprägtere Saisonalitäten aufweisen, was auch kulturelle Unterschiede widerspiegeln kann. Zudem ist die Definition des Stadtgebiets relevant,

Kreis	Trainingsdaten Stadt München	Trainingsdaten gesamtes Gebiet	Zensus 2021 Bevölkerung
Fürstenfeldbruck	+5145	-26721	218579
LK München	+76116	-64861	349837
Ebersberg	-6846	-14971	144562
Dachau	-10765	-16738	155449
Freising	-30847	-2795	181144
Erding	+1845	+9793	139622
Starnberg	+2996	+22909	136747
Stadt München	+71525	-575276	1487708
Gesamt	+165727	-668660	2813648

Tabelle 3: Auf Kreis aggregierte Abweichung der Ergebnisse der Bevölkerungsschätzung zu den Zensusdaten. Einmal mit einem Modell trainiert mit Daten der Stadt München, einmal mit einem Modell trainiert mit Daten des gesamten Studiengebiets.

da die Einbeziehung ausschließlich innerstädtischer Hochhausbezirke zu anderen Ergebnissen führt als die Berücksichtigung auch von Vororten. Die Ergebnisse des Vergleichs in diesem Beitrag bestärken die Aussage, dass das Modell zwar die Bevölkerung schätzen kann, aber die richtige Datenbasis dafür benötigt.

4.3 VERGLEICH VERSCHIEDENER METRIKEN UND UHRZEITEN

Um herauszufinden, welche der vier Mobilfunkmetriken zu welcher Uhrzeit die besten Modellergebnisse erzeugt, wurde für jede Kombination an Metrik und Uhrzeit einer Woche ein Modell trainiert und dann wieder mithilfe einer Leave-One-Out-Kreuzvalidierung mit zehn Durchläufen der NRMSE und R^2 berechnet. Die Ergebnisse für den R^2 sind in Abbildung 4 zu sehen. Der Vergleich zeigt, dass die Modellergebnisse für Anruferanzahl und Anruferdauer eine höhere Varianz haben, aber auch die höchsten Korrelationen erzielen. Außerdem ist das Tagesmuster stärker ausgeprägt mit

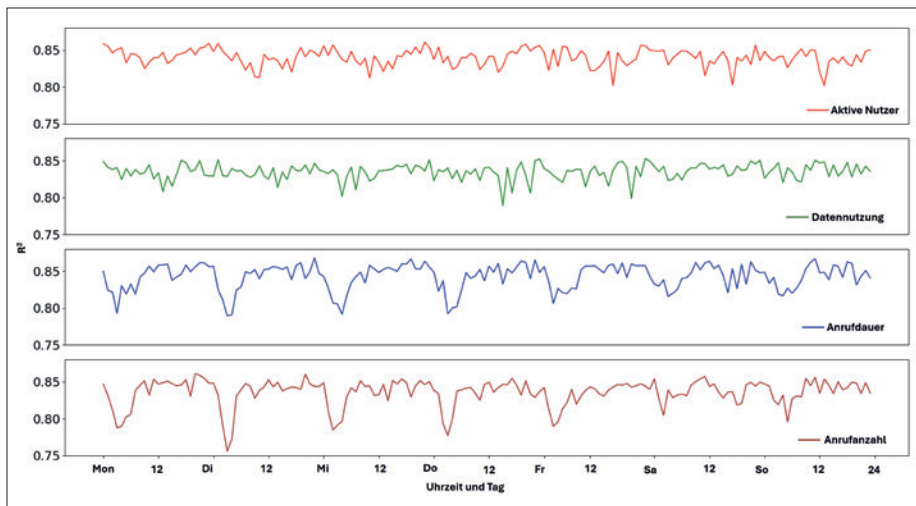


Abbildung 4: R^2 -Werte für Modelle, trainiert mit jeder der vier Metriken über den Verlauf einer Woche

niedrigen Werten zwischen 1 Uhr und 4 Uhr morgens und dem Maximum am Abend zwischen 20 Uhr und 23 Uhr. Die Varianz ist am Wochenende geringer als an den Wochentagen. Der höchste Wert wird, wie ursprünglich vermutet, am Dienstag um 22 Uhr für die Anruferdauer erreicht. Andere Studien nutzen unterschiedliche Kombinationen von Metriken und Zeiträumen, was auf einen fehlenden Konsens in der Literatur hinweist. Beispielsweise verwenden Khodabandelou et al. (2019) Daten zwischen 4 Uhr und 5 Uhr morgens, Bohnensteffen et al. (2021) Daten zwischen 20 Uhr und 23 Uhr und Deville et al. (2014) nutzen die Nachtstunden zwischen 20 Uhr abends und 7 Uhr morgens. Bei der Auswahl der Metrik ist außerdem die unterschiedliche Mobilfunknutzung nach Ort zu beachten. Eine Erklärung für die niedrigeren R^2 -Werte für die Datennutzung kann sein, dass Menschen zuhause WLAN für den Internetzugang nutzen, aber immer noch normale Anrufe tätigen. Langfristig wird die Wahl geeigneter Metriken durch den technologischen Wandel beeinflusst, wie etwa durch das Wachstum von VoIP und den

Rückgang klassischer SMS. Eine Kombination verschiedener Datenquellen, Metriken und Uhrzeiten, wie etwa die Berechnung einer mittleren Gesprächsdauer während eines längeren Zeitraums, könnte eine vielversprechende Alternative darstellen.

5 FAZIT

Die vorliegende Studie zeigt, dass geoinformatische Methoden in Kombination mit Mobilfunknetzdaten zur Schätzung der Bevölkerungsverteilung vielversprechende Ergebnisse liefern. Mittels linearer Regressionsmodelle, die auf Mobilfunk-Nutzungsdaten basieren und durch Gebäudeinformationen gewichtet sowie nach Landnutzungsklassen differenziert wurden, konnte eine Bevölkerungsabschätzung mit hoher räumlicher Auflösung realisiert werden. Dabei wurde untersucht, welche Mobilfunkmetriken (aktive Nutzer, Anruferdauer, Anruferanzahl und Datennutzung) und welche Zeiträume den besten Bezug zu aktuellen Zensusdaten aufweisen. Die Validierung erfolgte aggregiert auf Landkreis- und Stadteilebene.

Ein zentrales Ergebnis ist, dass trotz der hohen räumlichen Auflösung, die durch den Mangel an detaillierten Ground-Truth-Daten die Verifikation erschwert, die Methode grundsätzlich in der Lage ist, die räumliche Bevölkerungsverteilung abzubilden. Allerdings zeigte sich, dass das Modell tendenziell die tatsächliche Bevölkerungszahl unterschätzt. Besonders interessant ist die Erkenntnis, dass ein ausschließlich auf Daten der Stadt München trainiertes Modell präzisere Ergebnisse liefert, was darauf schließen lässt, dass Modelle regionenspezifisch sind, aber eine gewisse Übertragbarkeit haben. Die Verwendung einer statischen Trennung des Untersu-

chungsgebiets in Landnutzungsklassen und der Einsatz von Rohdaten erweisen sich dabei als praktikabel.

Zudem weisen Schöpflin et al. (2024) darauf hin, dass Mobilfunkdaten aus Datenschutzgründen oft nur mit einer beschränkten raumzeitlichen Auflösung zur Verfügung stehen. Die Integration zusätzlicher räumlicher Informationen, etwa aus Gebäude- oder Landnutzungsdaten, kann diese Limitationen jedoch teilweise ausgleichen und somit präzisere Analysen ermöglichen.

Künftige Arbeiten könnten sich unter anderem der Frage widmen, welche Mobilfunknummern sich am besten für Bevölkerungsschätzungen eignen, vor allem im Hinblick auf sich wandelnde

Nutzungsmuster. Ebenfalls vielversprechend ist die Weiterentwicklung der initialen dasymetrischen Verteilung, bei der untersucht werden sollte, welche zusätzlichen Daten die Qualität der Verteilung am stärksten verbessern.

Zusammenfassend zeigt der Beitrag, dass Mobilfunkdaten grundsätzlich als Basis für bevölkerungsbezogene Analysen genutzt werden können. Die Ergebnisse deuten darauf hin, dass weitere Untersuchungen zur Optimierung der Modellgenauigkeit und zur praktischen Anwendung, etwa in der Stadtplanung oder im Katastrophenschutz, sinnvoll sind.

Literatur

- Aasa, A.; Kamenjuk, P.; Saluveer, E.; Šimbera, J.; Raun, J. (2021): Spatial interpolation of mobile positioning data for population statistics. In: *Journal of Location Based Services* 15 (4), S. 239–260.
- Ahas, R.; Aasa, A.; Yuan, Y.; Raubal, M.; Smoreda, Z.; Liu, Y.; Ziemlicki, C.; Tiru, M.; Zook, M. (2015): Everyday space-time geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn. In: *International Journal of Geographical Information Science* 29 (11), S. 2017–2039.
- Bayerisches Landesamt für Statistik (2021): Fortschreibung des Bevölkerungsstandes. https://www.statistik.bayern.de/statistik/gebiet_bevoelkerung/bevoelkerungsstand/index.html#link_2 (6/2023).
- Bettencourt, L. M. A. (2013): The origins of scaling in cities. In: *Science* 340 (6139), S. 1438–1441.
- BKG – Bundesamt für Kartographie und Geodäsie (2023): Verwaltungsgebiete 1:5 000 000. <https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete/verwaltungsgebiete-1-5-000-000-stand-01-01-vg5000-01-01.html> (6/2023).
- Bohnensteffen, S.; Mülhan, J.; Saidani, Y. (2021): Mobilität während der Corona-Pandemie. In: *WISTA – Wirtschaft und Statistik* 3/2021, S. 89–105.
- Cecaj, A.; Lippi, M.; Mamei, M.; Zambonelli, F. (2020): Comparing deep learning and statistical methods in forecasting crowd distribution from aggregated mobile phone data. In: *Applied Sciences* 10 (18), 6580.
- Deville, P.; Linard, C.; Martin, S.; Gilbert, M.; Stevens, F. R.; Gaughan, A. E.; Blondel, V. D.; Tatem, A. J. (2014): Dynamic population mapping using mobile phone data. In: *Proceedings of the National Academy of Sciences of the United States of America* 111 (45), S. 15888–15893.
- Douglass, R. W.; Meyer, D. A.; Ram, M.; Rideout, D.; Song, D. (2015): High resolution population estimates from telecommunications data. In: *EPJ Data Science* 4 (1).
- European Union's Copernicus Land Monitoring Service information (2020): CORINE Land Cover 2020 (vector/raster 100 m), Europe, 6-yearly. <https://doi.org/10.2909/960998c1-1870-4e82-8051-6485205ebbac> (5/2023).
- Furno, A.; Fiore, M.; Stanica, R.; Ziemlicki, C.; Smoreda, Z. (2017): A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. In: *IEEE Transactions on Mobile Computing* 16 (10), S. 2682–2696.
- Järvi, O.; Tenkanen, H.; Toivonen, T. (2017): Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. In: *International Journal of Geographical Information Science* 31(8), S. 1630–1651.
- Khodabandelou, G.; Gauthier, V.; Fiore, M.; El-Yacoubi, M. A. (2019): Estimation of static and dynamic urban populations with mobile network metadata. In: *IEEE Transactions on Mobile Computing* 18(9), S. 2034–2047.
- Ricciato, F.; Coluccia, A. (2023): On the estimation of spatial density from mobile network operator data. In: *IEEE Transactions on Mobile Computing* 22 (6), S. 3541–3557.
- Schöpflin, F.; Farthofer-Oster, S.; Carmona, J.; Mittlböck, M. (2024): Added value and challenges of 5G technology in the field of geoinformatics – Presentation of results of the two research projects 5G-MOBIS and 5G-EXPS. In: *gis.Science* 1/2024, S. 27–35.
- Statistisches Amt München (2022): Indikatorenatlas München. <https://www.mstatistik-muenchen.de/indikatorenatlas/atlas.html?indicator=i51> (6/2023).
- Statistisches Bundesamt (Destatis) (2014): Ergebnisse des Zensus 2011 zum Download – erweitert. <https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html> (5/2023).
- Statistisches Bundesamt (Destatis) (2023): Bevölkerungsstand. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/_inhalt.html (5/2023).
- Tennekes, M.; Gootzen, Y. A. P. M. (2022): Bayesian location estimation of mobile devices using a signal strength model. In: *Journal of Spatial Information Science* 25, S. 29–66.
- Vester, D. (2023): Estimating dynamic population using cellular network data. https://github.com/daniel-vester/Estimating-dynamic-population-using-cellular-network-data/blob/main/Masterarbeit_Daniel_Vester_final_compressed.pdf (2/2025).
- Zhu, D.; Cheng, X.; Zhang, F.; Yao, X.; Gao, Y.; Liu, Y. (2020): Spatial interpolation using conditional generative adversarial neural networks. In: *International Journal of Geographical Information Science* 34 (4), S. 735–758.
- Zong, Z.; Feng, J.; Liu, K.; Shi, H.; Li, Y. (2019): Deepdpm: Dynamic population mapping via deep neural network. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (1), S. 1294–1301.