Kleinräumige Bevölkerungsschätzung mit Hilfe der "EEA Fast Track Service Precursor on Land Monitoring" und "CORINE Land Cover" Datensätze

Bachelorarbeit

Zur Erlangung des akademischen Grades des
Bachelor of Science
im Studiengang Geoinformatik

Vorgelegt im Prüfungsamt für Geowissenschaften
der Universität Münster
Erstgutachter: Professor Dr. Edzer Pebesma
Zweitgutachterin: Dipl. Geow. Dorothea Lemke

Von Lars Syfuß
Münster, 07.07.2014
Matrikelnummer: 381307

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Abstract

This bachelor thesis deals with an approach for the estimation of population in small areas. The approach was developed by Klaus Steinnocher (Steinnocher et al., 2006, Steinnocher et al., 2011) and relies on the assumption that the population density is proportional to the degree of soil sealing. Population data on the community level is disaggregated to a level of electoral districts for the whole region of Münster. The "EEA Fast Track Service Precursor on Land Monitoring" dataset represents the degree of soil sealing in 20x20m grid cells and is used as data basis for this approach. The CORINE Land Cover (CLC) dataset is used to mask those areas from the EEA dataset that are not used for residential purposes. For the same reason an OpenStreetMap dataset is used to mask the streets from the EEA dataset. The approach from Steinnocher is applied to the remaining areas which represent the living space as good as possible. The calculated population data is compared to the reference population data on a level of electoral districts. Differences and tendencies are discussed. Furthermore the realized approach from Steinnocher (Steinnocher et al., 2006, Steinnocher et al., 2011) is compared to another approach developed by Francisco Javier Gallego (Gallego et al., 2001, Gallego et al., 2010). Differences and characteristics of the two approaches are discussed.

2. Zusammenfassung

Diese Bachelorarbeit befasst sich mit dem von Klaus Steinnocher (Steinnocher et al., 2006 + 2011) entwickelten Ansatz zur kleinräumigen Bevölkerungsschätzung. Der Ansatzes beruht auf der Annahme, dass die Bevölkerungsdichte proportional zu der Bebauungs- bzw. Versiegelungsdichte ist. Bevölkerungsdaten auf Gemeindeebene sollen mit Hilfe dieses Ansatz im Regierungsbezirk Münster disaggregiert werden. Als Datengrundlage wird der "EEA Fast Track Service Precursor on Land Monitoring"-Datensatz (EEA, 2009), welcher die prozentuale Versiegelung in 20x20m Rasterzellen darstellt, verwendet. Aus diesem werden, mit Hilfe des CORINE Land Cover (CLC) Datensatzes, die bebauten aber nicht für Wohnzwecke genutzten Gebiete ausmaskiert. Weiterhin werden auch Straßen durch den OpenStreetMap-Datensatz ausmaskiert. Auf den verbleibenden Flächen, welche möglichst die wirkliche Wohnfläche repräsentiert, wird der Ansatz von Klaus Steinnocher angewendet. Abschließend werden die berechneten Bevölkerungsdaten mit Referenz-Bevölkerungsdaten auf Wahlbezirksebene verglichen und Abweichungen und Tendenzen herausgestellt und begründet. Zudem wird der Ansatz von Francisco Javier Gallego (Gallego et al., 2001, Gallego et al., 2010) zur Schätzung der Bevölkerungsdichte mit dem zuvor angewendeten Ansatz von Klaus Stein-

nocher (Steinnocher et al., 2006, Steinnocher et al., 2011) verglichen, wobei Unterschiede und Besonderheiten dargelegt und begründet werden.

3. Einführung

Die kleinräumige Bevölkerungsschätzung gewinnt zunehmend an Bedeutung. Viele Anwendungsgebiete benötigen Bevölkerungsdaten für kleine Flächen. Ein Anwendungsgebiet, welches zugleich Anstoß des Themas dieser Bachelorarbeit war, ist die räumliche Überwachung von auftretenden Krebsfällen. Während die nötigen Daten aus den Krebsregistern nahezu vollständig sind, mangelt es an Daten der Bevölkerung unter Risiko für kleinräumige Flächen (Lemke et al., 2013).

Um diese Daten zu erhalten wird das „dasymetric mapping" eingesetzt. Aggregierte Werte, welche sich auf einen größeren räumlichen Ausschnitt beziehen, werden unter Zuhilfenahme anderer Datenquellen (z.B. Landnutzungsdaten) so zerlegt, dass sich die Daten auf kleinere räumliche Ausschnitte beziehen lassen. Durch diesen Ansatz wird die fehlerhafte Annahme, dass sich die aggregierten Daten homogen auf eine größere Fläche beziehen lassen, verbessert. Der Ansatz wurde im frühen zwanzigsten Jahrhundert von Benjamin Petrovich Semenov-Tyan-Shansky entwickelt aber erst 1936 durch John Kirtland Wright in den Vereinigten Staaten bekannt gemacht (Mennis, 2003). Seitdem hat der Umfang an Hilfsdaten, wie Satellitenbilder oder vergleichbare Rasterdaten, welche insbesondere aus der Fernerkundung stammen, enorm zugenommen. So kam es, dass aus verschiedenen Bereichen wieder großes Interesse an dem Ansatz aufkam. Zum Beispiel im Bereich des Umwelt- und Gesundheitsschutzes (Zandbergen et al., 2006) oder aber auch im Bereich des Katastrophenschutzes (Thieken et al., 2006). Das Verfahren wird bisher allerdings nicht einheitlich angewendet. Stattdessen gibt es unterschiedliche Ansätze und noch kein standardisiertes Verfahren.

Einige der „dasymetric mapping“-Ansätze greifen auf Landnutzungsdaten zurück. Ursprünglich wurde das Verhältnis zwischen Landnutzung und Bevölkerung zu einem großen Teil subjektiv durch den jeweiligen Anwender festgelegt (Eicher et al., 2001). Um dieses Problem zu umgehen, wurden entsprechende statistische Ansätze entwickelt.

Zu nennen ist zunächst der Ansatz von Mennis (Mennis, 2003). Er schätzte die Bevölkerungsdichte auf Rasterzellen der ursprünglichen Raumeinheit (z.B. Zensusblock) durch eine Klassifizierung in drei Urbanisierungsklassen (hoch, klein, nicht-urban) und dem Verhältnis des urbanen Bereichs zum Gesamtbereich der Raumeinheit (Mennis, 2003). Da es mit den üblichen Landnutzungsdaten, gerade im städtischen Bereich, zu Fehleinschätzungen kam,

wählten einige Wissenschaftler als Hilfsdaten die Straßennetze. Hierdurch wurden deutlich bessere Ergebnisse erzielt. Der Ansatz beruht auf der Annahme, dass die Randzonen der Haupt- und Nebenstraßen im städtischen Bereich stark besiedelt sind (Maantay et al., 2007). Eine Übersicht weiterer Ansätze können (Krunic et al., 2011) entnommen werden.

Javier Gallego und Steve Peedell (Gallego et al., 2001) haben einen Algorithmus entwickelt, welcher die Bevölkerung in Europa mit Hilfe des CORINE Land Cover (CLC) Datensatzes auf kommunaler Ebene disaggregiert (Gallego et al., 2001, Gallego et al., 2010). Im Wesentlichen ist die Bevölkerungsschätzung nach diesem Ansatz abhängig von der Bevölkerungsdichte und der Landnutzungsklassifikation. Der Ansatz generiert iterativ Flächengewichte in Form von medianen Bevölkerungsdichten, wobei die verschiedenen Landnutzungsklassen unterschiedlich gewichtet werden. Der Datensatz mit den entsprechenden Bevölkerungsdichtewerten ist frei im Internet verfügbar (EEA, 2009).

Einen anderen Ansatz haben Klaus Steinnocher, Mario Köstl und Jürgen Weichselbaum (Steinnocher et al., 2006, Steinnocher et al., 2011) entwickelt, welcher in dieser Bachelorarbeit angewendet wird. Der Ansatz beruht auf der Annahme, dass die Bevölkerungsdichte proportional abhängig von der Bebauungs- bzw. Versiegelungsdichte ist (Steinnocher et al., 2011).

Für den speziellen Anwendungsfall liegen die Bevölkerungsdaten aggregiert auf Gemeindeebene vor. Als Hilfsdatensatz wird der 20x20 Meter EEA Fast Track Service Precursor on Land Monitoring Rasterdatensatz benutzt, welcher die prozentuale Versiegelung auf diesen Flächen angibt. Das Ziel ist es, die Bevölkerung der versiegelten 20 x 20 Meter Rasterzellen zu schätzen.

In dieser Arbeit soll die Bevölkerung im Regierungsbezirk Münster (Nordrhein-Westfalen), durch Benutzung des Ansatzes von Steinnocher (Steinnocher et al., 2006, Steinnocher et al., 2011), kleinräumig bestimmt werden. Die Ergebnisdaten sollen dabei helfen die räumliche Verteilung der Hintergrundbevölkerung, aus welcher z.B. Krebsfälle hervorgehen, zu beschreiben und zu analysieren. Hierdurch kann dann die räumliche Verteilung zwischen beobachteten und zu erwartende Krebsfälle untersucht werden.

4. Daten und Methoden

Im folgenden Kapitel werden die grundlegenden Daten, welche zur Bearbeitung der Aufgabe benutzt wurden, aufgezählt und beschrieben. Um den Ausgangsrasterdatensatz zu erstellen

waren einige Anpassungen nötig, welche hier weiter erläutert werden sollen. Im Weiteren wird die Funktionsweise des Programms, welches die Bevölkerungsdaten als Ergebnisraster ausgibt, beschrieben.

4.1. Das Untersuchungsgebiet

Das Untersuchungsgebiet ist der Regierungsbezirk Münster, welcher sich im Norden Nordrheinwestfalens befindet (Abb. 1).

Abb. 1 Lage des Regierungsbezirks Münster in Deutschland [1]

Im Jahr 2011 wurde die letzte Volkszählung durchgeführt und kam zu dem Ergebnis, dass in dem Untersuchungsgebiet insgesamt 2.571.195 Menschen leben (Information und Technik, 2014). Die Gesamtfläche des Gebiets wurde zuletzt zu dem Stichtag des 31.12.2012 mit 6.917,1614 km² angegeben (Information und Technik, 2014). Legt man die genannten Zahlen zugrunde, so leben im Untersuchungsgebiet durchschnittlich ca. 372 Menschen pro km². Das Gebiet ist in 78 Gemeinden unterteilt, die wiederum fünf Kreisstädten und drei kreisfreien Städten zugeordnet sind. Im südwestlichen Teil des Untersuchungsgebietes liegt das Ruhrgebiet, welches mit seinen städtischen Ballungsräumen im Vergleich zu den meisten anderen Gemeinden signifikant höhere Bevölkerungszahlen aufweist. Den größten Teil der Fläche macht das überwiegend ländlich geprägte Münsterland aus.

4.2. Rasterdatensätze und Shapefiles

Rasterdaten bestehen aus Zellen, denen Werte für ein oder mehrere Attribute zugeordnet werden können. Für die Arbeit wurde auf die „CORINE Land Cover“- (EEA, 2006) und „EEA

[1] online abgerufen am 06.05.2014 unter http://upload.wikimedia.org/wikipedia/commons/thumb/0/08/Locator_map_RB_MS_in_Germany.svg/506px-Locator_map_RB_MS_in_Germany.svg.png

Fast Track Service Precursor on Land Monitoring“-Rasterdatensätze (EEA, 2009) zurückgegriffen.

Shapefiles speichern die Positionen und Attributinformationen von geographischen Objekten. Als geographische Objekte können Punkte, Linien und Polygone verstanden werden. Durch Polygone können auch Flächen dargestellt werden. Die für die Arbeit verwendeten Shapefiles sind vom Typ Linie oder Polygon. Für die Arbeit wurde auf drei Shapefiles zurückgegriffen, welche die Gemeindegrenzen (GKZ), die Wahlbezirksgrenzen (WBZ) und das Straßennetz im Regierungsbezirk darstellen.

Im Folgenden sollen zunächst die verwendeten Rasterdatensätze und dann die Shapefiles im Einzelnen erläutert werden.

4.2.1. EEA Fast Track Service Precursor on Land Monitoring

Der „EEA Fast Track Service Precursor on Land Monitoring“ ist ein Rasterdatensatz, welcher den Anteil von versiegelten Flächen (0-100%) auf 20x20 Meter Rasterzellen in 38 europäischen Ländern beschreibt. Diese Zellen stellen also von Menschen bebaute Flächen für Wohnraum, Industrie, Verkehr oder sonstige Nutzungen dar. Der Datensatz wurde entwickelt um den Einfluss des Menschen auf die Umwelt sichtbar zu machen. Der Datensatz bezieht sich auf das Jahr 2006 und ist seit Ende 2009 auf der Internetseite der europäischen Umweltagentur verfügbar (EEA, 2009).

Abgeleitet wurde der Datensatz aus hochauflösenden Satellitenbildern mittels automatischer Klassifikation und anschließender visueller Überarbeitung. Der Versiegelungsgrad wurde durch Zuhilfenahme von kalibrierten Vegetationsindizes berechnet (Steinnocher et al., 2011).

Da sich dieser Datensatz auf ein großes Gebiet von Europa erstreckt, musste auch hier das Referenzsystem entsprechend angepasst werden. Dies wird im Abschnitt „Wahl des Raumbezugssystems“ (4.5.1) beschrieben.

4.2.2. CORINE Land Cover

CORINE steht für „coordination of information on the environment“. Bei dem CORINE Land Cover Datensatz handelt es sich um einen Datensatz, welcher die Landnutzung auf 100x100 Meter Rasterzellen in 36 europäischen Ländern beschreibt. Es liegen 44 verschiedene Klassifikationen vor, welche die Landnutzungen repräsentieren. Der Datensatz existiert seit 2006

und ist mittlerweile in der 17. Version frei auf der Seite der europäischen Umweltagentur verfügbar (EEA, 2006). Dies zeigt, dass es einen hohen Bedarf an diesen Daten gibt.

Der CORINE-Datensatz wurde zur Ausmaskierung von Flächen aus dem Ausgangsdatensatz benutzt, auf denen keine Wohnbebauung zu erwarten ist. Hiervon waren vor allem Industrie- und Verkehrsflächen betroffen. Da der Datensatz möglichst gut auf das Gebiet des Regierungsbezirks passen soll, musste das Referenzsystem entsprechend angepasst werden. Dies wird im Abschnitt „Wahl des Raumbezugssystems“ (4.5.1) beschrieben.

4.2.3. OSM Straßennetz (Regierungsbezirk Münster)

Das Open Street Map Straßennetz des Regierungsbezirks Münster liegt als Shapefile vom Typ Linie vor. Zu jeder Linie liegen beschreibende Attribute wie Straßennamen und Straßenklassifizierung vor. Die Daten werden üblicherweise täglich aktualisiert und sind frei über den Geofabrik-Downloadserver verfügbar (GEOFABRIK, 2014). Da sich Open Street Map auf das World Geodetic System 1984 bezieht, mussten auch hier Änderungen bezüglich des Referenzsystems berücksichtigt werden. Dies wird im Abschnitt „Wahl des Raumbezugssystems“ (4.5.1) beschrieben.

4.2.4. GKZ und WBZ

Sowohl das Gemeindekennziffer- als auch das Wahlbezirk-Shapefile liegen vom Typ Polygon vor und stammen aus dem Jahr 2009. Durch die Datensätze kann den Bevölkerungsdaten ein Raumausschnitt zugeordnet werden. Das Gemeindekennziffer-Shapefile speichert die Geometrien und Attribute der 78 Gemeinden des Regierungsbezirks Münster. Jede Gemeinde kann durch eine eindeutige Nummer identifiziert werden, welche als Attribut „KGS8“ vorliegt.

Das Wahlbezirk-Shapefile speichert die Geometrien und Attribute der 1.983 Wahlbezirke, des Regierungsbezirks Münster. Jeder Wahlbezirk kann durch seine eindeutige Nummer identifiziert werden, welches als Attribut „KGS22“ vorliegt. Zusätzlich wird jeder Wahlbezirksnummer auch eine Gemeindekennziffer und der Gemeindename zugeordnet.

Beide Shapefiles liegen im Raumbezug „ETRS_1989_UTM_Zone_32N“ vor und geben das Zielsystem vor, da dieses für den Regierungsbezirk das am besten geeignete Raumbezugssystem ist.

4.3. Bevölkerungsdaten

Die Bevölkerungsdaten liegen für die 78 Gemeinden aus dem Jahr 2005 vor. Die Daten liegen für Männern und Frauen im Alter von 0-65 Jahren getrennt vor und wurden von INFAS Geodaten, heute „NEXIGA next level geomarketing“, erworben (NEXIGA, 2014). Weiterhin liegen die Bevölkerungsdaten der 1.983 Wahlbezirke von 2005 vor. Die Wahlbezirke sind mit ca. 500 Haushalten pro Bezirk in etwa gleich frequentiert. Dies bringt eine bedeutend höhere räumliche Auflösung im Vergleich zu den Daten auf Gemeindeebene mit sich. Die mittlere Bevölkerungsdichte beträgt 1.533 Einwohner pro km² wobei das Minimum bei vier Einwohnern und das Maximum bei 13.615 Einwohnern pro km² liegt. Die Daten der Wahlbezirke dienen als Referenzdaten zur Auswertung der Ergebnisse.

4.4. Beschreibung des Ansatzes von Steinnocher (2011)

Steinnocher verfolgt einen Disaggregationsansatz auf der Grundlage eines Rasterdatensatzes, der den Versiegelungsgrad bebauter Flächen in Europa repräsentiert (Steinnocher et al., 2011). Der Ansatz beruht auf der Annahme, dass die Bevölkerungsdichte proportional zu der Bebauungs- bzw. Versiegelungsdichte ist. Der Vorteil dieses Ansatzes besteht in der hohen räumlichen Auflösung (20x20 Meter) des Landbedeckungsdatensatzes (EEA Fast Track Service Precursor on Land Monitoring) (siehe 4.2.1). Dieser dient als Datengrundlage für die räumliche Disaggregation der Bevölkerungsdaten.

Für die räumliche Disaggregation muss eine funktionale Beziehung zwischen den aggregierten (die zu verteilenden Daten) und den Hilfsparametern (die Daten mit der höheren Auflösung) vorliegen. Als Hilfsparameter wird hier die Bebauungsdichte herangezogen.

Der Ansatz basiert im Kern auf den folgenden drei Annahmen:

- die Bevölkerungsdichte ist proportional zur Bebauungsdichte
- die Bevölkerung tritt nur innerhalb der bebauten Flächen auf
- innerhalb einer Ausgangsregion (hier Gemeinde) ist das Verhältnis von Bevölkerungs- zu Bebauungsdichte konstant

Diese Annahmen können durch folgende Formeln beschrieben werden:

(1) $$BevDi = k * BebDi$$

(2) $$Bev = \sum_i A_i * k * BebDi_i$$

Folgende Tabelle beschreibt die Bedeutung der Variablen (Tab. 1).

Tab. 1 Bedeutung der Variablen in dem formalisierten Ansatz von Steinnocher

Variable	Bedeutung
$BevDi$	Die Bevölkerungsdichte des lokalen Zielgebiets
$BebDi$	Die Bebauungsdichte des lokalen Zielgebiets
k	Das Verhältnis von der gesamten Bevölkerungsdichte zur gesamten Bebauungsdichte des lokalen Zielgebiets.
Bev	Die Gesamtbevölkerung (des Untersuchungsgebiets)
A_i	Die Fläche mit der Bebauungsdichte i

Bei der Anwendung des Ansatzes wird zunächst die zweite Formel angewendet um den Faktor k zu bestimmen. Anschließend wird die erste Formel angewendet um die Bevölkerung für das lokale Zielgebiet zu bestimmen.

Die Bebauungsdichte wird aus der Versiegelungsdichte abgeleitet. Dabei ist zu beachten, dass alle vorhandenen versiegelten Flächen in dem Raster berücksichtigt werden. Für den Ansatz von Steinnocher (Steinnocher et al., 2011) soll jedoch nur die versiegelten Flächen mit einer Wohnnutzung berücksichtigt werden. Um dies zu ermöglichen, werden unter Zuhilfenahme des CORINE Landcover Datensatzes alle Rasterzellen ausmaskiert, denen keine Wohnnutzung zugeordnet werden kann. Diese umfassen insbesondere:

- Industriell und gewerblich genutzte Flächen
    - Industrie, Gewerbe und Transportflächen (CLC Klasse 1.2)
    - Abbauflächen, Deponien und Baustellen (CLC Klasse 1.3)
    - Städtische Grünflächen, Sport- und Freizeitanlagen (CLC Klasse 1.4)
- Verkehrsflächen
    - Gebufferte und gerasterte Straßennetzdaten

Erst dadurch ist das Raster für die Disaggregation geeignet. Flächen mit einem Versiegelungsgrad von 100%, die außerhalb von Siedlungen liegen, sind größtenteils industriell oder gewerblich genutzte Flächen und werden daher auch ausmaskiert. Das verbleibende Raster repräsentiert die Bebauungsdichte von Wohngebieten und ist als Hilfsparameter für die Disaggregation verwendbar.

4.5. Anpassung der Daten

Um mit den Ausgangsdaten arbeiten zu können mussten einige Anpassungen vorgenommen werden. Diese Anpassungen beziehen sich insbesondere auf das Raumbezugssystem und die Auswahl der versiegelten Zellen, die eine Wohnnutzung repräsentieren. Im Folgenden sollen diese Anpassungen beschrieben werden.

4.5.1. Wahl des Raumbezugssystems

Durch ein Raumbezugssystem ist es möglich, Punkte zu einem definierten Zeitpunkt durch ihre Lage untereinander zu beschreiben und diese durch Koordinaten abbildbar zu machen. Je nach Ausmaß eines Gebietes und nach dem Zweck der Abbildung kann es Sinn machen verschiedene Raumbezugssysteme zu benutzen. Je größer ein Gebiet ist, desto schwerer wird es alle Punkte in diesem Gebiet mit einer hohen Genauigkeit bestimmen zu können, da die Erdoberfläche sehr unregelmäßig ist.

Da die Ausgangsdaten in verschiedenen Referenzsystemen vorlagen (Tab. 2), musste zunächst entschieden werden in welchem Raumbezugssystem gearbeitet werden soll.

Tab. 2 Verwendete Datensätze und ihre Referenzsysteme

Datensatz	Raumbezug	Datum
CORINE Land Cover	ETRS_1989_LAEA_L52_M10	D_ETRS_1989
EEA Fast Track Service Precursor	ETRS_1989_LAEA	D_ETRS_1989
OSM Straßen Shapefile	GCS_WGS_1984	D_WGS_1984
GKZ Shapefile	ETRS_1989_UTM_Zone_32N	D_ETRS_1989
WBZ Shapefile	ETRS_1989_UTM_Zone_32N	D_ETRS_1989

Die „CORINE Land Cover“- und „EEA Fast Track Service Precursor on Land Monitoring“-Datensätze beziehen sich beide auf einen großen Teil Europas und haben daher den Raumbezug „ETRS_1989_LAEA“. ETRS 1989 steht für das Europäische Terrestrische Referenzsystem von 1989. LAEA steht für „Lambert Azimuthal Equal Area“. Das System dient der flächentreuen Darstellung der Daten in Europa.

Das OSM Straßen Shapefile hat den Raumbezug WGS 84. Dies steht für das World Geodetic System von 1984. Dieses System ist der Erdoberfläche überall auf der Erde bestmöglich an-

gepasst aber führt in lokalen Systemen, wie beispielsweise den Regierungsbezirk Münster, zu Ungenauigkeiten.

Die Gemeinde- und Wahlbezirk-Shapefiles haben den Raumbezug „ETRS_1989_UTM_Zone_32N“. Sie stellen im Vergleich die lokalsten Datensätze dar und sind daher auch mit einem lokaleren Raumbezug versehen. Der Regierungsbezirk Münster liegt vollständig in der UTM Zone 32. In diesem Referenzsystem sind die kleinsten Ungenauigkeiten zu erwarten, weshalb es als Zielsystem für alle erstellten Datensätze verwendet wird.

4.5.2. Bearbeitung der Rasterdaten

Für die Bearbeitungsschritte wurde das Programm ArcGis 10.1 von ESRI verwendet. Das Ziel der Anpassungen ist ein Rasterdatensatz auf Basis des „EEA Fast Track Service Precursor on Land Monitoring“, welcher folgende Anforderungen erfüllt:

- Nur Zellen innerhalb des Regierungsbezirks Münster sind von Interesse
- Die Corine Land Cover Klassen 1.2 - 1.4 (siehe 4.4 Beschreibung des Ansatzes von Steinnocher) sollen ausmaskiert werden, da sie keine Wohnfunktion erfüllen
    - Industrie-, Gewerbe- und Verkehrsflächen [CLC-Klasse 1.2]
    - Abbauflächen, Deponien, Baustellen [CLC-Klasse 1.3]
    - Künstlich angelegte, nicht landwirtschaftlich genutzte Flächen [CLC-Klasse 1.4]
- Die Verkehrsflächen sollen ausmaskiert werden

Um diesen Anforderungen gerecht zu werden, wurden verschiedene ArcGis Funktionen aus der ArcToolbox verwendet (Tab. 3).

Tab. 3 Verwendete ArcToolbox Funktionen

ArcToolbox Funktion	Navigation zur Funktion	Aufgabe
Nach Maske extrahieren	Spatial Analyst Tools → Extraktion → Nach Maske extrahieren	Ein Rasterausschnitt wird durch eine Vorlage ausgeschnitten
Reklassifizieren	Spatial Analyst Tools → Reklassifizieren → Reklassifizieren	Teilt ausgewählten Zellen einen neuen Wert zu
Mosaik zu neuem Raster	Data Management Tools → Raster → Raster-Dataset → Mosaik zu neuem Raster	Fügt mehrere Raster zu einem neuen Raster zusammen

Die wichtigsten Verarbeitungsschritte von den Ausgangsrasterdatensätzen bis zum Raster, welches die bewohnten Rasterzellen darstellt, sind im folgenden Flowchart (Abb. 2) dargestellt.

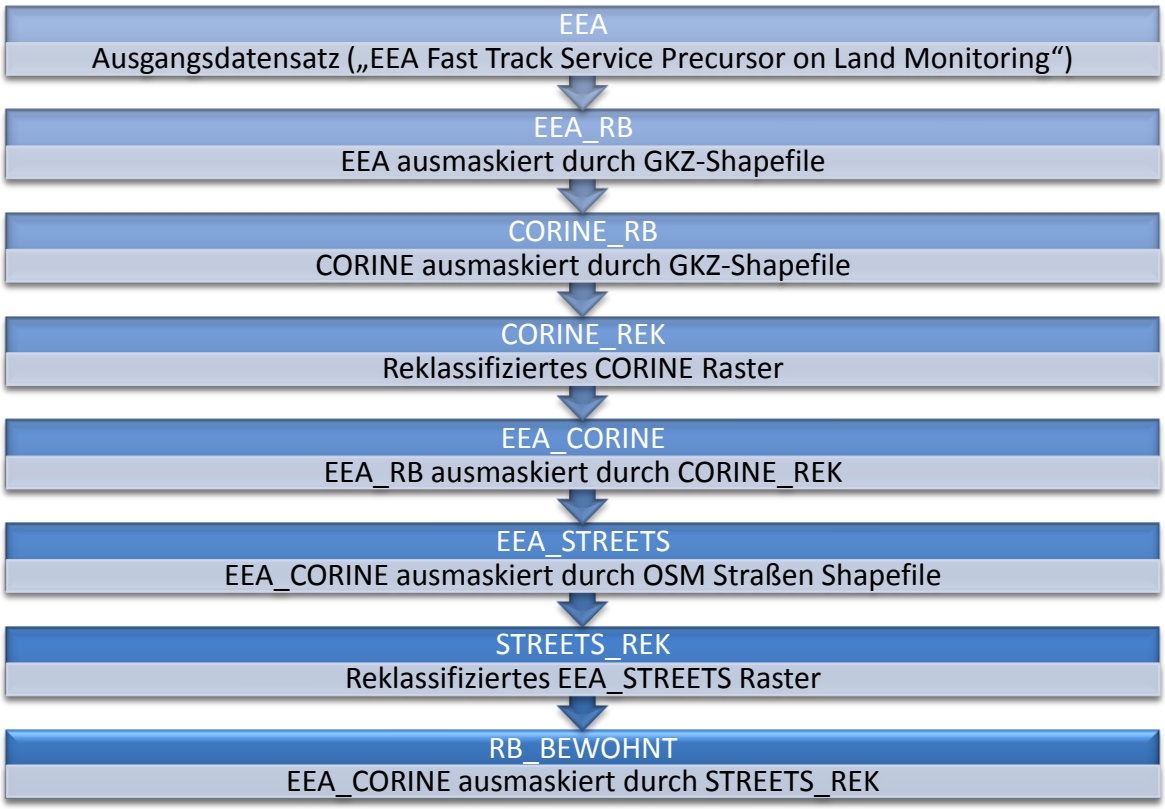

Abb. 2 Die wichtigsten Verarbeitungsschritte zur Erstellung des Ausgangsdatensatzes

Diese Arbeitsschritte werden im Folgenden genauer erläutert. Zunächst wurde der „EEA Fast Track Service Precursor on Land Monitoring“-Rasterdatensatz in ArcGis eingeladen. In dem Flowchart wurde der Datensatz als „*EEA*“ bezeichnet (Abb. 2). Der Datensatz wird in folgender Abbildung dargestellt (Abb. 3). Die verschiedenen Farben der Zellen, welche die verschiedenen Werte repräsentieren, sind in diesem Maßstab schlecht zu erkennen jedoch soll hier das Ausmaß des Datensatzes deutlich werden.

Abb. 3 Der "EEA Fast Track Service Precursor on Land Monitoring"-Datensatz bedeckt 38 europäische Staaten. (Screenshot in ArcGis)

Der Datensatz sollte zunächst auf die Größe des Regierungsbezirks angepasst werden. Dafür wurde die ArcToolbox Funktion *Nach Maske extrahieren* benutzt, wobei der „EEA Fast Track Service Precursor on Land Monitoring“ Datensatz als Eingaberaster und das GKZ-Shapefile als Vorlage zur Ausmaskierung benutzt wurde. Dem Ausgaberaster (Abb. 4) wurde der Raumbezug „ETRS_1989_UTM_Zone_32N“ zugewiesen. In dem Flowchart wurde der Datensatz als „*EEA_RB*“ bezeichnet (Abb. 2).

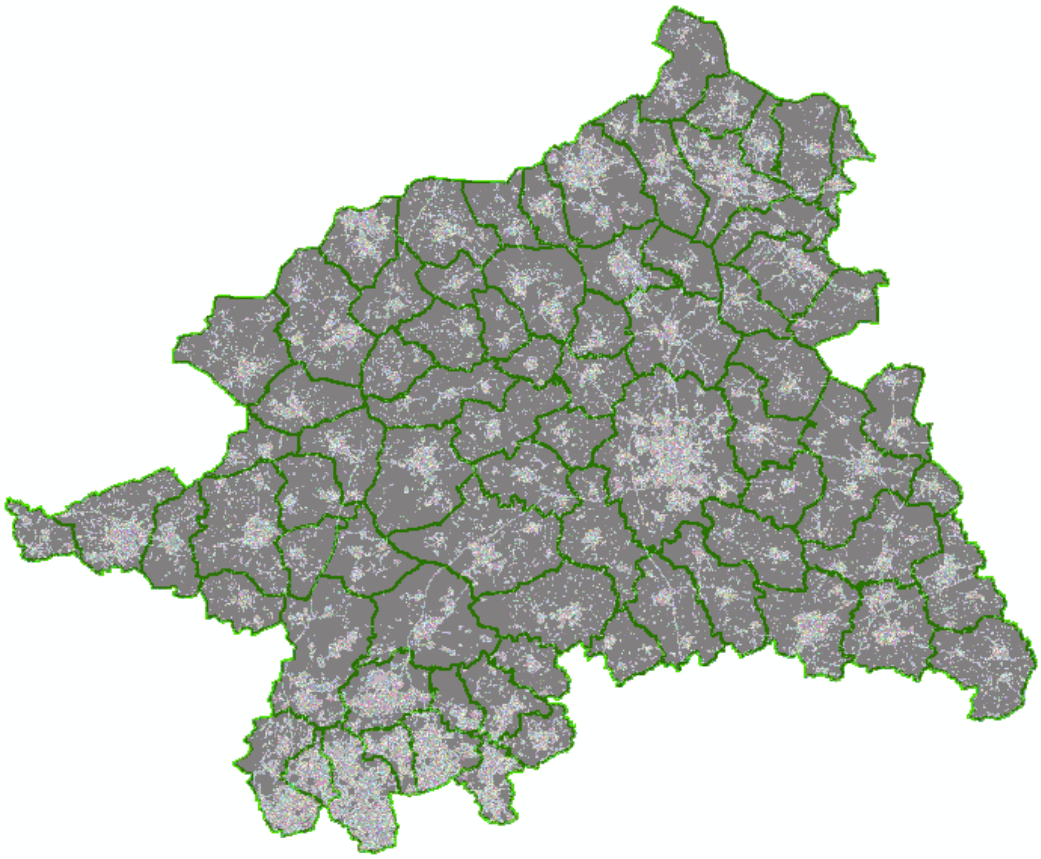

Abb. 4 Der "EEA Fast Track Service Precursor on Land Monitoring"-Datensatz angepasst auf die Größe des Regierungsbezirks Münster. Die Gemeindegrenzen sind in grün dargestellt. (Screenshot in ArcGis)

Genauso wurde der CORINE Land Cover Datensatz bearbeitet um ihn auf die Ausmaße des Regierungsbezirks anzupassen. Auch diesem wurde, wie allen folgenden Datensätzen, der Raumbezug „ETRS_1989_UTM_Zone_32N“ zugewiesen. In dem Flowchart wurde das Raster „*CORINE_RB*“ genannt (Abb. 2).

Die nicht bewohnten CLC-Klassen mussten als nächstes reklassifiziert werden, sodass sie zur Ausmaskierung benutzt werden konnten. Zu diesen gehören die zu Beginn dieses Kapitels genannten CLC-Klassen 1.2 - 1.4 (4.5.2 Bearbeitung der Rasterdaten).

Mit der ArcToolbox Funktion *Reklassifizieren* wurde diesen Rasterzellen der Wert „NoData“ zugewiesen. In dem Flowchart wurde der Datensatz als „*CORINE_REK*“ bezeichnet (Abb. 2). Die Folgende Abbildung zeigt das Ergebnis (Abb. 5). Die Rasterzellen mit „NoData“-Werten werden in weiß dargestellt. Es ist bereits gut zu erkennen, dass der Großteil der versiegelten Flächen in den Stadtkernen zu finden ist.

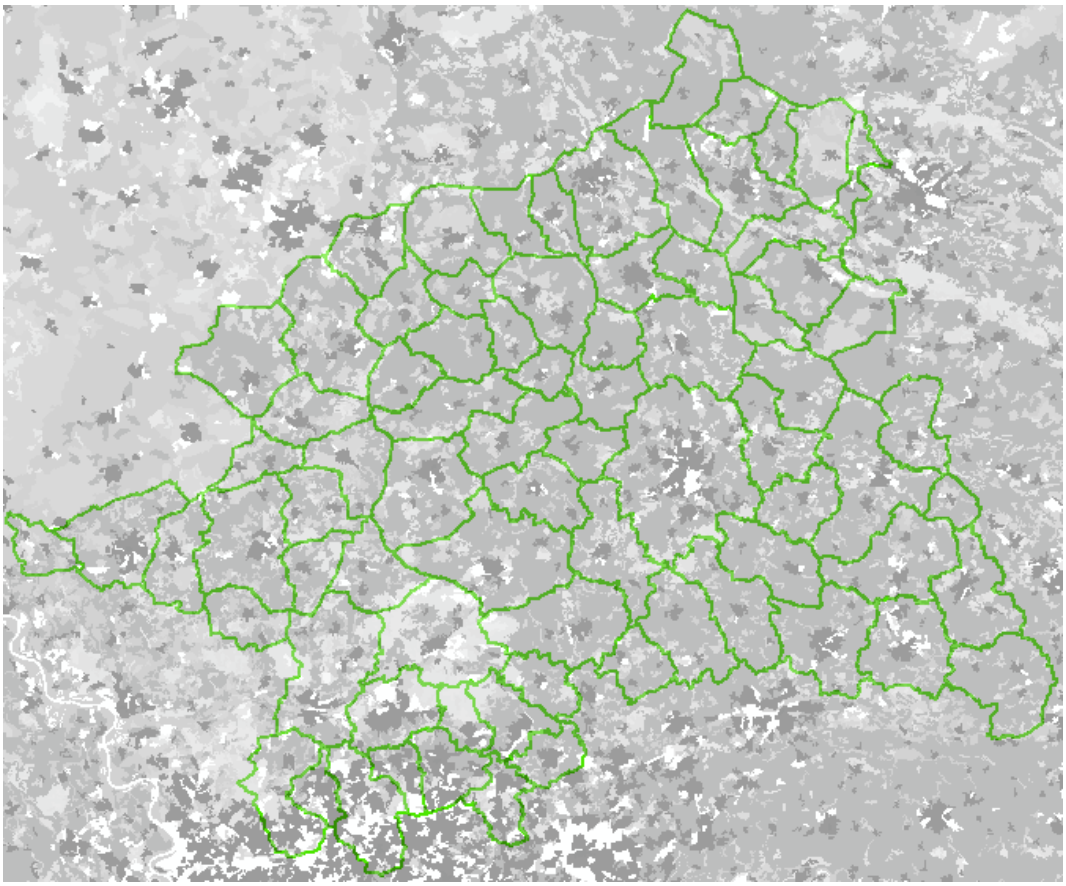

Abb. 5 CORINE Land Cover Datensatz mit reklassifizierten Rasterzellen."NoData"-Werte sind in weiß und die Gemeindegrenzen in grün dargestellt. (Screenshot in ArcGis)

Mit diesem Raster konnten anschließend mit der bereits bekannten ArcToolbox Funktion *Nach Maske extrahieren* nur die Zellen aus dem zuvor erzeugten Raster „*EEA_RB*“ maskiert werden, welche nicht den Wert „NoData“ haben. Hierzu wurde das zuvor erzeugte Raster „*CORINE_REK*“ benutzt. In dem so erzeugten Ausgaberaster waren also alle 20x20 Meter Rasterzellen des *„EEA_RB“*-Datensatzes, die von keinem der ausmaskierten CLC-Klassen geschnitten wurden (Abb. 6). In dem Flowchart wurde der Datensatz als „EEA_*CORINE*“ bezeichnet (Abb. 2).

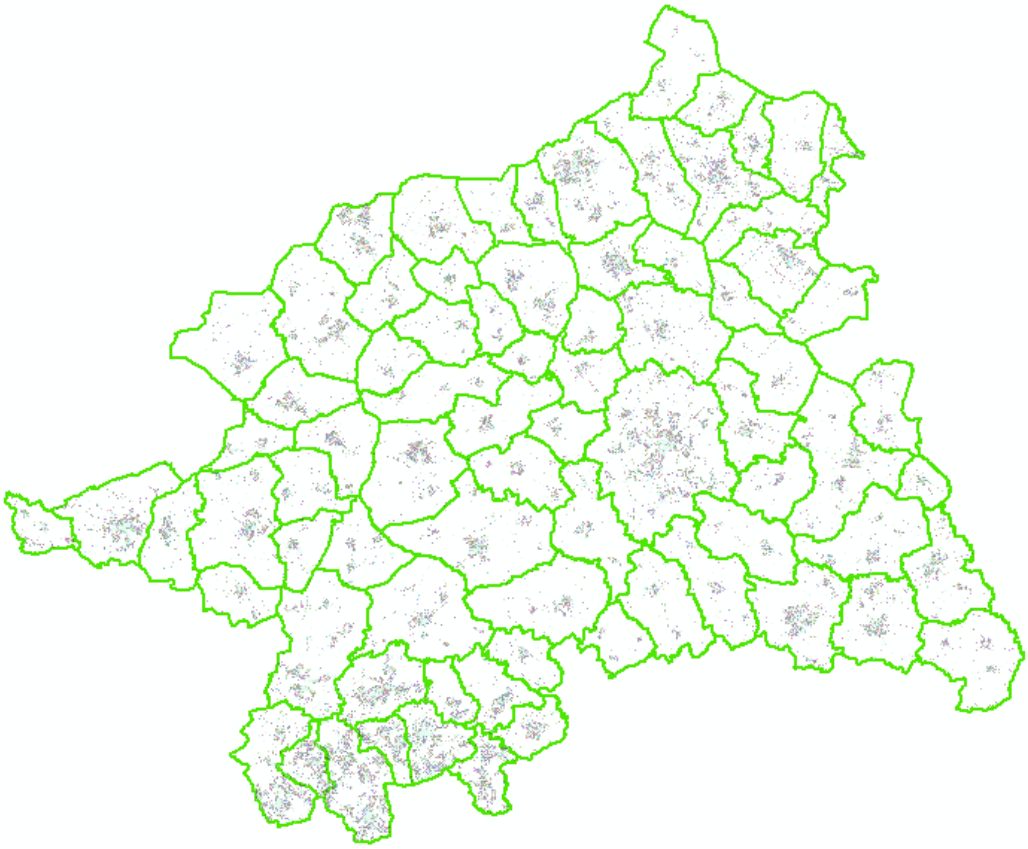

Abb. 6 Bewohnte Zellen des RB Münster."NoData"-Werte sind in weiß und die Gemeindegrenzen in grün dargestellt. (Screenshot in ArcGis)

Da auffällig viele Zellen von Straßen geschnitten wurden, sollten in einem letzten Schritt aus dem zuletzt erzeugten Raster „*EEA_CORINE“* auch alle Zellen ausmaskiert werden, die von Straßen geschnitten werden. Hierzu dient das OSM Straßen Shapefile, mit dessen Hilfe alle Straßen aus dem Regierungsbezirk dargestellt werden können.

Zunächst wurden mit der ArcToolbox Funktion *Nach Maske extrahieren* alle Rasterzellen extrahiert, die von Straßen geschnitten werden. In dem Flowchart wurde der Datensatz als „EEA_*STREETS*“ bezeichnet (Abb. 2). Einen Ausschnitt aus diesem Raster zeigt der folgende Screenshot, welcher auf einer hohen Zoomstufe erzeugt wurde (Abb. 7). Es ist deutlich zu erkennen, dass viele versiegelte Flächen von Straßen geschnitten werden.

Abb. 7 Ausmaskierte bewohnte Straßen. Die Straßen des OSM Shapefiles sind als schwarze Linien zu erkennen und die ausmaskierten Rasterzellen als farbige Quadrate. (Screenshot in ArcGis)

Da bei der Ausmaskierung alle Zellen auf „NoData“ gesetzt werden, die von „NoData“-Zellen geschnitten werden, musste das „*EEA_STREETS*“-Raster invertiert werden. Mit der bereits bekannten ArcToolbox Funktion *Reklassifizieren* haben folglich alle Zellen einen Wert bekommen, die zuvor den Wert „NoData“ hatten und umgekehrt (Abb. 8).

Abb. 8 Das invertierte Raster der ausmaskierten Straßen. Die Straßen des OSM Shapefiles sind als schwarze Linien zu erkennen. (Screenshot in ArcGis)

In dem Flowchart wurde der Datensatz als „*STREETS_REK*“ bezeichnet (Abb. 2). Dieser konnte zur Ausmaskierung der Straßen benutzt werden.

In dem letzten Schritt wurden mit der ArcToolbox Funktion *Nach Maske extrahieren* also alle Rasterzellen aus dem „*EEA_CORINE*“-Raster ausmaskiert, die zuvor Werte hatten und von Straßen geschnitten wurden. Das Ausgaberaster ist der fertig bearbeitete Rasterdatensatz (Abb. 9) und wird im Flowchart als „*RB_BEWOHNT*“ bezeichnet (Abb. 2). Natürlich kann auf Grund der groben Daten nicht für alle Zellen eine Wohnnutzung garantiert werden. Jedoch enthält dieser Datensatz annähernd nur die versiegelten Flächen mit einer Wohnnutzung und kann als Ausgangsdatensatz für das Programm benutzt werden.

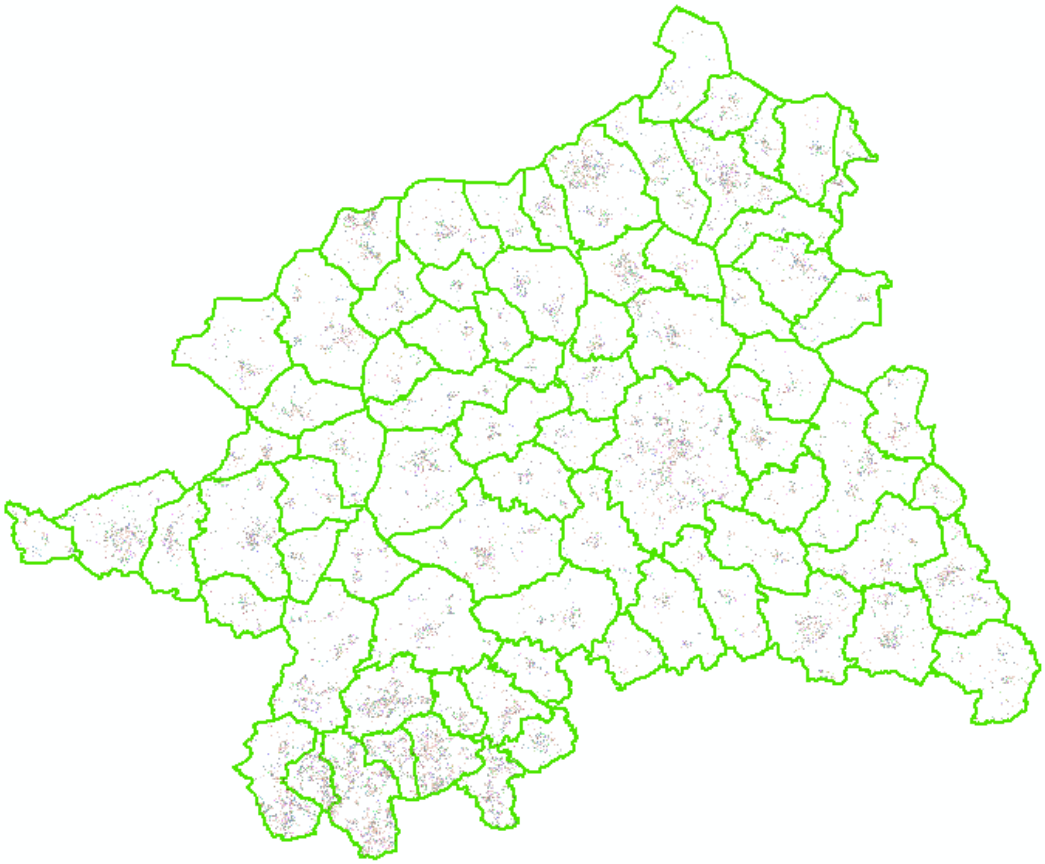

Abb. 9 Fertig bearbeiteter Rasterdatensatz. Die Gemeindegrenzen sind in grün dargestellt. (Screenshot in ArcGis)

4.6. Funktionsweise des Programms

Um die Bevölkerung auf die Rasterzellen zu verteilen und die Ergebnisse anschließend überprüfen zu können, wurde ein Programm in der Programmiersprache R (R, 2014) geschrieben. Programmiert wurde in RStudio (RStudio, 2014) (Version 0.98.501). Zur Umsetzung wurde auf vorhandene Packages zurückgegriffen (Tab. 4).

Tab. 4 Verwendete R-Packages

Package	Benutzt für…
raster	Arbeiten mit Rasterdaten
Rgdal	Arbeiten mit Shapefiles
gdata	Arbeiten mit Exceldateien
ggplot2	Erstellung der Grafiken (zur Auswertung)

Der Ausgangsrasterdatensatz repräsentiert den Grad der Versiegelung in seinen 20x20m Rasterzellen. Das Raster wurde so erstellt, wie es im vorangegangenen Kapitel (4.5.2 Bearbeitung der Rasterdaten) beschrieben wurde. Die Funktionsweise des Programms soll der Einfachheit halber durch Pseudocode verdeutlicht werden. Kommentare werden durch eine Raute (#) eingeleitet und grün dargestellt. Der komplette Code ist im Anhang zu finden.

Erstellung des disaggregierten Rasters:

```
Lade den Ausgangsrasterdatensatz
Lege Listen für die zu speichernden Daten an # GKZ-Nummer, GKZ-Name, berechneter Faktor K, berechnete Einwohnerzahl, Anzahl der Zellen mit einem Wert != NoData
Für jede Gemeinde des Untersuchungsgebiets # aus dem Gemeinde-Shapefile (GKZ)
    Lese die Einwohnerzahl aus Excel
    Lese Gemeindename aus Excel
    Wähle das Polygon der entsprechenden Gemeinde aus dem Shapefile
    Maskiere die Gemeinde aus dem Ausgangsraster aus
    Zähle Zellen des ausmaskierten Gemeinderasters, welche einen Wert haben
    Summiere die Zellwerte des ausmaskierten Gemeinderasters
    Berechne den Faktor K
    Erstelle das neue Populationsraster # ausmaskiertes Gemeinderaster * Faktor k
    Hänge zu speichernde Attribute an die entsprechenden Listen an
Schreibe zu speichernde Daten in eine Datei
```

Die Populationsraster werden für jede Gemeinde einzeln gespeichert. Mit der ArcGis Funktion *Mosaik zu neuem Raster* (Tab. 3) lassen sie sich leicht zusammenfügen. Dieses Vorgehen wurde so gewählt, weil es den Zugriff auf einzelne Gemeinden erleichtert.

Um die Ergebnisraster überprüfen zu können, mussten die Daten noch mit den Referenzdaten auf Wahlbezirksebene verglichen werden. Hierzu mussten zunächst die Daten auf Wahlbezirksebene aus dem zuvor erzeugten Raster ausgelesen werden. Auch hierfür wurde eine entsprechende Funktion geschrieben.

Berechnung der Bevölkerung auf Wahlbezirksebene:

```
Lade den Rasterdatensatz mit den Populationswerten
Lege Listen für die zu speichernden Daten an # Wahlbezirksnummer, berechnete Einwohnerzahl
Für jeden Wahlbezirk des Untersuchungsgebiets # aus dem WBZ-Shapefile
    Wähle das Polygon des entsprechenden Wahlbezirks aus dem Shapefile
    Maskiere den Wahlbezirk aus dem Populationsraster aus
    Berechne die Einwohnerzahl des ausmaskierten Wahlbezirksrasters
    Hänge zu speichernde Attribute an die entsprechende Liste an
Schreibe zu speichernde Daten in eine Datei
```

Die so erhaltenen Daten können mit den vorliegenden Bevölkerungsdaten (siehe 4.3 Bevölkerungsdaten) verglichen werden um die Daten auszuwerten. Die Auswertung wird im folgenden Abschnitt behandelt.

4.7. Fehlerberechnung

Die absolute Abweichung ergibt sich indem die berechnete Population von den Referenzwerten der Population subtrahiert wird. Je nach tatsächlicher Population ist dies natürlich nicht besonders aussagekräftig, weshalb auch der relative Fehler ermittelt wurde.

Der relative Fehler gibt an, welchen Anteil die berechnete Population an den Referenzwerten aus dem Jahr 2005 hat. Formal kann dies beschrieben werden durch:

(1) $Abw_{abs} = Pop_{ber} - Pop_{2005}$

(2) $RE = \frac{Abw_{abs} * 100}{Pop_{2005}}$

Folgende Tabelle beschreibt die Bedeutung der Variablen (Tab. 5).

Tab. 5 Bedeutung der Variablen zur Berechnung der absoluten und relativen Abweichung

Variable	Bedeutung
Abw_{abs}	Die absolute Abweichung
Pop_{ber}	Die berechnete Einwohnerdichte (Population)
Pop_{2005}	Die Referenzdaten der Einwohnerdichte (aus dem Jahr 2005)
RE	Der relative Fehler

5. Auswertung / Ergebnisse

Das Auswertungskapitel gliedert sich in zwei Teile: Zuerst wurden die berechneten Parameter auf Gemeindeebene ausgewertet. Danach werden die Ergebnisse des Vergleichs zwischen dem berechneten Populationsraster und den Referenzdaten auf Wahlbezirksebene dargestellt. Diese Auswertungen sollen im Folgenden beschrieben werden.

5.1. Auswertung der Ergebnisse auf Gemeindeebene

Folgende Parameter wurden auf Gemeindeebene für die kleinräumige Bevölkerungsschätzung berechnet:

- Faktor K
- Anzahl der Zellen mit vorhandenem Versiegelungswert (ungleich NoData)
- Die Mittelwerte des Versiegelungsgrads (nur Zellen mit vorhandenem Versiegelungsgrad wurden berücksichtigt)

Zunächst soll die Bedeutung des Faktors K herausgestellt werden. Der Faktor K wurde berechnet indem die Gesamtbevölkerung einer Gemeinde durch die Summe der Versiegelung einer Gemeinde dividiert wurde (siehe auch 4.4 Beschreibung des Ansatzes von Steinnocher (2011)). Somit drückt der Faktor K die Beziehung zwischen Bevölkerungs- und Bebauungsdichte aus. Um die Bevölkerung für eine Zelle berechnen zu können, wird der gegebene Versiegelungsgrad mit dem Faktor K der Gemeinde multipliziert.

Ein kleiner Faktor K bedeutet also, dass ein entsprechend hoher Versiegelungsgrad vorhanden sein muss, damit einer Zelle (20x20 Meter) in der jeweiligen Gemeinde Einwohner zugeordnet werden können. Bei einem Faktor K von 0,02 muss beispielsweise ein Versiegelungsgrad von 50% vorhanden sein um einer Zelle einen Einwohner (50*0,02) zuordnen zu können.

Maximal können bei einem Faktor von 0,02 und einem Versiegelungsgrad von 100% nur zwei Einwohner (100*0,02) pro Zelle vorhanden sein. Ein großer Faktor K ermöglicht hingegen eine entsprechend höhere Einwohnerzahl pro Zelle.

Probleme können sich dort ergeben, wo die Bevölkerung sehr unregelmäßig verteilt ist. Wenn die Bevölkerung auf versiegelten Flächen im städtischen Bereich sprunghaft anwächst, wie dies in Großstädten durchaus passieren kann, reicht ein kleiner Faktor K nicht aus um der tatsächlichen Einwohnerdichte zu genügen und die Bevölkerung wird unterschätzt. Andersherum kann es auch vorkommen, dass außerhalb des städtischen Bereichs die Bevölkerung sprunghaft abnimmt und ein großer Faktor K in diesen Gegenden zu Überschätzungen führen kann. Die folgende Abbildung (Abb. 10) zeigt die Verteilung des Faktors K im Regierungsbezirk Münster.

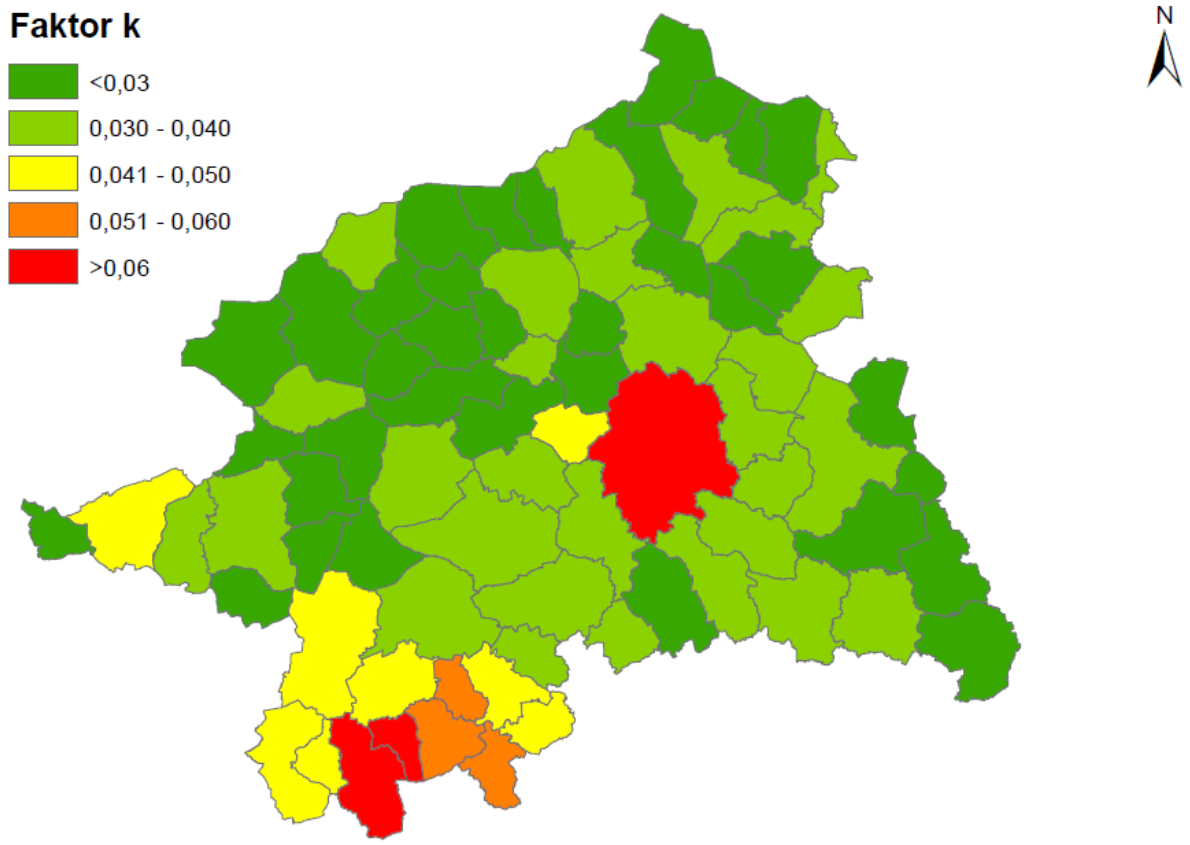

Abb. 10 Räumliche Verteilung des Faktors K im Regierungsbezirk Münster

Wie zu erwarten ist das Ruhrgebiet als Ballungsgebiet im Süden mit einem entsprechend hohen Faktor K zu erkennen. Im Westen ist noch Bocholt mit einem erhöhten Faktor vertreten. Münster sticht im Zentrum mit einem sehr hohen Faktor K heraus und westlich davon hat auch die Gemeinde Havixbeck einen erhöhten Faktor K. Die anderen Gemeinden im Regie-

rungsbezirk sind eher ländlich geprägt und haben, wie es zu erwarten war, einen entsprechend kleinen Faktor K.

Mit Hilfe des Faktors K konnte dann das Ergebnisraster, welches die Einwohnerdichte pro versiegelter Rasterzelle enthält, erstellt werden. Die Einwohnerdichte reicht von null bis sechs Einwohner pro Rasterzelle. Zur Übersicht wurde eine Karte mit den kategorisierten Einwohnerzahlen erstellt (Abb. 11):

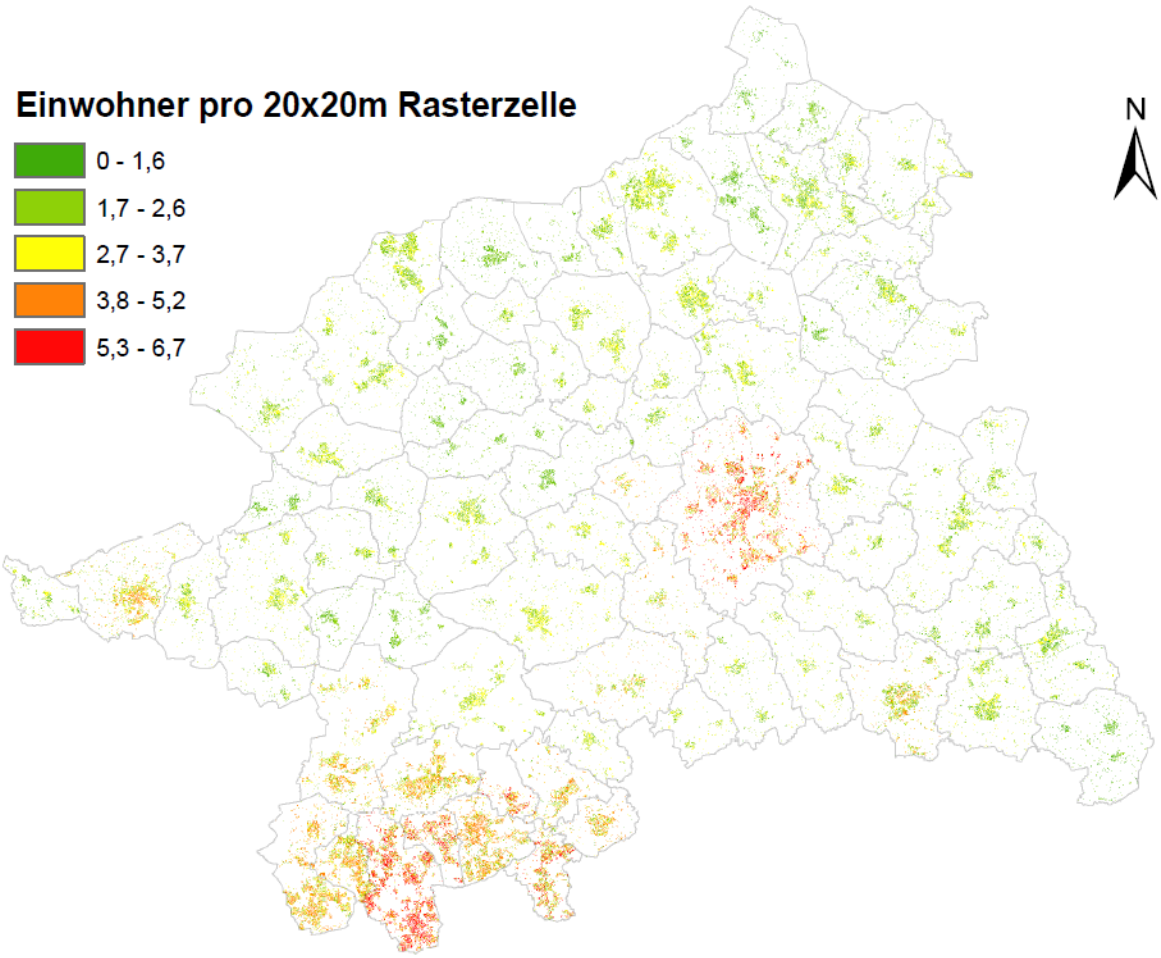

Abb. 11 Ergebnisraster - Verteilung der Bevölkerung im Regierungsbezirk Münster

Die Parallelen zu der Verteilung des Faktors K im Regierungsbezirk (Abb. 10) werden sofort deutlich. Die Population pro Rasterzelle ist dort besonders hoch, wo auch der Faktor K entsprechend groß ist. Besonders viele Einwohner pro Rasterzelle sind im Ruhrgebiet und in den Gemeinden Bocholt und Münster zu finden.

Folgender Scatterplot zeigt die Verteilung des Faktors K auf die Anzahl der versiegelten Zellen und die Mittelwerte des Versiegelungsgrads (Abb. 12).

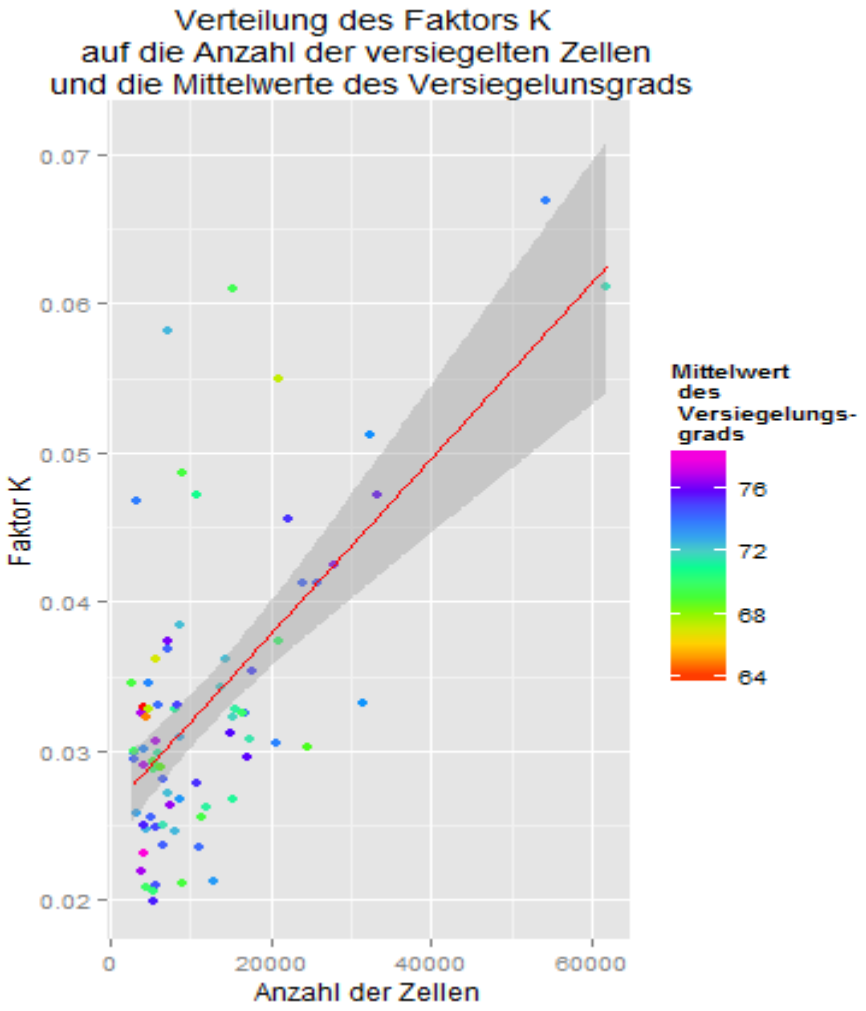

Abb. 12 Verteilung des Faktors K auf die Anzahl der Versiegelten Zellen und die Mittelwerte des Versiegelungsgrads

Der maximale Faktor K von rund 0,07 ist in Gelsenkirchen zu finden. Die Anzahl der Zellen liegt hier bei 54.304 und der Mittelwert des Versiegelungsgrads bei 74,8%. Gelsenkirchen gehört zum Ruhrgebiet und stellt einen Ballungsraum dar, was die großen Werte für den Faktor K und die Anzahl der Zellen erklärt. Der minimale Faktor K von rund 0,02 liegt in Heek. Die Anzahl der Zellen liegt hier bei 5.567 und der Mittelwert des Versiegelungsgrads bei 75,6%. Heek liegt im westlichen Münsterland und ist sehr ländlich geprägt, was die kleinen Werte für den Faktor K und die Anzahl der Zellen erklärt. Eine berechnete Zelle repräsentiert 400 Quadratmeter. Mit dem minimalen Faktor von ca. 0,02 können einer Zelle bei 100% Versiegelung nur zwei Einwohner zugeordnet werden. Wie dieses Beispiel zeigt, führt ein kleiner Faktor K im städtischen Bereich tendenziell zu Unterschätzungen.

Die minimale Anzahl an Zellen mit vorhandener Versiegelungsdichte, die auf Wohnbebauung zurückzuführen ist, ist mit 2.663 Zellen in Laer zu finden. Der Faktor K liegt hier bei 0,03 und der Mittelwert des Versiegelungsgrads bei 69,1%. Laer liegt im ländlich geprägten Münster-

land und hat einen Dorfcharakter, was die geringe Wohnbebauung erklärt und somit auch die geringen Werte für den Faktor K und die Anzahl der Zellen. Die maximale Anzahl an Zellen mit vorhandener Versiegelungsdichte ist mit 61.809 Zellen in Münster zu finden. Der Faktor K liegt hier bei 0,06 und der Mittelwert des Versiegelungsgrads bei 71,7%. Münster hat knapp 300.000 Einwohner und ist das Oberzentrum des Münsterlandes, was die große Wohnbebauung erklärt und somit auch die hohen Werte für den Faktor K und die Anzahl der Zellen.

Es liegt eine tendenzielle Abhängigkeit zwischen dem Faktor K und der Anzahl der Zellen mit einer Wohnnutzung in einer Gemeinde vor. Je mehr versiegelte Flächen in einer Gemeinde vorhanden sind, desto größer ist der Faktor K. Dies stellt auch die rote Trendlinie dar (Abb. 12). Damit dieser Zusammenhang stimmt, muss natürlich auch die Gesamtbevölkerung der Gemeinde entsprechend hoch sein. Zumeist geht eine hohe Wohnfläche auch mit einer hohen Gesamtbevölkerung einher.

Der Mittelwert der Versiegelung hat dahingegen wenig Aussagekraft. Es lässt sich kein Bezug zum Faktor K oder zu der Anzahl der Zellen herstellen (Abb. 12).

5.2. Auswertung der Ergebnisse auf Wahlbezirksebene

Aus dem Ergebnisraster (Abb. 11) konnte für jeden einzelnen Wahlbezirk die Population berechnet werden. Durch diesen Wert konnten durch entsprechende Auswertungen weiterhin folgende Werte abgeleitet werden:

- Abweichungen der gegebenen und der berechneten Bevölkerung
    - absolute Abweichung
    - der relative Fehler

Die berechnete Einwohnerdichte wurde als Gleitkommazahl gespeichert. Wäre die Einwohnerdichte pro berechneter Zelle auf ganze Zahlen gerundet worden, hätte dies auf Grund der vielen Rasterzellen zwangsläufig zu einer hohen Unter- oder Überschätzung geführt. Im Folgenden sollen die Ergebnisse der Abweichungen dargestellt werden.

Die absolute Abweichung ist zum Teil bis auf eine Person genau gelungen. Das Minimum wurde mit -1.965 Personen und das Maximum mit 5.042 Personen ermittelt. Die wichtigen statistischen Kenngrößen des relativen Fehlers zeigt die folgende Tabelle 6.

Tab. 6 Die wichtigsten statistischen Kenngrößen des relativen Fehlers

Statistische Kenngröße	Wert
Minimum	-97,9
Maximum	1.181,0
Mittelwert	7,1
Median	-11,5
Standardabweichung	72,6
Interquartilsabstand	66,2

Das folgende Histogramm zeigt die Verteilung des relativen Fehlers auf die Wahlbezirke (Abb. 13).

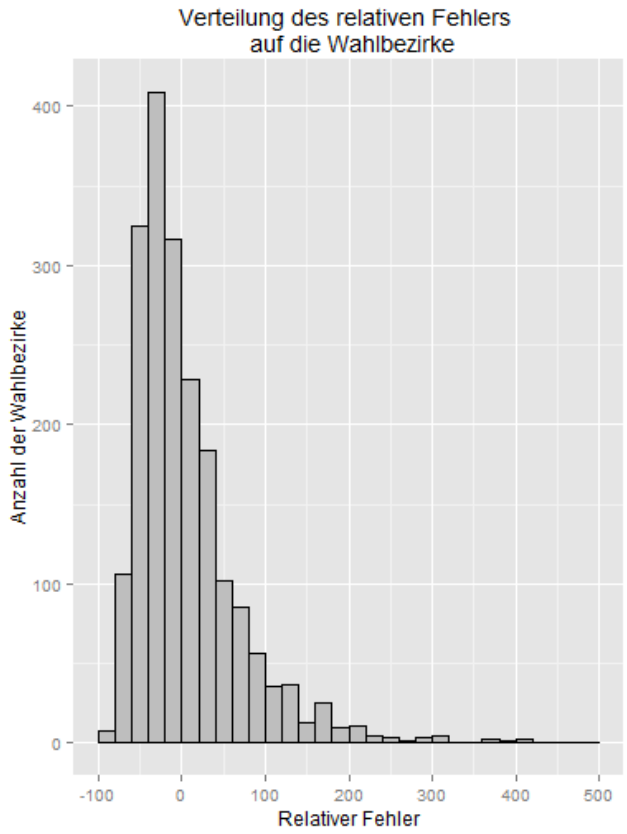

Abb. 13 Verteilung des relativen Fehlers auf die Wahlbezirke

Die Anzahl der Wahlbezirke mit einem relativen Fehler größer als 500% ist so minimal, dass die Darstellung eingegrenzt wurde. Nur 8% der relativen Fehler liegen außerhalb des Intervalls von -100% bis 100%. Der relative Fehler wurde kategorisiert. Anhand der Kategorisierung wurde die Verteilung des relativen Fehlers im Regierungsbezirk Münster visualisiert (Abb. 14).

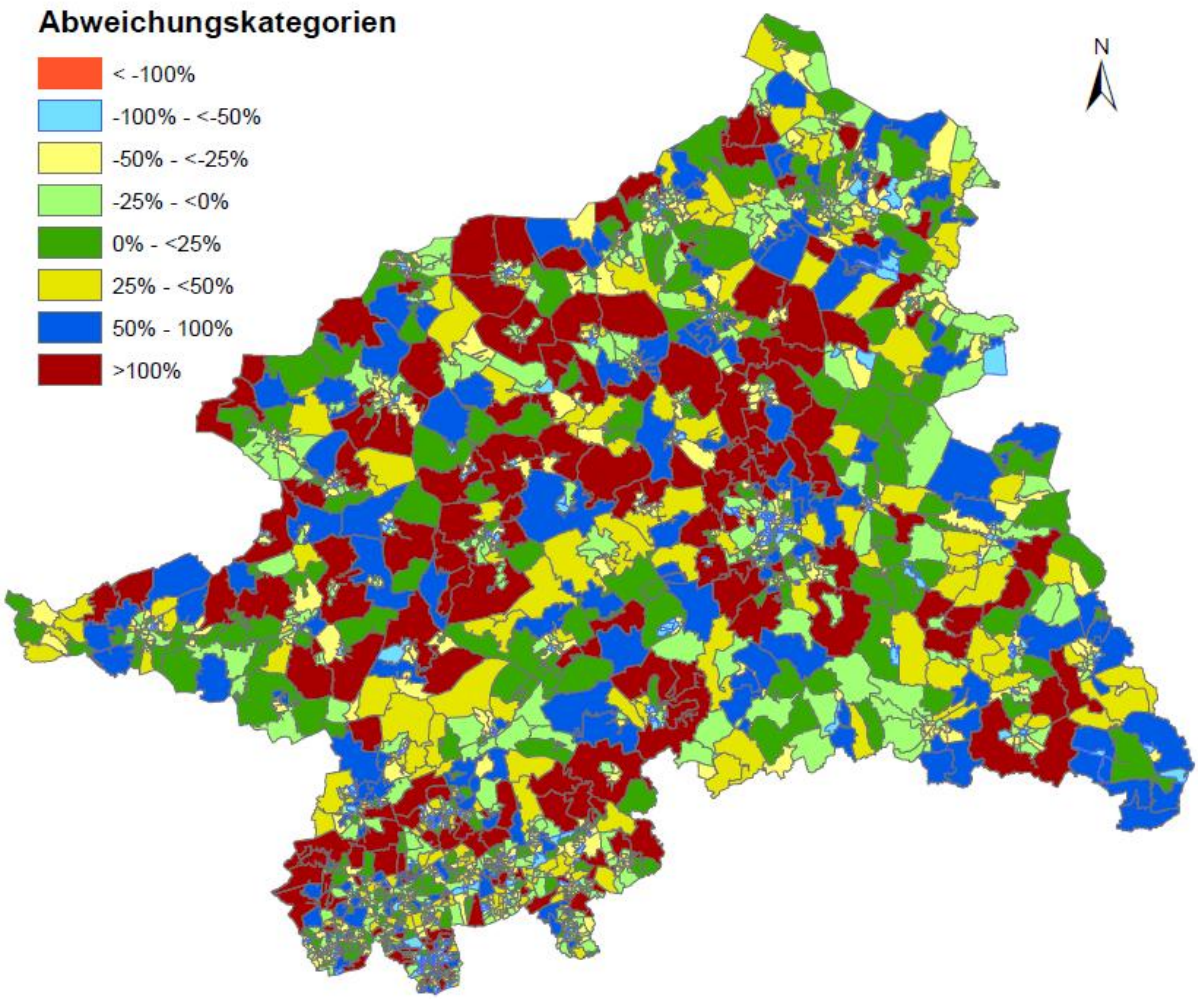

Abb. 14 Verteilung des relativen Fehlers in Kategorien als Kartendarstellung.

Die Farben wurden so gewählt, dass die absolute Abweichung durch einen Farbton (grün - gelb - blau - rot) ersichtlich wird (Abb. 14). Durch die Helligkeit der Farbe ist ersichtlich, ob die Abweichung positiv (dunkel) oder negativ (hell) ist.

Die Fehlerkategorie >100% hat einen großen Anteil an der Fläch, tatsächlich sind zahlenmäßig jedoch nur wenige Wahlbezirke betroffen (Abb. 14). Insgesamt ist der Fehler nicht einseitig verteilt sondern über das gesamte Untersuchungsgebiet gestreut. In nahezu jeder Gemeinde lassen sich Wahlbezirke mit Über- und Unterschätzungen finden.

Weiterhin wurde auch die Einwohnerdichte der Wahlbezirksebene in vier Kategorien von ländlich bis großstädtisch unterteilt. Diese orientieren sich an den Quantilen (25%, 50% und 75%) und haben daher in etwa gleiche Einwohnerzahlen. Auf die Fehlerkategorien sind die Einwohnerzahlen unregelmäßiger verteilt. Zu jeder Fehlerkategorie kann man ablesen wie sie sich auf die Kategorien von ländlich bis großstädtisch verteilt und umgekehrt. Folgende Tabelle zeigt die Ergebnisse (Tab. 7).

Tab. 7 Verteilungsmatrix der Populationsdichte und der relativen Fehler in Kategorien

Kategorie	ländlich (<= 926)	kleinstädtisch (927 - 1247)	städtisch (1248 - 1658)	großstädtisch (> 1658)	Summe	Anteil (in %)
< -100%	0	0	0	0	0	0,00
-100% - -50%	58	69	70	66	263	13,2
-50% - -25%	85	103	138	147	473	23,8
-25% - 0%	84	126	105	113	428	21,6
0% - 25%	69	68	64	82	283	14,3
25% - 50%	47	54	48	45	194	9,8
50% - 100%	63	38	47	30	178	9,0
>100%	90	38	23	13	164	8,3
Summe	496	496	495	496	1983	
Anteil in %	25,0	25,0	25,0	25,0		100,0

Die Summen der Kategorien lassen sich sowohl horizontal (ländlich - großstädtisch) als auch vertikal (relativer Fehler) auf die 1.983 Wahlbezirke aufsummieren. Aus der Tabelle lässt sich also ablesen wie sich der relative Fehler insgesamt verteilt.

Teilt man die Tabelle gedanklich jeweils horizontal und vertikal in der Mitte, so lässt sich über die Verteilung der Werte sagen, dass Gebiete mit einer niedrigen Einwohnerzahl tendenziell unterschätzt werden und Gebiete mit einer hohen Einwohnerzahl tendenziell überschätzt werden (Tab. 7).

Ohne Berücksichtigung, ob der relative Fehler positiv oder negativ ist, lässt sich zunächst sagen, dass mit ca. 36% mehr als ein Drittel der relativen Fehler in der erstrebenswerten Kategorie 0%-25% liegen. Damit liegt in 64% der Wahlbezirke ein relativer Fehler von mehr als 25% vor. Ca. 30% der Schätzungen liegen mit einem relativen Fehler von mehr als 50% in einem nicht mehr akzeptablen Bereich. Davon weisen 8 % sogar eine Abweichung von mehr als 100% zur tatsächlichen Bevölkerung auf.

Berücksichtigt man zudem ob sich der relative Fehler positiv oder negativ auswirkt, kann man sagen, dass es keine Unterschätzung der Bevölkerung von mehr als 100% gibt. Eine Überschätzung der Bevölkerung von mehr als 100% liegt hingegen mit 8% vor. In allen anderen Kategorien, aber vor allem in dem Bereich von 0-50%, liegt eine tendenzielle Unterschätzung vor.

Folgende Plots zeigen die Verteilung des relativen Fehlers auf die Einwohnerdichte pro km² (Abb. 15).

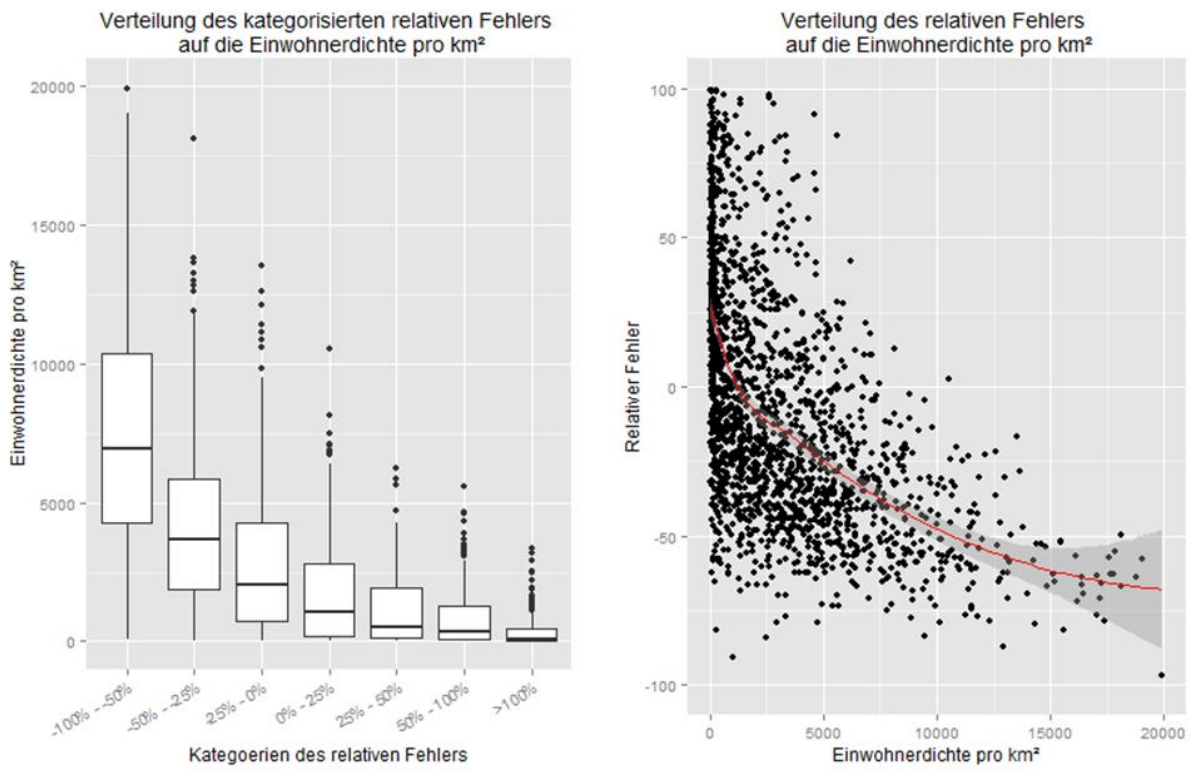

Abb. 15 Verteilung des relativen Fehlers auf die Einwohnerdichte

Der relative Fehler ist in dem Scatterplot (Abb. 15) im Bereich von -100 bis 100% eingegrenzt, da 92% der Werte in diesen Bereich liegen (Tab. 7). Es ist schnell zu erkennen dass der relative Fehler, absolut betrachtet, tendenziell mit der Einwohnerdichte wächst. Dies verdeutlicht auch die rot eingezeichnete Trendlinie (Loess-Glättung). Da jedoch sehr viele Punkte sehr eng beieinander liegen, ist es schwer die genaue Verteilung einzuschätzen.

Der Boxplot (Abb. 15) ist besser geeignet um das Verhältnis zwischen der Einwohnerdichte pro km² und dem relativen Fehler geordneter zu betrachten. In dem Boxplot ist die Streuung der Kategorien sehr gut zu erkennen. Außerdem lässt sich besser erkennen, dass Gebiete mit einer hohen Einwohnerdichte tendenziell unterschätzt werden. Gebiete mit einer niedrigen Einwohnerdichte werden dahingegen überschätzt. Dies lässt sich auch sehr gut aus den Medianen ablesen, welche in allen Boxen zum 25%-Quantil tendieren.

6. Diskussion und Ausblick

Die wichtigsten Erkenntnisse aus dem Auswertungskapitel sollen zunächst noch einmal zusammengefasst werden:

- Gebiete mit einer niedrigen Einwohnerzahl werden tendenziell unterschätzt wohingegen Gebiete mit einer hohen Einwohnerzahl überschätzt werden.
- Der Großteil (92%) der relativen Fehler liegt in dem Intervall von -100% bis 100%. Außerhalb des Intervalls gibt es nur Überschätzungen und keine Unterschätzungen. In allen anderen Kategorien, aber vor allem in dem Bereich der absoluten Abweichung von 0-50%, liegt hingegen eine tendenzielle Unterschätzung vor.
- Der relative Fehler ist räumlich über das ganze Gemeindegebiet Verteilt: In nahezu jeder Gemeinde lassen sich Wahlbezirke mit Über- und Unterschätzungen finden.
- 36% der Schätzungen haben eine maximale Abweichung von 25% zum Referenzwert.
- Es liegt eine tendenzielle Abhängigkeit zwischen dem Faktor K und der Anzahl der Zellen mit einer Wohnnutzung in einer Gemeinde vor. Je mehr versiegelte Flächen in einer Gemeinde vorhanden sind, desto größer ist der Faktor K.

Im Folgenden soll der angewendete Ansatz von Steinnocher (Steinnocher et al., 2011) mit einem Ansatz von Gallego (Gallego et al., 2001) verglichen werden und die Vor- und Nachteile der Ansätze diskutiert werden. Abschließend sollen mögliche Verbesserungen des verwendeten Ansatzes aufgezeigt werden.

6.1. Vergleich der Ergebnisse mit dem Ansatz von Gallego

Steinnocher kommt mit seinem Ansatz (Steinnocher et al., 2011) zu dem Schluss, dass die Disaggregationsergebnisse besser sind als die Ergebnisse anderer Produkte, aber dennoch systematische Fehler aufweisen. Diese Fehler zeichnen sich dadurch aus, dass es eine generelle Tendenz gibt, die geringer bevölkerten Zellen zu überschätzen und die höher bevölkerten Zellen zu unterschätzen. Dies begründet er zum einen damit, dass die lineare Beziehung zwischen Versiegelungsgrad und Bebauungsdichte nicht immer gültig ist. Insbesondere gilt das für urbane Zentren, in denen die Bevölkerung aufgrund fehlender Informationen über die Gebäudehöhen unterschätzt wird. Weiterhin begründet er dies damit, dass die Elimination von Flächen, die keiner Wohnnutzung unterliegen, nicht vollständig möglich ist und somit fälschlicherweise auch Verkehrs-, Industrie- und Gewerbeflächen Bevölkerung zugewiesen wird.

In dieser Arbeit können die gleichen Rückschlüsse gezogen werden. In der letzten Abbildung des Auswertungskapitels (Abb. 15) ist sehr gut zu erkennen, dass Gebiete mit einer hohen Einwohnerdichte auch in dieser Arbeit tendenziell unterschätzt werden, wohingegen Gebiete

mit einer niedrigeren Einwohnerdichte tendenziell überschätzt werden. Dies untermauert die Behauptung von Steinnocher (Steinnocher et al., 2011).

Weiterhin kommt es zu einigen Fehleinschätzungen bei der Ausmaskierung, welche sich nicht so leicht verhindern lassen. In Gebieten, die als Feld- oder Waldflächen ausgewiesen sind, können beispielsweise auch versiegelte Flächen für Lagerstätten vorhanden sein. Eine wichtige Aussage in diesem Zusammenhang ist, dass eine Wohnbebauung natürlich immer mit einer entsprechenden Versiegelung zusammenhängt. Umgekehrt kann aber nicht aus einer versiegelten Fläche automatisch auf eine Wohnbebauung geschlossen werden. Da die Ausmaskierung von Versiegelungsflächen im Detail nicht präzise genug ist, führt dies natürlich zu Fehlern.

Wie die Auswertung zeigt, wird die Bevölkerung in den städtischen Bereichen nur unterschätzt. Bei der Ausmaskierung der Straßen kommt es zu Fehlern, da die Straßen nicht überall vollständig abgebildet sind und die Lagegenauigkeit nicht überall perfekt ist. Hierdurch werden folglich nicht alle Straßenteile ausmaskiert und diesen Zellen wird im Zuge der Disaggregation eine Bevölkerung zugewiesen. Außerdem kann es vorkommen, dass einzelne Wohnbebauungen wie beispielsweise Bauernhöfe, nicht in dem Versiegelungsdatensatz enthalten sind. Dies führt wiederum zu Unterschätzungen.

Leichte Ungenauigkeiten bei der Auswertung können weiterhin daher entstanden sein, dass sich die Datensätze auf verschiedene Jahre beziehen. Während sich der „CORINE Land Cover“- und der „EEA Fast Track Service Precursor on Land Monitoring“-Datensatz auf das Jahr 2006 beziehen, liegen die Bevölkerungszahlen aus dem Jahr 2005 vor. Weder Bevölkerungszahlen noch die Flächennutzungen und Bebauungen ändern sich jedoch so rasant, als das die dadurch entstandenen Fehler kritisch sein könnten.

Um einen Vergleich mit Gallego ziehen zu können, muss der Ansatz von Gallego (Gallego et al., 2001) zunächst in seinen wesentlichen Zügen erläutert werden. Als Datengrundlage dienen auch für diesen Ansatz Bevölkerungsdaten und der CORINE Land Cover Datensatz. Formal beschrieben wird der Ansatz durch folgende zwei Gleichungen:

$$(1) \qquad Y_{cm} = U_{ch} * W_m$$

$$(2) \qquad W_m = \frac{X_m}{\sum_c S_{cm} * U_{ch}}$$

Folgende Tabelle beschreibt die Bedeutung der Variablen (Tab. 8).

Tab. 8 Bedeutung der Variablen in dem formalisiertem Ansatz von Gallego

Variable	Bedeutung
Y_{cm}	Einwohnerdichte in Abhängigkeit von der Landnutzungsklasse *c* und der Gemeinde *m*
U_{ch}	Koeffizient einer Landnutzungsklasse *c* in Abhängigkeit der Bevölkerungsdichte *h* [dicht, weniger dicht, dünn besiedelt] (Dieser wird iterativ durch Annäherung bestimmt indem die Differenz der tatsächlichen zur berechneten Bevölkerungsdichte minimiert wird)
W_m	Korrekturfaktor um sicher zu stellen, dass die berechnete Gesamtbevölkerung einer Gemeinde *m* gleich dem gegebenem Bevölkerungswert dieser Gemeinde *m* ist
X_m	Gegebener Bevölkerungswert einer Gemeinde *m*
S_{cm}	Fläche der Landnutzungsklasse *c* in Gemeinde *m*

Das Modell setzt die folgenden zwei Annahmen voraus:

(a) Die Einwohnerdichte ist für alle Pixel innerhalb der gleichen CLC-Klasse einer Gemeinde immer gleich.

(b) Das Verhältnis zwischen der Einwohnerdichte von zwei Landnutzungsklassen ist in allen Gemeinden gleich.

Im Wesentlichen ist die Bevölkerungsschätzung nach dem Ansatz von Gallego also abhängig von der Einstufung bezogen auf die Bevölkerungsdichte und der entsprechenden Landnutzungsklasse. Berücksichtigt werden von dem CORINE Land Cover Datensatz 26 Landnutzungsklassen, welche wiederum in 16 Gruppen eingeteilt werden (Gallego et al., 2001). Auf den anderen Landnutzungsklassifikationen wird eine Wohnnutzung ausgeschlossen. Dieses Ausschlusskriterium wurde also wie auch in dieser Arbeit berücksichtigt.

Gallego kommt mit seinem Ansatz (Gallego, 2010) zu dem Schluss, dass sich die erzielten Ergebnisse bisheriger Anwendungen den tatsächlichen Bevölkerungsdaten annähern aber noch weit von einer guten Lösung entfernt sind. Weiterhin werden Einwohnerzahlen für nicht urbane Landnutzungsklassen mit seinem Ansatz allgemein überschätzt. Auch bemängelt er, dass es zu fehlerhaften Landnutzungsklassifikationen kommt und begründet dies mit der minimalen Kartiereinheit („MMU“) von 25ha. Außerdem bemängelt Gallego, dass die Qualität der disaggregierten Werte für Gebiete schlechter ausfallen, in denen die Größe der Kommunen sehr heterogen ist. Dies begründet er zum Teil dadurch, dass die CORINE Daten ursprünglich Vektordaten waren und die Umwandlung in grobe Rasterdaten insbesondere auf kleinere Gemeinden einen negativen Einfluss haben kann.

Die Situation vor Ort bleibt bei Gallegos Ansatz zu wenig berücksichtigt. Zellen aus der gleichen Gemeinde und der gleichen Klassifikation haben sehr ähnliche Einwohnerzahlen, obwohl vor Ort ganz unterschiedliche Eigenschaften vorherrschen. Dies mag statistisch für eine große Fläche richtig sein, aber verzerrt das Bild einzelner Zellen beträchtlich.

Mit seinen 100x100 Meter Rasterzellen ist der CORINE Land Cover Datensatz sehr grob. Problematisch ist auch, dass bei mehreren vorhandenen Landnutzungen nur eine deutlich überwiegende Landnutzungsklasse oder eine Misch-Landnutzungsklasse gewählt wird. Dies ist auf die minimale Kartiereinheit („MMU“) von 25ha zurückzuführen. Dadurch wird das Ergebnis insgesamt natürlich verzerrt. Wenn beispielsweise ein Wohngebiet zusammen mit einer anderen, flächenmäßig überwiegenden Landnutzung in eine Zelle fällt, ist es möglich, dass die Klassifikation nicht auf Wohnnutzung schließen lässt.

Der CORINE Land Cover Datensatz bringt eine tendenzielle Unterschätzung mit sich. Dies kann zum einen damit begründet werden, dass einzelne Wohnbebauungen in ländlich geprägten Gegenden zu wenig Fläche einnehmen und daher für die Klassifizierung zu wenig Gewicht haben. Das hat zur Folge, dass Zellen eine ländliche Klassifikation erhalten und entsprechend gering gewichtet werden. Zum anderen unterscheidet der Datensatz bei der städtischen Klassifikation nicht die Art der Wohnbebauung. Es fehlen Informationen darüber, ob es sich beispielsweise um Einzelhäuser, dichte Reihenhäuser oder sogar hohe Wohnkomplexe handelt. Diese Informationen hätten natürlich großen Einfluss auf die Schätzung der Einwohnerdichte, da sie unterschiedlich gewichtet werden müssten.

Auch wenn der Ansatz von Steinnocher noch lange kein perfektes Ergebnis liefert, so ist er verglichen mit dem Ansatz von Gallego wohl die bessere Alternative, da er mit einem großen Vorteil einhergeht: Die Auflösung des „EEA Fast Track Service Precursor on Land Monitoring“-Datensatzes ist mit seinen 20x20 Meter Zellen deutlich höher als der CORINE Land Cover Datensatz, was mit einer deutlich höheren Genauigkeit verbunden ist. Aufgrund der höheren räumlichen Auflösung ist der Ansatz von Steinnocher für die Disaggregation auf Wahlbezirksebene besser geeignet. Der Ansatz von Gallego ist für diese Aufgabe dahingegen nicht so gut geeignet, da er für größere Flächen ausgelegt ist und die Qualität der disaggregierten Werte unter den heterogenen Wahlbezirksgrößen leidet.

Im Folgenden sollen Überlegungen beschrieben werden, die den verwendeten Ansatz weiter verbessern und zuverlässigere Ergebnisse bringen können.

6.2. Mögliche Verbesserungen des genutzten Verfahrens

In der Auswertung wurde sichtbar, dass nur ca. 36% der Daten um bis zu 25% von den tatsächlichen Werten abweichen. Es stellt sich also die Frage wie man das Verfahren weiter verbessern kann um bessere Ergebnisse zu erhalten.

Schon bei der Erstellung des Ausgangsrasters wird klar, dass insbesondere der CORINE Land Cover Datensatz, durch die niedrige räumliche Auflösung sehr grob ist. Der CORINE-Land Cover Datensatz maskiert die tatsächliche Bebauung ohne Wohnfunktion mit 100x100 Meter Zellen natürlich verhältnismäßig ungenau aus. Könnten diese Ungenauigkeiten verhindert werden, würden natürlich entsprechend bessere Ergebnisse zu erwarten sein. Der Zugriff auf genauere Flächennutzungsdaten würde zu einer großen Verbesserung führen. Die Genauigkeit kann durch höher auflösende Daten gesteigert werden.

Eine Möglichkeit wäre die Nutzung von Flächennutzungsplänen, aus denen die entsprechenden Flächennutzungen viel genauer hervorgehen. Diese Daten müssten aber von jeder Gemeinde einzeln angefordert werden. Zudem sind die Daten nicht mit Standardprogrammen zu verarbeiten und müssten zunächst entsprechend angepasst werden.

Eine weitere Möglichkeit wäre der Gebrauch von Katasterdaten. Deutschlandweit wird auf das System ALKIS (amtliches Liegenschaftskataster Informationssystem) umgestellt, welches das zuvor genutzte System aus ALK (amtliche Liegenschaftskarte) und ALB (amtliches Liegenschaftsbuch) zusammenführt. In Nordrhein Westfalen ist dieses System bereits eingeführt.

Nach Auskunft einer Mitarbeiterin von Geobasis NRW können Objekte tatsächlicher Nutzung gebildet werden, die eine zusammenhängende Fläche bilden und einheitliche Nutzungseigenschaften aufweisen (GeobasisNRW, 2014). Dort gibt es einen Objektkartenbereich „Tatsächliche Nutzung“, welcher folgende Objektkartengruppen enthält:

- Gewässer
- Siedlung
- Vegetation
- Verkehr

Die Objektkartengruppe mit der Bezeichnung „Siedlung“ beinhaltet die bebauten und nicht bebauten Flächen, die durch die Ansiedlung von Menschen geprägt werden oder zur Ansiedlung beitragen.

Die Objektkartengruppe umfasst wiederum folgende Objektarten:

- Wohnbaufläche
- Industrie- und Gewerbefläche
- Halde
- Bergbaubetrieb
- Fläche gemischter Nutzung
- Sport-, Freizeit- und Erholungsfläche
- Friedhof

Die Flurstücke sind flächendeckend für NRW verfügbar. Die bereitgestellten Daten werden aus dem Sekundärdatenbestand beim Geodatenzentrum selektiert, welcher in der Regel einmal jährlich aktualisiert wird. Die verfügbaren Daten haben eine Aktualität von Dezember 2013 bis April 2014 und sind im NAS Format verfügbar. Für wissenschaftliche Zwecke kann unter bestimmten Bedingungen eine gebührenfreie Nutzung ermöglicht werden.

Dieser Datensatz würde eine vorherige Ausmaskierung von Flächen unnötig machen. Die Wohnbaufläche des Datensatzes könnte zur Ausmaskierung der Versiegelungsflächen genutzt werden und der Gebrauch der groben CORINE-Daten wäre unnötig. Gröbere Fehler im Zuge der Ausmaskierung werden damit ausgeschlossen. Das daraus resultierende Raster würde die Versiegelung der tatsächlichen Wohnbebauungen widerspiegeln.

Weiterhin werden allerdings die Informationen über die Höhe von Gebäuden fehlen. Hierdurch werden auch weiterhin systematische Fehler auftreten, da Mehrfamilienhäuser tendenziell unterschätzt und einfache Wohnhäuser überschätzt werden. Durch den Gebrauch von Software wie Google Earth kann die Höhe von Gebäuden annähernd eingeschätzt werden. Dies ist allerdings mit einem enormen Arbeitsaufwand verbunden.

Prinzipiell könnten auch Plattformen geschaffen werden, auf denen die Öffentlichkeit bei der Datenbeschaffung hilft. Vorstellbar wäre beispielsweise, dass Daten, wie die Anzahl der Geschosse von Wohngebäuden, von der Öffentlichkeit angegeben werden können. So könnte die Information über die Anzahl der Geschosse eines Hauses flächendeckend gesammelt werden. Allerdings würde ein solcher Ansatz, der auf freiwillige Mithilfe angewiesen ist, vermutlich zu lückenhaft bleiben, um ihn sinnvoll nutzen zu können.

In Zukunft wird es mit großer Wahrscheinlichkeit noch höher auflösende Rasterdatensätze als den CORINE Land Cover Datensatz und den EEA Fast Track Service Precursor on Land Monitoring geben. Dies wird eine bessere Klassifizierung ermöglichen und die Datenqualität weiter verbessern.

Literaturverzeichnis

Eicher, Cory / Brewer, Cynthia (2001):
Dasymetric mapping and areal interpolation: Implementation and evaluation.
In: Cartography and Geographic Information Science 28 (2): S. 125-138.

Gallego, Francisco (2010):
A population density grid of the European Union.
In: Population and Environment, 31, S. 460-473.

Gallego, Javier / Peedell, Steve (2001):
Using CORINE Land Cover to map population density.
In: Towards Agri-environmental indicators, Topic report 6/2001 European Environment Agency, Copenhagen, S. 92-103.

Krunic, Nikola / Bajat, Branislav / Kilibarda, Milan / Tosic, Dragutin (2011):
Modelling the spatial distribution of Vojvodina's population by using dasymetric mapping..
In: SPATIUM (24): S. 45-50.

Lemke, Dorothea / Mattauch, Volkmar / Heidinger, Oliver / Pebesma, Edzer / Hense, Hans-Werner (2013):
Detecting cancer clusters in a regional population with local cluster tests and Bayesian smoothing methods: a simulation study.
In: International Journal of Health Geographics 2013, 12 (54).

Maantay, Juliana / Maroko, Andrew / Herrmann, Christopher (2007):
Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS).
In: Cartography and Geographic Information Science, 34 (2): S. 77-102.

Mennis, Jeremy (2003):
Generating Surface Models of Population Using Dasymetric Mapping*.
In: The Professional Geographer, 55 (1), S. 31-42.

Steinnocher, Klaus / Köstl, Mario / Weichselbaum, Jürgen (2011):
Kleinräumige Bevölkerungsmodellierung für Europa – räumliche Disaggregation auf Basis des Versiegelungsgrades.
In: Angewandte Geoinformatik 2011.

Steinnocher, Klaus / Köstl, Mario / Weichselbaum, Jürgen (2006):
Linking remote sensing and demographic analysis in urbanised areas.
In: First Workshop of the EARSeL SIG on Urban Remote Sensing “Challenges and Solutions", March 2-3.

Thieken, A. / Müller, M. / Kleist, L. / Seifert, I. / Borst, D. / Werner, U. (2006):
Regionalisation of asset values for risk analyses.
In: Nat. Hazards Earth Syst. Sci., 6, S. 167-178.

Zandbergen, Paul / Chakraborty, Jayajit (2006):
Improving environmental exposure analysis using cumulative distribution functions and individual geocoding.
In: International Journal of Health Geographics 2006, 5 (23).

Onlinequellen:

EEA (2006): Europäische Umweltagentur - CORINE:
Online unter: http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3#tab-gis-data (abgerufen am: 07.05.2014).

EEA (2009): Europäische Umweltagentur - Fast Track Service Precursor on Land Monitoring - Degree of Soil sealing:
Online unter: http://www.eea.europa.eu/data-and-maps/data/eea-fast-track-service-precursor-on-land-monitoring-degree-of-soil-sealing/#tab-european-data (abgerufen am: 07.05.2014).

EEA (2009): Europäische Umweltagentur - Raster data on population density using Corine Land Cover 2000 inventory:
Online unter: http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2 (abgerufen am: 30.06.2014).

Bezirksregierung Köln: Abteilung 7 - Geobasis NRW (E-Mail-Verkehr mit Frau Michèle Schütte):
Online unter: http://www.bezreg-koeln.nrw.de/brk_internet/organisation/abteilung07/index.html (abgerufen am: 26.06.2014).

Geofabrik-Downloadserver - Daten-Auszüge aus dem OpenStreetMap-Projekt - Straßennetz RB Münster:
Online unter: http://download.geofabrik.de/europe/germany/nordrhein-westfalen/muenster-regbez.html (abgerufen am: 08.05.2014).

Information und Technik Nordrhein-Westfalen - Statistiken:
Online unter: http://www.it.nrw.de/statistik/a/index.html (abgerufen am: 06.05.2014).

Nexiga next level geomarketing:
Online unter: http://www.nexiga.com/startseite/ (abgerufen am: 08.05.2014).

R Website: The R Project for Statistical Computing:
Online unter: http://www.r-project.org/ (abgerufen am: 03.07.2014).

RStudio Website:
Online unter: http://www.rstudio.com/ (abgerufen am: 03.07.2014).

Anhang

Quellcode des R-Programms:

```
# @author: Lars Syfuß
# All the needed data paths
excel_path = "path_to_excel_file.xls"    # Excel file with polulation data
shp_path = "path_to_shapefile_folder"  # Path to the Shapefiles with community polygons (GKZ +
                                         WBZ)

regbez_path = "path_to_tif_file.tif"       # 20x20m sealingraster (tif) of RegBez Muenster

result_raster_path = "path_to_tif_file.tif"  # Raster with population data (previously created with
                                               start_disaggregation-function and merged in ArcGis)

raster_mask_path = "path_to_folder"     # The path to the folder where the function
                                         maskFrom_RegBezRaster_saveFile stores the result files
raster_pop_path = "path_to_folder"       # The path to the folder where the function create_popRaster
                                         stores the result files

# Excel sheets with population data
gkz_2005 = "GKZ_EW0_65plus2005"     # Population data (0-65) of communities (2005)
wbz_2005 ="WBZ_EW_0_65plus2005"  # Population data (0-65) of residential districts (2005)

result_gkz_path = "path_to_gkz_results.txt"      # Results of the GKZ calculation are stored in this txt
result_wbz_path = "path_to_wbz_results.txt"    # Results of the WBZ calculation are stored in this txt

# Loading the needed packages
library(rgdal)              # To work with shapefiles
library(gdata)              # To work with Excel files
library(raster)             # To work with raster files

### Work with a shapefiles ##################################################################
# Load the shapefile (GKZ)
communities_shape = readOGR(dsn = shp_path, layer = "GKZ")
# Load shapefile (WBZ)
residentialDistricts_shape = readOGR(dsn = shp_path, layer = "WBZ")

# Create the shape for the required community
create_gkzShape = function(gkz){
        print("creating the gkz shapefile...")
        gkz_shape = communities_shape[communities_shape$GKZ==gkz,]
        gkz_shape
}

# Create the shape for the required residential district
create_wbzShape = function(kgs22){
        print("creating the wbz shapefile...")
        wbz_shape = residentialDistricts_shape[residentialDistricts_shape$KGS22==kgs22,]
        wbz_shape
}

### Work with Excel data ####################################################################
# Load the file with needed sheet
pop_excel = read.xls(excel_path, sheet = wbz_2005)
```

```
# Calculate the population sum in a residential district (KGS22)
calc_pop_wbz = function (param_KGS22){
	print("calculating the population sum in the residential district...")
	# Load the Excel file with needed sheet (wbz_2005)
	pop_wbz = read.xls(excel_path, sheet = wbz_2005)
	# get the community data
	comm_data = pop_wbz[pop_wbz$KGS22==param_KGS22,]
	 # sum male population
	pop_m = sum(comm_data$X0_65m)
	# sum female population
	 pop_f = sum(comm_data$X0_65w)
	# calculate population sum
	 pop_sum = pop_m + pop_f
	 print(paste("population sum is: ", pop_sum))
	 pop_sum
}

# Calculate female, male, or complete population in GKZ
# Works for 2005 sheet! Accepts "male", "female" or "complete" as gender.
calc_pop_gkz_2005 = function (param_KGS8, gender){
	if(gender =="male" || gender=="female" || gender =="complete"){
		print(paste("calculating the ", gender, " population in the community..."))
		# Load the Excel file with needed sheet (gkz_2005)
		pop_gkz = read.xls(excel_path, sheet = gkz_2005)
		comm_data = pop_gkz[pop_gkz$KGS8==param_KGS8,]
		# Sum male population
		pop_m = sum(comm_data$SUM_0_65m)
		# Sum female population
		pop_f = sum(comm_data$SUM_0_65w)
		# Calculate population
		if(gender=="male"){
			pop=pop_m
		}
		if(gender=="female"){
			pop=pop_f
		}
		if(gender=="complete"){
			pop= pop_m + pop_f
		}
		print(paste("population (", gender, ") is: ", pop))
		pop
	}
	else{
		# Error message
		stop("Attribute gender (in function calc_pop_gkz_2005) is not valid")
	}
}

# Get the name of a community by the gkz number
get_communityName = function(param_KGS8){
	print("Getting the community name...")
	# Load the Excel file with needed sheet (gkz2005)
	pop_gkz = read.xls(excel_path, sheet = gkz_2005)
	comm_data = pop_gkz[pop_gkz$KGS8==param_KGS8,]
	# Community name
```

```
        print(paste("community name: ", comm_data$G_Name))
        comm_data$G_Name
}

### Work with raster data ##################################################################

# Load the RegBez raster
regbez_raster = raster(regbez_path)

# Make sure that all needed files use the same projection
#projection(regbez_raster)
#projection(communities_shape)
#projection(residentialDistricts_shape)

# Mask out a polygon (shapefile) from the RegBez raster and save that file.
# Name must be name.tif
maskFrom_RegBezRaster_saveFile = function(shape, name){
        print("creating the masklayer...")
        f_name = paste(raster_mask_path, name)
        # Use the crop function first to have a smaller raster
        regbez_crop = crop(regbez_raster, shape)
        mask(regbez_crop, shape, filename=f_name, overwrite=TRUE)
        # Return the filename for further work
        f_name
}

# Mask out a polygon (shapefile) from the given raster
mask_fromRaster = function(shape, raster){
        print("creating the masklayer...")
        # crop first to have a smaller raster
        raster_crop = crop(raster, shape)
        mask(raster_crop, shape)
}

# Calculate sum of raster cells
calc_rasterSum = function(raster){
        print("calculating the sum of all raster values...")
        rasterSum = sum(getValues(raster),na.rm=TRUE)
        print(paste("the sum of all raster values is: ", rasterSum))
        rasterSum
}

# Count raster cells that have a value (not NA)
count_valueCells = function(raster){
        print("counting the number of raster cells (not NA)...")
        length(Which(!is.na(raster), cells=TRUE))
}


# Calculate the mean sealing for a raster
calc_rasterMean = function(raster){
        print("calculating the mean of all raster values...")
        rasterMean = mean(getValues(raster),na.rm=TRUE)
        print(paste("the mean of all raster values is: ", rasterMean))
        rasterMean
}
```

```
# Calculate how often which cell-value occurs
calc_valueFrequency = function(raster){
	print("calculating the frequency of all raster values...")
	res_freq = freq(raster, digits=4)
	res_freq
}

# Calculate factor K
calc_k = function(pop, sealing){
	print("calculating the factor k...")
	k = pop/sealing
	print(paste("factor k is: ", k))
	k
}

# Calculate new raster with population data for each cell. The name must be name.tif!
create_popRaster = function(k, raster, name){
	print("creating the population raster...")
	raster = raster*k
	f_name = paste(raster_pop_path, name)
	writeRaster(raster, filename=f_name, overwrite=TRUE)
	# Return the filename for further work
	f_name
}

# Start the disaggregation and create raster for each community with its population
# The raster_path links to the raster file with all the communities
# The shapefile delivers the polygons
# The Excel file delivers the population data
# The gender can be "male", "female" or "complete"
# The save_mask attribute is a boolean to choose wether to save the mask files or not
start_disaggregation = function(raster_path, shape, gender, save_mask){
	rb_raster = raster(raster_path)
	# Store the variables gkz (number), gkz name and factor k for each loop
	res_gkz=c()
	res_gkz_name=c()
	res_k=c()
	res_gkz_value_cells=c()
	res_gkz_mean_sealing=c()
	# Build in a counter to see the program status
	i = 1
	# For each community (inside the shapefile)...
	for(gkz in communities_shape$GKZ){
		print(paste("Running disaggregation in community number: ", i))
		# Calculate the population
		gkz_pop = calc_pop_gkz_2005(gkz, gender)
		# Create the shape
		gkz_shape = create_gkzShape(gkz)
		# Get the name of the gkz
		gkz_name = paste(get_communityName(gkz), ".tif")
		# Mask out the raster
		if(save_mask==TRUE){
		  gkz_raster_mask_path = maskFrom_RegBezRaster_saveFile(gkz_shape, gkz_name)
		  gkz_raster_mask = raster(gkz_raster_mask_path)
		} else {
```

```
                    gkz_raster_mask = mask_fromRaster(gkz_shape, rb_raster)
                }
                # Count the raster cells that are not NA (have value)
                gkz_value_cells = count_valueCells(gkz_raster_mask)
                # Calculate the sum of the community sealing
                gkz_sealing = calc_rasterSum(gkz_raster_mask)
                # Calculate the mean of the community sealing
                gkz_mean_sealing = calc_rasterMean(gkz_raster_mask)
                # Calculate the factor K
                gkz_k = calc_k(gkz_pop, gkz_sealing)
                # Create the new raster with population values per cell
                create_popRaster(gkz_k, gkz_raster_mask, gkz_name)
                #Store the results
                res_gkz=c(res_gkz, gkz)
                res_gkz_name=c(res_gkz_name, gkz_name)
                res_k=c(res_k, gkz_k)
                res_gkz_value_cells=c(res_gkz_value_cells, gkz_value_cells)
                res_gkz_mean_sealing=c(res_gkz_mean_sealing, gkz_mean_sealing)
                # Increase the counter
                i=i+1
        }
        # Create a dataframe as a result and write it to the result textfile
        res = data.frame(KGS8=res_gkz, G_NAME=res_gkz_name, K=res_k,
         CELLS_COUNT=res_gkz_value_cells, MEAN_SEALING=res_gkz_mean_sealing)
        write.fwf(res, file=result_gkz_path, append=FALSE, sep=";")
        print("start_disaggregation run completed!")
}

# Run the disaggregation
start_disaggregation(regbez_raster, communities_shape, "complete", FALSE)

# Function to calculate the population in the residential districts (from the result raster)
# The raster stores the population data
calc_wbzPopulation = function(raster_path){
        pop_raster=raster(raster_path)
        # Store the variables wbz(number), and the calculated population for each loop
        res_wbz=c()
        res_pop=c()
        res_sealing=c()
        # Build in a counter to see the program status
        i = 1
        # For each community (inside the shapefile)...
        for(wbz in residentialDistricts_shape$KGS22){
                print(paste("Running calculation of population in wbz nr.: ", i))
                # calculate the population (from excel)
                wbz_pop = calc_pop_wbz(wbz)
                # Create the shape
                wbz_shape = create_wbzShape(wbz)
                # Create the wbz raster
                wbz_raster = mask_fromRaster(wbz_shape, pop_raster)
                # Calculate the wbz population from raster
                wbz_raster_pop=calc_rasterSum(wbz_raster)
                # Calculate mean sealing
                wbz_mean_sealing = calc_rasterMean(wbz_raster)
                #Store the results
                res_wbz=c(res_wbz, wbz)
```

```
                res_pop=c(res_pop, wbz_raster_pop)
                res_sealing=c(res_sealing, wbz_mean_sealing)
                # Increase the counter
                i=i+1
        }
        # Create a dataframe as a result and write it to the result textfile
        res = data.frame(KGS22=res_wbz, POPULATION_BERECHNET=res_pop,
         MEAN_SEALING=res_sealing)
        write.fwf(res, file=result_wbz_path, append=FALSE, sep=";")
        print("calc_wbzPopulation run completed!")
}

# Calculate the population and mean sealing per WBZ
calc_wbzPopulation(regbez_path)
```

Quellcode der R-Auswertung:

```
# @author: Lars Syfuß
# Analysis

# Loading the needed packages
library(ggplot2) # to work with scatterplots

# csv filepath
results_gkz_path = "path_to_GKZ_results.csv"    # GKZ results
results_wbz_path = "path_to_WBZ_results.csv"  # WBZ results

# Load csv files
results_gkz = read.csv(results_gkz_path, header=TRUE, sep=";")
results_wbz = read.csv(results_wbz_path, header=TRUE, sep=";")


# Access the data
results_k = results_gkz$K
results_sealing = results_gkz$MEAN_SEALING
results_cells = results_gkz$CELLS_COUNT

results_kategorie = results_wbz$Kategorie..Error.
results_pop_dens = results_wbz$Einwohner_km_orig
results_relative_error = results_wbz$relative.Error
results_pop_dens = results_wbz$Population..orig..

# mean, median, SD und IQR for relative error
mean(auswertung_relative_error)
median(auswertung_relative_error)
sd(auswertung_relative_error)
IQR(auswertung_relative_error)

# Calculate quantiles for population
quantile(results_pop_dens, c(0, 0.25, 0.5, 0.75, 1))

# SCATTERPLOT (relative error - population)
qplot(results_pop_dens, results_relative_error,
        ylim=c(-100,100),
```

```
          xlim=c(0,20000),
          main="Verteilung des relativen Fehlers \n auf die Einwohnerdichte pro km²",
           ylab="Relativer Fehler", xlab="Einwohnerdichte pro km²"
)+geom_smooth(method = "loess", colour = "red")

# SCATTERPLOT (number of cells - factor K)
qplot(auswertung_k, auswertung_cells,
     main="Verteilung des Faktors K \n auf die Anzahl der versiegelten Zellen",
     ylab="Anzahl der Zellen", xlab="Faktor K"
)+geom_smooth(method = "auto", colour = "red") # method local smooths

# SCATTERPLOT (number of cells - factor K - mean degree of soil sealing in colours)
qplot(auswertung_cells, auswertung_k,
     color = auswertung_sealing,
     #scale_colour_gradientn(colours = c("red","yellow","green"), values=auswertung_sealing),
     main="Verteilung des Faktors K \n auf die Anzahl der versiegelten Zellen \n und die Mittelwerte
               des Versiegelunsgrads",
     ylab="Faktor K", xlab="Anzahl der Zellen"
)+geom_smooth(method = "lm", colour = "red")+
  scale_colour_gradientn(colours = rainbow(7), name = "Mittelwert des Versiegelungsgrads")

# BOXPLOT (population distribution - category of the relative error)
qplot(results_kategorie, results_pop_dens,
          main="Verteilung des kategorisierten relativen Fehlers \n auf die Einwohnerdichte pro km²",
           ylab="Einwohnerdichte pro km²", xlab="Kategoerien des relativen Fehlers",
          ylim=c(0,20000),
          geom = "boxplot"
)+scale_x_discrete(limits=c("-100% - -50%", "-50% - -25%", "-25% - 0%", "0% - 25%",
                            "25% - 50%", "50% - 100%", ">100%"))+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

# HISTOGRAM relative error
qplot(results_relative_error,
          main="Verteilung des relativen Fehlers \n auf die Wahlbezirke",
          xlab="Relativer Fehler", ylab="Anzahl der Wahlbezirke",
)+geom_bar(colour="black", fill="grey")+
  scale_x_continuous(breaks=c(-100,0,100,200,300,400,500), limits=c(-100,500))

# HISTOGRAM category of relative error
qplot(results_kategorie,
          main="Verteilung des relativen Fehlers \n auf die Wahlbezirke",
          xlab="Relativer Fehler", ylab="Anzahl der Wahlbezirke",
)+scale_x_discrete(limits=c("-100% - -50%", "-50% - -25%", "-25% - 0%", "0% - 25%",
                            "25% - 50%", "50% - 100%", ">100%"))+
  geom_bar(colour="black", fill="grey")+theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

Plagiatserklärung des Studierenden

Hiermit versichere ich, dass die vorliegende Arbeit selbstständig verfasst worden ist, dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien – dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

_____

(Datum, Unterschrift)

Ich erkläre mich mit einem Abgleich der Arbeit mit anderen Texten zwecks Auffindung von Übereinstimmungen sowie mit einer zu diesem Zweck vorzunehmenden Speicherung der Arbeit in eine Datenbank einverstanden.

_____

(Datum, Unterschrift)