Datenqualitätssicherung für den IRB-Katalog

Güleş, Antje

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Datenqualitätssicherung für den IRB-Katalog

Antje Güleş

An der Datensammlung Innerstädtische Raumbeobachtung (IRB[1]) arbeiten jährlich mindestens 200 Personen, die unterschiedlich qualifiziert sind und unterschiedliche Anforderungen an die kleinräumige Statistik stellen. Bei 30 Tabellen und rund 400 Merkmalen in zurzeit 51 Städten der IRB-Kooperation bleibt es daher nicht aus, dass Fehler entstehen. Sei es beim Kopieren der Daten, beim Datenabruf selbst oder bei der Zusammenstellung der Daten innerhalb der Städte und im BBSR. Auch Neuzuschnitte in den Stadtteilen, gesetzliche Änderungen mit Auswirkungen auf die Melde- oder Sozialstatistik und unterschiedliche Formen der Anonymisierung führen in den Daten zu Inkohärenzen zwischen den einzelnen Städten sowie im zeitlichen Vergleich. In diesem Beitrag werden die Anforderungen an den Datenkatalog IRB und die Prüfung, Nachfrage, Korrektur und Dokumentation der Daten und Metadaten dargestellt.

1 Gütekriterien für die IRB

Zunächst zu den Kriterien der Güte amtlicher Statistik, wie sie sowohl in den Kommunen als auch im Bund angelegt werden. Im ersten Beitrag dieses Heftes (Gutfleisch/Sturm) werden die Kriterien bereits ausführlich diskutiert. An diesem orientiert gebe ich einen kurzen Überblick der Kriterien, die für die IRB zutreffen. Auf die wichtigsten wird in diesem Beitrag näher eingegangen.

Die *Relevanz* der Daten ergibt sich aus den Aufgaben der wissenschaftlichen Politikberatung. Wanderungsdaten, Daten zur Bevölkerungsstruktur und zur Beschäftigung auf kleinräumiger Ebene können zu Indikatoren gefasst werden, die Entwicklungen innerhalb der Städte abbilden. Gerade im Vergleich mit anderen Städten können Handlungsbedarfe in den einzelnen Städten eruiert und realistisch eingeschätzt werden. Der Datenkatalog wird in der Kooperationsgemeinschaft abgestimmt, regelmäßig überprüft und gegebenenfalls angepasst. So wird die Relevanz der Daten gesichert.

Um Stadtvergleiche zu ermöglichen, sind die Ansprüche an die *Genauigkeit* der Daten der IRB relativ hoch. Dabei kommt es jedoch nicht auf jede einzelne Zahl in den Zellen an, sondern auf das genügend genaue Abbild, das die Variable für die Stadt bzw. den Stadtteil liefert. Die weiter unten dargestellten Kriterien Kohärenz und Eindeutigkeit sind Voraussetzungen für die Genauigkeit der Daten.

Die *Aktualität* der Daten wird über den im Kooperationsprojekt festgelegten Zeitrahmen abgesichert. So sind die Daten nicht tagesaktuell, sondern geben den Stand zum 31.12. des Vorjahres wieder. Die Daten werden zum Oktober eines Jahres geliefert, sodass dazwischen eine Zeitspanne von zehn Monaten liegt. Dazu kommen die Zeiten der Aufbereitung und der Nachlieferungen. Der Datensatz liegt daher in der Regel erst zum Beginn des Folgejahres vollständig vor. Stadtvergleichende Auswertungen können im Rahmen der IRB mit einem Vorvorjahresdatenstand vorgenommen werden. Dies ist insofern hinreichend aktuell, als allgemeine Trends sich nur langsam innerhalb der Städte durchsetzen und die Analyseziele im Stadtvergleich sich eher auf grundlegende Veränderungen beziehen.

Als weiteres Qualitätsmerkmal ist der Punkt *Ressourcen* aufgeführt. Das Engagement des jeweiligen statistischen Amtes im Kooperationsprojekt hängt immer von den gegebenen Ressourcen und Priorisierungen in der Kommune ab. Je nach Aufgaben der jeweiligen Statistikstelle und der Arbeitsteilung innerhalb der Städte, kann es aufgrund mangelnder Ressourcen zum Ausfall eines Themenbereiches oder gar einer ganzen Jahreslieferung kommen. Insbesondere, wenn mehrere Wahlen anstehen oder aufwändige Zusatzerhebungen alle Arbeitskräfte bündeln, kommt es vor, dass einzelne Städte gar nicht oder erst sehr verspätet ihre Daten für die IRB liefern können. Dem Unsicherheitsfaktor Ressourcen kann nur dadurch entgegengewirkt werden, dass der Mehrwert der IRB auch für die einzelne Stadt deutlich wird. Hierfür muss einerseits

Antje Güleş
Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Bundesamt für Bauwesen und Raumordnung
Deichmanns Aue 31–37
53179 Bonn
E-Mail: antje.gueles@bbr.bund.de

in der Stadt selbst die IRB einen selbstverständlichen Platz in der statistischen Berichterstattung einnehmen, andererseits muss aber auch seitens des BBSR der Mehrwert der IRB für den Bund deutlich werden. Durch eine Aufwandsentschädigung wird zumindest ein Teil der in den Kommunen eingesetzten Ressourcen zurückerstattet. Die Aufbereitung, Prüfung und Dokumentation der IRB-Daten, aber auch die Vernetzung und Organisation der IRB-Kooperation nimmt einen guten Teil der Arbeitszeit der mit der IRB betrauten Personen im BBSR ein. Hinzu kommt der Aufwand, die Daten auszuwerten, zu analysieren und die Befunde zu veröffentlichen. Im BBSR sind insbesondere die Ressourcen für die Aufbereitung der IRB inklusive der Datenplausibilisierung begrenzt, weshalb hier ergänzend auf die Fehlerrückmeldung der Datennutzer, insbesondere der IRB-Städte, gesetzt wird. Gerechtfertigt ist der Mitteleinsatz jedoch durch die dabei gewonnene Qualität der Ergebnisse und ihrer Reichweite in den Aussagen für deutsche Großstädte. Das Analysepotenzial der IRB ist in dieser Hinsicht noch nicht ausgeschöpft.

Eindeutigkeit und Kohärenz der Daten der innerstädtischen Raumbeobachtung sind bezogen auf jede Stadt und zwischen den Städten prüfbar. Hier kann man mehrere Ebenen unterscheiden. Zum einen wird die Ebene der Variablen über die Kooperation zwischen den Städten so abgestimmt, dass die Variablen eindeutig definiert und auch zum gleichen Stand geliefert werden. Die Aufnahme neuer Variablen erfolgt in Probeläufen, bis eine einheitliche Erhebung vorliegt. Jüngstes Beispiel für diese Vorgehensweise bildet die Erhebung zum Migrationshintergrund. Wegen der unterschiedlichen Erhebungsmethoden in den Städten ist für dieses Merkmal noch keine Kohärenz erreicht. Das Konzept des Migrationshintergrundes ist durch die Mikrozensusauswertungen des statistischen Bundesamtes bekannt. Im Mikrozensus wird seit 2005 der Migrationshintergrund im Rahmen einer Befragung erhoben. Auf dieser Grundlage kann nach der Definition des statistischen Bundesamtes[2] der Migrationshintergrund für die befragten Personen ermittelt werden. Demgegenüber wird in den Städten ein Registerauszug erstellt. Da der Migrationshintergrund selbstverständlich nicht Bestandteil des Meldewesens ist, wird in den Städten jeweils indirekt aus Daten des Melderegisters auf den Migrationshintergrund geschlossen. Die Verfahren hierzu sind von Stadt zu Stadt unterschiedlich und schließen jeweils unterschiedliche Gruppen ein. Von der sehr engen Fassung der Personen mit (ggf. zusätzlicher) ausländischer Staatsbürgerschaft bis hin zur Interpretation von Namen sind Lösungsansätze entwickelt worden. So sind die Daten zum Migrationshintergrund bislang weder zwischen den Städten noch zu den Mikrozensusergebnissen kohärent. Eine Kohärenz zwischen den Städten wird weiterhin angestrebt. Vorreiter ist hier die KOSIS-Gemeinschaft Koordinierte Haushalts- und Bevölkerungsstatistik HHStat. Sie entwickelte mit MigraPro ein Tool zur Ermittlung des Migrationshintergrundes, das in einigen Städten bereits angewendet wird. Bis dahin wird die vergleichende Interpretation der Zahlen zum Migrationshintergrund nur mit Kenntnis der entsprechenden Ermittlungsalgorithmen der Städte möglich sein.

Weiterhin ist die Anforderung der Kohärenz für einige Variablen in der Zeitreihe schwierig zu realisieren. Beispielsweise führen Gesetzesänderungen und Verfahrensänderungen innerhalb der Statistiken der Bundesagentur für Arbeit im zeitlichen Verlauf zur unterschiedlichen Erfassung arbeitsloser Personen. Um in diesem Kontext Zeitreihen bilden zu können, ist ein umfassendes Wissen der Erhebungsumstände vonnöten. Der Dokumentation von Variablen kommt daher zukünftig eine hohe Bedeutung zu.

Innerhalb der Städte sind die Variablen eindeutig definiert. Mehrdeutigkeiten ergeben sich in der Zeitreihe auf Stadtteilebene, wenn Beobachtungsräume neu zugeschnitten werden. Auf das Problem inkohärenter räumlicher Einheiten und daraus resultierender Mehrdeutigkeit im Zeit- und Städtevergleich wird in Kapitel 5 dieses Beitrags eingegangen.

Kohärenz besteht im übergeordneten Rahmen nur teilweise zu den Daten auf Kreis- bzw. Landesebene. So lassen sich etwa die Bevölkerungsstrukturen nicht unmittelbar mit den vom Land gelieferten Daten innerhalb der Laufenden Raumbeobachtung des BBSR vergleichen. Die Daten, die auf Landesebene für einzelne Kreise und kreisfreie Städte geliefert werden, beruhen auf der

(1)
Zur Geschichte und Einbettung der Datensammlung siehe Gutfleisch/Sturm in diesem Heft.

(2)
Zur Bevölkerung mit Migrationshintergrund zählen alle Personen, die nach 1949 auf das heutige Gebiet der Bundesrepublik Deutschland zugezogen sind, alle in Deutschland geborenen Ausländer/-innen und alle in Deutschland mit deutscher Staatsangehörigkeit Geborene mit zumindest einem zugezogenen oder als Ausländer in Deutschland geborenen Elternteil (vgl. Statistische Ämter des Bundes und der Länder 2013, Glossar).

Bevölkerungsfortschreibung der Volkszählung von 1987. Die Daten der IRB werden jedoch aus den Einwohnerregistern erzeugt. Auf diese Art und Weise ist der kleinräumige Vergleich erst möglich. Differenzen zwischen den Zahlen der Fortschreibung und den Einwohnermelderegistern sind teils erheblich und wurden auch durch die aktuelle Zensuserhebung deutlich.

2 Der Aufbau des IRB-Datenkatalogs

2.1 Datensammlung

In der jährlichen IRB-Mitgliederversammlung wird die Aktualisierung des Katalogs besprochen und die nächste Erhebungsrunde abgestimmt. Da in den Städten auch die abgeleiteten Statistiken (Haushalte, Migrationshintergrund – siehe Infokasten) auf unterschiedliche Art und Weise erhoben werden, werden in der Abstimmung auch die jeweiligen Verfahren thematisiert. Zudem werden die zu erhebenden Variablen geprüft, ihre Aktualität betrachtet und der Merkmalskatalog[3] angepasst. In diesem Abschnitt wird der Prozess der Sammlung und Aufbereitung der Daten der IRB dargestellt.

Die Datensammlung der innerstädtischen Raumbeobachtung erfolgt jährlich für den Stand 31. Dezember des Vorjahres. Die Daten werden im Frühsommer angefordert. Im Herbst des Jahres sind die meisten Datenlieferungen erfolgt. Um eine hohe Datenqualität zu gewährleisten, erfolgt die Sammlung standardisiert in Excel-Tabellen. Sie umfassen 30 Tabellen mit insgesamt über 400 Spalten bzw. Variablen. Diese sind so erstellt, dass – bei korrekter Eingabe der Daten – eine automatisierte Zusammenstellung ermöglicht wird. Im Jahr 2010 wurde eine Arbeitshilfe erstellt, die sowohl die Abläufe als auch die Datenqualität verbessert hat. Die in der Arbeitshilfe geforderten Standards sind hierfür zwingend vorausgesetzt. Tabellenlieferungen, die nicht den formulierten Standards entsprechen, werden in der Regel erneut angefordert. Zu den Standards gehört der einheitliche Umgang mit fehlenden Werten, ebenso der weitgehende Verzicht auf die Anonymisierung (siehe Infokasten) der Daten. Die Stadtteile sollen aufsteigend sortiert und vollständig sein. Die Tabellen sollen, abgesehen von „-1“ für fehlende Werte, nur positive Zahlen oder „0“ enthalten. Als zusätzlicher Stadtteil existiert für alle Städte der Stadtteil 99999, der die nicht zuordenbaren Fälle enthält. Die statistischen Ämter plausibilisieren die zu liefernden Daten vor Ort. Da dies nicht immer gewährleistet werden kann, kommt diesem Punkt innerhalb des BBSR eine umso größere Bedeutung zu.

Die Vollständigkeit der für den Stichtag 31.12.2011 gelieferten Tabellen der einzelnen Städte ist in Abbildung 1 wiedergegeben. Deutlich wird, dass wenige Städte (aktuell sind es vier) alle Tabellen bedienen können. Eine Stärke der IRB ist jedoch, dass alle anderen über 75 % der angeforderten Inhalte liefern können. Seit 2002 haben sich 45 Städte an der IRB beteiligt. Weitere sind in 2003, 2004, 2006 und 2011 hinzugekommen. Eine Stadt ist 2009 ausgeschieden. Aktuell sind 51 Städte in der IRB vertreten, wobei eine Stadt für 2011 noch keine Daten einspeisen konnte. Die folgenden Darstellungen beziehen sich daher auf die 50 in 2011 aktiven Städte.

Haushalte

Haushalte und deren Struktur werden im Einwohnermelderegister nicht erfasst. Um trotzdem einen Überblick über die Haushalte zu bekommen, werden Verfahren entwickelt, wie Einzelpersonen an Wohnadressen zusammengefasst werden können. In einigen Städten wird hierzu der „steuerrechtliche Personenverband“ ermittelt.

Das Haushaltegenerierungsverfahren HHGEN der KOSIS-Gemeinschaft HHSTAT fasst Haushalte in einem mehrstufigen Verfahren zusammen. Genutzt werden hierbei vorhandene Verknüpfungen zwischen Ehegatten sowie zwischen Kindern und deren Elternteilen. Auch Namensübereinstimmungen im Bereich der Familiennamen und demografische Merkmalskonstellationen, die mit anderen Indizien auf bestimmte familiäre Beziehungen hinweisen, werden verwendet.

Im ersten Schritt werden die Personen nach Ihrer Stellung im Haushalt klassifiziert, hernach werden nicht-eheliche Paare generiert und dann die Nachkommen zu Vorfahren zugeordnet. In weiteren Schritten wird so jede Person in eine Beziehung zu anderen Personen in dem betreffenden Haushalt gesetzt. Die einzelnen Stufen können auf den Seiten der KOSIS-Gemeinschaft HHSTAT unter www.staedtestatistik.de nachgelesen werden.

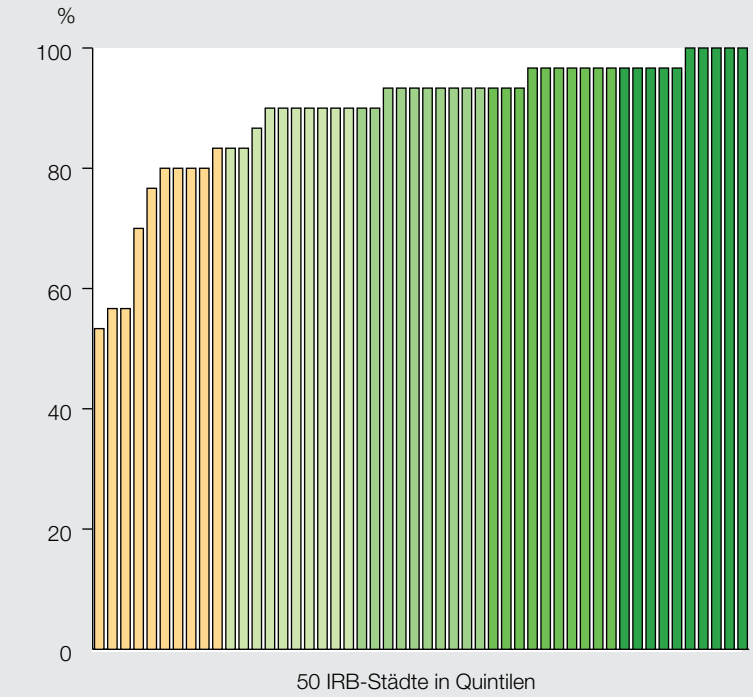
Anonymisierung

Aus datenschutzrechtlichen Gründen müssen Verfahren zur Anonymisierung gefunden werden, um nicht aus den Merkmalen und Merkmalskombinationen auf einzelne Personen im Stadtteil schließen zu können. Zwar ist dies aufgrund der nur gebietsbezogenen Daten in der IRB kaum möglich. In Einzelfällen kann es aber durchaus vorkommen, dass etwa die einzige ausländische Familie in einem kleinen ostdeutschen Stadtteil mit drei Personen plötzlich in den IRB-Daten „zu sehen“ ist.

Die meisten Städte verwenden eigene Anonymisierungsverfahren. In der Regel werden die Daten der IRB unanonymisiert an das BBSR weitergeleitet. Dies ist notwendig, da sonst die Zeilen- und Spaltensummen in den Datentabellen fehlerhaft werden. Um also einer Identifikation dieser Familie und ihrer Einzelpersonen zu vermeiden, müssen die Daten der IRB vor ihrer Weitergabe an Dritte anonymisiert werden. Dies geschieht in einfacher Form. Zellen mit Häufigkeiten <4 werden entweder in „0“ oder „4“ recodiert. Daraus resultieren in kleinen Stadtteilen Abweichungen, die sich jedoch im Gesamtvergleich und bei Analysen wieder relativieren

(3) Siehe Anhang: Merkmalskatalog der aktuellen Erhebung (Stand 2012).

Abbildung 1
Vollständigkeit der IRB-Lieferung 2011



Datenbasis: Innerstädtische Raumbeobachtung des BBSR, Kommunalstatistiken der IRB-Städte

In einzelnen Jahren gab es spezifische Ausfälle, diese sind jedoch begrenzt und dem zeitweisen Fehlen von Arbeitskapazitäten in der jeweiligen Stadt geschuldet. Für die Jahre 1980 bis 1997 sind die Daten sehr lückenhaft. Auch waren in den 80er und 90er Jahren die EDV-technischen Voraussetzungen nicht gegeben, die in einem angemessenen Zeitaufwand die Verfügbarkeit aller heute gelieferten Variablen ermöglicht hätten. In den Jahren 1998 bis 2001 wurde aufgrund des parallelen Aufbaus des KOSIS-Datensatzes KOSTAT auf die Weiterführung der IRB verzichtet. Daher ist es in den meisten Analysekontexten sinnvoll, auf den Datensatz ab 2002 zuzugreifen. Nur in Einzelfällen lohnt sich eine Analyse mit den langen Zeitreihen und ist hauptsächlich für die aus KOSTAT importierbaren Variablen möglich. Hinsichtlich der grundlegenden Bevölkerungszahlen lässt sich die Lücke zwischen 1998 und 2001 hiermit schließen. Zudem ermöglicht die Sondertabelle IRB-2000 „Historische Daten“, dass insbesondere der Kooperationsgemeinschaft neu beitretende Städte einen grundlegenden Fundus von zwölf Variablen für frühere Jahre nachmelden können.

2.2 Datenaufbereitung

Die Aufbereitung erfolgt in mehreren Schritten per Excel und SPSS. Zunächst werden die Tabellen auf ihre formale Vollständigkeit und Korrektheit überprüft. Ist an dieser Stelle nichts zu beanstanden, kann mittels eines Excel-Makros die Tabellensammlung einer Stadt zusammengefasst werden. Eine zwingende Voraussetzung ist, dass die Reihenfolge der Stadtteile in den Tabellen aufsteigend sortiert ist und gleich bleibt, da sonst alle verschobenen Stadtteile falsch zugeordnet werden. Dies zeigt sich spätestens, wenn in der Plausibilisierung unlogische Werte erreicht werden.

Die per Makro zusammengefassten Tabellensammlungen werden nun mit SPSS eingelesen und zusammengefügt. Einige häufig verwendete Variablen, etwa die Altersgruppe der erwerbsfähigen Bevölkerung, werden berechnet. Variablen und Werte (die Gemeindekennziffer) werden gelabelt und der Datensatz mit den Vorjahresdaten verknüpft. Abhängig von der Analysengrundlage der einzelnen Städte wird die IRB-Bevölkerung definiert. Einige Städte verwenden in ihren Analysen die wohnberechtigte Bevölkerung, andere die Hauptwohnbevölkerung als Bezugsgröße. Deshalb dient die Definition der IRB-Bevölkerung der Kohärenz zu den in den einzelnen Städten vorgenommenen Analysen. Eine weitere Identifikationsvariable „idirb“ wird erstellt, die eine Vergleichbarkeit und Verknüpfbarkeit über die Jahre gewährleistet. Zusätzlich zu den Datentabellen werden Anmerkungen und Kommentare zu Änderungen in den Städten (etwa Neuzuschnitte von Stadtteilen oder die Einführung der Zweitwohnsitzsteuer) und Schwierigkeiten bei den Tabellen zentral dokumentiert. So geht keine Information zur Interpretation der Daten verloren. Der fertige Datensatz wird im nächsten Schritt plausibilisiert.

3 Plausibilisierung

In den Vorjahren wurden kleinere Fehler nicht weiter beachtet, da sich in der Regel auf der Ebene der innerstädtischen Lage (siehe Kapitel 5 in diesem Beitrag), auf der im BBSR ausgewertet wird, und bei Stadtteilen mit mehr als 50 Einwohnern die Fehler herausgemittelt haben. Dies hatte zur Folge, dass in einigen Städten noch Fehler

aus den Vorjahren vorhanden waren. Inzwischen sind die Anforderungen an die Daten jedoch gestiegen, da verstärkt Analysen auf Stadtteilebene (etwa Korrelationen zwischen Indikatoren, Ermittlung von Segregationsindizes) durchgeführt werden.

Bei der Prüfung einzelner Werte werden keine Konfidenzintervalle angelegt, die Städte haben hierfür zu unterschiedliche Zuschnitte und weisen je nach Variable Besonderheiten auf. Trotzdem lassen sich die Fehler auch ohne dieses Instrument finden. Extremwerte in einzelnen Stadtteilen oder aber völlig unerwartete Werte für die Gesamtstadt sind Hinweise auf fehlerhafte Daten. Vor allem starke Schwankungen der Werte zwischen unterschiedlichen Jahren lassen einen Fehler vermuten. Zwar ist bei einzelnen Variablen auch mit Schwankungen zu rechnen (etwa wenn aufgrund der Einführung der Zweitwohnsitzsteuer die Anzahl der Hauptwohnbevölkerung in studentischen Vierteln stark steigt), jedoch lassen sich diese häufig auf Änderungen in der Stadt oder allgemeine gesellschaftliche Entwicklungen zurückführen. Lässt sich keine plausible Erklärung für ausreißende Werte finden, ist dies ein Grund, die Einzelwerte der betreffenden Stadt zu prüfen. Stadtteile mit kleinem Zuschnitt und wenig Bevölkerung sind von Extremwerten häufiger betroffen als Stadtteile mit hoher Bevölkerungsdichte. Da die Werte ungewichtet in die Plausibilisierung eingehen, können auch kleine Stadtteile für entsprechende Fehlerspannen sorgen.

Ziel bei der Plausibilisierung ist, möglichst viele Fehlerquellen aufzudecken, ohne dabei die eigentliche Arbeit mit der IRB – Auswertung und Analyse der Daten – zu vernachlässigen. Um die Plausibilisierung möglichst effizient zu gestalten, wird ein Methodenmix aus Überprüfung einzelner Variablen, Bildung von sinnvollen Summen und Differenzen sowie Bildung von Indikatoren zur Überprüfung mehrerer Tabellen angewendet. Die Daten werden hauptsächlich im Zeitvergleich und auf Ebene der Städte geprüft. Eventuelle Abweichungen über die Zeit lassen sich so leichter auffinden. Im grafischen Abgleich wird aktuell auch der Vergleich einzelner Stadtteile möglich. Die folgenden Abschnitte stellen die Plausibilisierung exemplarisch dar.

3.1 Einzelne Variablen

Im ersten Schritt werden einzelne Variablen ausgewertet. So wird die Gesamtbevölkerung der Städte ermittelt, die Anzahl der Stadtteile über die Jahre verglichen und die Verteilung der Stadtteile auf die innerstädtischen Lagetypen überprüft. Hierbei kann es, bedingt durch Gebietstandsänderungen, immer mal wieder zu Abweichungen kommen, die dann durch Rückfragen und Abgleich mit den Anmerkungen zur Datenlieferung geklärt werden. Die Überprüfung einzelner Variablen, wie Einwohnerzahl oder Verteilung der Stadtteile über Lagetypen, bildet gleichzeitig die Grundauszählung für die IRB.

3.2 Summen und Differenzen über Tabellen

Ein guter Teil des IRB-Merkmalskatalogs ist durch die Aufteilung in Altersgruppen geprägt. Die Überprüfung der Tabellen wird durch Aufsummieren der einzelnen Altersgruppen und Differenzbildung zur Gesamtheit der Einwohner vorgenommen. Etwa bei den Einwohnern am Hauptwohnsitz: Hier sind von „0 bis 3 Jahre“ bis „85 und mehr Jahre“ 13 Altersgruppen abgebildet. Die Differenz zur Einwohnerzahl am Hauptwohnsitz muss dabei in allen Stadtteilen Null ergeben. In den Fällen, wo einzelne Stadtteile anonymisiert geliefert werden, ist dabei mit Abweichungen zu rechnen, was eine der häufigsten Fehlerquellen gerade bei kleinen Stadtteilen darstellt. Daher wird angestrebt, die Daten auch in kleinen Stadtteilen nicht anonymisiert zu erhalten und erst bei Abgabe an Dritte zu anonymisieren.

Auch bei den Wanderungen lassen sich Differenzen über Zu- und Fortzüge bilden und so die einzelnen Tabellen überprüfen. Am einfachsten gelingt dies bei den Wanderungen innerhalb der Stadt. Hier muss die Summe über alle Stadtteile Null ergeben. Bei den Wanderungssalden über die Stadtgrenzen gibt es wiederum große Unterschiede zwischen den Städten, sodass hier eine Plausibilisierung nicht ohne Weiteres an einem bestimmten Wert gelingt. Weisen jedoch Städte mit Wachstums- oder Schrumpfungstendenz einen dem Trend entgegenstehenden Wert auf, kann auch hier von einem Fehler ausgegangen werden.

3.3 Indikatoren

Der wichtigste und umfangreichste Schritt ist die Bildung und Prüfung von Indikatoren. Dies hat den Vorteil, mehrere Variablen der IRB zu plausibilisieren. Geprüft wird die Ausprägung der Indikatoren sowohl inhaltlich als auch über die Zeit. Verwendet werden unter anderem die Indikatoren, die als Auszug in der unten stehenden Tabelle samt ihren erwarteten Werten dargestellt werden. Dabei werden die Indikatoren für alle Stadtteile berechnet und der Mittelwert nach Stadt und Jahr wiedergegeben. Abweichungen vom erwarteten Wert müssen dabei entweder durch Spezifika der jeweiligen Stadt erklärbar sein oder durch Extremwerte bei Stadtteilen mit niedriger Zellbesetzung hervorgerufen werden. Sollte dies nicht der Fall sein, liegt der Verdacht nahe, dass es in einer der für die Indikatorbildung verwendeten Variablen einen Fehler gibt. Indikatoren, die Anteile wiedergeben (z. B. Frauenanteil in der Bevölkerung) müssen darüber hinaus zwischen 1 und 100 % liegen. Das Wissen für die Wertebereiche in einzelnen Städten wird einerseits durch den Vergleich mit den Vorjahreswerten und andererseits durch die Verifizierung bei den verantwortlichen Statistikstellen abgesichert. Die Fehlersuche erfolgt im Anschluss durch Tabellenabgleich und Identifikation von Extremwerten für die Indikatoren in den Stadtteilen.

3.4 Grafischer Abgleich

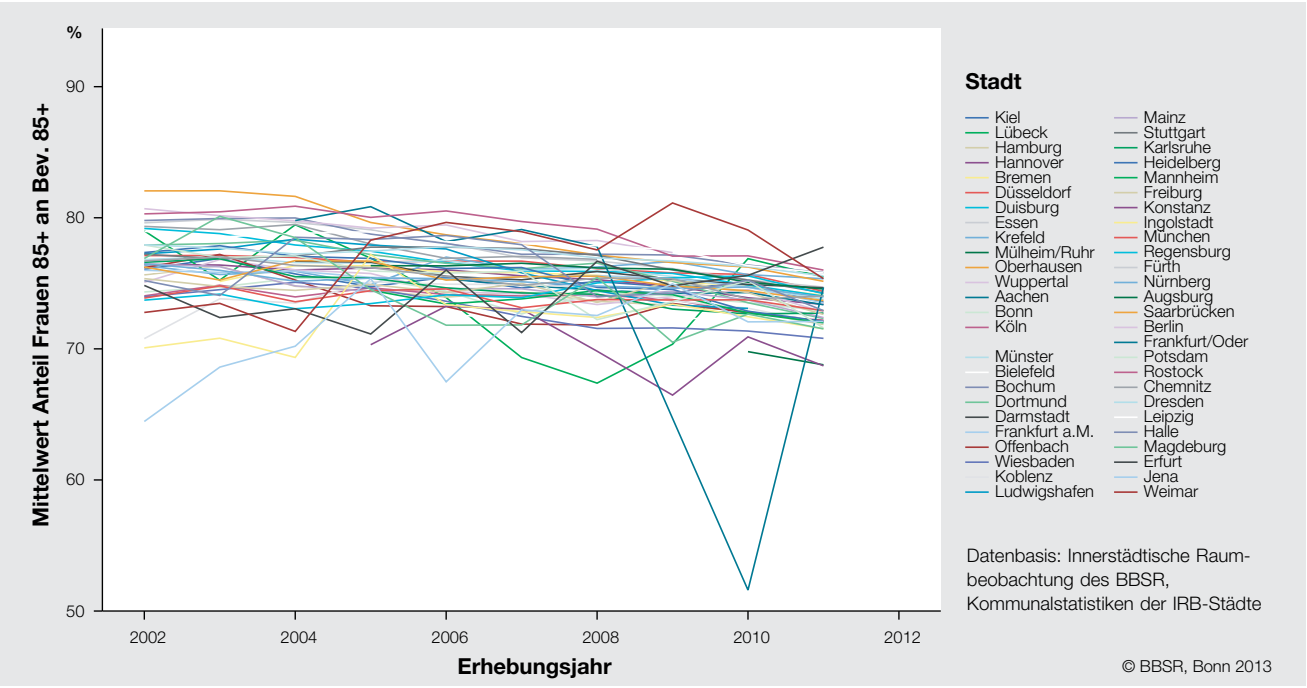
Über einen grafischen Abgleich der einzelnen Indikatoren lassen sich Ausreißer auf einen Blick identifizieren. Hierfür wird der Mittelwert des Indikators über alle Stadtteile einer Stadt gebildet. In Abbildung 2 wird beispielsweise ein Absinken des Indikators „Anteil der Frauen 85+ an Bevölkerung 85+“ in Frankfurt/Oder auf etwas über 50% ersichtlich. Schwankungen sind hier allgemein nicht in diesem Ausmaß zu erwarten. Ein genauerer Blick offenbart die Fehlerquelle: In Frankfurt/Oder wurde die Lieferung für das auffällige Jahr auf eine kleinräumigere Aggregatebene verändert. Zudem lebten in einigen Stadtteilen nur sehr wenige hochbetagte Personen, sodass der Durchschnitt über alle Raumeinheiten sehr schwankt, auch wenn nur einzelne Personen sterben. Eine weitere mögliche Fehlerquelle für diesen Indikator wäre das Vertauschen der Bevölkerungszahlen für die weibliche oder männliche Bevölkerung gewesen. Dann wäre der entsprechende Indikatorwert sprunghaft auf etwa 25% gesunken.

Eine jüngere Entwicklung stellt im BBSR das Softwaretool Tableau dar. Mit Hilfe dieses Tools lassen sich die einzelnen Stadtteile für jede Stadt schnell im grafischen Vergleich darstellen und auf einen Blick

Tabelle 1
Auszug von Indikatoren und deren Prüfwerten für die IRB

Indikator	Variablen, die überprüft werden	Algorithmus	Prüfwert
Anteil Hauptwohnbevölkerung an wohnberechtigter Bevölkerung	Hauptwohnbevölkerung gesamt, wohnberechtigte Bevölkerung gesamt	antewhg = ewhg/ewwg*100.	Je Stadt und Zweitwohnsitzsteuerregelung unterschiedlich
Anteil Frauen an Bevölkerung	Frauen gesamt, IRB-Bevölkerung gesamt	antfrau=frag/ irbg*100	~ 50 %
Anteil hochbetagter Frauen an Hochbetagten	Frauen 85 und mehr Jahre, IRB-Bevölkerung 85 und mehr Jahre	antfrau3=fra85um/irb85um*100.	~ 70 % bis 75 %
Anteil Personen ohne deutsche Staatsangehörigkeit an Bevölkerung	Ausländer gesamt IRB-Bevölkerung gesamt	antaus=ausg/ irbg*100	~ 8 bis 30 % im Westen ~ 3 bis 14 % im Osten der Republik (inkl. Berlin)
Bevölkerungsdichte	IRB-Bevölkerung gesamt, Fläche gesamt	bevdi=irbg/fl_ges.	~ 20 bis 80 EW/ha
Beschäftigte je Einwohner im Alter von 15 bis < 65 Jahre	Sozialversicherungspflichtige Beschäftigte, IRB-Bevölkerung 15 bis unter 65 Jahre	besch=besges/irb1565*100.	50 bis 70 %, Konjunkturabhängig
Anteil älterer Arbeitsloser	Arbeitslose 55 und mehr Jahre, Arbeitslose gesamt	aloalt=alo55um/aloges*100.	~ 10 bis 25 %, Tendenz momentan steigend
Euro je SGB II-Leistungsberechtigte	Ausgaben für SGBII in Euro, Anzahl der Leistungsempfänger SGBII	sgb2_fin=sgb2_eur/ sgb2_ges.	~ 450 bis 600 €, abhängig vom Wohnungsmarkt der Stadt
Anteil Einpersonenhaushalte	Haushalte alleinlebender, Haushalte insgesamt	anthhs=hh_1/hh_ges*100.	~ 50 %
Anteil Haushalte mit Kindern an Haushalten	Haushalte mit Kindern, Haushalte insgesamt	anthhk=hhk_ges/hh_ges*100.	< 20% mit Ausnahmen
Personen je Wohnung	IRB-Bevölkerung, Anzahl der Wohnungen gesamt	antwohn=irbg/ wohnges	~ 2

Abbildung 2
Anteil der Frauen 85+ an Bevölkerung 85+

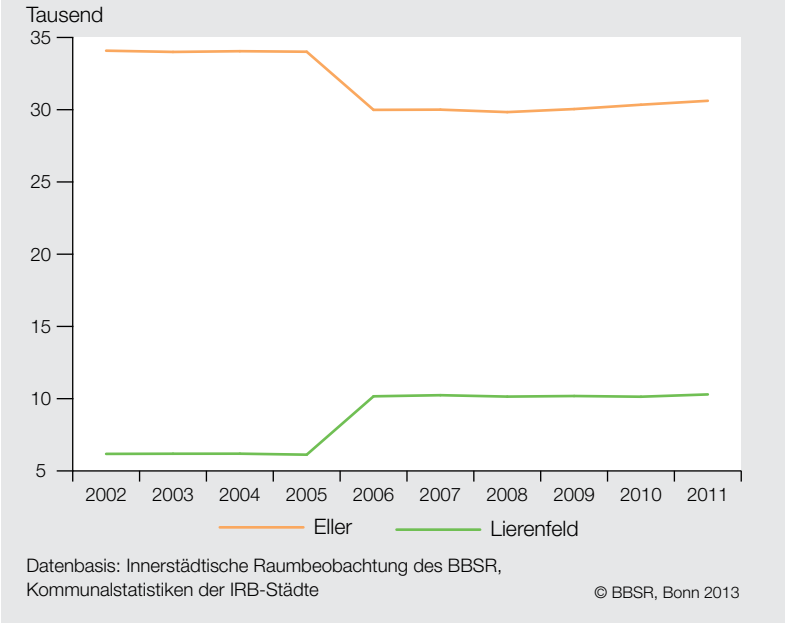


erfassen. Beispielsweise werden Änderungen im Stadtteilzuschnitt dadurch deutlich, dass etwa der Verlauf für die Anzahl der Einwohner zwischen zwei Stadtteilen symmetrisch wechselt. In dem einen Stadtteil nimmt die Zahl der Einwohner sprunghaft zwischen zwei Jahren zu, während sie in einem anderen Stadtteil um etwa den gleichen Betrag sinkt (vgl. Abb. 3). Ausreißer lassen sich durch Peaks im Jahresverlauf identifizieren. Änderungen in der Stadtteilbenennung werden sichtbar, weil Jahresverläufe nicht in einer Linie fortgeführt werden. Die Aufarbeitung der auf diese Art und Weise deutlich gewordenen Unstimmigkeiten hilft, auch ältere Unstimmigkeiten zu identifizieren.

4 Rückmeldung und laufende Korrekturen

Sind beim Überprüfen der Werte Auffälligkeiten festgestellt worden, wird zunächst der Fehler in den Daten der zusammengesetzten Dateien und der Originallieferung gesucht. Fehler, die durch die Verarbeitung im BBSR entstanden sind, können in der Regel schnell identifiziert und behoben werden. Fehler, die durch Datenverschiebungen zwischen einzelnen Tabellen oder

Abbildung 3
Detailansicht der Veränderung der Einwohnerzahl in Lierenfeld und Eller, Landeshauptstadt Düsseldorf



Stadtteilen in der Datenlieferung entstehen, können meist durch auffällige Werte und Abweichungen von der Originallieferung identifiziert werden. Die Originallieferung wird in diesen Fällen entsprechend angepasst (etwa die Stadtteile in die richti-

ge Reihenfolge gebracht und dann entsprechend aktualisiert). Wenn die Fehler nicht so einfach zu identifizieren und im BBSR zu beheben sind, werden sie in die Städte zurückgemeldet und entsprechende Korrekturen angefordert. Die Korrekturen werden eingearbeitet, sodass im Laufe des Jahres neue Versionen der IRB entstehen. Dies macht eine Versionierung der IRB-Datensammlung für den jeweiligen Jahresstand notwendig. Mit der aktuellen Datenlieferung wurde die Versionierung eingeführt.

Da bei über 400 Variablen und inzwischen zehnjähriger Neuauflage der IRB nie alle Fehler restlos behoben werden können, kann es sein, dass bei bestimmten Berechnungen auffällige Werte resultieren. Diese werden wie bei der Plausibilisierung zu ihrem Ursprung zurückverfolgt und gegebenenfalls korrigiert. Auch werden die zuständigen Mitarbeiterinnen und Mitarbeiter in den Städten selbst auf Fehler in ihren Daten aufmerksam und senden entsprechend korrigierte Tabellen zu. All dies führt zu einem Datensatz, dessen Qualität stetig zunimmt. Imputationen werden für die IRB nicht vorgenommen, da dies je nach Analysezweck unterschiedlich geschehen müsste. Bei einigen Analysen werden einzelne fehlende Werte ersetzt, um ein etwas vollständigeres Bild zu erreichen. Dies erfolgt von Fall zu Fall und findet keinen Eingang in den Datensatz.

Ein weiterer wichtiger Punkt für die Qualitätssicherung und die Verwertung der IRB-Daten ist die Nutzung der IRB durch wissenschaftliche Einrichtungen. Auch hier werden immer wieder Fehler oder Unstimmigkeiten in den Daten entdeckt und zurückgespielt. Gleichzeitig lassen die Ergebnisse etwa von Diplom-Arbeiten Rückschlüsse auf die IRB und deren Datenqualität und Auswertungsmöglichkeiten zu. Das Nachhalten der Datenabfragen wissenschaftlich forschender Dritter ist somit die letzte wichtige Säule für die Qualitätssicherung in der IRB.

(4) Zum allgemeinen Problem der Gebietsreformen auf Kreis- und Gemeindeebene sowie zur Umschätzung von Daten siehe BBSR 2010d.

5 Umgang mit Gebietstandsänderungen in der IRB

Neben den reinen Datenbeständen in der IRB werden in den Städten immer mal wieder Neuzuschnitte der statistischen Bezirke vorgenommen. Je nachdem, um welche Art des Zuschnitts es sich handelt, müssen Anpassungen erfolgen, um die Daten in der Zeitreihe vergleichbar zu halten und mit aktuellen Geografien verknüpfen zu können. Aufgrund der komplexen Datenstruktur und der Wirkung innerstädtischer Neuzuschnitte von Stadtteilen ist in der IRB bisher keine Umrechnung der vorherigen Gebiete möglich. Hierin besteht ein wesentlicher Unterschied zu den Umschätzungen in der laufenden Raumbeobachtung.[4] Der Referenzierung von Gebietstandsänderungen kommt daher eine eigene Bedeutung zu.

5.1 Referenzierung und Gebietstandsänderungen

Die Stadtteile sind einerseits mit der Gemeindekennziffer und andererseits mit der IRB-Stadtteilkennziffer referenziert. Über diese Kennziffern lassen sich Verknüpfungen zu den Katalogen KOSTAT und Urban Audit herstellen, soweit es sich um ein bestimmtes Jahr handelt. Auch eine Verknüpfung mit Geobasisdaten ist prinzipiell möglich, erfordert allerdings innerstädtische Geometrien und erhebliche Anpassungsleistungen seitens der GIS-Nutzenden, sofern nicht die städtischen Originaldaten verwendet werden. Bei Änderungen der statistischen Einheiten seitens der Städte werden auch die IRB-Stadtteile verändert. So kommt es auch zu einer Neunummerierung der Stadtteile und damit zu einer neuen Referenz. Um zu verhindern, dass Stadtteile im Zeitvergleich zusammengebracht werden, die nicht identisch sind, sind im BBSR folgende Anpassungsregeln entwickelt worden:

- Die alten Stadtteile werden bei einfacher Umbenennung und gleichbleibender Identität der aktuellen Kennziffer zugeordnet.
- Bei Neuzuschnitten bekommen die aktuellen Stadtteile die aktuelle Kennziffer; die Stadtteile, die in den Vorjahren bestanden und nun nicht mehr identisch existieren, erhalten für die Vorjahre eine Kennziffer, die sich eindeutig von den

Kennziffern der anderen Stadtteile unterscheidet.

- Die Änderungen werden in einer Datenbank dokumentiert, sodass sie jederzeit verfügbar sind.
- Seit dem Lieferstand 2011 werden die Änderungen nur berücksichtigt, sofern sie mehr als 1 % der Bevölkerung betreffen und nicht in einem neu entwickelten Stadtgebiet liegen.

Beispielsweise wurden im Jahr 2008 in Hamburg einige Stadtteile per Senatsbeschluss im Zuschnitt geändert und eine Neunummerierung vorgenommen. Im Bezirk Altona ist dabei der neue Stadtteil Sternschanze entstanden, der Teile von St. Pauli, Altona-Altstadt, Eimsbüttel und Rotherbaum enthält. Dies hat für die IRB-Kennziffern zur Folge, dass nicht nur die neuen Kennziffern vergeben wurden, sondern für die Stadtteile, die je einen Teil zur Sternschanze „verloren“ haben, für die Vorjahre eindeutig unterschiedene Kennziffern vergeben werden mussten.

Auf Stadtteilebene sind zeitliche Verläufe daher nur für diejenigen Stadtteile abzubilden, die während der abzufragenden Zeitspanne keinen Neuzuschnitt erfahren haben. Eine Dokumentation der Änderungen in den Stadtteilen erfolgt über eine Accessdatenbank. Hier werden auch alte und neue vergebene Kennziffern vermerkt, sodass die Änderungen jederzeit abrufbar und mit der IRB referenzierbar sind. Umgekehrt lässt sich über die Gültigkeit von Variablen ein möglichst günstiger Auswertungszeitraum für bestimmte Raumeinheiten festlegen.

Entwicklungen, die auf Grundlage der Stadtteile gemessen, aber für die Gesamtstadt oder die innerstädtischen Lagetypen ausgegeben werden, können in den meisten Fällen unabhängig von kleinräumigen Gebietsstandsänderungen betrachtet und interpretiert werden. Da Zeitvergleiche und Vergleiche zwischen den Städten in BBSR-Analysen regelmäßig auf Grundlage der innerstädtischen Lagetypen durchgeführt werden, gleicht die höhere Aggregatebene die mit den innerstädtischen Gebietsstandsänderungen verbundenen Fehler weitgehend aus.

5.2 Innerstädtische Lagetypen

Die innerstädtischen Lagetypen[5] der IRB dienen im BBSR der stadtvergleichenden Analyse. Sie unterscheiden die innerstädtischen Räume nach ihrer Zentralität. Bei der Beschreibung der Stadtstruktur orientieren sie sich an den historisch gewachsenen Vorgaben (vgl. BBR 2007). Stadtteile werden den folgenden Lagetypen zugeordnet:

- „City“ und „Cityrand/sonstige Gebiete der Innenstadt“, für die ein Teil der Städte die Daten gleich unter der Bezeichnung „Innenstadt“ zusammenfasst;
- „Innenstadtrand“ bzw. innenstadtnahe Stadtteile, die zusammen mit den beiden vorherigen Lagetypen die „Innere Stadt“ bilden;
- „Stadtrand“, oft auch „Äußere Stadt“ genannt (durch Eingemeindungen treten sich leicht ändernde Bezugsgrößen auf);
- zu den vier innerstädtischen Lagetypen kommt die Bestimmung eines „Nahbereichs“ der Städte, der für die Differenzierung der überkommunalen Wanderungen bedeutsam ist. Da die Kommunen in ihrer Behörde mit unterschiedlichen Umlandabgrenzungen arbeiten, beziehen sich auch die Daten für die Umlandwanderungen entweder auf einen eng gefassten Nahbereich, der meist nur die angrenzenden Gemeinden umfasst, oder auf einen stadtregionstypisch weiter gefassten Nahbereich (vgl. BBSR 2012 b).

In Absprache mit den IRB-Städten wurde als Innenstadt der historische Kern definiert. In dieser Lage finden sich die typische verdichtete Infrastruktur und die zentralen Geschäftslagen. Innenstadtgebiete sind durch eine hohe Bevölkerungsdichte sowie eine hohe Fluktuation als „Integrationsdrehscheibe“ gekennzeichnet (vgl. BBSR 2010b).

Um die Innenstadt wird dann ein Ring angrenzender Stadtteile als Innenstadtrand definiert. Dieser enthält häufig die gründerzeitliche Stadterweiterung. Die Bevölkerungsdichte ist hier etwas geringer als in den Kernstadtgebieten. Die dann verbleibenden Stadtgebiete werden als Stadtrand charakterisiert. Der Stadtrand ist sehr unterschiedlich geprägt. Einerseits findet man

(5) Siehe hierzu die Abbildung 1 im vorhergehenden Beitrag.

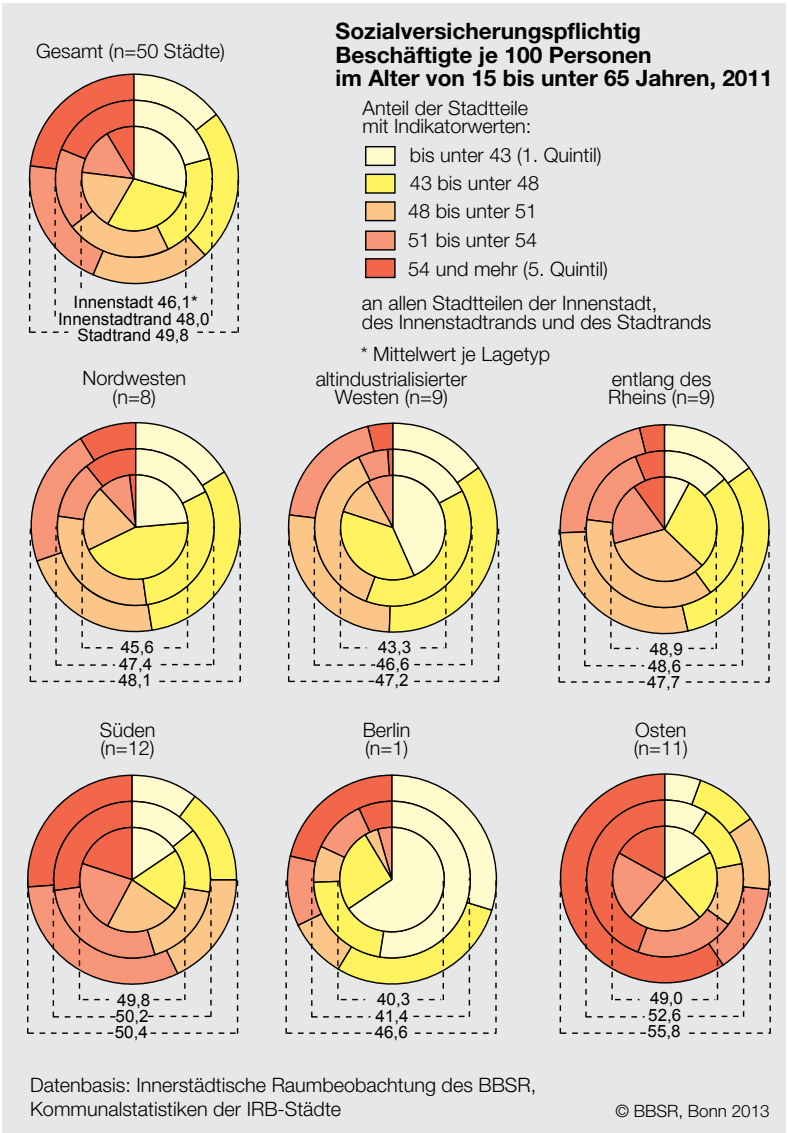
durch Eingemeindungen dörflich gebliebene Strukturen und Einfamilienhausgebiete, andererseits können hier auch die Großwohnsiedlungen der Außenbezirke liegen. Entsprechend sind hier Auswertungen zur Bevölkerungs- und Sozialstruktur sehr genau für die einzelnen Städte zu betrachten.

Insgesamt trägt die grobe Untergliederung in drei Lagetypen jedoch eine Vielzahl vergleichender Analysen und spiegelt sozialräumliche Strukturen in den Städten sehr gut wieder. Als Beispiel ist in Abbildung 4 eine Auswertung der sozialversicherungspflichtig Beschäftigten am Wohnort nach innerstädtischer Lage und regionaler Zuordnung der Städte wiedergegeben. In der Innenstadt findet man die meisten Stadtteile mit niedrigen Indikatorwerten, im Stadtrand liegen die Stadtteile, die hohe Indikatorwerte aufweisen. Deutlich wird, dass die angestellten Beschäftigten eher am Stadtrand wohnen und in Westdeutschland generell niedrigere Beschäftigungsquoten erreicht werden als im Osten der Republik. Letzteres lässt sich vor allem dadurch begründen, dass in den ostdeutschen Städten eine höhere Erwerbsbeteiligung der Frauen besteht.

(6) www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/lor/ [abgerufen am 29.05.2013].

(7) Ein von dem räumlichen Zuschnitt der einzelnen Einheiten abhängiger Index für die Ungleichverteilung einer Bevölkerungsgruppe über die Stadt; vgl. Duncan/Duncan 1955.

Abbildung 4
Anteil der sozialversicherungspflichtig Beschäftigten am Wohnort je 100 Einwohner im Erwerbsalter (15–65 Jahre) nach innerstädtischer Lage und Region 2011



Datenbasis: Innerstädtische Raumbeobachtung des BBSR, Kommunalstatistiken der IRB-Städte

Die Auswertung auf Ebene der innerstädtischen Lage bietet einen guten Überblick über städtische Gegebenheiten, ohne allzu anfällig für kleinere Fehler oder Änderungen von Stadtteilzuschnitten zu sein:

Ändert sich in einer Stadt die Aggregat-Ebene für die IRB-Lieferung oder ganz grundsätzlich die kleinräumige Gliederung, so werden für alle neuen Raumeinheiten der Stadt neue Kennziffern vergeben. Ein zeitlicher Vergleich ist dann kleinräumig nicht möglich, da dieser eine Zuordnung zu den ehemaligen Stadtteilen voraussetzt. Deutlich wird dies am Beispiel Berlin: Die Lieferung erfolgte bis zum Stand 2009 auf Ebene der 195 statistischen Gebiete. Da Berlin bereits seit 2006 mit den Lebensweltlich orientierten Räumen (LOR) auf der Ebene von 447 Planungsräumen[6] arbeitet, wurde die Lieferung für den Stand 2010 hierauf umgestellt. Zugleich wurden die Planungsräume gemäß den innerstädtischen Lagetypen der IRB klassifiziert und die zuvor gelieferten statistischen Gebiete falls notwendig nachträglich angepasst. So sind für Berlin weiterhin Zeitreihenbetrachtungen auf Ebene der Lagetypen möglich. Allerdings sind auch gesamtstädtisch die Folgen für die berechneten Kennwerte bei veränderten Raumzuschnitten zu beachten: Der für Berlin berechnete Segregationsindex[7] für Leistungsberechtigte im SGBII erhöht sich beispielsweise allein durch den Ebenenwechsel von 0,23 in 2009 auf 0,27 in 2010.

6 Ausblick

Die Datensammlung IRB ist ein einmaliger Datensatz in der Bundesrepublik. Die Möglichkeit, mehr als 400 Variablen in mehreren Themenbereichen für 50 Städte kleinräumig in der Zeitreihe auswerten zu können, birgt ein enormes Analysepotential. Gleichzeitig ist der Datensatz für Außenstehende aufgrund seiner komplexen Struktur und spezifischer Einschränkungen schwer zu handhaben. Analysen der Ergebnisse aus den Daten bedürfen tiefgreifender Kenntnisse bezüglich Merkmalsdefinitionen, Aggregatebenen und kommunaler Besonderheiten. Einigen Problemen kann die Datenqualitätssicherung abhelfen. Je weniger Fehler der Datensatz enthält und je besser die Dokumentation der Daten erfolgt, desto leichter ist es, Auswertungen mit der IRB durchzuführen. Eine hohe Datenqualität ist mit allen an der IRB Beteiligten erreichbar und erfordert ein entsprechend abgestimmtes Verfahren des Qualitätsmanagements.

Die zeitlichen Gültigkeiten für einzelne Variablen sollten ebenso dokumentiert werden wie Änderungen ihrer Definition (etwa bei den Variablen zum Arbeitsmarkt und der Grundsicherung). Auch Unterschiede in den Ermittlungsmethoden der Städte für einzelne Variablen (etwa Migrationshintergrund oder Haushaltszusammensetzung) sollten festgehalten werden. Denkbar und für den Bereich „Migrationshintergrund“ im Aufbau ist eine Definition der Variablen für die jeweilige Stadt, sodass Datennutzende sehr schnell sehen können, wodurch Abweichungen von erwarteten Werten begründet sind. So können jeweils eigene Korrekturverfahren je Analysezweck entwickelt werden.

Eine Dokumentation der einzelnen Variablen erfolgt bisher über die Liefertabellen bzw. den Merkmalskatalog. Die zusammenfassende Dokumentation von Variablen und Metadaten wird zukünftig eine wichtigere Rolle spielen.