Georeferenzierung

Geokodierung mit dem RBS-Geocoder

von **Renee Lin**

Der Geoservice des Amtes für Statistik Berlin-Brandenburg (AfS) pflegt die Geodaten, das heißt die Koordinaten für Berliner Adressen, Straßen, Blöcke und verschiedene Raumbezüge wie Wahlbezirke oder Einschulbereiche. Der Berliner Straßenschlüssel wird ebenso vom AfS vergeben. Im Fortschreibungsverfahren werden Adressen sowie deren Zuordnungen zu Blockseiten, Blöcken, Straßenabschnitten und Straßen in das Regionale Bezugssystem (RBS) aufgenommen. Basierend auf diesen Daten wurde im AfS ein Geokodierungsverfahren implementiert und ein RBS-Geocoder mit einer grafischen Benutzeroberfläche entwickelt.

Geokodierung

Raumbezogene Analysen haben in den letzten Jahren mehr an Bedeutung gewonnen und werden in fast allen Bereichen des öffentlichen Lebens, der Wirtschaft und in vielen Wissenschaftsdisziplinen benötigt. Voraussetzung für eine solche Analyse ist die Transformation von Sachdaten in ein räumliches Bezugssystem. Dieses Verfahren der Geokodierung erfolgt mit einer Software, dem sogenannten Geocoder. Dieser besteht aus miteinander verzahnten Operationen, Algorithmen und Datenquellen (Goldberg et al. 2008), die zwischen eingegebener Adresse und georeferenzierten Daten vermitteln (Schulte et al. 2010). Im Geokodierungsprozess wird eine Anschrift in eine Position auf der Erdoberfläche umgewandelt und dieser dann raumbezogene Informationen, zum Beispiel in welchem Ortsteil die eingegebene Anschrift liegt, zugewiesen. Der komplette Ablauf besteht aus mehreren Schritten, bei denen innerhalb einer Referenzdatenbank die zutreffendste Koordinate zur Adresse gesucht wird. Die resultierenden Datensätze werden als geografische Objekte mit Attributen ausgegeben, die zu Zwecken der Zuordnung oder für räumliche Analysen verwendet werden können. In der amtlichen Statistik werden nachgeokodierte Daten zur Bearbeitung und Auswertung sowie zur Ergebnisdarstellung und Visualisierung verwendet.

Referenzdaten

Referenzdaten spielen eine wichtige Rolle im Geokodierungsprozess. Der Umfang und die Aktualität der Referenztabellen beeinflussen die Ergebnisqualität erheblich. Referenztabellen im RBS-Geocoder bestehen aus den folgenden drei Teilen:

- aktuelle Adresse
- historische Adresse
- Objekt

Für die interne Nutzung sowie für die Bereitstellung der Daten, zum Beispiel dem Landesamt für Bürger- und Ordnungsangelegenheiten (LABO) oder der Senatsverwaltung für Stadtentwicklung und Wohnen Berlin, wird monatlich eine Datensicherung und Koordinatentransformation im RBS durchgeführt. Für historische Adressen werden die gelöschten Adressen von der Datenbank ausgelesen. Die Objekttabelle ist eine Sammlung von Bahnhöfen, Plätzen und Flughäfen ohne Hausnummerzuordnung. X/Y-Koordinaten des Objekts sind Mittelpunkte der Straßen oder Objekte; eventuell ist keine Zuordnung zu einigen Gebieten (zum Bespiel Block) möglich. Auch die Referenzdaten für den RBS-Geocoder werden monatlich aktualisiert.

Für die Referenztabellen werden nicht nur Straßenname, Hausnummer, Postleitzahl und Koordinaten aufgenommen, sondern auch die Zuordnungen der verschiedenen Raumbezüge integriert, wie Planungsraum, Verkehrszelle, LAEA-Gitterzelle[1] usw. (Tabelle 1).

RBS-Geocoder

Der RBS-Geocoder ist eine lokale Java-Anwendung im AfS, die die Zuordnung von Koordinaten und Raumbezügen zu einer Liste von Adressen erlaubt. Abbildung a zeigt die Benutzeroberfläche des RBS-Geocoders. Der komplette Ablauf der Geokodierung besteht von der Eingabe bis zur Ausgabe aus vier Teilen (Abbildung b). Zu Beginn wird eine Adressenliste mit den notwendigen Angaben für

1 LAEA (Lambert azimuthal equal-area) ist eine Kartenprojektion, in der die gesamte Kugeloberfläche wiedergegeben werden kann. LAEA-Gitterzellen basieren auf dieser flächentreuen Azimutalprojektion. Als europaweit einheitliche geografische Gitter werden sie in quadratischen Zellgrößen vom 100 m bereitgestellt. Sie dient der Darstellung und Analyse statistischer Sachverhalte.

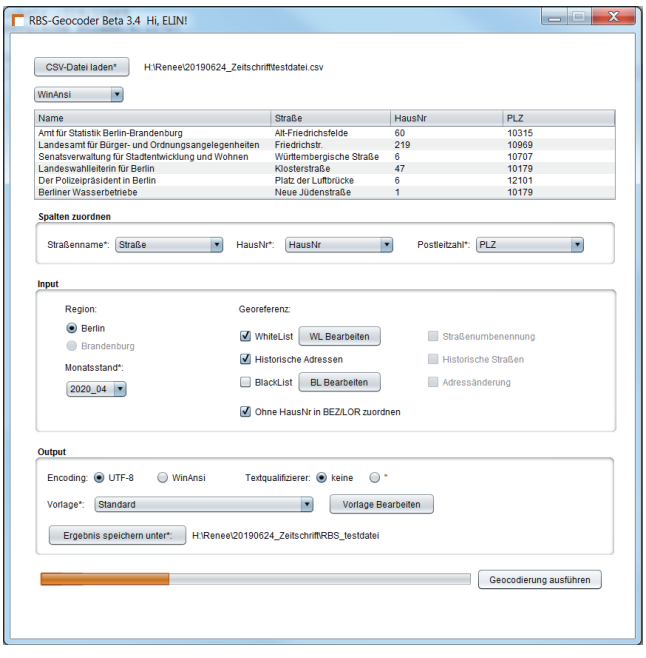
Straßenname, Postleitzahl und Hausnummer (inkl. Hausnummerzusatz) aus einer csv-Datei hochgeladen und der Stand der Referenzdaten sowie eine Ausgabevorlage erfasst. Danach werden Sonderzeichen und Abkürzungen in den Adressen normalisiert. In diesem Vorgang werden fehlende Datensätze von der Eingabedatei ausgefiltert. Daraufhin folgt der dritte Schritt, der den Algorithmus der Adresszuordnung umfasst. Alle Schritte des Prozesses der Adresszuordnung werden in folgender Reihenfolge ausgeführt:

1. Die gesuchte Adresse wird über eine Referenzliste (Blacklist) zu einer benutzerdefinierten Adresse zwingend erforderlich zugeordnet. Das heißt, dass Nutzende ein bestimmtes Ergebnis steuern können. Die Liste kann zukünftig von den Nutzenden selbst gepflegt werden.
2. Hier wird eine gesuchte Adresse mit Hausnummerbereich erkannt und eine eindeutige Hausnummer berechnet. Ein Treffer wird für die erste Hausnummer bzw. die zweite Hausnummer erzielt. Der Algorithmus unterscheidet aufgrund der Berliner Nummernvergabe mit dem gebräuchlichen Zickzack-System (Jacobs 2016) zwischen geraden, ungeraden und fortlaufenden Hausnummern. Bei geraden/ungeraden Bereichen zählt der Algorithmus in Zweierschritten, bei fortlaufenden Bereichen in einem Schritt. Hausnummerzusätze (einzelner Buchstabe) werden nur hochgezählt, wenn in beiden Bereichen Hausnummerzusätze existieren und sie der Reihenfolge des Alphabets entsprechen.
3. Die gesuchte Adresse wird einer aktuellen RBS-Adresse zugeordnet.
4. Die gesuchte Adresse wird einer historischen RBS-Adresse zugeordnet.
5. Die gesuchte Adresse wird einem Bahnhof, einem Platz oder einem Flughafen zugeordnet.
6. Die gesuchte Adresse wird über eine Referenzliste (Whitelist) zugeordnet. In der Whitelist werden Straßennamen oder Adressen mit typischen Rechtschreibfehlern, Abkürzungen, Straßenumbenennungen usw. oder Objektbezeichnungen (zum Beispiel Rotes Rathaus) zu aktuellen RBS-Adressen zugeordnet. Die Liste ist nicht vollständig und kann von den Nutzenden selbst gepflegt werden.

1 | Übersicht über die Berliner Raumbezüge des RBS-Geocoders

Attribut	Beschreibung
AGB	Amtsgerichtsbezirk
AKZ	Aktive Zentren
ARB	Arbeitsamt
ATY	Adresstyp
AWK	Abgeordnetenhauswahlkreise
BEZ	Bezirk
BEZA	Bezirk (alt)
BLK	Blocknummer
BLSFolgeNr	Blockseitefolgenummer
BWB	Briefwahlbezirk
BWK	Bundestagswahlkreise
EGM	Erhaltungsgebiete Milieuschutz
EGS	(EGS) Erhaltungsgebiet städtebauliche Eigenart
ESA	Einschulbereich aktuelles Schuljahr
ESP	Einschulbereich
EUF	EU-Fördergebiet 2000–2006
EWK	Einwohnerkategorie
FIN	Finanzamt
GAF	Gemeinschaftsaufgabe Fördergebiet 2007–2013
GeoGrid_100m_ID	Gitter-ID im Koordinatensystem LAEA Europe (EPSG: 3035)
GRS	Großsiedlung
GRW	Fördergebiet Gemeinschaftsaufgabe „Verbesserung der regionalen Wirtschaftsstruktur" 2014 –2020
KBA	Teilverkehrszellen (Kraftfahrt-Bundesamt)
LOR	Lebensweltlich Orientierte Räume (Planungsraum)
OT	Ortsteil
OTA	Ortsteil alt
OWT	Ost-West politische Teilung
PKB	Polizeikontaktbereich
QM	Quartiersmanagement neu
QMA	Quartiersmanagement (historisch)
SAN	Sanierungsgebiet
SDS	städtebaulicher Denkmalschutz
STG	Statistisches Gebiet
STRNr	Straßennummer
STU	Stadtumbau
SVE	Spielplatzversorgungseinheit
TVZ	Teilverkehrszelle
TVZA	Teilverkehrszellen bis 12/2012
UR2	EU-Gemeinschaftsinitiative (Urban II)
URA	Senstadt Monitoring (UrbanAudit)
UWB	Urnenwahlbezirk (Stimmbezirk)
VBWB	Volksentscheid Briefwahlbezirk
VKZ	Verkehrszelle
VKZA	Verkehrszellen bis 12/2012
VUWB	Volksentscheid Urnenwahlbezirk (Abstimmbezirke)
WOL	Wohnlage
WSG	Wasserschutzgebiet
X_25833	X-Koordinate im ETRS89/UTM Zone 33N (EPSG: 25833)
X_3035	X-Koordinate im LAEA (EPSG: 3035)
Y_25833	Y-Koordinate im ETRS89/UTM Zone 33N (EPSG: 25833)
Y_3035	Y-Koordinate im LAEA (EPSG: 3035)

a | Benutzeroberfläche des RBS-Geocoders



7. Die gesuchte Adresse konnte in oben genannten Schritten nicht gefunden werden. Falls die angegebene Straße und die Postleitzahl jedoch einem Straßenabschnitt zugeordnet werden können, der eindeutig in einem Bezirk verortet oder einem Planungsraum (LOR) zugewiesen ist, werden diese Felder (sofern angefordert) gefüllt.

Es ist optional, ob eine gesuchte Adresse über Blacklist, Whitelist, historische Referenzdaten oder ohne Hausnummer in Bezirken oder den LOR ermittelt werden soll. Die Benutzeroberfläche des RBS-Geocoders ermöglicht eine individuelle Einstellung dieser Referenzdaten.

Als letzter Schritt werden nach der Geokodierung eine Ergebnistabelle und ein Protokoll automatisch generiert. Die Ergebnistabelle wird von der eingegebenen Adressliste mit Qualitätskriterien bzw. Qualitätscodes, Bemerkungen, gefundenen Adressen, Koordinaten und ausgewählten Attributen erweitert. Im Protokoll sind folgende Informationen enthalten:

- Nutzer, Fachgebiet, Uhrzeit und Datum der Ausführung
- Eingabeparameter wie Dateiname, Stand, Ausgabevorlage usw.
- Geokodierungsergebnisse mit Qualitätsauswertung, Gesamtlaufzeit usw.

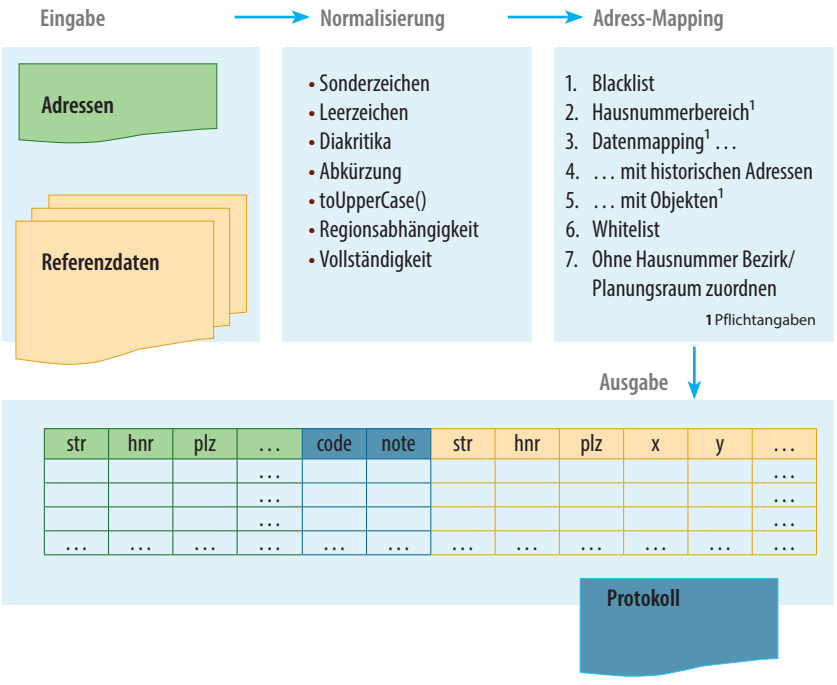
Geokodierungsqualität

Die Qualität der Geokodierung kann durch unterschiedliche Variablen, einschließlich falscher Angaben der Anschriften, fehlender Adresskomponenten und abweichenden Schreibweisen beeinträchtigt werden (McDonald et al. 2017). Schwerwiegend kann es sein, wenn nur einzelne Teile einer Anschrift unvollständig oder fehlerhaft sind. Der Geokodierungsprozess kann unter diesen Umständen möglicherweise zu einer ungenauen Analyse und ungültigen Konklusion führen. Des Weiteren sieht sich die Geokodierung mit den nachstehenden Herausforderungen konfrontiert (Xu et al. 2012):

1. Adressen können unterschiedlich geschrieben werden, beispielsweise B.-Bästlein-Str. statt Bernhard-Bästlein-Straße oder U-Bhf Britz-Süd statt U-Bhf. Britz-Süd. Dieselbe Adresse kann in völlig unterschiedlichen Schreibweisen dargestellt werden. Das Vergleichen von zwei Adressen erfordert daher ein komplexes Verfahren statt eines einfachen Vergleichs von Zeichenfolgen.
2. Die Adressstandardisierung könnte die Lösung für die oben genannte Herausforderung sein. Allerdings ist es sehr aufwendig, alle Schreibweisen in einem „Adresswörterbuch" zu sammeln. Außerdem könnte eine fehlerhafte Kombination der Adresskomponenten zu Konflikten führen, wenn zum Beispiel Alt-Friedrichsfelde 60, 10315 mit einer ungültigen Postleitzahl 10317 geführt wird. In diesem Fall könnte die Geokodierung mittels einer raumbezogenen Abfrage den Fehler erkennen.
3. Die Geokodierung ermöglicht komplexe Funktionen bei der Verknüpfung von räumlichen Datensätzen. Dabei könnten potenzielle Nachbarschaften, Distanzinterpretationen von zwei Adressen, historische Daten, Straßenumbenennungen etc. berücksichtigt werden. Infolgedessen hängt die Genauigkeit der geokodierten Ergebnisse davon ab, welche Datenquelle als Referenzdaten zur Verfügung steht.

Seit Juni 2018 steht der RBS-Geocoder im AfS für interne Zwecke zur Verfügung. Die aktuelle Version lautet Beta 3.4. Die Aktualisierung der Datengrundlagen erfolgt monatlich. Der RBS-Geocoder enthält vielfältige Daten, wie unter anderem Kfz- und SGB-

b | Ablauf der Geokodierung im RBS-Geocoder



II-Daten, Baufertigstellungen, Unternehmensregisterdaten, Wahllokale und Spielhallen. Etwa 270 000 Firmenadressen aus dem Berliner Unternehmensregister werden regelmäßig über den RBS-Geocoder geokodiert und danach für die Qualitätsprüfung und statistische Analyse verwendet. Das Ergebnis liegt im Durchschnitt bei 99,8 % gefundenen Adressen. Die errechnete mittlere Laufzeit beträgt 12,5 Minuten. Abbildung c zeigt, dass Qualität und Plausibilität der Geokodierung im Zeitraum von November 2018 bis Oktober 2019 nahezu konstant geblieben sind. Auch für große Datenmengen, wie die Kfz-Daten mit 1,5 Mill. Adressen, hat sich der RBS-Geocoder bewährt und liefert zuverlässige Ergebnisse.

Fazit

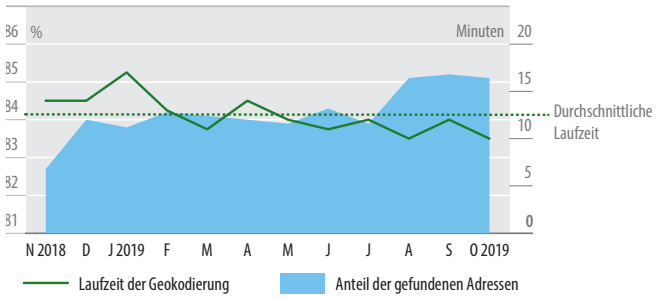
Die Geokodierung ist ein viel diskutiertes Thema. Sie gewinnt nicht nur in der Forschung zunehmend an Aufmerksamkeit, sondern wird auch in § 14 des E-Government-Gesetzes[2] ausdrücklich gefordert. Der Geoservice des AfS pflegt derzeit fast 400 000 Berliner Adressen sowie umfangreiche historische Daten. Auf Basis dieser vollständigen und amtlichen Referenzdaten wurde der RBS-Geocoder im AfS für die Geokodierung entwickelt. Folgende Kernpunkte und Funktionen sind im RBS-Geocoder berücksichtigt:

1. Vielfältige Informationen zu den Anschriften: Das Ergebnis liefert nicht nur die zu den Anschriften zugehörigen Geokoordinaten, sondern zusätzlich die raumbezogenen Informationen (zum Beispiel Block, LOR, LAEA-Gitterzelle usw.).
2. Berücksichtigung der historischen Adresse und von Objekten: Die gesuchte Adresse kann einer historischen RBS-Adresse sowie einem Bahnhof, Flughafen oder Platz zugeordnet werden.
3. Unterschiedliche Referenzdatenstände stehen zur Verfügung: Die Monatsstände werden rückwirkend bis Juni 2018 angeboten und regelmäßig am Monatsanfang für den vergangenen Monat aktualisiert.
4. Nachbearbeitung über Whitelist und Blacklist: Die gesuchte Adresse kann über individuelle Referenzlisten einer aktuellen RBS-Adresse zugeordnet werden. Damit haben Nutzende die Möglichkeit, die Ergebnisse zu verbessern.
5. Betrachtung der Adressenbesonderheit: Für die gebräuchliche Zickzack-Hausnummerierung in Berlin unterscheidet der Algorithmus zwischen geraden, ungeraden und fortlaufenden Hausnummern.
6. Nutzerorientierung: Über eine Nutzerverwaltung werden Whitelist, Blacklist und Ausgabevorlage Fachbereichen zugeordnet und gespeichert.
7. Nachvollziehbares Ergebnisprotokoll: Alle Eingabeparameter, Daten der Nutzenden, Ausführungsdatum und -uhrzeit, die Gesamtlaufzeit, Ausgabepfad sowie die Ergebnisse mit einer statistischen Information werden protokolliert.

Bisher wurde ein Geokodierungsprozess für alle Berliner Adressen aufgebaut und für die praktische Anwendung der RBS-Geocoder realisiert. Als nächste Schritte sind eine Visualisierungsfunktion zur Darstellung und Verortung, eine Weiterentwicklung zur Web-Anwendung und die Erweiterung auf Brandenburger Adressen geplant. Anschließend kann der RBS-Geocoder auch externen Kunden zur Verfügung gestellt werden.

Renee Lin ist Referentin im *Geoservice* des Amtes für Statistik Berlin-Brandenburg.

c | Anteil der gefundenen Adressen und Laufzeit der Geokodierung im RBS-Geocoder



2 Gesetz zur Förderung der elektronischen Verwaltung (E-Government-Gesetz – EGovG) vom 25. Juli 2013 (BGBl. I S. 2749), das zuletzt durch Artikel 15 des Gesetzes vom 20. November 2019 (BGBl. I S. 1626) geändert worden ist.

Literaturverzeichnis

- Goldberg, D. W.; Swift, J. N.; Wilson, J. P. (2008): Geocoding Best Practices: Reference Data, Input Data and Feature Matching. Reaserchgate. Online unter: https://www.researchgate.net/publication/239919647_Geocoding_Best_Practices_Reference_Data_Input_Data_and_Feature_Matching [Abgerufen am 26.02.2020].
- Jacobs, St. (2016): Wie das Chaos bei den Berliner Hausnummern entstand, Online unter: https://www.tagesspiegel.de/berlin/datenanalyse-wie-das-chaos-bei-den-berliner-hausnummern-entstand/13426854.html [Abgerufen am 26.02.2020].
- McDonald, Y. J.; Schwind, M.; Goldberg, D. W.; Lampley, A.; Wheeler, C. M.(2017): An analysis of the process and results of manual geocode correction. Geospat Health; 12(1): 526. doi:10.4081/gh.2017.526.
- Schulte, B.; Lippmann, F.; Schweikart, J. (2010): Geokodierung mit Webkartendiensten – Möglichkeiten, Unterschiede und Grenzen. In: Strobl, J., Blaschke, T.; Griesebner, G. (Hrsg): Angewandte Geoinformatik 2010. Beiträge zum 21. AGIT-Symposium Salzburg. Heidelberg: Wichmann, 773–778.
- Xu, S.; Flexner, S.; Carvalho, V. (2012): Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data. In conjunction with the seventh International Conference on Geographic Information Science 2012 (GIScience 2012).