Bild: stock.adobe.com_sosnovskaya

„OpenStreetMap“ und die Frage nach der Vollständigkeit

Vom Modell zur Datenvollständigkeit bei Nahverkehrs-Busstrecken

Auf die Frage: „Wie vollständig sind die Daten?“ antwortet „OpenStreetMap – Deutschland“ gleich mit einer Gegenfrage: „Was ist „vollständig“? Alle Autobahnen und Bundesstraßen? Oder alle Radwege und Briefkästen? Jede einzelne Hausnummer und jede Parkbank? – Der Detaillierungsgrad der OSM-Daten ist regional sehr unterschiedlich.“ Und weiter lautet die Antwort: „In vielen Städten sind wir schon besser als die meisten proprietären Karten – aber anderswo ist bei uns ein weißer Fleck oder nur eine Durchgangsstraße, wo eigentlich ein ganzer Ort hingehört. Jeder, der unsere Daten einsetzen will, muss sich selbst ein Bild davon machen, ob sie für den anvisierten Zweck ‚vollständig genug‘ sind“ [1].

Autor: **Oliver Fritz**

Damit zeigt sich schon von Haus aus: Die Qualität der Geodaten auf „OpenStreetMap“ (OSM) zeichnet sich durch Heterogenität aus. Die Krux: Verlässliche Referenzdaten sind nur stellenweise verfügbar. Ergebnisse herkömmlicher Ansätze zur Qualitätsbewertung, die auf dem Vergleich mit Referenzdaten beruhen, können aufgrund dieser Heterogenität nicht auf den gesamten Datenbestand übertragen werden. Um dennoch ein Bild der Vollständigkeit von OSM-Daten zu erhalten, können „Referenzdaten“ über ein Regressionsmodell erzeugt werden. Ein Beispiel der Anwen-

dung sind Nahverkehrs-Busstrecken. So wird auf Basis demografischer und sozioökonomischer Indikatoren sowie sporadisch verfügbarer GTFS-Daten die Anzahl von Nahverkehrs-Busstrecken auf die Zellen eines globalen Hexagon-Rasters vorhergesagt, um die Ergebnisse mit dem OSM-Datenbestand abzugleichen.

Das Regressionsmodell

Auf Grundlage der nicht flächendeckend verfügbaren GTFS-Daten und der in globalen Rastern verfügbaren demografischen und sozioökonomischen Indikatoren wurde ein Modell zur Vorhersage der realen weltweiten Verbreitung von Busstrecken entwickelt. Damit sollten die großen Lücken in der Verfügbarkeit von Referenzdaten durch Verwendung von vorhergesagten Werten geschlossen und eine Vollständigkeitsanalyse für den globalen Datenbestand ermöglicht werden. Für 3 438 ISEA3-Hexagone der Auflösungsstufe 10 konnte die Anzahl der Busstrecken aus GTFS-Daten extrahiert werden. In drei Fällen fehlt der Wert mindestens einer der unabhängigen Variablen. Die verbliebenen 3 435 Hexagone (ca. 1,9 % aller Hexagone mit Landanteil) wurden in einen Trainingsdatensatz (n = 2 750; 80 %) und einen Testdatensatz (n = 685; 20 %) aufgeteilt. Es wurden fünf unterschiedliche Regressionsmodelle zur Vorhersage der Anzahl der Busrouten angepasst. Zwei einfache generalisierte lineare Modelle jeweils für eine Quasi-Poisson- (GLM-QP) und eine negative Binomial-Verteilung (GLM-NB) in der Antwortvariable dienen als Benchmark. Um den möglichen Effekt der Kollinearität der Regressoren zu minimieren, wurde ein regularisiertes generalisiertes Modell angepasst. In den vorgenannten Fällen wurden die Regressoren einer Yeo-Johnson-Power-Transformation unterzogen, um sie der Normalverteilung anzunähern. Weder das generalisierte additive Modell noch das Random-Forest-Modell benötigen diese Transformation. Beim GAM-Modell ist die Antwortvariable linear von Glättungsfunktionen der Regressoren abhängig. RF-Modelle beruhen auf Ensembles aus einer Vielzahl von Entscheidungsbäumen. Sie setzen nicht die Linearität der Variablenbeziehungen voraus, sind robust, aber weniger unmittelbar interpretierbar als lineare Regressionsmodelle. Die Anpassung der Modelle erfolgte jeweils durch Optimierung der Wurzel der mittleren Fehlerquadratsumme (RMSE) im Rahmen einer fünfmal wiederholten zehnfachen Kreuzvalidierung (CV). Im Anschluss wurden die Modelle durch Vorhersage auf die Testdaten geprüft.

Modell	Hyperparameter	RMSE (in-sample CV)	RMSE (out-of-sample)	RMSLE (out-of-sample)
GLM-QP	–	140,80	181,59	1,50
GLM-NB	–	154,21	182,63	1,47
GLMNET	alpha = 1 lambda = 0,1	141,07	187,61	1,73
GAM	select = TRUE method = GCV.Cp	138,58	187,86	–
RF	mtry = 2 splitrule = extratrees min.node.size = 5	134,17	185,67	1,42

Tabelle 1: Zusammenfassung der Regressionsmodelle

ÖPNV-Infrastruktur und Bevölkerungszahl, Urbanität und Wirtschaftsleistung

Es bestehen jeweils statistisch signifikante und moderat starke Korrelationen zwischen den unabhängigen und der abhängigen Variable, welche die Hypothese eines Zusammenhangs der ÖPNV-Infrastruktur mit Bevölkerungszahl, Urbanität und Wirtschaftsleistung untermauern. Die Beziehungen erscheinen nach logarithmischer Transformation als annähernd linear. Die starke Korrelation der unabhängigen Variablen untereinander ist sachlich plausibel und auch dadurch bedingt, dass diese Datensätze ihrerseits Ergebnisse von Modellen sind, in die meist mehrere der jeweils anderen Variablen eingeflossen sind.

Tabelle 1 gibt eine Zusammenfassung der Ergebnisse der Regressionsmodelle wieder. Aufgrund der stark rechtsschiefen Verteilung der Antwortvariablen ist der RMSE-Wert der Modelle allein wenig aussagekräftig, da er übermäßig durch große absolute Abweichungen der Vorhersagewerte in Gebieten mit ungewöhnlich vielen Busstrecken beeinflusst wird. Als Grundlage für die Vorhersage der Anzahl der Busstrecken wird daher das RF-Modell ausgewählt, das bessere Ergebnisse in der Wurzel der mittleren logarithmischen Feh-

Abbildung der Heterogenität

Um die Heterogenität der Objektvollständigkeit im globalen Datenbestand angemessen abbilden zu können, wurde ein globales Raster in der „Icosahedral Snyder Equal Area Aperture 3 Hexagon"-Projektion der Auflösungsstufe 10 erstellt. Das zerlegt die Erdoberfläche in 590 492 Hexagone. Die resultierenden Hexagone sind mit einer Fläche von je 863,80 km² gleich groß. Ermittelte Zählwerte können daher ohne weitere Normalisierung verglichen werden. Die Hexagon-Größe der gewählten Auflösungsstufe entspricht am ehesten der durchschnittlichen Fläche eines Busnetzwerks in den extrahierten Daten der General Transit Feed Specification (GTFS). Für die Vorhersage der Anzahl der Busstrecken wurden nur Hexagone mit Landanteil (Anzahl: 180 978) berücksichtigt.

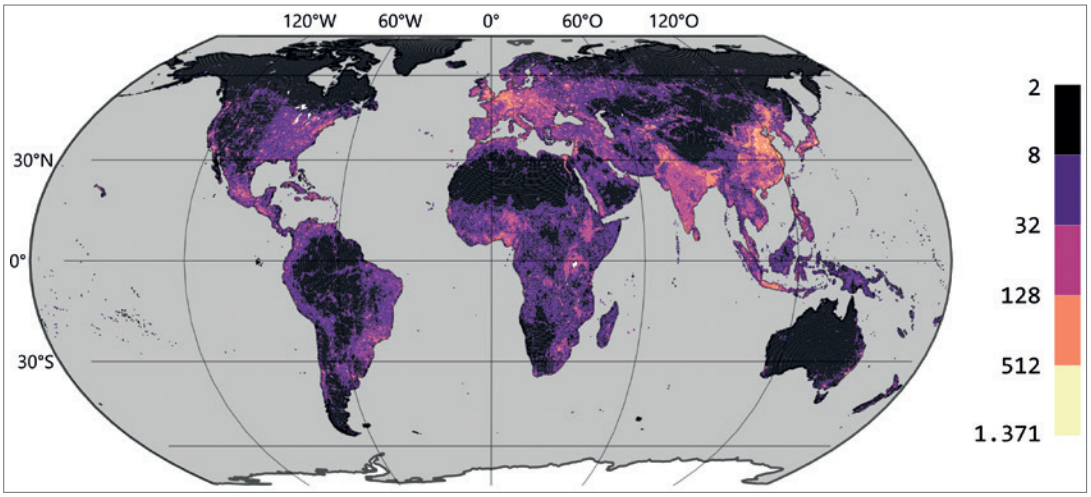

Bild: O. Fritz

Abb. 1: Vorhergesagte Anzahl der Busstrecken je ISEA3-Hexagon

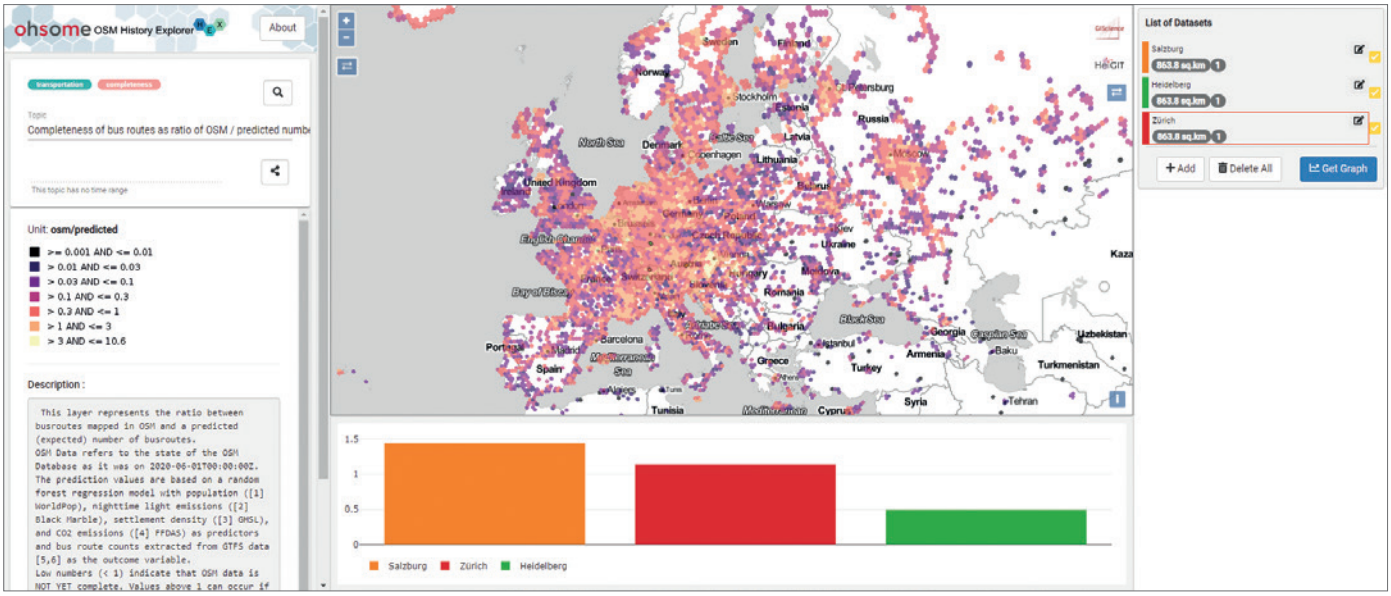

Bild: „ohsome History Explorer"

Abb. 2: Visualisierung des Verhältnisses der OSM-Busstrecken zur vorhergesagten Anzahl

lerquadratsumme (RMSLE) erreicht, die Aufschluss über den relativen Vorhersagefehler gibt.

Auf Grundlage des ausgewählten Regressionsmodells kann die Anzahl der Busstrecken je ISEA3-Hexagon vorhergesagt werden (siehe Abb. 1). Zu beachten ist, dass sich wegen der fehlenden Gebiete ohne ÖPNV in den Trainingsdaten eine Mindestanzahl von zwei Busstrecken ergibt. Für die Ableitung von Aussagen zur Vollständigkeit im OSM-Datenbestand bedeutet dies, dass diese gerade in Gebieten, in denen sie aufgrund der Nichtexistenz entsprechender Objekte in der Realität erreicht ist, stark unterschätzt wird.

Nach Erzeugung der flächendeckenden Referenzdaten kann die Vollständigkeit des OSM-Datenbestands beurteilt werden, indem die Anzahl der OSM-Objekte mit der vorhergesagten Objektzahl abgeglichen wird. Neben der einfachen Differenzbildung bietet sich die Ermittlung des Verhältnisses zwischen OSM-Datenbestand und Vorhersage für den Vergleich an, zumal dies die relevantere relative Abweichung wiedergibt. Die Ergebnisse können in der Webapplikation „ohsome History Explorer" (ohsomeHeX) [2] erkundet werden (Abb. 2). Die interaktive Visualisierung vermag einen Überblick darüber zu geben, wo und in welchem Maße Lücken im OSM-Datenbestand anzunehmen sind. Die Ergebnisse stehen so zur Qualitätseinschätzung und zur Planung gezielter Datenerfassung durch die Gemeinschaft der OSM-Mitwirkenden zur Verfügung.

Quellen:

[1] www.openstreetmap.de/faq.html
[2] ohsome.org/apps/osm-history-explorer

Kontakt:

Oliver Fritz
Heidelberg Institute for Geoinformation Technology (HeiGIT)
E: oliver.fritz@heigit.org

Wichmann

gis.Radio

hier gibts Geo-IT aufs Ohr!

Der Geo-IT-Podcast.

Immer hintergründig, immer aktuell, mit Beiträgen, Reportagen und Interviews.

Jetzt reinhören:
www.gispoint.de/gisradio

www.gispoint.de/gisradio

Bild: adobestock_9dreamstudio