LIfBi WORKING PAPERS

Jonas Klingwort, Rainer Schnell und Michaela Sixt

GEO-MASKING VON KOORDINATEN DER BILO BEFRAGTEN FÜR ZUKÜNFTIGE DATENSCHUTZGERECHTE DISTANZ-BERECHNUNGEN

LIfBi Working Paper No. 87
Bamberg, März 2020

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE e.V.

Working Papers of the Leibniz Institute for Educational Trajectories (LIfBi)
at the University of Bamberg

The LIfBi Working Papers series publishes articles, expert reports, and findings related to data collected and studies conducted at the Leibniz Institute for Educational Trajectories—first and foremost, the National Educational Panel Study (NEPS) in Germany.

LIfBi Working Papers are edited by the LIfBi Board of Directors and the Heads of the LIfBi Departments. The series started in 2011 under the name "NEPS Working Papers" and was renamed in 2017 to broaden the range of studies which may be published here.

Papers appear in this series as work in progress and may also appear elsewhere. They often present preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the LIfBi management or the NEPS Consortium.

The LIfBi Working Papers are available at www.lifbi.de (see section "Institute > Publications"). LIfBi Working Papers based on NEPS data are also available at www.neps-data.de (see section "Data Center > Publications").

Contact:
Leibniz Institute for Educational Trajectories
Wilhelmsplatz 3
96047 Bamberg
Germany
contact@lifbi.de

Geo-Masking von Koordinaten der BiLO Befragten für zukünftige datenschutzgerechte Distanzberechnungen

Jonas Klingwort, Universität Duisburg-Essen
Rainer Schnell, Universität Duisburg-Essen
Michaela Sixt, Leibniz-Institut für Bildungsverläufe e.V. (LIfBi)

E-Mail-Adresse:

jonas.klingwort@uni-due.de

Bibliographische Angabe:

Klingwort, J., Schnell, R. & Sixt, M. (2020). *Geo-Masking von Koordinaten der BiLO Befragten für zukünftige datenschutzgerechte Distanzberechnungen* (LIfBi Working Paper No. 87). Bamberg, Deutschland: Leibniz-Institut für Bildungsverläufe.

Geo-Masking von Koordinaten der BiLO Befragten für zukünftige datenschutzgerechte Distanzberechnungen

Zusammenfassung

Die räumliche Distanz zu Bildungsangeboten ist eine relevante Information im Bereich der Bildungsforschung. Allerdings gelten Informationen über personenbezogene Geo-Koordinaten als sensitive Daten und müssen oftmals nach Projektende gelöscht werden. Potenzielle Distanzberechnungen zur Beantwortung relevanter Forschungsfragen sind somit nur zeitlich begrenzt möglich. In diesem Papier verwenden wir eine neue Methode zur Berechnung der approximierten Entfernungen zwischen anonymisierten räumlichen Daten, wobei die Koordinaten der Wohnorte der BiLO Befragten und der Adressen von Bildungseinrichtungen verwendet wurden. Zur Approximation dieser Entfernungen wurden sich überschneidende Sets von Gitterpunkten (ISGP) verwendet. Die Methode ist ein neues Verfahren zur Sicherung des Datenschutzes bei der Verarbeitung von georeferenzierten Datensätzen. Die in dieser Anwendung anonymisierten Koordinaten der BiLO Befragten können in zukünftigen Studien für approximative Distanzberechnungen verwendet werden.

Schlagworte

Geographische Daten, geo-referenzierte Daten, record-linkage, ISGP

Abstract

The geographical distance to educational facilities is a relevant information in the field of educational research. However, information on personal geo-coordinates is considered sensitive data and often must be deleted after the end of the project. Potential distance calculations to answer relevant research questions are therefore only possible for a limited period of time. In this paper we use a new method to calculate the approximate distances between anonymized spatial data using the coordinates of BiLO respondents' places of residence and the addresses of educational institutions. Overlapping sets of grid points (ISGP) were used to approximate these distances. The method is a new procedure to guarantee data privacy when processing georeferenced data sets. The anonymized coordinates of the BiLO respondents can be used for approximate distance calculations in future studies.

Keywords

Geographical data, geo-referenced data, record-linkage, ISGP

1. Einleitung

Das vorliegende Arbeitspapier beschreibt die Arbeits- und Analyseschritte im Projekt „Anonymisierung von Geo-Koordinaten (Ano-Geo)“, welches im Auftrag des Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) durch Prof. Dr. Rainer Schnell und Jonas Klingwort in Zusammenarbeit mit Dr. Michaela Sixt durchgeführt wurde. Im Rahmen dieses Projekts wurden Geo-Koordinaten, die im Rahmen des Projekts „BildungsLandschaft Oberfranken (BiLO)“ erhoben wurden anonymisiert, um zukünftige Distanzberechnungen unter Sicherung des Datenschutzes zu ermöglichen.

Dazu wird im Folgenden zunächst das BiLO Projekt skizziert, um den inhaltlichen Hintergrund zu den verwendeten Daten zu erläutern (Abschnitt 2). Anschließend wird in Abschnitt 3 das Projekt „Ano-Geo“ und dessen Ziel erläutert. In Abschnitt 4 werden die Daten, sowie deren Aufbereitung beschrieben. Die verwendete Methode wird in Abschnitt 5 kurz vorgestellt, wobei zu technischen Details auf weiterführende Literatur verwiesen wird. In Abschnitt 6 und 7 werden die zentralen Analyseschritte im Projekt „Ano-Geo“ und deren Ergebnisse beschrieben. Das Arbeitspapier schließt in Abschnitt 8 mit einem Ausblick auf zukünftige Anwendungen.

2. Das Projekt „BildungsLandschaft Oberfranken“ (BiLO)

Die Bedeutung regionaler Disparitäten beim Zugang zu Bildungschancen hat angesichts des demographischen Wandels, zurückgehender Bevölkerungszahlen in ländlichen Regionen und damit verknüpftem Infrastrukturabbau (wieder) stärker wissenschaftliche Aufmerksamkeit erfahren. Das von der Oberfranken Stiftung geförderte Projekt „BildungsLandschaft Oberfranken (BiLO)“ (2014-2018) untersucht insbesondere die schichtspezifische Bedeutung regionaler Angebotsstrukturen bei individuellen Bildungsentscheidungen in zentralen Bildungsphasen im Lebensverlauf (Sixt et al., 2017). Dazu war in der ersten Projektphase vorgesehen, zum einen Daten zum lokal verorteten, objektiven Bildungsangebot über Adressrecherchen, Sekundärdaten der amtlichen Statistik und projekteigenen Onlinebefragungen zusammenzutragen. Zum anderen wurden Individualdaten über das subjektiv wahrgenommene Bildungsangebot für anstehende und tatsächlich getroffene Bildungsentscheidungen mit einer auf einer Zufallsstichprobe (Einwohnermeldeamt-stichprobe) basierenden Bevölkerungsbefragung für Oberfranken erhoben. Indem die beiden Datenquellen in der zweiten Projektphase verknüpft wurden, konnten individuelle Distanzmaße zu objektiv verfügbaren und subjektiv wahrgenommenen Bildungsangeboten in die Auswertungen spezifischer Fragestellungen im Bereich der Frühkindlichen Bildung, des Übergangs in die Sekundarstufe, in Ausbildung und Studium, der Erwachsenenbildung und der Kulturellen Bildung einfließen (Sixt et al., 2018).

3. Das Projektziel von „Ano-Geo“

Die im Projekt BiLO erhobenen personenbezogenen Geo-Koordinaten müssen nach Abschluss des Projekts (30.09.2020) aus Gründen des Datenschutzes gelöscht werden. Als Folge dessen sind zukünftige Distanzberechnungen mit den personenbezogenen Geo-Koordinaten der BiLO Befragten nicht mehr möglich. Das Ziel des Projekts „Ano-Geo“ ist es, zukünftige Distanzberechnungen unter Wahrung des Datenschutzes weiterhin zu ermöglichen, obwohl die personenbezogenen Geo-Koordinaten der BiLO Befragten gelöscht wurden.

4. Daten

Die amtlichen Grenzen des Regierungsbezirks Oberfranken sind die geographische Datengrundlage der Analyse. Weiterhin wurden die Geo-Koordinaten der Wohnorte der BiLO Befragten verwendet. Zudem wurden Geo-Koordinaten der Adressen der Bildungseinrichtungen, welche im Rahmen des Projekts BiLO aus verschiedenen Quellen recherchiert bzw. erworben wurden, verwendet. Dabei lag das Augenmerk insbesondere auf Einrichtungen der frühkindlichen Bildung (Frühki; Kindertagestätten, Kindergärten u.ä.), allgemein bildenden Schulen (Grundschulen und Schulen des Sekundarbereichs des Bildungswesens) sowie beruflichen Schulen (Berufsfachschulen und Schulen des Gesundheitswesens). Die Aufbereitung der Geo-Koordinaten der Befragten und Institutionen wird im Folgenden beschrieben.

4.1 Datenaufbereitung

Der Datensatz mit Geo-Koordinaten der BiLO Befragten enthält 8.672 Befragte. Die Analyse ergab, dass 25 Befragte fehlende Werte für die Geo-Koordinaten aufwiesen. Weitere 27 Befragte hatten Geo-Koordinaten, welche außerhalb von Oberfranken liegen. Da die BiLO Grundgesamtheit als die Wohnbevölkerung von Oberfranken definiert ist, wurden diese Befragte für die vorliegende Analyse aus dem Datensatz entfernt, sodass 8.620 Befragte verbleiben. Abbildung 1 zeigt die amtlichen Grenzen von Oberfranken und eine grobe Visualisierung der Koordinaten der BiLO Befragten, die keine Rückschlüsse auf den Adressstandort der Befragten zulässt. Der Datensatz mit Geo-Koordinaten der Institutionen enthält 6.259 Institutionen. Die Institutionen sind in folgende Gruppen unterteilt:

- G1: FrühKi & Grundschulen, n=3.040
- G2: Sekundarschulen, n=2.253
- G3: Berufliche Schulen (Berufsfachschulen und Schulen des Gesundheitswesens), n=966

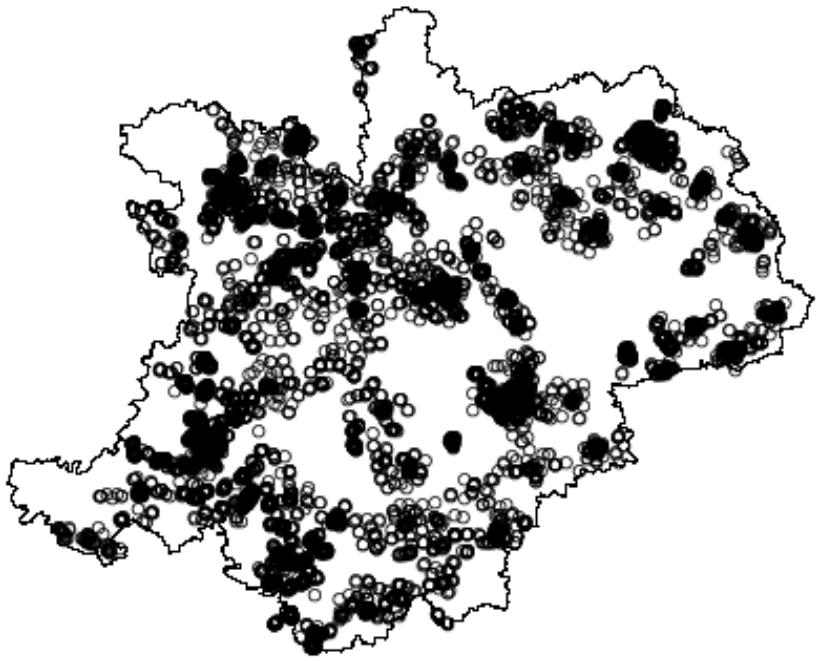

Abbildung 1. BiLO Koordinaten in Oberfranken.

Im nächsten Schritt werden die geographischen Grenzen von Oberfranken durch künstlich (wesentlich erweiterte) geographische Grenzen ersetzt. Dies ist notwendig um Institutionen die gegenwärtig, aber auch bei zukünftigen Berechnungen, außerhalb von Oberfranken liegen in der Analyse verwenden zu können. Darüber hinaus wird dadurch die Qualität der Approximationen von Koordinaten die nah an der ursprünglich geographischen Grenze liegen verbessert, da so gewährleistet wird, dass die Radii, die um die Punkte gelegt werden, ausreichend Punkte beinhalten (siehe dazu Abschnitt 5). Die geographische Struktur von Oberfranken und die relative Anordnung aller enthaltenen Koordinaten wird durch diesen Schritt nicht verändert. Abbildung 2 zeigt die künstliche Erweiterung der geographischen Grenzen in Form eines Rechtecks, sowie die Koordinaten der Befragten (in schwarz), von G1 (in rot), von G2 (in grün) und von G3 (in blau).

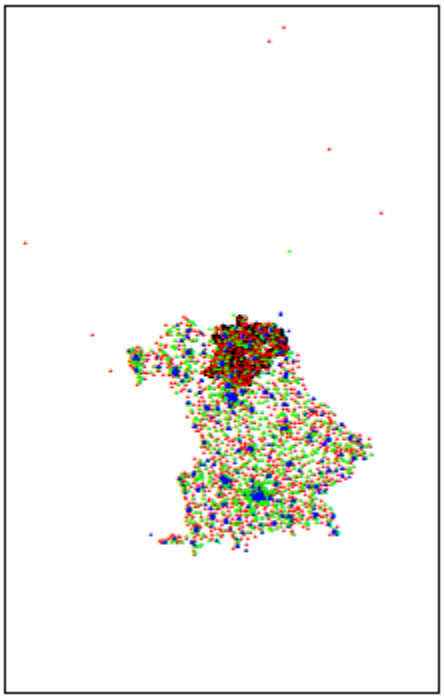

Abbildung 2. BiLO Koordinaten in Oberfranken (in schwarz) und Koordinaten der Institutionen G1 (in rot), G2 (in grün) und G3 (in blau).

5. Beschreibung der Methode

Die verwendete Methode basiert auf einer Idee aus einem unpublizierten Vortrag von Farrow (2014), die von Farrow (2017) patentiert wurde. Mittlerweile wurde die Idee von Schnell & Klingwort (2017) und von Farrow et al. (2020) weiterentwickelt. Die Methode approximiert die Distanz zwischen zwei räumlichen Punkten in einem zweidimensionalen Raum, ohne dabei Informationen über die exakte Lage der Punkte zu verwenden. Dabei werden 'intersecting sets of grid points' (ISGP) verwendet, wobei ISGP die Gitterpunkte innerhalb der Schnittfläche zweier Radii sind, welche jeweils um zwei Punkte (in diesem Fall personenbezogene Geo-Koordinaten der BiLO-Befragten) gelegt werden. Diesen Punkten werden Zufallszahlen zugewiesen. Es werden ausschließlich diese Zufallszahlen gespeichert und keine personenbezogenen Koordinaten. Dadurch werden die Werte in der Matrix zensiert, da nur Distanzen die kleiner als zweimal der gewählte Radius sind approximiert werden. Distanzen die größer sind als zweimal der gewählte Radius werden durch eine Konstante, nämlich das Doppelte des Radius, ersetzt. Weitere technische Details zu der Methode werden in diesem

Arbeitspapier nicht beschrieben. Hierzu verweisen wir auf die bereits genannte weiterführende Literatur.

6. Ermittlung des optimalen Radius getrennt nach Institutionen

Im Folgenden wird der optimale Radius zur Approximation der Distanzen getrennt für alle drei Gruppen ermittelt. Tabelle 1 stellt die Parameter der Simulation dar.

Tabelle 1: Parameter der Simulationen zur Bestimmung des optimalen Radius, getrennt nach Gruppe

Gruppe	Stichprobengröße Befragte	Stichprobengröße Institutionen	Anzahl Gitterpunkte	Radius-set (m)
G1	5% (n=431)	5% (n=152)	29.866	(5.000,20.000,1.000)
G2	5% (n=431)	5% (n=112)	29.866	(25.000,42.000,1.000)
G3	5% (n=431)	5% (n=96)	29.866	(30.000,5.0000,1.000)

Für alle Gruppen wurden jeweils 5%-Stichproben aus den Befragten und den Institutionen gezogen. Für die Institutionen von G3 wurde aufgrund der kleinen Gruppengröße eine 10% Stichprobe gezogen. Die Stichproben der Befragten sind identisch, da der identische Startwert des Zufallszahlengenerators verwendet wurde. Aufgrund der Form und Größe der darzustellenden Fläche werden 29.866 Gitterpunkte in die Fläche gelegt. Die Schrittweite der Radii zur Ermittlung des optimalen Radius basieren auf den empirischen Radii, die durch das Projektteam von BiLO ermittelt wurden. Die empirisch ermittelten Radii basieren auf Angaben der BiLO Befragten darüber, welche Einrichtungen sie generell kennen. Für die empirisch ermittelten Radii wurden die Distanzen gewählt, die vom Wohnort der Befragten am weitesten entfernt liegen und der Durschnitt gebildet. Dieser beträgt für G1 = 8,5km, G2 = 30km und G3 = 38,9km.

Als optimaler Radius wird jener gewählt, der die geringste absolute mittlere Differenz zwischen approximierter und tatsächlicher Distanz aufweist. Dazu wird zunächst für jeden Befragten (i) in Gruppe (j) die Differenz (Δ_{ij}) zwischen der berechneten Approximation ($\hat{\alpha}_{ij}$) relativ zur tatsächlichen Distanz (α_{ij}) berechnet. Anschließend wird pro Radius die mittlere relative Differenz aller (Δ_{ij}) berechnet. Die berechnete mittlere relative Differenz wird in absolute Werte umgerechnet. Die Ergebnisse der Simulationen sind in den Abbildungen 3 – 5 im folgenden Abschnitt dargestellt.

6.1 Approximationen und tatsächliche Distanzen

Die Abbildungen 3 – 5 zeigen die Ergebnisse der Simulationen getrennt nach Größe des Radius und Institution. Dabei werden die approximierten und die tatsächlichen Distanzen (in m) dargestellt. Deutlich sichtbar ist der starke Zusammenhang zwischen den approximierten und tatsächlichen Distanzen, unabhängig von der Institution. Die Konstanten in allen drei Abbildungen kommen dadurch zustande, dass oberhalb eines Schwellenwertes die Approximation auf zweimal Radius gesetzt wird (siehe Abschnitt 5).

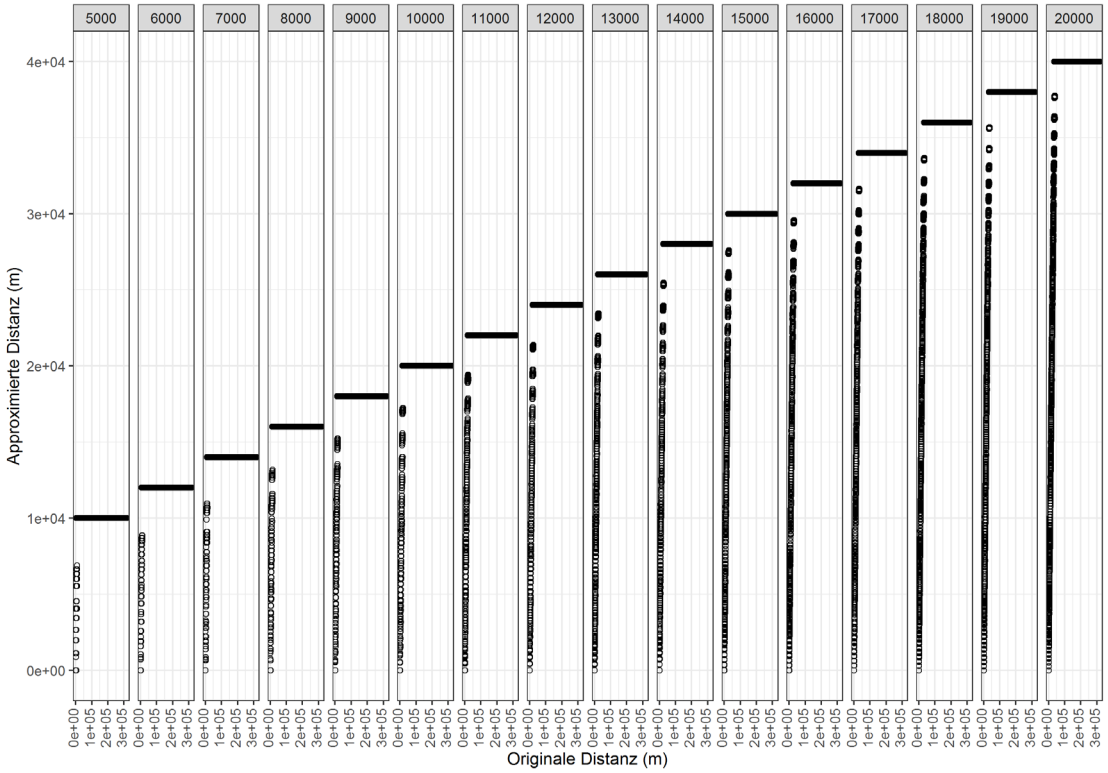

Abbildung 3. Tatsächliche und approximierte Distanzen für G1. Fälle > 2r sind enthalten.

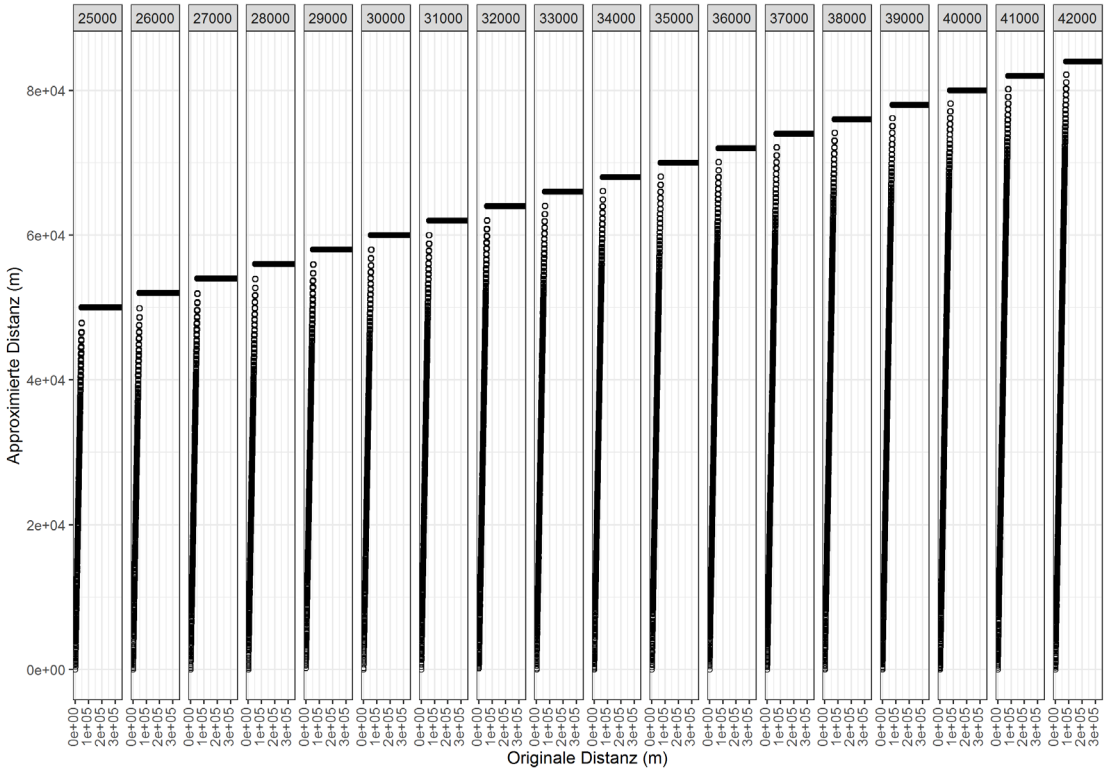

Abbildung 4. Tatsächliche und approximierte Distanzen für G2. Fälle > 2r sind enthalten.

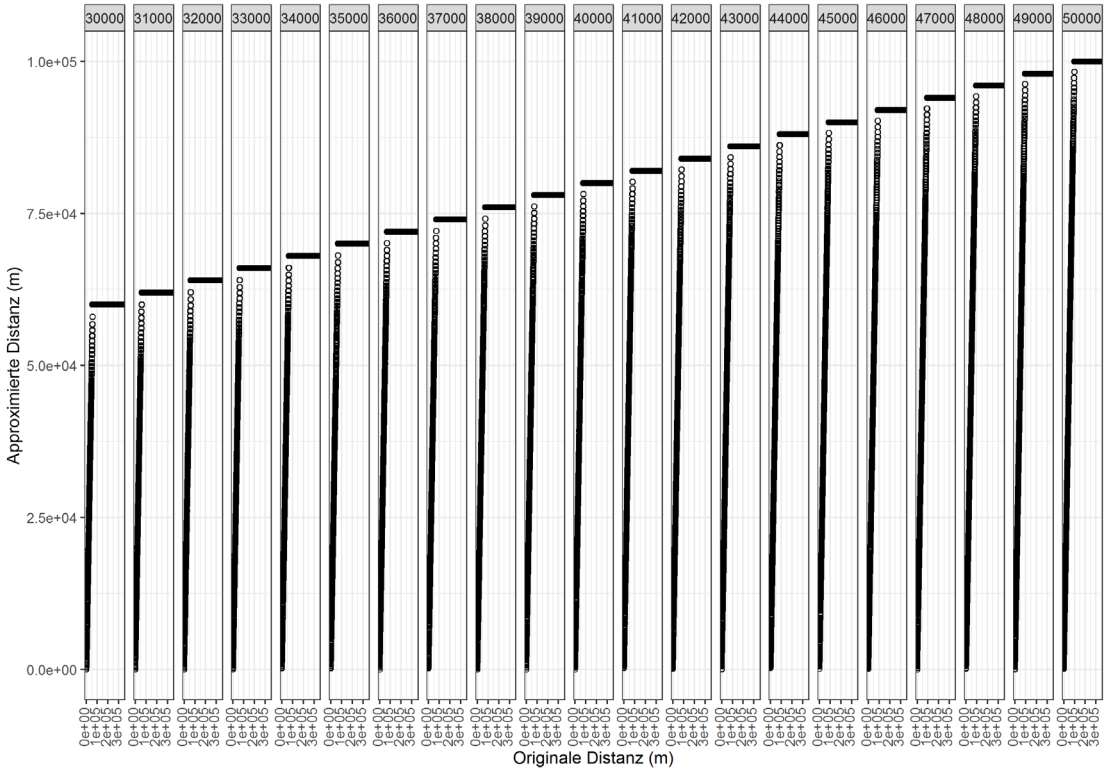

Abbildung 5. Tatsächliche und approximierte Distanzen für G3. Fälle > 2r sind enthalten.

6.2 Mittlerer absoluter relativer Fehler getrennt nach Radius

Die ermittelten optimalen Radii sind für G1 11km, für G2 42km und für G3 43km, was jeweils inhaltlich sinnvollen Werten entspricht. Nachfolgen sind die relativen Fehler nach gewähltem Radius zur Ermittlung des optimalen Radius gruppenweise in den Abbildungen 6 – 8 dargestellt. Generell sind die Unterschiede zwischen tatsächlicher und approximierter Distanz zwischen den gewählten Radii minimal, sodass mit jedem Radius aus dem Set an gewählten Radii akkurate Approximationen möglich wären.

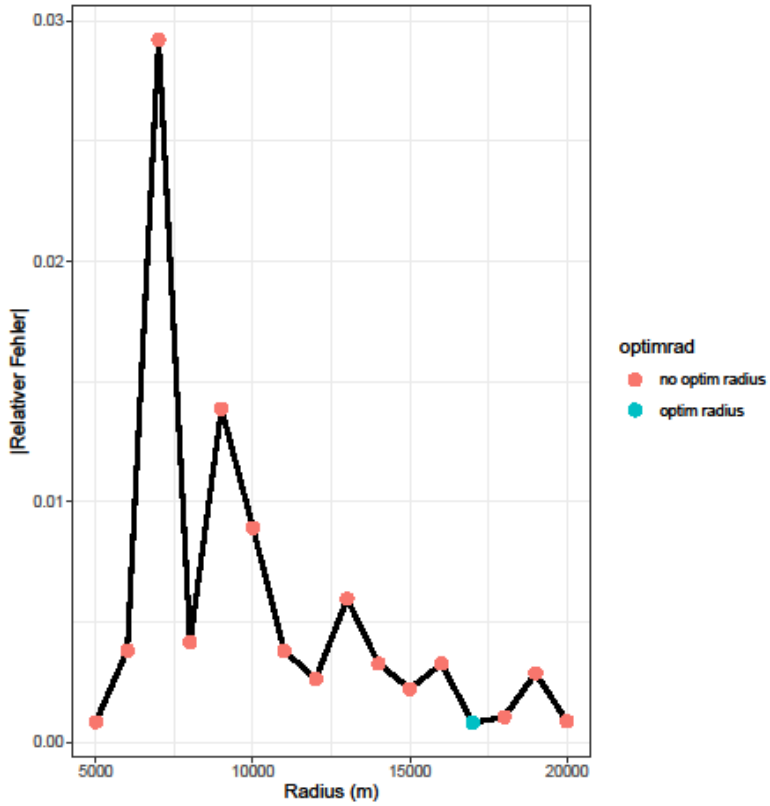

Abbildung 6. Mittlerer absoluter relativer Fehler getrennt nach Radius für Gruppe 1. Fälle > 2r sind nicht enthalten.

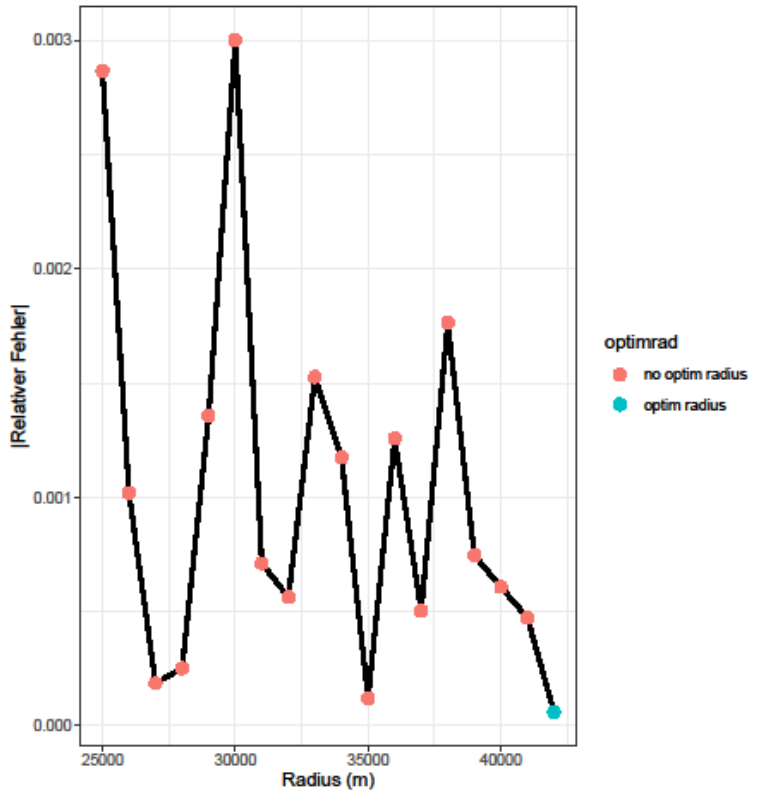

Abbildung 7. Mittlerer absoluter relativer Fehler getrennt nach Radius für Gruppe 2. Fälle > 2r sind nicht enthalten.

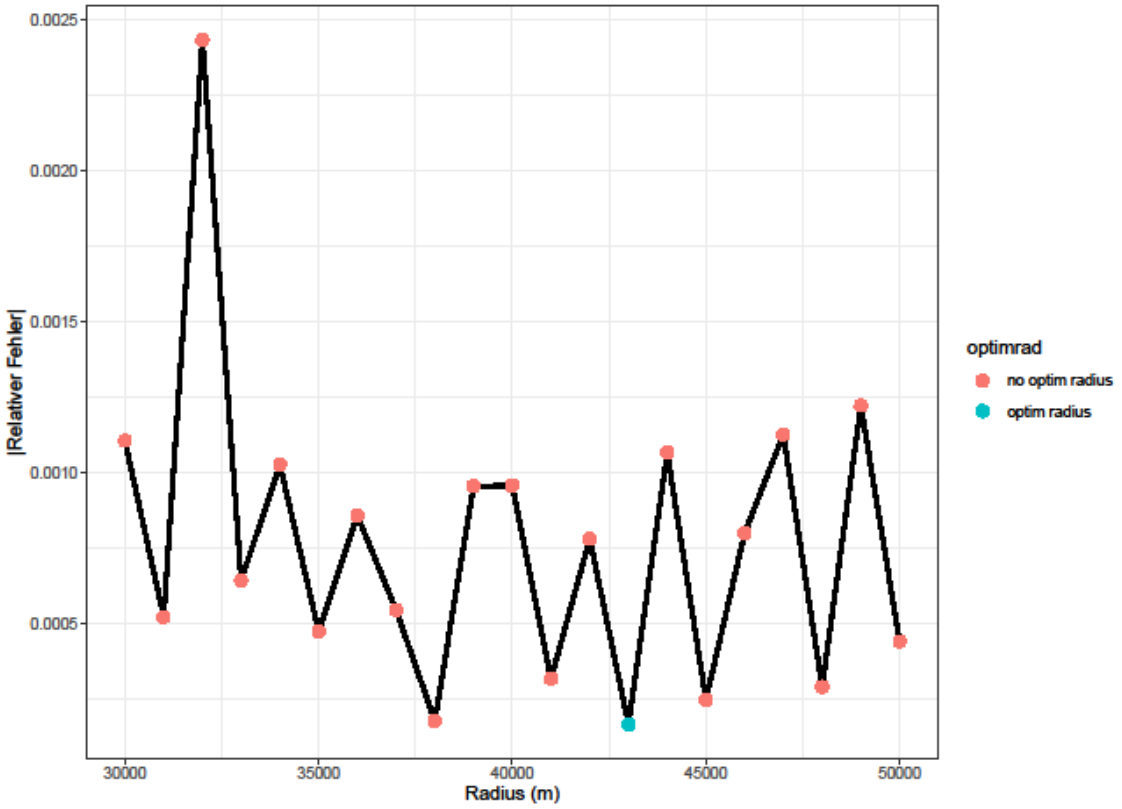

Abbildung 8. Mittlerer absoluter relativer Fehler getrennt nach Radius für Gruppe 3. Fälle > 2r sind nicht enthalten.

6.3 Bestimmung der Hash-Werte für die Koordinaten der BiLO Befragten

Unter Verwendung der zuvor ermittelten optimalen Radii wurden die Hash-Werte der Befragten, getrennt nach G1, G2 und G3, erstellt. Bei den Hash-Werten handelt es sich um Zufallszahlen, die innerhalb der Schnittfläche von zwei Radii zweier Punkte liegen (siehe dazu Abschnitt 5). Eine Replikation dieser Zahlen ist ausschließlich dem LIfBi möglich, da nur das LIfBi über das notwendige Passwort zum Setzen des Startwert für die Erzeugung der Zufallszahlen verfügt. Die im Rahmen dieses Projekts erstellten Hash-Werte können für zukünftige approximative Distanzberechnungen durch das LIfBi verwendet werden.

7. Zukünftige Anwendungen

Mit den erstellten Hash-Werten können zukünftig unter Verwendung der hier beschriebenen Methode Distanzen zu den BiLO Befragten approximiert werden. Ein Anwendungsszenario könnte zum Beispiel sein, dass das LIfBi Geo-Koordinaten weiterer Institutionen erhält. Für diese Geo-Koordinaten müssen ebenfalls Hash-Werte unter Verwendung der geographischen Einheit, der gewählten Radii und des Passworts erstellt werden. Mit den beiden Sets an Hash-Werten können dann die Distanzen zwischen den neuen Institutionen und den BiLO Befragten approximiert werden.

8. Literaturverzeichnis

Farrow, J. (2014). Privacy Preserving Distance-Comparable Geohashing. Second International Health Data Linkage Conference 2014, Vancouver April 2014.

Farrow, J. (2017). Method and System for Comparative Data Analysis. Patent Application Publication (Pub. No.: US 2017/0039222 A1).

Farrow, J., Schnell, R. und Klingwort, J. (2020). Locational Privacy Preserving Distance Computations with Intersecting Sets of Randomly Labelled Grid Points. Universität Duisburg-Essen, Research Methodology Group, Working Paper.

Schnell, R. und Klingwort, J. (2017). Putting People on the Map Without Revealing Their Location: Privacy Preserving Locational Distance Computations. European Survey Research Association (ESRA) 2017, Lissabon, Portugal.

Schnell, R. und Klingwort, J. (2020a). Geomasking with Intersecting Sets of Grid Points (ISGP). Gastvortrag am Leibniz-Institut für Bildungsverläufe e.V (LIfBi); Bamberg, Deutschland.

Schnell, R. und Klingwort, J. (2020b). Privacy Preserving Locational Distance Computations – Application in R. Workshop am Leibniz-Institut für Bildungsverläufe e.V (LIfBi); Bamberg, Deutschland.

Sixt, M., Baur, H.-R., Gerbig, F., Hofmann, J., Müller, D., Stöhr, I., Thürer, S., Zeichner, C. & Bayer, M. (2017). Das Projekt „BildungsLandschaft Oberfranken (BiLO)“ – eine Skizze (LIfBi Working Paper No. 71). Bamberg, Deutschland: Leibniz-Institut für Bildungsverläufe.

Sixt, M., Bayer, M. & Müller, D. (Hrsg.) (2018). Bildungsentscheidungen und lokales Angebot. Die Bedeutung der Infrastruktur für Bildungsentscheidungen im Lebensverlauf. Münster: Waxmann.