Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

Methodological aspects of labour market data

09/2012
EN

Geocoding of German Administrative Data

The Case of the Institute for Employment Research

Theresa Scholz,
Cerstin Rauscher,
Jörg Reiher,
Tobias Bachteler

Bundesagentur für Arbeit

Geocoding of German Administrative Data

The Case of the Institute for Employment Research

Theresa Scholz (Institut für Arbeitsmarkt- und Berufsforschung),
Cerstin Rauscher (Institut für Arbeitsmarkt- und Berufsforschung),
Jörg Reiher (Universität Duisburg-Essen),
Tobias Bachteler (Universität Duisburg-Essen)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract

Economic research addressing spatial questions is often constrained by the availability of data at a sufficiently fine regional scale. For this reason the 2009 cross-section of the Integrated Employment Biographies (IEB) of the Institute for Employment Research was geocoded. The IEB are collected from administrative processes of the Federal Employment Agency and include employees and their employing establishment, unemployed, jobseekers, benefit recipients and participants of employment measures. The geocodes assigned to the IEB serve as a basis to generate small regions that represent neighbourhoods. Additionally, the geocoded data itself allows for the application of research methods at a regional level that can be chosen independent of administrative areas. The corresponding data protection regulations have to be adhered to.

Zusammenfassung

Der Erforschung von wirtschaftswissenschaftlichen Fragestellungen mit räumlichem Bezug wird oft durch Verfügbarkeit von Daten auf kleinräumigem Niveau Grenzen gesetzt. Aus diesem Grund wurde der 2009er Querschnitt der Integrierten Erwerbsbiographien (IEB) des Instituts für Arbeitsmarkt- und Berufsforschung georeferenziert. Die IEB stammen aus administrativen Prozessen der Bundesagentur für Arbeit und umfassen Beschäftigte und deren beschäftigenden Betrieb, Arbeitslose, Arbeitssuchende, Leistungsempfänger und Maßnahmeteilnehmer. Die Punktkoordinaten, die der IEB zugewiesen werden, dienen als Basis zur Generierung kleinräumiger Regionen, die Nachbarschaften darstellen. Zusätzlich erlauben die georeferenzierten Daten selber der Forschung Methoden auf einem regionalen Niveau anzuwenden, das unabhängig von administrativen Gebieten ist. Dabei müssen die entsprechenden Datenschutzbestimmungen beachtet werden.

Keywords: geocoding, administrative data, record linkage

1 Introduction

Empirical labour market research is often constrained by the availability of data at a sufficiently fine regional scale. Spatial analysis based on German administrative data usually cannot go beyond administrative areas like counties or (in few cases) municipalities. One example are models of social interaction that analyze, whether the behaviour and/or composition of a regionally constrained reference group – for example a school or a neighbourhood – affect the behaviour of a single member. A neighbourhood with high unemployment will for instance exchange less information on employment possibilities, and unemployment may be less stigmatized, which in turn may influence the individual labour market outcome of the group members (Manski 1993, Manski 2000).

For the purpose of analyzing these so called neighbourhood effects, however, administrative areas prove to be too coarse. Even postcode areas are only partially useful, since their geographic size as well as the number of inhabitants varies strongly. For this reason data of the Institute for Employment Research (IAB) is geocoded within the context of the project "Neighbourhood Effects: The Analysis of individual-rational behaviour in a social context"[1]. We aim at generating new, small-area regions in order to represent neighbourhoods. The geocoded data itself additionally offers regional labour market research the possibility to apply methods at a regional level that can be freely chosen and is independent of administrative areas.

2 Approach

The generation of the new neighbourhood areas is conducted in two steps.

First, we select all 36.2 Million persons and 2.5 Million establishments contained in the Integrated Employment Biographies (IEB) of the Institute for Employment Research on June 30th 2009.[2] The IEB are collected from administrative processes of the Federal Employment Agency of Germany. The data set comprises all persons registered with the Federal Employment Agency due to the following: (i) employment subject to social security or marginal part-time employment, (ii) receipt of benefits in accordance with Social Code Book II or III, (iii) job search and (iv) planned or actual participation in an employment or training measure. Additionally, for jobholders the employing establishment is included. We link the selected data to geocoded data from the Federal Agency for Cartography and Geodesy (*Georeferenzierte Adressdaten Bund - GAB*[3]) using record linkage techniques. The GAB contains approximately 22 Million addresses of German buildings and the corresponding geographic

[1] The project is financed by the *Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e.V* within the context of the *Pakt für Forschung.* Aside from the IAB, project partners are the Rheinisch-Westfälisches Institut für Wirtschaftsforschung RWI (main applicant) and the German Socio-Economic panel SOEP. The work was also supported by the „German Record Linkage Center", which is funded by the DFG.

[2] The exact numbers amount to 36,241,017 Persons and 2,472,604 establishments.

[3] Geoinformationen © **Bundesamt für Kartographie und Geodäsie** (www.bkg.bund.de)

coordinate. The GAB data was collected between December 2008 and August 2010, hence the cross-section drawn from the IEB should be covered by the GAB. Due to the enormous number of records to be geocoded, execution of the program codes is expected to require several days. In order to keep the runtime to a minimum, we therefore proceed in two sub-steps. At first, the pre-processing of the address material as well as a deterministic linkage is implemented directly within the database management system (DBMS) in the server environment of the German Federal Employment Agency. The second sub-step consists of a distance based matching, which is applied to the remaining records by means of the record linkage software Merge Toolbox (MTB) (Schnell/Bachteler/Reiher 2005) as well as the data-mining software KNIME (Berthold et al. 2007).

Second, the micro data is aggregated to small-area regions on the basis of the geocoded addresses in order to represent neighbourhoods. Geographic grid cells may serve as an alternative representation of neighbourhood areas. In the following, the outlined steps will be described in more detail.

3 Linkage of the IAB data with geocoded address data

3.1 Pre-processing and deterministic linkage

Both the pre-processing and the linkage are conducted using a DBMS[4]. We first turn to the description of the pre-processing of the GAB data. Postal community names are pre-processed by separating them into two parts, where abbreviations for the word “Ortsteil” (part of town) are removed from the second part. The variable street name is prepared for the linkage by substitution of umlauts (ä → ae and so forth); ß is substituted by ss. Subsequently, all characters are converted to upper cases and STRASSE is replaced by STR if it occurs at the end of the string or is separated by a blank. The variable house number is dissected into the actual house number and a house number annex. Finally, all non-ASCII upper case characters are removed from each variable.

The IEB address material is pre-processed differently than the GAB data. This is because the data are collected for different purposes than the GAB and are hence stored differently. Especially, we need to change the preparation of street names, since street name and house number are in part contained within one variable in the IEB. This variable is separated into a leading part up to the first occurring number. The remaining part is saved as house number in a newly generated variable. The already existent house number is replaced by the new one if the newly generated house number is not empty. Then, house number is separated into a first part up to the first non-numeric character and a house number annex containing the residual part. The remaining pre-processing procedures correspond to the ones applied to the GAB.

After the GAB and IEB have been pre-processed, the two data sets are merged on the address information. For this purpose, we use a deterministic linkage, meaning that only strings

[4] Microsoft SQL Server 2008; the program code is SQL.

exactly identical in postal code, place, street and house number are defined to be a match. The deterministic linkage assigns a coordinate to 93.0 percent of all establishments and 89.6 percent of all persons selected from the IEB[5].

3.2 Distance-based linkage

When the linkage using the DBMS was finished, about 10.4 percent of the person-records and 7.0 percent of the establishment-records could not be assigned to a record from the GAB. To deal with possible typographical errors in the linkage keys street name and place, two distance-based record linkage runs are performed between the hitherto unmatched records and the GAB using the linkage software MTB. In both runs, zip-code serves as a blocking variable and street name, place, house number and house number annex are used as linkage keys. Street name and place are compared using the Damerau-Levenshtein distance function (Hall/Dowling 1980), respectively, thus allowing for some typographical variation in assigned record pairs. To avoid a huge number of false-positive assignments we are, however, forced to postulate the exact agreement of the remaining linkage keys in both runs. For this reason, even though we are able to assign quite a few additional records, the relative contribution of the distance-based linkage proves somewhat unfruitful. Concerning the establishment-records, an additional 5,270 (0.2 percent) can be assigned; concerning person-records we receive an additional 156,838 (0.4 percent) assignments. This means that after the distance-based matching there are 90.0 percent assigned person-records and 93.2 percent assigned establishment-records.

In the course of the routine check of the linkage results an idiosyncratic problem of the data at hand is revealed, namely the fact that IEB person records contain suffixes to place names as in "Dortmund (Westf.)" wherever applicable, whereas "Dortmund" without the suffix would be typical for the names of larger cities in the GAB records. Originally it was decided to leave such suffixes in the place names in order to exploit their discriminatory power, because there are many place names in Germany differing only in such suffixes. Whereas this strategy paid in most instances, in the described case records could not possibly be matched. The IEB records are therefore pre-processed a second time deleting such suffixes. Following that, we simply repeat the exact joins as described in section 3.1. Both the second pre-processing and the join are performed using the information mining software KNIME. As a result, we are able to assign an additional 1,628,509 person records of the IEB to an address from the GAB. This leaves us with overall matching rates of 94.6 percent of the person records and 93.2 percent of the establishment records. Quality checks of the linkage result show that the remaining records are nearly randomly distributed across Germany.

[5] In this paper matching rates refer to **all** persons and establishments selected from the IEB. When referring only to the number of persons and establishments for whom the zip code was filled, we obtain slightly higher matching rates.

4 Aggregation to small-area regions

4.1 Generation of neighbourhood regions

As a result of the record linkage between the IEB and GAB data we receive geocoded addresses of persons and establishments from the IEB. In the second step, these are to be aggregated to small-area regions in order to represent neighbourhoods.

To accomplish that, we use as a first layer the 8,226 German zip-code areas occurring in the data. Within the zip-code areas, at least 15 geocoded units are pooled at a time based on their spatial proximity. For the pooling process we use the Maximum Distance to Average Vector (MDAV) Algorithm by Domingo-Ferrer/Mateo-Sanz (2002). Applied to our data, the algorithm produces 2,280,864 neighbourhood regions in 8,226 zip-code areas.

4.2 Alternative to neighbourhood regions: grid cells

In order to assign the places of work and residence to a grid cell with 1,000 meters edge length, the geographic coordinates need to be projected into a two dimensional space. In accordance with the INSPIRE guide line (EU 2007) we project the geographic information to the Lambert Azimuthal Equal Area (LAEA) coordinate reference system (Annoni et al. 2003) using a specifically developed Stata code. The LAEA coordinates provide the distance in meters of the respective point from the origin of the coordinate reference system. Each point coordinate can hence easily be assigned to its corresponding grid cell. The coordinate of the lower left corner of the cell serves as a grid cell identifier. After this step has been conducted, we dispose of a grid cell identifier for each geocoded person and establishment.

5 Conclusion

The explained steps generate a novel data set that for the first time allows for labour market analysis to be conducted below the municipality level, while still profiting from the high quality of administrative data. Within the context of the project “Neighbourhood Effects: The Analysis of individual-rational behaviour in a social context” the generated regions enable us to truly examine neighbourhoods instead of administrative areas. Furthermore, the newly generated data represent an opportunity for social and economic research as well as a challenge for data protection.

However, many data protection issues are still unsolved as the field of geocoded administrative data for social and economic research is relatively new, at least in Germany. Data which assigns point coordinates to person or establishment information may in no case be published and will not be stored permanently at the Research Data Centre. The aggregation of the individual data to neighbourhood regions or grid cells is one approach to ensure data privacy.

6 Authors Contributions

The record linkage procedure was conceived by Tobias Bachteler and Jörg Reiher. This included the choice of the pre-processing steps, the formulation of the regular expressions and the decisions on the classification procedures. Cerstin Rauscher was responsible for the programming of the pre-processing and deterministic linkage implemented within the DBMS in the server environment of the Federal Employment Agency. Jörg Reiher developed the KNIME nodes to implement the pre-processing of the IEB and the join described in section 3.2. Additionally, he decided on the micro-aggregation algorithm and did the programming used to generate the neighbourhood regions. Theresa Scholz was responsible for checking the quality of the linkage results as well as the generation of the geographic grid cells. All authors contributed to the manuscript, Theresa Scholz wrote the final version. All authors approved the final version of the manuscript.

References

Annoni, Alessandro; Luzet, Claude; Gubler, Erich; Ihde, Johannes (2003): Map Projections for Europe, EU Report 20120 EN.

Berthold, Michael R.; Cebron, Nicolas; Dill, Fabian; Gabriel, Thomas R.; Kötter, Tobias; Meinl, Thorsten; Ohl, Peter; Sieb, Christoph; Thiel, Kilian; Wiswedel, Bernd (2007): KNIME: The Konstanz Information Miner, In: Data Analysis, Machine Learning and Applications - Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.. (Studies in Classification, Data Analysis, and Knowledge Organization). Berlin, Germany: Springer, 319–326.

Domingo-Ferrer, Josep; Mateo-Sanz, Josep M. (2002): Practical Data-Oriented Microaggregation for Statistical Disclosure Control. IEEE Transactions on Knowledge and Data Engineering, 14(1), 189–201.

European Union (EU) (2007): Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, L108.

Hall, Patrick A. V.; Dowling, Geoff R. (1980): Approximate string matching. ACM Computing Surveys, 12(4), 381–402.

Manski, Charles F. (1993): Identification of Endogenous Social Effects: The Reflection Problem. Review of Economic Studies, 60(3), 531-542.

Manski, Charles F. (2000): Economic Analysis of Social Interactions. Journal of Economic Perspectives, 14(3), 115-136.

Schnell, Rainer; Bachteler, Tobias; Reiher, Jörg (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung, ZA-Information / Zentralarchiv für Empirische Sozialforschung 56, 93–103.

PID: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-131793.

Imprint

FDZ-Methodenreport 9/2012

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2012/MR_09-12_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Theresa Scholz,
The Research Data Centre (FDZ),
Regensburger Str. 104,
90478 Nürnberg
Telefon: 0911 / 179-5809
E-Mail: theresa.scholz2@iab.de