Generating a Pseudo Resident Registration Register by Using Open Data

Dominik Visca, Max Hoppe, Pascal Neis
Department of Technology
Mainz University of Applied Sciences
Mainz, Germany
dominik.visca@hs-mainz.de, max.hoppe@hs-mainz.de, pascal.neis@hs-mainz.de

Abstract— **The paper offers a possibility for a (partial) reconstruction of a resident registration dataset combining and linking open (geo-) data with respect to data protection regulations. Here, a method is proposed that simulates a building-level georeferenced resident registration register as a pseudo-derivative. Understood as a potential supplement for research projects, the increasing relevance and discussion of datasets for analyzing individual dynamics on a small scale is taken up with easy-to-generate datasets without facing data protection issues. Complete reconstruction and disaggregation are not possible due to anonymization techniques and data available, but further improvements are conceivable with the latest census and additional open data. This procedure thus reflects trends of Urban-Geo to expand and better visualize small-scale demographic analyses, and also highlights the potential value of opening up register data to science.**

Keywords-Resident registration register; disaggregation; census data; open data; register-based research

I. INTRODUCTION

Public sector data (federal, state, local) are essential sources for small-scale analyses. However, data from statistical offices are usually aggregated at the municipal level or higher. These numbers do not provide an adequate framework for answering many critical questions in science and society, as it is impossible to conclude individual dynamics from aggregated data [1]. Therefore, applicability is limited. Especially in Social Sciences, the need for combined datasets for modeling research questions on a preferably small-scale level, e.g. for the investigation of local inequalities, of lifestyles or the residential location choice of certain groups, is in high demand [2]. As a consequence, also under the impression of the Corona pandemic, the discussion about opening up register data for research is very present.

A large number of private companies offer micro geographic data variables that are used in the commercial sector, e.g. to analyze customer potential, for target-group-specific advertising campaigns, or for location and branch network planning. Some of these variables are available at an address level, but at least at the level of settlement units or urban neighborhoods. They contain information on household sizes and household type, but also derived indicators on social status, income classes or purchasing power [3]. However, the underlying methods used by private companies to produce fee-based micro geographic data are generally not revealed. For scientific work, this means a limitation in terms of in-depth quality control. Therefore, transparent methods are needed being compatible with the requirements of data protection and yet flexible enough for modeling different research questions as well as other applications [2].

In compliance with data protection regulations and within concessions to quality, timeliness, and data resilience, expedient solutions for observations below the community level can be offered as a potential complement for research projects. This paper discusses a method for a (partial) reconstruction of a building-level georeferenced resident registration register using open data for the use as a small-scale geomonitoring based on demographic analyses.

Before the method is outlined in Section III, the paper introduces background and challenges regarding register-based research and resident registration data in Section II. In Section IV, results for a test area are discussed and evaluated. The paper ends in Section V with a conclusion and an outlook regarding potentially available data in the future.

II. BACKGROUND AND CHALLENGES

In Germany, certain information about residents is held in a register covered by the German Federal Registration Act (Bundesmeldegesetz, BMG). Currently, the registration authorities have to keep 19 attributes and 11 details for individuals who are required to register. Among the address, it contains information like date of birth, gender, or when moving in and out. The initial law dates back from May 3, 2013, and entered into force on November 1, 2015 [BGBl. I, 2015]. With this legislation, the federal government used its exclusive legislative competence, which was transferred to the state as part of a federalism reform. Affecting federal states now only have regulatory authority if they are explicitly entitled [4]. To its legally defined extent, the resident registration register is neither made for science nor a source for small-scale geomonitoring [5]. However, the overall positive effects of register-based research are reflected, for example, in Recital 157 of the EU General Data Protection Regulation [2016] (GDPR): "By coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer, and depression. On the basis of registries, research results can be enhanced as they draw on a larger population. Within social science, research on the basis of registries enables

 ISBN: 978-1-61208-983-6

researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. “

For example, Othengrafen, Linda, and Greinke [6] discuss the potential of resident registration data concerning multilocality in rural areas. Schmoigl and König [7] point out the value of opening register data to science if regulated using three examples from different fields. Also highlighting the need of population data with a fine spatial resolution, Pajares, Muñoz Nieto, Meng and Wulfhorst [8] propose an approach to hybrid population disaggregation using open and widely available data. As reference data for comparison, they use information from a resident registration register.

Emphasizing a more detailed spatial observation in terms of a small-scale geomonitoring, Schaffert and Höcht [5] draw attention to the possibilities of resident registration data. By assigning geocoordinates to data, referred to as georeferencing or geocoding, spatial observations can be related to each other. At the same time, concrete analyses such as calculations of geographical distances between points or a more precise determination of supply needs and land consumption become possible [5][9].

Initially, addresses held in resident registration data allow buildings to be referenced. They do not necessarily consist of personal data with concern to the GDPR [9]. However, the assignment of an address to a single building could become critical if only one person inhabits the building. The identity of this individual might then be revealed [9]. In case geocoded address data are combined with further information (e.g., simple survey data without name reference), it is crucial to pay attention to data protection, as the spatial location poses a re-identification risk or may allow de-anonymization of individuals [9][10]. This difficulty is also mentioned by producers of German official statistics working on integrating statistical and spatial data: The more small-scale statistical data are provided and illustrated, the problem of detection risk becomes more relevant [11]. With regard, Section 16 of the Federal Statistics Act (2016) states that individual data on personal and factual circumstances must be kept secret resp. that re-identification of individuals is prohibited.

III. METHOD AND PROCEDURE

Attributes to be kept in resident registration registers are listed in Section 3 of the BMG. Due to privacy and data regulations, a complete reconstruction of a resident registration register using open data is impossible since attributes such as names, ID card numbers, or tax IDs are not available. With some loss of information, other attributes can be reconstructed. For example, open data offers age classes instead of using the date of birth as listed. But focusing on geomonitoring based on demographic analyses, a complete reconstruction is not necessary. For the method presented here, open data of the 2011 Census [12] and free-to-use geodata from OpenStreetMap [13] will be utilized. Table 1 shows an overview of attributes relevant to demographic analyses that are kept in resident registration registers and the data sources used as underlying input for deriving.

TABLE I. RELEVANT ATTRIBUTES AND THEIR DERIVATION FROM OPEN DATA

Attribute	Open Data Source
Date of birth	2011 Census (age groups)
Current address	OpenStreetMap
Indication, whether spouse or life partner exists	2011 Census (family types)
Indication, whether minor children exist	2011 Census (family types)

Small-scale, localized results of the 2011 Census are available in statistical tables for linking to a geographic grid system. These new data sets were only made possible by a 2013 amendment to the Federal Statistics Act (Bundesstatistikgesetz, BStatG). They are not available for all information collected in the 2011 Census [14]. To assign data geographically, it is necessary to use a corresponding grid dataset.

The Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodäsie, BKG) provides geographic grids in different spatial resolutions. According to INSPIRE implementing regulations regarding interoperability of spatial data sets and services [15], a georeferenced ETRS89-LAEA 100m grid is decisive for using with 2011 Census data. But not all attributes of the 2011 Census are available within a spatially resolved 100m grid system. Available data can be accessed through a web portal that provides what are called "grid cell-based results" for "population," "demographics," "families," "households," "buildings," and "housing." Handed as CSV tables, this data can then be joined to geographic grids by assigning the ID [16].

An overview of all single processing steps generating a resident registration register from open data can be seen in Figure 1. The area for which a derived resident registration register is to be generated can be bounded by a polygon. First, this area is intersected with all census grid cell data. Then, attributes contained are linked to the grid cells. Structurally, the table for "population" is different from others. Result tables for "demographics," "families," "households," "buildings," and "dwellings" contain a separate row for each existing characteristic value of each characteristic per grid cell. Tables need to be pivoted to enable a 1:1 join between grid cells and result tables. In conclusion, tables now contain one row for each grid cell and one column for each existing combination of the characteristic and characteristic value. In contrast, the result table "population" contains for each grid cell a column with the population number of the respective cell, which eliminates the need for preprocessing.

 ISBN: 978-1-61208-983-6

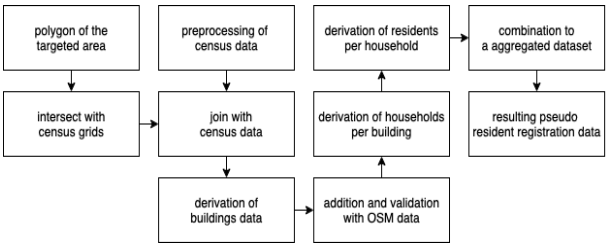

Figure 1. Schematic illustration for all single processing steps

Next, datasets of the grid cell-based results are linked to the geographic grid cells based on their grid ID. It is essential to ensure that the geographic grid is projected correctly, as the grid ID is derived from the respective cell coordinates [17]. An assignment to the 2011 census results is only possible when applying an ETRS89-LAEA projection for the geographical grids. As for further processing, all single cells of the geographic grid are the decisive spatial reference unit. I.e., the following work steps take place for each grid cell individually. The result of each iteration leads to a complete data set (see Figure 2). A data structure is filled for each iteration of the following processing steps. The grid ID is passed as metadata and spatial reference information to the data structure. After that, the total number of buildings and households are transferred from the linked tables. This serves as a control panel.

The census grid cell-based results' building type (size) attribute, which describes the building type e.g. “detached single house” or “apartment building”, served as a critical comparison feature for the OpenStreetMap dataset. This attribute classifies, among other things, whether buildings are single-family or multi-family houses. The summed number of buildings per attribute constitutes the total population of residentially inhabited buildings in a grid cell. Single buildings are enriched with address information from OpenStreetMap. For this purpose, the Overpass API tool is used to extract addresses from a spatially bounded area, i.e., the extent within the current grid cell [18]. In addition to spatial bounding, the Overpass Query Language (Overpass QL) also allows thematic selection based on OpenStreetMap's data model and tagging guidelines [18]. In OpenStreetMap, buildings are provided as *way* to which thematic information such as the address or the type of a building is kept in the form of *tags*. As a result of free editability, incorrect or missing data can be part of OpenStreetMap datasets [19]. Sometimes, address information is not attached to a building polygon but to a point object within the building polygon. Therefore, *address tags* of all *nodes* and *landuse* tags of all *ways* in respective grid cells are also requested. Data sets with an *amenity* tag or an incompatible *landuse* value (i.e., other than residential or settlement area) are filtered out, meaning commercially used buildings are not considered. This allows more accurate matching of buildings listed in the census dataset, containing only buildings with (partial) residential occupancy [20].

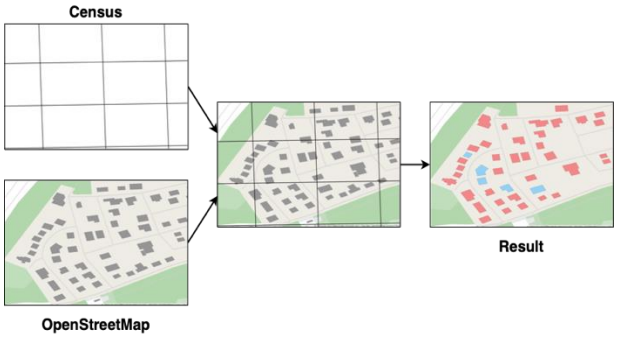

Figure 2. Derivation of the pseudo-derivative, © OpenStreetMap contributors

If possible, the assignment of address data from OpenStreetMap and building information within the census datasets proceeds via building types, otherwise randomized. Based on the building type and the census data attribute "number of apartments in the building,” apartments or households are distributed across the buildings. The population of households is founded on the summed households belonging to the characteristic "type of private household,” to which adequate factual data of the characteristic "size of private household" was assigned in each case. In this way, households with a corresponding number of individuals are assigned to a building. A classification of whether the household members are adults or children is done by the household type, i.e., "single-person household" or "couples with children.” Afterward, a certain number of people datasets are generated for each household according to the household size attribute. These are assigned to appropriate age classes from the census data. After successful iteration, the final cell data structure is added to a comprehensive data set.

IV. EVALUATION AND DISCUSSSION

The method presented here was applied to the city of Herborn as a test area (Germany, state of Hesse) shown in Figure 3. The area covers an area of approximately 4.52 km2 with 257 residential grid cells. 1,992 houses were derived and enriched with attributes relevant for demographic analyses.

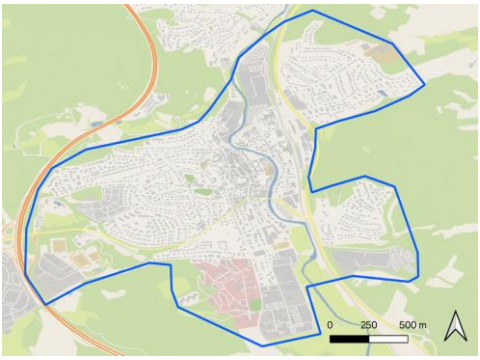

Figure 3. Overview across the test area of the city of Herborn, © OpenStreetMap contributors.

Figure 4 shows a classification by building type (size) based on the census data. All nine residential buildings listed in OpenStreetMap were detected within the grid cell, eight of them as a single-family (marked in red) and one as a multi-family house (marked in blue). A manual comparison with the keys maintained in OpenStreetMap shows seven single-family and two multi-family houses for this area. A field check also reveals a corresponding result.

In total, 25 residents are distributed across twelve households in this grid cell. More specifically, three households consist of couples with one child, two households with a single parent and one child, two single-person households, and five couples with no children. For each resident, the characteristics listed in Table 1 are available. The grid cell shown is characterized by its relatively homogeneous settlement structure. This is a factor of a high level of coherence between Census and OpenStreetMap data. However, this needs further investigation.

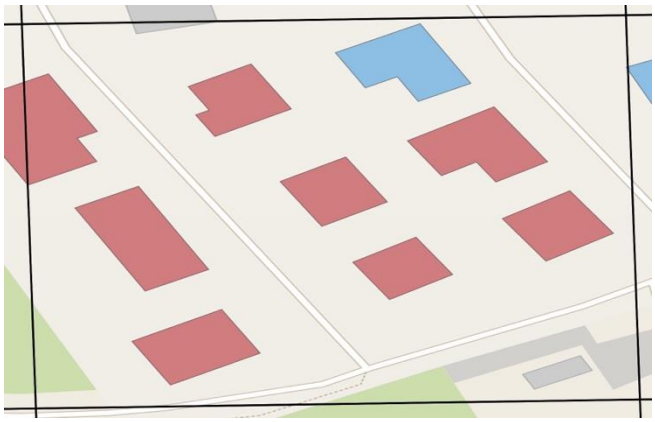

Figure 4. Classification result by building type (size), © OpenStreetMap contributors.

Parts of the census dataset suggest two additional residential buildings for this grid cell, a fact interesting to point out. The anonymization technique underlying the production of census data leads to a scenario in which summations across different characteristics will not necessarily produce identical results [20]. Summing up, all expressions of the characteristic "building type (size)" returns only 1,805 buildings. This is a difference of about 9%. In total, 161 of the 257 grid cells, meaning almost 63 %, showed a deviation with a mean value of 1.75 buildings and a standard deviation of 0.89 of these two characteristics, which can vary either positively or negatively. These deviations are also found for other attributes of the census data set and put a theoretically achievable accuracy of the disaggregation into perspective. As such, it also fulfills the purpose of the SAFE (Secure anonymization for individual data) technique for anonymization [20]. Figure 5 below shows a result of an anonymization technique. According to OpenStreetMap, the grid cell contains four buildings; the census dataset indicates 76 inhabitants, but any information on buildings or households is missing. An in situ field comparison shows the existence of three multi-story buildings with elderly residences and a school building (left part of Figure 5). So far, the method described here cannot account for these 76 residents without information on buildings and households.

Figure 5. Lack of mapping due to anonymization of census data, © OpenStreetMap contributors

Furthermore, the results of the presented method correlate with the address completeness in OpenStreetMap. The selected test area shows a medium to good completeness of about 71 % in relation to the official real estate cadastre (Amtliches Liegenschaftskatasterinformationssystem, ALKIS). Building-related accuracy is expected to be marginal for areas with low address completeness. Additionally, the timeliness of the datasets also affects the results. While the census dataset reflects the state of 2011, OpenStreetMap tends to contain a continuously updated dataset. Depending on urban dynamics and population fluctuations, smaller or larger deviations are expected. In addition, the intended area of application must also be taken into account. As mentioned above, both of these factors need to be investigated in more detail in a follow-up study.

V. CONCLUSION AND OUTLOOK

This paper proposes a method that generates a georeferenced resident registration register at a building level as a pseudo-derivative based on openly available spatial and public-statistical data. Complete reconstruction and disaggregation of the census data is not possible due to the anonymization techniques but is usually not necessary for research purposes. In this case, addresses were assigned to demographic characteristics with relevant age classes. The accuracy of the assignment can be improved by considering additional census data characteristics, OpenStreetMap tags, and different open data sets. The census dataset offers other attributes such as gender, religious affiliation, or citizenship. This can increase the applicability of the pseudo-derivative. The lack of internal consistency of the census data caused by the underlying SAFE-anonymization technique and the data quality of OpenStreetMap requires a more comprehensive evaluation of the quality of the dataset generated by the method outlined.

Regarding the 2022 Census in Germany, it is possible to continue providing easy-to-generate datasets for research purposes without facing data protection issues. Looking at

 ISBN: 978-1-61208-983-6

ways of enriching or combining data, the German government's open data strategy published in 2021 [21] gives reason to expect that more data from public administrations will be made available and that its quality will steadily improve. In addition, the availability of previously inaccessible data from the economy, science, and civil society is expected to increase. The State of Hesse, or rather the Hessian Administration for Land Management and Geoinformation (Hessische Verwaltung für Bodenmanagement und Geoinformation, HVBG), where the method presented here was applied using the example of the city of Herborn, has also been making additional free geodata available since 2022.02.01 [22]. The products of the real estate cadastre, the state measurement, and the real estate valuation are made available via the store component, the download center, as Web Map Service (WMS) and Web Feature Service (WFS). The data is continuously updated and permanently available.

The Council for Social and Economic Data (Rat für Sozial- und Wirtschaftsdaten, RatSWD) [23] also recommends in a position paper for the 20th legislative period of the German Bundestag to facilitate access to register and administrative data for scientific purposes. An envisaged Research Data Act is intended to strengthen science in Germany and open up research into socially relevant issues, even in politically sensitive areas.

ACKNOWLEDGMENT

The work on this article was done as part of the project "Spatial Intelligence for the Integrated Care of Seniors in Rural Neighborhoods (RAFVINIERT)", which is funded by the Carl Zeiss Foundation in the program "Transfer - Intelligent Solutions for an Aging Society".

REFERENCES

[1] W. Redaktion, „Registry Research: Opportunities, Risks, and Challenges,“ Wirtschaft Und Gesellschaft, vol. 46(3), pp. 315–328, November 2020.

[2] B. Heldt and D. Heinrichs, „Using the 2011 Census to model households' residential location choices under conditions of secrecy,“ Zeitschrift für amtliche Statistik Berlin Brandenburg, pp. 29-33, 2017.

[3] A. Milbert and S. Fina, „Small town research methods: definitions, data, and spatial analysis,“ in A. Steinführer, L. Porsche and M. Sondermann (Eds.), Compendium of Small Town Research, pp. 24-49, Hannover: Forschungsberichte der ARL 16, 2021.

[4] E. Ehmann, Dealing with Data from Population Registers Correctly, 3rd ed., Stuttgart: Boorberg, 2017.

[5] M. Schaffert and V. Höcht, „Geocoded Data from Population Registers as a Source for Needs-Based Planning in Rural Municipalities and Regions,“ Raumforschung und Raumordnung, vol. 76(5), pp. 421–35, 2018.

[6] F. Othengrafen, L. Linda and L. Greinke, Temporary arrivals and absences in rural areas: effects of multilocal lifestyles on land and society. Wiesbaden, Wiesbaden: Springer Fachmedien, 2021.

[7] R. König and L. Schmoigl, „Successful Registry Research in Austria. What additional value does the regulated opening of registry data generate for scientific research? A presentation based on three examples,“ Österreichisches Institut für Wirtschaftsforschung (WIFO) und Institut für höhere Studien (IHS), pp. 1-16, 2020.

[8] E. Pajares, R. Muñoz Nieto, L. Meng and G. Wulfhorst, “Population Disaggregation on the Building Level Based on Outdated Census Data“, ISPRS Int. J. Geo-Inf., 10(10), 662, pp. 1-21, 2021, doi.org/10.3390/ijgi10100662.

[9] S. Müller, "Spatial Linking of Georeferenced Survey Data with Geospatial Data: Opportunities, Challenges, and Practical Recommendations," in U. Jensen, S. Netscher and K. Weller, Research Data Management of Social Science Survey Data, pp. 211-29, Barbara Budrich, 2019.

[10] M. Van Der Meer, F. Meissner, M. Merten and D. Münderlein, "Development and Potentials of Digital Spatial Research. Ethical Issues and Impulses for University Teaching," RaumPlanung, vol. 2/3(196), pp. 20–27, 2018.

[11] S. Schnorr-Bäcker and M. Etienne, "Challenges and Possible Solutions in Combining Statistical and Spatial Data from the Perspective of Federal Statistics," Stadtforschung und Statistik: Zeitschrift des Verbandes Deutscher Städtestatistiker, vol. 3(1), pp. 63–69, 2018.

[12] Statistische Ämter des Bundes und der Länder. Zensus 2011. [Online]. Available from: https://www.zensus2011.de/ [retrieved: 04.2022].

[13] OpenStreetMap. OpenStreetMap. [Online]. Available from: https://www.openstreetmap.org [retrieved: 04.2022].

[14] M. Neutze, „Grid-based Evaluations of the 2011 Census," Stadtforschung und Statistik, pp. 64-67, February 2015.

[15] COMMISSION REGULATION (EU) No 1089/2010 of 23 November 2010 Implementing Directive 2007/2/EC of the European Parliament and of the Council as Regards Interoperability of Spatial Data Sets and Services. [Online]. Available from: https://eur-lex.europa.eu/eli/reg/2010/1089 [retrieved: 04.2022].

[16] Statistische Ämter des Bundes und der Länder. Results of the 2011 Census for Download - Extended. [Online]. Available from: https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html [retrieved: 04.2022].

[17] Bundesamt für Kartographie und Geodäsie. Documentation. Geographic Grids for Germany. GeoGrid. [Online]. Available from: https://sg.geodatenzentrum.de/web_public/gdz/dokumentation/deu/geogitter.pdf [retrieved: 04.2022].

[18] R.M. Olbricht, “Data Retrieval for Small Spatial Regions in OpenStreetMap,” in J. Jokar Arsanjani, A. Zipf, P. Mooney, M. Helbich, Eds. OpenStreetMap in GIScience. Lecture Notes in Geoinformation and Cartography, pp. 101-122. 2015.

[19] P. Neis and D. Zielstra, “Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap,” Future Internet, vol. 6, pp. 76-106, 2014.

[20] Statistische Ämter des Bundes und der Länder. 2011 Census. Methods and Procedures. [Online]. Available from: https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2015_06_MethodenUndVerfahren.pdf;jsessionid=48260D25A514027445F421D903862F47.live422?__blob=publicationFile&v=2 [retrieved: 04.2022].

[21] Bundesministerium des Innern, für Bau und Heimat. Open Data Strategy of the Federal Government.

[22] Hessische Verwaltung für Bodenmanagement und Geoinformation. Introduction of Open Data by 2022.02.01 [Online]. Available from: https://hvbg.hessen.de/open-data [retrieved: 04.2022].

[23] Rat für Sozial- und Wirtschaftsdaten. RatSWD Position Paper on the Federal Government's Data Strategy.