AGILE: GIScience Series, 4, 38, 2023. https://doi.org/10.5194/agile-giss-4-38-2023
Proceedings of the 26th AGILE Conference on Geographic Information Science, 2023.
Editors: P. van Oosterom, H. Ploeger, A. Mansourian, S. Scheider, R. Lemmens, and B. van Loenen.
This contribution underwent peer review based on a full paper submission.


# Benefits of using address-based dasymetric mapping in micro-level census disaggregation

Denis Reiter [1], Mathias Jehling [1], and Robert Hecht [1]

[1]Leibniz Institute of Ecological Urban and Regional Development, Dresden, Germany

Correspondence: Denis Reiter (D.Reiter@ioer.de)

**Abstract.** Dasymetric mapping is a well-known technique when attempting to refine census data spatially and/or temporally. Existing approaches in micro-level census disaggregation make use of building areas or volumes in the mapping process. In an empirical error comparison it is shown that using additional address data rather than only building footprints or 3D models can substantially reduce dislocation of residential population. We propose the use of address points as a geometric representation unit for a more refined census disaggregation method in the future.

**Keywords.** census disaggregation, dasymetric mapping, population density, address data

## 1 Introduction

Information about the human population, its spatial distribution and change over time is an essential asset for various scientific disciplines such as geography, history and social science (Monteiro et al., 2019). Common sources for such data are governmental databases, statistics and most important, census reports. Unfortunately, the data scarcity of such products can be a problem: they can have low spatial resolution and/or long update cycles (Wu et al., 2020), which makes small-scale analysis complicated or even impossible. The general need for more accurate data, spatially and temporally, is long since well-known (Smith et al., 2002), which can be addressed by using dasymetric mapping methods (Eicher and Brewer, 2001), especially considering various approaches using building data to achieve micro-level population estimations (Hecht et al., 2018; Schug et al., 2021).

While making use of census data for dasymetic mapping, it must be considered that available data is generally produced by aggregating data collected on household levels into bigger spatial units through different approaches. Aggregation of data can lead to a modifiable area unit problem (or MAUP, see Openshaw (1983)), which has been proven to affect census data (Flowerdew, 2011). While not in the same magnitude as with administrative units, the MAUP is also present for raster data (Lyn, 2001), which is a common format for census data. Beyond that, census data collection and especially the representation unit is not uniformly regulated across nations: the data can be attributed to actual building footprints/3D models (an object representing an area or volume) or to addresses/housing units (a single-point object). Considering these characteristics poses substantial challenges for dasymetric mapping, as the available aggregated census data needs to be again disaggregated to relevant units (Mennis, 2003; Monteiro et al., 2019; Wu et al., 2020; Huang et al., 2021) when considering various applications that need very fine grained population data. In particular, for urban planning activities such as active mobility (walking and biking) transport solutions, land use planning or in various risk management applications such as flood prevention, very detailed population data is needed, down to the level of individual buildings (Calka et al., 2017; Pajares et al., 2021).

When the original collected census data is aggregated for the publishing of finished datasets, each census object is assigned to a raster cell, census tract, etc. The object type (e.g. address point or building footprint) which is used for representing or "storing" the collected data is very important in this process: For any point object type the assignment is unique, but for for areal object representations there may be multiple possibilities to assign them, since buildings can overlap multiple raster cells. This is especially relevant when using existing, persistent rasters, like standardized European grids, since they do not adapt to given structural features and overlapping happens frequently. The assignment of areal objects can then be done in various ways, for example by largest overlap, by using an anchoring point (the location of the actual physical entrance to a building) and various other methods. Technically it would be possible to split the areal objects so that overlaps are removed, but this would lead to dislocation of population in the mapping process, because the original census does not split buildings either. Also the building

area/volume as a dasymetric mapping factor would be altered by splitting, which would lead to proportionally less remapped population per building. For the most accurate remapping it is therefore necessary to aissign each building completely to a single cell.

Considering all the aforementioned problems, disaggregation approaches for census data thus should always consider the preceding aggregation techniques. This is especially crucial when remapping to a different unit than the one the data collection was done with (e. g. remapping to buildings when data collection was done using address points), to avoid incorrect spatial matching of objects and subsequently biases in the resulting spatial patterns in micro-level population distribution. It is also very important for any scenario where disaggregated population and building data is used to train models in the field of GeoAI or machine learning.

It is the aim of this paper to express the benefits of using address-based dasymetric mapping in micro-level census disaggregation. Taking the example of disaggregation from raster-based population data to building geometries, this paper contrasts two different remapping variants in an empirical comparison: the first one solely looking at the largest overlap between building footprints and census cells, the second one using the building address coordinates as anchoring points for the building footprints. For both variants, a remapping of the census grid population to the building footprints by leveraging the building volume as a mapping factor is then performed. The federal state of Nordrhine-Westfalia in Germany serves as case study to describe the general principles of the method. The overall goal of the proposed method is solving the allocation problem of assigning a building footprint to the same raster cell to which it was originally assigned in the census, if possible, to get a reconstruction of the non-available original census data on a building level as perfect as possible.

## 2 Used data

This section provides information about the used data and the proposed method. To achieve the best possible accuracy for the remapping and error estimation, official data products issued by German authorities have been used, since they are subject to thorough quality controls.

### 2.1 INSPIRE grid

The Infrastructure for Spatial Information in Europe (INSPIRE) defines standard geographic grids for Europe (INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems, 2014). For this paper the 100m resolution Equal Area Grid based on the ETRS89-LAEA coordinate system was chosen as the spatial basis. The German subsection of the grid as well as all other used data sets were obtained via the Federal Agency for Cartography and Geodesy (BKG). The latest grid reference is from 2019 (BKG, 2021b).

### 2.2 Census data

The latest officially issued census data available for Germany is from 09.05.2011 (Federal Statistical Office, 2018). It is available as a table including the INSPIRE cell code for identification. Therefore it is possible to directly join the census data and the 100m INSPIRE grid and load the population data into a geographic information system (GIS) for further analysis.

### 2.3 3D building models

Using building geometries or 3D models is a common practice for disaggregating census data (Biljecki et al., 2016; Wu et al., 2020; Huang et al., 2021).In the proposed work, a building volume based approach was chosen to dasymetrically map the census population from the census grid onto the buildings. The most suited product in Germany containing building volumes is the "3D-Gebäudemodelle Level-of-Detail-2 Deutschland" or LoD2-DE (BKG, 2021a). Problematic was the fact that the LoD2-DE was not available for the year 2011, so it was needed to project data from the available year 2020 onto a different data set (see 2.5).

### 2.4 Address coordinates

Since the census in Germany is conducted using address data (Federal Statistical Office, 2015), it was obvious to choose address coordinates as one of the disaggregation features. The official address coordinates HK-DE (BKG, 2012a) were available for the year 2012.

### 2.5 Building footprints

As a projection target for the LoD2-DE data the official building footprints HU-DE (BKG, 2012b) were chosen, since they were also available for the year 2012. Despite the fact that the census was conducted in 2011, it was concluded to use the building footprints from 2012. The reasoning behind this was to make sure that the least possible buildings were missing, since the target date of the census was in the middle of the year 2011, using a building stock from 2012 could ensure that less data was missing while still being close to the the census year. Moreover, using the 2012 footprints resulted in a better matching with the year the address coordinates were available for (which were not available for 2011).

## 3 Methods

An overview of the developed workflow is shown in Fig. 1. The process is divided into two main steps: data prepa-

ration and the actual population disaggregation, which was done in two different variants. Variant A uses only building footprints, Variant B also utilizes address point data to assign the building footprints to an INSPIRE grid cell. The census data was joined to the INSPIRE grid using the provided cell codes. Before applying the method, all data was projected into the ETRS89-extended / LAEA Europe coordinate system (EPSG: 3035).

### 3.1 Data preparation

Data preparation began with the projection of the LoD2-DE building volume onto the HU-DE building footprints. Firstly, all buildings smaller than 50 $m^2$ were deleted from the data set, so that small sheds, garages etc. were excluded. The LoD2 was then converted into points (centroids) and spatial joined to the building footprints. In the next step, all non-residential building footprints were deleted using the usage information from the LoD2 where available, otherwise a classification scheme based on the Authorative Topographic-Cartographic Information System (ATKIS) was used (Hartmann et al., 2016). Since there were a number of building footprints which had no matching counterpart in the LoD2, their volume needed to be interpolated. This was done using a k-means method based on the nearest 10 neighboring buildings. At the end of the data preparation, about 3.8 million building footprints were left.

### 3.2 Disaggregation Variant A

For Variant A using only the building footprints, the grid with the census population was spatial joined to the footprints using the largest overlap method, buildings outside of the grid were deleted. After that, the total building volume per cell was calculated. With that, the volume percentage share of each individual footprint inside of a cell could be determined. This was then multiplied with the total cell population, which resulted in the mapped population for each individual building.

### 3.3 Disaggregation Variant B

The main difference between variants A and B is the way the buildings were assigned their respective cell codes. Instead of assigning the cell to a building by the largest overlap, the census grid was first spatial joined to the address coordinates (HK-DE), which were then spatial joined again to the building footprints (HU-DE). Each building was then assigned the cell code which belonged to the corresponding address point, regardless if the building was physically located mostly inside of another cell. For buildings where no address point could be assigned, the largest overlap principle was applied again. The rest of the process was identical to Variant A.

In Fig. 2 a constructed example is shown to elaborate on the difference between both disaggregation variants. With Variant A, the building would be assigned to the right raster cell and thus the population of the right cell would be mapped to the building. With Variant B, the building would be assigned to the left cell. Depending for which cell the total cell buildings (and with them their inhabitants) are then summarized/aggregated, this may introduce small-scale population shifts if the aggregation is performed on the "wrong" cell.

### 3.4 Data and software availability

Data described under 2.1 and 2.2 is openly available under the licence "dl-de/by-2-0". All other data is not publicly available and needs to be individually licensed from BKG. All described processing steps were done in ArcGIS Pro 3.0.3 (ESRI, 2023) using built-in functions and tools. The software ArcGIS Pro 3.0.3 is available through licensing via the manufacturer ESRI.

## 4 Results

Within the study area, there were 3.738.826 buildings identified as residential. Regarding these buildings, a difference in cell allocation due to the different mapping variants could be observed in 162.637 (4,35 %) cases. This is equivalent to a potential number of 978.038 (5,57 %) dislocated inhabitants out of a total 17.555.418 inhabitants.

Since there is no building-level population reference data available for 2011, the evaluation of the method was done by re-aggregating the mapped population again and comparing the results to the original census grid. The re-aggregation uses the same raster resolution as the disaggregation was done with, since the overall goal is to reconstruct the original census as perfect as possible. Examples for the generated error maps are shown in Fig. 3. It can clearly be seen that the Variant A using the largest overlap method leads to frequent dislocation of population, while Variant B does much less often (see also Tab.1). This is due to the fact that the point-based approach of B does not allow for buildings to be located into "wrong" cells, since there is no way of misinterpreting the physical location of a point. With Variant A, buildings can be misplaced more easily, since they are actual areal objects, often overlapping multiple cells. While the overall population on the state level does not change when re-aggregating, the important micro-level population distribution does differ between both approaches. The much better performance of Variant B can also be attributed to the fact that the census data collection in Germany is address-based (see 2.4 and therefore it is to be expected that an address-based remapping outperforms another variant even more than under normal circumstances.

It is important to note that the errors in Variant B are almost solely attributed to the lack of building footprints in the corresponding grid cells. This explains why the values are always at -100 % and below, since there was no pop-

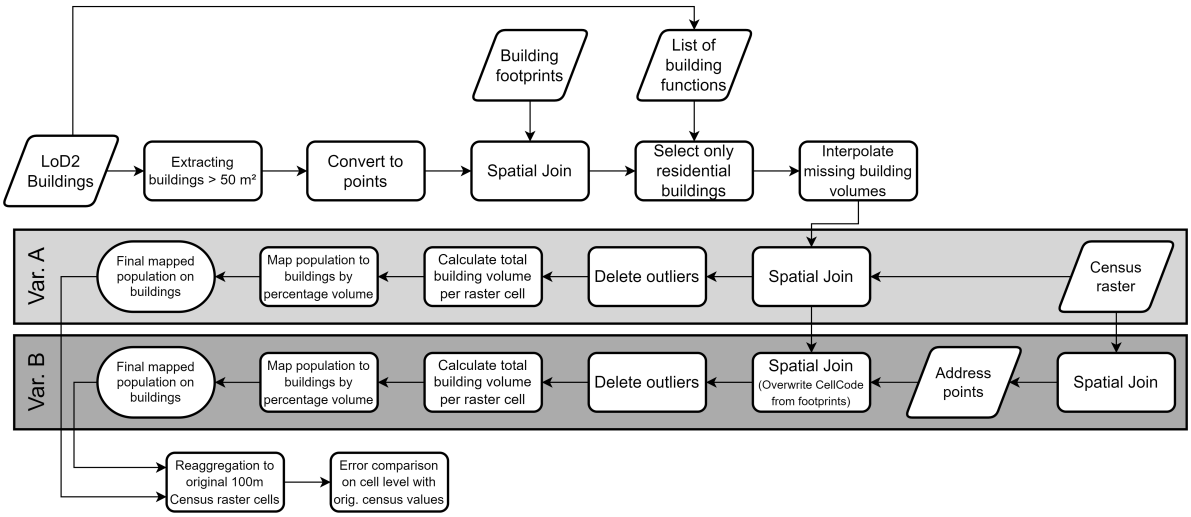


**Figure 1.** Detailed workflow schema of the applied methodology in this paper.

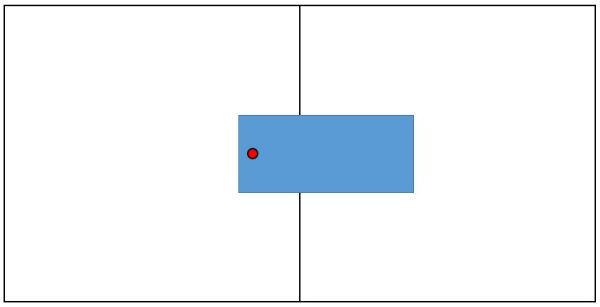

**Figure 2.** Constructed example showcasing the basic allocation problem.

**Table 1.** Number of cells in relative error categories of both mapping variants. (* These values were removed since they resulted from projection errors as described in the text).

| Relative error to original census [%] | No. of cells (Variant A) | No. of cells (Variant B) |
|---|---|---|
| <= -100 | 15.632 | 12.706 |
| <= -50 | 4.840 | 0 (1*) |
| <= -25 | 15.159 | 0 (5*) |
| < 0 | 122.853 | 0 (3*) |
| No difference | 3.109.899 | 3.398.566 |
| <= 25 | 107.583 | 0 |
| <= 50 | 18.012 | 0 |
| <= 100 | 10.834 | 0 |
| > 100 | 6.460 | 0 |

ulation disaggregated in these cells at all. Values of less than -100 % arise because empty cells in the census (and therefore in the disaggregated data too) have a value of -1 inhabitants, not 0. Moreover, there where 9 cells where the projection into the ETRS89-extended / LAEA Europe coordinate system shifted address points lying very close to a cell border in a way that they fell into a neighboring cell. These cells were excluded from all analysis, since this was no true error, but a technical one.

For a first attempt to explore on where these errors occur the most, a cell-level comparison of the respective total building volume and the relative error between the disaggregation Variant A and the original census is done. The results are shown in Fig. 4, and it is clear that a smaller total building volume in a given cell results in a higher error average as well as a bigger margin of errors.

## 5 Discussion

The presented results are a first indication that census representation units do in fact matter when attempting to dasymetrically map census data from raster cells to buildings. However, the method has limitations that will be discussed below.

There are several sources of error that have an impact on the overall accuracy of the mapping: first and foremost is a misclassification of the residential buildings, which leads to the fact that there were a lot of cells where buildings were missing. This resulted in no population remapped at all in such cases, and therefore partly or completely empty cells where the original census showed inhabitants. Tied into this problem was also that there were some cases where the address point did not intersect the building footprint, which let to same result that no population was remapped. The second problem could be handled by applying a matching of address points and building footprints beforehand.

Another source of error is the fact that building footprints could have multiple address points, and that those were in some cases located in different cells Fig. 5. With the proposed approach, the entire building is assigned to a single

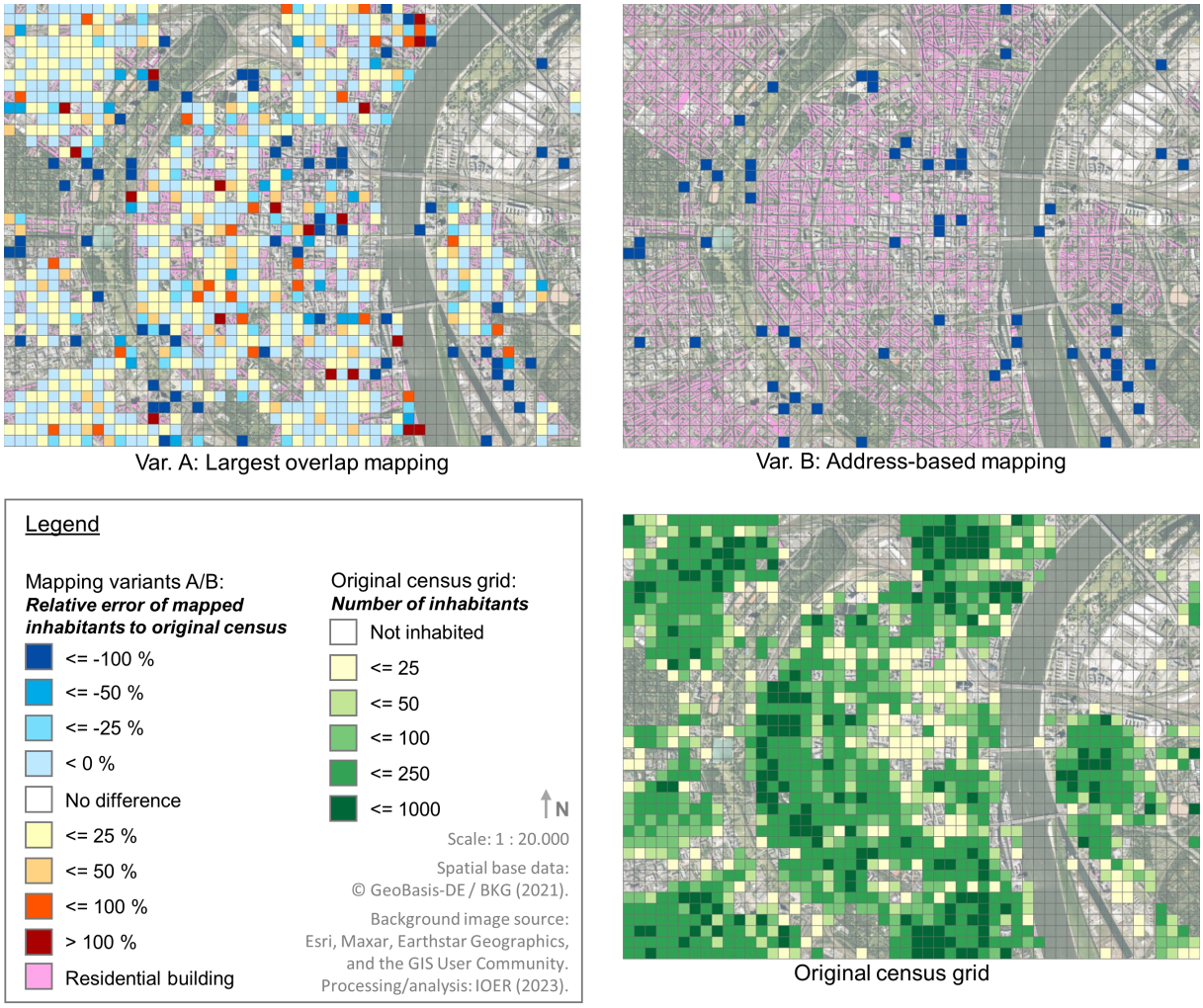


**Figure 3.** Comparison of errors of both disaggregation methods to the original census grid. As an example area the city center of Cologne is shown.

address point, which leads back to the original problem of possibly locating buildings in a wrong cell. Adding to this problem is the fact that the LoD2 and the building footprints geometry do not match in most cases. By using the original LoD2, which is geometrically split into multiple building parts mostly, the problem with multiple address points could be reduced by a fair amount, although not completely. It may be necessary to develop a splitting mechanism beforehand, so that every building is attributed with exactly one address. This could improve the disaggregation further.

In an attempt to find possible indicators for where errors occur more frequently, results showed that the total building volume per cell might be one such indicator. Cells with lesser total building volume had a significantly higher relative error than cells with higher building volumes. This could be related to the fact that small building volumes in a cell typically indicate a lesser number of buildings. If errors occur in a cell with only one or two buildings in total, higher relative errors are more likely to be observed, since there is no "evening-out-effect" with neighboring buildings.

## 6 Conclusion and further research

Considering potential problems in dasymetric census mapping on a micro-level in general, this paper shows benefits of using an address-based dasymetric mapping approach for micro-level census disaggregation by comparing two methods of population disaggregation. As a conclusion it proposes to utilize address point based approaches rather than purely building footprint based ones, at least when dealing with raster-based census data. This can further be emphasized for countrys using an address-based census data collection, where the correct assignment of buildings to their corresponding raster cells is even more important. First results are promising that a more accurate building level population data set could be derived by further developing this method. It could also be observed that building size/number in a local neighborhood could have an effect on the margin of error.

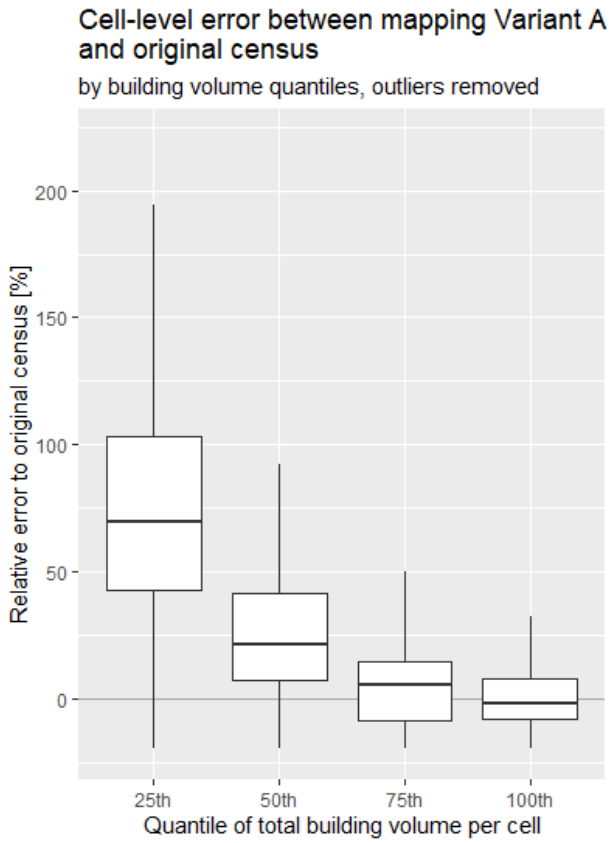


**Figure 4.** Relative error between the disaggregation Variant A and the original census, cell-level. Positive error means too many inhabitants mapped, negative error means too few.

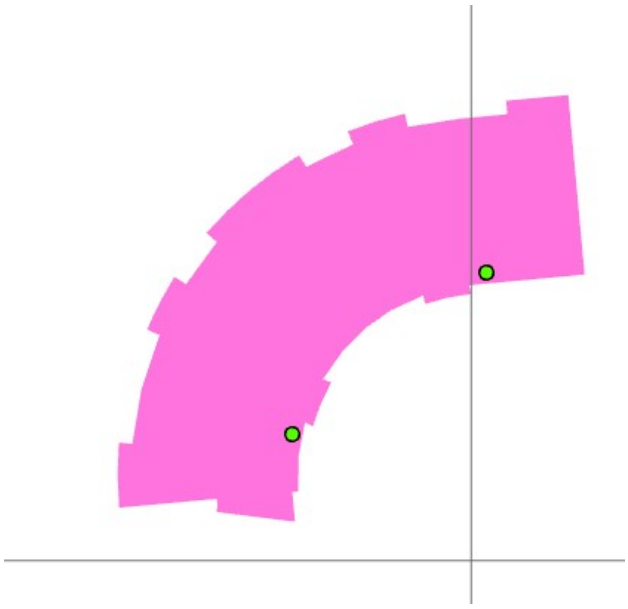

**Figure 5.** Example building footprint with multiple address points locatated in different grid cells, which potentially leads to spatial dislocation even when using the proposed method. Spatial base data: © GeoBasis-DE / BKG (2021)

Further research and refinement of the method includes improvement of the residential building classification process, as well as the building part separation and matching of address points to them. Also a look into additional indicators for errors, such as built-up area per cell, municipality type and more is advisable.

Following through on the findings of this paper, it would also be thinkable to implement a machine learning algorithm that learns relations between building features and inhabitants. The presented results contribute an important groundwork for such an algorithm, since a lot of very accurate training data on the building level would be needed, which this paper aims to provide a methodology for obtaining. Having a reproduction of the census data as accurate as possible is probably one of the key influencing factors on the results of a machine learning process.

A possible application of such an algorithm could be the closing of the massive time gap between census reports, since official building data are generally available in Europe. In case they might not be openly accessible buildings from Bing Maps or OpenStreetMap could be used instead. With precisely produced input data, high accuracy interpolations of micro-level population data sets in between census years by using building related data only could be obtained in the future.

## References


Biljecki, F., Arroyo Ohori, K., Ledoux, H., Peters, R., and Stoter, J.: Population Estimation Using a 3D City Model: A Multi-Scale Country-Wide Study in the Netherlands, PLOS ONE, 11, e0156 808, https://doi.org/10.1371/journal.pone.0156808, 2016.

BKG: Amtliche Hauskoordinaten Deutschland (HK-DE), https://gdz.bkg.bund.de/index.php/default/amtliche-hauskoordinaten-deutschland-hk-de.html, 2012a.

BKG: Amtliche Hausumringe Deutschland (HU-DE), https://gdz.bkg.bund.de/index.php/default/amtliche-hausumringe-deutschland-hu-de.html, 2012b.

BKG: 3D-Gebäudemodelle LoD2 Deutschland (LoD2-DE), https://gdz.bkg.bund.de/index.php/default/3d-gebaudemodelle-lod2-deutschland-lod2-de.html, 2021a.

BKG: Geographische Gitter für Deutschland in Lambert-Projektion (GeoGitter Inspire), https://gdz.bkg.bund.de/index.php/default/inspire/sonstige-inspire-themen/geographische-gitter-fur-deutschland-in-lambert-projektion-geogitter-inspire.html, 2021b.

Calka, B., Nowak Da Costa, J., and Bielecka, E.: Fine scale population density data and its application in risk assessment, Geomatics, Natural Hazards and Risk, 8, 1440–1455, https://doi.org/10.1080/19475705.2017.1345792, 2017.

Eicher, C. L. and Brewer, C. A.: Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation, Cartography and Geographic Information Science, 28, 125–138, https://doi.org/10.1559/152304001782173727, 2001.

ESRI: ArcGIS Pro: Release 3.0.3, https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview, 2023.

Federal Statistical Office: Zensus 2011 Methoden und Verfahren, https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2015_06_MethodenUndVerfahren.pdf?__blob=publicationFile&v=2, 2015.

Federal Statistical Office: Bevölkerung im 100 Meter-Gitter, https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=559100#Gitter, 2018.

Flowerdew, R.: How serious is the Modifiable Areal Unit Problem for analysis of English census data?, Population Trends, 145, 106–118, https://doi.org/10.1057/pt.2011.20, 2011.

Hartmann, A., Meinel, G., Hecht, R., and Behnisch, M.: A Workflow for Automatic Quantification of Structure and Dynamic of the German Building Stock Using Official Spatial Data, ISPRS International Journal of Geo-Information, 5, 142, https://doi.org/10.3390/ijgi5080142, 2016.

Hecht, R., Herold, H., Behnisch, M., and Jehling, M.: Mapping Long-Term Dynamics of Population and Dwellings Based on a Multi-Temporal Analysis of Urban Morphologies, ISPRS International Journal of Geo-Information, 8, 2, https://doi.org/10.3390/ijgi8010002, 2018.

Huang, X., Wang, C., Li, Z., and Ning, H.: A 100 m population grid in the CONUS by disaggregating census data with open-source Microsoft building footprints, Big Earth Data, 5, 112–133, https://doi.org/10.1080/20964471.2020.1776200, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/20964471.2020.1776200, 2021.

INSPIRE Thematic Working Group Coordinate Reference Systems & Geographical Grid Systems: D2.8.I.2 Data Specification on Geographical Grid Systems – Technical Guidelines, https://inspire.ec.europa.eu/id/document/tg/gg, 2014.

Lyn, U. E.: MAUP: Modifiable Areal Unit Problem in raster GIS datasets. Raster pixels as modifiable areas, GIM International, 15, 43–45, http://pubs.er.usgs.gov/publication/70023728, 2001.

Mennis, J.: Generating Surface Models of Population Using Dasymetric Mapping, The Professional Geographer, 55, 31–42, https://doi.org/10.1111/0033-0124.10042, 2003.

Monteiro, Martins, Murrieta-Flores, and Moura Pires: Spatial Disaggregation of Historical Census Data Leveraging Multiple Sources of Ancillary Information, ISPRS International Journal of Geo-Information, 8, 327, https://doi.org/10.3390/ijgi8080327, 2019.

Openshaw, S.: The Modifiable Areal Unit Problem, in: Classification using information statistics, no. no. 37 in Concepts and techniques in modern geography, Geo Books, Norwich [Norfolk], 1983.

Pajares, E., Muñoz Nieto, R., Meng, L., and Wulfhorst, G.: Population Disaggregation on the Building Level Based on Outdated Census Data, ISPRS International Journal of Geo-Information, 10, 662, https://doi.org/10.3390/ijgi10100662, 2021.

Schug, F., Frantz, D., van der Linden, S., and Hostert, P.: Gridded population mapping for Germany based on building density, height and type from Earth Observation data using census disaggregation and bottom-up estimates, PLOS ONE, 16, e0249 044, https://doi.org/10.1371/journal.pone.0249044, 2021.

Smith, S. K., Nogle, J., and Cody, S.: A regression approach to estimating the average number of persons per household, Demography, 39, 697–712, https://doi.org/10.1353/dem.2002.0040, 2002.

Wu, T., Luo, J., Dong, W., Gao, L., Hu, X., Wu, Z., Sun, Y., and Liu, J.: Disaggregating County-Level Census Data for Population Mapping Using Residential Geo-Objects With Multi-source Geo-Spatial Data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1189–1205, https://doi.org/10.1109/JSTARS.2020.2974896, 2020.