REGION
The Journal of ERSA
Powered by WU

Volume 13, Number 2, 2026, R15–R27
DOI: 10.18335/region.v13i2.620

journal homepage: region.ersa.org
ISSN: 2409-5370

Geographic data conversion using grid-level data: The case of Austrian municipalities and ZIP code areas

Dieter Pennerstorfer[1]

[1] Johannes Kepler University Linz, Linz, Austria

Received: 11 July 2025/Accepted: 26 October 2026

Abstract. Researchers often need to convert variables between non-hierarchical regions. This article presents a method for calculating non-binary, population-based conversion weights for Austrian municipalities and ZIP code areas using 250-meter grid-level population data. Due to the increasing availability of small-scale grid data, this approach can be applied to other countries as well. To evaluate the accuracy of the data conversion, the article compares the converted data with the observed data for population, employment, retirees, unemployment, and supermarkets, finding median absolute percentage errors ranging from 0.25 % to 2.47 %. This approach substantially outperforms naive binary assignments, especially when the source regions are split relatively evenly among the target regions. Year-specific weights are provided for the years 2013 to 2020, along with implementation examples for Stata and R, to facilitate adoption.

Data availability: The data supporting this study have been deposited in Zenodo and are publicly available at https://doi.org/10.5281/zenodo.23474497.

1 Introduction

In regional economics, applied researchers often need data at a different level of spatial aggregation than what is available. There are various reasons for this. (i) Researchers may want to combine data available at different levels of aggregation (Arntz, Wilke 2007). (ii) The geographic boundaries of regional units change over time and one wants to construct a consistent time series of variables (Norman et al. 2003, Norman 2010, Logan et al. 2014, Vasan et al. 2018). (iii) Researchers may want to analyze spatial data at a different level of aggregation than the level at which the data are provided. This is important, because the results of regional analyses can be sensitive to the level of spatial aggregation, a phenomenon known as the 'Modifiable Area Unit Problem' (MAUP).[1] (iv) Regional entities differ vastly between countries.[2] For cross-country comparisons, researchers may wish to homogenize regional units to a standardized spatial structure (Netrdová et al. 2020). (v) Researchers may simply want to present data or results at a different level of aggregation than the data provided (Simpson 2002). Consequently, converting geospatial data from one regional aggregation level to another is a common challenge for researchers and practitioners, requiring a precise, easily applicable solution.

Geography conversion tables (Simpson 2002) are one practical solution to this issue, providing weights to convert variables from one geography to a different one. While

[1] For more information, see Openshaw, Taylor (1979).

[2] For example, the average population size of municipalities in the European Union in 2025 ranges from 1,710 in Czechia to 161,097 in Ireland (Prorok, Calzola 2026).

the conversion almost always involves approximation, the applied weighting criterion is essential for minimizing the errors in the transformation. The empirical literature has provided different ways to calculate these weights and evaluated its preciseness (see Section 2 for a literature discussion).

In this article, I use detailed information on the spatial distribution of the population at the 250 m grid cell level, as well as the spatial extent of municipalities and ZIP code areas, to calculate the necessary weights for converting variables between these two regional entities. I apply this method to convert variables from the municipal to the ZIP code level (and vice versa) in Austria from 2013 to 2020. Using data from both levels of aggregation, I can evaluate the quality of the data conversion, discuss sources of error, and quantify their impact on the conversion's precision. As part of this article, I am also providing the weights and the code for the statistical software programs Stata and R.

The purpose of this article is twofold. First, I contribute to the empirical literature on data conversion between different geographic areas by drawing on detailed grid-level population data. Comparing the converted data with the actual data suggests that this approach works reasonably well for a wide variety of applications. Note that population data at small-scale grid levels are becoming increasingly available in many developed countries. For instance, Eurostat and all national statistical agencies in EU member countries provide such data (Bergs, Budde 2022). Therefore, this approach can easily be used in many countries and applications. Second, I provide weights for converting variables between municipalities and ZIP code areas in Austria, similar to Simpson (2002) for the UK or Logan et al. (2014) for US census tracts, along with instructions for using them. Since much of the data is only available at one of the two regional levels of aggregation, these weights can be helpful for researchers and practitioners studying Austrian regions.

The remainder of the article is organized as follows: Section 2 outlines the methodology for calculating the weights for data conversion and reviews the relevant literature. Section 3 describes the data sources and addresses the challenges of using grid-level data for this purpose. Section 4 evaluates the error associated with data conversion, drawing on data available at both regional levels. Section 5 provides data accessibility and instructions for using the data in Stata and R, and Section 6 concludes.

2 Methodology

When converting variables from one level of regional aggregation to another, there is usually a finite number of source regions (S) and a finite number of target regions (T). No specific assumptions are made about the sizes of the source and target regions, the level of overlap, or the level of hierarchy. Suppose we have a count or frequency variable, y, that we observe only in the source regions (e.g., population, unemployment, or university graduates). To estimate the value of the variable in a target region t, denoted by $\widehat{y}_t$, we calculate a weighted sum of the variable's values in the source regions, s:

$$\widehat{y}_t = \sum_s \omega_{st} y_s \tag{1}$$

with ω_{st} as the corresponding weights, such that $\sum_t \omega_{st} = 1$.

This data conversion is exact, rather than an approximation, in two situations. First, the relationship between s and t is hierarchical, with region t as the higher level. In this case, $\omega_{st} = 1$ if region s is fully contained within region t, and 0 otherwise. Second, we observe the variable of interest y at each intersection of regions s and t, denoted by y_{st}. Following the literature (Gregory 2002, Gotway, Young 2002, Schroeder 2007), the intersection of regions s and t, st, is called an 'atom'. This is typically only possible if we have access to geo-referenced individual records (Simpson 2002). In this case, the weight, $\omega_{st} = y_{st}/y_s$, is the exact share of the variable y_s located in the target region t, and the data conversion is without error—we simply add up all the atoms belonging to the target region t:

$$\widehat{y}_t = \sum_s (y_{st}/y_s) y_s = y_t \tag{2}$$

As the true weights $\omega_{st} = y_{st}/y_s$ are usually not observed, they have to be replaced by estimated ones, $\widehat{\omega}_{st}$, in equation (1). There are different ways in the literature to estimate these weights. A simple method is to use the area of source region s, A_s, and the area of the atom of regions s and t, A_{st}, such that the weights $\widehat{\omega}_{st} = A_{st}/A_s$. This procedure is labeled as areal weighting (e.g., Goodchild, Lam 1980, Gregory, Ell 2005, Netrdová et al. 2020) and implicitly assumes that the variable of interest y is distributed uniformly in the source region (Gotway, Young 2002) or at least has the same density in each atom, which is a strong and usually unrealistic assumption (Gregory 2002).

To avoid this assumption of homogeneity within source regions, researchers use ancillary data to better predict the intra-regional distribution of the variable of interest to receive more accurate results when converting the variable between non-hierarchical regions. These ancillary data are usually available at finer spatial scale than the source region (Leyk et al. 2019) and this approach is commonly referred to as dasymetric mapping (e.g., Arntz, Wilke 2007, Leyk et al. 2019, Gregory 2002). The ancillary data can include land cover (Logan et al. 2014), imperviousness, road networks, and nighttime light (Zandbergen, Ignizio 2010), school locations (Zhang, Qiu 2011) or population density in the target region (Schroeder 2007), among others (see Mennis 2009, for an overview).[3]

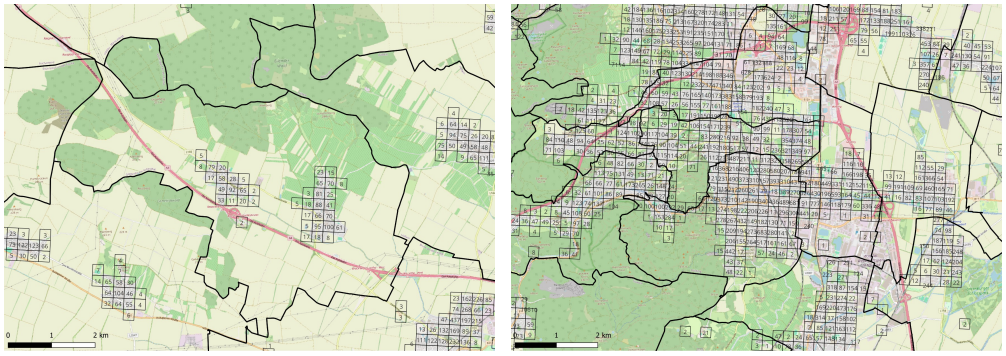
A related approach is to use a single supplementary variable, denoted by z, which (i) is available at a fine spatial scale and can be assigned to each atom, thus enabling the calculation of corresponding weights, and (ii) more precisely predicts the spatial distribution of the variable of interest than areas do. I use the term z_{st} if the supplementary variable can be precisely assigned to each atom, and I use the term $\widehat{z}_{st}$ if this process is imprecise, meaning that $\widehat{z}_{st}$ is only an estimate of the true but unobserved value z_{st}. Simpson (2002), Norman et al. (2003), and Norman (2010) use the number of residential addresses in postcode locations for this purpose. Since the information for each postcode includes the regions in which it lies across different geographies, it can be assigned to each atom. The error in data conversion will be the smaller the more precisely the supplementary variable can be located to the atoms (i.e., $corr(\widehat{z}_{st}, z_{st}) \to 1$) and the more strongly the supplementary data and the variable of interest are correlated (i.e., $corr(z_{st}, y_{st}) \to 1$). This leads to estimated weights $\widehat{\omega}_{st}$ that are similar to the true, but unobserved weights ω_{st}:

$$\widehat{\omega}_{st} = (\widehat{z}_{st}/\widehat{z}_s) \approx (y_{st}/y_s) = \omega_{st} \tag{3}$$

Please note that when converting variables in the form of shares, such as percentages, rather than count variables, the calculation of the weights changes to $\widehat{\omega}_{st} = \widehat{z}_{st}/\widehat{z}_t$ (Gregory 2002).

In this article, I take a similar approach, using small-scale, 250-meter by 250-meter grid-level population data to generate estimated weights. Grid-level population data is aggregated at the atom ($\widehat{pop}_{st}$) and at the source region level ($\widehat{pop}_s$) to calculate the weights $\widehat{\omega}_{st} = \widehat{pop}_{st}/\widehat{pop}_s$. Grid cells are ideal for calculating the weights and providing geographic conversion tables for three reasons. First, grid cells are small, so they can be assigned to each area with a reasonably small error, similar to UK postcodes or US census blocks. Second, small-scale grids are becoming increasingly available in many countries and are provided by national statistical agencies in all EU member countries (Bergs, Budde 2022). Thus, a similar approach can be applied in other countries as well. Third, statistical agencies typically provide socioeconomic variables in addition to population counts at the grid level. For example, Statistics Austria provides data on the resident population by age group and gender, occupational status, country of birth, number of workplaces, and number of self-employed and employed persons by economic sector. They also provide data on the number of buildings by construction period. In this application, I use the residential population data to calculate weights. However, if the population is only loosely correlated with the variable of interest, a variety of supplementary variables are available for data conversion purposes.

[3] Rather than converting variables directly, researchers can alternatively generate a smooth map of the geographical distribution of the variable of interest, called smooth pycnophylactic interpolation (see Tobler 1979, for details), or even simulate the geo-coordinates for each unit (Groß et al. 2020). These smooth maps or estimated locations can then be aggregated at any potential target region.



(a) Unambiguous assignment (b) Ambiguous assignment

Figure 1: Precision of assignment of grid cells to regions

Notes: The black lines indicate municipality borders. The numbers in the grid cells represent the residential population in 2015. Grid cells without a resident population are omitted.

3 Data sources and description

I apply the proposed procedure to Austrian municipalities and ZIP code areas as source and target regions. Three data sources are used to calculate the weights needed to convert variables between these regions. First, information on the spatial distribution of the population at the 250 m grid cell level is provided by Statistics Austria for the year 2015. Statistics Austria places these grid units over the entire territory of Austria, and the cells are independent of administrative boundaries. Each person is assigned to exactly one cell based on the postal address of their primary residence ('Hauptwohnsitz').[4] Information on the residential population is also provided at both municipality and ZIP code level. Out of 8,584,926 inhabitants, only 4,609 (0.05 %) could not be assigned to a grid cell.

Second, annual maps of the geographic locations of Austrian municipalities are provided by Statistics Austria annually for the period between 2013 and 2020 via their website.[5] According to the Nomenclature of Territorial Units for Statistics (NUTS) of the EU (see Eurostat 2019, for details), municipalities correspond to the level of local administrative units (LAU).[6] Third, annual data on ZIP code areas are provided by the company GeoMagis for the period between 2013 and 2020 at the grid cell level.

As described in Section 2, one source of error in converting variables is that the supplementary variable (in this case, the population) cannot be precisely assigned to the atoms. This is the case if a grid cell is not fully contained in both one municipality and one ZIP code area. Figure 1 shows two map sections where all populated grid cells can be allocated to a single municipality (left panel) and a suburban area where the allocation of grid cells to municipalities is ambiguous for a substantial number of cells. In total, 11.06 % of all populated grid cells with 9.59 % of population could not be unambiguously assigned to municipalities.

There are different methods for assigning grid cells to regions when they are not entirely contained within one region. Grid cells can be assigned to regions based on their centroid or the largest proportion of their area within a region. Alternatively, cells can be assigned to regions based on their area overlap (areal weighting), or ambiguous cells

[4] Anyone moving into or out of a residence in Austria must report this to the relevant registration office within three days. Therefore, the information provided should be complete and up to date. The population includes individuals living in private households (98.37 % of the population) and in institutional households (e.g., student dormitories, nursing homes, monasteries, prisons; 1.60 %), as well as homeless individuals who are registered as homeless or recorded as homeless through a facility of a homeless assistance center (0.03 %, see Statistics Austria 2024). However, the homeless population is likely to be underreported (Musil et al. 2024).

[5] The maps are available on this website: https://data.statistik.gv.at/web/meta.jsp?dataset=-OGDEXT_GEM_1.

[6] Although Vienna is a single municipality and a single LAU, it is included at the district level. It is divided into 23 districts.

Table 1: Precision of aggregating grid cells to regions

region	number of units	type of allocation	absolute error			absolute percent error		
			median	p75	p90	median	p75	p90
ZIP code area	2,212	centroid	11.84	34.68	86.98	0.63	1.68	3.72
municipality	2,124	centroid	13.91	38.37	99.38	0.73	1.81	4.02
municipality	2,124	largest area	13.43	37.64	95.60	0.69	1.72	3.61
municipality	2,124	areal weighting	13.76	35.34	85.80	0.71	1.77	3.65
municipality	2,124	ambiguous excluded	117.51	238.08	529.24	7.23	10.22	14.50

Notes: The mean (median) population of ZIP code areas and municipalities is 3,881.07 (1,716) and 4,041.87 (1,807), respectively.

can be excluded. In any case, the population of the grid cells is scaled so that the total population of cells used in the analysis equals the total regional population.

Table 1 compares the actual number of inhabitants with the population resulting from allocating grid cells to regions. Since maps (shapefiles) are available for both municipalities and grid cells, all of the aforementioned options for linking grid cells to municipalities (i.e., matching by centroid or largest area, areal weighting, and excluding ambiguous cells) can be implemented. These processes were carried out using the geographic information system QGIS (version 3.44). Due to data restrictions, grid cells can only be assigned to ZIP code areas via their centroids.[7] The median absolute and percentage errors when aggregating grid cell populations to the ZIP code level are approximately 12 inhabitants and 0.63 %, respectively. These errors increase to 1.68 % and 3.72 %, respectively, at the 75^{th} and 90^{th} percentiles; however, they remain reasonably small for most regions. The error for municipalities is similar when cells are assigned based on their centroids, the largest proportion of their area, or area overlap (areal weighting). However, excluding ambiguous cells substantially worsens the outcome. The allocation would be more precise if the grid cells were smaller or the regions under scrutiny were larger (see Simpson 2002, for a discussion). In what follows, I will assign grid cells to municipalities based on their centroids, consistent with their allocation to ZIP code areas.

4 Conversion quality

In order to evaluate the quality of the data conversion, I use the methodology discussed in Section 2 and the weights specified there to convert variables that are available at both the Austrian ZIP code area level and the municipal level. This allows me to compare the converted variables $\widehat{y}_t$ with the observed variables y_t in the target regions.

Austrian municipalities and ZIP code areas are similar in size, with an average (median) population of approximately 4,000 (1,800) in 2015. However, only about 30 % of all municipalities and ZIP code areas were identical. To compare the geographic overlap between these regions further, I follow Simpson (2002) and calculate the degree of hierarchy and fit between these regional entities. The degree of hierarchy is defined as the share of source regions fully contained within a target region and is calculated as $\sum_{s,t}(\omega_{st} = 1)/S$. The degree of hierarchy is 55 % for ZIP code areas and 54 % for municipalities as source regions. However, many source regions are almost fully contained within a target region. The degree of fit recognizes this and is calculated as $\left(\sum_s \max_t \omega_{st}\right)/S$. The degree of fit is substantially higher, taking values of 94 % and 93 % for ZIP code areas and municipalities, respectively, as source regions. This suggests that most source regions have a dominant share contained in a single target region.

Table 2 shows the error associated with converting the residential population between ZIP code areas and municipalities when using 'non-binary' weights $\widehat{\omega}_{st} = \widehat{pop}_{st}/\widehat{pop}_s$. Since the weights depend on the variable of interest, errors only stem from populations in grid cells being incorrectly assigned to regions because they are not fully contained within a ZIP code area or municipality, respectively, as discussed in Section 3. The results

[7] Rather than providing maps of ZIP code areas, GeoMagis provides the ZIP code area of each populated grid cell in tabular form based on the cell's centroid. Grid cells whose centroids are located outside of Austria, comprising 908 inhabitants (0.01 % of the total population), are excluded from the analysis.

Table 2: Conversion error for different region types

type	# target regions	weights	median pop.	absolute error			absolute percent error		
				median	p75	p90	median	p75	p90
ZIP code area (source region) ⟶ municipality (target region)									
all	2,124	non-binary	1,807	5.17	22.83	63.83	0.25	1.11	2.73
all	2,124	binary	1,807	53.00	420.00	1,129.00	2.27	20.86	100.00
clear	696	non-binary	1,611	0.00	0.00	2.00	0.00	0.00	0.07
clear	696	binary	1,611	0.00	0.00	2.00	0.00	0.00	0.07
rel. clear	692	non-binary	2,049	11.79	30.99	73.71	0.54	1.38	2.79
rel. clear	692	binary	2,049	53.00	166.00	411.00	2.32	6.95	20.81
ambiguous	736	non-binary	1,868	17.04	43.65	94.11	0.82	2.07	4.38
ambiguous	736	binary	1,868	586.50	1,226.00	2,040.00	32.77	100.00	100.00
municipality (source region) ⟶ ZIP code area (target region)									
all	2,212	non-binary	1,716	4.99	21.39	58.45	0.26	1.10	2.57
all	2,212	binary	1,716	68.00	487.00	1,143.00	3.30	31.59	100.00
clear	689	non-binary	1,617	0.00	0.00	2.00	0.00	0.00	0.07
clear	689	binary	1,617	0.00	0.00	2.00	0.00	0.00	0.07
rel. clear	663	non-binary	2,050	10.07	30.01	74.90	0.50	1.42	2.82
rel. clear	663	binary	2,050	49.00	136.00	342.00	2.32	7.12	100.00
ambiguous	860	non-binary	1,637	13.59	36.23	76.39	0.77	1.80	3.79
ambiguous	860	binary	1,637	588.00	1,112.50	1,829.50	39.07	100.00	100.00

Notes: Conversion errors are calculated for all target regions and separately for 'clear', 'relatively clear', and 'ambiguous' target regions. 'Clear' cases include only target regions t for which all weights $\widehat{\omega}_{st} \in \{0, 1\}\ \forall\ s$, i.e., those with a hierarchical relationship to the source region. 'Ambiguous' cases are target regions with at least one source region such that $\widehat{\omega}_{st} \in (0.1, 0.9)$, meaning that there is at least one source region where a substantial share is assigned to the target region t as well as to other target regions. All other cases are classified as 'relatively clear'. When using 'binary' weights, each source region is assigned to exactly one target region based on the largest weight. Non-binary weights are based on the population shares in each atom, $\widehat{\omega}_{st} = \widehat{pop}_{st}/\widehat{pop}_{s}$.

show that the median absolute error is approximately five people or 0.25 %, irrespective of whether the variable is converted from ZIP code areas to municipalities or vice versa. The error increases to 1.11 % and 1.10 % for the 75$^{\text{th}}$ percentile and to 2.73 % and 2.57 % (depending on the direction of the conversion) for the 90$^{\text{th}}$ percentile, suggesting that the conversion is relatively precise in most cases.

A simple and commonly used ad hoc method is to use binary weights and assign each source region to exactly one target region based on the greatest intersection (Goodchild, Lam 1980), which Arntz, Wilke (2007) refer to as 'naive binary weights'. When the source regions are fully assigned to target regions based on the largest weight, i.e. $\widehat{\omega}_{st} = 1$ if $\widehat{pop}_{st}/\widehat{pop}_{s} > \widehat{pop}_{sk}/\widehat{pop}_{s}\ \forall\ k \neq t$, the absolute median error increases about tenfold. Approximately 11 % of all municipalities and 15 % of all ZIP code areas are not assigned a single source region, resulting in an estimated value of $\widehat{y}_t = 0$ and an error of 100 %. Despite the relatively high fit between ZIP code areas and municipalities, the error associated with converting variables using binary weights is considerable.

The error when converting variables between non-hierarchical regions depends on whether and to what extent the source regions are divided among the target regions. I classify the target regions into three categories: Target regions consisting only of source regions that are fully contained (i.e., $\widehat{\omega}_{st} \in \{0, 1\}\ \forall\ s$) are referred to as 'clear' cases in Table 2. If at least one weight for a target region falls between 0.1 and 0.9, the target regions are classified as 'ambiguous'. In this case, for at least one source region, a relatively large proportion of at least 10 % is located both within and outside the target region. A relatively high degree of approximation may be involved in data conversion for this group. The remaining observations are classified as 'relatively clear' cases. For the clear cases, the errors are identical for non-binary and binary weights (by definition) and are only marginally larger than zero in a few cases, arising from inaccurate grid cell assignment to atoms. However, the error for the ambiguous cases differs greatly depending on whether non-binary or binary weights are used. While the median error increases slightly to 0.82 % and 0.77 % with non-binary weights, it is as high as 32.77 % and 39.07 % with binary weights, depending on the direction of the data conversion. Thus, naive binary weights perform poorly when the source regions are split (and particularly

relatively evenly) between target regions, while the non-binary, population-based weights still perform reasonably well.

To further evaluate the accuracy of the data conversion, I use variables other than population count that are available at both the ZIP code and municipality levels. I use data on employed, unemployed, and retired residents from 2015, which were provided by Statistics Austria.[8] Furthermore, I use data on the number of supermarkets in both regions. Information on the locations of supermarkets at the individual level was collected in 2015 by Herold Business Data GmbH, a telephone directory containing firm addresses and industry codes. This information was then cross-validated through general Internet research (see Pennerstorfer, Yontcheva 2021, for details). Each geocoded address is linked to both ZIP codes and municipalities.

The accuracy of the data conversion is reported in Table 3. As a benchmark, the error for converting the residential population is stated first, while the other variables are sorted by frequency. When focusing on converting data from ZIP code areas to municipalities, as well as on the 75$^{\text{th}}$ percentile, the error increases modestly from 1.11 % for population to 1.89 % for employed residents and to 2.73 % for retirees. The error associated with converting the number of unemployed individuals is substantially higher, reaching 6.84 % for the 75$^{\text{th}}$ percentile. Please note that the values of this variable are much smaller than those previously discussed, with a median of just 45, and it is less strongly correlated with the population.[9] For supermarkets, with a median of only one outlet per municipality, the error is more than twice as high (13.74 %). The results on the accuracy are qualitatively similar when converting data from municipalities to ZIP code areas. These results suggest that the data conversion performs particularly well when the variable of interest is strongly correlated with the population (or generally, the variable used to calculate the weights) and occurs in large numbers (employed residents and retirees). The error is substantially larger if the correlation between the population and the variable being converted is weaker, as is the case with the unemployed, or if the variable mostly takes on small values, as is the case with supermarkets.

Table 3: Conversion error for different variables

variable	median	absolute error			absolute percent error		
		median	p75	p90	median	p75	p90
ZIP code area (source region) ⟶ municipality (target region)							
population	1,807	5.17	22.83	63.83	0.25	1.11	2.73
employed	920	5.33	19.66	45.09	0.48	1.89	4.79
retirees	397	3.48	12.12	30.90	0.70	2.73	7.02
unemployed	45	0.75	3.49	10.13	1.47	6.84	20.15
supermarkets	1	0.03	0.29	0.80	1.88	13.74	34.65
population (2020)	1,843	7.21	28.84	70.96	0.34	1.37	3.36
municipality (source region) ⟶ ZIP code area (target region)							
population	1,716	4.99	21.39	58.45	0.26	1.10	2.57
employed	873	5.36	19.73	44.80	0.55	2.07	5.45
retirees	377	3.81	13.42	29.97	0.81	3.37	8.47
unemployed	44	0.81	3.47	9.05	1.59	7.47	21.49
supermarkets	1	0.04	0.29	0.73	2.47	15.24	36.21
population (2020)	1,737	7.81	30.09	73.66	0.42	1.62	4.03

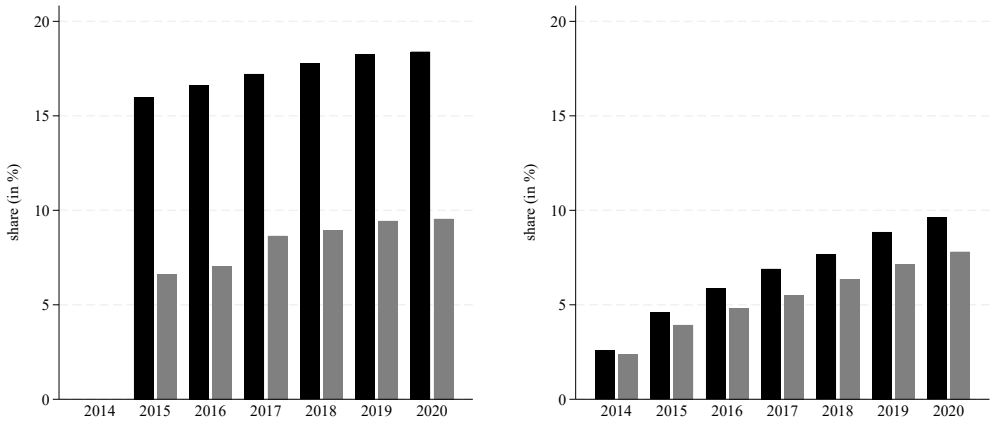
Notes: Unless otherwise stated, all variables are from 2015. To calculate the absolute percentage error, it was necessary to exclude regions with zero cases of the relevant variable. For retirees, this includes two ZIP code areas. For the unemployed, this includes six ZIP code areas and three municipalities. For supermarkets, this includes 581 ZIP code areas and 551 municipalities.

The final variable is the population of ZIP code areas and municipalities in 2020, rather than 2015. This information was also provided by Statistics Austria. The 2015 weights cannot be used directly, because both municipalities and ZIP code areas change

[8] Please note that the grid-level population data, grid cell allocation to ZIP codes, and municipality maps are from January 1, 2015, while the data on (un)employment and retirees are from October 31, 2015. Regions affected by boundary changes since January 1 are excluded from the corresponding data conversion.

[9] The correlation between population and the number of unemployed is approximately 0.95 for both ZIP code areas and municipalities, while the correlation between population and the number of employed residents or retirees is greater than 0.99.

over time. Between 2013 and 2020, the number of municipalities decreased substantially, dropping from 2,376 to 2,117. Conversely, the number of ZIP code areas increased slightly, rising from 2,203 to 2,221. In addition to changes in the number of entities, the borders of regional units also vary over time. Figure 2 shows the percentage of regional units that changed their borders compared to 2013, as well as the population affected by these changes. For example, approximately 10 % (8 %) of the population in 2020 resided in a municipality (ZIP code area) that did not correspond to the regional unit in 2013. While a reform of the municipal structure in Styria (see Stötzer et al. 2020, for details) dominated the time series on municipalities, reducing the number of municipalities in this state from 542 to 287 in 2015, the ZIP code areas were found to adjust more continuously over time.



(a) Municipalities (b) ZIP code areas

Figure 2: Changes in regional entities over time

Notes: The black bars show the percentage of municipalities (left panel) and ZIP code areas (right panel) that have changed their borders since 2013. Gray bars indicate the percentage of the population living in these regional entities.

However, if the population distribution does not change substantially, it may be possible to calculate the weights using outdated grid-level population data, as well as the current geographical extent of the source and target regions. Table 3 reports the conversion error for the 2020 population. The results show that using outdated grid-level population data from 2015 increases the error only marginally. For instance, the median error rises from 0.25 % to 0.34 % when converting population data from ZIP code areas to municipalities and from 0.26 % to 0.42 % when converting it in the opposite direction. These results suggest that using grid data from a different time period will introduce only a small additional error, provided that the time difference is not too large. This is an important finding because grid-level data may be more expensive to obtain than regional maps, which are often available free of charge. Therefore, I provide weights based on the 2015 grid-level population, as well as the current boundaries of ZIP code areas and municipalities, for all years between 2013 and 2020, not just 2015.

5 Data availability and implementation in Stata and R

In this section, I provide the necessary data for converting variables and give examples of implementation in the statistical software programs Stata and R. I focus on these programs because Stata is widely used for social science data analysis, has simple and intuitive syntax, and has comprehensive documentation. I also include R because it is fully open source, its user community and popularity are growing rapidly, and it has a large ecosystem (see Gordon 2023, Preface, for a discussion on the relative strengths of both programs).

The data required to convert variables between ZIP code areas z and municipalities m have been deposited in Zenodo and are publicly available (see Pennerstorfer 2026). They are also linked in Table 4. The data are aggregated at the atom level, i.e., at the intersection between these two regional entities, mz, for each year from 2013 to 2020. The data are available in Stata format and named 'weights_municipalities_zip_2013.dta' to 'weights_municipalities_zip_2020.dta'. Combinations of municipalities and ZIP code areas without common inhabitants (where $\widehat{pop}_{mz} = 0$ and thus $\widehat{\omega}_{zm} = \widehat{\omega}_{mz} = 0$) are omitted for brevity. Relevant information on the data can be accessed by using the Stata command `notes` or in the following data documentation. To convert a variable from the ZIP code level to the municipality level in 2015, for example, the ID for the ZIP code areas should be named 'plz_2015' ('Postleitzahl—PLZ'). In the opposite case, the variable comprising the municipality IDs should be named 'gkz_2015' and should include the official ID ('Gemeindekennziffer—GKZ'), which is assigned by Statistics Austria.

Below, I provide an example of how to convert a count variable from one regional level of aggregation to another. Suppose you have a variable labeled '`unemployed_zip`' that represents the number of unemployed people in each ZIP code in 2015. The data file 'unemployed_2015_zip.dta' includes this variable, as well as a variable to identify the ZIP codes, named '`plz_2015`'. The weights for the data conversion are provided in Table 4 with 'weights_municipalities_zip_2015.dta' as the data file and '`weight_z_to_m_2015`' as the weights relevant for 2015. The following code can be used to import the unemployment variable at the ZIP code level, link it to the weights, and convert it to the municipality level:

Stata: ZIP code areas ⟶ municipalities

```
use unemployed_2015_zip.dta, clear
merge 1:m plz_2015 using weights_municipalities_zip_2015.dta
notes
gen unemployed_mun_tilde = unemployed_zip * weight_z_to_m_2015
collapse (sum) unemployed_mun_tilde, by(gkz_2015)
```

R: ZIP code areas ⟶ municipalities

```
library(haven)
library(dplyr)

unemployed_zip <- read_dta("unemployed_2015_zip.dta")
weights_2015 <- read_dta("weights_municipalities_zip_2015.dta")
merged <- unemployed_zip |>
    left_join(weights_2015, by = c("plz_2015" = "plz_2015"))
merged <- merged |>
    mutate(unemployed_mun_tilde = unemployed_zip *
       weight_z_to_m_2015)
result <- merged |>
    group_by(gkz_2015) |>
    summarise(unemployed_mun_tilde = sum(unemployed_mun_tilde,
       na.rm = TRUE), .groups = "drop")
```

The variable '`unemployed_mun_tilde`' denotes the number of unemployed converted from the ZIP code to the municipality level.

On the other hand, suppose that the information on the number of unemployed is available at the municipal level, labeled '`unemployed_mun`', and stored in 'unemployed_2015_mun.dta', along with a municipality ID variable named '`gkz_2015`'. The following code can be used to import the municipal-level variable, link it to the weights '`weight_m_to_z_2015`' from data file 'weights_municipalities_zip_2015.dta' provided in Table 4, and convert it to the ZIP code level:

Stata: municipalities ⟶ ZIP code areas

```
use unemployed_2015_mun.dta, clear
merge 1:m gkz_2015 using weights_municipalities_zip_2015.dta
notes
gen unemployed_zip_tilde = unemployed_mun * weight_m_to_z_2015
collapse (sum) unemployed_zip_tilde, by(plz_2015)
```

R: municipalities ⟶ ZIP code areas

```
library(haven)
library(dplyr)

unemployed_mun <- read_dta("unemployed_2015_mun.dta")
weights_2015 <- read_dta("weights_municipalities_zip_2015.dta")
merged <- unemployed_mun |>
    left_join(weights_2015, by = c("gkz_2015" = "gkz_2015"))
merged <- merged |>
    mutate(unemployed_zip_tilde = unemployed_mun *
       weight_m_to_z_2015)
result <- merged |>
    group_by(plz_2015) |>
    summarise(unemployed_zip_tilde = sum(unemployed_zip_tilde,
       na.rm = TRUE), .groups = "drop")
```

The variable '`unemployed_zip_tilde`' denotes the number of unemployed converted from the municipal to the ZIP code level.

If the variables to be converted are shares rather than counts (e.g., percentages), the weights `weight_m_to_z_2015` must be used when converting variables from ZIP code areas to municipalities, and the weights `weight_z_to_m_2015` must be used when converting variables from municipalities to ZIP code areas (see Section 2). Alternatively, one can first transform the share variables into levels and then apply the above-outlined procedure to convert the information from one regional entity to another.

Table 4: Data access: weights for data conversion

Year	Resource
2013	weights_municipalities_zip_2013.dta
2014	weights_municipalities_zip_2014.dta
2015	weights_municipalities_zip_2015.dta
2016	weights_municipalities_zip_2016.dta
2017	weights_municipalities_zip_2017.dta
2018	weights_municipalities_zip_2018.dta
2019	weights_municipalities_zip_2019.dta
2020	weights_municipalities_zip_2020.dta

6 Summary and conclusions

This article presents a method for calculating population-based conversion weights for Austrian ZIP code areas and municipalities using 250 m grid cell population data. This grid-level population data is assigned to both regional levels to calculate the population at each intersection of ZIP code areas and municipalities in order to determine the weights for the data conversion. This approach is based on the assumption that the spatial distribution of the variable of interest is proportional to the spatial distribution of the residential population.

Validation against observed data in both geographies shows high accuracy for variables closely aligned with population size. For example, the median absolute percentage

error when converting the number of employed individuals from ZIP code areas to municipalities is only 0.48 %. However, the error is higher for variables with a weaker correlation to population size, such as unemployment, as well as for variables that mostly take small values, such as the number of supermarkets. Conversion accuracy decreases when source regions are split relatively evenly among target regions. This suggests that applying the proposed approach to regions with a lower degree of hierarchy and fit than Austrian ZIP code areas and municipalities may result in higher conversion errors. Interestingly, using outdated grid data to calculate the weights only marginally decreases the precision of the data conversion. The validation exercise suggests that this approach clearly outperforms naive binary assignments. To facilitate adoption in the specific case of Austria, year-specific weights are provided from 2013 to 2020, along with implementation examples for Stata and R.

Converting variables between two non-hierarchical regions is a common challenge in regional economics. It occurs when researchers want to combine data from different regional levels (Arntz, Wilke 2007), construct consistent time series when regional boundaries change (Logan et al. 2014), or analyze data at different levels of spatial aggregation. In this article, I demonstrate that using grid cell data to calculate weights for data conversion works reasonably well. Fine-scale grid data have two advantages: they are standardized, and they are becoming increasingly available in many countries (Bergs, Budde 2022). This makes it possible to apply the proposed approach in other countries as well. Validation exercises demonstrate that conversions are most accurate when the variable of interest strongly correlates with the weighting variable (e.g., population). Since grid data provided by statistical agencies often include other socioeconomic variables, these variables can be used instead of population data to construct weights, if appropriate, which can further increase precision.

Acknowledgements

I thank Astrid Pennerstorfer for sharing the data for this project.

References

Arntz M, Wilke RA (2007) An application of cartographic area interpolation to German administrative data. *AStA Advances in Statistical Analysis* 91: 159–180. CrossRef

Bergs R, Budde R (2022) The potential of small-scale spatial data in regional science. *Review of Regional Research* 42: 97–110. CrossRef

Eurostat (2019) *Methodological manual on territorial typologies: 2018 edition*. Publications Office of the European Union. CrossRef

Goodchild MF, Lam NSN (1980) Areal interpolation: A variant of the traditional spatial problem. *Geo-processing* 1: 297–312

Gordon RA (2023) *Applied statistics for the social and health sciences*. Routledge. CrossRef

Gotway C, Young L (2002) Combining incompatible spatial data. *Journal of the American Statistical Association* 97: 632–648. CrossRef

Gregory IN (2002) The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26: 293–314. CrossRef

Gregory IN, Ell PS (2005) Breaking the boundaries: Geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society Series A: Statistics in Society* 168: 419–437. CrossRef

Groß M, Kreutzmann AK, Rendtel U, Schmid T, Tzavidis N (2020) Switching between different non-hierachical administrative areas via simulated geo-coordinates: A case study for student residents in Berlin. *Journal of Official Statistics* 36: 297–314. CrossRef

Leyk S, Gaughan AE, Adamo SB, de Sherbinin A, Balk D, Freire S, Rose A, Stevens FR, Blankespoor B, Frye C, Comenetz J, Sorichetta A, MacManus K, Pistolesi L, Levy M, Tatem AJ, Pesaresi M (2019) The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth System Science Data* 11: 1385–1409. CrossRef

Logan JR, Xu Z, Stults BJ (2014) Interpolating us decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer* 66: 412–420. CrossRef

Mennis J (2009) Dasymetric mapping for estimating population in small areas. *Geography Compass* 3: 727–745. CrossRef

Musil R, Schnell P, Dlabaja C (2024) Machbarkeitsstudie: Datenbasis zu Obdachlosigkeit, Wohnungslosigkeit und prekärem Wohnen. Bundesministerium für Soziales, Gesundheit, Pflege und Konsumentenschutz (BMSGPK)

Netrdová P, Nosek V, Hurbánek P (2020) Using areal interpolation to deal with differing regional structures in international research. *ISPRS International Journal of Geo-Information* 9. CrossRef

Norman P (2010) Identifying change over time in small area socio-economic deprivation. *Applied Spatial Analysis and Policy* 3: 107–138. CrossRef

Norman P, Rees P, Boyle P (2003) Achieving data compatibility over space and time: Creating consistent geographical zones. *International Journal of Population Geography* 9: 365–386. CrossRef

Openshaw S, Taylor P (1979) A million or so correlated coefficients: Three experiment on the modifiable areal unit problem. In: Wrigley N (ed), *Statistical applications in spatial sciences*. Pion

Pennerstorfer D (2026) Geographic data conversion using grid-level data: The case of Austrian municipalities and ZIP code areas. Supplementary data, version v1, Zenodo. CrossRef

Pennerstorfer D, Yontcheva B (2021) Local market definition in competition analysis: An application to entry models. *Economics Letters* 198: 109678. CrossRef

Prorok T, Calzola S (2026) European local government report 2025: Structures, finances and functions of local governments in the European Union. KDZ – Centre for Public Administration Research, Vienna, Austria, https://www.kdz.eu/system/files/downloads/2026-03/European Local Government Report 2025.pdf

Schroeder JP (2007) Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39: 311–335. CrossRef

Simpson L (2002) Geography conversion tables: A framework for conversion of data between geographical units. *International Journal of Population Geography* 8: 69–82. CrossRef

Statistics Austria (2024) Machbarkeitsstudie Nutzung von Verwaltungsregistern zur zeitnahen Bestimmung von Personengemeinschaften insbesondere für zielgerichtete Förderungen. Statistics Austria

Stötzer S, Stangl M, Brödl C (2020) Auswirkungen von Gemeindestrukturreformen – Eine Einzelfallstudie einer Gemeindefusion in der österreichischen Steiermark. *dms – der moderne staat – Zeitschrift für Public Policy, Recht und Management* 13: 217–240. CrossRef

Tobler WR (1979) Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74: 519–530. CrossRef

Vasan S, Baker J, Alcántara A (2018) Use of kernel density and raster manipulation in GIS to predict population in New Mexico census tracts. *Review of Economics Finance* 14: 25–38. CrossRef

Zandbergen PA, Ignizio DA (2010) Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science* 37: 199–214. CrossRef

Zhang C, Qiu F (2011) A point-based intelligent approach to areal interpolation. *The Professional Geographer* 63: 262–276. CrossRef

© 2026 by the authors. Licensee: REGION – The Journal of ERSA, European Regional Science Association, Milan, Italy. This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).