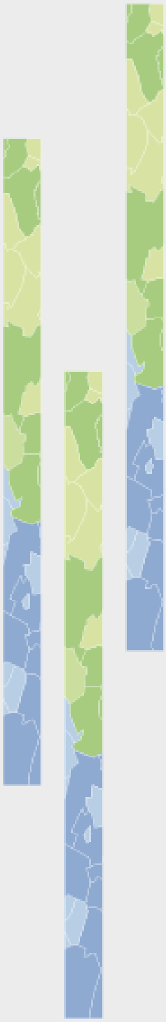
TECHNICAL REPORT

ESPON



USING DOWNSCALED POPULATION IN LOCAL DATA GENERATION

A COUNTRY-LEVEL EXAMINATION

CONTENT

- Research Context and Approach. This part outlines the background to and methodology of the examination of downscaled population data.
- A Country-Level Examination. This part presents the results of a Swedish examination of downscaled population, focusing on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.
- Summary and Discussion. This part points out that while there are obvious limitations to downscaled population data, it is a quite reasonable tool for certain purposes. In particular, fairly good estimations of UMZ population can be obtained.

ESPON 2013 DATABASE
MARCH 2011

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

16 pages

LIST OF AUTHORS

Magnus Strömgren, Dept. of Social and Economic Geography, Umeå University

Einar Holm, Dept. of Social and Economic Geography, Umeå University

Contact

magnus.stromgren@geography.umu.se
einar.holm@geography.umu.se

tel. + 46 90 786 52 58

TABLE OF CONTENTS

Introduction

In the ESPON db context, there is a need to utilize or present population data with a high degree of spatial resolution. For instance, in the disaggregation of socioeconomic data to grid level, detailed local population data is required for a proper downscaling of certain variables. Similarly, in reporting population figures for geographical subdivisions such as Urban Morphological Zones (UMZs), NUTS or even Local Administrative Units (LAU), level 2 population figures won't suffice. The approach that has been taken is to make use of downscaled population data—a population grid produced by the Joint Research Centre (JRC). This dataset, "Population density disaggregated with CORINE land cover 2000", distributes LAU, level 2 population data to a grid, mainly using CORINE land cover data.

However, there are a limited number of tests of the suitability of using the population grid for different purposes, as well as of its reliability in different settings. This technical report presents the results of a country-level examination of the population grid, using Swedish register population data.

1 Research Context and Approach

In addition to exploring the role of survey data, an important ESPON db activity for the Department of Social and Economic Geography, Umeå University is to carry out comparisons between Swedish data and data with EU coverage. The department has access to Swedish register data, which not only covers the entire population of Sweden for a substantial time period, but also has a high degree of spatial resolution. This resource makes possible a broad range of exploratory studies and evaluations.

This technical report presents the results of a country-level examination of downscaled population for the EU. In the study, Swedish register population data is used to examine the JRC population grid "Population density disaggregated with CORINE land cover 2000". The population grid—which allocates LAU, level 2 2001 census population data to 100 m^2 squares, mainly using CORINE land cover data—is an important tool in the ESPON db project. First, it is part of the workflow to disaggregate socioeconomic data into a grid structure. This is presented in more detail in the technical report "Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase". Second, it is utilized in order to assign population to Urban Morphological Zones (UMZs).

However, the suitability of using the population grid for different purposes, as well as its reliability in different settings, has not been subject to much scrutiny. Still, some validations of the population grid have been performed. For instance, a comparison with Austrian reference data at the km^2 level showed an overall reduction by 50 percent in the disagreement with reference data, when compared to a non-weighted distribution of the population.[1] Against this background, it is not without interest to examine how the population grid compares to Swedish register data.

1.1 Methodology

Since the population grid departs from population figures for LAU, level 2—in the Swedish case, municipalities—grid population summarized at that level can be expected to largely correspond to register data. However, at the local level, it may be more or less reliable. Similarly, the performance of the grid in estimating population figures for other geographical subdivisions (e.g., UMZs) is unclear. Taking into account how the population grid is employed in the ESPON db project, the examination focuses on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.

The basis for the **first test**, concerning grid population estimates in varying local settings, is a calculation of residuals. This is carried out by comparing an aggregation of the population grid to square kilometers with corresponding Swedish register data. Absolute residuals are then summarized at the municipality level (n=290). In addition to looking at the results per municipality, results are also categorized by "municipality groups"—a classification of Sweden's municipalities in nine different groups, created in 2005 by the Swedish Association of Local Authorities and Regions. The municipality group classification aims at defining homogenous regions, which share similar characteristics in terms of for instance population size, commuting patterns and

[1] Gallego J., Downscaling population density in the European Union with a land cover map and a point survey, JRC-Ispra.

employment profile. There are nine municipality groups, presented in Figure 1 and Table 2. This first test considers not only the absolute residual sum, but also the residual sum in relation to municipality area (expressed in km^2) and population size. When relating residual sum to population size, initial figures are multiplied by 50 in order to get a more gini-style estimation of the overall discrepancy between grid and register data.

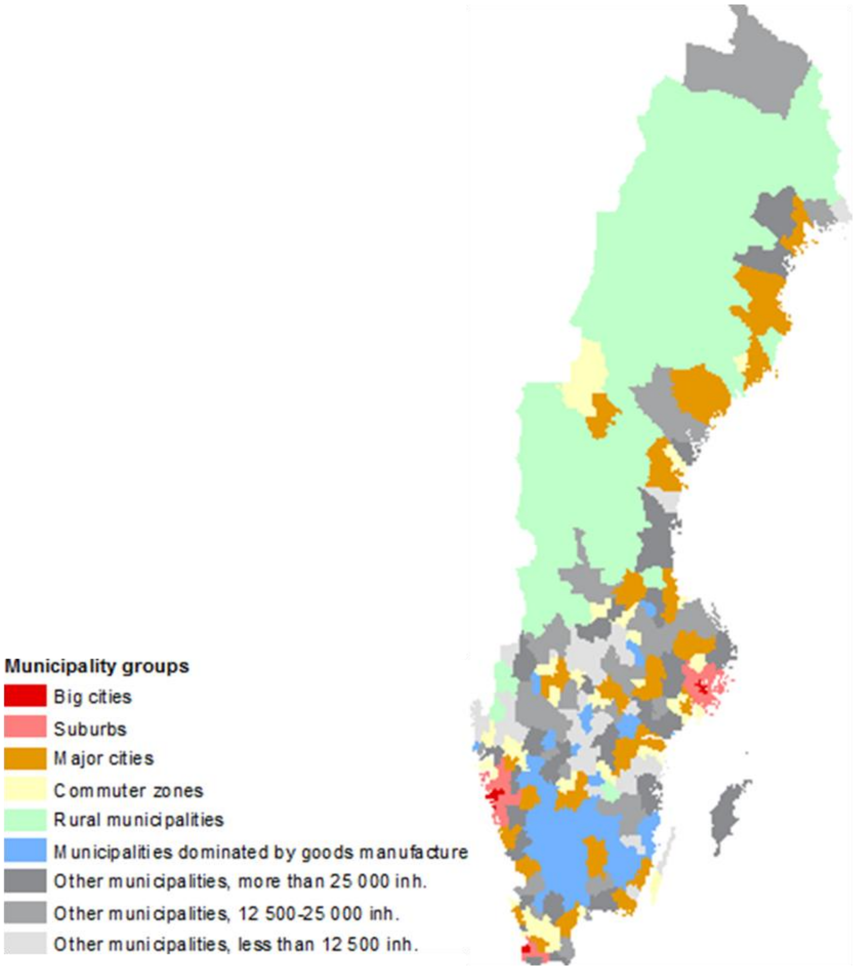

Municipality groups
Big cities
Suburbs
Major cities
Commuter zones
Rural municipalities
Municipalities dominated by goods manufacture
Other municipalities, more than 25 000 inh.
Other municipalities, 12 500-25 000 inh.
Other municipalities, less than 12 500 inh.

Figure 1: *Municipality group map*

ID	Category	Number of municipalities
1	Big cities	3
2	Suburbs	38
3	Major cities	27
4	Commuter zones	41
5	Rural municipalities	39
6	Municipalities dominated by goods manufacture	40
7	Other municipalities, more than 25,000 inhabitants	34
8	Other municipalities, 12,500–25,000 inhabitants	37
9	Other municipalities, less than 12,500 inhabitants	31

Table 1*: Municipality groups: IDs and frequencies*

The **second test** concerns using the population grid to estimate the population of Urban Morphological Zones (UMZs)—a delimitation of urban areas with EU coverage. In this test, overall population figures for each UMZ are calculated using both the original population grid and register data, which then are used for calculation of per-UMZ residuals. Thus, in contrast to the first test—which is based on the sum of absolute square residuals—this test focuses is the overall predictive capabilities of the grid, when it comes to UMZs of different sizes. Grid residuals within each UMZ have also been produced, primarily for purposes of trying to clarify patterns of over- and underestimation.

2 A Country-Level Examination

2.1 Population Estimates in Varying Local Settings

The first test of the population grid concerns population estimates in varying local settings. Clearly, the way discrepancies between grid and register data is associated to the local context depends on whether absolute residuals are just summarized or related to area or population. In Figure 2, the ten municipalities with highest (red) and lowest (green) absolute residuals are displayed, using three different measures: residual sum (left) as well as residual sum related to municipality area (middle) and population size (right). Municipalities with a comparatively large population (e.g., Malmö) tend to fare quite bad regarding residual sum and residual sum related to area, but pretty good when residual sum is related to population size. For municipalities with a comparatively small population (e.g., several municipalities in the inland of Northern Sweden), the situation is the opposite.

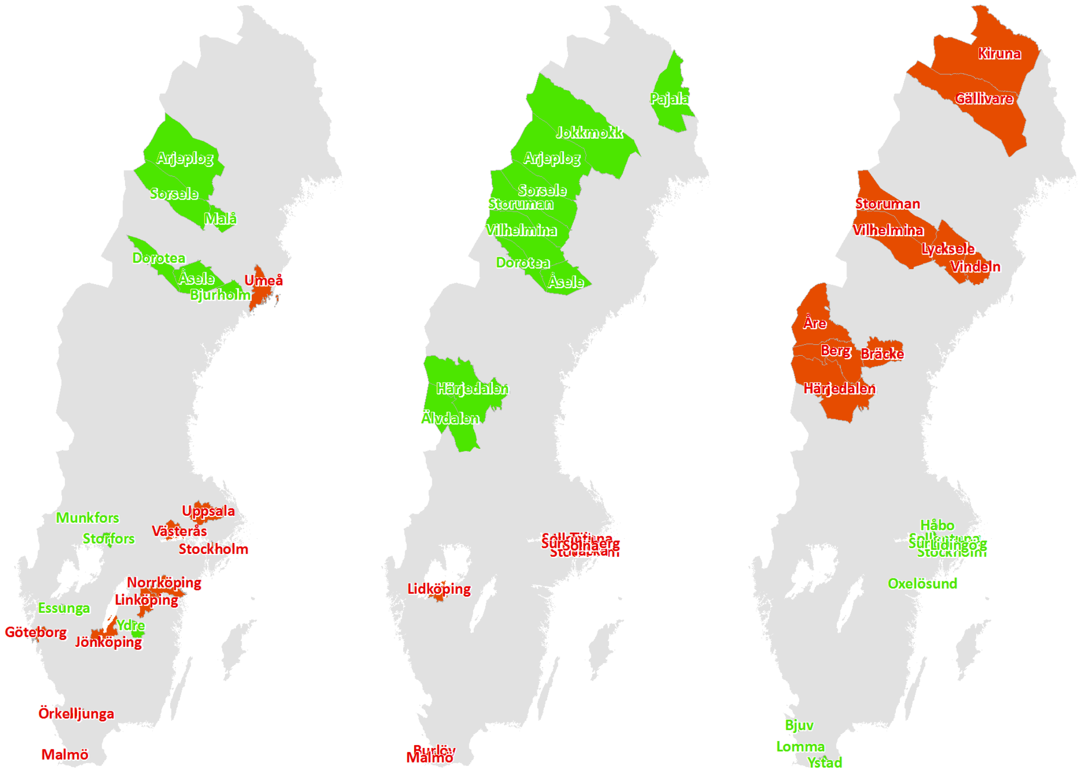


***Figure 2**: The ten municipalities with highest (red) and lowest (green) absolute residuals summarized (left) and in relation to area (middle) and population size (right)*

In Figure 3, all 290 municipalities are displayed in a scatterplot, with population size on the x-axis and absolute residuals by population size on the y-axis. The scale on the x-axis is logarithmical. There is a clear relationship between the two dimensions. As municipality population size increases, residual sum relative population size tends to decrease. However, this overall relationship is not without exceptions. In particular, for municipalities with a population of about 10,000 inhabitants, there are considerable variations in the level of overall discrepancy between grid and register data.

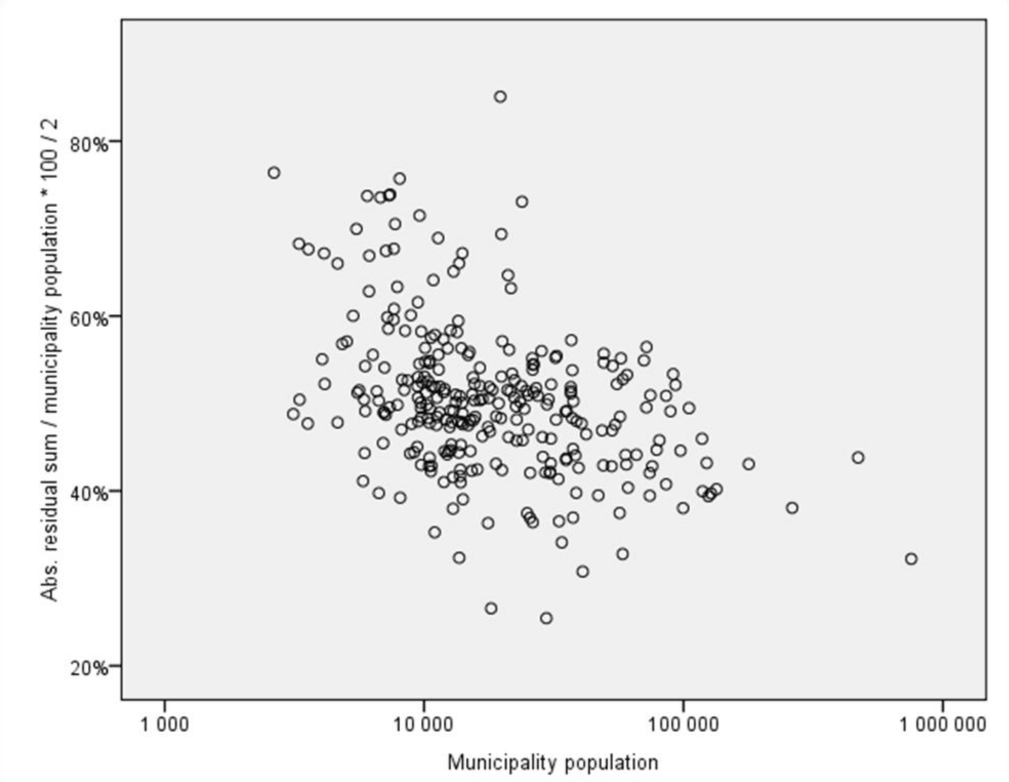


***Figure 3**: Municipality population size compared to absolute residual sum in relation to municipality population size*

2.1.1 Results by Municipality Group

In order to get a better understanding of how the three residual measures are related to the local context, municipalities are categorized by the municipality group to which they belong (see Figure 1 and Table 1). By the use of boxplots to graphically present the results, variations within and between these varying kinds of local settings becomes apparent. The first boxplot (Figure 4) displays absolute residual sum. The by far highest median error can be found in group 1, “big cities”. Municipality groups 3 (“major cities”) and 7 (“other municipalities, more than 25,000 inhabitants”) also exhibit comparatively large median errors (cf. Figure 2, left). It should be noted that the scale on the y-axis is logarithmical.

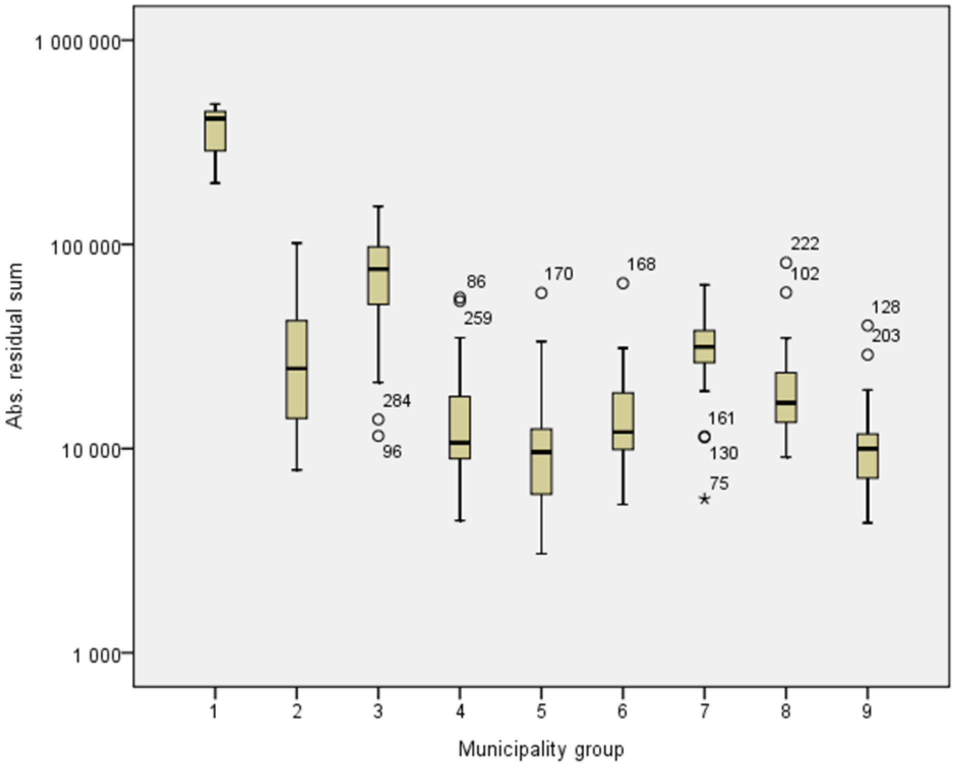


Figure 4*: Absolute municipality residual sum subdivided by municipality group*

In relation to area (Figure 5), a somewhat similar pattern of differences between municipality groups emerges. The big cities category (1) exhibits the largest median error; rural municipalities (5) clearly the smallest (cf. Figure 2, middle). When it comes to municipality group 2, which represents suburban municipalities, there is a substantial internal variation. It should be noted that the scale is the y-axis is logarithmical. For the gini-style measure of residual sum related to municipality size, the pattern is quite different (Figure 6). Rural municipalities exhibit the largest median error; big cities by far the smallest (cf. Figure 2, right). There are substantial variations within not only suburban, but also rural municipalities.

Like the results presented in Figure 2 and Figure 3, the municipality group comparison indicates that there is a relationship between the three residual measures and population size. In addition, it reveals substantial variations within certain municipality groups. In the case of suburban municipalities, it is easy to see why such diversity may arise. The settlement structure in suburban areas varies considerably, ranging from spacious residential area to crowded housing estates. Presumably, the mix of suburban housing in certain municipalities produces a population distribution more in line with the figures of the population grid.

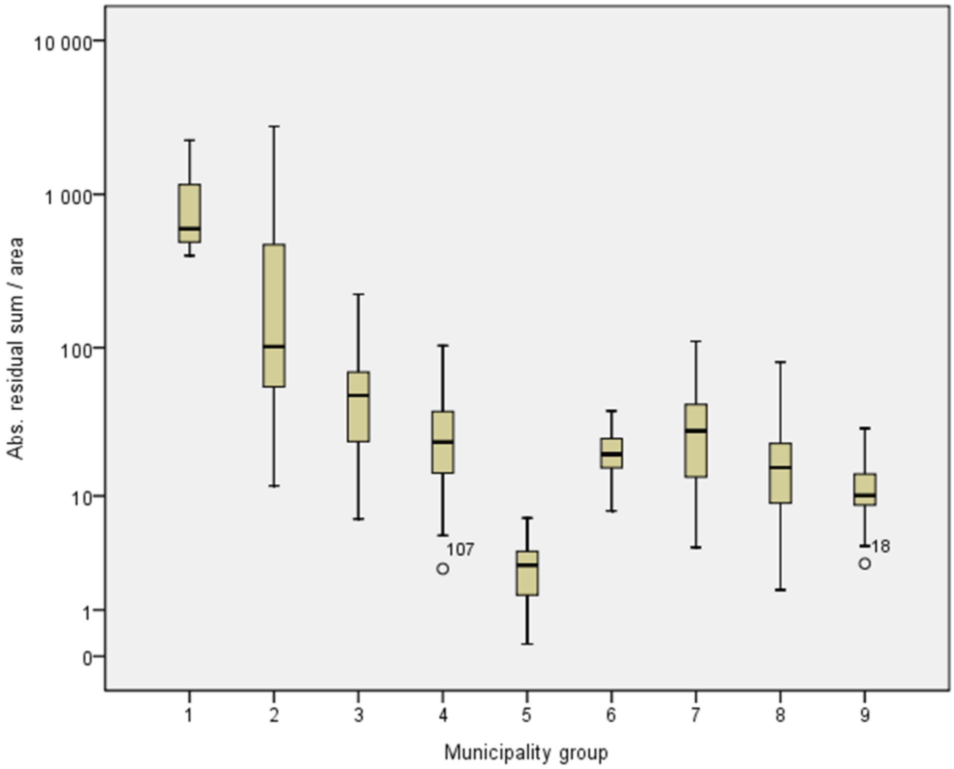


***Figure 5** : Absolute municipality residual sum in relation to area, subdivided by municipality group*

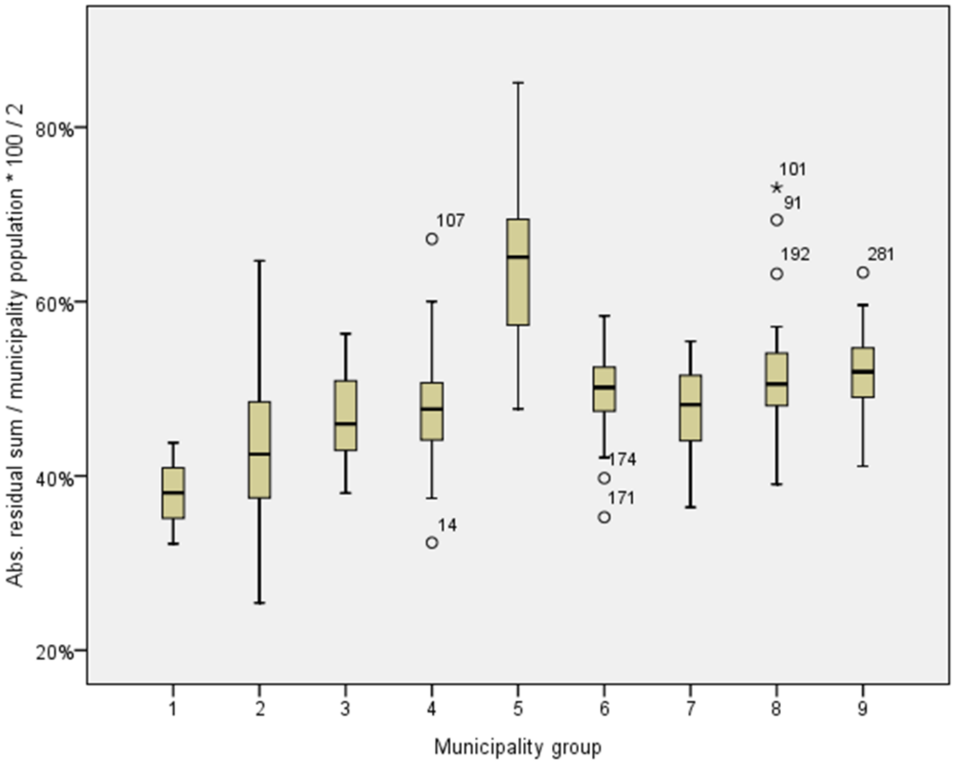


***Figure 6**: Absolute municipality residual sum in relation to population size, subdivided by municipality group*

2.1.2 The Residual Map

In the interpretation of the results, the underlying km^2 residual map may yield some clues. Figure 7 shows a residual map for Sweden as a whole; Figure 8 two close-ups of the residuals, representing a part of Southern (left) and Northern (right) Sweden. In these maps, red color means that the grid population is larger than the register population. Conversely, blue color indicates that the grid population is smaller than the register population. It should be noted that these maps only show squares that are inhabited in register data. As can be seen in Figure 7 and—even more clearly—Figure 8, there is a tendency for the grid to underestimate the population size of inhabited squares in the inland of Northern Sweden. Primarily, this is due to the assignment of population figures to many actually uninhabited squares. Naturally, this is a phenomenon that is likely to be more pronounced in sparsely populated areas, such as the rural municipalities in the inland of Northern Sweden (cf. Figure 1).

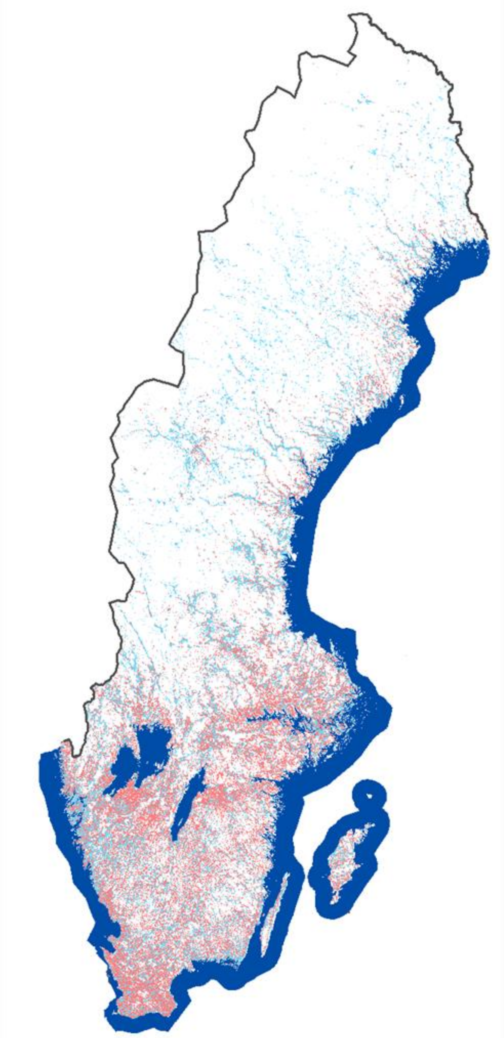


***Figure 7**: Absolute km^2 square residuals*

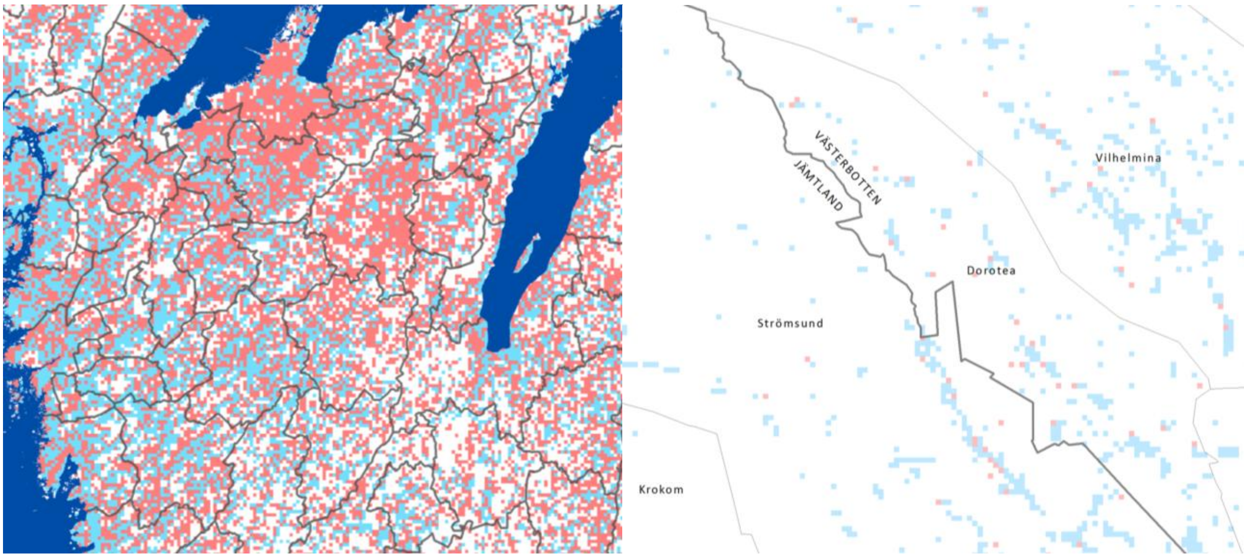


Figure 8: *Close-up of km^2 residuals in Southern (left) and Northern (right) Sweden*

2.2 Estimations of UMZ Population

The second test of the population grid, which concerns estimations of overall UMZ population, reveals an intriguing pattern of varying degrees over- and underestimation depending on UMZ size. Figure 9 displays, in scatterplot form, UMZ register population size on the x-axis, and overall residuals (register population – grid population) on the y-axis. For both the x- and the y-axis, the scale is logarithmical. In addition, the observations are binned: the larger the dots, the more UMZs are located in and around that point in the scatterplot. All in all, the number of over- and underestimated UMZs are about equal. Two separate clusters—one of which exhibits a linear trend—are clearly evident. First, for most UMZs with about 1,000 inhabitants and more according to register data, population is underestimated, and the underestimation increases with UMZ size. Second, the population of many UMZs with a register population below 1,000 inhabitants is—more or less—overestimated. In the boxplot that makes up Figure 10, this phenomenon is evident by the large range and many outliers in the UMZ size classes "200-999" and "1000-9,999", respectively. Generally, the amount of over- and underestimation is quite modest, especially when viewed in the light of actual population size.

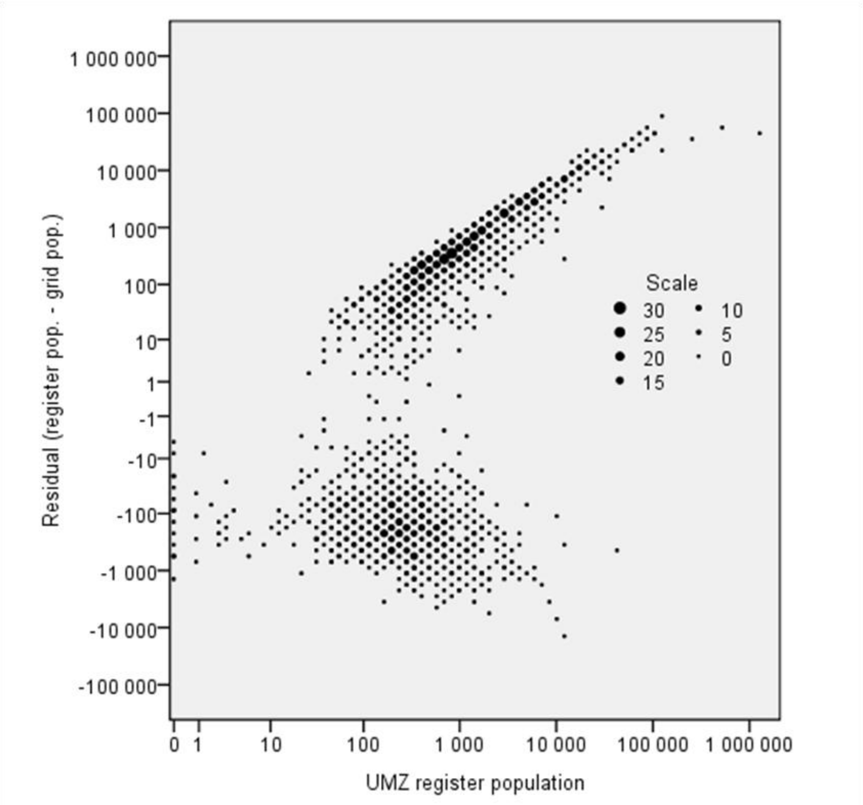


***Figure 9**: UMZ residuals by UMZ register population*

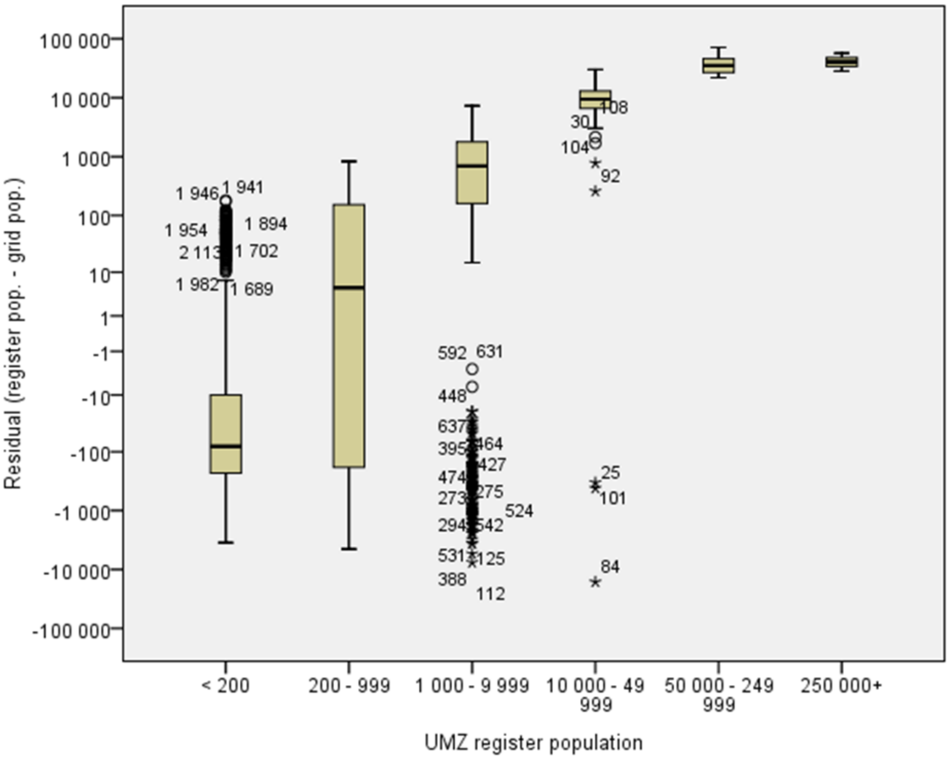


***Figure 10**: UMZ residuals by UMZ register population classes*

When it comes to the overestimation of many small UMZs, a possible explanation could be that the population grid overestimates areas with many buildings but small resident population. Spatial agglomerations of second homes, which are quite common in Sweden, are obvious examples of this kind of area. In Table 2, the UMZ boundaries are related to on the one hand the Swedish definition of urban localities (*tätort*)—basically, any agglomeration of 200 or more inhabitants—and on the other hand a delimitation of "other concentrated settlement". For UMZs that intersect any urban locality, 40 percent are overestimated. By contrast, for UMZs intersecting either only "other concentrated settlement" or neither category, the overestimation figure rises to about 80 percent. In practice, the "other concentrated settlement" category is largely made up by second home areas. Clearly, then, this finding lends some support to the notion of second home areas being responsible for the cluster of overestimated UMZs. Still, 80 percent of UMZs actually overlap urban localities, and a substantial proportion of those UMZs are also overestimated. In other words, the pattern of overestimation may also be a question of UMZ size.

Relation to Swedish delimitations	% of UMZs	% of UMZs overestimated
UMZ intersects neither urban locality (*tätort*) nor ”other concentrated settlement”	10	82
UMZ intersects only ”other concentrated settlement”	10	79
UMZ intersects urban locality (*tätort*)	80	40

Table 2*: UMZs in relation to Swedish definitions of settlements*

Concerning the general and increasing underestimation of large UMZs, there is harder to find an explanation for the phenomenon. Figure 11 shows grid residuals for UMZs in the Stockholm area, including residuals for Stockholm UMZ—the largest UMZ in Sweden in terms of population size, and also among the most underestimated using the population grid. In this map, red and—in particular—dark red color means that the grid population is larger than the register population. Conversely, the two shades of blue indicate that the grid population is more or less smaller than the register population. In the city center, there is—not surprisingly—a clear tendency for the grid to overestimate the population, while suburban areas generally exhibit a mixed pattern of over- and estimation. While this overall residual pattern is likely to occur in many other larger UMZs, it gives no obvious clue as to the reasons for the overall trend towards increased population underestimation with increased UMZ size.

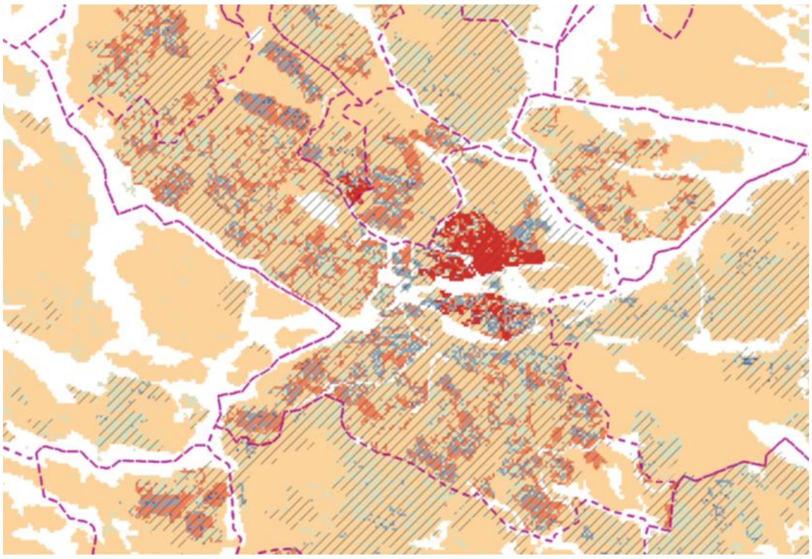


Figure 11*: Close-up of 100 meter^2 residuals in UMZs in the Stockholm area*

3 Summary and Discussion

In the ESPON db context, there is a need for population data with a high degree of spatial resolution. For instance, when it comes to disaggregating certain socioeconomic variables, or reporting population figures for geographical subdivisions such as UMZs, available data with good spatial coverage (e.g. for NUTS 2 or LAU, level 2 regions) have obvious limitations. Therefore, downscaled population data—specifically, a JRC population grid covering the entire EU—has been employed for these purposes. Against this background, the dataset has been subject to a country-level examination using Swedish register population data. Taking into account how the population grid is employed in the ESPON db project, the examination focuses on 1) population estimates in varying local settings, and 2) the estimation of overall population for UMZs of different sizes.

The first test summarizes absolute local residuals for Sweden's 290 municipalities (i.e., LAU, level 2 subdivisions), using three different measures: residual sum as well as residual sum related to municipality area and population size. Results are also presented categorized by municipality group—a classification of municipalities in nine different groups according to their characteristics. The results indicate that there is a relationship between municipality population size on the one hand, and the three residual measures on the other. Municipalities with a large population are associated with low discrepancies between grid and register data when residual sum is related to population size, but high discrepancies in terms of residual sum and residual sum related to area. For small municipalities, such as many rural municipalities in Northern Sweden, the situation is the opposite. A map of the actual local residuals reveals that there is a tendency for the population grid to underestimate the population size of inhabited squares in such settings. Primarily, this is due to the assignment of population figures to many actually uninhabited squares. In an EU perspective, this is likely to be less of an issue. The municipality group comparison reveals substantial variations within certain municipality groups, particularly regarding the category representing suburban municipalities. Presumably, this reflects the considerable diversity in settlement structure and population distribution that exist in suburban areas.

The second test concerns using the population grid to estimate the population of Urban Morphological Zones (UMZs). When overall residuals are related to UMZ population size, the about equal number of over- and underestimated UMZs form two separate clusters. For large UMZs the number of inhabitants tends to be underestimated, and the underestimation increases with UMZ size, while the population of many small UMZs is—more or less—overestimated. A plausible explanation for the latter phenomenon is that the population grid overestimates areas with many buildings but small resident population, such as second home areas. Generally, the amount of overall over- and underestimation is quite modest, especially when actual UMZ population size is taken into account.

In this country-level examination of downscaled population data, the discrepancies between downscaled and register data varies depending on local setting, and is also highly influenced by the way residuals are expressed. In any case, it is hardly a stretch to conclude that local grid population estimates often are quite unreliable. Still, using the population grid to downscale socioeconomic data is a likely to enhance to quality of data, and—and least for now—no better alternative exists. In the estimation of overall UMZ population size, the population grid works quite well—at least in the Swedish context. Consequently, while there are obvious limitations to downscaled population data, the JRC population grid is a quite reasonable tool for the enhancement of certain ESPON datasets.