Improving the quality of official statistics with geographical disaggregation based on dasymetric mapping: Two Eurostat experiments on tourism and population statistics

Julien Gaffuri, Eurostat - GISCO, julien.gaffuri@ec.europa.eu

Abstract

Official statistics are often reported on statistical units which are too large to depict properly the geographical distribution of the underlying phenomenon. In the European context for example, most statistics are produced only at national level (NUTS 0) and do not allow a true understanding of the spatial pattern at more local scales. Geographic resolution is a crucial component of quality in official statistics which should be better addressed. This article describes two experiments carried out at Eurostat for disaggregating statistics using auxiliary geographic data. These experiments are both based on so-called dasymetric mapping method: input statistical values are first distributed at the level of geographical features; these new statistical values are then re-aggregated at the level of the target statistical units with a finer resolution.

A first experiment is the disaggregation of tourism statistics over Europe from NUTS 2 to NUTS 3 and a 10km resolution grid. The auxiliary geographic information used is a dataset containing the location of around 160000 touristic accommodations over Europe. The outcome reveals a striking image of touristic activity over Europe with spatial patterns which cannot be revealed at NUTS 2 level.

The second experiment is on the disaggregation of mobile phone data over Belgium in order to assess population distribution over a 1km resolution grid. Mobile phone data are collected at antenna level, whose reception zones are extremely irregular in shape and size, especially in rural areas. Cadastral information on the location and volume of each single building over Belgium is used to define precisely the position of mobile phone users around built-up areas.

Both experiments show the pertinence of using geographic information with dasymetric mapping method to improve the overall quality related to geographical resolution. This method has been implemented in the generic library EuroGeoStat[1] and is intended be applied to other domains.

Keywords: Geographic disaggregation, geographic resolution, scale, dasymetric mapping, geographic information, tourism statistics, population statistics, mobile phone data, Eurostat.

1. Context and motivation

[1] https://github.com/eurostat/EuroGeoStat

Statistics are often reported on geographical regions, which are not always suitable to depict properly their geographical distribution. In the European context for example, most statistics are produced at country level and allow only simplistic rankings of countries instead of a true understanding of the spatial patterns at more local scales. Showing inter-countries disparities is not always relevant (and can even be misleading) especially when intra-country disparities are higher. Statistics reported on unsuitable geographical regions could thus lead to misleading interpretations, analyses and… decisions. The Modifiable Areal Unit Problem, MAUP (Openshaw, 1979) is a well-known illustration of this problem. Geographic resolution is therefore a crucial component of quality in official statistics and should better be addressed. On the other hand, many geospatial datasets are available nowadays, offering extremely detailed and rich descriptions of the European territories for various thematic areas. For example, national topographic datasets contain geometrical descriptions at building level showing various physical infrastructures for a large number of human activities.

The objective of the present work is to combine both statistical and geospatial information to produce statistics with a better geographical granularity. This idea is not new and has already been popularised through the dasymetric mapping method by (Tobler, 1979; Kim, 2010; Petrov, 2012). The principle of this method is to disaggregate input statistical values at the level of geographical features, which are supposed to represent fairly the geographical location of the underlying statistical phenomena, and then re-aggregate this information at the level of target statistical units with a finer resolution. Dasymetric mapping has already been applied in many cases, mainly for the geographical disaggregation of population statistics based upon land cover geospatial information, satellite images or addresses (Mennis, 2009; Gallego, 2011; Li, 2011; Zandbergen, 2011; Stevens, 2015; Pavía, 2016). More recently Batista (2018) has also focussed on tourism statistics decomposed using several geographical 'big' data sources.

This article describes two new experiments carried out at Eurostat for disaggregating statistics with auxiliary geospatial datasets, from a given NUTS level to finer ones and to a 10km resolution grid.

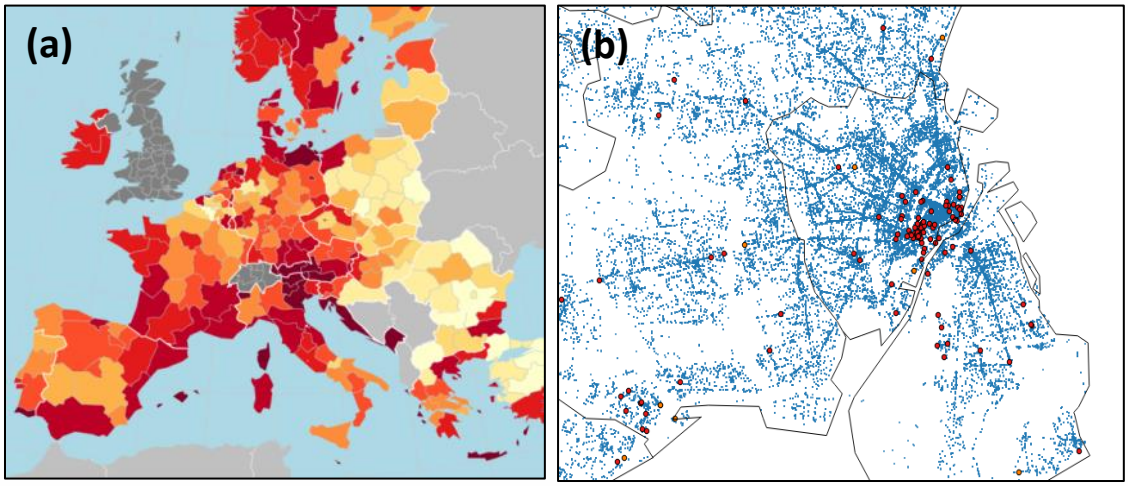
2. First experiment on tourism

2.1. Input data

The input statistical data is the Eurostat dataset on the "number of nights spent at tourist accommodation establishments" (Reference code: *tour_occ_nin2*) which is published annually at NUTS 2 level (Figure 1.a). The figures are decomposed by type of touristic accommodation: Hotel, holiday accommodation and camping ground.

For the disaggregation of this dataset, we used the TomTom MultiNet database, which contains the geographical location of around 160'000 touristic accommodations over Europe (Figure 1.b). These touristic accommodations are classified by different types (camping ground, holiday rental, hotel or motel, residential accommodation) which can easily be mapped to Eurostat codes. An important missing information is the size of these accommodations (the number of beds for example), which could help refining the method. This information is available from other data sources only for some specific countries. Batista (2018) has successfully addressed this limitation with the collection of additional information on Internet websites.

Figure 1. Input data – (a) Eurostat statistics at NUTS 2 level. (b) Geographical location of points of interest from TomTom MultiNet database (in blue) with a focus on touristic accommodations (in red) around Copenhagen. Gray lines are NUTS 3 boundaries.



2.2. Methodology[2]

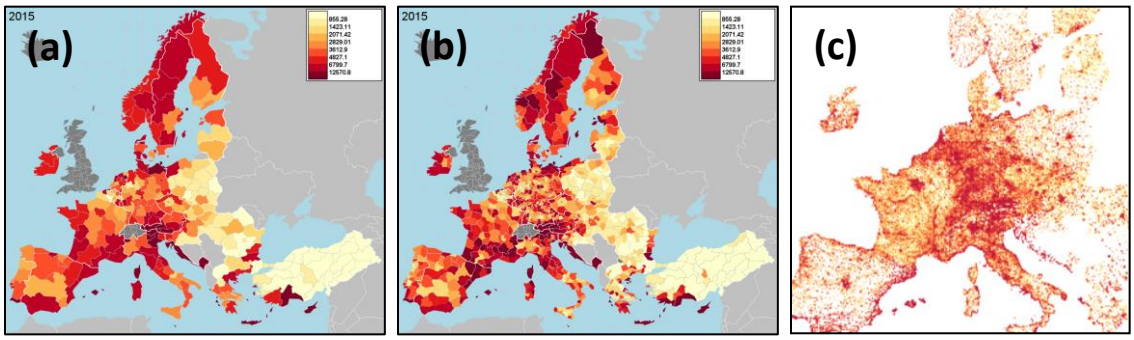
[2] The source code of the process can be retrieved from: https://github.com/eurostat/EuroGeoStat

Dasymetric mapping technique has been used from this input data for a geographical disaggregation to NUTS3 level and to a 10km resolution grid. The first step consists in linking the touristic accommodations to the NUTS 2 region they belong to. The statistical values are then transferred at the level of each touristic accommodation. Finally, a new aggregation at the level of the target statistical units is performed.

2.3. Results

Figure 2 presents the main results obtained for different levels of disaggregation.

Figure 2. Outcome for 2015, for all touristic accommodation types: (a) Input statistical data at NUTS 2 level. (b) Result of the disaggregation at NUTS 3 level. (c) Result of the disaggregation on a 10km grid.



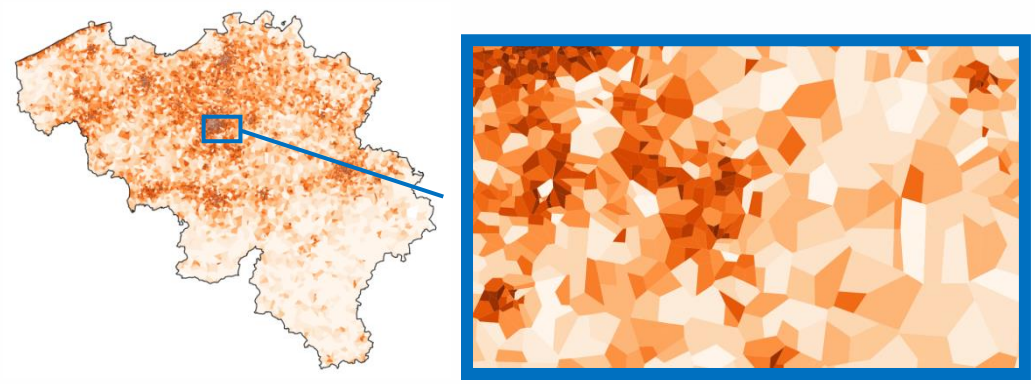
As expected, these results show a more detailed spatial distribution of the touristic attractiveness. Some new spatial patterns appear which lead to a totally different interpretation of the statistical data:

1. Concentration along the coast is stronger, as revealed along the German and Polish coast on the Baltic Sea, on the Black Sea coast and also in the south west coast of Turkey.
2. Some inland regions have some particular touristic attractions, such as the regions along the Norwegian and Swedish border, and also in the south west of France).
3. Some strong touristic poles appear within large regions, such as the surroundings of Krakow, in the Romanian Carpat or capital cities such as Ankara.

3. Second experiment on population

The second experiment aims at integrating mobile phone and cadastral data for the estimation of population distribution over Belgium. Mobile phone data has already been used to analyse population dynamics (Ahas, 2015; Deville, 2014; De Meersman 2016; Kamenjuk, 2017). An identified limitation of this type of data is its irregular geographical granularity mobile phone data is usually available at the level of signal reception antennas. Indeed, this data is located on reception zones around these antennas which are extremely irregular in shape and size, especially in rural areas (Figure 3). The objective is to assess the benefit of using cadastral geographical data with dasymetric mapping to overcome this limitation.

Figure 3. Input raw mobile phone data over Belgium reported on irregular reception zones.



3.1. Input data

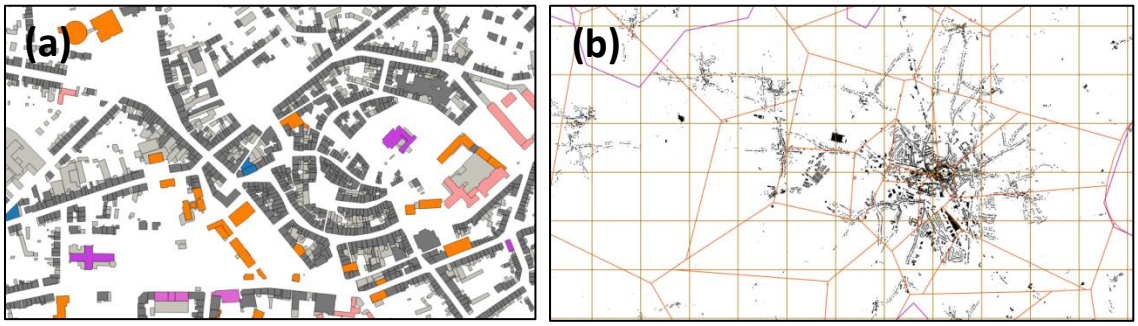
The mobile phone data have been provided by Proximus[3] mobile phone company over Belgium in the context of the Eurostat big data task force activities. It consists of the number of mobile phone users detected in a reception zone every 15 minutes. To focus on residential population, we have selected the data collected during the night, from 4:00 to 4:15. The geographical information on buildings has been retrieved from the three different regional mapping authorities of Belgium. The three data sources[4] contain information on the polygonal representation of each building (Figure 1.a).

[3] https://www.proximus.be

[4] For Brussel region: https://urbisonline.brussels. For Flanders: GRBgis database https://download.agiv.be/. For Wallony region: PICC database http://geoportail.wallonie.be/catalogue/b795de68-726c-4bdf-a62a-a42686aa5b6f.html

Information on the volume of the building (height and/or number of floors) as well as the usage type (residential, commercial, industrial, etc.) can also be made available depending on the source.

Figure 4. (a) Input geographical data on buildings, in Arlon. (b) Overlay of the difference spatial units: Buildings in black, mobile phone reception zones in orange, 1km resolution grid in yellow, municipalities in purple.



The motivation for this experiment is based on the common sense assumption that residential population is usually located in residential buildings. The idea is thus to combine both mobile phone and building data to determine the geographical distribution of the population. We expect significant improvements especially in rural areas, where mobile phone reception zones are large and irregularly populated over space.

3.2. Methodology[5]

The methodology consists in applying dasymetric mapping with the input data as described in the following steps:

Step 1: The building "liveable" areas are computed. Only buildings with housing or residential usage are selected. The number of floors, when not directly available, is estimated from the building height. The building liveable area is simply computed as its area multiplied by the number of floors.

[5] The source code of the process can be retrieved from: https://github.com/jgaffuri/mapreduceEBD/tree/master/src/main/java/eu/ec/eurostat/bd/proximus

Step 2: The liveable building area of each mobile phone reception zone is estimated as the weighted average of the building liveable areas (computed in step 1) it contains. The weight is the share of the building intersecting the zone. This weight is equal to 1 for buildings fully encompassed within the zone. The result of this step is used to compute the density of liveable area for each reception zone. This density value can be greater than 1 for densely urbanised zones, e.g. with buildings with a lot of floors.

Step 3: The number of mobile phone users is disaggregated at the level of the buildings. First, the density of mobile phone users is computed for each reception zone as the ratio between the mobile phone population information and the total liveable building area computed in step 2. The density value is assumed to be homogeneous for all buildings of the zone and is assigned to every building in the zone. For buildings intersecting several zones, a weighted average is computed based on the intersection areas of the building geometry and the intersecting zones. Each building population is then computed from its density and liveable area. The result is the number of mobile phone users for each single building as shown on figure 5.

Figure 5. Estimation of mobile phone data at building level (city of Arlon). The mobile phone reception zone limits are shown in black.

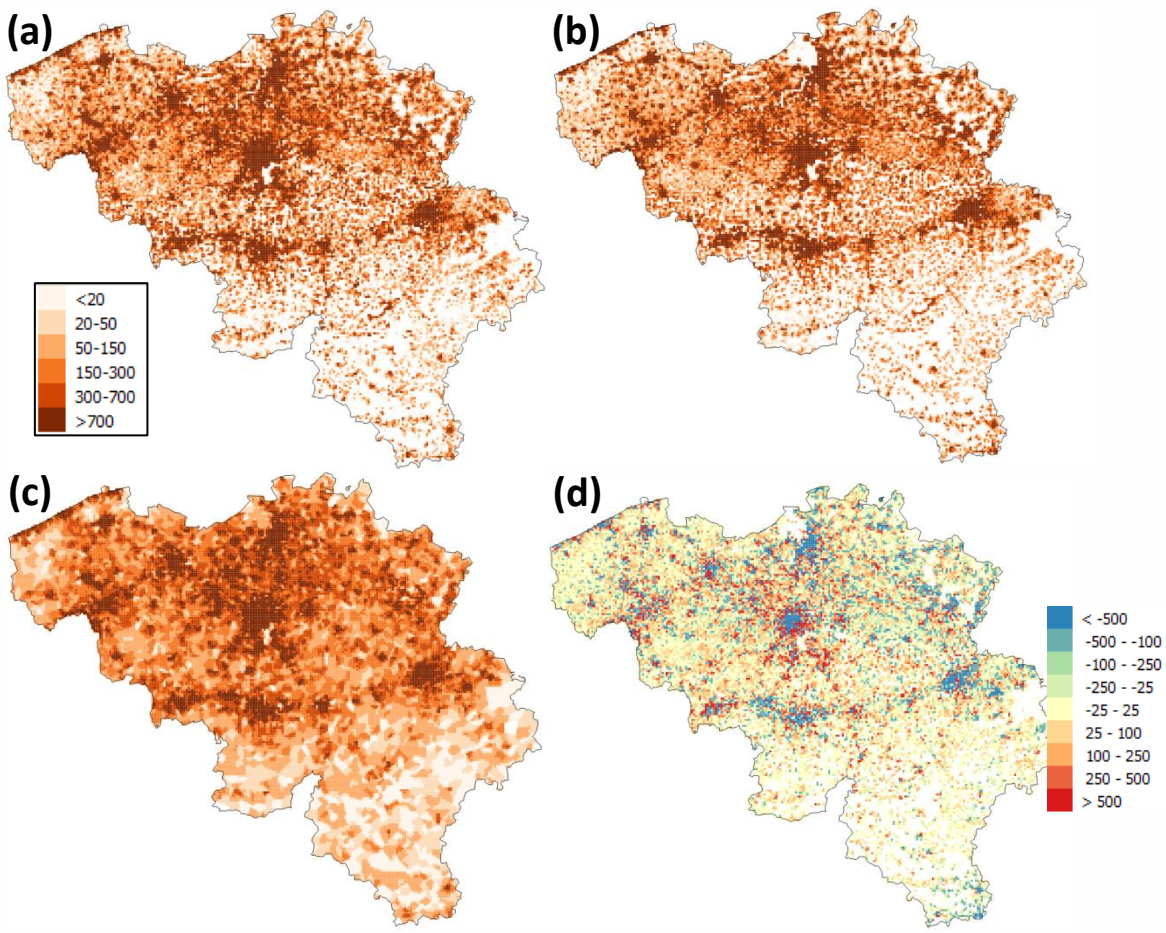
Step 4: The mobile phone data at building level is finally aggregated at 1km grid cell level as the sum of the building population the zone contains. For buildings intersecting several grid cells, a weighted average based on area intersections is used. Note that the aggregation method applied here is the same as for step 2.

Step 5: The final step is a normalisation of the grid figures based on an estimate of the total population of Belgium in 2015.

3.3. Results

Figure 6 shows an overview of the final result (a), compared with data derived from other sources of information. The comparisons show the benefits of using geographical location of building for the disaggregation of mobile phone users' data.

Figure 6. (a) Final result over Belgium . (b) Geostat population 2011. (c) Raw mobile phone data by reception zone. (d) Difference between final result (a) and geostat population in 2011 (b).



4. Conclusion

These two experiments, together with the works presented in (Mennis, 2009; Gallego, 2011; Li, 2011; Zandbergen, 2011; Stevens, 2015; Pavía, 2016, Batista, 2018), show both the feasibility but also the high relevance of spatially disaggregated

statistics. Integrating geospatial data and these statistics with dasymetric mapping has a high potential and should be generalised. A more systematic approach could be adopted, consisting in listing available geospatial datasets and matching their information to statistical domains, in the same way TomTom MultiNet dataset has been matched to the Eurostat dataset on the number of nights spent at tourist accommodation establishments. Dasymetric mapping could then be considered for these pairs of matched datasets.

The methodology applied for these experiments could be generalised for statistical productions based on various data sources and auxiliary geospatial data. Dasymetric mapping deals with the integration of three types of datasets:

- Input statistical information such as the Eurostat data on tourism at NUTS2 and mobile phone user counts.
- Statistical unit datasets such as the NUTS 2 and 3 regions, the mobile phone reception zones, the 10 and 1km grids.
- Auxiliary geospatial information, such as points of interests or buildings.

Table 1 is an attempt to show the complementary strengths of the different data sources involved and the benefit of integrating them to produce a new dataset balancing these strengths.

Table 1. Overall comparison of the strengths and weaknesses of the input and output datasets.

	Statistical relevance	**Temporal granularity and timeliness**	**Spatial granularity**
Input statistical information	++++	++++	+
Auxiliary geospatial information	+	+(+)	++++
Output geo-disaggregated statistical information	+++	++++	+++

Dasymetric mapping offers a powerful method to transfer the high temporal granularity and timeliness of statistical data and the high spatial granularity of auxiliary geospatial data to a final statistical product balancing these two crucial quality components.

5. References

R. Ahas, et al. (2015). `Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn'. International Journal of Geographical Information Science 29(11):2017-2039.

F. Batista, et al. (2018). `Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources'. Tourism Management 68:101-115.

F. De Meersman, et al. (2016). `Assessing the quality of mobile phone data as a source of statistics'. In European Conference on Quality in Official Statistics. Eurostat.

P. Deville, et al. (2014). `Dynamic population mapping using mobile phone data'. Proceedings of the National Academy of Sciences 111(45):15888-15893.

F. J. Gallego, et al. (2011). `Disaggregating population density of the European Union with CORINE land cover'. International Journal of Geographical Information Science 25(12):2051-2069.

P. Kamenjuk, et al. (2017). `Mapping changes of residence with passive mobile positioning data: the case of Estonia'. International Journal of Geographical Information Science pp. 1-23.

H. Kim & X. Yao (2010). `Pycnophylactic Interpolation Revisited: Integration with the Dasymetric-mapping Method'. Int. J. Remote Sens. 31(21):5657-5671.

T. Li & J. Corcoran (2011). `Testing dasymetric techniques to spatially disaggregate the regional population forecasts for South East Queensland'. Journal of Spatial Science 56(2):203-221.

J. Mennis (2009). `Dasymetric Mapping for Estimating Population in Small Areas'. Geography Compass 3(2):727-745.

S. Openshaw & J. Taylor (1979). `A million or so correlation coefficients: three experiments on the modifiable areal unit problem'. Statistical applications in the Spatial Sciences pp. 127-144.

J. M. Pavía & I. Cantarino (2016). `Can Dasymetric Mapping Significantly Improve Population Data Reallocation in a Dense Urban Area?'. Geogr Anal p. n/a.

A. Petrov (2012). `One Hundred Years of Dasymetric Mapping: Back to the Origin'. The Cartographic Journal 49(3):256-264.

F. R. Stevens, et al. (2015). `Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data'. PLOS ONE 10(2):e0107042+.

W. R. Tobler (1979). `Smooth pycnophylactic interpolation for geographical regions.'. Journal of the American Statistical Association 74(367):519-530.

P. A. Zandbergen (2011). `Dasymetric Mapping Using High Resolution Address Point Datasets'. Transactions in GIS 15:5-27.