Pavía, JM.; Cantarino Martí, I. (2017). Can dasymetric mapping significantly improve population data reallocation in a dense urban area?. Geographical Analysis. 49(2):155-174. doi:10.1111/gean.12112.

# Can dasymetric mapping significantly improve population data reallocation in a dense urban area?

## 1. Introduction

As required by national, supranational and federal laws, the vast majority of social variables distributed by statistical agencies are offered aggregated by areal units because of confidentiality issues (see, e.g., UK Parliament 2007; OJEU 2013; USC 2002). This is the case, for example, with census figures, which are only available as geographically accumulated data in order to ensure that it is impossible to identify a particular person, either directly or indirectly in connection with any other published information. In fact, according to the US Census Bureau (2015), before issuing small-area aggregated statistics single records are even "switched with similar records from a neighboring area" and/or randomly perturbed to preserve individuals' privacy.

Social variables are disseminated in a great variety of geographic entities. Some of them are well-established administrative units, such as states, provinces, counties or cities. Others are more arbitrary and easily modifiable, such as state legislative and congressional districts used in US elections (which are redrawn every 10 years following US censuses) or census blocks and precincts (which are instances of the smallest geographical units for which social data are made public).

For the first group of geographic entities, large series of data are increasingly available in developed countries and temporal comparisons and longitudinal studies can be performed directly from the published data. For the second group of spatial units, however, implementing longitudinal analyses is not so simple; the complexity of the problem even growing as the scale of geographical aggregation reduces. In the case of relatively large geographic entities (such as congressional districts), rebuilding the history of the variables of interest is not as a rule so complex, particularly when the data of the smallest geographic units that constitute it (such as precincts or census tracts) are available (Pavía and López-Quilez 2013). The real problem arises when we deal with smaller areas (such as census blocks). In the latter, when the data come from enumerations, surveys or administrative registers with detailed spatial references (like

postal addresses or GPS coordinates), the figures of the target variables corresponding to the new redrawn areas could, after some additional workings, be theoretically restored from the microdata.

Unfortunately, ordinary analysts hardly ever have access to such detailed information and, what's more, in many cases complete geographical indicators are not available (even for statistical producers), as when dealing with historical data or with electoral outcomes or with surveys in which only rough spatial references are collected. Despite this, disciplines like economics, political science, sociology and other social sciences do not shy away from using census figures, unemployment rates or party supports to examine social trends, evaluate policy impacts or test social theory. Hence, it is not surprising that many approaches have been proposed in the literature attempting to overcome the limitations that the modifications of the spatial unit boundaries impose on performing longitudinal studies.

In particular, to solve the problem of reallocating data from a set of geographical administrative units onto another overlapping but non-hierarchical set of spatial units (i.e., in a context where the scale of analysis is fixed and only the shape of the aggregation units is changed), a number of methods of progressively growing in complexity have been suggested over time. They have moved from simple areal weighting interpolation (see, e.g., Goodchild and Lam 1980)—which are equal to performing an Euler-Venn geometric approach—and point-based areal interpolation procedures (e.g., Martin 1989; Bracken and Martin 1989; Fisher and Langford 1995) to more complex strategies based on using dasymetric mapping with ancillary sources of information. Because social data are related to population, dasymetric methods use auxiliary variables to guide in an intelligent fashion the redistribution process (Wright, 1936). Among the variables tested we can find data about land uses (Mennis 2003; Giordano and Cheever 2010), satellite imagery (Robinson et al. 2002; Holt et al. 2004), road networks and night-time lights (Reibel and Bufalino 2005; Briggs et al. 2007; Zandbergen and Ignizio 2010), address points (Zandbergen 2011), the spatial distribution of built structures (Maantay et al. 2007), LiDAR-derived building heights (Sridharan and Qiu 2013), or a combination of spatial methods and the Expectation-

Maximization (EM) algorithm (Flowerdew and Green 1994; Gregory and Ell 2005; Sridharan and Qiu 2013).

The first methods are conceptually simple and do not require an in-depth knowledge of spatial methods. They are quite intuitive and can be implemented easily. The dasymetric approaches, nevertheless, are more complex and data demanding and ask for a higher understanding and ability in the use of GIS tools. They entail employing more spatial layers and combining data from more sources. The aim of this work is to examine whether and to what extent a more complex approach is worthwhile in an instance in which *a priori* its usefulness can be put into question.

The rest of the paper is distributed as follows. Section 2 introduces the case study and sets the problem. Section 3 describes the sources utilized and the spatial methods tested in this research. In addition to the baseline approaches of point-areal interpolation and areal weighting, ten additional methods based on dasymetric techniques are analyzed. In Section 4 the different reallocations obtained are compared and their relative merits assessed. Finally, Section 5 summarizes and discusses the findings.

## 2. Case Study

Barcelona is the second-largest city in Spain with more than 1.6 million of inhabitants. Its metropolitan area, with a population of around 4.7 million people, makes up the largest European urban area on the Mediterranean Sea. Barcelona, located on the northeast coast of the Iberian Peninsula (see Figure 1), is a polycentric and spatially complex city (Catalán et al. 2008) with a marked social structure (Broner 2010) that extends for a total area of 170 $km^2$, of which 101 are occupied by the city itself and just 48.6 are classified as residential areas.

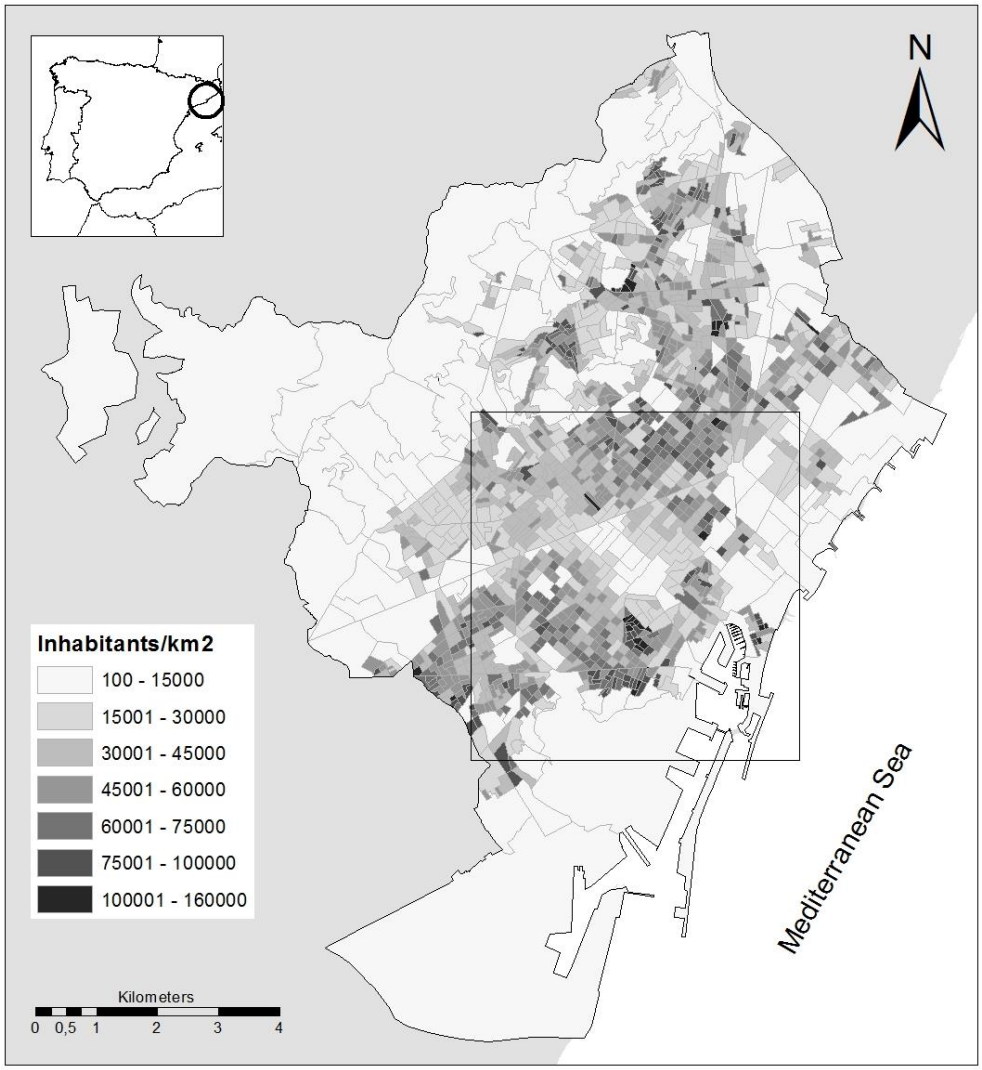


**Figure 1**. Study area, Barcelona (Spain). Population density for the 2009 census sections (source units) depicted. The square delimits the area displayed in Figure 2.

Barcelona is the focus for the study mainly because it constitutes, from both theoretical and empirical standpoints, an instance in which *a priori* we can expect the dasymetric allocating approaches not to perform any better than the simpler methods. From a theoretical perspective, it could be argued that when units are small enough to be relatively homogeneous with regards to their inner distribution of population and, moreover, when big differences between contiguous units are not expected, areal weighting should be sufficient and that any auxiliary data employed will result in marginal, if any, improvements. Indeed, some previous studies (Brinegar and Popick 2010; Zandbergen and Ignizio 2010) point to this for relatively densely populated areas. Barcelona with an average population density of 16,000 inhabitants per $km^2$, which rises to 34,000 if only to residential areas are considered, represents a very densely and

continuously populated urban area (see Figure 1), with a statistically significant positive spatial correlation. Actually, taking the 2009 census sections as spatial units, the Moran index (Cliff and Ord 1981) of Barcelona population density is around 0.08 (p-value: 0.0000), an indicator that population density does not change abruptly at the boundaries of the source units. What is more, the Moran index rises to 0.26 (p-value: 0.0000) when we just consider the core of the metropolitan area (the census sections depicted in Figure 2-left), which comprises more than half the Barcelona total population (just under one million inhabitants). As a measure of urban compactness (Tsai, 2005), this value signals a very highly compact urban form, characteristic of the historic centers of the European cities.

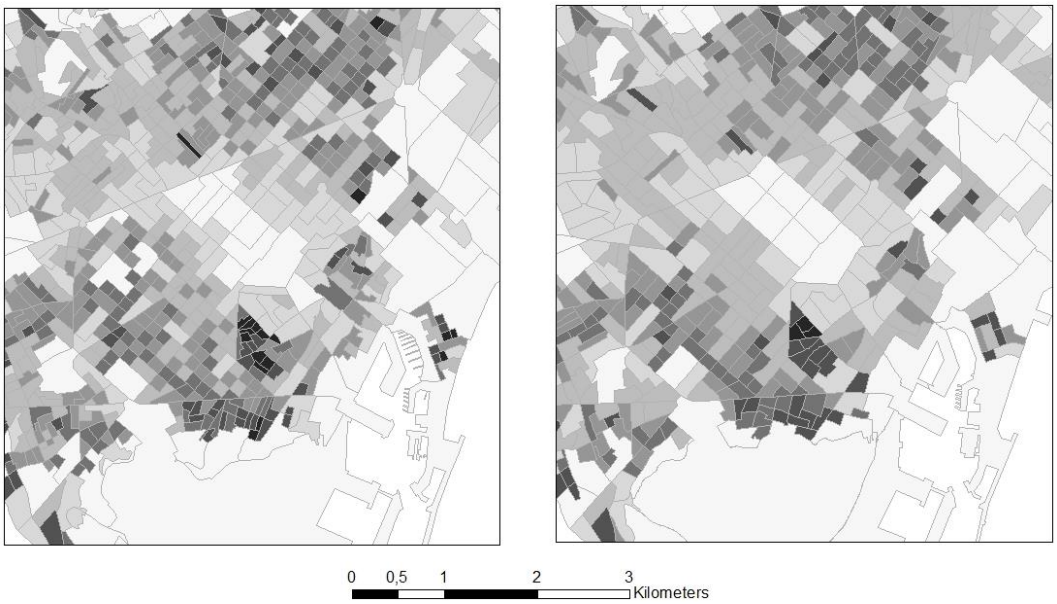


**Figure 2**. Extract of Barcelona (Spain) division in census sections during the 2009 (left panel) and 2010 (right panel). The same zone (area around Catalonia Square, see Figure 1) depicted in both panels. Population densities by census section shaded in the figures.

From an empirical perspective, in urban areas, and mainly in big cities, spatial breakdowns are usually performed by local expert analysts with an in depth knowledge of the region. The boundaries of small areas tend to be established in a thoughtful and intelligent way, respecting area barriers such as rivers, wide avenues, roads or municipalities' boundaries, which can sometimes also act as social barriers. Indeed, in the Barcelona breakdowns the integrity of block buildings are always respected, with the boundaries of the majority of census sections being placed on streets. This, together with the clear spatial patterns that many social variables show, suggests that the

transitions between old and new units should be smooth and consequently begs the question whether in these circumstances dasymetric reallocation techniques would provide significant improvement in accuracy.

This question is pertinent because the majority of previous studies have been focused on big areas less densely populated, where moreover the population is sparsely distributed in the territory. For example, the study area in Eicher and Brewer (2001) encompasses 159 counties, covering four US states, with a maximum population density of 3,424 inhab/km$^2$. Mennis (2009) considers Delaware County, an area of 490 km$^2$ with 550,864 people in the year 2000 and 144 census tracts. Zandbergen (2011) handles 16 counties in Ohio, comprising 1 million people distributed in 230 census tracts across an area of 19,000 km$^2$. And, Reibel and Bufalino (2005) focus their study on Los Angeles County: 9.5 million people in the year 2000 distributed in 10,570 km$^2$. A similar picture is found for European studies. For example, Suárez et al. (2008) analyze the population distribution in the Gran Canaria island (Spain)—with 304,000 people distributed in 1,560 km$^2$—, while Goerlich and Cantarino (2013) and Batista et al. (2013) deal with, respectively, the whole of Spain and the whole of Europe.

Every year, and referenced to the first of January, the Spanish Official Statistical Agency (*Instituto Nacional de Estadística*, INE) publishes (among many other variables) the population by age, broken down into five year groups, for each census section. These data come from the Municipal register, where municipality inhabitants are recorded, and can be downloaded free from the INE website (http://www.uv.es). Census section data published refers to the spatial census section breakdown in force in each period. Hence, for Barcelona and at census section level, the official population data were delivered using different breakdowns in 2009 and 2010. Therefore, to link them we would need to express them in the same breakdown.

Once a variable is reallocated from the 2009 breakdown to the 2010 breakdown, a problem emerges when one tries to compare the resulting approximations to the official statistics available for the 2010 breakdown. The official data available for the 2010 breakdown correspond to the first of January of 2010, whereas the reallocated data is referred to the first of January of 2009. In addition to the differences resulting from the

reallocation process, divergences can also occur because of natural and migration annual population movements registered during the year of 2009. Fortunately, on this occasion, single records are available and, under request and on payment, INE agents were kind enough to compute for us the actual data that they would have published if the 2010 breakdown had been in force on the first of January, 2009. In particular, INE agents provided us with a file containing for each 2010 census section the number of people living there the first of January of 2009 broken down into eighteen age groups: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, and >85. These have been the variables reallocated in our analysis.

We assess whether dasymetric reallocation techniques would provide significant improvement in the case of a densely and continuously populated urban area (taking Barcelona as an ongoing example), by reallocating the population figures available for the 2009 breakdown into the 2010 breakdown and by comparing the resulting estimates to the INE values that would have been observed if the 2010 breakdown had been in force the first of January of 2009.

In addition to performing our scrutiny in absolute values as is usual in the reallocating spatial interpolation literature, we have also applied the reallocating approaches to the variables measured in percentages: calculated as a ratio over the total population of each census section. Many health and social variables are observed as rates or proportions and, therefore, in our opinion more effort should be devoted to finding out how simple and dasymetric interpolation methods would work when faced with the problem of reallocating rates or proportions from a set of geographical administrative units onto another, overlapping but non-hierarchical, set of spatial units.

## 3. Data sources and methods

In this section, we describe and explain the details of the reallocating approaches assessed in this research. Grouped in pairs, fourteen methods have been tested. The first two pairs of proposals (grounded on point-based interpolation and areal weighting interpolation) are considered simple: they do not require auxiliary variables. They are

used as baseline methods to gauge the value of the other ten alternatives, which are based on dasymetric techniques and therefore are more complex. The point and areal weighting interpolation solutions just require the files providing the polygon attribute tables of the spatial breakdowns to be executed. The dasymetric mapping approaches need additional sources of information to guide the reallocation process: auxiliary variables related to the distribution of the population. In what follows, we first present the ancillary sources and variables employed in the dasymetric approaches, subsequently we introduce the methods considered.

### *3.1. Ancillary sources of information*

In addition to data from INE, which provides the census variables analyzed in this research, we have dealt with geographical objects provided by four additional institutions: the City Council of Barcelona (*l'Ajuntament de Barcelona*), the Spanish Geographic Institute (*Instituto Geográfico Nacional*), the European Environment Agency (EEA) and the Spanish Cadastral Agency (*Dirección General del Catastro*). We also examined information from the OpenStreetMap project. However, because in that project data are collected by different teams using different standards, we had to discard it due to its limited adaptability and wide heterogeneity in terms of quality.

*CartoBCN* is the official website for cartographic information of *l'Ajuntament de Barcelona* (http://w20.bcn.cat/cartobcn/default.aspx?lang=en), from which many geographic files can be downloaded after registering free. From this source, we obtained (i) the shape (shp) files corresponding to January 2009 and January 2010 section breakdowns of Barcelona (according to which INE official census variables are delivered), and (ii) the Barcelona city street map (*Callejero*). *Callejero* contains information on the streets of Barcelona and their intersections. In *Callejero*, each stretch of road has associated with it, among other information, INE and city council specific codes, the street name and the type of road.

Files corresponding to 2009 SIOSE, LiDAR and MDT05 databases are available in the Download Center (http://centrodedescargas.cnig.es/CentroDescargas/) of the *Instituto Geográfico Nacion*al. Land Cover and Land Use Information System of Spain (SIOSE) is a

unique database of Spain produced at 1:25,000 scale combining topographic maps, satellite imagery, aerial photography and cadastral registers. This database contains information about heterogeneity of land use within any given Spanish polygon (Goerlich and Cantarino 2013). SIOSE comprises 2.5 million polygons with nearby 820,000 different land cover categories, obtained after combining the land cover elements with different weights (Cantarino *et al*. 2014). Each SIOSE polygon contains information about what percentage of the surface in the polygon corresponds to each attribute. Inhabited buildings are identified with four different type attributes in SIOSE. They correspond to compact apartment blocks (A1), isolated apartment blocks (A2), terraced houses (H1), and detached houses (H2). According to Goerlich and Cantarino (2013), taking 1.00 as reference value of the relative population density for a standard spatial unit of detached houses (H2), the values of relative population densities for compact apartment blocks (A1), isolated apartment blocks (A2) and terraced houses (H1) are, respectively, 9.23, 6.67 and 1.83.

Digital files with altimetry information from the LiDAR (which stands for Light Detection and Ranging) cloud of points are distributed in Spain in files of a 2x2 km grid. The download is done by municipality. Point clouds have been captured by flights using LiDAR sensor with a density of 0.5 points/$m^2$ and automatically classified and colored by RGB obtained from orthophotos with a pixel size of 25 or 50 cm of the Spanish National Aerial Orthophotography Plan (PNOA). The points, besides elevation, contain information about nine classes of terrain attributes.

MDT05 is a digital elevation model (DEM) with 5 meters resolution, which has been obtained, depending on the map sheet, using one of two main procedures. In Barcelona MDT05 data have been computed by interpolation from LiDAR flights of PNOA by selection of class type “Ground”, with sub-metric precision.

The European Environment Agency (EEA) provides the Urban Atlas database. This offers pan-European comparable land use and land cover data for Functional Urban Areas. It can be downloaded (http://www.eea.europa.eu/data-and-maps/data/urban-atlas) for the main European cities, including Barcelona. Available at a 1:10,000 scale with a position accuracy of +/-5 m, its date of production is 2009. The layer called “Urban

Fabric", which classified the polygons into categories according to their percentage of soil sealing or sealing levels, has been the one used in this research. Under the principle that the larger the percentage of soil sealing in a polygon the larger is the percentage of residential area of the polygon, we have considered as proportions of residential areas in each polygon the average of soil sealing of its class: 0.05, 0.20, 0.40, 0.70 and 0.90 for, respectively, the categories "Discontinuous Very Low Density Urban Fabric", "Discontinuous Low Density Urban Fabric", "Discontinuous Medium Density Urban Fabric", "Discontinuous Dense Urban Fabric" and "Continuous Urban Fabric" and "Isolated Structures".

The Spanish cadastral database is an administrative register that is supported by the Ministry of Finance and Public Administration and contains the description of all the rustic, urban and special feature properties in Spain. Cadastral information is a free service (https://www.sedecatastro.gob.es/), but an electronic certificate is required to access it. Shapefile (shp) and alphanumeric (cat) data can be downloaded by municipality. The shp files define the boundaries of the cadastral parcels. The cat files contain information about the use (residential or other types), year of building and living area per dwelling. Shapefile files have no information about the use of buildings or surface of housing, but thanks to cadastral references they can be linked with the data of cat files.

### *3.2. Spatial reallocating approaches*

This subsection offers details of the fourteen reallocating approaches considered in our scrutiny. The first two methods (M01-rIDW and M02-rNN) are point-based approaches, the Tobler's approach (M12-TPYC) is rasted-based, and the other eleven methods are vector-based procedures. The point-based methods have been included for the sake of comparison. Besides being inefficient, they do not verify the pycnophylactic condition of volume preservation (Tobler 1979). All the other twelve approaches are volume-preserving (also called mass-preserving). Table 1 displays a summary of the methods with some details regarding the way they have been computed. The auxiliary variables (and sources) employed as well as the acronyms with which we are going to identify them are also included. Following the classical terminology (Goodchild and Lam, 1980),

the units in which the variable of interest is available are called source units (or polygons) and the alternative units where the reallocated data are required are called target units.

**Table 1**. Summary of interpolation methods considered.

| | | Ancillary information | |
|---|---|---|---|
| **Name** | **Procedure of computation** | **Source** | **Variable** |
| M01. IDW point Interpolation: rIDW. | Target census section (CS) values are obtained from a population density surface constructed applying IDW interpolation on values located in source CS centroids. | No | n/a |
| M02. NN point Interpolation: rNN. | Similar to rIDW. The density surface is obtained by Natural Neighbors, instead of IDW. | No | n/a |
| M11. Areal weighting interpolation: AW. | Areal weighting interpolation. Weighted sum of source values, with source-target area intersections as weights. | No | n/a |
| M12. Pycnophylactic reallocation: TPYC. | Tobler's pycnophylactic interpolation method (Tobler 1979). | No | n/a |
| M21. SIOSE residential: SIOSE-RSD. | Weighted sum of source values, with source-target intersections, restricted to SIOSE residential areas, as weights. | SIOSE | Residential Area |
| M22. SIOSE population: SIOSE-Pop. | Similar to SIOSE-RSD, with weights weighted by relative population densities (Goerlich and Cantarino 2013). | SIOSE | Relative Population |
| M31. Urban Atlas residential: UA-UF. | Weighted sum of source densities, with source-target intersections, restricted to Urban Fabric areas, as weights. | Urban Atlas | Urban Fabric areas |
| M32. Urban Atlas residential: UA-SS. | Similar to UA-UF, with Urban Fabric areas weighted by relative soil sealing densities. | Urban Atlas | Urban Fabric soil sealing % |
| M41. Volume buildings: UA-VB. | Weights by volume (high) of buildings according to LiDAR, in the Urban Fabric areas defined by Urban Atlas. | U. Atlas LiDAR | LiDAR volume buildings |
| M42. Adjusted volume buildings: UA-VBadj. | Similar to UA-VB, with LiDAR data amended in class "high vegetation". | U. Atlas LiDAR | LiDAR volume buildings |
| M51. Road network length: CBCN-L. | Weights determined by the length of the streets, according to *Callejero*, within source-target intersections. | Carto BCN | Length of streets |
| M52. Buffer of streets: CBCN-SB. | Buffer of streets of *Callejero*, and source-target intersection, restricted to the buffer areas, as weights. | Carto BCN | Area buffer of streets |
| M61. Cadastral residential: C-RSD. | Weighted sum of source values, with source-target intersections, restricted to cadastral footprints of residential buildings, as weights. | Cadaster | Footprints of residential buildings |
| M62. Cadastral area homes: C-AH. | Weights by total area of housing in cadastral residential buildings. | Cadaster | Dwelling area |

M01-rIDW and M01-rNN are point-based approaches. These procedures (i) identify each source unit with its centroid, to which is assigned the observed value in the source unit, (ii) create a smooth prediction raster surface of the target variable, and (iii) estimate the variable of interest in each target polygon by averaging the estimated surface on it. The implementation of this process requires specifying a grid resolution and an interpolation procedure. We have worked with cells of size 50x50 meters and have used the methods

of inverse distance weighting (IDW) and of natural neighbor (NN) (as default in ArcGis® 10.2) as interpolation procedures.

In addition to being inefficient by reducing all the information of each source area to a point, the point-based approaches are not mass-preserving. That is, these procedures are not reversible: if we go back from target to source units, the outcomes will not coincide with the original values. The other twelve proposed methods are mass-preserving and share approach. All of them use as allocation function an estimator of the form given by equation (1), where $\hat{P}_j$ is the estimated population in target unit $j$, $P_i$ is the population of source unit $i$, $S$ is the total number of source units and $w_{ij}$ is the weight assigned to source unit $i$ in estimating the population of target unit $j$. The methods differ in the way the weights $w_{ij}$ are computed.

$$\hat{P}_j = \sum_{i=1}^{S} w_{ij} P_i \qquad (1)$$

The areal weighting (M11-AW) approach, also known as polygon overlay (Markoff and Shapiro 1973), is one of the most widely used methods (Goodchild and Lam 1980) and the most popular choice when ancillary information is not available. In this method, the source and target units are overlaid to obtain intersections and the (imputation) weights, $w_{ij}$, are determined by the ratio between the area of the intersection between the source unit $i$ and the target unit $j$ and the total area of unit $i$. Areal weighting is considered a simple method because, as is the case with point-based methods, it does not require any additional data besides source and targets units.

The implicit assumption of the M11-AW approach is that the variable of interest is homogeneously distributed in each source unit, which is quite unlikely. Hence, as an alternative, Tobler (1979) proposed the pycnophylactic method (M12-TPYC) that, maintaining the mass-preserving condition, assumes that the attribute values should not change abruptly at the boundaries of the source units (Kyriakidis 2004). This allows a different value to be assigned to each cell of each source unit, from which $w_{ij}$ weights are calculated. To preserve the pycnophylactic property, Tobler proposed an iterative procedure, with Mennis (2003) as an alternative. This method is also categorized as a

simple interpolation approach because no ancillary data is used to transfer the variable of interest from the source units to the target units.

The above methods (initially) handled the space of each source unit equally, irrespective of where the population is located and how this is distributed within it. Dasymetric mapping exploits ancillary sources to provide insights on how the population is spatially distributed. The most popular methods track where population is located in each source unit using two-dimensional (2-D) areal measures as auxiliary variables. From richer databases, three-dimensional (3-D) volume measures can be constructed to additionally know how the population is distributed inside of the source unit (Sridharan and Qiu 2013). Examples with one-dimensional (1-D) length measures can also be found in the literature (e.g., Reibel and Bufalino 2005).

The simplest 2-D dasymetric mapping methods are based on a binary classification of land uses: residential and nonresidential (Eicher and Brewer 2001; Holt et al. 2004; Sridharan and Qiu 2013). In the binary approach, nonresidential land use areas are considered unpopulated and consequently omitted (or zero weighted) in the distribution process. Although this refinement is expected to improve population allocations, it is not without flaws. On the one hand, residential areas usually include parts of a landscape (such as roads, footways and yards) where people do not reside. On the other hand, this approach implicitly assumes that within each source unit the population is evenly distributed across its residential areas, which again is seldom true (Maantay et al. 2007). This assumption often results in population underestimates in areas with high-rise buildings and in overestimates in areas with low-rise buildings (Harvey 2002; Sridharan and Qiu 2013). As alternatives, the 2-D multiclass (or polycategorical) dasymetric approaches and the 3-D methods try to amend these misestimates by accounting for the vertical distribution of the population. The 2-D multiclass approaches classify residential areas in more than a class with different population densities and the 3-D methods account for the height, the volume or the total area of the residential buildings. Note that the 2-D multiclass methods could be also be cataloged as 3-D procedures observing relative population densities as heights.

The 2-D binary methods that we have considered maintain quite a resemblance to the areal weighting procedure. In particular, the source and target units are again overlaid but this time restricted to residential areas and the weights, $w_{ij}$, determined by the ratio between the corresponding area of the intersection between the source unit $i$ i and the target unit $j$ and the residential area of unit $i$. That is, the numerator of the ratio is determined by the area of the intersection among the source unit $i$, the target unit $j$ and the residential land use polygons. The different methods diverge in the source employed to classify polygons as residential. M21-SIOSE-RSD method utilizes SIOSE, M32-UA-UF employs Urban Atlas and M61-C-RSD uses the footprints of the buildings classified as residential in the Cadaster. The details of the method M52-CBCN-SB are discussed in the next paragraph.

In 1-D length methods, it is assumed that the density of population across a unit is directly related to the density of road/power/streetlights network segments across the unit. Hence, under this assumption, the 1-D length methods calculate the weights, $w_{ij}$, as the ratio between the length of the network segments within the overlapping area between the source unit $i$ and the target unit $j$ and the total length of network segments within the unit $i$. The M51-CBCN-L method uses the Callejero road/street network. In Barcelona, however, the boundaries of the majority of census sections are placed on streets, so it is not uncommon (given the one-dimensional nature of street networks) that the majority of Callejero network segments located on street boundaries are completely subsumed in just one census section, when in these cases half of the corresponding length should be apportioned to each side of the street. To amend this, the M51-CBCN-SB method constructs a buffer of 10 meters centered in the road/street lines and proceeds as the other 2-D binary methods do, after viewing the buffer polygons as residential areas. This method could be classified as 1.5-D as it combines a 1-D auxiliary variable with a 2-D strategy.

Land use data are particularly useful as a means to distinguish residential areas from non-residential areas, but 2-D binary methods do not discriminate by residential attributes. In 2-D polycategorical dasymetric procedures, residential areas are divided into different groups (for example, from “Discontinuous Very Low Density Urban Fabric”

to “Continuous Urban Fabric”) and a different weight per area unit is assigned to each type according to their relative population density. In particular, denoting by $K$ the number of residential types, $d_k$ the relative population (or the relative soil sealing surface) of residential polygons of type $k$ and $w_{ijk}$ the area of the intersection among the source unit $i$, the target unit $j$ and the residential polygons of type $k$, we have that the $w_{ij}$ weights are reached after dividing $\sum_{k=1}^{K} d_k w_{ijk}$ by $\sum_{k=1}^{K} d_k w_{ik}$, where $w_{ik}$ is the area of the intersection between the source unit $i$ and the residential polygons of type $k$. The M22-SIOSE-Pop method employs the classification in apartment blocks, isolated apartment blocks, terraced houses and detached houses used by SIOSE to identify inhabited buildings and the M32-UA-SS procedures makes use of the five Urban Fabric categories defined in the Urban Atlas database.

The M41-UA-VB, M42-UA-VBadj and M62-C-AH approaches are 3-D methods. They all consider the vertical dimension of residential buildings to compute the weights. The M41-UA-VB and M42-UA-VBadj methods are really close to the methods employed in Qiu et al. (2010) and Sridharan and Qiu (2013), being the M42-UA-VBadj approach a refinement of the M41-UA-VB method.

In the M41-UA-VB method, we consider the residential areas defined by Urban Atlas and construct (LiDAR)-derived residential building volumes as product of LiDAR height measures and LiDAR areas corresponding to points confined in Urban Atlas residential areas and classified as buildings in the LiDAR database. In our approach, a 5x5 m cell is classified as a building when it contains any point classified as building. The relative heights of buildings from the ground are extracted from the LiDAR point cloud taking as reference MDT05 as digital elevation model. Once residential building volumes are calculated, we compute the $w_{ij}$ weights as the ratio between the total volume of the buildings located within the intersection of the source unit $i$ and the target unit $j$ and the total volume of unit $i$.

After a random revision of LiDAR point clouds, however, we were aware that many points classified as “high vegetation” in LiDAR should have been classified as building. Hence, to build the M42-UA-VBadj weights we reclassify as building cells those cells that being adjacent to a building cell and being located in an Urban Fabric polygon have a

height higher than 12 meters and occupy a plot with an area of more than 75 $m^2$. Afterwards we proceed in the same way as in the M41-UA-VB approach.

The M62-C-AH method follows a different strategy. This approach constructs weights exploiting the more detailed information available to the original unit used by INE to collect population data: the household. Among other issues, the cadaster contains information about the postal address, floor and square meters of each existing property in Spain. Thus, we combine the cadastral shp and cat files to compute in each cadastral polygon its total housing area to then compute the $w_{ij}$ weights as the ratio between the total housing area placed within the intersection of the source unit $i$ and the target unit $j$ and the total housing area of unit $i$.

All the spatial computations have been completed in ESRI ArcGIS® 10.2 (ESRI 2014), using its geo-processor ArcPy to create scripts. The calculation routines have been performed with Python 2.7.3 (Python Software Foundation, 2014).

## 4. Assessing allocations

A total of thirty-six variables (eighteen population variables in absolute values and the same number in percentages) have been allocated using fourteen methods for the 1,061 census sections defining the 2010 breakdown (i.e., more than half a million values have been calculated). This section evaluates the closeness of allocations and actual values. The results are quite clear and robust. The same order of preference emerges among groups of methods for all the variables and with all the measures of closeness calculated. The method M62-C-AH, which allocates variables using as ancillary data the total area of the homes in each residential building, is the one producing by far the most accurate results, whereas the point-based methods and the 1-D approach (M51-CBCN-L) are clearly the less accurate.

To evaluate which areal interpolation algorithm generates the most accurate estimates, three classical measures for summarizing the closeness between imputations, $\widehat{\mathrm{P}}_\mathrm{j}$, and actual values, $\mathrm{P}_\mathrm{j}$, in the $T$ target units have been computed for each combination of variable and method: the root of the mean square error (RMSE), the mean absolute

percentage error (MAPE) and the Pearson correlation coefficient (CORR); see Table 2. Although other measures may also be computed—such as the mean squared error (Sadahiro 2000), the adjusted RMSE (Hawley and Moellering 2005) or the value weighted MAPE (Qiu et al. 2012)—the conclusions would have remained the same. RMSE and MAPE statistics are distance measures. Thus, the smaller the RMSE and MAPE distances, the closer allocated values and actual values are. In many applications, however, more important than closely approximating actual values is to dispose of a variable really alike in correlation terms: a variable that may be used in a regression in place of the unobserved variable. In this sense, the closer CORR is to one, the more related (correlated) are allocated and actual values.

**Table 2**. Closeness measures between imputations ($\widehat{P}_j$) and actual values ($P_j$).

| Description | Acronyms | Equations |
|---|---|---|
| Root Mean Square Error | RMSE | $\sqrt{\frac{1}{\mathrm{T}}\sum_{\mathrm{j}=1}^{\mathrm{T}}\left(\widehat{\mathrm{P}}_{\mathrm{j}}-\mathrm{P}_{\mathrm{j}}\right)^2}$ |
| Mean Absolute Percentage Error | MAPE | $\frac{100}{\mathrm{T}}\sum_{\mathrm{j}=1}^{\mathrm{T}}\frac{\left\lvert\widehat{\mathrm{P}}_{\mathrm{j}}-\mathrm{P}_{\mathrm{j}}\right\rvert}{\mathrm{P}_{\mathrm{j}}}$ |
| Pearson Correlation | CORR | $\frac{\mathrm{T}\sum_{\mathrm{j}=1}^{\mathrm{T}}\widehat{\mathrm{P}}_{\mathrm{j}}\mathrm{P}_{\mathrm{j}}-\sum_{\mathrm{j}=1}^{\mathrm{T}}\widehat{\mathrm{P}}_{\mathrm{j}}\sum_{\mathrm{j}=1}^{\mathrm{T}}\mathrm{P}_{\mathrm{j}}}{\sqrt{\mathrm{T}\sum\widehat{\mathrm{P}}_j^2-\left(\sum\widehat{\mathrm{P}}_{\mathrm{j}}\right)^2}\sqrt{\mathrm{T}\sum\mathrm{P}_j^2-\left(\sum\mathrm{P}_{\mathrm{j}}\right)^2}}$ |

Table S1 (available in the supplementary online appendix) presents, by method and variable, the values of the distances (computed using RSME and MAPE as closeness measures) and of the correlations (calculated using CORR) between the amount of the people (in whole numbers) living in the 2010 census sections at the first of January, 2009 and the estimates obtained from the official values available in the 2009 census sections. The same statistics are displayed in Table S2 (see the online supplementary appendix) for the variables measured in percentages.

From the analysis of the numbers in both tables, several patterns clearly emerge. Firstly, the point-based interpolation strategies, in addition to not being volume-preserving, clearly present the worst approximations. In our application, the M51-CBCN-L method also shows similar figures. Secondly, as a rule, dasymetric methods are clearly preferable to simple methods. Thirdly, within dasymetric approaches, 3-D approaches are

preferable to 2-D procedures. Finally, the differences narrow when we work in relative terms (percentages) instead of in absolute values.

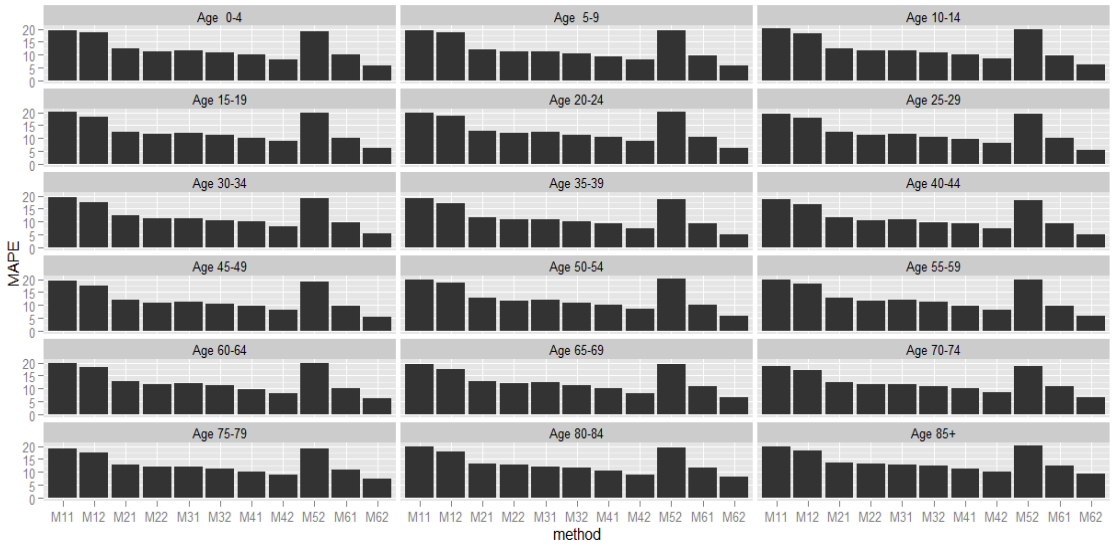


**Figure 3**. Differences by method, measured using MAPE, between the actual and allocated population absolute values for different age groups. The M01, M02 and M51 methods have been excluded to avoid that their large values dominate the scrutiny.

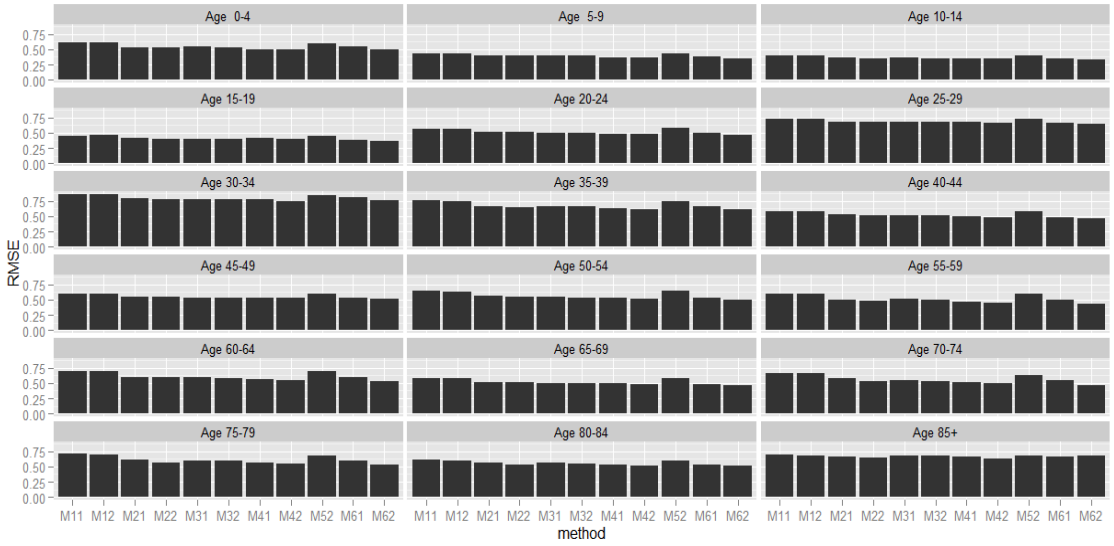


**Figure 4**. Differences by method, measured using RSME, between the actual and allocated population percentage values for different age groups. The M01, M02 and M51 methods have been excluded to avoid that their large values dominate the scrutiny.

To make these patterns more evident, Figure 3 and Figure 4 display by variable, excluding the three worst methods (M01, M02 and M51), some of the numbers

available in Tables S1 and S2. Figure 3 depicts MAPE distances for the allocations in absolute values and Figure 4 portrays RMSE distances for the estimates in percentages.

A closer look at the results even highlights some preferences within the different strategies. Focusing on the allocations of the population figures in whole numbers, the order of preferences we find is: M62 $\succ$ M42 $\succ$ M41 $\succ$ M32 $\succcurlyeq$ M22 $\succcurlyeq$ M31 $\succcurlyeq$ M61 $\succ$ M21 $\succ$ M12 $\succ$ M52 $\succ$ M11 $\succ$ M51 $\succ$ M02$\succ$ M01. That is, within the 3-D approaches, which are clearly the best as a group, the method M62-CBCN-SB, whose auxiliary variable best approaches the way the population information is originally collected, reveals itself as the procedure generating the most accurate results. Furthermore, we notice that the refinement that method M42-UA-VBadj represents over method M42-UA-VB also yields its fruits.

Likewise, although the differences among the 2-D dasymetric methods are not so evident at first glance, a deeper look reveals a clear pattern among 2-D subgroups. The more complex multiclass 2-D methods (M32-UA-SS and M22-SIOSE-Pop) generate as a rule more accurate results than the binary 2-D approaches. This should not be a surprise, given that in our theoretical disquisition we already note that these methods could even be considered as 3-D. Following multiclass 2-D procedures, the list continues in order of accuracy with the three binary 2-D methods, with the M21-SIOSE-RSD method placed in the third position of the trio, probably consequence of the higher quality of Urban Atlas and Cadaster databases compared to SIOSE database. Among the non-dasymetric approaches, areal weighting procedures are preferable to point-based interpolation, with the Tobler procedure generating better outputs than the simple areal weighting approach. It is interesting to observe that the method M52-CBCN-SB, which can be thought of as dasymmetric 1.5-D, presents levels of accuracy alike to those of the areal weighting approaches. Finally, the 1-D method M51-CBCN-L and the two point-based procedures are clearly placed towards the end of the list.

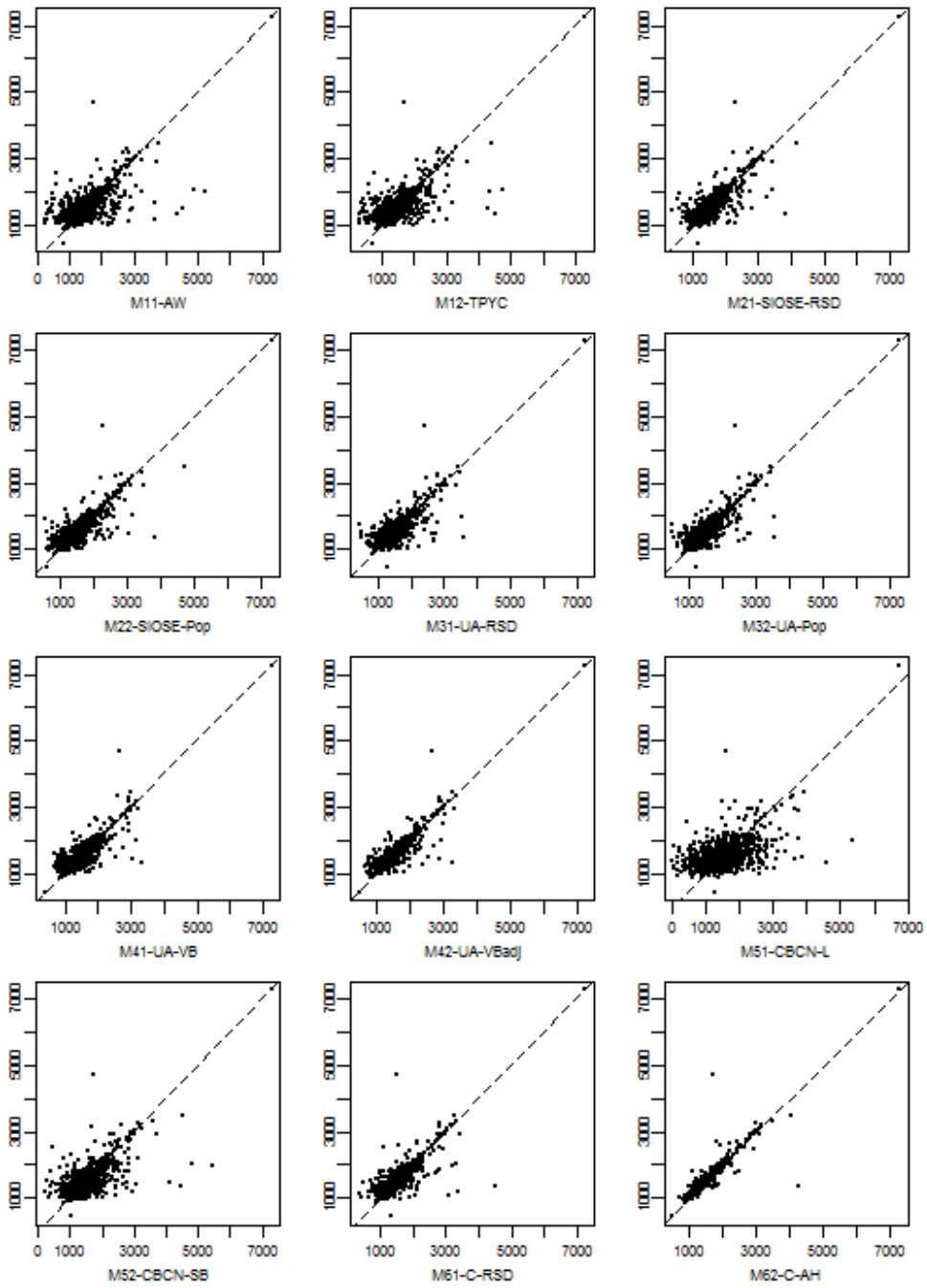


**Figure 5**. Comparing allocated estimates (horizontal axes) and actual values (vertical axes) of the 2009 total population of Barcelona (Spain) in the 2010 census section breakdown. The distance from the 45º line indicates how far apart allocations and actual values are. The number of data points in each scatterplot is 1,061. Details of the allocating methods can be found in Table 1. The point-based methods have been excluded given their limited quality.

The above conclusions are also manifestly visible observing Figure 5, where comparisons at the level of census section for the total population are displayed for all the methods but the point-based ones. Among panels of Figure 5, the panel corresponding to the M62-C-AH unambiguously stands out as the best and the one corresponding to M51-

CBCN-L unmistakably the worst. The rest of the estimates are in an intermediate position.

When we look at the results of closeness measures corresponding to the allocations of the variables in percentages (Table S2 and Figure 4), the results are not so evident. Nevertheless, although the differences reduce significantly, the order of preference among the methods is almost the same as when we allocate whole numbers. The order of preferences we find in this case is: M62 $\succ$ M42 $\succ$ M41 $\succ$ M61 $\succcurlyeq$ M22 $\succcurlyeq$ M32 $\succcurlyeq$ M31 $\succ$ M21 $\succ$ M52 $\succcurlyeq$ M12 $\succcurlyeq$ M11 $\succ$ M51 $\succ$ M02$\succ$ M01. Despite there being some changes in the relative order of some methods, the general picture remains: 3-D methods are preferable to polycategorical 2-D methods and these again preferable to binary 2-D methods, these are followed by areal weighting and 1-D methods and point-based interpolation procedures are again towards the end of the list.

## 5. Conclusions and final remarks

The question of reallocating population figures from a set of geographical administrative units onto another set of units is an issue that has received a great deal of attention in the literature. Every other day, a new procedure exploiting a previously unused ancillary source of information is introduced in the literature, claiming that it outperforms alternative algorithms. Given that an accurate location of population is crucial to answering many practical questions of social interest, the introduction in the reallocation process of new auxiliary variables through dasymetric mapping is the route commonly followed. Unfortunately, when the new (usually more complex) methods are applied to a new instance, the improvements achieved are sometimes dubious and just marginal. The tradeoff cost-effectiveness of each solution tends to be case-dependent: it usually depends on both the data and the geography.

In this sense, it is of unquestionable interested to know what approaches would generate satisfactory solutions under each group of circumstances since many studies have shown that different interpolation approaches can yield outcomes with really different implications. To evaluate which areal interpolation algorithm is the most appropriate for a given application, a significant majority of studies have been focused

on large areas with really heterogeneous population densities, the general conclusion being that as a rule more sophisticated methods are worth the extra effort they entail.

From a theoretical point of view, it could be proved that with a target variable uniformly distributed in the territory, areal weighting interpolation would produce perfect allocations. In the extreme opposite case, if people were concentrated in just a bunch of small subareas of the whole space, knowing the geographical distribution of the population would be a necessary condition to (depending on the exact breakdowns) yield accurate interpolations. Obviously, infinite in-between situations are possible. It could be argued, however, that when we work with (relative) homogeneous small source units whose variable to be allocated varies gradually between contiguous units, areal weighting interpolation could be enough and that any ancillary variable employed would yield marginal improvements. The spatial distribution of population in census sections of Barcelona in Spain meets these requirements. We have studied whether in these circumstances dasymetric reallocation techniques would provide significant improvement.

Our study shows that even under the above conditions the most sophisticated approaches clearly produce the better results. In particular, the method that allocates people using as ancillary variable the total dwelling area in each residential building is by far the one yielding the most accurate outcomes. In general, our study shows the 3-D methods generating the better outcomes followed, in order, by the multiclass 2-D procedures, the binary 2-D approaches and the areal weighting and 1-D algorithms. The point-based interpolation procedures are by far the ones producing the worst estimates.

Finally, it should be noted that, although some previous studies suggest that when working with small source units the use of more complex methods may be unnecessary when the variable to be distributed presents a strong spatial correlation (such as in the case of partisan vote proportion distributions), our analysis does not completely support this conclusion. Indeed, despite the differences between the methods showing a clear narrowing (for instance, from using areal weighting to using our best method the MAPE improvement gap reduces from 50% to 20%), we reach similar conclusions when we deal with the variables in percentages as well as in absolute numbers. Further research,

therefore, should be done to gauge to what extent more sophisticated approaches could be helpful in the process of reallocating a rate or a proportion in an urban area. In particular, it would be interesting to find out how the relative performance of simple and complex methods could be affected by the type of variable to be distributed, the final aims of the reallocation process and/or the urban structure under study. A possible line of future research would consist in studying whether spatial measures of urban morphology (such as the Moran index, as a summary of the combined effect of the variable to be distributed and the urban geography) are related to the accuracies of the areal distribution methods.

**References**


Batista e Silva, F., Gallego, J. and Lavalle, C. (2013). A high-resolution population grid map for Europe. *Journal of Maps*, 9(1), 16–28.

Bracken, I. and Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environmental and Planning A*, 21(4), 537–543.

Briggs, D.J., Gulliver, J., Fecht, D. and Vienneau, D.M. (2007). Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sensing of Environment*, 108, 451–466.

Brinegar, S.J. and Popick, S.J. (2010). A Comparative Analysis of Small Area Population Estimation Methods. *Cartography and Geographic Information Science*, 37(4), 273–284

Broner, S.J. (2010). *Análisis Espacial de Datos Electorales. Aplicación al Municipio de Barcelona*. Universitat Politécnica de Catalunya: PhD Dissertation.

Catalán, B., Saurí, D. and Serra, P. (2008). Urban sprawl in the Mediterranean? Patterns of growth and change in the Barcelona Metropolitan Region 1993–2000. *Landscape and Urban Planning*, 85(3-4), 174–184.

Cantarino, I., Torrijo, F.J., Palencia, S. and Gielen, E. (2014). Assessing residential building values in Spain for risk analyses – application to the landslide hazard in the Autonomous Community of Valencia. *Natural Hazards and Earth System Science*, 14(11), 3015–3030.

Cliff, A.D. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion Limited.

Eicher, C. and Brewer, C. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28(2), 125–138.

ESRI (2014). *ArcGIS Desktop: Release 10.2*. Redlands, CA: Environmental Systems Research Institute.

Fisher, P.F. and Langford, M.J. (1995). Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, 27(2), 211–224.

Flowerdew, R. and Green, M. (1994). Areal interpolation and types of data. In *Spatial Analysis and GIS* (eds A. S. Fotheringham and P. A. Rogerson). London: Taylor and Francis, 121–145.

Giordano, A. and Cheever, L. (2010). Using dasymetric mapping to identify communities at risk from hazardous waste generation in San Antonio, Texas. *Urban Geography*, 31(5), 623–647.

Goerlich, F.J., and Cantarino, I. (2013). A population density grid for Spain. *International Journal of Geographical Information Science,* 27(12), 2247–2263.

Goodchild, M. and Lam, N.S.N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing*, 1, 297–312.

Gregory, I.N. and Ell, P.S. (2005). Breaking the boundaries: geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society A*, 168(2), 419–437.

Harvey, J.T. (2002). Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing*, 23(10), 2071–95.

Hawley, K. and Moellering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, 32, 411–423.

Holt, J.B., Lo, C.P. and Hodler, T.W. (2004). Dasymetric estimation of population density and areal interpolation of census data. *Cartography and Geographic Information Science*, 31(2), 103–121.

Kyriakidis, P.C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3), 259–89.

Maantay, J.A., Maroko, A.R. and Herrmann, C. (2007). Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science,* 34(2), 77–102.

Markoff, J. and Shapiro G. (1973). The linkage of data describing overlapping geographical units. *Historical Methods Newsletter,* 7(1), 34–46.

Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, 14(1), 90–97.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55, 31–42.

Mennis, J. (2009). Dasymetric mapping for estimating population in small areas. *Geography Compass*, 3/2, 727–745

OJEU (2013). Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002 (1). *Official Journal of the European Union*, 56, L164/16–L164/19.

Pavía, J.M. and López-Quilez, A. (2013). Spatial vote redistribution in redrawn polling units. *Journal of the Royal Statistical Society A*, 176(3), 655–678.

Python Software Foundation (2014). *Python Language Reference, version 2.7.3*. http://www.python.org

Qiu, F., Sridharan, H. and Chun, Y. (2010). Spatial autoregressive models for population estimation at the block level using LIDAR derived volume information. *Cartography and Geographic Information Science*, 37(3), 239–57.

Qiu, F., Zhang, C. and Zhou, Y. (2012). The development of an areal interpolation ArcGIS extension and a comparative study. *GIScience & Remote Sensing*, 49(5), 644–663.

Reibel, M. and Bufalino, M.E. (2005). Streetweighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environment and Planning A*, 37, 127–39.

Robinson, S., Langford, M. and Tate, N. (2002). Modelling population distribution with OS LandLine.Plus data and Landsat imagery. In Proc. GIS Research UK 11th A. Conf. (eds S. Wise, P. Brindley, Y.-H. Kim and C. Openshaw). Sheffield: University of Sheffield, 320–325.

Sadahiro, Y. (2000). Accuracy of count data estimated by the point in polygon method. *Geographical Analysis*, 32, 64–89.

Sridharan, H. and Qiu, F. (2013). A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geographical Analysis*, 45, 238–258.

Suárez Vega, R., Santos Peñate, D.R. and Dorta González, P. (2008). Generación de un modelo superficial de la población de Gran Canaria. In Tecnologías de la Información Geográfica para el Desarrollo Territorial (eds L. Hernández and J.M. Parreño). Las Palmas de Gran Canaria: Servicio de Publicaciones y Difusión Científica de la ULPGC, 183–193.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519–30.

Tsai, Y.-H. (2005). Quantifying urban form: compactness versus 'sprawl'. *Urban Studies*, 42(1), 141–161.

UK Parliament (2007). *Statistics and Registration Service Act 2007 (c. 18)*. London: The Stationery Office Limited.

USC (2002). Confidential information protection and statistical efficiency act of 2002. *Public Law 107-347 "E-Government Act"*. 116 STAT.2962–116 STAT.2970.

US Census Bureau (2015). *Confidentiality of Public Use of Microdata Sample (PUMS)*. http://www.census.gov/acs/www/data_documentation/pums_confidentiality/

Wright, J.K. (1936). A method of mapping densities of population with Cape Cod as an example. *The Geographical Review*, 26(1), 103–110.

Zandbergen, P.A. and Ignizio, D.A. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science*, 37(3), 199–214.

Zandbergen, P.A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, 15(s1), 5–27.

**SUPPLEMENTARY (ONLINE) APPENDIX**

**Table S1**. Closeness of actual and allocated population absolute values by method and variable.

| | | Variable: Age Group | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | 0-4 | 5-9 | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | >85 | Mean |
| Root Mean Square Error | M01 | 45.0 | 44.7 | 44.9 | 45.2 | 44.2 | 44.0 | 44.8 | 44.8 | 44.4 | 44.4 | 44.6 | 44.6 | 44.4 | 44.9 | 45.2 | 46.2 | 47.1 | 47.3 | 45.0 |
| | M02 | 43.4 | 43.3 | 43.4 | 43.8 | 42.7 | 43.1 | 43.8 | 43.5 | 43.2 | 43.1 | 42.9 | 42.9 | 43.2 | 43.9 | 44.2 | 45.0 | 45.7 | 45.4 | 43.7 |
| | M11 | 19.4 | 19.5 | 20.1 | 20.0 | 20.0 | 19.4 | 19.4 | 19.0 | 18.8 | 19.4 | 20.0 | 19.7 | 20.0 | 19.5 | 18.7 | 19.1 | 19.6 | 19.8 | 19.5 |
| | M12 | 18.7 | 18.7 | 18.1 | 18.3 | 18.5 | 17.9 | 17.5 | 17.3 | 16.9 | 17.6 | 18.5 | 18.2 | 18.5 | 17.7 | 16.9 | 17.5 | 17.9 | 18.2 | 17.9 |
| | M21 | 12.3 | 12.2 | 12.6 | 12.7 | 12.9 | 12.6 | 12.5 | 11.9 | 11.7 | 12.1 | 12.7 | 12.7 | 12.7 | 13.0 | 12.4 | 12.7 | 13.2 | 13.5 | 12.6 |
| | M22 | 11.3 | 11.3 | 11.7 | 11.9 | 12.2 | 11.5 | 11.3 | 10.8 | 10.7 | 11.1 | 11.8 | 11.8 | 11.8 | 12.1 | 11.7 | 12.2 | 12.8 | 13.3 | 11.7 |
| | M31 | 11.7 | 11.3 | 11.9 | 12.1 | 12.3 | 11.7 | 11.4 | 10.9 | 10.7 | 11.5 | 12.1 | 12.1 | 12.1 | 12.3 | 11.6 | 12.1 | 12.2 | 13.0 | 11.8 |
| | M32 | 10.9 | 10.4 | 11.0 | 11.2 | 11.3 | 10.7 | 10.5 | 10.0 | 9.8 | 10.5 | 11.1 | 11.3 | 11.4 | 11.5 | 10.8 | 11.2 | 11.5 | 12.4 | 11.0 |
| | M41 | 10.1 | 9.6 | 10.1 | 10.2 | 10.4 | 9.8 | 10.0 | 9.2 | 9.2 | 9.7 | 10.0 | 9.7 | 9.6 | 10.1 | 10.2 | 10.3 | 10.7 | 11.4 | 10.0 |
| | M42 | 8.3 | 8.1 | 8.6 | 8.9 | 8.9 | 8.2 | 8.1 | 7.5 | 7.3 | 8.1 | 8.6 | 8.2 | 8.4 | 8.3 | 8.5 | 8.8 | 9.0 | 10.1 | 8.5 |
| | M51 | 40.6 | 39.5 | 39.9 | 40.9 | 41.6 | 40.1 | 40.2 | 42.0 | 40.0 | 40.4 | 40.8 | 40.3 | 40.7 | 40.9 | 40.3 | 40.0 | 40.0 | 41.0 | 40.5 |
| | M52 | 19.2 | 19.3 | 19.8 | 19.8 | 20.2 | 19.5 | 19.0 | 18.7 | 18.4 | 19.0 | 20.1 | 20.0 | 19.9 | 19.6 | 18.8 | 18.9 | 19.6 | 20.2 | 19.4 |
| | M61 | 10.2 | 9.8 | 10.0 | 10.2 | 10.5 | 10.1 | 9.9 | 9.5 | 9.3 | 9.6 | 10.2 | 9.9 | 10.3 | 10.8 | 10.9 | 11.0 | 11.6 | 12.5 | 10.3 |
| | M62 | 5.9 | 5.9 | 6.2 | 6.3 | 6.2 | 5.6 | 5.7 | 5.1 | 5.0 | 5.4 | 5.9 | 5.8 | 6.2 | 6.6 | 6.8 | 7.3 | 8.0 | 9.3 | 6.3 |
| Mean Absolute Percentage Error | M01 | 26.5 | 23.1 | 22.3 | 25.3 | 37.8 | 59.0 | 63.6 | 53.3 | 45.5 | 41.6 | 37.1 | 35.0 | 34.0 | 27.8 | 28.5 | 29.0 | 22.9 | 20.5 | 35.2 |
| | M02 | 26.7 | 23.1 | 22.0 | 24.9 | 36.7 | 57.7 | 62.9 | 53.2 | 45.5 | 41.1 | 36.4 | 34.2 | 33.6 | 27.6 | 28.4 | 28.8 | 22.5 | 20.1 | 34.7 |
| | M11 | 23.5 | 18.7 | 16.8 | 18.7 | 28.9 | 42.3 | 45.5 | 41.7 | 34.9 | 32.1 | 29.2 | 24.4 | 23.4 | 18.2 | 18.2 | 18.0 | 12.9 | 11.6 | 25.5 |
| | M12 | 22.7 | 18.1 | 15.4 | 17.2 | 27.8 | 40.7 | 42.9 | 39.1 | 32.3 | 30.0 | 27.3 | 22.6 | 21.4 | 16.3 | 16.0 | 16.0 | 11.6 | 10.7 | 23.8 |
| | M21 | 15.6 | 12.5 | 11.3 | 13.1 | 21.3 | 31.5 | 32.8 | 28.9 | 23.8 | 23.0 | 20.4 | 16.2 | 15.7 | 12.7 | 12.6 | 12.7 | 9.4 | 8.9 | 17.9 |
| | M22 | 15.3 | 11.8 | 10.6 | 12.3 | 20.6 | 30.6 | 31.6 | 27.9 | 22.9 | 21.6 | 19.2 | 15.4 | 15.0 | 11.9 | 11.7 | 12.0 | 9.5 | 9.7 | 17.2 |
| | M31 | 14.8 | 12.3 | 11.1 | 12.2 | 20.1 | 30.1 | 30.7 | 26.9 | 22.5 | 21.2 | 18.5 | 15.7 | 15.4 | 12.4 | 12.2 | 12.5 | 9.5 | 9.8 | 17.1 |
| | M32 | 14.1 | 11.7 | 10.4 | 11.5 | 19.6 | 29.5 | 29.8 | 25.9 | 21.5 | 20.1 | 17.6 | 14.9 | 14.7 | 11.6 | 11.4 | 11.8 | 9.2 | 9.9 | 16.4 |
| | M41 | 13.3 | 10.5 | 9.5 | 10.1 | 18.4 | 29.0 | 29.9 | 24.5 | 20.0 | 18.2 | 15.5 | 12.6 | 12.5 | 10.0 | 10.5 | 11.0 | 8.8 | 10.0 | 15.2 |
| | M42 | 11.2 | 9.7 | 8.7 | 9.5 | 17.1 | 26.3 | 25.9 | 21.6 | 18.0 | 17.1 | 14.3 | 11.5 | 11.3 | 8.7 | 9.2 | 9.4 | 7.4 | 8.8 | 13.6 |
| | M51 | 25.3 | 21.7 | 20.5 | 23.2 | 36.1 | 53.5 | 55.5 | 48.4 | 42.0 | 39.8 | 37.3 | 34.9 | 34.5 | 26.2 | 25.7 | 24.6 | 18.8 | 17.1 | 32.5 |
| | M52 | 20.7 | 17.2 | 16.2 | 18.6 | 29.4 | 42.4 | 43.3 | 38.3 | 32.6 | 31.8 | 29.5 | 25.2 | 23.7 | 17.9 | 17.5 | 17.3 | 12.9 | 11.7 | 24.8 |
| | M61 | 13.6 | 10.8 | 10.0 | 11.3 | 23.6 | 36.7 | 35.7 | 28.7 | 22.8 | 20.0 | 18.2 | 14.9 | 14.7 | 12.2 | 12.3 | 12.4 | 9.4 | 10.3 | 17.6 |
| | M62 | 7.5 | 6.5 | 6.1 | 6.7 | 19.1 | 31.2 | 28.1 | 20.9 | 16.3 | 13.7 | 11.4 | 8.4 | 8.4 | 6.9 | 7.1 | 7.9 | 7.6 | 10.0 | 12.4 |
| Pearson Correlation | M01 | .677 | .658 | .633 | .593 | .639 | .680 | .651 | .613 | .556 | .534 | .552 | .626 | .568 | .502 | .526 | .500 | .546 | .646 | .594 |
| | M02 | .661 | .630 | .613 | .583 | .638 | .667 | .622 | .586 | .531 | .528 | .549 | .620 | .552 | .482 | .498 | .475 | .530 | .633 | .578 |
| | M11 | .697 | .725 | .745 | .744 | .745 | .785 | .780 | .723 | .698 | .679 | .662 | .773 | .754 | .754 | .777 | .783 | .828 | .851 | .750 |
| | M12 | .711 | .737 | .776 | .777 | .765 | .802 | .803 | .751 | .730 | .708 | .695 | .804 | .793 | .802 | .827 | .827 | .859 | .872 | .780 |
| | M21 | .840 | .856 | .869 | .856 | .853 | .875 | .875 | .846 | .831 | .805 | .810 | .890 | .872 | .861 | .878 | .877 | .898 | .906 | .861 |
| | M22 | .844 | .870 | .884 | .873 | .863 | .882 | .884 | .856 | .844 | .827 | .829 | .902 | .884 | .878 | .896 | .891 | .896 | .889 | .872 |
| | M31 | .854 | .859 | .872 | .876 | .870 | .886 | .889 | .866 | .847 | .831 | .843 | .896 | .877 | .866 | .886 | .882 | .898 | .885 | .871 |
| | M32 | .865 | .871 | .885 | .888 | .876 | .890 | .895 | .874 | .858 | .846 | .857 | .906 | .886 | .881 | .900 | .894 | .903 | .883 | .881 |
| | M41 | .878 | .893 | .904 | .914 | .891 | .894 | .895 | .886 | .875 | .874 | .890 | .934 | .919 | .912 | .915 | .908 | .911 | .882 | .899 |
| | M42 | .913 | .908 | .918 | .924 | .905 | .913 | .921 | .911 | .898 | .887 | .905 | .945 | .932 | .932 | .934 | .931 | .937 | .908 | .918 |
| | M51 | .651 | .651 | .652 | .635 | .616 | .673 | .691 | .651 | .606 | .560 | .515 | .569 | .520 | .560 | .611 | .646 | .672 | .709 | .622 |
| | M52 | .746 | .754 | .754 | .739 | .730 | .781 | .796 | .756 | .724 | .675 | .643 | .743 | .736 | .756 | .787 | .794 | .823 | .844 | .755 |
| | M61 | .878 | .891 | .898 | .895 | .821 | .831 | .852 | .847 | .843 | .850 | .849 | .908 | .889 | .869 | .883 | .883 | .900 | .875 | .870 |
| | M62 | .961 | .958 | .960 | .962 | .882 | .877 | .906 | .915 | .915 | .927 | .939 | .971 | .963 | .957 | .960 | .951 | .933 | .881 | .934 |

**Table S2**. Closeness of actual and allocated population percentages by method and variable.

| | | Variable: Age Group | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | 0-4 | 5-9 | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 | 50-54 | 55-59 | 60-64 | 65-69 | 70-74 | 75-79 | 80-84 | >85 | Mean |
| Root Mean Square Error | M01 | 10.7 | 9.6 | 9.9 | 10.4 | 10.0 | 8.5 | 8.8 | 8.5 | 7.2 | 7.4 | 9.6 | 11.0 | 11.5 | 11.2 | 11.8 | 13.5 | 14.8 | 16.6 | 10.6 |
| | M02 | 10.3 | 9.4 | 9.7 | 10.1 | 9.5 | 8.2 | 8.4 | 8.2 | 7.1 | 7.2 | 9.4 | 10.7 | 11.0 | 11.0 | 11.6 | 13.2 | 14.5 | 16.1 | 10.3 |
| | M11 | 6.9 | 6.6 | 6.6 | 6.6 | 6.5 | 5.5 | 5.8 | 5.3 | 4.6 | 5.1 | 6.2 | 6.7 | 7.3 | 7.8 | 8.4 | 9.1 | 10.0 | 10.7 | 7.0 |
| | M12 | 6.9 | 6.5 | 6.6 | 6.6 | 6.4 | 5.5 | 5.7 | 5.2 | 4.5 | 5.1 | 6.1 | 6.6 | 7.2 | 7.7 | 8.4 | 9.0 | 9.9 | 10.5 | 6.9 |
| | M21 | 6.3 | 6.0 | 6.2 | 6.0 | 6.0 | 5.1 | 5.3 | 4.7 | 4.2 | 4.7 | 5.6 | 5.9 | 6.5 | 7.1 | 7.6 | 8.4 | 9.4 | 10.2 | 6.4 |
| | M22 | 6.2 | 5.9 | 6.1 | 6.0 | 6.0 | 5.1 | 5.3 | 4.6 | 4.1 | 4.6 | 5.6 | 5.7 | 6.5 | 7.1 | 7.5 | 8.2 | 9.2 | 10.1 | 6.3 |
| | M31 | 6.2 | 5.9 | 6.0 | 5.9 | 5.8 | 5.1 | 5.2 | 4.6 | 4.1 | 4.6 | 5.5 | 5.8 | 6.3 | 7.0 | 7.3 | 8.2 | 9.0 | 9.9 | 6.2 |
| | M32 | 6.2 | 5.8 | 6.0 | 5.8 | 5.8 | 5.0 | 5.1 | 4.6 | 4.1 | 4.6 | 5.4 | 5.8 | 6.3 | 7.0 | 7.2 | 8.1 | 9.0 | 9.8 | 6.2 |
| | M41 | 6.1 | 5.7 | 6.0 | 6.0 | 5.7 | 5.0 | 5.1 | 4.4 | 4.0 | 4.6 | 5.4 | 5.6 | 6.3 | 6.9 | 7.2 | 7.9 | 8.8 | 9.8 | 6.2 |
| | M42 | 5.8 | 5.5 | 5.7 | 5.7 | 5.5 | 4.8 | 4.9 | 4.3 | 3.8 | 4.4 | 5.2 | 5.3 | 6.0 | 6.6 | 6.9 | 7.5 | 8.5 | 9.4 | 5.9 |
| | M51 | 8.8 | 8.5 | 8.5 | 8.8 | 8.8 | 7.1 | 7.4 | 7.0 | 6.1 | 6.6 | 8.1 | 9.3 | 10.1 | 10.2 | 10.7 | 11.7 | 13.1 | 14.3 | 9.2 |
| | M52 | 7.0 | 6.7 | 6.9 | 6.8 | 6.8 | 5.7 | 5.9 | 5.5 | 4.7 | 5.2 | 6.4 | 7.0 | 7.5 | 8.0 | 8.5 | 9.4 | 10.4 | 11.2 | 7.2 |
| | M61 | 6.1 | 5.6 | 5.9 | 5.7 | 5.6 | 4.8 | 5.1 | 4.5 | 4.0 | 4.5 | 5.3 | 5.6 | 6.1 | 6.7 | 7.2 | 8.1 | 8.9 | 9.8 | 6.1 |
| | M62 | 5.6 | 5.3 | 5.5 | 5.3 | 5.2 | 4.5 | 4.7 | 4.0 | 3.7 | 4.1 | 4.9 | 4.9 | 5.5 | 6.2 | 6.5 | 7.2 | 8.0 | 9.3 | 5.6 |
| Mean Absolute Percentage Error | M01 | .823 | .549 | .504 | .607 | .792 | 1.000 | 1.145 | 1.055 | .755 | .764 | .897 | .914 | .924 | .739 | .824 | .906 | .776 | .810 | .821 |
| | M02 | .772 | .538 | .490 | .584 | .760 | .969 | 1.085 | 1.002 | .740 | .738 | .858 | .878 | .905 | .716 | .795 | .872 | .754 | .792 | .792 |
| | M11 | .613 | .439 | .405 | .456 | .562 | .734 | .862 | .759 | .581 | .598 | .639 | .605 | .701 | .583 | .663 | .707 | .610 | .692 | .623 |
| | M12 | .607 | .436 | .401 | .457 | .556 | .729 | .853 | .750 | .576 | .598 | .636 | .601 | .695 | .578 | .658 | .697 | .596 | .679 | .617 |
| | M21 | .532 | .406 | .366 | .409 | .515 | .686 | .791 | .664 | .529 | .550 | .567 | .505 | .599 | .522 | .575 | .620 | .564 | .671 | .560 |
| | M22 | .527 | .395 | .357 | .400 | .509 | .678 | .781 | .647 | .520 | .541 | .553 | .484 | .594 | .517 | .526 | .570 | .529 | .653 | .543 |
| | M31 | .542 | .402 | .362 | .402 | .499 | .682 | .781 | .666 | .521 | .536 | .540 | .507 | .591 | .503 | .539 | .597 | .556 | .683 | .551 |
| | M32 | .538 | .397 | .356 | .394 | .494 | .679 | .776 | .661 | .512 | .532 | .528 | .492 | .586 | .506 | .532 | .589 | .552 | .678 | .544 |
| | M41 | .506 | .374 | .357 | .419 | .485 | .674 | .772 | .631 | .491 | .534 | .530 | .470 | .568 | .492 | .508 | .570 | .531 | .655 | .531 |
| | M42 | .492 | .367 | .346 | .400 | .484 | .659 | .752 | .619 | .477 | .527 | .516 | .451 | .555 | .478 | .490 | .547 | .511 | .628 | .517 |
| | M51 | .643 | .492 | .461 | .539 | .691 | .852 | .967 | .836 | .673 | .713 | .745 | .787 | .865 | .702 | .743 | .807 | .721 | .788 | .724 |
| | M52 | .597 | .438 | .400 | .445 | .576 | .735 | .847 | .743 | .577 | .598 | .646 | .604 | .697 | .584 | .637 | .685 | .603 | .680 | .616 |
| | M61 | .544 | .391 | .353 | .389 | .503 | .668 | .815 | .668 | .489 | .524 | .535 | .491 | .589 | .489 | .542 | .601 | .537 | .669 | .544 |
| | M62 | .493 | .356 | .335 | .369 | .469 | .643 | .754 | .606 | .465 | .507 | .495 | .429 | .527 | .462 | .472 | .527 | .510 | .675 | .505 |
| Pearson Correlation | M01 | .809 | .864 | .870 | .829 | .822 | .875 | .870 | .825 | .828 | .782 | .817 | .848 | .851 | .852 | .846 | .829 | .842 | .828 | .838 |
| | M02 | .825 | .864 | .872 | .839 | .827 | .883 | .884 | .841 | .822 | .792 | .830 | .857 | .850 | .858 | .857 | .842 | .844 | .829 | .845 |
| | M11 | .880 | .906 | .909 | .899 | .898 | .931 | .925 | .903 | .883 | .861 | .898 | .929 | .907 | .902 | .897 | .892 | .887 | .867 | .899 |
| | M12 | .883 | .908 | .911 | .899 | .900 | .932 | .926 | .905 | .886 | .861 | .900 | .931 | .909 | .904 | .900 | .895 | .893 | .873 | .901 |
| | M21 | .911 | .920 | .926 | .919 | .916 | .940 | .937 | .927 | .904 | .883 | .921 | .951 | .933 | .923 | .924 | .918 | .904 | .875 | .919 |
| | M22 | .913 | .925 | .930 | .922 | .918 | .942 | .939 | .931 | .908 | .887 | .925 | .955 | .934 | .924 | .937 | .932 | .916 | .882 | .923 |
| | M31 | .907 | .922 | .928 | .922 | .921 | .941 | .939 | .926 | .907 | .889 | .929 | .951 | .935 | .928 | .934 | .925 | .907 | .871 | .921 |
| | M32 | .909 | .924 | .930 | .925 | .923 | .942 | .940 | .927 | .911 | .891 | .932 | .954 | .936 | .928 | .935 | .927 | .909 | .873 | .923 |
| | M41 | .920 | .933 | .930 | .916 | .925 | .942 | .940 | .934 | .918 | .890 | .931 | .958 | .940 | .931 | .941 | .931 | .916 | .881 | .927 |
| | M42 | .925 | .936 | .934 | .923 | .926 | .945 | .943 | .937 | .923 | .893 | .935 | .961 | .943 | .935 | .945 | .937 | .922 | .891 | .931 |
| | M51 | .867 | .881 | .880 | .854 | .841 | .907 | .904 | .881 | .840 | .794 | .860 | .877 | .854 | .856 | .870 | .857 | .838 | .823 | .860 |
| | M52 | .887 | .907 | .911 | .903 | .893 | .932 | .928 | .907 | .885 | .860 | .897 | .930 | .908 | .903 | .906 | .900 | .890 | .871 | .901 |
| | M61 | .907 | .927 | .932 | .927 | .919 | .943 | .933 | .925 | .919 | .895 | .930 | .954 | .935 | .932 | .933 | .923 | .914 | .876 | .923 |
| | M62 | .924 | .940 | .938 | .935 | .931 | .948 | .943 | .939 | .927 | .902 | .940 | .965 | .949 | .940 | .949 | .942 | .923 | .873 | .934 |