Cantarino Martí, I.; Goerlich, FJ. (2013). A population density grid for Spain. International Journal of Geographical Information Science. 27(12):1-17. doi:10.1080/13658816.2013.799283.

A population density grid for Spain.

Goerlich, Francisco; Cantarino Martí, Isidro (2013).
International Journal of Geographical Information Science. 1-17.
doi:10.1080/13658816.2013.799283

Abstract

This paper describes a high resolution land cover data set for Spain and its application to dasymetric population mapping (at census tract level). Eventually this vector layer is transformed into a grid format. The work parallels the effort of the Joint Research Centre (JRC) of the European Commission, in collaboration with Eurostat and the European Environment Agency (EEA), in building a population density grid for the whole of Europe, combining CORINE Land Cover with population data per commune. We solve many of the problems due to the low resolution of CORINE Land Cover, which are especially visible with Spanish data. An accuracy assessment is carried out from a simple aggregation of georeferenced point population data for the region of Madrid. The bottom-up grid constructed in this way is compared to our top-down grid. We show a great improvement over what has been reported from commune data and CORINE Land Cover, but the improvements seem to come entirely from the higher resolution data sets and not from the statistical modelling in the downscaling exercise. This highlights the importance of providing the research community with more detailed land cover data sets, as well as more detailed population data. The dasymetric grid is available free of charge from the authors upon request.

Keywords: Population density, Dasymetric mapping, Downscaling, CORINE Land Cover, Land Cover and Use Information System (SIOSE).

JEL Classification: J10, R14.

1. INTRODUCTION

It is widely known that the best way to produce gridded demographic maps of population density is collecting individual data with coordinates of the residential addresses and counting the number of people in each grid cell. This is called a bottom-up approach. However, very few European countries have georeferenced point population data available to pursue this approach: from individual data to the grid. In general, only areal unit population data is available, so creating a population surface should use areal interpolation techniques.

One of the most widely used techniques in this context is dasymetric mapping (Wright 1936). A dasymetric map depicts quantitative areal data using boundaries that divide the mapped area into zones of relative homogeneity with the purpose of best portraying the underlying statistical surface. Dasymetric zones are usually generated by using ancillary information.

Hence, when the bottom-up approach is not available, the alternative is to use spatial disaggregation methods; that is, to downscale the population data by means of auxiliary information and statistical techniques. This is known as the top-down approach, which generally uses land cover information to locate people more precisely and statistical methods to redistribute the population into selected polygons (Fisher and Langford 1995; Eicher and Brewer 2001; Holt, Lo and Hodler 2004).

There are a large range of possible approaches for downscaling. In a classic paper, Eicher and Brewer (2001) report three basic methods, which are all based on dasymetric mapping principles:

- The old and widely known binary method, which assigns the whole population of the administrative boundary to one cover class, either urban or artificial surface, depending on data availability (Langford and Unwin 1994; Langford 2007).

- The three-class method, which in addition to urban cover has information on agricultural or other surfaces capable of holding residential population, so that they receive some inhabitants (Mennis 2003). This method naturally extends to more classes, with differential densities per class, given enough detail in the land cover auxiliary data set.

- The limiting variable method which initially attributes the same density to all classes within each administrative unit; and subsequently densities are modified by applying different thresholds to each land cover class and redistributing the excess to other classes (Eicher and Brewer 2001).

In the study by Eicher and Brewer (2001), the limiting variable method gave the best results. A modified version of this method, which also incorporates regression techniques, is used by Gallego, Batista, Rocha and Mubareka (2011). Among the many methods used by these authors the modified limiting variable method appears to be best, though no single method seems to outperform others in all situations for the countries considered.

The aim of Gallego (2010) and Gallego, Batista, Rocha and Mubareka (2011) was to produce a dasymetric population map for Europe in grid format, so the zones are geographically determined, instead of taken as given by the administrative boundaries of the communes, which are delineated for historical and political reasons. Hence, in addition to a downscaling problem, they face the problem of transforming geographic data from one set of boundaries to another. The elaboration of this grid resulted from the efforts of the Joint Research Centre (JRC) of the European Commission, in collaboration with Eurostat and the European Environment Agency (EEA). Databases for this exercise were commune data from the last available census of population at that time, 2001, and CORINE Land Cover (CLC) as the auxiliary information for the spatial

disaggregation. These authors perform their exercise in raster format, obtaining a population density grid of 1-ha resolution according to European standards.

The general conclusion from Gallego, Batista, Rocha and Mubareka (2011) and their validation exercise was that, although the grid has some limitations, it is an improvement over choropleth maps at commune level and can provide a useful tool for researchers in different fields to carry out geographical analysis linking population with territorial features. However, downscaling is far from perfect, and the density attributed to non-urban classes is generally overestimated. In fact, these authors show that different statistical methods have limited capability to improve the final results. The analysis therefore confirms a well-known fact highlighted by Martin, Tate and Langford (2000, p.-358) more than a decade ago: "...input data quality is more important that algorithmic detail."

Given that there is room for improvement, the goal of this paper is to present a high resolution land cover data set for Spain and its application to the problem of dasymetric population mapping. With the exception of the binary method, a typical problem in using land cover information to redistribute population among classes is determining the threshold densities to be employed in the downscaling algorithm, given that typically there is no information on this available from own data sources. Using census tract population data and the Land Cover and Use Information System of Spain (SIOSE), we are able to get this information by standard GIS operations, so a very accurate population vector layer can be obtained by modelling densities per land cover class using regression methods. In this way, we build a dasymetric map in which inhabited zones are bounded within each census tract. Eventually this polygon vector layer is transformed into a grid format for a validation exercise with the region of Madrid, for which we have a bottom-up grid, and for a comparison with the validation results provided by Gallego, Batista, Rocha and Mubareka (2011) for other European countries.

Our contribution to the literature on dasymetric mapping is twofold. On the one hand, we optimally use an object oriented land cover data set that contains information about heterogeneity of land use; this data set is unique to Spain. On the other hand, contrary to the limiting variable method that starts with a uniform density, then sort classes by density and proceeds sequentially in the population redistribution per class; we use regression methods, that allow us to attribute population to all classes simultaneously.

The structure of the paper is as follows. Section 2 explains why the grid distributed by the EEA gives particularly bad results for Spain, as compared with other European countries. This section also introduces the Land Cover and Use Information System of Spain (SIOSE), and reports on the population data used. Section 3 presents our exercise,

describing the methods adopted, results obtained and the outcome of a validation exercise. A final section offers concluding comments.

2. LAND COVER INFORMATION AND THE DISTRIBUTION OF POPULATION

2.1. CORINE land cover and the distribution of population in Spain

Land cover information is a critical aspect of geographical and environmental information. In 1985, the European Commission launched the Corine Land Cover (CLC) project, to produce land cover databases for the whole of Europe. There are three versions of CLC according to the reference year (1990, 2000 and 2006), which are public and available in the EEA data service. We use the vector 2006 format. CLC can be considered the standard European land cover map and has been produced under common rules in all the countries of EU by means of photo-interpreting Landsat ETM+ images. As is widely known, CLC´s nomenclature has a hierarchical structure with 44 classes at level 3. Artificial surfaces comprise 11 classes, 2 of which are urban zones: continuous urban fabric and discontinuous urban fabric, where it can be expected that most of the people will reside.

Most of the inaccuracy problems when using CLC in downscaling population can be traced back to its poor resolution for this type of exercise: in particular the minimum mapping unit is 25 ha. Smaller patches are included in polygons labeled with the dominant land cover type. If there is no clearly dominant land cover class in the polygon, it is coded as "heterogeneous". This occurs in 11% of the EU area, but in 18% of the Spanish surface.

Other problems are related to the heterogeneous size and density of Spanish communes.[1] Table 1 gives a general description of the size of communes in terms of surface and 2010 population figures. While the average size is 62 km^2, the median size is almost half, 35 km^2. Heterogeneity in sizes is not strikingly different from what we find across Europe, but population concentration is. It is worth noting that in only 60 communes (0.7%), covering 9% of the national surface (which is less than half of what we find in Europe for the > 500 km^2 interval) live 14% of the population, more than twice the European percentage.[2] At the other end of the distribution, 5% of the population lives in the smallest 10% of the communes. These figures show a clear concentration towards big cities and away from small communes, some of which are tiny: 60% of the municipalities have a resident population of less than 1,000 inhabitants, and 13% less than 100. The tendency towards population concentration can be seen

[1] We use the Nomenclature of Territorial Units for Statistics (NUTS) developed by Eurostat. Hence, we use interchangeably the terms commune and municipality, and they refer to the Local Administrative Units 2 (LAU2) in the European terminology. For regions we use the smallest ones, the NUTS 3 regions. In Spain we have 50 NUTS 3 regions and 2 autonomous cities in northern Africa.

[2] European comparisons are based on Gallego (2010), using the 2001 census population figures and CLC2000, whereas we use 2010 population from the municipal registry and CLC2006.

historically in Spain for more than a century, and still continues with people moving to the coast, to the valleys and to the regional capitals, away from the mountains and rural areas (Goerlich and Mas 2008, 2009).

Table 1. Heterogeneity of commune (LAU2) sizes and population concentration in Spain

Size (km^2)	Communalities		Area		Population 2010	
	Number	%	km^2	%	Inhabitants	%
(0, 10]	771	9.5%	4,931	1.0%	2,562,385	5.4%
(10, 100]	6,054	74.6%	231,903	46.0%	23,119,824	49.2%
(100, 500]	1,229	15.1%	221,696	43.9%	14,799,103	31.5%
> 500	60	0.7%	46,105	9.1%	6,539,719	13.9%
Total	**8,114**	**100.0%**	**504,636**	**100.0%**	**47,021,031**	**100.0%**

	Median	Mean	Standard deviation
Area (km^2)	34.9	62.2	92.3
Population 2010	582	5,795	47,527

Source: Own elaboration from INE (http://www.ine.es). Municipal Registry and Territory.

The small communes and the asymmetry in the population size distribution are clearly seen if we compare the average size, 5,795 inhabitants, with the median size, 582 residents. In addition, the population is not scattered over the territory in most of Spain, but is heavily concentrated in small urban nuclei. Given the 25 ha minimum mapping unit of CLC, this implies that in many communalities CLC does not report any urban area, because they do not contain any urban patch larger than 25 ha within their boundary. Gallego (2010, p.-463) reports that this happens in 29% of the cases for Europe with CLC2000, but in Spain things are much worse. At national level, CLC2006 does not report urban area (classes 1.1.1. and 1.1.2.) in more than half of the municipalities, 57%, though they cover only 4% of the population. Obviously urban areas exist in these communes. At NUTS 3 level the situation is very heterogeneous: in 10 regions CLC2006 does not report any urban area in more than 75% of their municipalities, and in 2 regions this figure rises to 90%.

These communes require a separate treatment for population mapping. In practice, what is done is to define strata according to commune density and urban area reported by CLC, so that communes without urban areas form a separate stratum, and modeling is then done separately for the different strata.

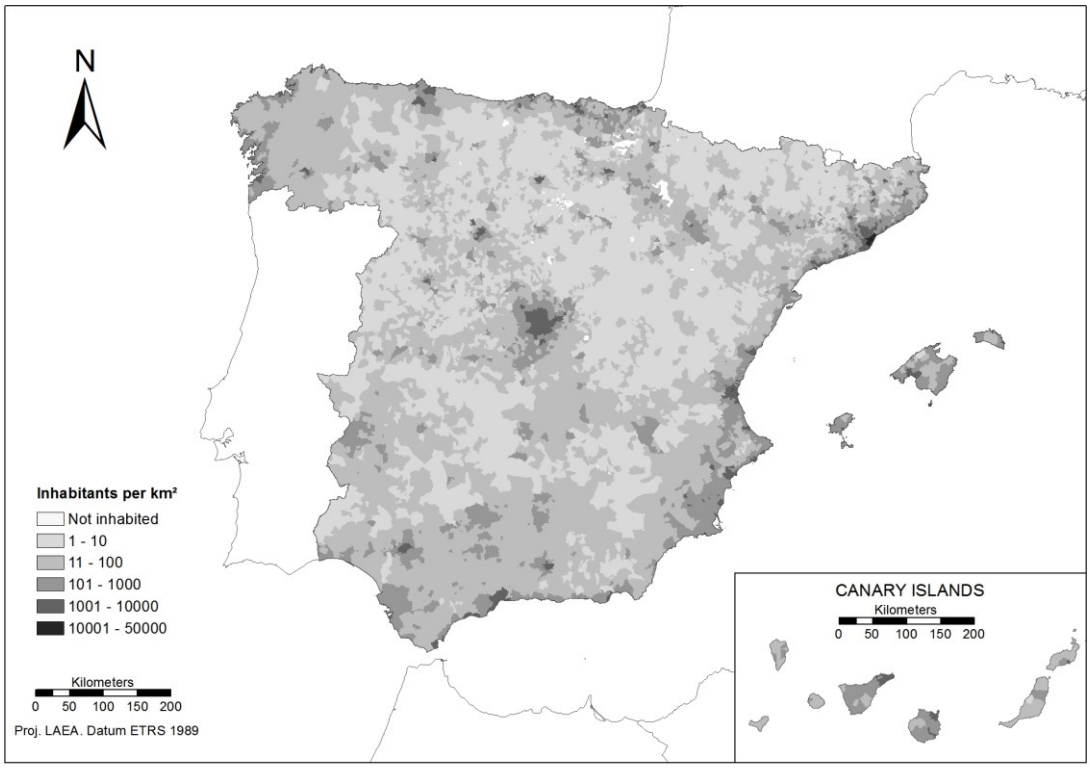
These figures are even more striking if we restrict ourselves to artificial surfaces. In half of the communalities CLC2006 does not report any of the artificial land cover classes. At NUTS 3 level, in 6 regions this happens in more than 75% of their municipalities. Given the pycnophylactic constraint (Tobler 1979) at commune level, the absence of urban areas in many of them implies that population is spread over other non-urban land cover classes, and thus eventually population in agricultural and forest

classes is overestimated, even if the initial assumed densities for these classes are rather small. Moreover, the effect is quite heterogeneous across regions.

Figure 1 shows a map of municipal density, and figure 2 shows the EEA population grid aggregated to 1 km^2 resolution, though the original resolution of the published grid is 1 ha. Differences are clearly visible at this level. The grid representation shows some empty spaces corresponding to land cover classes in which it is assumed that there is no resident population, but in figure 2 the population is clearly over-dispersed in many parts of Spain. In fact, 85% of grid cells contain population, which does not agree with observation of the Spanish geography, especially in the center and south of the country. The main reason for this distribution is the absence of urban land cover classes in many communes according to the CLC dataset.

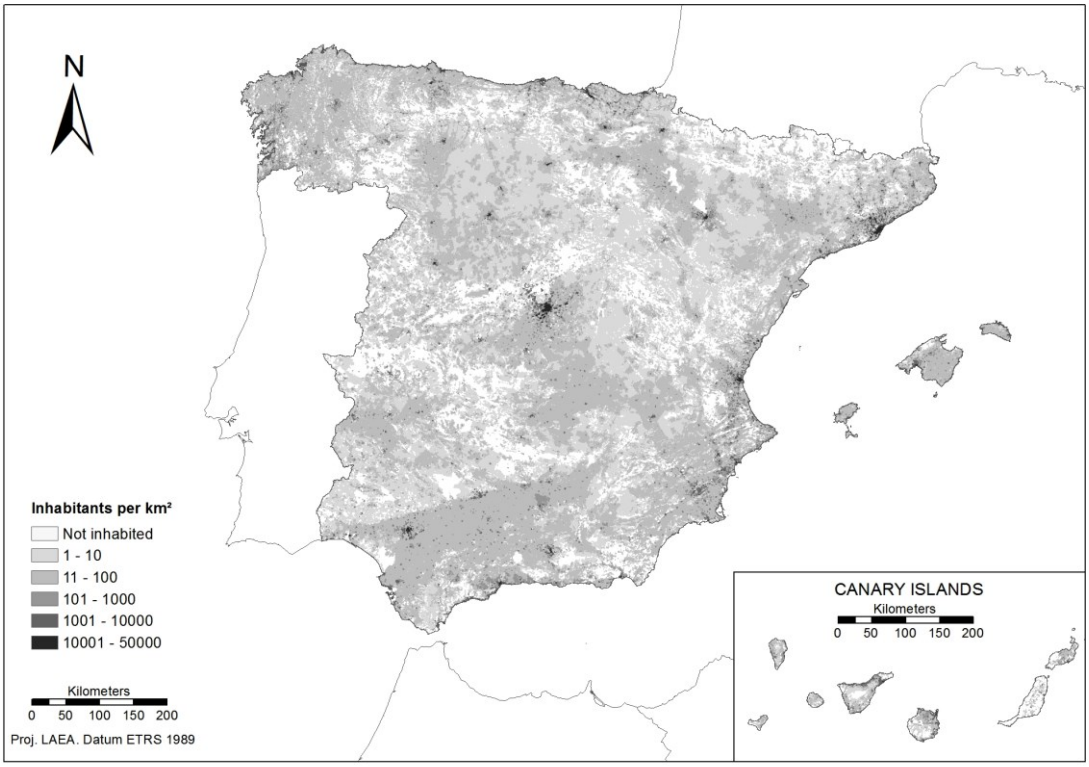
Given this lack of resolution, a new version of the grid, using a modified version of CLC2006 with the soil sealing layer of the EEA (Kopecky and Kahabka 2009) and other auxiliary information, is currently under development (Batista e Silva, Lavalle and Koomen 2012). The objective is to provide a more realistic population distribution across Europe when bottom-up grids are not available. As we shall see in the sequel, such a refinement of CLC is not necessary for Spain, since we have available a much higher resolution land cover database with very rich information. This land cover dataset can be used to locate precisely the population within administrative boundaries, which in turn can be used to obtain a grid at any desired resolution.

Figure 1. Commune (LAU2) population density. 2010.



Source: Municipal Registry: Instituto Nacional de Estadística; Instituto Geográfico Nacional and own elaboration.

Figure 2. 1km^2 population density grid. 2001.



Source: Original data: Joint Research Center – Eurostat – European Environment Agency. Census 2001 population data: Instituto Nacional de Estadística. Aggregation to 1km^2 resolution: European Forum for Geostatistics and own elaboration.

2.2 Land cover and use information system of Spain (SIOSE)

Building on the previous experience from CLC, the National Geographical Institute (IGN) developed a new Land Cover and Use Information System of Spain (SIOSE) with the same reference date as the last version of CLC, 2006. SIOSE is aimed at solving most of the problems of CLC and was produced by means of photo-interpreting SPOT5 satellite images, aerial photography, topographic maps at 1:25,000 scale and cadastral information. It is based on the INSPIRE principle: "spatial data collected at one level of public authority to be shared between all the different levels of public authorities" (SIOSE 2010). SIOSE is public and can be downloaded free of charge from the IGN warehouse (http://centrodedescargas.cnig.es/CentroDescargas/index.jsp).

From our point of view, there are two main characteristics of SIOSE that deserve consideration. The first is its resolution; the scale is 1:25,000 as against 1:100,000 in CLC, which translates into a minimum mapping unit that varies according to the land cover class:

- Agricultural land, Forest and Natural areas: 2 ha.
- Urban fabric and Water bodies: 1 ha.
- Wetlands, Beaches, Riverside vegetation and Coastal cliffs: 0.5 ha.

Therefore, in terms of population distribution, where urban fabric is the main cover to be considered, we have a minimum mapping unit of 1 ha against 25 ha in CLC. At this level of resolution all communes have some degree of urban fabric cover and thus a precise place in which to locate people within administrative boundaries is always found.

However, the most innovative feature of SIOSE is its data model. CLC is a hierarchical land cover data set: each polygon is assigned a unique cover from a fixed nomenclature that comprises 44 classes at level 3 of disaggregation. SIOSE is an object oriented land cover data set (Villa *et al* 2008): the polygon is the working unit and its coverage is homogeneous over the surface covered. The aim of SIOSE is, however, not to ascribe a cover to the polygon from a given nomenclature, but to describe its contents. The SIOSE data model consists of objects, attributes, relationships and consistency rules that integrates in a digital vector geographic data base to describe every polygon in which the national territory is divided. Contrary to CLC, SIOSE is only available in vector format, and cannot be translated to raster form easily.

Given a complex structure that starts from a simple list of 40 land cover elements, the SIOSE data model makes it possible to attribute one or more land cover for each polygon, with some restrictions. Polygons may therefore have a "simple coverage", when only one land cover element is present (i.e. buildings, water, rock or vegetation…); but more generally, they have a "composite coverage", when they have a combination of land cover elements (i.e. urban fabric, artificial agricultural settlement, or recreational park…). Because this combination of land cover elements is obtained using percentages, their sum must always be equal to 100%. In addition, attributes assigned to simple land cover elements give us further information (i.e. buildings can be isolated blocks or detached houses). In addition to the land cover elements, SIOSE includes a list of "predefined composite coverages". This list is characteristic of Spanish territory (i.e. family orchard, artificial agricultural settlement, urban fabric: old quarter, enlargement or discontinuous…), and is neither exhaustive of the surface nor a closed list, but facilitates extraction and manipulation of the database. The SIOSE´s data model makes different and potentially infinite combinations of land cover elements possible.

It is clear that the information provided by SIOSE is much more detailed than that provided not only by CLC (given its higher resolution), but by any other hierarchical land cover data model. In fact, for CLC2006 we have about 150 thousand polygons in Spain, with an average size of 3.3 km^2, belonging to one of the 44 different classes. In SIOSE we encounter about 2.5 million polygons, with an average size of 0.20 km^2, and around 820 thousand different land cover categories, in the sense of different combinations of land cover elements (different polygons).

It is worth noting that, given this structure, and in relation to residential areas and calculation of population densities, we have two surfaces available: the surface of the

polygon where the buildings are located and the net built-up area occupied by the residential buildings. In the next section, for the downscaling exercise the surfaces to be calculated and model population densities refer to this last surface. Thus, we expect to isolate (at the maximum), the distortionary effect of density calculation over heterogeneous surface area. Note that, even if the minimum mapping unit for urban fabric is 1 ha, land cover elements are represented within a polygon as long as they represent at least 5% of the polygon surface. We therefore expect residential developments of 500 m^2 or more to be represented in SIOSE.

2.3. Additional data

We use census tracts, or enumeration districts, for demographic variables. We have about 8,000 communes, but around 35,000 census tracts. Because most communes, 72%, have only one census tract we gain nothing in the smallest rural areas by going from commune to census tract population data; they represent about 6% of the population. However, they lead to a great improvement in medium and big cities, especially in city centers, as we shall see in the next section. In fact, 60% of the census tracts have a surface of less than 1 km^2. Again, a great deal of heterogeneity can be seen in census tract sizes. Mean size is 14.2 km^2, but median size is just 0.22 km^2, because of the small size in cities. Standard deviation is almost three times the mean, reaching a value of 38.7 km^2.

We use the year 2010 as the reference year for the population, dated 1st January, and the criteria to locate people is residence (night population). These data come from an administrative registry which is public, and can be downloaded free of charge from the Spanish NSI website (INE, http://www.ine.es/).

A vector shape file in GIS format for census tracts is available from INE and is supplied on demand at a given fee. This is our source for the original geography of the population. Because the 2006 file was full of topological errors, we decided to use 2010 as the reference year for the population, since this vector layer was free of topological errors.[3] The errors introduced by different reference dates between the land cover and population datasets are difficult to estimate and a better match would only be possible with the new release of SIOSE. In any case, the final result is much closer to reality than the standard choropleth maps.

Eventually, to transform the dasymetric map at census tract level after the intersection with SIOSE into a grid format we need a vector structure to assemble the final disaggregated population figures. The grid vector layer comes from the European Forum for Geostatistics website (EFGS, http://www.efgs.info/data/eurogrid), identical

[3] The 2010 file was free of topological errors, but it had some geometric inaccuracies when aggregated to commune boundaries with respect to the boundaries supplied by the *IGN*. These inaccuracies are the main reason why about 0.5% of the total population is missing in the downscaling process described in the next section.

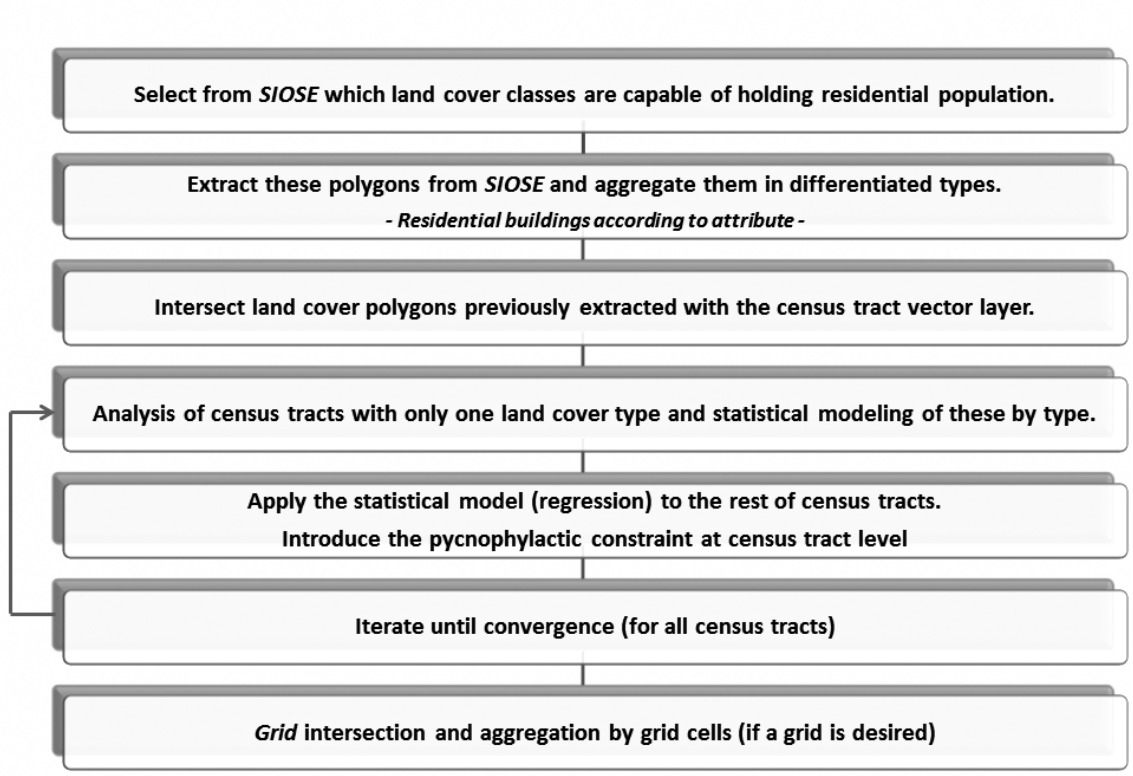
to the grids distributed by the EEA, and it is INSPIRE (2010a, 2010b) compliant. Because the final goal in transforming the disaggregated vector layer data into a grid format is to compare our results with the standard at the European level, we use a 1 km^2 resolution, but a finer grid could also be obtained.

3. DOWNSCALING POPULATION FROM CENSUS TRACTS AND SIOSE

The method used in downscaling population is a regression method (Yuan, Smith and Limp 1997) that, thanks to the high resolution of our datasets, does not require external reference data for implementation. Instead of starting from a uniform density, and then redistribute population by applying thresholds progressively after ranking classes by density, as it is done in the limiting variable method (Eicher and Brewer 2001); we proceed simultaneously after statistically modelling densities per class at census tract level. We start from the simple definition of population density, written as $Population = Density \times Surface$, and statistically model *Density* by regression methods, given available information on population and land use. The simplified work-flow can be seen in figure 3.

Finally, a validation exercise is performed with a bottom-up grid for the region of Madrid. This was produced by counting people with a census address in each grid cell, resolution is 1 km^2 and reference year is 2010. The original georeferenced point population data was kindly supplied for this purpose by Madrid´s regional statistical institute. All our work is carried out with vector layers.

Figure 3. Simplified work-flow diagram.



Given the structure of SIOSE, we choose as the support for residential population any polygon with buildings of non-industrial type, within the complex predefined land cover class of Artificial Agricultural Settlements or Urban Fabric of any of the types considered in SIOSE. Because of the available attribute information, buildings must be of the following types: apartment blocks, isolated or not, and houses, detached or terraced. Note that the chosen polygons can be urban or rural, since the defining characteristic for the polygon to be eligible as the support for residential population is to have buildings in it of any of the four types considered. Of course, we may exclude some land cover classes that perhaps hold residential population, but allowing people in other classes may disperse the population in excess. In fact, for our georeferenced point population data for Madrid we have checked that 97.8% of the population falls within our SIOSE polygons and it is unclear whether extending the list will improve the results.

We extract these polygons from SIOSE and aggregate them according to the type of building, so we eventually have four different types. The criterion for aggregating according to the type of building is that this is the characteristic which most affects population density. An analysis of variance for census tracts with only one type of building revealed this. Average population density was very heterogeneous according not only to the type of building, but also for different regions and strata by commune

size. However, the most variability was shown by type of building, and so we finally decided to keep only four differentiated types in the downscaling algorithm.

Next, land cover polygons extracted from SIOSE were intersected with the vector layer of census tracts. After this intersection many census tracts only had one type of building in them. This is true for the four types considered: apartment blocks, isolated or not, and houses, detached or terraced. These are called pure census tracts, and play a key role in the downscaling exercise. With CLC there are no pure classes, given its resolution and commune data, but in our case we have 56% of pure census tracts (most of them in cities, but also in the smallest municipalities for the terraced houses type). In total, they cover 51% of the population. For these cases, there is no problem of redistribution; SIOSE fixes the population within the census tract accurately. Statistically modelling the pure census tracts is essential to redistribute the population for the rest of the classes.

Given a set of land cover classes, *c*, with different densities per administrative unit (census tract), *m*; total population can be written as

$$P^m = \Sigma_c P_c^m = \Sigma_c d_c^m \times S_c^m \quad (1)$$

where d_c^m represents the density of class *c* in census tract *m* and S_c^m its surface, which in our case is the net built-up area, as mentioned in the previous section.

Equation (1) is an identity. For pure census tracts $c = 1$, and there is no redistribution problem; for the rest $c > 1$. We know S_c^m, from the geometric intersection between SIOSE and the census tract vector layer, and different methods of estimating d_c^m give us different redistribution algorithms. In addition, the algorithm should incorporate the volume constraint (Tobler 1979), so $P^m = \Sigma_c P_c^m$. This population is known from the municipal registry.

The simplest case is to assume that d_c^m is constant for each *c*, up to a scale factor,

$$d_c^m = \theta_c \times \lambda^m \quad (2)$$

where θ_c depends on class only, for example the type of building, and can be inferred from the information on pure census tracts for each *c*. A natural candidate for θ_c is the density per class. These densities can be applied to the rest of the census tracts to get a population estimate given S_c^m. The factor λ^m ensures that, at the end of the process, the pycnophylactic constraint is satisfied, so eventually $P^m = \Sigma_c P_c^m$. Substituting (2) into (1)

$$P^m = \sum_c \theta_c \times \lambda^m \times S_c^m = \lambda^m \times \sum_c \theta_c \times S_c^m \qquad \Rightarrow \qquad \lambda^m = \frac{P^m}{\Sigma_c \theta_c \times S_c^m} \tag{3}$$

which represents a proportional redistribution of the initial discrepancies for a given θ_c.

In this way, since we have pure census tracts for all classes, initial values are available to start the downscaling algorithm. These starting values coincide with densities for pure census tracts per class. Eventually, the final discrepancies are adjusted to the known population figure at the census tract level, and we end up with the densities $d_c^m = \theta_c \times \frac{P^m}{\Sigma_c \theta_c \times S_c^m}$. An identical process can be implemented if we stratify the classes by municipal size. An ANOVA model justifies stratification, showing significant different densities by class and stratum. These methods are comparable to the one introduced by Mennis (2003, 2009), and it is essentially the CLC-iterative method initially tried by Gallego and Peedel (2001), suitably adapted to the structure of our data.

These are sometimes called fixed-ratio methods (Gallego, Batista, Rocha and Mubareka 2011), since a simplifying assumption is that the ratio between population densities of two classes is constant for all census tracts in the same stratum. To escape from this restriction, we can relax (2), and postulate

$$d_c^m = \theta_c^m \times \lambda^m \tag{4}$$

where λ^m plays the same role as before, and θ_c^m is a redistribution threshold by census tract and class, obtained by modeling densities for pure census tracts.

Given that S_c^m are net built-up surfaces, there should be a clear positive relation between densities at census tract level by class and density at commune level, since in this case the heterogeneity in sizes plays no role. Using logs, simple correlations vary from 0.68 for isolated blocks to 0.91 for terraced houses. This suggests estimating a log-log relationship for pure census tracts of the form

$$\log d_c^m = \alpha_c + \beta_c \log d^n + u_c \tag{5}$$

where d^n is commune density in terms of net built-up surface, $d^n = \frac{P^n}{S^n}$. This model is estimated for each class separately, and its projection for non-pure census tracts is used as an estimate of θ_c^m in (4).

However, we have available more information at census tracts level that can be incorporated into regression (5), so it is natural to extend this regression with additional explanatory variables

$$\log d_c^m = \alpha_c + \beta_c \log d^n + \boldsymbol{\gamma}_c' x^m + u_c \qquad (6)$$

The variables included in x^m are the share of people of 65 years old and above, the share of foreigners, and the share of residents in the census tract who were born in the commune to which the census tract belongs. They were significant and eventually included to allow for more generality in the redistribution thresholds.

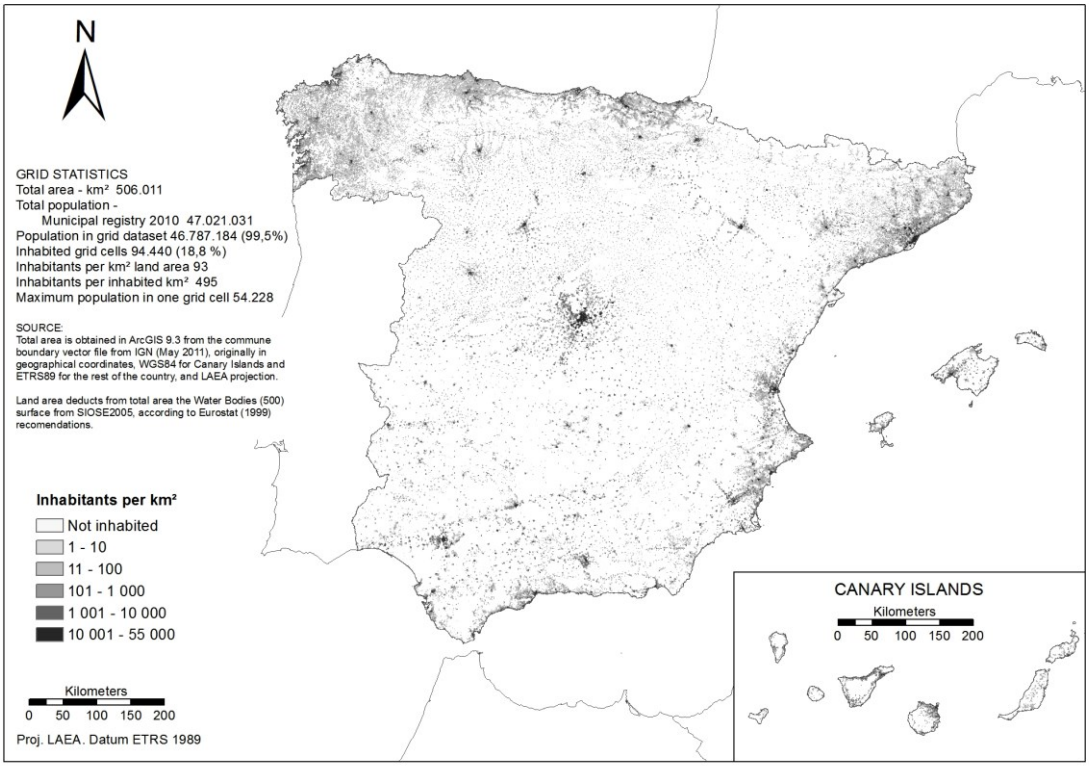
Summing up: (*i*) Estimate (6) for pure census tracts by class; this is, by different type of buildings. (*ii*) Project this model to obtain initial redistribution densities by census tract and class, θ_c^m; these are applied to non-pure census tracts. (*iii*) Discrepancies with respect to the census tract known population value are adjusted with λ^m, (3) (pycnophylactic constraint, Tobler 1979). (*iv*) Finally, once we have estimated population values for all classes and census tracts, we re-estimate the regression equation (6), now using all census tracts, and repeat the redistribution until convergence of the estimated regression coefficients. In this way, we build a vector layer of inhabited polygons within each census tract. In this way we model our own data sets instead of relying on external information, like thresholds from bottom-up grids for other countries. This is an innovative feature of our method, that it is possible given the high resolution of our databases.

After this process has been completed, gridding at any desired resolution is straightforward. Intersect the resulting vector layer of census tract population by class with the grid in vector format and aggregate population by areal weighting. The resulting 1 km^2 grid, with the corresponding statistics, is shown in figure 5.[4] According to our calculations the inhabited surface barely reaches 20% of land area,[5] and whereas the average density for Spain is about 93 inhabitants per km^2, the density per inhabited km^2 is notably higher, 495 inhabitants. Figure 5 offers a more realistic picture of the population distribution than the one obtained by a *choropleth* map at commune level, or even at census tract level.

[4] Even though a finer grid can easily be obtained by the same methods, we use this resolution because it is the standard one in applied work at the European scale (Van Eupen *et al* 2012), and we can compare it with other grids.

[5] Following Eurostat (1999) recommendations, land area deducts water bodies from total area, where these are estimated from SIOSE.

Figure 4. 1km^2 population density grid. 2010.



Source: Municipal Registry: Instituto Nacional de Estadística; SIOSE: Instituto Geográfico Nacional and own elaboration.

3.1. Validation

Our results are validated against a 1 km^2 bottom-up grid for the region of Madrid for 2010, which contains 98.8% of the population in the municipal registry. We use the total absolute error calculated by Gallego (2010)

$$\Delta = \sum_j \left| P_j - P_j^{ref} \right| \tag{7}$$

where j indexes the grid cells, the super-index *ref* refers to the reference grid. The value of Δ varies from 0, perfect coincidence, and twice the population, no coincidence at all. Note that from a spatial point of view this is quite naive, since it does not take into account the distance by which we misallocate people: the contribution to the value of the index is the same if we misallocate a person to the contiguous cell or to the other side of the country. We use (7) for comparability.

Given the dependency of (7) on population size it is natural to scale it to the [0, 1] interval to make it independent of population size. Thus, we shall use the total relative error

$$\delta = \frac{\Delta}{2.\Sigma_j P_j} = \frac{\Sigma_j \left| P_j - P_j^{ref} \right|}{2.\Sigma_j P_j} \tag{8}$$

In this way, 100×δ can be interpreted as the share of population that is incorrectly located, given the resolution of the grid. Table 2 shows the reported total relative error results from Gallego, Batista, Rocha and Mubareka (2011), for different European countries and downscaling methods. Generally speaking, the best method seems to be the limiting variable of Eicher and Brewer (2001), as modified by Gallego, Batista, Rocha and Mubareka (2011). In any case, discrepancies rarely fall below 20%, and the performance of the different methods vary for the different countries. It seems sensible to affirm that there is no single algorithm that performs uniformly better in all cases.

Table 2. Total relative error, 100× δ, against bottom-up grids in several European countries. Grid JRC - Eurostat - EEA

Spatial disaggregation method	Austria	Denmark	Finland	Sweden	Netherlands	Northern Ireland	Estonia
Country population (millions)	*8.03*	*5.35*	*5.18*	*8.88*	*15.99*	*1.69*	*1.31*
Communes (*choropleth*)	55.79	56.82	65.54	70.27	58.54	33.14	58.02
CLC – iterative	28.33	38.04	52.51	45.33	22.30	21.30	35.88
CLC - LUCAS simple	27.33	37.10	48.84	45.55	28.24	21.01	38.93
CLC - LUCAS *logit*	27.09	36.92	48.55	45.44	22.14	18.93	37.40
CLC - EM	28.02	37.20	49.42	45.50	29.05	20.12	39.31
CLC - Limite Variable simple	27.15	22.06	32.63	37.78	19.20	19.23	37.79
CLC - Limite Variable modified	27.21	20.47	30.89	35.70	19.32	17.75	35.50
				##.##	Minimum value for each country		

Source: Gallego, Batista, Rocha and Mubareka (2011).

Table 3 shows 100×δ for our methods and also for *choropleth* maps at commune and census tract level. For the *choropleth* map at commune level we obtain comparable results to the ones reported by Gallego, Batista, Rocha and Mubareka (2011); second line on table 3. In this case, discrepancies are of the order of 70%. For the *choropleth* map at census tract level discrepancies are much lower, of the order of 12%. In this case, no land cover information is used, and results show clearly the benefits from making census tract population data publicly available. Note that because this discrepancy is lower than in all cases reported in table 2, disseminating demographic information for small population units has a high pay-off.

The next row incorporates land cover information using a binary dasymetric method (Langford and Unwin 1994; Langford 2007). In this last case, discrepancies are slightly higher than 5%, which is almost a quarter of the best methods shown in table 2. We can interpret the reduction in the total relative error from the *choropleth* map at census tract level to the dasymetric binary method as the gain associated with the land cover auxiliary information used, for a given resolution. Under this interpretation we can see the importance of a having available a high resolution land cover data set as auxiliary information in downscaling.

Table 3. Total relative error, 100× δ, against a *bottom-up grid in the region of Madrid.*

Population (persons)	Municipal Registry	*Grid population*	
	6,458,684	6,437,936	99.7%
		Georeferenced population	
		6,378,775	98.8%

Method	**Index**
Communes (*choropleth*)	69.02
Census tracts (*choropleth*)	12.03
Dasymetric binary	5.35
Simple density by class and stratum	4.40
log-log density regression: equation (6)	4.41
	Chosen grid

Source: Own elaboration from INE, IEM, IGN and SIOSE2005.

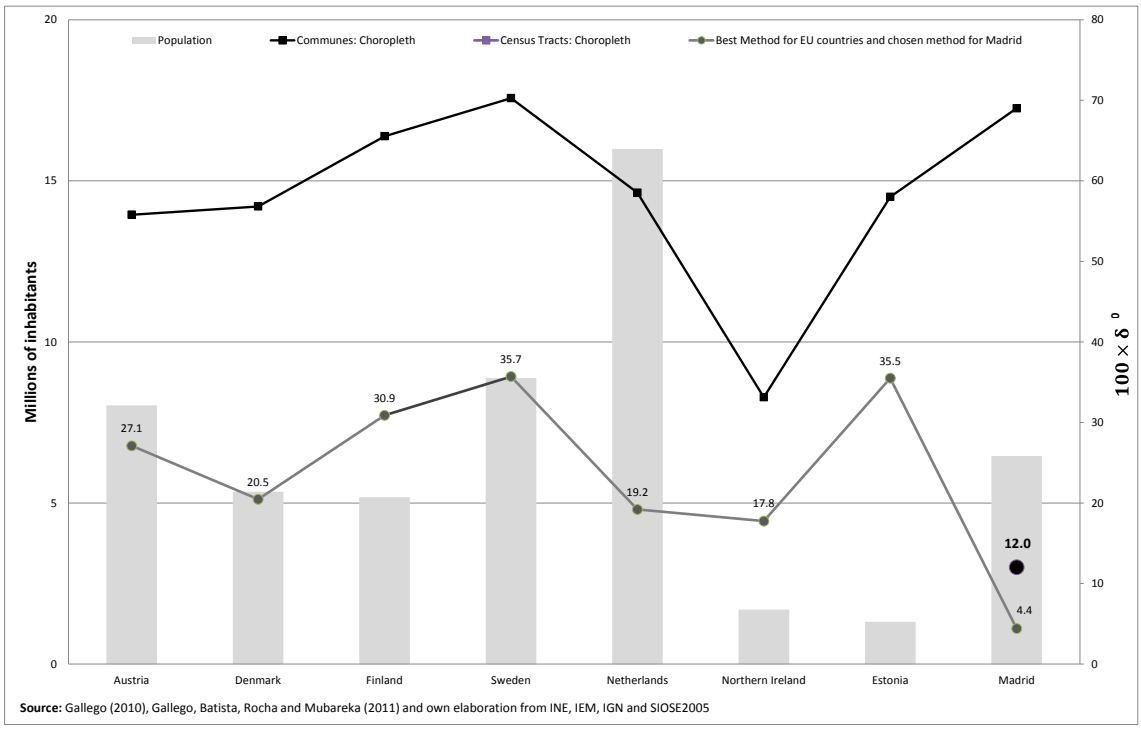
The last two rows show the benefits of disaggregating population using fixed-ratio or regression methods. We can see that all improvements from statistically modeling densities are marginal, since the improvements are very small. In addition, for this particular case, there is no clear gain from using one method over the other. However, given the special characteristics of applying a fixed-ratio method, and that regression methods are simple to apply, we opted finally for the grid obtained from equation (6), which is the one depicted in figure 4.

In addition, it is worth noting that Madrid is probably an atypical region: heavily urbanized, with a high density and population concentration. In fact, because 67% of the population resides in pure census tracts in this region, the redistribution problem affects only 33% of the population. This explains the good behavior of the dasymetric binary method in this particular case, while also limiting the potential gains from statistical modeling. This is another argument for preferring regression methods, since in this case the threshold applied in the downscaling algorithm is different for each class and census tract.

Figure 5 depicts graphically the numerical information from tables 2 and 3 and shows the great improvement from our higher resolution data sets.

The general conclusion from our validation exercise is clear and well known. Quality and resolution of the information is much more important than the choice of the downscaling algorithm (Martin, Tate and Langford 2000). This applies not only to the land cover auxiliary information, but also to the demographic data.

Figure 5. Comparative error against bottom-up grids: Total relative error.



Source: Gallego (2010), Gallego, Batista, Rocha and Mubareka (2011) and own elaboration from INE, IEM, IGN and SIOSE2005.

4. FINAL COMMENTS

A precise location of population is the key to many practical questions of social interest. We have presented a vector population layer built from census tract population data and a high resolution land cover data set with a complex structure, SIOSE. This basic layer can be aggregated into grid form at any desired resolution. A validation exercise, using a 1 km^2 grid, shows that our results are highly accurate, drawing attention to the importance of the resolution of the available data. In fact it seems that, beyond a certain point, statistical modeling is of little help in improving accuracy. This confirms Gallego's (2010) conclusion that, given a data set, the accuracy of the different statistical methods changes only moderately. A recommendation from this study is that efforts should be directed towards disseminating population data at a more detailed scale than municipalities and towards increasing the spatial and thematic resolution of the land cover data set available at European level.

Applications for this data set are huge, especially in the analysis of the relations between population and the environment (Verburg *et al* 2010), some of which are currently under research. Questions like accessibility of public services and infrastructures (Verburg *et al* 2004; Uchida and Nelson 2009), the definition of rural and urban areas (Eurostat 2010, 2012; Van Eupen *et al* 2012), and the inter-relationship between them (Dijkstra and Poelman 2008), fit quite naturally in this context. Clearly, more socio-demographic variables, such as sex, age or nationality, would be very

useful. Incorporating these variables is straightforward with our methods and publicly available information.

As a final conclusion we can say that, until bottom-up grids are widely available from National Statistical Offices, the population vector layer and its associated grid we have presented in this paper is highly accurate, and seems capable of answering good questions involving the real location of people. When new data from SIOSE are released, the error introduced by the different reference dates between the land cover and population datasets could be analyzed, and in addition changes in the population distribution at cell level could also be studied.

Even when, in the future, demographic data are widely available in grid format, our methods can be directed towards dynamic continuous models of population distribution (Martin 2010) or to downscaling economic data (Sachs, Mellinger and Gallup 2001; Nordhaus 2002, 2006, 2008), a type of information for which a top-down approach seems more feasible.

References

Batista e Silva, F.; Lavalle, C. and Koomen, E. (2012). "A procedure to obtain a refined European land use/cover map". *Journal of Land Use Science*, forthcoming: available on line 29 March 2012 [doi:10.1080/1747423X.2012.667450].

Dijkstra, L. and Poelman, H. (2008). Remote rural regions: How proximity to a city influences the performance of rural regions. *Regional Focus, 1/2008*. Brussels: DG-Regio.

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

Eurostat (1999). *Recommendations for a harmonized definition of calculation of surface area of territorial units*. Methods and Nomenclature. Theme 1. Luxemburg: European Commission.

Eurostat (2010). *Eurostat regional yearbook 2010. Eurostat Statistical Books*. Luxemburg: Publication Office of the European Union, Eurostat.

Eurostat (2012) *Eurostat regional yearbook 2012. Eurostat Statistical Books*. Luxemburg: Publication Office of the European Union, Eurostat.

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

Gallego, F. J. and Peedell S. (2001). Using land cover to map population density. In *Towards agri-environmental indicators. Integrating statistical and administrative data with land cover information.* Topic report nº 6. (pp. 94-105). Luxemburg: Eurostat. DG Agriculture. DG Environment. Joint Research Center. European Environment Agency.

Gallego, F. J. (2010). A population density grid of the European Union. *Population & Environment, 31(6), July,* 460-473.
Available at: http://www.springerlink.com/content/0199-0039/31/6/

Gallego, F. J.; Batista e Silva, F.; Rocha, C. and Mubareka, S. (2011) "Disaggregating population density of the European Union with CORINE land cover". *International Journal of Geographical Information Science*, 25, 12, (December), 2051-2069. Available at: http://dx.doi.org/10.1080/13658816.2011.583653

Goerlich, F. J. and Mas, M. (2008). Empirical evidence of population concentration in Spain. *Population-E, 63(4),* 635-650.

Goerlich, F. J. and Mas, M. (2009). Drivers of agglomeration: Geography *versus* History. *The Open Urban Studies Journal (TOUSJ), 2,* 28-42. Available at: http://www.bentham.org/open/tousj/openaccess2.htm

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.

INSPIRE (2010a). *D2.8.I.1 INSPIRE Specification on Coordinate Reference Systems – Guidelines*. Version 3.1. Brussels: INSPIRE Thematic Working Group Coordinate Reference Systems and Geographical Grid Systems (26 April 2010). Available at: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2.

INSPIRE (2010b). *D2.8.I.2 INSPIRE Specification on Geographical Grid Systems – Guidelines*. Version 3.0.1. Brussels: INSPIRE Thematic Working Group Coordinate Reference Systems and Geographical Grid Systems (26 April 2010). Available at: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2.

Kopecky, M. and Kahabka, H. (2009). *2006 GMES Fast track service precursor on land monitoring. Updated delivery report. European Mosaic*. Copenhagen: EEA. Available at: http://www.eea.europa.eu/data-and-maps/data/eea-fast-track-service-precursor-on-land-monitoring-degree-of-soil-sealing-100m-1

Langford, M. (2007). Rapid facilitation of dasymetric–based population interpolation by means of raster pixel maps. *Computers, Environment and Urban Systems*, *31(1), January,* 19-32.

Langford, M. and Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. *Cartographic Journal, 31(1),* 21-26.

Martin, D. (2010). "Progress report: 24-hour gridded population models". Presented at the *European Forum for Geostatistics 2010*, Tallinn (Estonia), 5-7 October 2010. Available at: http://www.efgs.info/workshops/efgs-2010-tallinn-estonia.

Martin, D., Tate, N. J. and Langford, M. (2000). Refining population surface models: Experiments with Northern Ireland Census data. *Transactions in GIS*, *3,* 285-301.

Mennis, J. (2003). "Generating surface models of population using dasymetric mapping". *Professional Geographer*, *55(1),* 31-42.

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

Nordhaus, W. D. (2002). "Alternative approaches to spatial rescaling". Version 2.2.2. Mimeo. New Haven (Connecticut): Yale University (28 February). Available at: http://gecon.yale.edu/research-papers

Nordhaus, W. D. (2006). Geography and macroeconomics: New data and new findings. *Proceedings of the National Academy of Sciences of the United States of America*, *103(10), March,* 3510-3517. Available at: http://gecon.yale.edu/research-papers .

Nordhaus, W. D. (2008). New metrics for environmental economics: Gridded economic data. *The Integrated Assessment Journal, Bridging Sciences & Policy*, *8(1),* 73-84.

Sachs, J. D.; Mellinger, A. D. and Gallup, J. L. (2001) "The geography of poverty and wealth". *Scientific American*, 284, (March), 70-75.

SIOSE (2010). *Land Cover and Use Information System (SIOSE). Technical Document Version 2.0*. Madrid: National Geographic Institute (IGN), 29 January 2010. Available at: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/241/referenceid/32883

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

Uchida, H. and Nelson, A. (2009) "Agglomeration index: Towards a new measure of urban concentration". Background paper for the World Bank's World Development Report 2009: *Reshaping Economic Geography*, pp.-16.
Available at: http://www.worldbank.org/.

van Eupen, M.; Metzger, M.J.; Pérez-Soba, M.; Verburg, P. H.; van Doorn, A.; and Bunce, R. G. H. (2012) "A rural typology for strategic European policies" *Land Use Policy* 29, 3, (July), 473-482.

Verburg, P. H.; Overmars, K. P. and Witte, N. (2004) "Accessibility and land use patterns at the forest fringe in the Northeastern part of the Philippines". *The Geographical Journal*, 170, 3, (September), 238–255.

Verburg, P. H.; van Berkel, D. B.; van Doorn, A. M., van Eupen, M. and van den Heiligenberg, H. A. R. M. (2010) "Trajectories of land use change in Europe: a model-based exploration of rural futures". Landscape Ecology, 25, 2, (February), 217–232.

Villa, G.; Valcarcel, N.; Arozarena, A.; Garcia-Asensio, L.; Caballero, M. E.; Porcuna, A. Domenech, E. and Peces, J. J. (2008). "Land cover classifications: An obsolete paradigm" *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII, Part B4. Beijing.* 609-614.

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

Yuan, Y.; Smith, R. M. and Limp, W. F. (1997). "Remodeling census population with spatial information from LandSat TM imaginery" *Computers, Environment and Urban Systems* 21, 3/4, (May/June), 245-258.