Downscaling population density in the European Union with a land cover map and a point survey

Francisco Javier Gallego
IPSC, JRC, 21020 Ispra (Varese), Italy
e-mail: javier.gallego@jrc.it
Phone: +39.0332.785101
Fax: 39.0332.783033

Summary

Population data in the European Union (EU) are available at commune level. More detailed data are available only for a few countries. CORINE Land Cover (CLC) provides land cover information with a medium resolution. This paper describes several approaches to combine commune population data with CLC to produce a EU-wide population density grid. The results are distributed by the European Environment Agency (EEA). The information provided by the point survey LUCAS (Land Use/Cover Area frame Survey) has been integrated to improve the coefficients of the model. An assessment, carried out for Austria with the help of a more accurate 1-km population density grid, suggests that CLC allows reducing by approximately 50% the inaccuracy of the homogeneous representation of population density per commune. A method based on logit regression gives the best results among the approaches tested, but the accuracy is similar for several approaches.

Keywords : Population density, Dasymetric mapping, downscaling, CORINE Land Cover (CLC), Modifiable Areal Unit Problem (MAUP), Land Use/Cover Area-frame Survey (LUCAS),

1 Introduction

Harmonised population density data for the European Union (EU) are available at the level of the commune. Some countries have more detailed geo-referenced data, but for EU-wide studies the communal level is the most detailed available. Data by commune may be insufficient for planning or modelling purposes. For example such data are not suitable to reply to questions of the type “how many inhabitants live within a distance of 2 km around industrial sites of a given type?”.

There is a need to downscale population density, i.e. to represent it in smaller geographical units, for example cells of 1 km^2 or 1 ha. There is a range of possible approaches for downscaling; Bierkens et al. (2000) mention a variety of downscaling methods for general purpose, several of them based on deterministic and stochastic functions, possibly combined with mechanistic models. Downscaling can be seen as a special case of the Modifiable Areal Interpolation Unit (MAUP), i.e. the transfer of data available for a given set of geographical units into an incompatible set (Openshaw, 1984)

Eicher and Brewer (2001) mention three types of methods to produce density (dasymetric) maps. The binary method (Langford, 1991) assigns the whole population to one land cover class (usually urban or artificial land cover). The three-class method attributes some density to agricultural and forest classes. The limiting variable method first uses simple areal weighting; densities are then modified by putting thresholds to land cover classes and redistributing exceeding population to other classes. In the example analysed in the paper by Eicher and Brewer, the limiting variable method gives the best results. Other methods (Flowerdew and Green, 1989, Yuan et al, 1997, Briggs et al, 2007) use a regression model to obtain the population density of each class that best matches the data. Coefficients are applied later to adjust the total population assigned to each administrative unit (commune) to the known population.

Wu and Murray (2005) use a cokriging method in a small test area in Ohio. This method has some advantages, such as giving a measure of the variance of the estimate in each location, given the underlying model, but presents computational and data availability problems with a large data set, as the one we consider here.

Some authors have produced more precise downscaled population density layers using streets and roads networks in a small area, such as a county (Xie, 1995). A similar approach is adopted by Reibel and Bufalino

(2005) and by Mrozinski and Cromley (1999), who assume that the population is concentrated in a buffer around a road network. This type of methods can be interesting at EU level with the help of navigation databases such as Tele-Atlas or Navteq, but it has not yet been tackled.
Another possible approach to the downscaling problem might be based on the EM algorithm (Dempster, 1977, Ambroise and Govaert, 1998). The EM algorithm has been tested with CLC data and we refer to it as CLC-EM in this paper. An alternative method, that we call here CLC iterative, is applied by Gallego and Peedell (2001). The "CLC-iterative" density map has been assessed by Thieken et al (2006), who conclude that the obtained density map gives realistic population figures for the areas flooded in Germany in the flood events of 1999 and 2002. However this paper finds, by comparing data with the population of the 5-digit postal codes, that the CLC-iterative method generally over-estimates the population in non-urban land cover classes in communes containing also an important urban nucleus.

Langford (2007) fairly claims that complex areal interpolation methods to produce dasymetric population maps is a major obstacle for the use of such methods by many users. The attempt of this paper is to document a ready-to-use raster layer of population density with 1 ha resolution. The user does not need to produce his own interpolation. The grid can be obtained free of charge, for non-commercial purpose, from the European Environmental Agency (EEA). More information on the way to obtain the grid can be found in the EEA data warehouse (http://dataservice.eea.europa.eu/). The methods used for the previous versions of the population raster map are also summarised. The result of each downscaling procedure presented here is a GIS layer in raster format with 100 m resolution. We attribute to each 1 ha pixel an estimated population density. This layer has some conceptual differences with the world-wide products LandScan™ population density grid (Dobson et al., 2000, Bhaduri et al., 2002) and the Gridded population of the world (GPW) of the Center for International Earth Science Information Network (CIESIN, 2005): In our case the area covered is smaller but the spatial resolution is finer. LandScan refers to the "ambient population", a time-weighted average of the number of people in a given area, while our grid locates each person in his/her dwelling (even if it is a circumstantial dwelling for a very short time). The population dataset presented by Bengtsson et al (2006) also includes projections of the spatial distribution of population until 2100.

A different approach to the problem, mainly for application at national or sub-national level, is designing the census enumeration areas optimizing its compatibility with different applications, to minimize the need of MAUP procedures (Martin, 1998).

2 Data

Several layers of information are combined for this exercise: Commune data (population and geographic boundaries), a land cover map, and a fine scale point survey:

2.1 Commune data.

The area covered by the study includes the 27 Member States of the European Union except Cyprus, peripheral islands and overseas territories are excluded. Croatia is included. Altogether an area around 4.3 million km^2 with more than 480 Million inhabitants. For Switzerland and Norway the population data were available, but the land cover information was missing in our database.
Population data from the 2001 census are available for each commune of the study area. The administrative units we refer to as "communes" in this paper correspond to the so-called Eurostat LAU-2 level (Local Administrative Units). Population data and commune boundaries were provided by Eurostat.

The number of communes in the study area is slightly over 114.000. The average area of a commune is about 36 km^2. The average commune area per country ranges from less than 15 km^2 in Slovakia, Czech Republic and France to more than 1500 km^2 in Sweden. Downscaling is more important for large communes. It is therefore meaningful to mention that, although only 6% of the communes have an area above 100 km^2, they represent 49% of the study area and 33% of the population.

2.2 CORINE Land Cover 2000

The land cover map we have used is CORINE Land Cover 2000 (CLC), produced by photo-interpretation of Landsat ETM+ satellite images (panchromatic + multispectral resampled with a resolution of 12.5 m.) with common rules in all the countries of the study area (CEC-EEA, 1993, JRC-EEA, 2005, Perdigão and Annoni, 1997). The nomenclature of CLC has 44 classes. The minimum mapping unit of CLC is 25 ha; smaller patches are included in polygons labeled with the dominant land cover type. If there is no clearly dominant

land cover type in a polygon, it is coded as “heterogeneous”. The classes labeled as “heterogeneous” are important (around 11% of the study area) due to the relatively coarse scale of CLC. A raster version with cells of 1 ha has been used in projection Lambert Azimuthal Equal Area with the parameters recommended for the EU by the INSPIRE initiative (Annoni et al., 2001)

For a large number of communes, CLC2000 does not report any urban area. In most cases this is because the communes do not contain any urban patch larger than 25 ha. This happens for 29% of the communes, that correspond to 16.6% of the total area and 2.9% of the total population. These communes may require a specific consideration when CLC2000 data are combined with population data.

2.3 LUCAS point data

The LUCAS-2001 sample has a two-stage systematic design (Delincé, 2001, Bettio et al, 2002). Primary Sampling units (PSU) are selected with a systematic grid of 18 km without stratification. Each PSU is a cluster of 10 points following a 5x2 rectangular pattern with a 300 m step. To be consistent with the ground work definitions, we can conceive the “point” as a circle of 3 m diameter. LUCAS-2001 only covers the 15 countries that were member states in 2001 (EU15).
LUCAS has a double nomenclature: each point has a land cover code (57 classes) and a land use code (14 classes). For this work the land use code has been used, focusing in particular on the class “residential”. 2245 LUCAS points were residential (2.4% of the total sample). Therefore LUCAS estimates the area with residential use in EU15 to be around 75,000 km2.

3 Areal weighting: an iterative method to estimate land cover coefficients.

We summarise in this section the method used by Gallego and Peedell (2001) to combine population data per commune from the 1991 census with CLC90 (the so-called CORINE Land Cover 1990, although the reference date changes from one country to another). This was the first version of the population grid distributed by the EEA. We considered several predefined categories of communes (strata), and we supposed that the population density can be expressed as:

$$Y_{cm} = U_{ch} W_m \qquad (1)$$

Where Y_{cm} is the density of population for land cover type c in commune m , that belongs to stratum h. The coefficient U_{ch} depends on the land cover class of CLC2000. W_m is a factor that ensures that the total population attributed to pixels in each commune matches the known commune population.
Communes have been stratified in each NUTS2 region applying a simple criterion:

1. Dense communes: population density higher than twice the average density in its region;
2. Less dense: population density lower than twice the average density in its region. Some urban area is reported in CORINE Land Cover;
3. Sparse population: No urban area reported in CORINE Land Cover.

This model is a version of modified areal weighting, and implies several simplifying assumptions:

- The population density is supposed to be the same for all pixels in the same commune and same CLC class. Other possible co-variables, such as altitude or the distance to an urban nucleus are not taken into account.
- the ratio between the population density of two land cover classes is supposed to be constant inside each stratum.

These assumptions are not fully realistic, but allow to improve in a simple way the representation of the population density.
If the coefficients U_c are known, we have

$$X_m = \sum_c S_{cm} Y_{cm} \qquad (2)$$

Where X_m is the population in commune m and S_{cm} is the area of land cover type c in commune m.
We easily get W_m: and Y_{cm}:

$$X_m = \sum_c S_{cm} U_{ch} W_m \qquad \Rightarrow \qquad W_m = \frac{X_m}{\sum_c S_{cm} U_{ch}} \qquad (3)$$

$$Y_{cm} = U_{ch} \frac{X_m}{\sum_c S_{cm} U_{ch}} \qquad (4)$$

This disaggregation had been previously carried out with an initial set of coefficients provided by the EEA for an aggregated nomenclature of CORINE Land Cover (Table 1).

grouped class	Initial coefficient U_c	CORINE Class	Label
1	32	111	Continuous urban fabric
2	25	112	Discontinuous urban fabric
3	1	12, 14	Industrial, commercial, transportation. Green urban, sport (urban infrastructure)
4	3	21, 231	arable land
5	5	22, 241, 242	Permanent crops, Complex cultivation patterns
6	3	243	Agriculture, with natural vegetation
7	1	244, 31, 32	Agro-forestry, Forest, woodland and natural vegetation
8	0	13, 33, 4, 5	Mine, dump and construction sites, sand, rock and burnt areas, glaciers, wetland and water

Table 1: grouped CORINE Land Cover classes and initial coefficients

3.1 Disaggregation of regional data to assess the validity of weighting coefficients.

The best way to assess the disaggregation of the commune populations would be comparing the results with data at infra-commune level, but such data are not available at EU level. One possible way to overcome this limitation is:

1. Consider a set of regions (larger than the communes), and pretend that we only know the data at regional level.
2. Disaggregate regional data with CLC using a given set of coefficients U_{ch}.
3. Consider now the communes and estimate the population of commune m by adding the population attributed in the previous step to all pixels in the commune.
4. Compare with the known population per commune and compute a disagreement indicator.
5. Modify the coefficients U_{ch} to reduce the disagreement and restart step 2 until the results become stable.

X_r is the population in region *r*.
S_{cr} is the area of land cover type *c* in region *r*.
Y_{cr} is the density of population we attribute to land cover type *c* in region *r* .
W_r is an adjustment factor to ensure that the total population in each region coincides with the known total.

Thus, $X_r = \sum_c S_{cr} U_{ch} W_r$ => The densities attributed are $Y_{cr} = U_{ch} \frac{X_r}{\sum_c S_{cr} U_{ch}}$ (7)

and the population attributed to each commune m in region r is

$$X_m^* = \sum_c S_{cm} Y_{cr} \qquad (8)$$

Now we can "remember" that we knew the real population X_m and we can compute the ratio between the attributed population and the known population

$$\psi_m = \frac{X_m^*}{X_m} \qquad (9)$$

and an aggregated difference between attributed and real population at regional or European level

$$\delta_r = \sum_{m \in r} \left| X_m^* - X_m \right| \qquad \delta = \sum_m \left| X_m^* - X_m \right| \qquad (10)$$

It can be easily checked that $\delta_r \leq 2X_r$. The maximum value of the deviation would happen when all the population is attributed to communes with real population 0.

For each region we can compute the correlation $\rho_{cr} = corr\left(\psi_m, \frac{S_{cm}}{S_m}\right)$ (11)

If the correlation $\rho_{cr} > 0$, this would mean that a too high population has been generally attributed to communes where the CORINE Land Cover class c has a relatively high proportion. We can try to compensate this tendency by reducing the coefficient for this region and land cover. A strong correction will be needed if the correlation ρ_{cr} is high and the disagreement δ_r is important compared with the population of the region X_r . We have empirically chosen the next formula to reduce the disagreement:

$$U'_{chr} = U_{ch}\left(1 - k\frac{\rho_{cr} \times \delta_r}{X_r}\right) \qquad (12)$$

Where k is a tuning coefficient: a small value of k makes the correction moderate, but avoids jumping from an over-estimation to an under-estimation. The coefficient U'_{cr} raises when the correlation is negative. The coefficient adjustment can be repeated in an iterative way until the difference indicator δ becomes stable. To avoid some extreme effects on the coefficients, limits have been introduced so that the ratio between the maximum and minimum density in a commune is constrained not to exceed 10,000, except for the CLC classes that are supposed to have no population.

3.2 Application to CLC90 and the 1991 population census.

This procedure was applied with CLC90 to a set of 13 countries (some regions missing) with data of the 1991 census. The total population of the area was 321×10^6 inhabitants. We have considered the regions known in the EU nomenclature as "NUTS2". The acronym NUTS stands for « Nomenclature des Unités Territoriales Statistiques » and the level NUTS2 corresponds to administrative regions with a size that usually contain from 100 to 1000 communes, with an area between 2,000 km2 and 50,000 km2 and a population between 500,000 and 5,000,000, although a number of outliers are larger or smaller. There are 272 NUTS2 regions in EU27.

Table 2: Disaggregation coefficients with three strata of communes.

	Urban dense	Urban discontinuous	Urban infrastructure	Arable	Permanent crops and complex	Pastures	Forest & natural vegetation
Stratum 1	1445.9	619.1	12.3	10.2	15.4	5.1	3.3
2	947.4	622.4	31.0	17.4	30.9	11.3	5.2
3				32.0	69.3	22.8	8.6

The representation with homogeneous density population in each region corresponds to a constant coefficient $U_c = U$. The total disagreement δ in this case was 322×10^6. With the initial coefficients in Table 1, the disagreement becomes $\delta \cong 241\times10^6$. The application of the algorithm described above reduces the disagreement, that becomes stable around $\delta \cong 137\times10^6$ without stratification of communes. With stratification, the disagreement has a further reduction until $\delta \cong 90\times10^6$. The coefficients U_{ch} are valid if they are multiplied by any constant *K* and the coefficient W_m is divided by *K*. The values of U_{ch} given in Table 2 correspond to a choice of *K* such that the median of W_m in each stratum is 1, so that U_{ch} correspond to the median density attributed to each land cover class in each stratum.

4 Reviewing parameter estimates with LUCAS data.

A CLC class, for example "arable land", is not pure for several reasons; the main reason is the scale limitation: patches smaller than 25 ha in an area where arable land is dominant will be included in an arable land polygon. If we "zoom" to a fine scale a certain percentage of the class "arable land" is residential. Overlaying the approx 96,000 points of the LUCAS sample on the CLC map, we get a contingency table

crossing CLC classes with fine scale land cover types (Gallego, 2003). From this contingency table we can estimate the proportion of each CLC class that has residential use. CLC classes were clustered to get a simplified nomenclature in 9 classes (table 3) that seems more suitable than the nomenclature previously used (table 1) on the basis of the proportion of residential use derived from LUCAS.

We make the assumption that, in non-urban areas, the population density is approximately proportional to the rate R_{cm} of residential area:

$$Y_{cm} \propto \xi_{cm} = \left\langle \frac{r_{cm}}{n_{cm}} \right\rangle \approx \frac{r_{kh}}{n_{kh}} \approx U_{ch} W_m \qquad (13)$$

Where n_{cm} is the number of LUCAS points in CLC class *c* , commune *m* , and r_{cm} is the number of them that are residential. However the size of the LUCAS sample is not large enough to compute such estimates in each commune. We can make a global estimation of the proportion of residential land in each CLC class (table 3); the coefficients in the right column where used for a version of the downscaled density map, that we call bellow "CLC-LUCAS simple". They are approximately derived from the % of residential land for the non-urban classes. For the urban classes, the % of residential area is not such a good proxy and we simply modified the coefficients derived in the previous version on the basis of subjective perception of the results in a number of known areas.

	LUCAS points			suggested
Land cover class	residential	total	% resid.	coefficients
Urban dense	60	132	45.4	2000
Urban discontinuous	1085	2609	41.6	500
Other urban	72	748	9.6	150
Artificial non residential	2	241	0.8	0
Agricultural	576	31956	1.8	30
Heterogeneous	272	9492	2.9	50
Forest and agroforestry	142	30826	0.5	8
Natural vegetation	19	13339	0.14	2
Open spaces and water	15	7087	0.21	0
Total	2243	96430	2.3	

Table 3: proportion of residential area in CLC classes using LUCAS

5 Application of logit regression

We can expect that the same non-urban CLC class (e.g. "agricultural") has more dense population in areas with higher average density, i.e. the commune coefficients W_m are higher for communes with higher average density D_m but W_m does not grow linearly with the average population density. Let us consider the next question: If we select a point at random and he have some information about this point (CLC class=LC_i, commune, altitude, etc.), which is the probability that this point has a residential land use? A good answer to this question would helpful improve the population density mapping. Here we only consider the land cover class according CLC and the average population density D_m of the commune m that contains the point. We assume again that, excluding the CLC class "urban dense" ($c \geq 2$), the population density Y_{cm} is approximately proportional to the proportion of the territory with residential use.

$Y_{cm} = \lambda_m \, p_{cm}$ with $p_{cm} = p\left(LUCAS\ LandUse_i = \text{"Residential"}\right) = f\left(LC_i, D_m\right)$ (14)

A usual way of modelling this type of probabilities is the logit regression:

$$logit\left(p_{cm}\right) = \log\left(p_{cm} / \left(1 - p_{cm}\right)\right) = \alpha + \sum_c \beta_c J_c + \gamma \log\left(D_m\right) + \varepsilon_{cm} \qquad (15)$$

Where J_c is an indicator function of CLC class c. (1 if the point is in class c and 0 otherwise).
For non-urban classes, the values of p_{cm} are generally small and the logit function is close to the simple logarithm. The model (15) is similar to a multiplicative model:

$$p_{cm} = \exp\left(\alpha + \beta_c + \varepsilon_{cm}\right) \times D_m^{\gamma} \qquad (16)$$

A value γ=0 would suggest that the population density in a point depends only on the CLC class and not on the average density of the commune; this assumption is implicit in the Poisson model considered by Flowerdew et al (1991) in which the higher population density in certain communes would be explained only by the larger area of more populated land cover types (urban in particular), while the same land cover type in

different communes would have approximately the same density with random variations. A value γ=1 would correspond to a population density in each CLC class that would be proportional to the average density of the commune.

A few attempts to fit a model with the available LUCAS data showed that the behaviour is slightly different for different types of communes (strata). The stratification used for the iterative method reported above has been tested, but a better fit was obtained with a slightly different definition of strata:

- Stratum a: Communes in which the CLC class "urban dense" is present.
- Stratum b: Communes with some CLC artificial area, but no "urban dense".
- Stratum c: Communes without any CLC artificial area.

The explanation can be that the absence of the "urban dense" class in CLC usually indicates that the urban nucleus (or nuclei) of the commune does not reach the CLC size threshold (25 ha), but it still exists: it has been integrated in a "heterogeneous" polygon or in another dominant class (agricultural, forest…). Therefore the probability that a LUCAS point, that appears to be agricultural or forest in CLC, falls in residential areas is increased. Communes without any artificial area reported in CLC are also a separate type.

Geographical positioning of the land cover information in CORINE and of LUCAS points is not perfect. As a consequence, a residential LUCAS point in an urban area, but close to the boundary, can wrongly appear to fall in another class (agricultural, forest, etc). In order to limit the disturbances due to relative mislocation, we have excluded for the logit regression the LUCAS points that are less than 100 m far from CLC polygon boundaries. The classes "artificial no residential" and "natural vegetation" have been removed because, having no or very few residential points, they prevented the logit algorithm from converging.

Table 4 reports the parameters obtained for the logit regression (15). We can make several comments on this table:

- For communes without any artificial area, the residential density for a given CLC class strongly depends on the average population density of the commune (γ=0.67). The dependence is much smaller, but still significant (γ=0.23, γ=0.18) for communes with CLC artificial areas.
- The central columns of table 4 provide information on communes in which the class "urban dense" is absent, but there is some other artificial area. If we take as a reference the residential density in the class "urban discontinuous", for the same type of commune, the residential density is approximately 34 times lower for the CLC class "agricultural", 19 times lower for CLC "heterogeneous", and more than 160 times lower for CLC "forest". For communes with some "urban dense" class these ratios become around 65, 34, and 175. This suggests that the ratios estimated in section 4 might have been underestimated. The ratios for the other classes (artificial, natural vegetation) are based on a small number of residential; points and are consequently weaker.
- These ratios are valid for CLC and cannot be directly extended to other land cover maps, especially if they have a different spatial resolution, although the same methodology can be applied.
- The residential density (proportion of the territory with residential use) is a proxy for population density, but both densities are not exactly proportional: the average residential surface per inhabitant may be larger in agricultural or forest areas than in urban areas, even if we consider "urban discontinuous". It has to be also taken into account that LUCAS points are coded as residential if they fall on secondary houses, where people are generally not censed.

Communes	without Logit param.	CLC artificial $\exp(\beta_c)$	With Logit param.	$\exp(\beta_c)$	With dense Logit param.	$\exp(\beta_c)$
α	-9.2	1.0 e-4	-7.3	6.4 e-4	-7.9	3.5 e-4
γ	0.67		0.23		0.18	
β: Urban disc.			5.9	354.4	6.7	799.5
β: Other artificial			3.8	42.7	4.5	91.9
β: Agricultural	2.8	17.3	2.4	10.5	2.5	12.4
β: Heterogeneous	3.6	36.8	2.9	18.2	3.2	23.8
β: Forest	0.2	1.3	0.8	2.2	1.5	4.6

Table 4: Parameters obtained for the Logit model.

More than 94% of the communes covering 85% of the territory do not have any pixel of the CLC class "urban dense" $\left(S_{urbandense,m}=0\right)$. For these communes the density per class is computed as

$p_{cm}=\lambda_m \exp(\alpha+\beta_c)\times D_m^{\gamma}$ with the dasymetric constraint.

$$X_m = \sum_{c=2}^{8} S_{cm} \lambda_m \exp(\alpha + \beta_c) D_m^{\gamma} \qquad \Rightarrow \qquad \lambda_m = \frac{X_m}{\sum_{c=2}^{8} S_{cm} \exp(\alpha + \beta_c) D_m^{\gamma}} \qquad (17)$$

For the communes in which the CLC class "urban dense" is present we need a different strategy, since the approach described above does not attribute any specific density to the CLC urban dense class. We first attribute a population density Y'_{cm} for the other land cover classes (c≥2, i.e. excluding urban dense) by simply averaging the densities $Y_{cm'}$ for the same land cover class c in neighbouring communes without urban dense class. The remaining population of the commune $X'_{1m} = X_m - \sum_{c>1} Y'_{cm} S_{cm}$ is attributed to the class "urban dense with a density $Y'_{1m} = X'_{1m} / S_{1m}$. This provisional computation attributes in some cases an unrealistic value for the density Y'_{1m} compared with the density Y'_{2m} attributed to the class c=2 "urban discontinuous" (notice that the procedure computes a value of Y'_{2m} even if the class c=2 does not exist in the commune m). We have introduced the empirical rule that the density Y'_{1m} should be between 4 and 10 times higher Y'_{2m}.

5.1 Some anomalies and correcting actions

Some anomalies appear in the results obtained with the methods reported above, in particular the density attributed to non-urban classes is too high for a certain amount of communes. We have considered anomalous an attributed density beyond a threshold (table 5) for each CLC class. Thresholds have been selected on the basis of the results of the disaggregation obtained with the iterative algorithm of section 2 and the observation of communes with estimated Y_{cm} values laying in the queues of the distribution. 9471 communes have at least one Y_{cm} value beyond these thresholds and we consider the result anomalous. In stratum C we find 2392 anomalous communes. The likely reason in most cases is that urban areas are too small to be reported in CLC, and the population living in the non-reported urban areas have to be represented in the non-urban classes as reported by CLC. Stratum B contains 6948 anomalous communes and stratum A 131 anomalous communes. In strata A and B the most frequent explanation seems to be that dense urban areas have been photo-interpreted in CLC as discontinuous and the density in these areas has been underestimated with a consequent overestimation of the density in non-urban areas. Some corrections have been applied by reallocating population to a different class in order to stay within the threshold as far as possible. In stratum C densities for agricultural, forest and natural vegetation classes have been brought within thresholds, by allowing a density up to 500 inh/km2 in the "heterogeneous" class. After reallocations the number of anomalies is reduced to 3345: 47 in stratum A, 1657 in stratum B and 1641 in stratum C, interpreted as a consequence of the limitations of CLC for which densities had to be above the thresholds of table 5. The introduction of these thresholds makes the method closer to the traditional "limiting variable" method (Eicher and Brewer, 2001).

Land cover class	Maximum density
Urban dense	100,000
Urban discontinuous	20,000
Other urban (green?)	2,000
Artificial no residential	1000
Agricultural	100
Heterogeneous	300
Forest and agroforestry	30
Natural vegetation	10

Table 5: thresholds applied for the density in different CLC classes.

6 Application of the EM algorithm

Flowerdew et al (1991) propose to apply the EM algorithm to estimate disaggregation coefficients. The EM method (Dempster, 1977) assumes an underlying probabilistic model. Here we follow the suggestion of Flowerdew et al. assuming that the population X_{mc} in land cover class c for the commune m follows a Poisson distribution with parameter $\mu_{mc} = U_c S_{cm}$; all distributions for different communes and land cover classes are assumed to be independent. Each iteration of this algorithm has two steps: the E step (expectation) and the

M step (maximum likelihood). The E step gives an estimate of X_{mc} from the disaggregation coefficients obtained in the previous M step:

$$\hat{X}_{mc}^{(t)} = \frac{U_c^{(t-1)} S_{cm} X_m}{\sum_c U_c^{(t-1)} S_{cm}} \qquad (18)$$

This step actually ensures that the total population attributed to pixels in the commune equals the known population of the commune.

In the M step we estimate the values of U_c that give maximum likelihood to the $\hat{X}_{mc}^{(t)}$ calculated in the previous E step. We have

$$p\left(X_{mc} = \hat{X}_{mc}^{(t)}\right) = \frac{e^{-\mu_{mc}} \mu_{mc}^{\hat{X}_{mc}^{(t)}}}{\left(\hat{X}_{mc}^{(t)}\right)!} \qquad (19)$$

And the likelihood is

$$L(U) = \prod_{cm} \frac{e^{-\mu_{mc}} \mu_{mc}^{\hat{X}_{mc}^{(t)}}}{\left(\hat{X}_{mc}^{(t)}\right)!} \qquad (20)$$

For the purpose of maximization we can use the log-likelihood and disregard the denominator, since it is constant, although each term of the product is not. The maximum is obtained for

$$U_c^{(t)} = \frac{\sum_m \hat{X}_m^{(t)} \log\left(S_{cm}\right)}{\sum_m S_{cm}} \qquad (21)$$

This procedure has been applied with the CLC classes grouped as in the previous section and with the same strata. Table 6 reports the coefficients obtained.

Table 6: Disaggregation coefficients with an EM algorithm and 3 strata.

	Communes		
	Without artificial	Without urban dense	With urban dense
Urban dense			470.1
Urban discontinuous		106.6	305.6
Infrastructure.		97.5	217.2
Unpopulated urban		0.11	0.00
Agricultural	1.74	2.53	0.84
Heterogeneous	3.06	3.99	0.0007
Forest	0.77	0.073	0.00
Natural vegetation	0.15	0.00	0.01
Bare land, water	0.22	0.00	0.00

7 Validation in Austria

The performance of the disaggregation procedures presented above have been compared with the help of reference data provided by the Austrian Statistical Institute. The Austrian reference data were presented as a 1 km resolution grid in UTM (zone 33) coordinates. They have been obtained by aggregation of individual dwellings on the basis of the 2005 population census.

We have compared the reference data with the next dasymetric maps with 1 ha resolution:

- Communes: the average population density of each commune is attributed to the whole commune in a uniform way.
- CLC-Iterative: disaggregation with the method presented in paragraph 3.
- Simple CLC-LUCAS: disaggregation with the method presented in paragraph 4.
- CLC-LUCAS logit: disaggregation with the method presented in paragraph 5.
- CLC-EM: disaggregation with the coefficients obtained with the EM algorithm as reported in paragraph 6.

These 5 maps were first produced as raster grids in the INSPIRE-recommended Lambert-Azimuthal projection with 1 ha resolution, then re-projected into UTM to make them compatible with the Austrian data and generalized to 1 km^2 with the same cell boundaries of the reference grid.

The disagreement indicator was computed as:

$$\Delta_m = \sum_j \left| Y_{j,m} - Y_{j,ref} \right| \qquad (22)$$

The values obtained for the disagreement are reported in table 7. This table indicates that disaggregation of commune-level population density with the help of CLC roughly reduces by 50% the disagreement with reference data. The level of improvement is likely to change if we consider a different country and a different resolution of the reference data, but this comparison gives a valuable indication. The improvement changes very little with the disaggregation procedure. The introduction of LUCAS data to tune the behaviour of CLC classes improves the results, but only slightly. The introduction of the logit model provides a further tiny improvement.

Table 7: disagreement of different dasymetric maps with reference data in Austria

Dasymetric map	*disagreement*
Communes (non disaggregated)	$8.96 * 10^6$
CLC-iterative	$4.55 * 10^6$
CLC-LUCAS simple	$4.39 * 10^6$
CLC-LUCAS logit	$4.35 * 10^6$
CLC EM	$4.50 * 10^6$

8 Detecting geographic patterns of scattered population.

In some European areas we can see numerous scattered houses outside the urban nuclei. In other areas the population is concentrated in large or small nuclei with very few houses in between. The density of LUCAS residential points in non-artificial CLC classes can help us to map these different landscape types, although the sampling density is relatively low.
Figure 1 represents the LUCAS points of scattered residential land, defined as land with residential use in CLC-non-artificial areas. The shaded background represents the density obtained by smoothing the proportion with a moving window of 200 km. Such smoothing is imposed by the low density of the sample, but may introduce some distortion, eliminating in particular landscape types that cover an area smaller than the window width.

The residuals of the logit model can be smoothed to produce a slightly different mapping of the phenomenon: a light-coloured area in Figure 2 indicates a low density of scattered houses without taking into account the population density of the region. A light-coloured area in Figure 2 represents a density of scattered housing lower than what would be expected for the population density of the region. For example the density of scattered housing is low in northern Scandinavia in absolute terms, but it is relatively high taking into account the low population density of this region. In both maps the meaning of “scattered” relates to the CLC specifications, i.e. it includes small agglomerations of less than 25 ha.

9 Discussion

This paper illustrates the method used to produce the population density grid of EU27 (plus Croatia) with a resolution of 1 ha that is currently distributed by the EEA without charge for non-commercial purposes. Population data of the 2001 census have been merged with CORINE Land Cover to produce a finer scale representation. The point survey LUCAS was used to tune in different ways the coefficients for the downscaling models. LUCAS-2001 data are only available for EU15. This means that the likely ratio between population density in different CLC classes in EU15 has been applied to the rest of EU27. This choice is debatable, but was imposed by the limitations of data availability.

The validation by comparison with a reference population density grid available for Austria, with 1 km resolution, has shown that the inaccuracy of the homogeneous density representation per commune is reduced by roughly 50% thanks to the introduction of CLC. The validation results suggest however that the inaccuracy does not change much with the different methods tested. This confirms the observation made by Martin et al (2000): the quality of the land cover map is more important than the choice of the downscaling algorithm.

Figure 1: Density of scattered residential land use derived from LUCAS and CLC

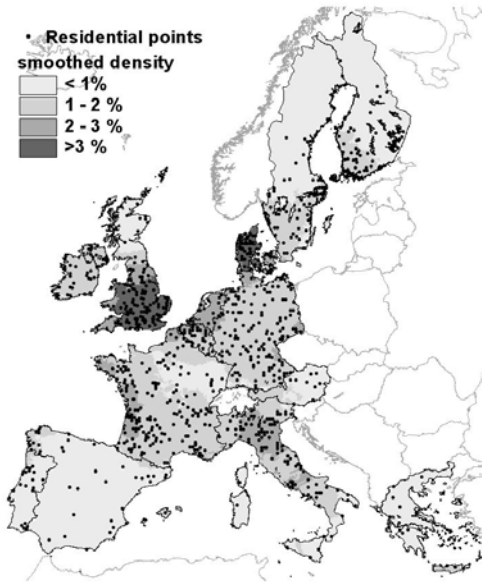
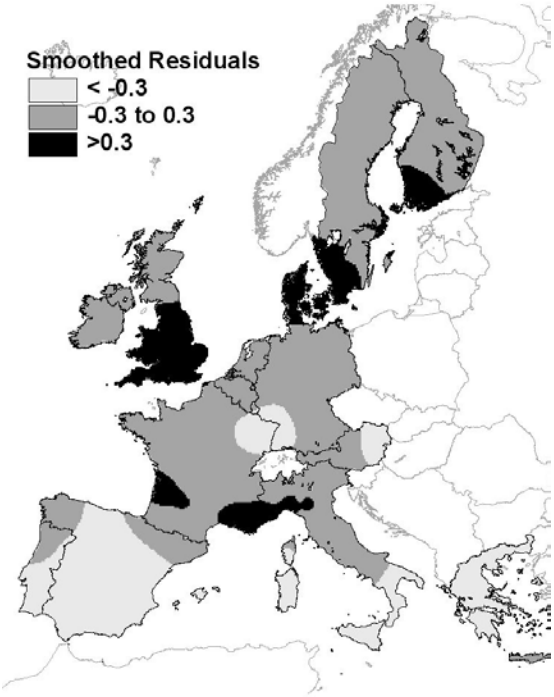


Figure 2: Smoothed residuals of the logit regression for the non-urban residential area



The two-class method (Langford, 2007), that attributes all the population to the urban class is not directly applicable because more than 30,000 communes do not have any urban CLC class, but an adaptation would be possible for further comparison. The geographical analysis of the residuals of the logit model (section 8) can give additional hints to improve the dasymetric maps.

Further analysis might be needed on the comparison with reference data to understand the limitations of the downscaling procedure, but the first assessment indicates that the population density can be very different for different grid cells in the same commune and the same CLC class. The way to improve the results might be the introduction of new information of layers, such as night-time light emissions, as made by Briggs et al (2007), although the coarse resolution of the available products of night-time light is probably a serious limitation.

The results reported in this paper relate to CLC. For example when we talk of population in agricultural land, this refers to the area reported as agriculture by CLC. Part of the inaccuracy that cannot be removed is due to the limitations of CLC.

ACKNOWLEDGEMENTS:
Eurostat provided most of the data for the study. We are particularly grateful to César de Diego, Albrecht Wirthmann, Daniel Rase, Torbiorn Carlquist and Edwin Schaaf. Statistics Austria provided reference data for validation through the "European Population Grid Club" specially Kaminger, Rina Tammisto and Lars Backer. Roger Milego, Oscar Gómez, Stefan Kleeshule and Tomas Soukup, from the European Topic Centre of Terrestrial Environment (ETC/TE) made useful suggestion. Mette Lund managed the distribution of the results through the EEA dataservice.

REFERENCES

Ambroise, C., Govaert, G. (1998). Convergence of an EM-type algorithm for spatial clustering, Pattern Recognition Letters 19 (10), 919-927

Annoni A., Luzet C., Gubler E., Ihde J. (2001) Map Projections for Europe. Report EUR 20120 EN, JRC-Ispra (Italy), 131 pp.

Bhaduri, B., Bright E., Coleman, Ph., Dobson, J., 2002. LandScan: Locating People is What Matters. *Geoinformatics* 5 (2), 34-37. http://www.ornl.gov/sci/landscan/.
Bengtsson M., Shen Y., Oki T., (2006), A SRES-based gridded population dataset for 1990-2100, Population and Environment, 28(2) 113-131.
Bettio M., Delincé J., Bruyas P., Croi W., Eiden G. (2002) Area frame surveys: aim, principals and operational surveys. Building Agri-environmental indicators, focussing on the European Area frame Survey LUCAS. EC report EUR 20521, pp. 12-27. http://agrienv.jrc.it/publications/ECpubs/agri-ind/
Bierkens M.F.P. Finke P.A., de Willigen P. (2000) Upscaling and downscaling methods for environmental research. Kluwer, Dordrecht, 190 pp.
Briggs D.J., Gulliver J., Fecht D., 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.
CEC-EEA (1993). CORINE Land Cover; technical guide. Report EUR 12585EN. Office for Publications of the European Communities. Luxembourg,. 144 pp. http://dataservice.eea.eu.int/dataservice/
Center for International Earth Science Information Network (CIESIN), Columbia University, 2005. Gridded Population of the World (GPW), Version 3. CIESIN, Columbia University, Palisades, NY. Available at http://sedac.ciesin.org/plue/gpw.
Delincé J. (2001). A European approach to area frame survey. Proceedings of the Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR), June 5-7, Vol. 2 pp. 463-472 available at http://www.ec-gis.org:/
Dempster A.P., Laird N.M., Rubin D.B., 1977, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B, 39 1-38.
Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000, LandScan: A global population database for estimating populations at risk *Photogrammetric Engineering and Remote Sensing*, 66 (7), pp. 849-857.
Eicher C., and Brewer, C. (2001). Dasymetric mapping and areal interpolation: implementation and evaluation, Cartography and Geographic Information Science **28** 125–138
Flowerdew R., Green M., Kehris E. (1991) Using areal interpolation methods in GIS; Papers in regional science, 70 (3) 303-315.
JRC-EEA (2005) Image2000 and CLC2000; Products and methods; Report EUR 21757 EN. JRC-Ispra.
Langford M. (2007) Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. Computers, environment and urban systems, 31 19-32.
Martin D. (1998) Optimizing census geography: the separation of collection and output geographies, International Journal of Geographical Information Science, 12 (7), 673-685.
Martin D., Tate, N.J., and Langford M. (2000). Refining population surface models: experiments with Northern Ireland Census data. Transactions in GIS 4(4), 343-360.
Mrozinski, R., Cromley R., 1999, Singly - and doubly - constrained methods of areal interpolation for vector-based GIS . Transactions in geographical Information systems, 3, 285-301.
Openshaw S., 1984, The Modifiable Areal Unit Problem, CATMOG nº 38, Norwich: Geo books.
Perdigão V., Annoni A., (1997). Technical and methodological guide for updating CORINE Land Cover data base. Report EUR 17288 EN. Office for Publications of the European Communities. Luxembourg,. 124 pp
Reibel M., Bufalino M. (2005). Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems, Environment and Planning A, 37(1) 127-139.
Thieken A., Müller M., Kleist L., Seifert I., Borst D., and Werner U. (2006), Regionalisation of asset values for risk analyses. Natural Hazards and Earth System Sciences, 6 167-178.
Wu Ch., Murray A. T. (2005), A cokriging method for estimating population density in urban areas, Computers, Environment and Urban Systems, 29(5) 558-579.
Xie, Y. (1995). The overlaid network algorithms for areal interpolation problem. Computers, Environment and Urban Systems, 19(4) 287-306