STATISTICAL COMMISSION and Working Paper No. 21
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Tallinn, Estonia, 25-28 September 2001)

Topic (iii): Spatial analysis

USING LAND COVER INFORMATION TO MAP POPULATION DENSITY

Submitted by ISTAT, Italy [1]

Invited paper

I. INTRODUCTION

1. Population density data are available to the European Commission (EC) at the level of the commune. The size of communes is heterogeneous across the EU, thus this level of spatial resolution may be insufficient in many cases for planning or modelling purposes or to assess the impact of EU policies. Data per commune cannot provide an appropriate answer to questions such as "How many people live at a distance of less than 200 m from the trajectory of a projected highway?". In general, national Administrations have more detailed population data, for example at census section level, but in most cases the coordinates of census sections are not suitably organised as a GIS layer.

2. CORINE Land cover (CLC) gives useful geo-referenced information for disaggregation. This geographic database provides information that is spatially more detailed than the commune limits. A certain commune may contain, for example, dense urban nucleus, agricultural land with some sparse population, and natural vegetation areas with very little or no population. Population data can be disaggregated, imputing different densities to land cover categories.

3. The problem of improved mapping of population density is not a recent one: it was tackled in early papers dealing with modern geography (Wright, 1936). Some papers focus on giving a smooth representation of the population density (Tobler, 1979, Slater, 1993). Our target is rather to provide a more detailed representation to enhance potential analysis. One possible approach to tackle this problem might be based on the EM algorithm (Dempster, 1977, Ambroise and Govaert, 1998), although it is not certain whether the underlying parametric model is suitable. Here we test a more empirical method.

4. An additional target of the study is to test population data methods that might be used for other problems, such as mapping the use of fertilizers or the Nitrogen surplus. Population density has the advantage of easier testing, at least in areas where georeferenced detailed information is available. This can provide an idea of the suitability of methods for other problems.

II. DATA: COMMUNES AND CORINE LAND COVER

5. The latest population data available to the Commission at commune level correspond to 1991, therefore the commune boundaries of the SABE database, version 1991, have been used (http://www.megrin.org/SABE/Sabe.html). The CLC data have been used as raster data (1 ha pixel).

---

[1] Prepared by Javier Gallego.

6. For the computations hereafter, the average population density of the commune is attributed to areas with missing CORINE Land Cover data. CLC was not available for Sweden and Finland, and the SABE commune boundaries were not available for Scotland. These areas have been excluded from the study (the column UK in the tables corresponds to England, Wales and Northern Ireland). The French département "Seine St Denis" has been also excluded because of code errors in the version of CLC used. The set used for the analysis included 87166 communes.

7. In some countries, such as France, where most communes comprise a small area (approx. 15 km2 in the average), the resolution may be sufficient for some types of analysis, but it is clearly insufficient in other countries with larger communes. In Denmark the average size of a commune is more than 150 km2. The number of communes in the study area larger than 100 km2 is relatively small (3221 communes, i.e. 3.7% of the total), but they represent more than 26% of the area. 78 communes are larger than 500 km2.

III. COMMUNES WITHOUT URBAN AREA IN CORINE LAND COVER

8. A large number of communes appear as not having any urban area according to CORINE Land Cover. There are several possible explanations:

- The built area is below the polygon size threshold for CORINE Land Cover (25 ha). This is the explanation in the great majority of cases.
- Geometric inconsistencies between CORINE Land Cover and SABE. For example, all the urban area of a commune may be in a thin linear development along the coast, but appears to fall outside the commune because of location inaccuracy. This can happen in theory but no case has been identified in practice.
- Errors in the data (CORINE Land Cover or SABE). In a few cases, crossing population data with commune borders can help focus attention on inconsistencies between land cover and population density. Some of them might turn out to be errors in data sets. Anomalies have been detected for several communes of Seine St Denis on the outskirts of Paris. Although this area has been excluded, similar errors may still remain in the data.

9. The 12.1 million inhabitants in this category of communes represent less than 4% of the 327 million inhabitants of the area under study, but it is a significant part of the rural population; so this case must be considered specifically. The most populated communes without CLC urban classes correspond to areas with very scattered settlements and no major error could be detected in CLC. Nearly 27,000 communes had no urban area reported by CLC; 51 of them had more than 5000 inhabitants (Table 1).

	<1000 inhab		1000-5000 inhab		>5000 inhab		Total	
	N communes	population (*1000)	N communes	population (*1000)	N communes	population (*1000)	N communes	population (*1000)
AT	488	301	772	1353	5	40	1265	1695
DE	1860	656	185	268	0	0	2045	924
ES	1269	247	151	333	22	157	1442	737
FR	12860	2699	80	108	2	11	12942	2819
GR	2841	865	110	164	5	77	2956	1107
IE	1973	767	109	143	0	0	2082	910
IT	886	410	357	617	6	48	1249	1074
LU	12	6	1	1	0	0	13	7
NL	2	1	2	2	0	0	4	4
PT	1585	742	624	1107	10	63	2219	1912
UK	106	79	509	878	1	5	616	961
total	23882	6773	2900	4975	51	403	26833	12151

Table 1: Communes with no urban area in CORINE Land Cover

IV. AREAL WEIGHTING WITH GIVEN COEFFICIENTS

10. Population data can be disaggregated assuming that the ratio between the population density of two given land cover classes is the same for any commune. This is a simplified version of modified areal weighting. We can initially assume that the coefficients are known:
Xm : population in commune m.
Scm : area of land cover type c in commune m.
Ycm : density of population for land cover type c in commune m. Inside each commune Ycm is assumed to be proportional to given coefficients Uc for each land cover type.
Wm : adjustment factor to ensure that the total population in each commune matches the administrative data. .

$$Y_{cm} = U_c W_m \qquad X_m = \sum_c S_{cm} Y_{cm} \qquad (1)$$

and it was assumed that $X_m = \sum_c S_{cm} U_c W_m \quad \Rightarrow \quad W_m = \frac{X_m}{\sum_c S_{cm} U_c}$ (2)

hence the densities are computed in this case as $Y_{cm} = U_c \frac{X_m}{\sum_c S_{cm} U_c}$ (3)

This approach has been applied with an initial set of coefficients provided by the European Environment Agency (EEA) for an aggregated CORINE Land Cover nomenclature (Table 2).

grouped class	Initial coefficient U_c	CORINE Class	Label
1	32	111	Continuous urban fabric
2	25	112	Discontinuous urban fabric
3	1	121	Industrial or commercial units
4	1	122, 123, 124	Road and rail networks, ports, airports
5	1	141, 142	Green urban areas, Sport and leisure facilities
6	3	211	Non-irrigated arable land
7	3	212	Permanently irrigated land
8	1	213	Rice fields
9	5	22	Permanent crops
10	3	231	Pastures
11	5	241	Annual and permanent crops associated
12	5	242	Complex cultivation patterns
13	3	243	Agriculture, with natural vegetation
14	1	244	Agro-forestry areas
15	1	31, 324	Forest and woodland
16	1	321, 322, 323	Other natural vegetation
17	0	13, 33, 4, 5	Mine, dump and construction sites, sand, rock and burnt areas, glaciers, wetland and water

Table 2: grouped CORINE Land Cover classes and initial coefficients

11. A simple regression model model

$$X_m = \sum_c S_{cm} U_c W + e_m \qquad (4)$$

gives completely unacceptable results with several negative coefficients and very high values for classes such as green urban or sport areas. Separating urban and rural areas is not sufficient to render coefficients homogeneous. Other attempts to define subsets of homogeneous sets of communes through cluster analysis have also failed to yield reasonable coefficients through linear regression.

V. ESTIMATION OF COEFFICIENTS BY DISAGGREGATING REGIONAL DATA

12. The proper way to assess the disaggregation of population by commune would be to compare the results with data at infra-commune level, but such data are currently not available to the EC except for those from some small test sites. Since global assessment requires knowledge of data at two levels of aggregation, we can follow the scheme below:

- Aggregate data for NUTS2 regions. Pretend for a moment we do not have communal data;
- Disaggregate regional data with CLC using a given set of coefficients;
- Reaggregate the attributed population to the territory of each commune;
- Compare the attributed with the known population per commune and compute a disagreement indicator;
- Modify the coefficients to reduce the disagreement.

13. We assume $X_r = \sum_c S_{cr} U_c W_r$, with

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

therefore we attribute densities $Y_{cr} = U_c \dfrac{X_r}{\sum_c S_{cr} U_c}$ (5)

and the population attributed to commune m in region r is $X_m^* = \sum_c S_{cm} Y_{cr}$ (6)

14. The differences between the attributed and the real population of the commune can be aggregated at regional or European level

$$d_r = \sum_{m \in r} \left| X_m^* - X_m \right| \qquad\qquad d = \sum_m \left| X_m^* - X_m \right| \qquad (7)$$

This aggregated difference is an indicator of the quality of the disaggregation from regional data. The maximum possible value of this indicator is twice the population: $d_r \le 2X_r$.

15. If the ratio $y_m = \dfrac{X_m^*}{X_m}$ between the attributed population and the known population of the commune is generally high for communes with a high proportion of land cover c, i.e. if the correlation $r_{cr} > 0$, this probably means that the population attributed to this class was too high. We can compensate this tendency by reducing the coefficient

$$U'_{cr} = U_c \left(1 - \frac{r_{cr} \times d_r}{2 \times X_r} \right) \qquad \text{where} \qquad r_{cr} = corr\left(y_m, \frac{S_{cm}}{S_m} \right).$$

The term $\dfrac{d_r}{2 \times X_r}$ is introduced to moderate the modification of the coefficient when the attribution is close to the real population.

16. The coefficient adjustment can be iterated until the disagreement indicator **d** becomes stable. To avoid some extreme effects on the coefficients, thresholds have been introduced for the ratio between the maximum and minimum density in a commune. The total deviation **d** with the same coefficient to all CLC classes gives an indicator of 322×106 for a total population of 321×106 inhabitants. This

corresponds to the inaccuracy of representing the average population density by NUTS2 compared with the representation by communes. With the coefficients in Table 2, the disagreement indicator goes down to around 241×106 . The application of the algorithm described above improves the deviation, so that results become stable under 137×106.

17. Looking at the average values of the coefficients (Table 3) by region, we observe:

- The three arable land classes have similar average coefficients;
- The coefficients for permanent crops are close to the coefficients of two of the complex classes (2.4);
- Agro-forestry presents similar coefficients to those of forest and semi-natural vegetation.

This suggests that some groups of land cover classes might be regrouped.

18. Mapping the disagreement per commune between the result of disaggregating the NUTS2 population and the communal data shows a pattern closely linked with urban agglomerations: population has been underestimated in urban areas and overestimated in rural areas. This could be interpreted as an inadequate ratio between land-cover coefficients, but is due rather to the fact that the underlying assumption of equal density for the same land cover class in each NUTS2 region is not acceptable: urban areas in Rome are more densely populated than the urban nucleus in a rural town in Lazio. This is also the main reason why the global disagreement cannot be brought below certain values.

19. The assumption of homogeneous population density for each land cover type in each administrative unit is much weaker when applying the disaggregation from the commune level rather than from the NUTS2 level, but is still not completely adapted to reality. This must be seen as a scale limitation of the approach.

Land Use	
Urban dense	198.41
Urban discontinuous	176.56
Industrial and commercial	9.56
Transport	3.71
Green urban	3.42
Arable non irrigated	2.98
Arable irrigated	3.36
Rice	2.90
Permanent crops	4.95
Pastures	2.99
Arable with permanent crops	5.40
Complex agricultural	5.92
Agricultural and natural	3.24
Agroforestry	0.86
Forest	0.94
Natural vegetation	0.63

Table 3: Average coefficients after 40 iterations

VI. STRATIFICATION AND FURTHER GROUPING OF CORINE LAND COVER CLASSES

20. The ratio between densities in different land cover classes is not the same in densely populated areas and in more rural areas. Therefore, communes have been stratified in each region by applying a very simple criterion:

- Stratum 1: Dense communes: population density higher than twice the average in its NUTS2 region;

- Stratum 2: Less dense: population density lower than twice the average density in its NUTS2 region, but urban area reported in CORINE Land Cover;
- Stratum 3: Communes without urban area reported in CORINE Land Cover.

21. A cluster analysis of the 16 CLC classes based on the table of coefficients by NUTS2 gives an indication of the CORINE Land Cover classes that have a similar behaviour for different region typologies. Taking into account the results of cluster analysis and the meaning of classes, the classes have been regrouped into 8 (Table 4) and the iterative algorithm has been rerun. The classes "Industrial and commercial", "transport" and "green urban and sport facilities" have been aggregated with the class "urban discontinuous" with fixed weights inspired in Table 3:
Urban 2 = Urban discontinuous + 0.05 * Industrial_commercial + 0.02*(transport+green urban)
Therefore, 6 classes have been kept for further analysis. Applying the iteration algorithm with 3 strata and 6 CORINE Land Cover classes, the deviation indicator **d** approaches 90×106 .

grouped class	CORINE Class	Label
1	111	Continuous urban fabric
Urbdisc (2a)	112	Discontinuous urban fabric
ind (2b)	121	Industrial or commercial units
otha (2c)	122	Road and rail networks and associated land
otha (2c)	123	Port areas
otha (2c)	124	Airports
otha (2c)	141	Green urban areas
otha (2c)	142	Sport and leisure facilities
3	211	Non-irrigated arable land
3	212	Permanently irrigated land
3	213	Rice fields
4	221	Vineyards
4	222	Fruit trees and berry plantations
4	223	Olive growes
5	231	Pastures
4	241	Annual and permanent crops associated
4	242	Complex cultivation patterns
5	243	Agriculture, with natural vegetation
6	244	Agro-forestry areas
6	311	Broad-leaved forest
6	312	Coniferous forest
6	313	Mixed forest
6	321	Natural grassland
6	322	Moors and heathland
6	323	Sclerophyllous vegetation
6	324	Transitional woodland-shrub

Table 4 New grouped classes

22. The link between the disagreement $d_m = \left(X_m^* - X_m\right)$ per commune and urban agglomerations can still be observed in spite of the stratification. This may indicate that the stratification is not sufficient. However, too many strata would lead to unstable coefficients. Improving the stratification criteria remains an open question. This might be achieved through a more sophisticated indicator of "rurality", possibly taking into account the spatial structure (distance to urban nuclei, etc.).

VII. SUGGESTED DISAGGREGATION

23. The algorithm gives weighting coefficients for each stratum in each NUTS2 region. For most regions, the coefficients are similar, but some outliers appear. Clustering NUTS2 regions does not show a clear grouping linked with population settlement styles. Therefore, a disaggregation rule has been proposed based on a set of coefficients for each stratum that are the same for all NUTS2 regions.

24. The population density we attribute to land cover class *c* in commune *m* is computed as $Y_{cm} = U_c W_m$. The coefficients U_c keep their meaning if they are multiplied by any constant *K* and the coefficient W_m is divided by *K*. The values of U_c given in Table 5 correspond to a choice of K such that the median of W_m in each stratum is 1. They still cannot be interpreted as "population density for land cover class *c*", but as median density for each land cover class in each stratum.

Table 5: Disaggregation coefficients with 6 CLC classes and three strata of communes

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

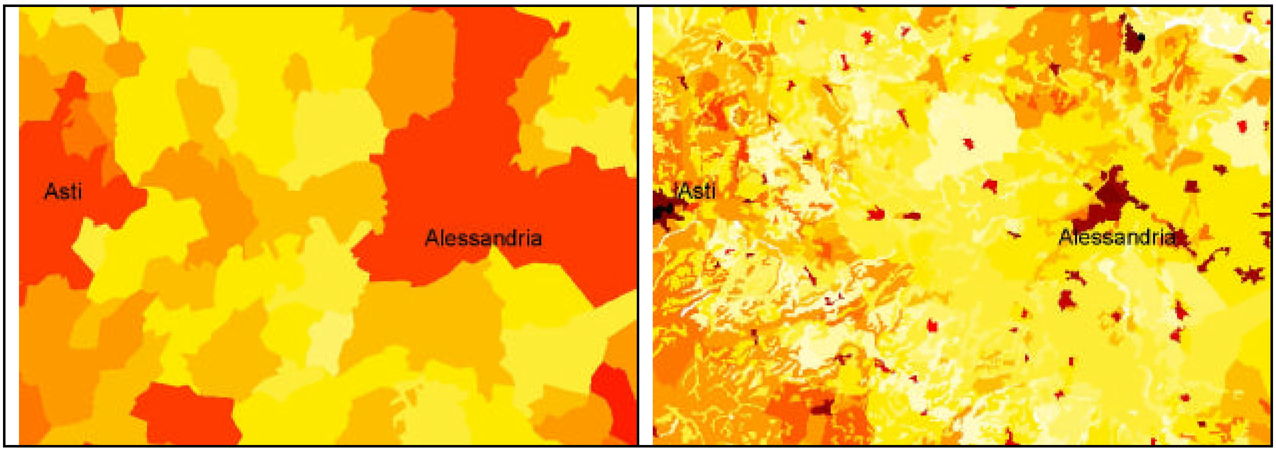

Figure 1: population density map before and after disaggregation

25. Table 6 gives the % of population that has been attributed to each CLC class in each country. The meaning of this table might need some clarification. For example, "28.9" in the cell Ireland-pasture does not mean that 28.9% of the Irish population live in pasture fields, but that this amount of population has been attributed to areas coded as "pasture" by CLC. This may correspond to a large number of small urban nuclei inside the CLC "pasture" class.

Table 6: % of population in each CLC class with the suggested disaggregation

	AT	BE	DE	DK	ES	FR	GR	IE	IT	LU	NL	PT	UK	All
Urban dense	11.8	5.7	2.5	10.1	54.1	13.2	33.2	8.4	23.8	14.8	85.3	18.3	16.1	21.9
Urban discontinuous	50.2	86.4	80.0	67.4	18.2	64.6	33.9	50.2	50.2	66.8	2.6	40.6	71.8	56.5
Indust. & commer.	0.2	0.6	0.6	0.3	0.6	0.8	0.3	0.4	0.7	0.7	0.4	0.4	0.3	0.6
Transport	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.1
Green urban	0.0	0.1	0.1	0.2	0.0	0.1	0.0	0.2	0.0	0.0	0.1	0.0	0.2	0.1
Arable non irrigated	8.6	2.4	8.1	16.5	6.2	7.1	4.4	4.1	7.9	2.5	3.1	6.1	4.8	6.7
Arable irrigated	0.0	0.0	0.0	0.0	2.2	0.0	2.7	0.0	0.5	0.0	0.0	0.1	0.0	0.4
Rice	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	0.1
Permanent crops	0.6	0.1	0.5	0.0	4.8	1.6	7.2	0.0	4.3	0.3	0.1	7.2	0.0	2.1
Pasture	6.6	0.7	2.2	0.2	0.3	2.5	0.0	28.9	0.3	1.4	3.7	0.0	5.1	2.4
Arable & perm crops	0.0	0.0	0.0	0.0	0.4	0.0	0.2	0.0	1.3	0.3	0.0	10.0	0.0	0.6
Complex agricultural	11.3	3.1	3.4	2.9	6.3	6.7	8.5	2.1	5.6	9.1	3.8	5.7	0.8	4.7
Agric. & natural veg.	0.5	0.4	0.4	1.4	1.3	0.8	2.3	3.0	1.6	1.4	0.3	3.7	0.1	0.9
Agroforestry	0.0	0.0	0.0	0.0	0.5	0.0	0.2	0.0	0.0	0.0	0.0	0.4	0.0	0.1
Forest	8.7	0.5	2.2	0.8	2.7	2.0	3.1	1.3	2.8	2.4	0.4	5.8	0.4	2.1
Natural vegetation	1.5	0.0	0.0	0.1	2.0	0.3	4.0	1.3	0.6	0.0	0.1	1.3	0.4	0.7

VIII. ARC-VIEW TOOL FOR COEFFICIENT TUNING

26. These coefficients can be seen as a starting point for a manual tuning procedure, which takes into account additional knowledge on the population settlement in specific areas. An Arc-view tool has been developed with the following functionalities:

- Viewing the attributed population density for a subset of land cover classes;
- Interactive selection of a subset of communes based on a set of criteria that includes a geographic window, average density or % of urban area;
- Attributing new coefficients to the selected set of communes.

IX. QUALITY ASSESSMENT OF THE DISAGGREGATION IN THE TEST SITE OF AREZZO (ITALY)

27. A quality assessment has been made by comparing the results of disaggregating commune data with data at sub-communal level (census sections) for a test site in Arezzo, an area of about 2000 km2 including 27 communes subdivided into 1656 census sections. These data are available as a geo-referenced layer in a GIS environment with geographic limits and population. The total population of the test site is 235,630 inhabitants.

28. For this exercise census sections and density maps have been rasterized with a resolution of 25 m. The quality of a density map ${}^{k}Y_i$ can be assessed by comparing the disaggregated density Y_{cm} (commune data + CLC) with the density by census section ${}^{1}Y_i = {}^{1}Y_s = X_s / A_s$ (assumed to be the truth), where X_s and A_s are the population and the area of the census section *s* and *i* is any pixel in section *s*. The comparison has been made in two different ways:

- per pixel: $\boldsymbol{d}_{1k}^{pix} = \sum_i \left| {}^{1}Y_i - {}^{k}Y_i \right|$
- per census section: $\boldsymbol{d}_{1k}^{sect} = \sum_s \left| X_s - {}^{k}X_s \right|$ where ${}^{k}X_s$ =population attributed to section *s* by map ${}^{k}Y_i$

29. The following density maps have been compared with the "truth" ${}^{1}Y$

k=2 : Density map closest to ^{1}Y fulfilling the condition of having the same value for pixels in the same commune and the same CORINE Land Cover class with the classification into 17 classes. These maps cannot be computed using only commune data and CLC. They give an idea of the best possible result that can be achieved by maps with this condition (scale effect).

$$^{2}Y_{cm} = \frac{^{2}X_{cm}}{S_{cm}} = \frac{\sum_j {}^{1}Y_j \, S_{cj}}{\sum_j S_{cj}}$$

k=3 : similar density map with CLC grouped into 7 classes
k=4 disaggregation with the suggested coefficients (Table 5)
k=5: disaggregation with the initial coefficients (Table 2)
k=6: no disaggregation, i.e. attributing uniform density in each commune.

	d_{1k}^{sect}	d_{1k}^{pix}
k=2	205766	235042
k=3	206737	237661
k=4	201946	240380
k=5	298676	307115
k=6	383998	425274

Table 7: Disagreement between the population density map per census section and different maps per CLC class x commune

30. Table 7 indicates that, for this particular site:

- There is a major disagreement between the density per commune and per census section;
- The disagreement is still strong between the density per census section and the disaggregation of data per commune obtained with the coefficients of Table 7;
- Most of this disagreement is due to the different scales of CORINE Land Cover and the census sections: a) many census sections correspond to a small built area and are not represented in CLC, and b) different census sections in the same commune and same CLC class (specially urban) have very different population densities.
- The map we have obtained at EU-13 level (k=4) is not too far from the optimum that can be obtained combining commune data and CLC with the constraint of homogeneous density per CLC class in each commune.

31. Systematic over/underestimation for CLC classes can also be assessed by comparing:

- the total population that has been attributed to pixels in each CLC class; and
- the population appearing in the same pixels when population density is mapped by census sections.

32. The result of this comparison (see Table 8) indicates that the amount of population attributed to CLC urban classes is approximately in agreement with the data by census sections. The coefficients seem to be too high for pasture, forest and natural vegetation, and slightly underestimated for arable land, permanent crops and heterogeneous agriculture.

CORINE	Census sections	Attributed (communes +CLC)
Urban dense	9.2	8.7
Urban discontinuous	43.3	45.0
Transport infrast.	3.2	2.2
Gree urban, leisure	0.2	0.2
Arable land	10.1	7.5
Perm. crops and heterog.	24.1	22.3
Pastures	3.5	4.5
Forest and nat. veg.	6.4	9.6

Table 8: % of population appearing in each CLC class in two density maps

X. CONCLUSIONS

33. CORINE Land Cover can be used to improve the mapping of population density available to the European Commission at commune level. An algorithm has been developed that combines two levels of administrative units (NUTS2 and communes in this paper) to estimate reasonable weighting coefficients for each land cover class. Results can be improved if each region is stratified, grouping communes with similar characteristics. The estimation of coefficients becomes more robust if the nomenclature is simplified.

34. A first assessment, comparison with more detailed data for a test site in Arezzo (Italy), suggests that the estimated coefficients are approximately correct for this site, but additional checks are necessary for sites with different styles of population settlement.

35. Several aspects are still to be improved:

- stratification can be reviewed;
- strong density jumps between adjacent communes for the same land cover class should be avoided;
- additional information should be integrated such as population settlement style (presence/absence of sparse houses in agricultural areas).

REFERENCES

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

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. Vol. 39; pp. 1-38.

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

Slater P. B., 1993, World population distribution smoothed representations. Applied mathematics and computation, Vol. 53, No 2-3, pp. 207-223

Tobler, W.R., 1979, Smooth pycnophylatic interpolation for geographical regions. Journal of the American Statistical Association, 74, pp. 519-530.