Populating urban data bases with local data

(ESPON M4D, Géographie-cités, June 2013 delivery)

We present here a generic methodology for populating urban databases with local data, applied to the case of Urban Morphological Zones and LAU2. The correspondence table between UMZ and LAU2 (see deliverables December 2012) allows populating this data base with local data, as it gives the composition of the UMZ in terms of LAU2. The intensity of each link LAU2-UMZ may be characterized by two attributes, the share of the LAU2's area that is included in the UMZ (area's contribution), and the share of the LAU2's population that is included in the UMZ (calculated with the JRC population density grid). This will be called respectively LAU2 area contribution and LAU2 population contribution.
We present in a first part the methodology for populating the database as well as a validation procedure. In a second part we propose an illustration of its interest but also of potential problems of completeness, due to missing or inconsistent values in the SIRE database[1].

I/ Methodology: illustration of the different steps based on the example of SIRE Database

1. **Join** between the correspondence table and statistical attributes
   1.1. Join between the attributes of SIRE database and the correspondence table based on LAU2 ID
   1.2. Identification of the unmatched records from one side and the other. Two methods have been used to reduce these cases.
      1.2.1. Codes: work on the correspondence between SIRE LAU2 ID and UMZ LAU2 ID (for instance, adding or retrieving a 0 in the code is sometimes enough to allow the matching).
      1.2.2. Names: using the names instead of the ID for improving the matching of SIRE and UMZ LAU2.

   Finally, only one hundred LAU2 on a total of 23 021 did not match.
2. **Allocation** and aggregation of SIRE data in each UMZ
   2.1. Allocation of SIRE data according to the intensity of each link LAU2-UMZ. Two different methods were tested, the allocation according to the LAU2 area contribution and the allocation according to the LAU2 population contribution. The first one gave too much incoherent results, possibly because of the heterogeneity of the LAU2 area sizes in Europe. We retained the LAU2 population contribution method.
   2.2. Aggregation of SIRE data into each UMZ. Three different cases were considered.
      2.2.1. One UMZ is included inside one LAU2. Then, the step 2.1 is enough for getting the data at the scale of the UMZ.
      2.2.2. Different UMZ are included inside one LAU2. For instance, in the case of Roma, in Italy, 11 UMZ lay inside the LAU2 of Roma. The, we have added the

[1] Data for census 2001 in SIRE Database 2008, Eurostat, BSI. We also used Hampson, P., Raxis, P., 2008. "Database documentation, Managment of SIRE Data Base", Eurostat, BSI, 145p.

total population of the included UMZs and the allocation/aggregation was done on the basis of this total population.

2.2.3. The other UMZ. In this case, the allocation/aggregation was done on the basis of each LAU2 population contribution to the UMZ and we have aggregated all these contributions.

2.3. Measurement of the quality of the aggregation. We have identified, for each UMZ, the number of unmatched LAU2.

3. **Verifications** of the results

3.1. Choice of a variable for the test: we have chosen a very simple variable, the population, given first by the SIRE database and secondly by the population density grid (JRC).

3.2. Comparison of the results obtained with the SIRE database by aggregation of the LAU2 share of population (previous procedure) and by aggregation of the cells of the JRC population grid

3.3. Choice of a tolerance threshold. The value of 10% was chosen[2].

3.4. Results: only 95 UMZ (out of a total of 4304 UMZ) present deviations exceeding this threshold. These results are expected for 45 UMZ that are concerned by missing LAU2 (see 1.2 and 2.3). For the 50 remaining UMZ, it is to be noted that most of the deviations are very concentrated near 10%. However, there are some exceptions that are, for most of them, due to some inconsistency in the density grid. For instance in Latvia (for unknown reason), or for UMZ located at the frontier with Switzerland where there is no data in the density grid. For these 95 UMZ, we have added an indicator of validity that indicates to which extent the data coming from SIRE database have to be considered with caution.

II/ Populating UMZ data base: illustrations of potential problems due to elementary data completeness and consistency

1. **Choice of indicators** in SIRE database

Different indicators have been selected for testing the populating methodology. We present here some conclusions regarding their level of quality, considering their potential use for populating UMZ database.

- **Total population**: the data seem to be coherent and have been used to populate the UMZ database (see above part I §3).
- **Commuting** data: we have presented these data and the problems we met with them in the June 2012 deliveries[3].

[2] It corresponds more or less to the range delimited by the quartiles of the statistical distribution of the deviation between the two measures. The deviation is computed as the ratio (in %) between the difference between JRC population and SIRE population obtained by allocation/aggregation, and the JRC population.

[3] The data about commuters are incomplete and not harmonised. A test for the Paris zone has shown that SIRE information did not completely match the INSEE national information about commuters. In particular, the lowest flows from each LAU2 are not documented and the threshold used to select data varies from one LAU2 to another.

- **Total population per age class**: the data seem to be coherent and have been used to populate the UMZ database (see below, § 2). We have aggregated the data into four different age classes, 0-24 years, 25-39, 40-64 and 65 and over.
- **Level of education**: the data have been used but the results are not good. There are no data for Germany and Lithuania, and no consistent data for Sweden, Portugal, Italy, Greece, Austria, Bulgaria, Croatia, Ireland, Denmark, and Czech Republic (the total population is different from the addition of the population of the different classes)
- **Individual housing and collective dwellings**: the data have been used but the results seem to be incoherent for some countries (see below, §3).

2. A consistent indicator: population by age classes

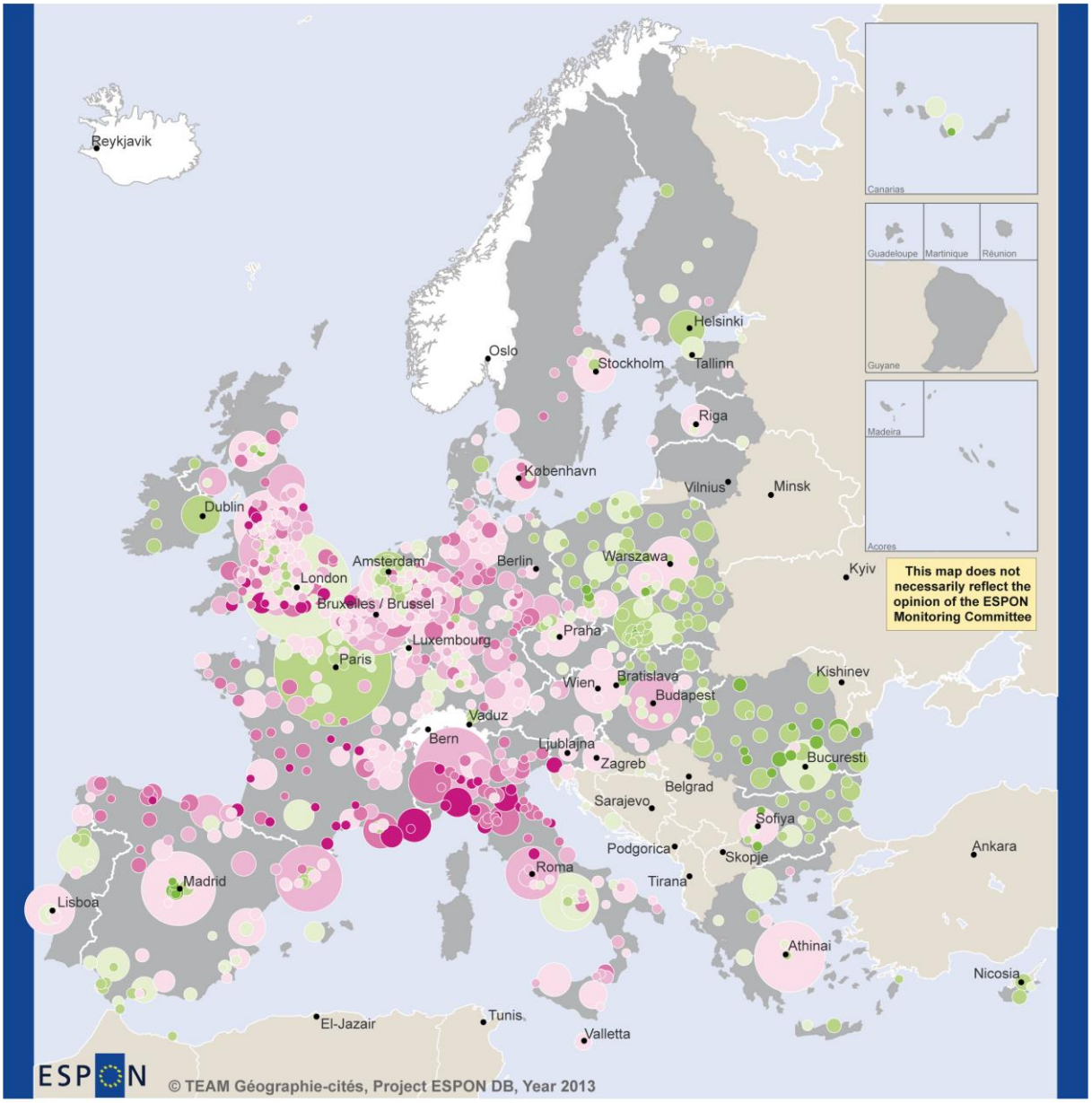
The total population per age class has been used to populate the UMZ database and the results seem to be coherent, as represented on Figures 1 and 2. On the first map, we have represented the old-age dependency ratio, which indicates the relationship between the working-age population and elderly persons (see for instance Eurostat Regional Yearbook 2011 or the ESPON project DEMIFER final report). On the second map, we have represented an indicator of demographic ageing, built as the ratio between the population older than 65 years and the population younger than 25 years. The second map may be seen as a projection in the future of the first one, making striker the contrasts between ageing regions and the other ones.

These maps enlighten *the interest of populating urban data bases by using local data*: the results are markedly different from those generally mapped with nuts 3 levels. These differences can be explained by two reasons.

First, the demographic indicator is not aggregated at Nuts3 level but in UMZ, most of them being very small and similar in size to one LAU2. Of course we recognize the national oppositions between countries with low demographic ageing (and high fertility rates, like France, Ireland…) and countries with high one (Nordic countries, north western and central and eastern countries, Mediterranean countries). But this fine scale also allows to enlighten interesting intra-regional contrasts, as the one between coastal regions (“rivieras”) and interior ones (south Spain, South England, South France, see Figure 2). Furthermore, it makes much more obvious intra-national contrasts, as the one between Scotland and the rest of England, or between the north and the south of Denmark, of Italy and of Spain.

Secondly, the UMZ database covers the whole urban hierarchy (4300 cities larger than 10 000 inhabitants), which allows to observe an opposition between small and large cities (especially capitals). The largest cities tend to be characterized by low demographic ageing (see Figure 2), probably because they are attractive for young adults, whereas the small cities of the rural counties or of the rivieras are characterized by more aged population, in particular retirees.

Figure 1: Old-age dependency ratio by European cities (UMZ) in 2001



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2013
Origin of data: Joint Research the European Environment Agency (UMZ 2000), SIRE database 2008.
© EuroGeographics Association for administrative boundaries

Old-age dependency ratio :
(Deviation from the mean of the ratios of the population aged 65 and over to the population aged 25 to 64 years)

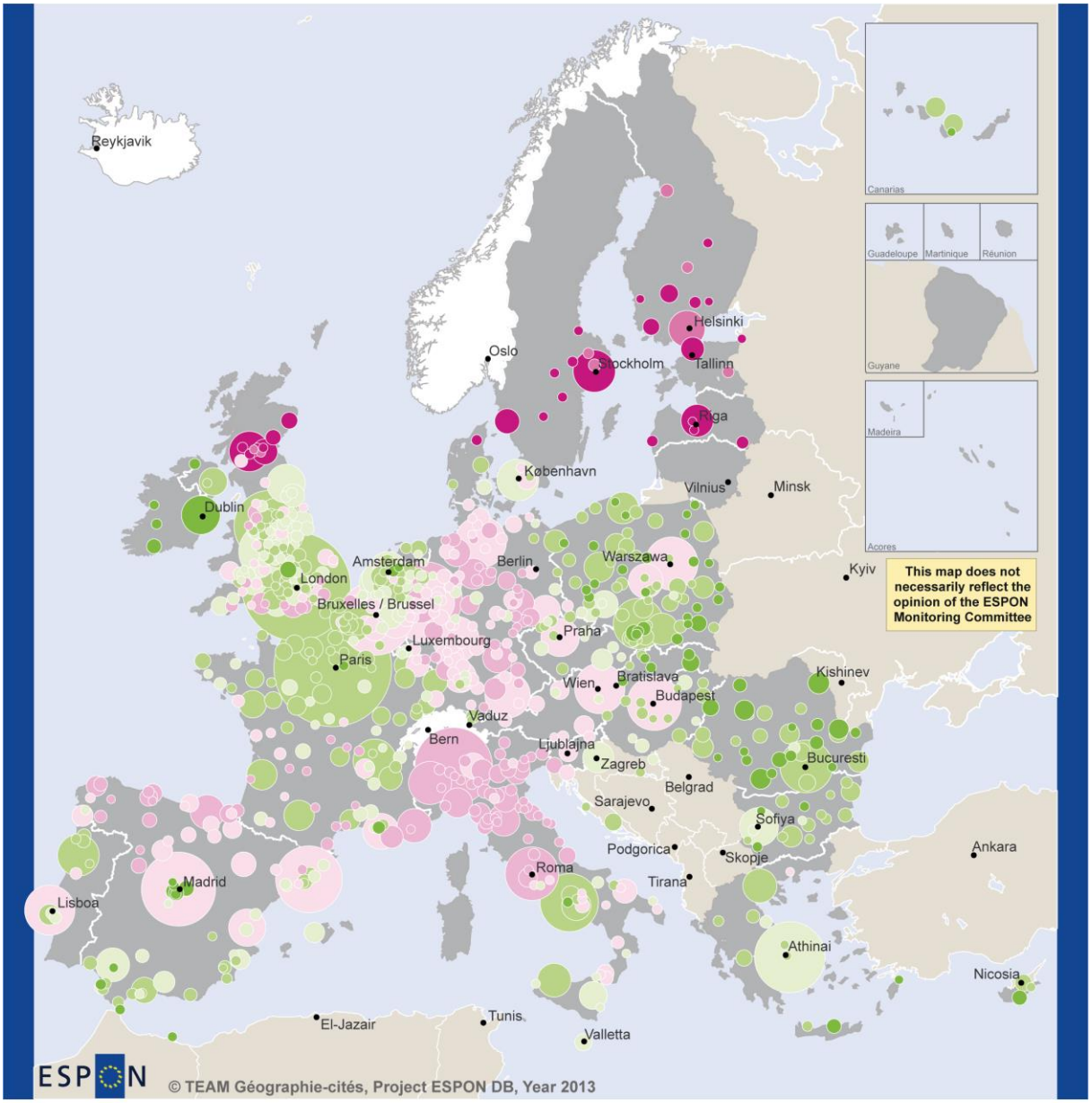
+0,34
+0,12
+0,05
+0,02
-0,02
-0,05
-0,12
-0,21

Total population :

9 700 000
2 300 000
150 000

No UMZ named
Out of Espon space

Figure 2: Demographic ageing by European cities (UMZ) in 2001



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2013
Origin of data: Joint Research the European Environment Agency (UMZ 2000), SIRE database 2008.
© EuroGeographics Association for administrative boundaries

Indicator of demographic ageing :

(Deviation from the median of the ratios of the population aged 65 and over to the population aged 0 to 24 years - In point per percent)

- +61
- +40
- +15
- +0.2
- +0.05
- -0.05
- -0.2
- -0.4

Total population :

- 9 700 000
- 2 300 000
- 150 000

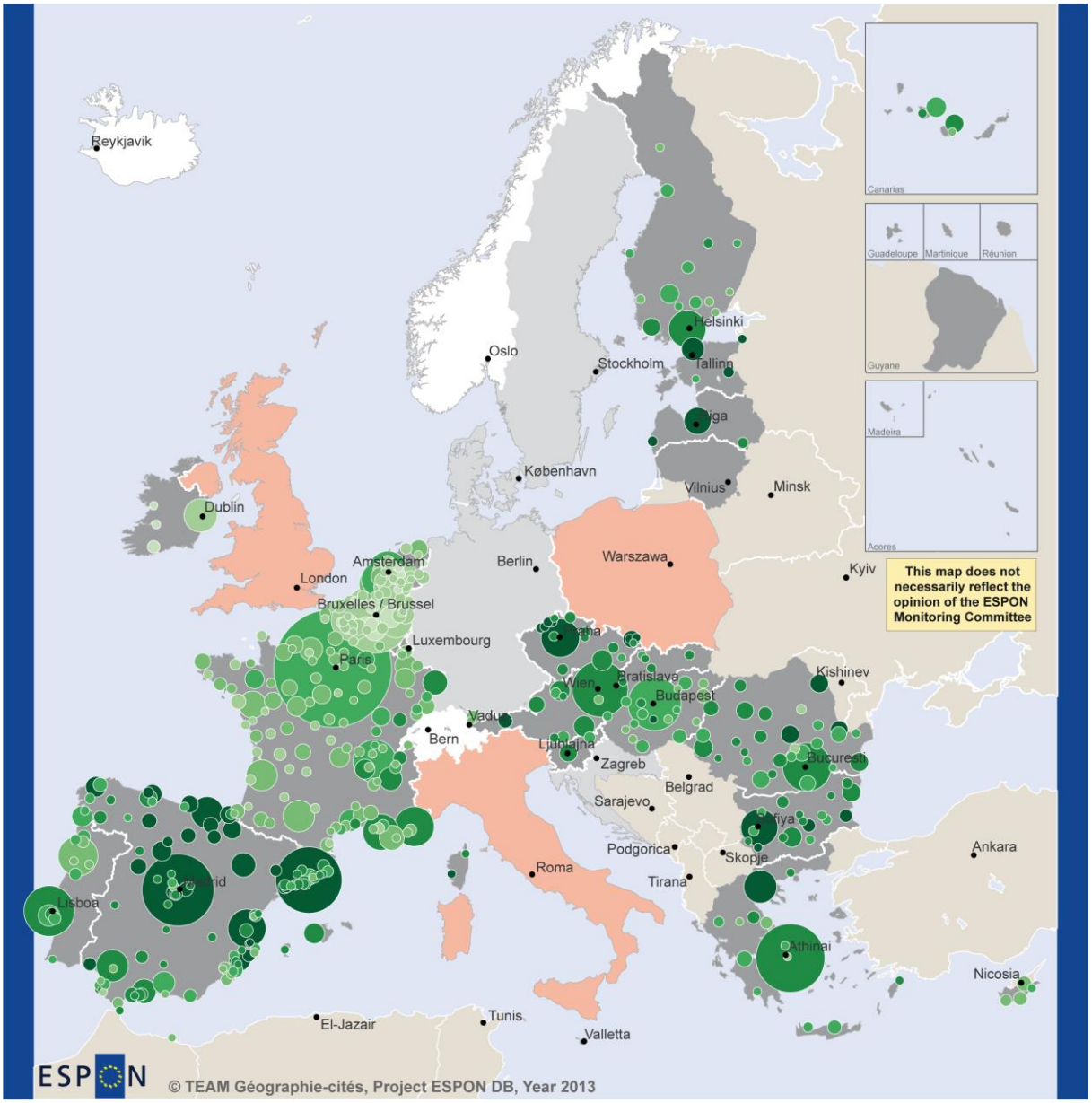
No UMZ named

Out of Espon space

3. Problems of consistency and completeness: collective dwellings

Starting from the SIRE indicators related to housing, we have chosen to represent the share of collective dwellings in European cities (Figure 3). The map enlightens two types of difficulties. First, a problem of completeness in SIRE database. Two countries (Germany and Sweden) did not send any data for collective dwellings. Secondly, a problem of consistency, easy to detect here but sometimes much more difficult to identify. In three countries, like United Kingdom, Poland and Italy, the share of collective dwelling is abnormally low (less than 0,7% whereas the average in the other countries is 81%).

Figure 3: Share of collective dwellings in European cities (UMZ) in 2001



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Regional level: NUTS 0
Source: ESPON DB, year 2013
Origin of data: Joint Research the European Environment Agency (UMZ 2000), SIRE database 2008.

Collectives dwellings (in %) for cities with 20 000 dwellings and over :

100
90
80
70
50
20
8

Total dwellings :

4 500 000
1 500 000
80 000

No UMZ named
No data (in Sire Database)
No consistent data (in Sire Database)
Out of Espon space

In order to conclude, UMZ database and the methods for populating it from LAU2 indicators are ready and the first results (demographic indicators) are very promising. We need now to collect other consistent indicators from SIRE database in order to complete the work.