TECHNICAL REPORT

ESPON

SPATIAL DATA CREATION FOR EUROPE

A TEST CASE

CONTENT

- Research problem: The report investigates how to enhance the availability of spatially fine-grained socio-economic data required for addressing urgent societal problems in Europe.
- Methodological development

  The report proposes a method for merging available tables from surveys and censuses with each other — and with remote-sense based population grid data — in order to gain a richer set of spatially fine-grained socio-economic information.
- Tentative outputs and future possibilities

  The report gives examples of the outcomes of a test version of the method for France and Sweden and discusses lessons learned for applying such methods for creating synthetic localized individuals in the entire Europe.

ESPON 2013 DATABASE
MARCH 2011

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

15 PAGES

LIST OF AUTHORS

Einar Holm, Dept. of Social and Economic Geography, Umeå University

Magnus Strömgren, Dept. of Social and Economic Geography, Umeå University

Contact

einar.holm@geography.umu.se
magnus.stromgren@geography.umu.se

tel. + 46 90 786 52 58

TABLE OF CONTENTS

1 Introduction

1.1 The Problem

For the foreseeable future, most available indicators of the state of the European population in terms of basic demography, family status, origin, education, work and income, etc. are highly aggregated. A large amount of information is compiled by Eurostat as national totals and averages. A few items are presented at the NUTS 2 and 3 levels. There is, however, the JRC (Joint Research Centre) ha population grid, which — despite some shortcomings — is a quite reasonable tool for certain purposes (Strömgren and Holm, 2010). This situation is prevailing despite the fact that substantial parts of concurrent social and economic problems of Europe are related to huge differences in the detailed spatial distribution of resources and conditions for local populations, as well as to the selective location of individuals with different socio-economic properties within the national and regional populations of Europe.

Conditions for integration of immigrants in a country are completely different if the newcomers are distributed as the general population or if they are concentrated to a few ghettos. The main development problem for remote areas in Europe is not lack of jobs but lack of qualified local specialists for the emerging jobs replacing the outdated ones — as well as for remaining qualified tasks required to be performed locally also in the future. This adds up into a national mismatch decreasing efficient use of human resources and hindering economic growth. The majority of the population and new economic activities are located in city regions covering a tiny fraction of space whereas a small share of the (aging) population is distributed over substantial parts of the countryside; containing economic activities receiving a large part of European subsidies and development support. Information for analysis and counteraction related to such urgent problems are effectively hidden by current aggregation levels in available data.

1.2 An Intermediate Solution

Since radical changes in availability of spatially fine-grained socio-economic data are far beyond the horizon, it is important to use current data as efficiently as possible. In this report, we propose and test a methodology based on merging available national and regional data tables from surveys and censuses with each other, as well as with remote-sense based population grid data. The idea is that the tables together contain more information about the joint distribution over space and attributes than do the single tables side by side.

One possible route is to start by creating an artificial synthetic individual database of the European population. In a first assessment, the attributes of the individuals are assigned more or less randomly, i.e. only conditioned by totals for nations and regions It is advisable to align the distribution of the start population based on at least two attributes, like age and region. After that, in the main step of the procedure, the artificial individuals successively exchange attribute values with each other. Each exchange is performed as long as it improves the total fit between the observed table cells and the corresponding cells in tables created by aggregating the current artificial

individuals. In the end, a synthetic individual database is created that is indistinguishable from the authentic (but unobserved) one when it comes to produce all the observed data tables together. The created synthetic individuals are jointly consistent with all information in the supplied tables (Figure 1).

After that, arbitrarily tables with dimensions corresponding to the attributes of the synthetic individuals can be computed. Many of those possible new tables were not contained in the set of used data tables, like population by age, sex, labor participation per km^2 square and so represent the new, created information. Of course, especially with just a few tables as the point of departure, the created set of attribute values for each individual is just an example out of many other distributions of sets of attributes that also, when aggregated, will be perfectly consistent with all supplied tables. The degrees of freedom for existence of such alternative populations however decreases with each additional supplied table used as constraint on the resulting population composition.

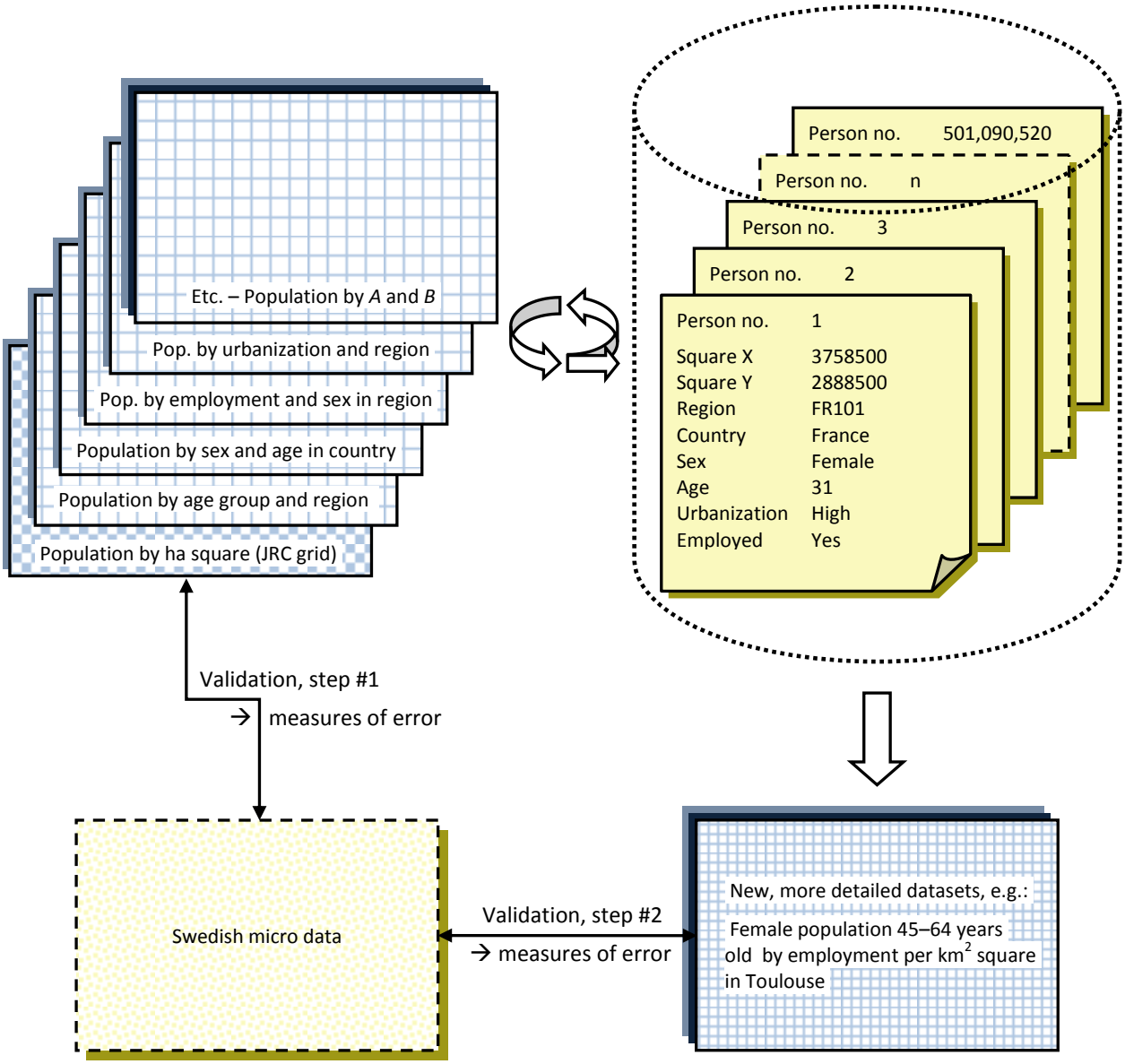


Figure 1. Using disaggregated population data together with different survey and census tables to produce a synthetic individual database — as well as new datasets.

1.3 On the Target Dataset

Most imaginable specific analyses requiring fine-grained spatial data not available today do not require the opposite extreme: micro data — information on individuals. For each of the problems exemplified above, the analysis would become greatly enhanced just by data at the km^2 level cross-classified with a few population attributes like age, sex, origin, etc. in addition to population size. As the intended applications for the created database increase, there will be additional requirements of attributes to cross-classify. Soon, the number of cells in the combined multi-dimensional table for each km^2 square will exceed the number of individuals in most km^2 squares; in the cross-classified version, most cells would be empty. So — based on Swedish experiences — as soon as the number of attributes/dimensions exceed seven or eight, the number of cells in the multidimensional representation exceed the number of individuals times attributes in the represented population. Thus, already from a resolution (and storage) point of view in all but the simplest applications, an individual representation of the target population is preferable even when no individual-level analysis is in the pipeline. In addition, many kinds of analyses would be enhanced by having access to the individual data, e.g. by being able to distinguish between individual-, group- and regional-level contributions (and interactions) in a multi-level analysis of variance components. Finally, if the quest were to feed an agent-based microsimulation with empirical content, individual data would be required. If not, the other supplied arguments for creating artificial individuals still stand.

1.4 Previous Experiences

The problem of scarce availability of spatially fine-grained socio-economic data is of course not a recently discovered one, but rather a long-standing severe constraint on country-based and European-wide spatial analysis faced by many researchers. As an example, U.K. geographers (e.g. Ballas *et al.*, 2007; Birkin and Clarke, 1989, 1998; Smith *et al.*, 2009; Voas and Williamson, 2000) have developed several examples of a methodology based on imputing the richer attribute information available in national surveys into the spatially more fine-grained census information on population by age and sex, etc. for local districts (wards), as constrained by the dimensions common for both the surveys and the census. Their efforts to create U.K. and Ireland local micropopulations containing more relevant information for current social research compared to what is available directly from published census tables target the same kind of problem as the one addressed for all countries of Europe in this report.

While U.K. geographers have pioneered the development of methods for generating synthetic populations, there are examples of such population generation from other countries as well. Using U.S. census data, Beckman *et al.* (1996) developed a method for creating synthetic populations for use in activity-based transportation models. In developing a microsimulation model for Southern France, a synthetic population dataset was produced (Aschan-Leygonie *et al.*, 2000; Holm *et al.*, 1999, 2000). In that case census tables for communes were available. Since communes constitute a quite small spatial unit, it was possible to construct a fairly detailed micropopulation with a high degree of spatial resolution. However, for the current endeavor targeting all of Europe it is not possible to presuppose the existence of such data sources for each country.

2 The Test Case: France and Sweden

In order to demonstrate the methodology, results from a test case with a somewhat constrained scope and methodology compared to what is required for a full-scale application for the entire Europe is presented in the following section. The test has been applied on data for two countries, France and Sweden, instead of for all European countries. The goal for the test is to fill a defined table with 26 combined age, sex and employment attributes for the population of each km^2 square. Data for the test is a small set of information contained in only five tables:

1. Total population per km^2. The table contains one row for each of the ca 860,000 inhabited km^2 squares of the two countries. In addition to total population per km^2 in year 2001, the table rows contain coordinates as well as codes for the NUTS 0, NUTS 2 and NUTS 3 region they belong to. Source: The JRC population grid.
2. Population by age in three age groups (0–14, 15–64 and 65+ years) in each NUTS 3 region year 2001. The table contains the number of persons in the different age groups in each NUTS 3 region of (mainland) France and Sweden — 96 departments and 21 counties, respectively. Source: Eurostat regional demographic statistics.
3. Population by sex in each NUTS 3 region year 2001. Source: Eurostat regional demographic statistics.
4. Employment by age in six age groups (15–24, 25–34, 35–44, 45–54, 55–64, and 65+ years) in each NUTS 2 region year 2001 — corresponding to 22 regions in (mainland) France and 8 national areas in Sweden. Source: Labour force survey.
5. Employment by sex in each NUTS 2 region year 2001. Source: Labour force survey.

The goal for the test model was set to "expand" table 1 into a result table distributing the total population of each km^2 square by age (six groups), sex and employment as accurately as possible given the information supplied by the five used tables. Figure 2 shows the first twenty rows of the result out of the ca 860,000 rows for each inhabited km^2 square in France and Sweden. The program creating the table works as follows:

1. One km^2 row is selected randomly with probability proportional to the number of remaining, not yet allocated individuals (random draw without replacement).
2. One of the 26 columns with all non-empty combinations of age group, sex and employment status is selected randomly with a probability reflecting the relative remaining "distance" to the constraint maximum in each relevant cell of the empirical tables — e.g. how close the created dataset is to the observed number of 45–54 years old employed females in a certain NUTS 2 region.
3. It is tested whether or not one person can be added to the chosen row and column cell without violating any constraint given by the observed cell values of the five tables. Firstly, total observed population of the chosen km^2 square should not be exceeded. Secondly, each of the 26 combined columns corresponds to specific cells in two or four of the remaining four tables. If e.g. column 5 was selected, then it is checked that current number of females allocated into the NUTS 3 area of the selected square in table 3 contains less

than the observed number in table 3. In the same way, it is tested that the current number of persons aged 15–24 in the same NUTS 3 area is less than observed. In addition, it is checked that the number of employed persons aged 15–24 in the NUTS 2 region of the chosen km^2 square in table 4 is not exceeded. Finally, it is checked that the number of employed females in the same NUTS 2 region in table 5 is not exceeded.

4. If all tests in step 3 were passed then one new person is added to the tested combination of km^2 square and age/sex/employment column in the result table. In addition, current versions of all observed tables are updated for use in next step 3 test.
5. Steps 1–4 are repeated a user-set number of times per person or until no further allocation was possible the last million test cycles.

Test/pers	100																															
Nrtest	2658830000	40.17506				66113481		Tot allc	6290564	5335892	2737488	2581597	1223475	1253171	2748383	2582130	1888851	1946631	2767687	3025452	1920538	1990201	2756235	2586686	1903558	1963821	2763268	3427766	1050422	1075806	5396186	473
Nralloc	66113363	99.89763						% alloc	9.514798	8.07081	4.14059	3.9048	1.85057	1.89548	4.15707	3.9056	2.85698	2.94438	4.18627	4.57615	2.90491	3.01028	4.16895	3.91249	2.87923	2.97038	4.17958	5.18467	1.58882	1.62721	8.16201	7.16
								Age6:	0-14	0-14	15-24	15-24	15-24	15-24	25-34	25-34	25-34	25-34	35-44	35-44	35-44	35-44	45-54	45-54	45-54	45-54	55-64	55-64	55-64	55-64	64+	64+
								Emp:	0	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	
								Sex:	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Square_ID	East_X	North_Y	NUTS0	NUTS2	NUTS3	Pop	Kumpop	CurPop																								
1	4530500	3587500	SE	SE22	SE224	25	25	23	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	1	0	0	0	
2	4531500	3587500	SE	SE22	SE224	25	50	25	0	17	0	0	0	0	0	3	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	
3	4532500	3587500	SE	SE22	SE224	25	75	25	0	9	0	0	0	0	0	5	0	9	0	0	0	1	1	0	0	0	0	0	0	0	0	
4	4533500	3587500	SE	SE22	SE224	57	132	57	0	25	0	0	0	0	0	0	0	22	0	1	1	2	1	0	0	0	0	0	0	0	5	
5	4534500	3587500	SE	SE22	SE224	605	737	605	44	81	28	9	0	9	3	1	9	17	46	80	50	6	34	12	0	0	28	11	3	50	52	
6	4535500	3587500	SE	SE22	SE224	806	1543	806	22	140	3	61	5	4	4	37	1	17	102	49	37	25	20	18	10	9	38	7	22	5	39	
7	4525500	3588500	SE	SE22	SE224	552	2095	552	21	47	1	43	12	9	46	18	4	32	5	6	3	9	23	6	27	1	59	12	0	17	36	
8	4526500	3588500	SE	SE22	SE224	267	2362	267	2	30	3	15	0	0	23	8	3	9	18	16	2	7	2	1	0	4	37	42	4	19	2	
9	4527500	3588500	SE	SE22	SE224	25	2387	25	0	1	0	2	0	0	3	1	1	0	0	0	0	0	0	0	0	0	3	5	1	1	0	
10	4528500	3588500	SE	SE22	SE224	25	2412	25	0	1	0	0	1	1	2	0	0	1	0	0	0	0	1	0	0	0	2	4	0	2	0	
11	4529500	3588500	SE	SE22	SE224	25	2437	25	1	1	0	5	2	0	3	0	0	0	0	0	0	0	0	0	0	0	3	4	1	0	0	
12	4530500	3588500	SE	SE22	SE224	25	2462	25	1	5	1	4	0	0	3	0	0	2	0	0	0	0	0	0	0	0	1	2	1	1	1	
13	4531500	3588500	SE	SE22	SE224	25	2487	25	1	5	0	3	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	5	0	3	1	
14	4532500	3588500	SE	SE22	SE224	25	2512	25	0	5	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	5	1	3	2	
15	4533500	3588500	SE	SE22	SE224	25	2537	25	0	1	0	0	0	0	2	0	1	0	0	0	0	0	2	0	0	0	1	5	0	2	3	
16	4534500	3588500	SE	SE22	SE224	25	2562	25	0	5	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1	0	2	2	0	2	3	
17	4535500	3588500	SE	SE22	SE224	25	2587	25	0	1	0	0	0	0	1	0	2	0	0	0	0	0	0	1	1	1	1	4	1	3	3	
18	4536500	3588500	SE	SE22	SE224	23	2610	23	0	6	1	0	0	0	2	0	3	0	0	0	0	0	0	1	0	1	1	2	0	2	1	
19	4537500	3588500	SE	SE22	SE224	421	3031	421	46	42	22	12	0	1	20	2	33	30	30	14	18	13	12	3	34	9	11	17	0	18	9	
20	4520500	3589500	SE	SE22	SE224	2023	5054	2023	122	188	40	52	104	16	99	58	113	7	83	94	84	85	68	67	38	74	62	112	61	161	48	

Figure 2. First 20 rows of a cross-tabulation of population by km^2 square and age, sex and employment status as created by the test model for France and Sweden.

For the test model experiment presented in Figure 2 and below, the number of tests per person was set to 100. That gives each of the 26 columns a fair chance to become a tested candidate for allocation of a new person for each individual in each km^2 square. The JRC population grid for France and Sweden contains 66.16 million persons. So, the program would at maximum perform 100 × 66.16 × 10^6 = 6.6 billion (66^9) test cycles 1–4 while trying to allocate artificial individuals into the result table. However, after 2.6 billion tests (40% of the possible 100/person), the last person possible to assign within one million tries was allocated into the result table. After that, it was not possible to allocate a single individual to any combination of km^2 square and age/sex/employment group with remaining capacity to receive another person due to at least one of the table constraints.

So, after all test cycles not 66.16 million persons were allocated but 65.11 million — 99.89% of those contained in the JRC grid. Why not all? There are two possible reasons: 1) the five used tables are not internally consistent, and 2) the allocation mechanism is oversimplified and blocks some possible solutions.

1. A first hint is given by just comparing population totals. Table 1, the JRC grid, contains 66.16 million persons, tables 3–5 68 millions. Consequently, probably a part of the error is due to inconsistencies between the utilized tables.

2. The employed allocation scheme in the test model application never reallocates a person from the first hit into another combination of km^2 square and age/sex/employment group. By accident it might happen that a combination becomes filled up to the total for the km^2 square at an early stage of the allocation process. That excludes later arrivals from entering otherwise still vacant positions in other combinations. Inspection of the result table reveals some cases with just a few columns filled especially for squares with small population in total. This calls for an allocation scheme in line with the intentions described above and in the discussion.

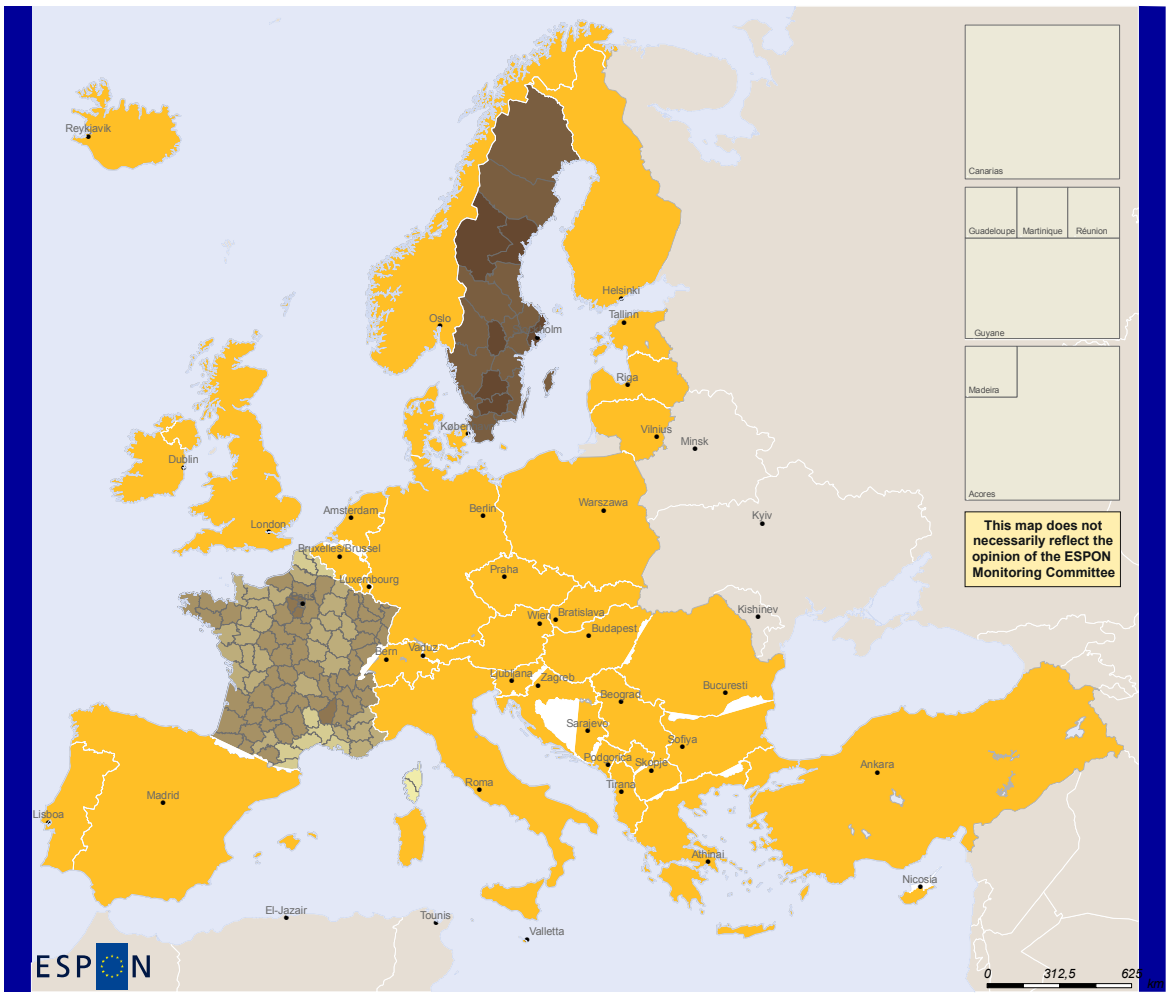
Despite the shortcomings of this simplified test model, it seems to enhance the information content possible to get from generally available socio-economic tables for Europe as demonstrated in the results section.

3 Results

The following maps illustrate some outcomes of the test. Map 1 shows work participation level for females aged 45–64 years per NUTS 3 region in France and Sweden as calculated by an earlier version of the test model. Map 2 demonstrates the same information for Sweden based on observed data. Comparing observed and calculated levels for Sweden reveals some errors in level and regional distribution. For instance, the differences between regions are smaller for calculated data compared to observed. This may reflect discrepancies between the used data sources — register data for the observed Swedish map; the Labour force survey and other Eurostat data for the calculated version — but also errors in the allocation algorithm.

The next two isarithm maps show work participation level for ages 15–64 years based on model-calculated values for km^2 squares (Map 3) and corresponding observed Swedish data (Map 4). Map 3 is created using the latest version of the test model. Comparing observed and calculated levels for Sweden demonstrates that the range of values correspond quite well with each other. However, the overall calculated employment levels are to low, due to a remaining error in the allocation algorithm. In this context, it should be mentioned that detailed maps would be become more spatially diverse and interesting, should the information-creation model also use tables that create differences between km^2 squares within regions.

Synthetic population



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

© EuroGeographics Association for administrative boundaries

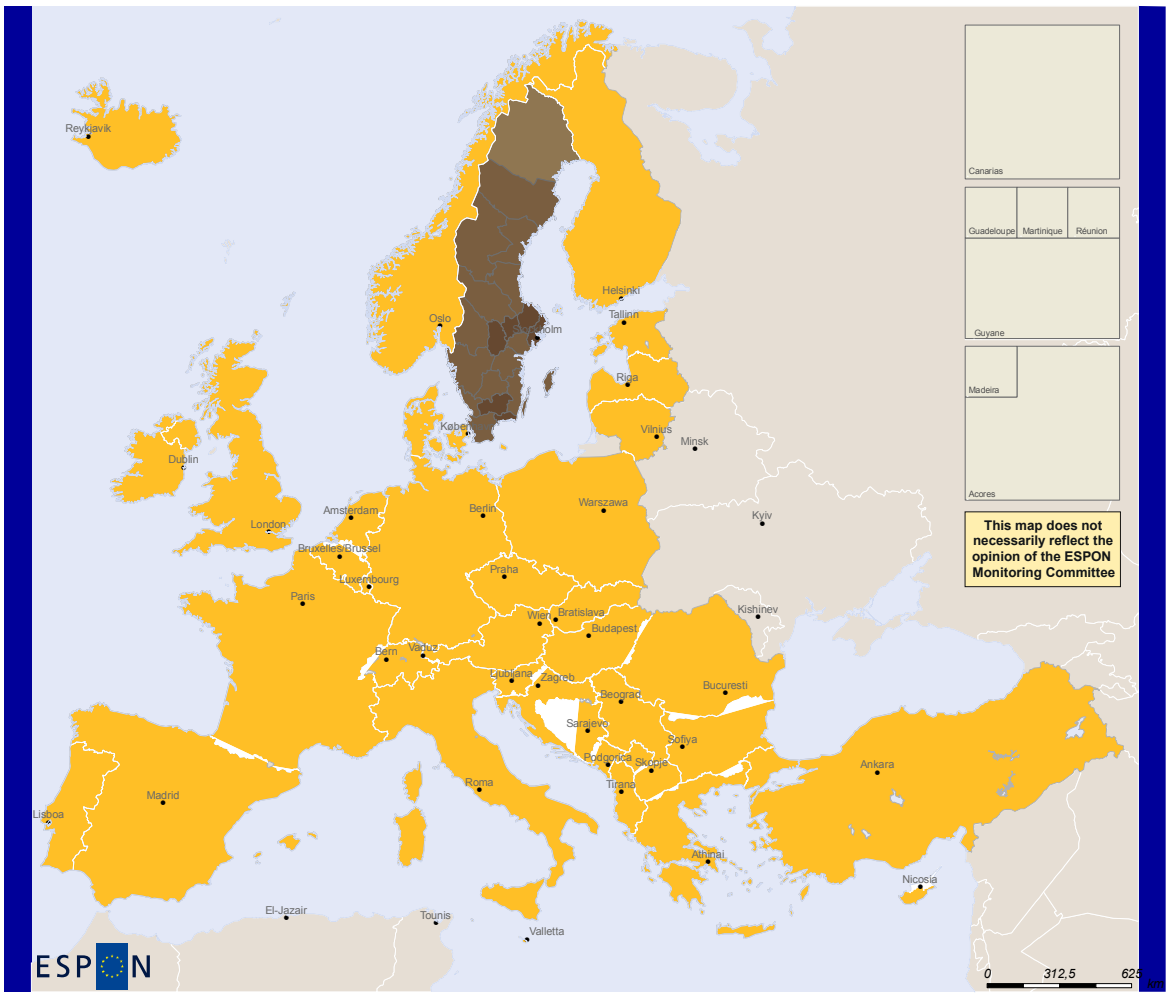
Legend

Work participation level 2001 (females 45–64 years)

- > 75
- 70 - 75
- 65 - 70
- 60 - 65
- 55 - 60
- 50 - 55
- < 50

Map 1. Work participation level for females aged 45–64 years per NUTS 3 region in France and Sweden as calculated by the test model (earlier version).

Observed population



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

© EuroGeographics Association for administrative boundaries

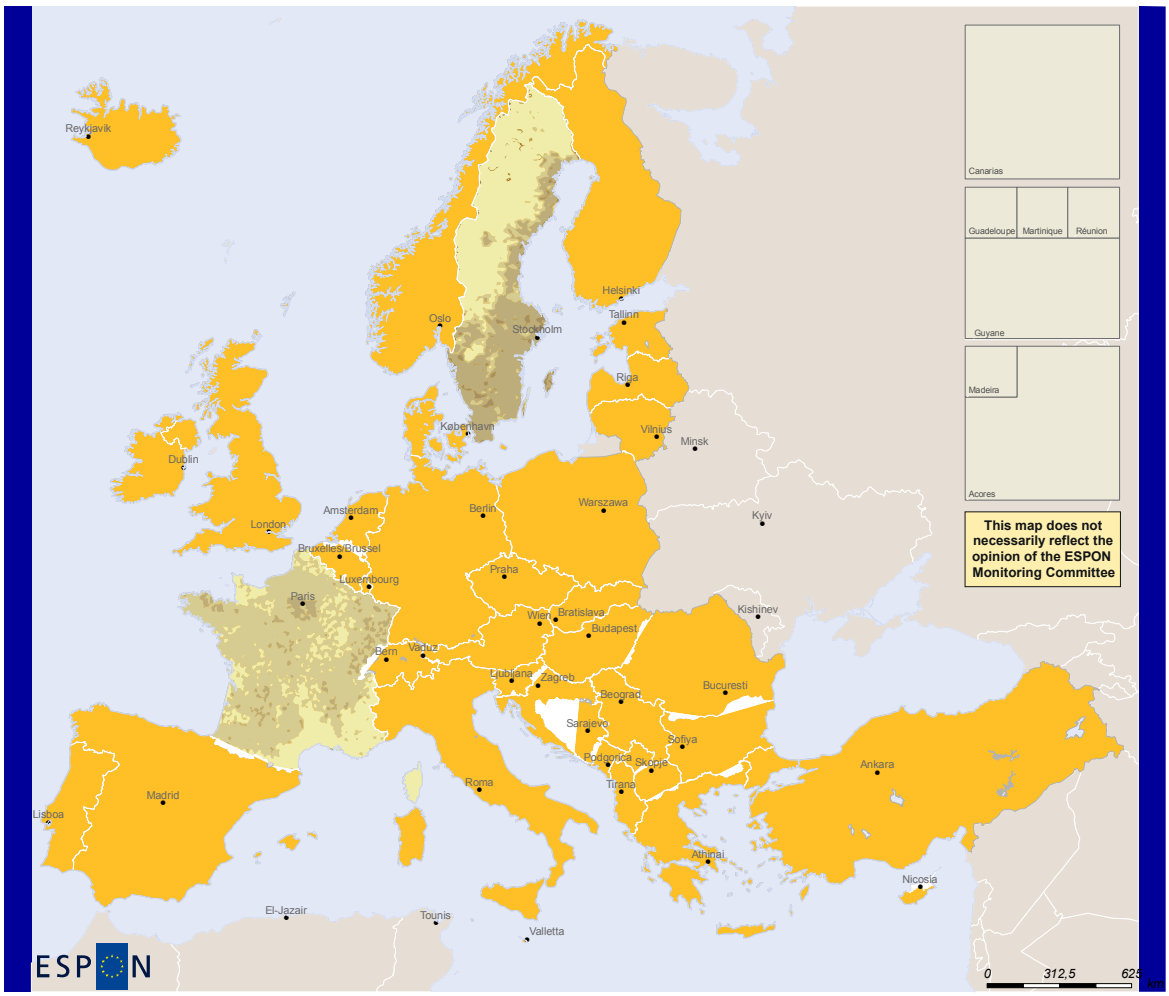
Legend

Work participation level 2001 (females 45–64 years)

- > 75
- 70 -75
- < 70

Map 2. Work participation level for females aged 45–64 years per NUTS 3 region in Sweden based on observed data.

Synthetic population



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

© EuroGeographics Association for administrative boundaries

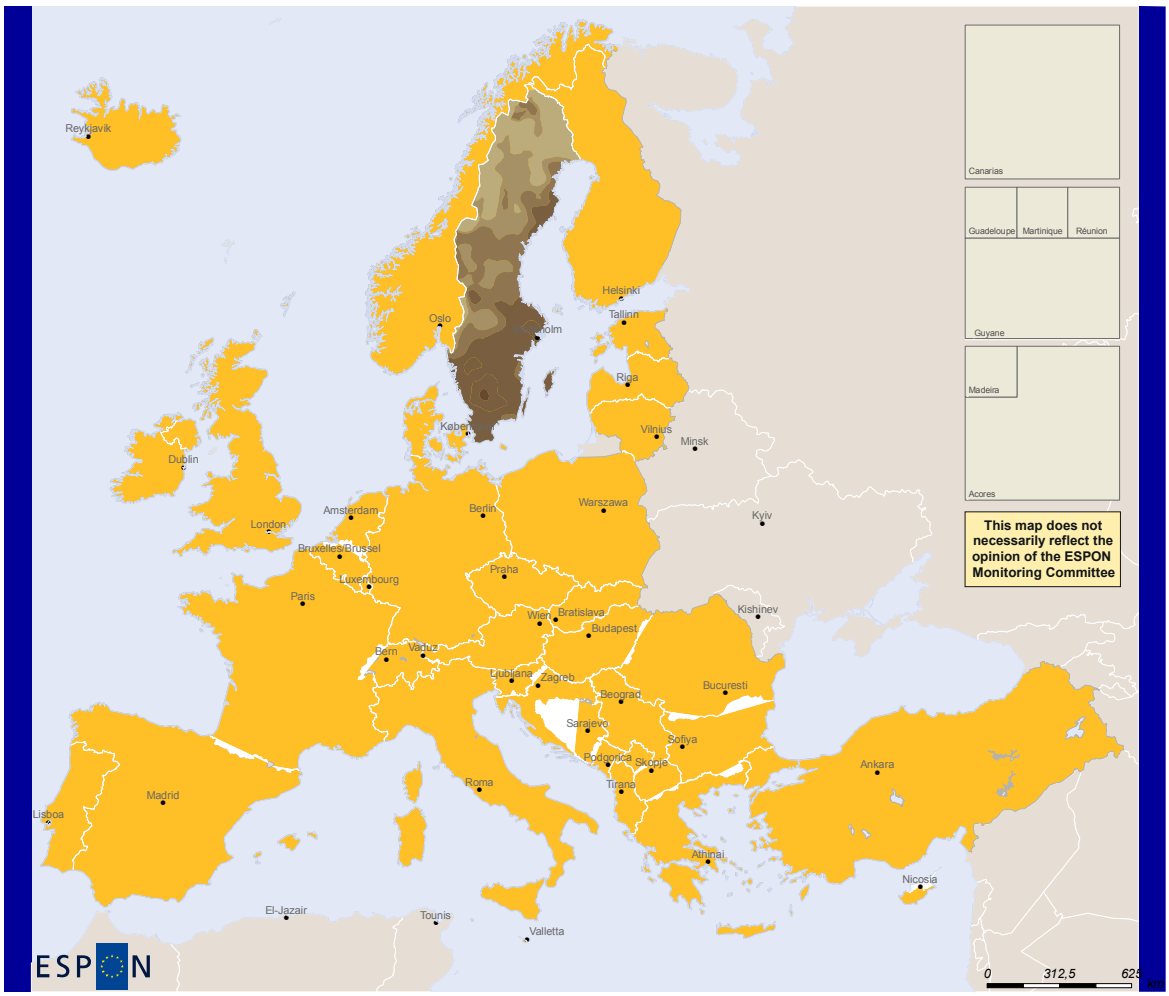
Legend

Work participation level 2001 (15–64 years)

- > 75
- 70 - 75
- 65 - 70
- 60 - 65
- 55 - 60
- 50 - 55
- < 50

Map 3. Work participation level for ages 15–64 years in France and Sweden as calculated by the test model (latest version).

Observed population



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

© EuroGeographics Association for administrative boundaries

Legend

Work participation level 2001 (15–64 years)

- > 75
- 70 - 75
- 65 - 70
- 60 - 65
- < 60

Map 4. Work participation level for ages 15–64 years in Sweden based on observed data.

4 Discussion

Some lessons can be learned from this test application of a method for enhancing the spatial resolution of available socio-economic data for Europe:

1. The presented test application demonstrates that it is possible to enhance the spatial and socio-economic detail of available regional data for Europe by simultaneously merging them based on the discussed simulation methodology.
2. If increased spatial resolution is the main target then at least one of the used tables must contain a dimension differentiating areas on the finest targeted level of resolution. This was one of the most obvious shortcomings of the test model. None of the used tables differentiated between any properties of the km^2 squares within a NUTS 3 region. Therefore, the produced differences between such squares become tiny and random. The obvious remedy is to find tables relating e.g. population density classes to age, education level, employment or anything else relevant for a targeted application and thereafter connect each of the smallest spatial units to the density class they belong to.
3. In this application, the synthetic individuals only appeared for the nanoseconds necessary to populate the defined, relatively small result table. With more input tables and potential output detail, the two-step procedure suggested initially becomes the main candidate. There would be something like the procedure demonstrated in this application in order to create a reasonably consistent start population with attributes. After that, there would be a procedure to systematically exchange attributes between the synthetic individuals in order to further improve the fit with the input tables.
4. In a real application with many tables, it becomes necessary to introduce an explicit idea about how to handle inconsistencies between the different tables. It might be the case that some tables are judged more reliable than others *á priori* and/or it might be the case that some variables have a more consistent representation in several tables compared to other variables. One approach would be to use such judgments as prior probabilities for the different tables' information in the spirit of Bayesian estimation. Then, conflicting information is averaged away by means of differing weights. Another route would be to promote one table as the master table and force all other tables to adopt while maintaining their relative distributions in attributes not contained in the master table.

References

Aschan-Leygonie, C., Mathian, H., Sanders, L. and Mäkilä, K. (2000): "A spatial microsimulation of population dynamics in Southern France: a model integrating individual decisions and spatial constraints", in Ballot, G. and Weisbuch, G. (eds.) *Applications of Simulation to Social Sciences*, pp. 109–125. Oxford: Hermes Science Publishing.

Ballas, D., Clarke, G., Dorling, D. and Rossiter, D. (2007): "Using SimBritain to model the geographical impact of national government policies", *Geographical Analysis*, vol. 39, pp. 44–77.

Beckman, R., Baggerly, K. and McKay, M. (1996): "Creating Synthetic Baseline Populations", *Transport Research A*, vol. 30, pp. 415–429.

Birkin, M. and Clarke, M. (1989): "The generation of individual and household incomes at the small area level using synthesis", *Regional Studies*, vol. 23, pp. 535–548.

Birkin, M. and Clarke, M. (1998): "SYNTHESIS: a synthetic spatial information system for urban and regional analysis: methods and examples", *Environment and Planning A*, vol. 20, pp. 1645–1671.

Holm, E., Lindgren, U., Mäkilä, K., Aschan-Leygonie, C., Baudet-Michel, S., Mathian, H., Sanders, L., Gautier (1999): "Micro-simulation of a population in a region with a strong urban growth", in A multiscalar investigation into the dynamics of land abandonment in Southern France, vol. 5 (contract ENV4CT95-0159, DGX11 of EU).

Holm, E., Lindgren, U. and Malmberg, G. (2000): "Dynamic microsimulation", in Fotheringham, S. and Wegener, M. (eds.) *Spatial Models and GIS: New Potential and New Models*, pp. 143–165. London: Taylor & Francis.

Smith, D. M., Clarke, G P. and Harland, K. (2009): "Improving the synthetic data generation process in spatial microsimulation models", *Environment and Planning A*, vol. 41, pp. 1251–1268.

Strömgren, M. and Holm, E. (2010): *Using Downscaled Population in Local Data Generation: A Country-Level Examination*. ESPON Technical Report.

Voas, D. and Williamson, P. (2000): "An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata", *International Journal of Population Geography*, vol. 6, pp. 349–366.