EUROPEAN FORUM FOR GEOSTATISTICS

TESTING AND QUALITY ASSESSMENT OF PAN-EUROPEAN POPULATION GRIDS

- BASED ON THE NORWEGIAN EXPERIENCE

Project	ESSnet project GEOSTAT 1B
Agreement No.	50502.2009.004-2009.860
WP	WP-1A Testing and quality assesment
Task	1A1-1A3
Deliverable	TESTING AND QUALITY ASSESSMENT OF PAN-EUROPEAN POPULATION GRIDS
Date	2012-12-20
Contributors	Statistics Norway Ola Nordbeck

1. BACKGROUND

In the Geostat project disaggregation techniques are important for completing the European map over population figures. Late 2011 a new version of a Pan-European dataset was released by Eurostat produced by the Austrian Institute of Technology (AIT).

AIT has used population data, from the census in administrative units, and spatial information on housing, derived from remote sensing. In terms of spatial disaggregation the global parameter is the total population of the region while the local parameter is the housing density derived from Earth Observation. Applying housing density as a proxy for population density allows estimating the local population distribution.

The dataset from AIT is not the only dataset that is based on this technique and more are to come. In the Geostat project some of the participants have contributed with their knowledge and data for the data producers in order to assess the quality and also to have some feedback on how the data can be improved.

Based on these experiences Geostat will here assemble the thoughts of the project partners.

2. AIT DATASET - BASED ON DISAGGREGATION

The following data were used in the AIT approach:

- EEA Fast Track Service Precursor on Land Monitoring - Degree of soil sealing 2006[1]
- Population per LAU2 (LAU1) 2006
- LAU2 borderlines
- Corine Land Cover 2006[2]
- Open Street Map data[3]

Based on these data AIT started of the work with creating non-residential masks with a scale 1:100000. In order to do so AIT used Corine Land Cover classes for Non-residential built-up areas (CLC 1.2, 1.3 and 1.4) and combined it with a rasterized version of road and rail networks from Open Street Map.

This combined mask was later on used on the Degree of soil sealing data resulting in a housing density layer. A linear disaggregation of LAU2 population data to Housing density layer was thereafter carried out resulting in a disaggregated dataset. In order to deliver 1km^2 this data was thereafter aggregated to the European ETRS89-LAEA grid.

For more information about the AIT approach, see:

Steinnocher K., Köstl. M., Weichselbaum J. (2011): Grid-based population and land take trend indicators - New approaches introduced by the geoland2 Core Information Service for Spatial Planning.

---

[1] http://www.eea.europa.eu/data-and-maps/data/eea-fast-track-service-precursor-on-land-monitoring-degree-of-soil-sealing-100m

[2] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-1

[3] http://www.openstreetmap.org/

Steinnocher K et al (2006): Linear disaggregation method estimating population density weights for residential CLC classes

3. QUALITY ASSESSMENT OF THE DATA USED FOR PRODUCING A DISAGGREGATED POPULATION DATASET

In order to assess the quality of the disaggregated dataset it is important to be aware of the quality of the various baseline data used in the disaggregation process of AIT. The quality assessment has in some cases been compared in urban respectively rural areas. Urban area is here the areas defined as urban settlements by Statistics Norway (see text box 1).

TEXT BOX 1. URBAN SETTLEMENT ACCORDING TO STATISTICS NORWAY

An urban settlement is according to Statistics Norway a hub of buildings inhabited by at least 200 persons and where the distance in between buildings does not exceed 50 metres. Exceptions are allowed for areas that cannot/are not to be occupied, for example parks, sport facilities, industrial areas or natural barriers such as rivers or arable land. Agglomerations that naturally belong to the urban settlement with up to a distance of 400 metres from the centre of the urban settlement are also included.

Source: http://www.ssb.no/en/beftett Data: http://www.efgs.info/data/european-datasets/norway

Below follow a quality description of the various datasets:

3.1 CORINE LAND COVER 2006:

In the case of CORINE Land Cover (CLC) three classes out of the 44 nomenclature classes are used in order to mask out non-residential areas. In the Norwegian CLC dataset there are only 31 classes are represented. The scale is 1:250'000[4].

Corine Land Cover is used by AIT for generating masks over areas that are not residential, to be more precise using the following classes:

- 1.2 Industrial, commercial and transport units

- 1.3 Mine, dump and construction sites

- 1.4 Artificial non-agriculture vegetated areas

In the case of Norway, more detailed national data have been used for generating CLC. The data used for establishing these three categories were mainly based on data from the Norwegian Mapping Authority with a scale of 1:50'000. Other more precise datasets such as port dataset, airport dataset and National cadastre data including building points were also used.

[4] http://www.skogoglandskap.no/filearchive/Rapport_10_10_corine_land_cover_2000-1..pdf

The data was transformed into pixels with a 25mx25m resolution and worked with in order to remove small areas smaller than (< 25 ha) and with a shape not broader than 100 m. This lowering of the resolution make it difficult to use this dataset for masking out many objects that are visible in the 1:50'000 dataset. The most important is here infrastructure where linear objects are removed. This is the reason why AIT decided to combine the CLC data with Open Street Map.

3.2 Open Street Map

In case The Open Street Map (OSM) data would be complete it would be important to have 2006 data, since the other datasets are from 2006. The date of the Open Street Map data used by AIT is not defined, but was probably downloaded in 2011. Below follow some comparisons between OSM and the data of the Norwegian Mapping Authority (NMA) for the road and railway network:

3.2.1 OSM Railway

Railway network is used by AIT for generating masks that are not residential.

Datasets 2011	Km of rails in total	Km of rails in urban areas	Km of underground rails in urban areas
OSM Rail	4671	2739	-
NMA Rail (scale 1:50'000)	4478	724 (210 km multi parallel rails)	83 underground (44 km multi parallel rails are underground)

The differences in the lengths of the railway network can be explained by the incomplete data specification of OSM where railways are mapped as railway lines as well as individual railway tracks. A railway according to NMA is mapped with one track only. The data specification to NMA for railways includes single or multi parallel railway lines to be compared with OSM where many tracks are mapped individually. It is therefore more reasonable to convert the data into areas and do comparisons on square kilometres instead of kilometres:

Datasets 2011	Km^2 of rails in total	Km^2 of rails in urban areas	Km^2 of rails in urban areas in tunnel/buildings
OSM Rail	39,3	7,8	-
NMA Rail (scale 1:50'000)	56,7	10,29	0,9

Completeness (%) of the OSM dataset	69,3	75,8	0

Based on 6 meter buffers around single railway lines or tracks and 11 meter around multiple railway lines.

According to the table above it is clear that the OSM dataset is not as complete as it seemed when comparing the lengths of railway data. The railway data is also more complete in urban areas than in rural areas. The completeness has an effect on the mask and built-up areas that are not covered will be considered as residential areas. The opposite effect occurs in the urban areas where the railway is underground. Since this information is not included in the attribute data some areas might be considered as non residential even though the residential area is built above the rails.

3.2.2 OSM ROADS

Road network is used by AIT for generating masks that are not residential. In order to carry out the comparison it was decided to find similarities in between the Open Street Map categories for roads and the standard used by the Norwegian Road Authority. Some of the road categories of Open Street Map were not comparable and the most important road categories were therefore chosen. Once the categories were linked it was possible to define an estimated buffer width based on Statistics Norway's experience from land use mapping:

Element according to Open Streetmap	Roadtype according to Norwegian Road Authority	Estimated buffer width (meters)*
Motorway	Europaveg	4
Motorway link	Europaveg	4
Trunk	Europaveg	4
Trunk link	Europaveg	4
Primary	Riksveg	4
Primary link	Riksveg	4
Secondary	Fylkesveg	2,5
Secondary link	Fylkesveg	2,5
Tertiary	Kommunalveg	2,5
Tertiary link	Kommunalveg	2,5
Living street	Kommunalveg/Privatveg	2,5
Residential street	Kommunalveg/Privatveg	2,5

*The estimated buffer width is a simplified version of the widths used in Statistics Norway's work with land use

In order to facilitate the assessment of the completeness of the road network in Open Street Map the N50 (1:50'000) dataset was used from the NMA. The data from the two datasets were thereafter converted into areas using the buffer widths as defined in the table above. Due to some computational errors it was only possible to compare motorways, motorway links, trunks, trunk links, primary roads and primary links in the table below:

Datasets 2011	Km^2 of roads in total	Km^2 of roads in urban areas	Km^2 of roads in urban areas in tunnel/buildings
OSM Road	81,6	12,09	
NMA Road (scale 1:50'000)	71,6	8,37	0,42
Completeness (%)			

3.3 SOIL SEALING DATA:

The EEA Fast Track Service Precursor on Land Monitoring is a raster dataset for built-up areas including continuous degree of imperviousness ranging from 0-100% in spatial resolution of 1ha. The dataset is also referred to as "High resolution soil sealing layer" (HR-SS). The data set is based on orthorectified high resolution satellite imagery (Image2006), acquired primarily in the reference year 2006 (+/- 1 year). Supervised classification techniques were used to automatically map built-up areas, followed by visual improvement of the classification results. The degree of soil sealing for the classified built-up was derived from calibrated NDVI (normalised difference vegetation index). The data set covers EU27 and neighbouring countries, in total 38 countries[5].

3.3.1 TEMPORAL RESOLUTION

When it comes to making use of satellite imagery and NDVI it is important to have in mind that the images used represents only a snapshot in time. The consequence is that built-up areas in green surroundings potentially with green canopy covering these areas might not be considered as built-up. It is therefore important to compare images of various dates before during and after the vegetation period. In other areas you might find opposite effect where some vegetation has been removed for later on grow up again.

[5] Steinnocher K., Köstl. M., Weichselbaum J. (2011)

A new soil sealing dataset for 2009 is about to be produced and in parallel the 2006 dataset is being modified. The modification seems to have only minor consequences for the classification where only of 0,5 % of the urban area of Oslo differ.

3.3.2 GEOGRAPHICAL RESOLUTION

The geometrical accuracy of the HR-SS datasets for Norway is 9,31 meters (Root Mean Square Error i x og y) and a good rendition of the geography of sealed areas in Norway when the

results are presented on small scale maps (1:5 00 000) where details are discarded and focus is on the overall national distribution of sealed land.

Norwegian Forest and Landscape Institute (NFLI) carried in 2008 out a Qualitative assessment "Verification of high resolution soil sealing layer" of the HR-SS image. They also recoded the data with areas having ≥ 80 % soil sealing represented as "1" and the remaining areas represented as "0". The recoded dataset is called HR-SS-80 and according to the assessment of NFLI following findings were important:

- HR-SS-80 is a good geographical representation of sealed land within "urban" (or densely built up) areas, including industrial sites.

-HR-SS-80 is an insufficient geographical representation of sealed land in "rural" areas and the outfields and lower mountains. Only a small fraction of these areas are sealed, but the inclusion of actually sealed areas in HR-SS-80 is coincidental. Furthermore, some of the areas included in HR-SS-

80 are not permanently "sealed" but instead areas where the vegetation is temporary removed.

- HR-SS-80 is of excellent quality for land cover types without any soil sealing, where the dataset correctly shows no soil sealing.

- The statistical quality of HR-SS and HR-SS-80 is unknown. Sealed land in "rural" areas is obviously underrepresented in HR-SS-80, but this may be levelled out by the representation of percentage sealing in HR-SS. Norway has databases, including 1:5 000 digital maps, a fairly complete register of buildings and a complete road database that can provide for compilation of much more reliable and accurate statistics than HR-SS. We therefore recommend that even though the data may be used for statistical purposes on the European level, statistical results should not be published for Norway or parts of Norway based on these data6.

[6] Norwegian Forest and Landscape Institute (2008): "Verification of high resolution soil sealing layer - Qualitative assessment"

These findings of Norwegian Forest and Landscape are important for the later comparison in between population grid of AIT and the point based population data of Statistics Norway.

4. QUALITY ASSESSMENT OF A DISAGGREGATED POPULATION DATASET

In order to assess the quality of a disaggregated dataset it was in the Norwegian setting important to compare the AIT data with aggregated population data from Statistics Norway. This was done on a square kilometre basis comparing urban and rural areas as illustrated in the following tabular setup:

	AIT, total	Norwegian population data, total	Difference (%)
Total population, grid based (354487 pixels)	4617792	4624530	- 0,15
Total population, within urban areas (208 pixels)	569597	599550	- 5
Total population, within rural areas (345028 pixels)	451082	853104	- 47,12
Total population, in areas where the urban outline is crossed by the population grid (9149 pixels)	3592536	3164965	13,51
Total population, in areas where the urban outline is crossed by the population grid and where the population is > 500 (1831 pixels)	2259217	2187680	3,17
Total population, in areas where the urban outline is crossed by the population grid and where the population is < 500 (6898 pixels)	1316876	975285	35,02

In the table above it is obvious that the populations in rural areas as well as in the most central urban areas are underestimated. The populations in the outskirts of urban areas that are not densely populated are overestimated in the AIT dataset

Based on the knowledge of the strengths and weaknesses of the baseline data used in the disaggregation process, make it easier to understand the various errors in the table above. In the case of Corine Land Cover and the soil sealing dataset the methodology is well specified and the errors are well defined while the completeness of OSM is more difficult to handle.

5. Possible ways to improve the disaggregated dataset based on data from and knowledge of the NSIs

Based on the quality assessment above it is clear that the use of national data can and will improve the disaggregation result. One example of how this is by generating a national infrastructure mask as the consequences of an incorrect infrastructure mask in urban areas (or in residential areas close to the railway/road or underground constructions) result in an underestimation of population in grid cells.

Another important improvement by using national data is to distinguish residential areas from rocky/stony areas, areas where vegetation is temporary removed or industrial areas. For NSIs with access to address or building data this can be done by filtering out grid cells without residential addresses or buildings.

The urban settlement dataset of Statistics Norway has proved to be very useful in order to determine areas for improvements. It is therefore recommended that similar studies are carried out by the NSIs in order to improve the datasets.