**Making census statistics more relevant – towards geo-enabled statistics**

Adam Wroński, Eurostat, adam.wronski@ec.europa.eu

David Thorogood, Eurostat, david.thorogood@ec.europa.eu

Fabian Bach, Eurostat, fabian.bach@ec.europa.eu

**Abstract**

*At a time when there is evidence of public distrust of statistics and of increasing distance between policy makers and the public, it is important that official data should relate closely to the realities of people's lives. One way to do this is to provide geographically detailed statistics that describe the situation at local level. In many contexts, statistics at the level of towns and neighbourhoods are the most relevant and most meaningful to the citizens, and are what is needed to support the local level policy decisions that are of importance to people's day-to-day lives.*

*For the 2021 EU population census, building on the success of the GEOSTAT 2011 grid and in response to growing needs for high resolution population statistics, work is advanced on collecting 13 census data items coded to a 1km² grid. This will be a new departure, allowing new types of statistical and geo-spatial analyses for areas that can be flexibly defined according to the needs of policy makers and researchers. In the future it will also assure geographical comparability over time. One use of such data is to ensure the efficient allocation of funds (e.g., European Cohesion Policy budget - roughly €50 billion/year) through the analysis of populations for small urban and rural areas, and assessing the accessibility of services for citizens.*

*This paper presents the current plans for developing the grid data collection for the 2021 population census and discusses the main legal, organisational and methodological issues that need to be addressed. It describes the innovative and complex development, requiring a separate legal act and a number of solutions to emerging technical issues. It also looks forward towards the post-2021 EU population censuses as the attractiveness of providing grid-coded population data made it one of the focal points of the developments.*

**Keywords:** population census, geo-coding.

**1. Introduction – the relevance of census**

Thirty-two European countries published their Census 2011 data on the European Census Hub. This system gives easy access to a vast amount of information on the population of these countries. However, it has been acknowledged that the timelier availability of the data, comparability between consecutive population and housing censuses (that traditionally happen every 10 years) and more small area data would have boosted their usage. The unique advantage of the census is that it represents the entire statistical universe, down to the smallest geographical units, of a country or region. Planners need these statistics for all kinds of development work, including: assessing demographic trends; analysing socio-economic conditions; designing evidence-based poverty-reduction strategies; monitoring and evaluating the effectiveness of policies; and tracking progress toward national and internationally agreed development goals.[[1]](#footnote-1) Unfortunately, the smallest traditional geographical units (e.g., LAU2) tend to evolve[[2]](#footnote-2) sufficiently fast to mean that the data from two consecutive censuses is largely non-comparable due to these changes. As a result censuses become isolated snapshots of the population. On the other hand, the public wants more of what we provide: faster statistics; in more fine-grained forms; and which directly address the issues of everyday life.[[3]](#footnote-3) As argued in a Guardian article by William Davies[[4]](#footnote-4), traditional methods and products of official statistics are nowadays coming under increasing pressure from new methods for decentralized and data-driven information that can be harvested by the internet. In particular, the article discusses in detail how long-standing national or even regional administrative aggregates fail to meet modern data analysis demands and thus fall short of public expectations.

This paper describes how the collection of population statistics at the common 1km² grid level can help address these issues for the EU population and housing census. The next section shows how the evolution of the census and the technology in other areas has made the 1km² grid data possible. Section 3 presents the approach taken by the EU Member States for the 2021 Census and the following section gives an insight into longer-term ambitions.

**2. The evolving background – an opportunity**

*2.1. Administrative data.*

Administrative data have been used in official statistics since many decades but only during the recent years has this trend become very prominent as advances in technology have facilitated the fast processing of large data sets. In parallel, there has been an increase in the demand for statistics on all aspects of our lives as well as growing concerns about the cost of the census and the increasing burden of survey response. For all these reasons, administrative data are increasingly used for population censuses. In fact, over 20 Member States will run purely administrative or combined administrative (with some other data collection) censuses in 2021; roughly double the number in 2011.[[5]](#footnote-5) This has fostered a very positive climate for the introduction of the geo-localisation of all dwellings that is a necessary prerequisite for the production of geo-coded statistics.

*2.2. INSPIRE directive.*

The INSPIRE[[6]](#footnote-6) Directive aims to create a spatial data infrastructure for the EU. This involves common technical standards intended to enable the sharing of spatial information between administrations and facilitating public access. A particular aim is to assist in policy making and evaluation across national boundaries by the use of EU-wide standards. The INSPIRE Directive came into force in 2007, with staged implementation to be completed by 2021. Under this Directive, EU Member States must ensure that spatial data falling under INSPIRE and produced by public authorities comply with the common INSPIRE standards relating to data, metadata and services. Consequently, any grid data produced as part of the EU census programme would need to comply with the requirements of the INSPIRE Directive. The technical requirements of the grid data, such as the definition of the grid and cell codes, as well as important features of the service infrastructure, are determined by INSPIRE.

Due to this legal obligation most NSIs were anyway independently planning to produce INSPIRE-compliant 2021 census statistics geo-coded to a grid as part of their national census programmes.

*2.3. Availability of geospatial data.*

During recent years, advances in technology have significantly improved the availability of geospatial data and many interesting use cases combining a census grid with geospatial data from different areas have emerged. These include: land cover / use and related surveys such as LUCAS including improvements to their sampling design, correlation with transport infrastructure, exposure to diseases; satellite data from the Copernicus programme[[7]](#footnote-7), and ecosystem services accounting.

These GIS applications need to employ the underlying population distribution to calculate exposure, risks, availability and cost of services. Consequently, a geographically precise population distribution is needed. A population census is the only statistical data collection that provides simultaneous, individual enumeration of the entire population of an area of interest.

For the census the GEOSTAT 2 report[[8]](#footnote-8) points to geocoding of individuals and households using the locations of either addresses or buildings or dwellings. Obtaining these locations is the task of a national geocoding infrastructure that should be provided as part of the national Spatial Data Infrastructure, usually by National Mapping Authorities. Hence, the 2015 report of the UN Global Geospatial Information Management (UN-GIMM) expert group highlighted that the 2020 round of censuses would be an important opportunity for the integration of statistical and geospatial data. The European Statistical System (ESS) also noted on many occasions over recent years the importance of developing geo-coded statistics: starting with the ESS Vision 2020 that cites the potential value of merging geospatial data with official statistics to provide better social and environmental information and, most recently, with the Budapest Memorandum in September 2017 stating the need '*to develop a basic set of census-related geo-referenced data covering a limited range of demographic and socio-economic characteristics, as a source of geographically detailed regional data on migrants'*.[[9]](#footnote-9) Among other European institutions, the Commission's Directorate General Regional and Urban Policy (DG REGIO) expresses the need for geo-coded population data to ensure the efficient allocation of funds (e.g., European Cohesion Policy budget - roughly €50 billion/year), the analysis of population for small urban and rural areas, assessing the accessibility of services for citizens, population exposure and emergency preparedness, etc.

As mentioned above, it is difficult to compare regional census data at the lowest level (LAU2 or municipalities) because the definition of these regions is subject to different national needs and because a high proportion of LAU2 units change over time. The statistics made available on a common European grid would be much more useful for regional and cross-border analysis. Moreover, sufficiently small grid cells could be freely clustered into regional aggregates according to individual analysis needs (e.g. functional areas describing environmental or agricultural aspects such as flooding zones). This would be unrelated to any administrative patterns or boundaries and thus stable over time.

Additionally the recently proposed definitions and analysis of territorial typologies[[10]](#footnote-10) such as urban and rural regions, border or coastal areas make direct use of 1km² grid population counts.

**3. The 2021 round of the EU population and housing census.**

In an effort to respond to these growing needs, taking advantage of technological advances and using this opportunity to increase the relevance of the census, the ESS is working together on geocoding existing census microdata to a common 1km2 grid. Selected technical information about the data collection and related methodological issues are presented below. These statistics would thus become a European pilot for the integration of statistical and geospatial information in a real production case bringing numerous additional possibilities: e.g. such information included in the survey frame could improve sampling design or lower the cost of data collection.

*3.1. The statistical data* *to geo-coded to the 1km² grid.*

The envisaged statistical data takes the form of a single two-dimensional table cross-tabulating the set of grid cells against 13 data items (selected categories of variables). In effect, 12 or 13 numbers have to be prepared for each grid cell:

* total population;
* sex (males, females);
* age (under 15, 15-64, 65 and over);
* employed persons (to be included only if available nationally);
* place of birth (in the reporting country, in another EU country, outside EU);
* usual residence 12 months before (unchanged, within reporting country, outside of the reporting country).

There would be no cross-tabulation of the different topics. Total population is to be provided within 12 months after the end of the census reference year (31 December 2022), the remaining data items 15 months later (31 March 2024).

*3.2. Statistical confidentiality and disclosure control*

Data at a 1km² grid level present particular confidentiality challenges that must be addressed as an integral part of the collection and dissemination process. The need for appropriate disclosure control methods was one of the key issues raised from the beginning of the work.

During 2016-17, a group of NSIs undertook a project ‘Harmonised protection of census data in the ESS’ under a Specific Grant Agreement (SGA) with Eurostat. This project has resulted in detailed methodological recommendations for a harmonised disclosure control methodology that can be used by NSIs to protect both the main data outputs of the 2021 EU census programme, and the census grid data.[[11]](#footnote-11) The use of these harmonised methods is recommended but voluntary - final decisions on the use of particular disclosure control methods rest with the NSIs.

The potential core data users (e.g. DG REGIO) maintain that one of the key factors impacting on the value of the census grid data is that information on whether a grid cell is truly populated or not (i.e. one or more persons truly observed in that grid cell or not) should be reported very accurately. However, this would potentially impact on the disclosure control measures that can be applied. Based on the SGA results, to overcome this problem it is proposed that a flag of ‘populated’ should be added to all grid cells where the observed population (before the application of any disclosure control measures) is not zero.[[12]](#footnote-12) This approach can straightforwardly be combined with the confidentiality methods recommended by the project and many national ones.

The overall result is a harmonised methodology (recommendations for best practices and implementation guidelines as well as actual software code) that the NSIs can choose to take up to protect their 2021 EU census programme outputs, specifically including non-nested 1km² grid statistics.[[13]](#footnote-13)

*3.3. Compliance with INSPIRE directive*

As discussed earlier the census grid data must respect the following technical requirements set out in the INSPIRE directive:

* sharing of census data using defined data models;
* a download service for the census grid data;
* a view service to display the grid data (the map viewing);
* INSPIRE metadata on data and on the above services.

Following detailed consultations within the Commission, an approach has been developed whereby the NSIs will be able to fulfil their INSPIRE data and metadata obligations by supplying the agreed set of European census grid data using the SDMX information model[[14]](#footnote-14). Eurostat will provide a central INSPIRE-compliant dissemination infrastructure, i.e. the download and view services for the census grid data, based on an extension of the Census Hub – EU population censuses dissemination system. This approach guarantees that by supplying the census grid data and metadata to Eurostat as defined, Member States are fulfilling their obligations under INSPIRE. It has been confirmed that the Census Hub, the INSPIRE Geoportal and the Eurostat view service infrastructure, could be extended to meet all four requirements listed above.

There are benefits for both Eurostat and the NSIs of this common centralised approach to the production of census grid data:

* efficiency and cost-effectiveness, because existing resources would be reused;
* more user friendly dissemination, since there will be one central access point for data and services instead of 28 different national access points;
* straightforward combination of national census grid datasets into pan-European, cross-border grid statistics.

In particular, for some NSIs, complying with the INSPIRE Directive is likely to pose significant extra burden. The use of this common INSPIRE-compliant approach greatly reduces this burden. For Eurostat, the marginal costs of creating the census grid data are reduced. This would be the first real-case model for the integration of statistical and geospatial information in a geospatial statistics product at European level. The linkages between SDMX and INSPIRE will serve as a useful test case for further developments.

Two other important metadata features worth mentioning are:

* As set out by INSPIRE and the planned TERCET regulation, the 1km2 grid to be used for European purposes is the ETRS89 / ETRS-LAEA (EPSG:3035) grid.[[15]](#footnote-15)
* The list of metadata on data items defines flags that may be applied to data items in certain circumstances. The flag ’confidential’ may not be applied to the 'total population' data items.

**4. Post-2021 census strategy of the ESS**

Improving both spatial and temporal comparability is one of the core objectives of the EU post-2021 census strategy, where the successful introduction of a European 1km2 grid based annual census data – unaffected by administrative changes and hence constant in space and time – would be an important milestone. Related geocoding infrastructure and geographic information system (GIS) technologies set up for the 2021 census would therefore be an investment generating considerable added value, as it would pave the way for the long-term future of (small area) census statistics. Once set up for 2021, this infrastructure could then be reused annually.

Subsequently, frequent and timely census information on a 1km2 grid would enable the ESS to provide more relevant, coherent and comparable geographically resolved demographic statistics and allowing for flexible use demographic statistics: in addition to the regional comparability over time, and in the light of ever increasing use of cost-efficient register-based census methods, there is realistically no other means of producing social statistics of this level of accuracy and geographical detail at a comparable cost – not through any conceivable survey programme.

In the long-term, when the technological, methodological (e.g. disclosure control) and data availability advances in Member States permit, one could consider also collecting:

* other demographical characteristics (e.g. births, deaths) or socioeconomic characteristics (e.g. educational attainment) or information on the dwellings/buildings (e.g. number, number of floors, use – residential, commercial, industrial);
* smaller size grids (e.g. 500m) in densely populated areas;
* in a more timely way.

**5. Conclusions**

Recent advances in technology and statistical methodology have created new opportunities for statisticians, analysts and researchers associated with geo-spatial data that can be combined with social and economic statistics. Researchers and analysts in various areas (incl. policy-making and evaluation) have become aware of these opportunities and this has resulted in increasing calls for statisticians to respond to needs for geo-enabled data. The population and housing census (and related dwelling and population registers) is a natural use case for these advances because of its universality and small area statistics. The ESS sees it as a priority.

Responding to these calls, the dissemination of geo-coded census data is in preparation for the 2021 round of EU Censuses: the harmonized EU-wide availability of selected census topics on a 1km2 grid to provide users with spatially resolved demographic information of the highest quality (completeness, accuracy, reliability) across the EU.

The work for the 2021 EU census can be seen as a model for further development in the post-2021 census era: a good starting point for annual 1km2 grid data updates. If this approach succeeds, it may also be possible to add on an incremental basis, additional variables without losing timeliness or grids of smaller size than 1km. To this end, Eurostat together with its ESS partners is continuing to address the different technical challenges such as disclosure control.

1. E.g.: http://www.unfpa.org/census [↑](#footnote-ref-1)
2. Around 2% annual change of around 100 000 LAU2s in EU 28 [↑](#footnote-ref-2)
3. Pullinger J. (2017) Letter to the Guardian. Available at: https://www.ons.gov.uk/news/statementsandletters/lettertotheguardian (Accessed: 2 May 2018). [↑](#footnote-ref-3)
4. <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy> [↑](#footnote-ref-4)
5. See dedicated Q2018 sessions on using administrative data 'Use of administrative data sources, data integration' as well as 'Quality of multisource data' [↑](#footnote-ref-5)
6. Directive 2007/2/EC (INSPIRE Directive) and its implementing rules harmonize the interoperability and accessibility of geospatial data on 34 themes relevant for the environment (among them population distribution) across the EU. [↑](#footnote-ref-6)
7. <http://www.copernicus.eu> [↑](#footnote-ref-7)
8. <https://www.efgs.info/geostat/geostat2/> for the final report of GEOSTAT 2 on a point based national geocoding infrastructure for statistics. [↑](#footnote-ref-8)
9. Budapest memorandum ['Population Movements and Integration Issues - Migration Statistics'](http://ec.europa.eu/eurostat/documents/7330775/7339365/FINAL%2BBudapest%2Bmemorandum.pdf/96a6db89-1395-44a5-8a46-85e8c49d576c), adopted by the ESSC on 22nd September 2017 [↑](#footnote-ref-9)
10. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0788 [↑](#footnote-ref-10)
11. Detailed results at: <https://ec.europa.eu/eurostat/cros/content/harmonised-protection-census-data_en>; see also the Q2018 contribution 'Statistical confidentiality: New initiatives in the European Statistical System', in particular section 2.4.1 [↑](#footnote-ref-11)
12. For details see F Bach 'Statistical Disclosure Control in Geospatial Data: The 2021 EU Census Example', to be published in J Döllner, M Jobst, P Schmitz (Eds.): 'Service Oriented Mapping' (chapter 18), <https://www.springer.com/gp/book/9783319724331> [↑](#footnote-ref-12)
13. Eurostat recently launched another SGA project which will (among other outputs) provide an implementation of the recommended methods inside the Argus software package for data protection, and also in the form of R code as part of the sdcMicro/sdcTable R packages; all software will be open-source and accessible via: <https://github.com/sdcTools> [↑](#footnote-ref-13)
14. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2017/mtg3/2017-UNECE-Standards-Workshop-INSPIRE-SDMX-paper-V2\_\_1\_.pdf [↑](#footnote-ref-14)
15. <http://spatialreference.org/ref/epsg/etrs89-etrs-laea/> [↑](#footnote-ref-15)