United Nations ECE/CES/2016/27

Economic and Social Council

Distr.: General
18 April 2016

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-fourth plenary session
Paris, 27-29 April 2016
Item 6 of the provisional agenda
Geospatial information services based on official statistics

Grid-based results from the 2011 census in Germany: products and use-cases of a new era

Note by the German Federal Statistical Office (Destatis)

Summary

This document discusses the genesis of grid-based census data in Germany. This includes the legislative process up to the publication of results of the 2011 census. It considers the advantages and disadvantages of grid-based data in comparison to administrative geographies and describes products and use-cases. Implications for disclosure control when using grid-based data are highlighted and future directions indicated.

The document is presented to the Conference of European Statisticians' seminar on "Geospatial information services based on official statistics" for discussion.

I. Introduction

1. In the previous century literature on geography one could find expressions like “unfortunately there is no geographical equivalent to the hour, minute or second”, thereby describing the diverse landscape of administrative regional units. Not only do they differ much in size and shape, they are mostly the outcome of historical processes. History does not stop and, therefore, there are not any stable boundaries for administrative units to expect. This leaves official statistics with broken time series and incomparable data.

2. The advent of powerful enough information technology (IT) resources that allow enriching any data point with geo-coordinates has changed the situation dramatically. Now that even individual observations can be precisely located on a map, who would not want to do this? Finally much more objective regional units could be built, namely grids. Europe is fortunate enough that the European Union (EU) came up with the Infrastructure for Spatial Information in Europe (INSPIRE) initiative in 2007, which means member states do not need to build their own grids but that there is a union-wide standardization.

3. However, location based data does not grow in and out of itself and so it took a rather long process until geo-coded statistical results for Germany could be made available. This is the story of how Germany became a provider of grid-based data, what it has learned so far and what is to expect in the near future.

4. The paper will showcase products from the 2011 census and will discuss how well the use-cases so far could reveal insights never before possible. Finally, since Germany shares borders with nine European countries and is itself a federation, the aspect of data provisioning across borders, especially for understanding commute patterns between countries or states has yet to come to fruition. But the underlying frameworks are already in place, namely INSPIRE for the grid definition, several EU directives aiming at comparable geo-coded data and the census being the most prominent example of course.

II. Grid-based results from the 2011 census in Germany

5. People like maps and the more detailed the better. But often it is underestimated that detailed maps based e.g. on the 1-km-grid require a complete enumeration while a lot of official statistics are based on samples. As the Census 2011 in Germany was conducted using a multiple source mixed mode design, some information (e.g. on education), that was only surveyed as part of a household sample cannot be analyzed on the grid level. Nevertheless, the census is a natural fit when it comes to grid level data analysis. And even then in densely populated countries like Germany there are plenty of grid cells with very low absolute numbers, making disclosure control challenging. Now let us imagine you have not done a census in a very long time and your privacy laws once even took down a whole census taking shortly before it was to begin.

6. Germany had not had a census for a quarter of a century, but now that it did take part in the 2011 census round, it strives for state-of-the-art provision of results. Among those are grid-based data products where the census broke ground for updating the German Statistical System with a legal base for geo-referencing. Thanks to the foresight of leading statisticians during the preparatory stages of the first German census after re-unification, all address-data was already geo-coded with precise coordinates that underwent thorough quality assurance. Together with the development on the European level, especially the INSPIRE initiative, Destatis could make the case for appending the legal provisions for disseminating census results with grid references down to the 100 meter level. This regulation now applies to all subject matter areas of official statistics in Germany.

7. To this day grid-based data suffers from the chicken-and-egg problem: Without the availability of publicly accessible statistics on the grid level, the purpose is difficult to prove – and without proving its purpose, resource-allocation for researching grid-based data analysis is difficult to acquire. Moreover a lot of real-world use-cases develop only once appropriate data has been made available to the public. It is our customers who finally show us, where grid-based census results will be used. But it is our task to get the ball rolling.

III. Products and use-cases of a new era

8. Since April 2015, Germany has been offering a variety of census results on the 1 km-grid and population counts even on the 100 m-grid, all freely available for commercial and non-commercial uses under a liberal "attribution only" license. As these data are difficult to grasp in raw numbers – on the 100 m-level (36 million cells for the whole of Germany) it is even beyond what modern spread sheet software can open – a visualisation tool is *de rigueur*.

A. Mapping grid-based data: Choropleth maps

9. Destatis, together with the statistical offices of the Länder, has thus prepared an online atlas showcasing choropleth maps of important demographic and housing indicators for the 1 km-grid[1]. The atlas is built on a tile-based mapping server from the Environmental Systems Research Institute (ESRI) and offers several topographic layers to combine with the grid-based census results. There are other products available that serve tile-based maps on the web but here an already developed solution for a previous atlas on grid-based results for statistics of agriculture has been re-used[2].

10. It may come as a surprise that this grid-based census atlas is not part of the census results database and was published two years after first census results were available. The reason lies in the legislative process. While the census was under way in 2011 it was not yet certain if ever there would be the possibility to use the geo-coded information to publish results on the grid. Until August 2013, the Federal Statistics Law allowed publication of statistical results only for administrative regions or city blocks. After that date the statistical law was finally extended to also include grids down to 100 m side length. As this provision was written into the Federal Statistics Law it applies to all Federal Statistics, if data are available on that level and data confidentiality can be guaranteed by appropriate disclosure control measures. This explains why the grid based census atlas looks a bit different from the other thematic maps. It is an add-on that needed to be developed while the publication process of the census results had already begun.

11. While the uniformity of the grid offers objective comparability, at the same time any resemblance with real world experience is lost. Therefore maps of the census data alone can be interpreted only when they refer to administrative boundaries and as long as the viewer is familiar with the geography at hand. Choropleth maps using uniform grid-cells however need at least some landmarks for interpretation (e.g. Rivers, Motorways) and then, usually once the user zooms in, a background map showing building blocks and road infrastructure is essential. It is important to choose only grey-scaled maps as background so to not interfere with the colour perception of the thematic data layer. Additionally, great care has to be taken in finding the proper level of transparency for the thematic layer to balance interpretation of the thematic layer against readability of the background map. If the

[1] https://atlas.zensus2011.de

[2] see http://www.atlas-agrarstatistik.nrw.de/ for the atlas on grid-based agriculture statistics

Geographic Information Systems (GIS) in use offers the 'multiply' blend mode, this has proven to serve both needs best.

12. A lot of challenges of thematic mapping in general apply to grid based maps as well: Classification of the data, appropriate colour schemes and when it comes to map servers, a robust infrastructure that can handle peak demands. Where mapping of grid based data differs in this process is the dependency of disclosure control and classification of the data. To maximise the areas for which colour-values can be shown, it was decided not to display exact values when hovering over a grid cell. Displaying only ranges of data – as is common with choropleth maps – works in favour of the disclosure control process. In fact the classification of the data had to be optimized not only in terms of appropriate differentiation (i.e. "Jenks Natural Breaks") but also to minimize the number of grid cells that cannot be displayed at all due to data confidentiality.

13. GIS are a niche and expert domain but the barriers to entry have been lowered significantly due to the widespread availability of powerful open source software (e.g. QGis). With the results from the 2011 census we saw demand from data journalists that are capable of building sophisticated visualisations and dealing with GIS data as part of their toolbox. To accommodate external GIS users, the data that is shown in the grid-based census atlas is also available as downloadable .csv files[3]. Here data are provided in two different specifications. On the one hand with the data-ranges identical to the map and on the other with precise values. The latter is affected a lot more by the disclosure control process, i.e. there are more grid cells where the value had to be suppressed for confidentiality reasons. So journalists or other interested parties can classify the gird-based indicators according to their needs but will have less grid-cells to work with in the process. All maps from the atlas are also provided as a Web Mapping Service (WMS) so that GIS users can easily add grid-based information of the census to their projects in cases where they can work with the provided data classification.

14. Given the diverse and multi-level administrative structure of Germany, the value of small and comparable output units that are available in densely populated urban areas as well as in rural ones, all at the same quality level, is yet to be fully appreciated. Once a newspaper described the grid-based maps as looking "pixelated" which they are of course, but it was not meant as a compliment. Given that a lot of political decisions refer to administrative boundaries, providing data for those and mapping them will probably always be the main regional resolution.

B. Aggregating grid-cells to provide data for arbitrary geographies

15. Beyond mapping census data on the 1km grid, a second use-case – that might turn out to be even more important – is aggregating 100 m grid cells of census data for any number of areas. Already noise emissions or other pollutants along airports or rail/road infrastructure are known to not spread along administrative boundaries. Those readings can be mapped and areas of similar pollution can then be joined with 100 m grid cells like pixels filling out arbitrary shapes. The affected population or even sub-populations (age, gender, etc.) and dwellings can be calculated for those areas very economically. At the same time those aggregates are large enough that any deviation due to confidentiality measures can be neglected. Aggregation on such a low level can also help fitting data to changes in administrative boundaries over time.

[3] https://www.zensus2011.de/SharedDocs/Aktuelles/Ergebnisse/zensusAtlas.html

16. Disclosure control makes it necessary that the aggregation process has to take place with the original data before disclosure control measure have been applied. Therefore, this can only be done inside statistical offices. In a typical scenario an engineering office may send a shapefile containing areas with similar noise levels around an airport that is a result of their measurements. The statistical office can then provide totals of census variables for the areas provided, be they population counts, demographic breakdowns or anything from the dwelling census. The whole process is tailor-made for each inquiry as it involves processing data in the Statistical Analysis Software (SAS) and ArcGIS and usually also involves several consultations with the customer.

17. Since these analyses will get charged on an hourly basis they are not available to all. It remains to be seen how much development in web based GIS systems with support for geographic objects in relational databases will take place so that it may be feasible to offer such calculations on a self-serving basis on the web. As an appetizer in this regard the German state of North Rhine-Westphalia published a geographic population calculator, where one can digitise simple geographic shapes (circles, polygons) on a map and calculate the aggregate population count within the digitised area[4]. This capability will be added to the grid based Census Atlas of Germany later in 2016. However this will not allow uploading one's own shapefile and for reasons of the disclosure control process it can only aggregate population counts and not demographic breakdowns.

18. Finally using grid-cells to approximate new geographies and therefore providing data for longer stretches of time than the stability of boundaries, will be a task for the next census. Shortly after census taking in Germany (9 May 2011) an amalgamation of communes took place in the German State of Mecklenburg-Western Pomerania (4 September 2011) and there are probably more to come until 2021.

IV. Conclusions and recommendations

19. So far working with grid based data from the 2011 census has been very well received. Many customers are drawn to it because of the newness and the promising possibilities it offers. Many of its advantages – like stability over time – are yet to reveal its potential in coming census rounds.

20. Dealing with grid-based census data are currently limited to special use cases and GIS specialists which most statisticians are not. Statistical offices need to cultivate the demand for such data. Example applications, like online Atlases, ideally with already integrated simple GIS tools, are of the essence.

21. Storing such data and making it accessible to users is another area of research. Destatis published a limited amount of data based on the INSPIRE grid as simple .csv downloads and fulfils many customized inquiries that will get charged.

22. So far we are in a phase of making grid-based data available and promote possible use-cases. It will take a few years until we can evaluate if and what kind of insights could be derived from grid-based data that could not with previous geographies like NUTS and LAU levels. We see a lot of grid-based maps that only reproduce population density patterns more or less. Instead mapping starts to reveal insights, when patterns become visible which do not follow the mere population distribution.

23. We see a huge potential in analyzing nodal regions which are intersected by administrative boundaries. Those might be the transitional areas between city and suburb or

[4] http://www.einwohner.nrw.de/

suburb and rural but can also be commuter regions across national borders. Here peoples' daily activities are spread over different administrative areas that themselves are very heterogeneous in shape and size. In other words, grid based data shines where administrative boundaries do not reflect the social land-use (any more). Similarly pollution or natural hazards also do not follow administrative boundaries in their spreading and the affected population can best be approximated using grid-cells.

24. Unfortunately, the grid may feel foreign, especially to the lay person. Visiting the prime meridian in Greenwich might be an accepted tourist activity and the nerdier amongst us might have heard of scavenger hunts for the intersections of integer latitude and longitude degree lines, but the INSPIRE grid on 1 km or 100 m resolution certainly bears no similarity in this respect.

25. Finally, there is a reason administrative areas are called this way: A lot of political decision making, namely collecting taxes and spending them, are restricted to exactly their administrative boundaries. The placement of industries or shopping malls quite often does not follow some central place theory of our geography classes but rather the boundaries of tax incentives.

26. There could not be a more exciting time when it comes to regional statistics: The census offers data on all regional levels, be they administrative or grid, and tools to explore them are freely available. Grid-based data certainly needs public relations efforts but as a young and data-hungry generation of journalists enters an Open-Data world, it will soon become a fast-selling item. Pairing the right choice of regional resolution with the appropriate investigative problem will lead to successful use-cases.

27. The next frontier is the timeliness of grid based data that stems from full enumerations and more frequently from registers.