A Cartography of the Effects of Disclosure Control on Small-scale Population Grids in the Context of Municipal Data Requirements in Northwest Germany.

Specht S*[1]

[1] OFFIS e.V. Oldenburg, Germany

January 17, 2022

Summary

Population grids, especially in rural areas, are vulnerable to the effects of statistic disclosure control methods. The paper proposes a way to map those effects and create awareness for users involved in the creation of those data sets. To achieve this, the difference between unaffected and protected grid cells is calculated and is being visualized. The effects of spatial aggregation and level of attribute details are being compared.

KEYWORDS: demography, population grid, statistical disclosure control, spatial aggregation

1. Background

Since their introduction in the German census in 2011, population data in equal-area cartographic grid cells were adopted as basis for planning, administration and analysis (Kaup & Riefel, 2013; Sigismund, 2014). As regular, time series-stable spatial units without historical genesis, as is the case with administrative areas, grid cells simplify comparisons between different municipalities and regions.

As a result of the WEBWiKo research project (Specht, Blohm, Handtke & Wette 2019), data on population size and migration based on micro-data from the registration offices of several neighbouring municipalities was transferred into a data warehouse for use by municipal planners. This data can only be handled anonymously and in aggregated form. From a spatial perspective, high-resolution demographic data, especially in sparsely populated areas, can generally be expected to show high before-and-after deviations through confidentiality procedures. The sum of the differences may be small in relation to the total number of the population, but can be very significant in relation to the distribution of the population. Not all stakeholders involved in the project were aware of this effect and the consequences of their data related requirements.

Meindl & Templ (2019) have presented SdcApp, an interactive user interface for the statistical confidentiality library "sdcMicro" implemented in the programming language R, which supports statistical non-disclosure tasks and their assessment. Unfortunately, a cartographic perspective is missing.

2. Aim

This article seeks to contribute to a decidedly spatial comparison of the effects of aggregation and confidentiality strategies. It is focused on their impact on the spatial distribution of the population.

3. Effects of confidentiality methods on spatial data

Confidentiality problems occur when the cells of data tables contain values that are formed from the characteristics of only one or two statistical objects (in maps: the values displayed on the basis of the areas) or within a data table column or row aggregates are formed from only one cell (in maps: the

* sebastian.specht@offis.de

number of areas displayed in a particular query or map). There are still significantly more complex secrecy issues (dominance problems, re-identification problems etc.) which are not discussed here. The most straightforward traditional method for suppressing small numbers of cases is cell blocking, which often amounts to a high loss of spatial information in sparsely populated areas. Spatial aggregation is an additional information-reducing method and will be considered below in the context of grid cells. Other existing approaches (rounding, quantisation, stochastic noise, spatial swapping methods, or smoothing) are not addressed here in an evaluated manner.

4. Cartographic visualisation of the confidentiality effects through subtraction

Within the scope of the research project, municipal data on population status and population movements were transferred to a data warehouse on the basis of personal microdata (B) from the residents' registration offices. The spatial information in the form of addresses of the registered persons was converted into geo-coordinates with the help of a geocoder. According to the desired characteristics of the target data set, these individual cases are being aggregated at the most detailed attribute level.

$$T = f(B) \tag{1}$$

The resulting aggregated table (T) does not contain personal data, but the number of persons aggregated according to target attributes. Occasionally, grid cells or regions may be filled with individual cases. Therefore, prior to any form of further use, a publishable table (P) must be created. For this purpose, the aggregated table (T) is re-aggregated and cleansed of individual cases by means of a confidentiality function (g).

$$P = g(T) \tag{2}$$

An evaluation of the confidentiality effects can be carried out after this step. In order to be able to assess the spatial effect of the confidentiality function, the difference between the aggregated and adjusted cell value is determined at the grid cell level for each available feature expression. The differences are summed up for each grid cell and are referred to as residuals (R_i) hereafter.

$$R_i = \sum_{j=1}^{k} (T_{i,j} - P_{i,j}) \tag{3}$$

with *i* = populated grid cell, *j* = feature expression, *k* = number of feature expressions

The residuals of the spatial units are set in relation to the respective sum of unchanged feature values and expressed as a residual ratio (Rv_i).

$$Rv_i = \frac{R_i}{\sum_{j=1}^{k} (T_{i,j})} \tag{4}$$

Beyond the cartographic visualisation of the spatial distribution of the confidentiality function's effects, an estimation of the effects for the entire considered space is of interest. A residual ratio (Rv) is formed in relation to the population size of the entire space.

$$Rv = \frac{\sum R_i}{\sum T_i} \tag{5}$$

The quotient (Fv) of the number of grid cells with a residual ratio Rv_i > 0.1 divided by the total number of inhabited grid cells is formed for the purpose of a first assessment of the affected area in relation to the entire region.

$$Fv = \frac{Number\ of\ Cells\ with\ Rv_i > 0{,}1}{Number\ of\ Populated\ Cells} \tag{6}$$

5. A cartography of the effects of confidentiality

The maps in **Figure 1** show the population density of a 120 km² area of the Lower Saxony/Bremen region and are overlaid with a visualisation of the local residual ratio. The area can be identified as a rural area on a transition to an urban area due to the many unpopulated cells (boundary of municipality in black). To demonstrate the approach, a simple cell blocking with threshold 5 is used as an approximation to a secrecy function in order. The rows of the map matrix show the effects of the secrecy functions on grid cells of different side lengths (1,000 m, 500 m and 100 m). The columns show the effects on feature dimensions with different levels of detail (two gender attributes vs. 20 age groups using a 5-year span).

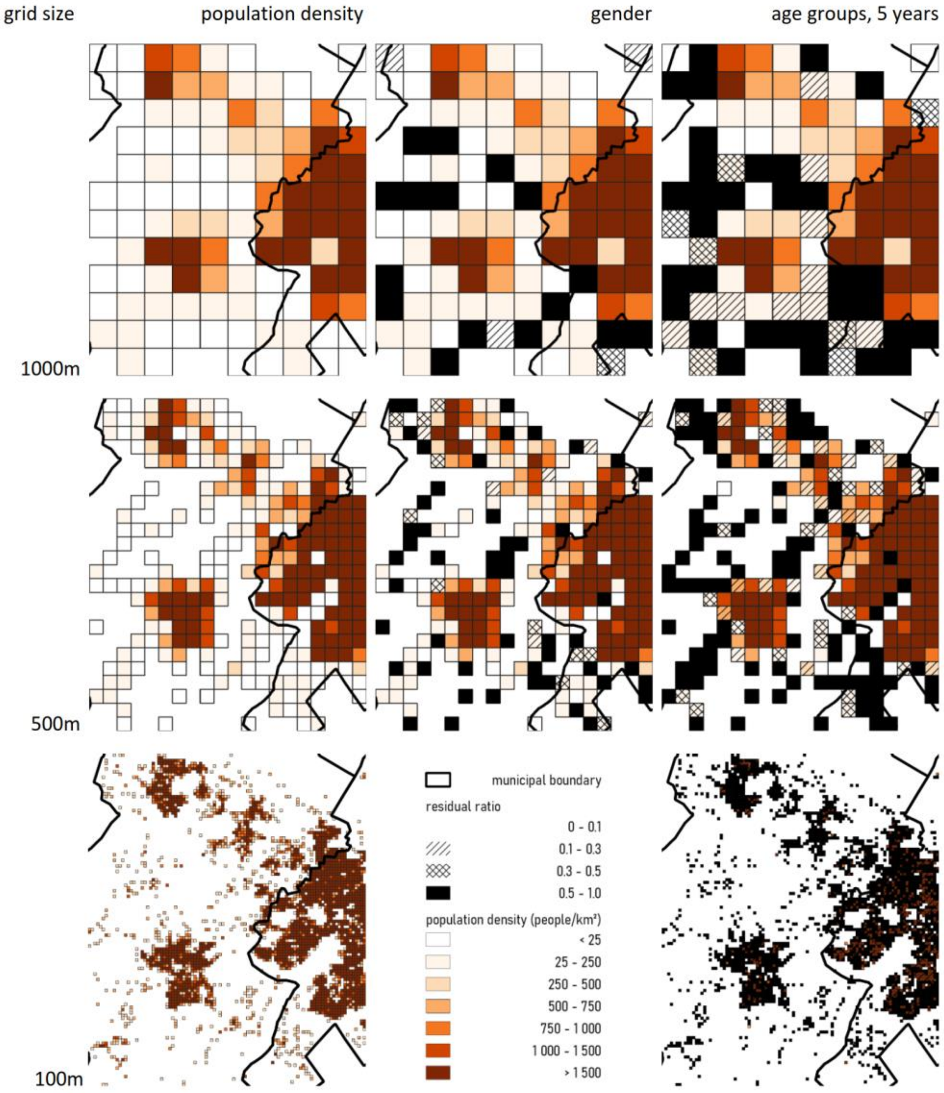

Figure 1 Effect of the confidentiality function (cell blocking with threshold value 5) on the grid cells of different edge length. (Own work, source: Research data from the WEBWiko project. Differentiation by gender is not available for the data in the 100 m grid).

All cells with a residual ratio Rv_i > 0.1 are shaded with different densities. A residual ratio of 0.1 means that 10% of the population value in such a grid cell is affected by the confidentiality function. The maps show: the more detailed the space and the feature dimension, the higher the change due to protection processes with sparsely populated areas being most affected. The extent of this relationship is shown in **Table 1**. However, the relationship between detail (grid size, granularity of characteristics) on the one hand and the number of inhabitants (Rv) and area (Fv) affected on the other hand is somewhat less to be expected. For instance, at the "1,000m age groups" level, only 1% of the number of inhabitants is affected, but almost 49% of the inhabited area. Despite the fourfold resolution of the 500m grid, the affected area is of a similar order of magnitude (Fv = 0.538) with still less than 3% of the population affected (Rv = 0.024).

Tabelle 1 Effect of the confidentiality method "cell blocking with threshold value 5" on the entire space (* for Rv_i > 0.1, ** different number of inhabitants due to varying data collection methods).

map	populated cells	...with residuals *	population **	residual	$\boldsymbol{Rv}$	$\boldsymbol{Fv}$
100m, age groups	2 829	2 802	73 328	43 835	0.598	0.990
500m, age groups	262	141	71 201	1 686	0.024	0.538
500m, sex	262	68	71 201	231	0.003	0.260
1000m, age groups	101	49	71 201	682	0.010	0.485
1000m, sex	101	20	71 201	72	0.001	0.198

6. Outlook

Of course, the basic relationship of "the more detailed the space and features, the higher the inaccuracy due to protection processes" is easily understood by municipal data users and probably trivial. The chance of this visualisation technique is mainly to make users aware of the effects of non-disclosure processes and their localisation in their specific region. In a data requirements assessment, the municipal experts can formulate their needs for the level of detail, assess the consequences for non-disclosure and possibly reformulate their requirements. The most important levers to minimise the impact, regardless of the confidentiality procedure used, remain the reduction of the attributes and their properties to what is actually necessary.

References

Kaup, S. & Rieffel, P. (2013): Rasterbasierte Regionalstatistik. *ILS-TRENDS*, (2). Dortmund.

Meindl, B. & Templ, M. (2019): Feedback-based integration of the whole process of data anonymization in a graphical interface. *Algorithms*, 12(9), 1-20. https://doi.org/10.3390/a12090191

Sigismund, M. (2014): KLASTER – Kleinräumiges Analyseraster für den Zensus. *IÖR Schriften*, 65, 159-167.

Specht, S., Blohm, K., Handtke, T., & Wette, L. (2019): Prognosen im Bevölkerungsraster für die interkommunale Kooperation – ein Experiment im Reallabor. *AGIT – Journal Für Angewandte Geoinformatik*, 5-2019, (5), 284–291. https://doi.org/doi:10.14627/537669027

Acknowledgements

The Author would like to thank Bernd Kramer and Andreas Lemke from our WEBWiKo project partner "regio gmbh" for making the analysis infrastructure available.

Biographies

Sebastian Specht studied cartography (1996 to 2000) and Digital Media (2004 to 2007, M.Sc.) and is working as a Data Science Software Engineer at OFFIS Oldenburg, department Data Management and Analysis for Health Services Research.