"Environmental Engineering" 10th International Conference
Vilnius Gediminas Technical University
Lithuania, 27–28 April 2017
http://enviro.vgtu.lt

eISSN 2029-7092 / eISBN 978-609-476-044-0

Article ID: enviro.2017.221
DOI: https://doi.org/10.3846/enviro.2017.221

Uncertainty Quantification of the Global Rural-Urban Mapping Project over Polish Census Data

Joanna Nowak Da Costa[1], Elzbieta Bielecka[2], Beata Calka[3]

Military University of Technology, Faculty of Civil Engeneering and Geodesy, Warszawa, Poland
E-mails: [1]joanna.nowakdc@wat.edu.pl (corresponding author); [2]elzbieta.bielecka@wat.edu.pl; [3]beata.calka@wat.edu.pl

Abstract. The aim of this study is to describe uncertainty of the Global Rural-Urban Mapping Project (GRUMP) data based on Polish population reference grid created by the Central Statistical Office of Poland, using INSPIRE grid coding system. The adopted population data uncertainty analysis methodology combined three different approaches, i.e. simple change detection algorithm to obtain discrepancies at the grid cell level, statistical analytical approach to investigate these discrepancies' frequency distribution, and GIS approach to analyse spatial pattern of distinguished population difference classes. The results showed significant differences in population count at the grid cell level. The maximum magnitude of GRUMP vs. Polish Reference Grid overestimation equals 4087 people per 1 sq. km, while the underestimation equals 20,086 people per 1 sq. km. Very few grid cell shows no difference in population count, i.e. 1.5% of total grid cell count. GRUMP data overestimates Polish total population by 0.15%, while it underestimates the average population density by 50%. The highest population underestimations were identified in the centers of the cities, while suburban areas were characterised by the large and regular population overestimations within GRUMP dataset. These GRUMP dataset imperfections can be attributed to country-specific administrative divisions and to the varying effectiveness of the urban centers delimitation mapping using the night sky light intensity, including blooming effects as well as not frequently illuminated small settlements.

Keywords: GRUMP, data uncertainty, population density, grid data, census population data.

Conference topic: Technologies of geodesy and cadastre (mapping technologies).

Introduction

With the spread of spatial information and popularization of GIS technology in business, strategic or public administration's decision-making, the issue of data quality and reliability started to be essential since data uncertainty propagates through spatial analysis (Heuvelink *et al.* 1989) and affects the final result and decisions-making process. Trends in population distribution and urbanization, in particular urban and rural population identification, are crucial in sustainable development and urban planning, healthcare development, crisis management and many others. Traditionally, population data are obtained during the census of population and housing performed by National Census Organizations on regular basis, usually every few years. Such detailed census data is subsequently related to predetermined statistical units, most often corresponding to the administrative division, and only then – published. In many applications this approach to modeling the population distribution is insufficient due to administrative units division variability and changeability (within time), and due to imposing of continuous geographical phenomena (population) into artificial spatial units. The latter issue is known in geography and spatial analysis as the modifiable areal unit problem (MAUP).

To diminish the shortcomings of representing population distribution in administrative units, scientists have developed a number of dasymetric modeling approaches that allow for population distribution representation into reporting units individually tailored to population density, and they elaborated some algorithms to calculate population within any reference unit. Regrettably, fine-scale population data availability is still very limited both on local and country or global level. Therefore the Global Rural-Urban Mapping Project (GRUMP), recently became publicly available free of charge, gained a lots of interests, e.g. health applications (Balk *et al.* 2006). However, only a few GRUMP quality evaluation exists, and none of them encompass Poland. Therefore, the rationale for this research is the quantification of uncertainty of GRUMP dataset over Polish territory. The comprehensive uncertainty study was proposed based on image differencing algorithm and GIS layers overlaying analysis. This is a step forward to GRUMP data usability and fit for purpose studies.

Methods of uncertainty quantification

Fundamentals of data quality evaluation

Diversity, variability and complexity of the real world obstructs its direct analysis. Abstracting the geographic reality, the useful and manageable geographic database can be created. However, there is a discrepancy between geographic

database and the geographic reality that the data are intended to represent, and uncertainty analysis assesses it. Goodchild *et al*. (1994) defines spatial data uncertainty notion as similar to data error one, however they differ in that uncertainty is a relative measure of the discrepancy while error tends to measure the value of the discrepancy. While according to ISO 5725-1 standard (1994), uncertainty is the component of a reported value that characterizes the range of values within which the true value is asserted to lie.

Probably the most obvious geographical feature uncertainty concerns its location in space and time. However, precise and accurate located set of features not accompanying with their proper attributes (attributive information) is practically valueless, and constitute descriptive and thematic uncertainty. The above mentioned uncertainties may arise during geographical feature measurement and cataloguing process. While vague definitions or ambiguous meaning may result in data uncertainty of conceptual origin. Let's take an example of a building. It can be defined as a roofed and walled construction or an open-wall building spanning a ground-level, and therefore it may or may not serve as a permanent shelter to protect persons or animals (Nowak Da Costa 2016b).

Although concepts of uncertainty in spatial data and spatial data quality are similar, the standards have been evolved only for the latter one, e.g. ISO 19157:2013 Data quality (ISO 2013) and ISO/TS 19158:2012 Quality assurance of data supply (ISO 2012). The issue how to evaluate and describe data uncertainty is still open, nevertheless data users are mostly interested in knowing if data fit their needs (Bielecka *et al.* 2014).

GRUMP uncertainty analysis and quantification approach

The proposed research involves execution of the following three research tasks: (1) to study uncertainty and understand the process of how uncertainty arises in GRUMP data; (2) to quantify uncertainty in GRUMP data; and (3) to visually represent uncertainty in an efficient and user-friendly way.

Due to availability of quality ground truth data, i.e. the Polish National Population Reference Grid (thereafter referred to as NPRG), and data format analogy between image scene and population grid, the authors employed a simple change detection algorithm (image differencing) to study and quantify GRUMP dataset uncertainty (direct evaluation). The *grid base population difference index* I_{PD} was calculated using the following formula:

$$I_{PD} = v_{NRPG_{ij}} - v_{GRUMP_{ij}}, \quad (1)$$

where: $v_{NRPG_{ij}}$ – population counts in National Reference Population Grid (NRPG), *ij* – position of grid cell in a grid matrix, and $v_{GRUMP_{ij}}$ – population counts in GRUMP data.

The resulting grid, namely Population Difference Grid (PDG), represents intensity of difference between population values attributed to grid cell of the same geographical location by Central Statistical Office of Poland (CSO) in Poland (actual population) and GRUMP (modelled population). Positive values of PDG represent GRUMP underestimation of population, while negative values represent overestimation of population counts with respect to national census data. To examine the discrepancies between two population datasets and their frequency distribution, both general analytical approach and graphic techniques were employed, i.e. basic descriptive statistical parameters of Population Discrepancy Grid together with its histogram and Normal Quantile-Quantile plot.

Due to nature of gridded population, its uncertainty refers also indirectly to vagueness of representation of phenomena that is continuous in its nature. In this case uncertainty may exists in both the population distribution model specification (here: a proportional allocation gridding algorithm) and the data used in the model application (e.g. data for statistical reporting units, the project collected population estimates, the approximate footprint for urban centers in each country). Therefore the knowledge of the GRUMP production process was also used to study and estimate its uncertainty (indirect evaluation). Spatial distribution of PDG difference classes was analysed using GIS approach, i.e. grid (PDG) and vector data (e.g. land uses) overlaying, since urban land use layer was used to create GRUMP dataset.

Finally, the cartographic visualization of the grid base population change difference index was proposed in a form of a choropleth maps. The presented approach conforms to the Goodchild's comprehension of geographic data uncertainty: a map depicting varying degrees of uncertainty associated with each of the features or phenomena represented in the data set (Goodchild *et al.* 1994).

Area and data used

Population and census data in Poland

The last three censuses of population and housing in Poland were carried out in: 1995, 2002, 2011. The changes in population totals between census years are minor and respectively equal: 0.39 thousands for 1995–2002 and 0.28 for 2002–2011 (see Table 1). It constitutes the decrease of approximately 1% in the 1995–2002 period, and increase of population of 0.7% during the 2002–2011 period. Population in urban area (oscillating between 61–62%) refer to those localities which have the status of towns.

Table 1. Population in Poland

Population* [in thousands]	1995	2002	2011
Total population	38,620	38,230	38,512
Population in urban areas (percentage of total population)	23,777 (61.6%)	23,610 (61.8%)	23,406 (60.8%)

*Based on CSO data in according to census 1995, 2002, 2011 (Demographic yearbook of Poland 2015)

National Reference Population GRID has been created in 2011, on the bases of 2011 census data, according to 'INSPIRE Data Specification on Population Distribution – Technical Guidelines v3.0'. The 1 km grid in Lambert Azimuthal Equal Area coordination system (LAEA) has been created following the requirements of COMMISSION REGULATION (EU) No 1089/2010 of 23 November 2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services and the INSPIRE Specification on Geographical Grid Systems – Guidelines (INSPIRE 2010).

GRUMP v1

The first major effort to generate a consistent global georeferenced population dataset was succeeded in 1995 with the creation of the Gridded Population of the World (GWP) by the National Center for Geographic Information Analysis (NCGIA), and subsequently by Columbia University Center for International Earth Science Information Network (CIESIN) – 2000 year update (Deichmann *et al.* 2001; Tobler *et al.* 1997; Balk, Yetman 2004; CIESIN 2005). GWP together with urban extents – as identified through NOAA's nighttime lights satellite imagery, settlement points (from Digital Chart of the World's Populated Places), administrative boundaries vector data, and census population data on low-level administrative division were the basis for the development of the next global dataset targeting at differentiation of population distribution between urban and rural areas, the Global Rural-Urban Mapping Project (GRUMP). The mass-conserving algorithm, applied iteratively, redistributed people into urban areas, within each administrative unit (Deichmann 1996; Balk *et al.* 2006).

GRUMP provides a 30 arc-second raster representation of estimates of human population for the three reference years, namely: 1990, 1995, and 2000. The comprehensive analysis of GRUMP data quality is provided by Hall *et al.* (2012) for Sweden. The authors compare gridded population data for part of Sweden (Scania) with high-resolution population records obtained from the Swedish National Registry. Surprisingly, the GRUMP data does not preserve population counts in the region. Moreover, it overestimate population in cities and underestimate in the transition zone between urban and rural areas. Similar conclusions are reported by Linard *et al.* (2010) for Somalia. The urban areas are wider in spatial extend than in reality due to the blooming effect on the NOAA's scenes. Das (2013) have noticed that the GRUMP data encompass the whole metropolitan urban area and thus clearly delineates urban and rural boundaries, and it has potential to provide an urban extent mask for some applications.

Results and discussion

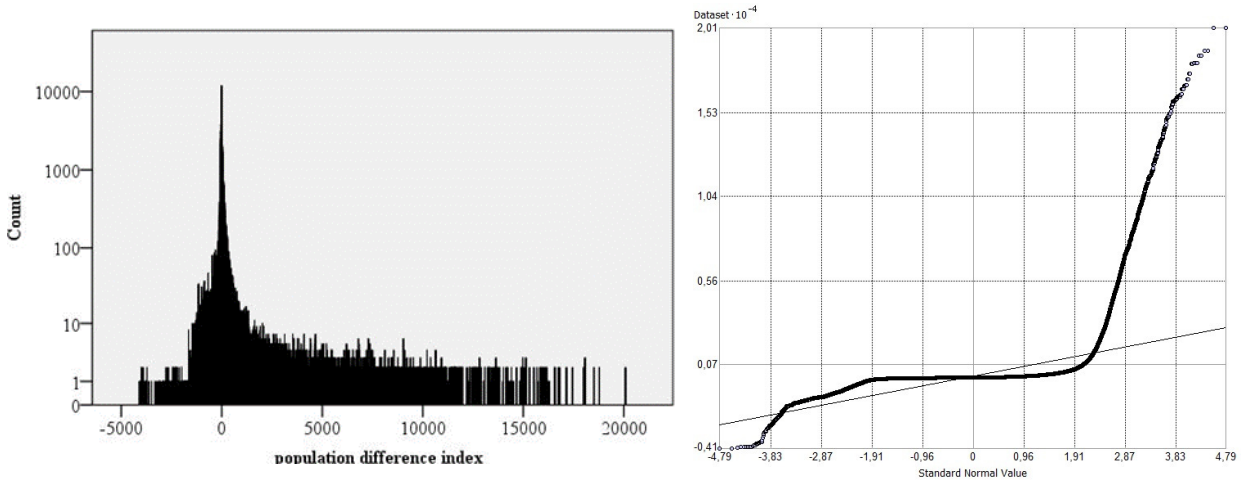
Population data comparative analysis

Basic summary statistics, calculated for Polish statistical (NRPG) data and GRUMP data, reveal some significant differences between them (Table 2). Standard deviation and variation show that variability in population counts in grid cells is more homogeneous in GRUMP than in NRPG, however median value is more than twice higher. Mode value suggests the overestimation of population in GRUMP data, even though total overestimation is only 0.15% (58,797 people). An overview of the nature of the relationship between these two resources is given by a relatively small R^2, coefficient of determination, equal to 0.28.

Table 2. Summary of descriptive statistics for NRPG and GRUMP population datasets

	Total sum	Max	Min	Average	Median	Mode	Standard deviation	Variation
NRPG	38,511,800	21531	0	123	12	0	658	432,661
GRUMP	38,569,597	4091	0	65	28	20	181	32,761

The examination of I_{PD} frequency distribution (Fig. 1a) reveals that only a few grid cells (1.5%) show no differences in population. The values on the right side of the histogram represent GRUMP population underestimation and they are not symmetric to the lying on the left (overestimation).



(a) (b)

Fig. 1. Frequency distribution (a) and Normal QQ plot (b) for Population Difference Grid

Both histogram shape (Fig. 1a) and PDG spread and shape measures (Table 3) indicate that *population difference index* I_{PD} is not normally distributed. It is confirmed through the normal QQ plot (Fig. 1b) where points, representing the I_{PD} variable, deviate from the reference line.

Table 3. Summary of descriptive statistics for Population Difference Grid (PDG)

dataset	Cell count	Max	Min	Average	Median	Std. dev.	Skewness	Kurtosis	1-st / 3-rd Quartile
PDG	589,861	20,086	–4,087	57.1	–11	583.3	13.6	250.6	–26 / 25

Deviations from the normal distribution are understood as an indication of existence of systematic character biases in the PDG dataset. It should be noted that distribution of the large values of population difference differ most significantly from the normal distribution.

The complex relationship between the NRPG and GRUMP datasets was further analysed superimposing Population Discrepancy Grid and vector data of land use, administrative boundaries, road network and forests. The study provided the following information on change (difference) rate and spatial distribution of difference classes (categories): large population underestimation mostly correspond to city centers, while the city suburbs population is often overestimated (Fig. 2). Moreover, there is noticeable correspondence between high value of the grid base population difference index and the residential settlements along major roads connecting towns and cities.

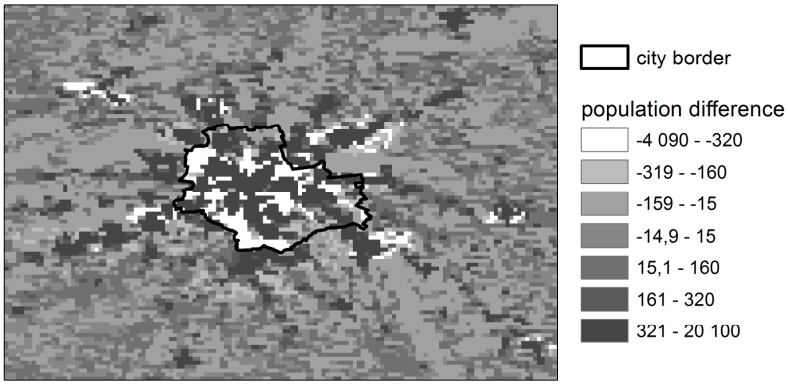


Fig. 2. An example of PDG cell value dependence on their geographical locations: underestimation in city center and overestimation in the suburbs

Frequent GRUMP population overestimation was also noticed in a grid cells that are within the territory of huge forests, especially if they adjoin big cites.

Uncertainty quantification

Several indexes were tested to quantify GRUMP uncertainty and to establish the thresholds. However, mainly due to particularly complex relation between two datasets, the issue is still open. Nevetheless, the authors proposed to way to distinguish the various degrees oof GRUMP population data uncertainty as follows. The average population in Poland was chosen as a meaningful threshold as regards one square kilometer grid population data uncertainty. The GRUMP population cell grids that do not differ more than ±20% from that value (Fig. 3, marked in grey), are considered as data of very low uncertainty (reliable data). They constitute 38.8% of all grid cells and they are located mostly in rural regions of Poland. The two cell classes characterised by the range of 20 to 100% of an average value of census population are considered as having medium uncertainty. The positive range class (Fig. 3, dark grey) encompasses 16.9% of Polish territory: dispersed small areas all over the country area, with a bit more instances in the central Poland. While the respective negative range class (light grey) constitutes a big fraction of 31.5% of the whole dataset, and covers mostly central and south part of Poland. The two last population difference classes refer to hugely underestimated (black, 9.7% of Polish territory) or overestimated grid cells (white, 3.1%) that lie respectively within city centers or its suburbs, and they both are described as highly uncertain GRUMP population data.

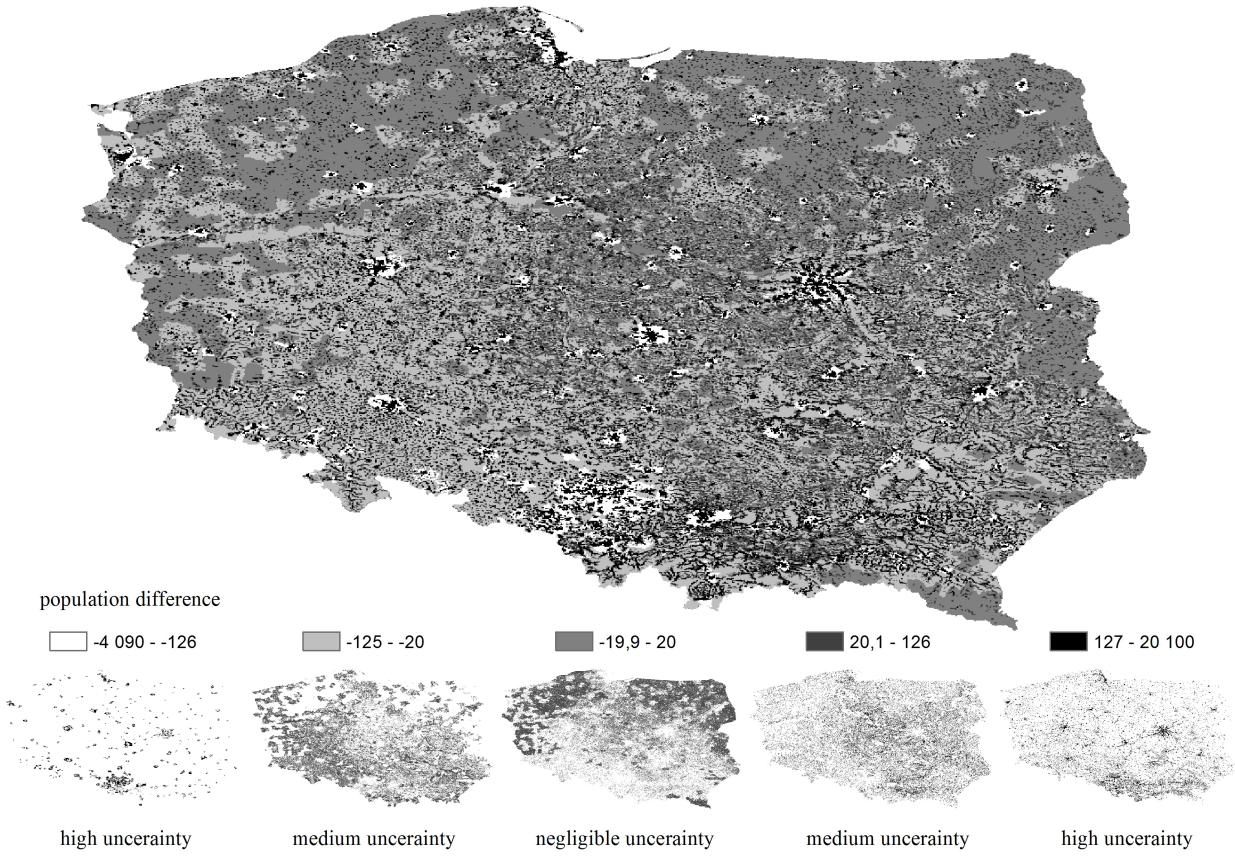


Fig. 3. Varying degrees of GRUM population data uncertainty (choropleth map, arbitrary manual class choice method)

These uncertainty levels may serve also as the choropleth map classes (Fig. 3). Therefore this final map visualises varying degrees of GRUMP population data uncertainty. Moreover, the geographic distribution of the population discrepancies reflects the characteristics of the algorithm used for GRUMP dataset creation.

Discussion

Many scientists, including (Nowak Da Costa 2016a, 2016b), underline that a proper comparison of different datasets requires thematic data semantic similarity assurance between corresponding entities in these databases. The studied NRPG and GRUMP population datasets show resemblance in grid cell size, however they dffer as far as the temporal validity is concerned (2000 vs. 2011 year). Moreover, the GRUMP population count derivation algorithm differs significantly from that used by CSO. The GRUMP population grid uses a proportional allocation rule (dasymetric mass-preserving algorithm), while CSO aggregate people based on their address points. These semantic discrepancies may affect the final analytical results.

The known and important GRUMP's limitation is an urban extents overestimation due to blooming effect on the night-time lights satellite scenes (Balk *et al.* 2006; Elvidge 1997, 2001). This characteristic is confirmed by the presented research, i.e. population overestimation in the city suburbs.

Due to different climate, economic and cultural aspects, different worlds regions are charecterised by different light use habits. These influence especially the rural settlments or less-densely populated areas that may not be detected on the night-time lights satellite scenes. The study confirmed the frequent GRUMP population underestimation in less-populated areas and rural regions of Poland.

Balk *et al.* (2006) also draw attention to the fact that urban extent were delineated five years earlier (i.e. 1994–1995) than the GRUMP dataset was created and other auxiliary data is dated (i.e. 2000). This inconsistence should also be a subject of further research. However, according to Corine Land Cover database, the changes in Polish built-up area, between 1994–2000 and 2000–2006, are very small and do not exceed 0.05% of Polish territory (Bielecka, Ciolkosz 2008). These changes mostly relate a built-up areas compaction than urban sprawl.

The other time validity issue to be tackle is the difference between the total Polish population year of reference for the National Reference Population GRID, i.e. 2011, and the Global Rural-Urban Mapping Project population grid, i.e. 2000. Due to a relatively stable population count in Poland, i.e the average annual rate of population change do not exceed 0.025% (compare tab.1), the total population differences between 2000 and 2011 are negligible.

Conclusions

The authors studied uncertainty of the Global Rural-Urban Mapping Project population count dataset in rela-tion to the Polish National Reference Population GRID over Polish teritory. The adopted data uncertainty analysis methodology combined three different approaches, i.e. simple change detection algorithm to obtain discrepancies at the grid cell level, statistical analytical approach to investigate these discrepancies' frequency distribution, and GIS approach to analyse spatial pattern of distinguished population difference classes. The authors found only 5,000 grid cells (1.5% of total) that showed no difference in population count and the ex-istence of systematic character biases in the Population Difference Grid (difference between the reference and the studied population data) were acknowledged.

The analysis of the spatial distribution of distinguished population difference classes, confirmed the results of previous studies: underestimations in the city centres and overestimations in the suburbs. Additionaly the frequent population underestimation along main roads was noticed. Finaly the varying degrees of GRUM population data uncertainty were defined and visualised on a choropleth map (Fig. 3).

The authors formulated some recommendations as for the further use and fit-for-purpose of the GRUMP population data as follows. The GRUMP dataset can be successfully used for the areas depicted as low uncertainty on the final map (Fig. 3) as the uncertainty associated with them is comparable with those of the national reference data. For other areas the GRUMP population dataset can be used only if higher quality data are unavailable, except if GRUMP is used for studies where the reference area constitutes a territorial unit above NUTS level 4 administrative division. This can be attributed to the specifics of the algorithm used for GRUMP dataset creation: the bigger study area the smaller GRUMP population count uncertainty.

Acknowledgements

The research would not have been possible without the access to the data hosted by the Center for International Earth Science Information Network – CIESIN – Columbia University, International Food Policy Research Institute – IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical – CIAT. 2011. Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Density Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4R20Z93. Accessed 12/03/2016

Funding

The study has been conducted under statutory research at the Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geodesy, No 933/2016

Disclosure statement

The authors declare no conflict of interest.

References

Balk, D.; Yetman, G. 2004. *The global distribution of population: evaluating the gains in resolution refinement, draft documentation for GPW v3* [online], [cited 22 August 2017]. Available from Internet: from http://beta.sedac.ciesin.columbia.edu/gpw

Balk, D. L.; Deichmann, U.; Yetman, G.; Pozzi, F.; Hay, S. I.; Nelson, A. 2006. Determining global population distribution: methods, applications and data, *Adv Parasitol.* 62: 119–156. https://doi.org/10.1016/S0065-308X(05)62004-0

Bielecka, E.; Ciołkosz, A. 2008. Land use mapping in Poland, *Geodesy and Cartography* 57(1): 21–29.

Bielecka, E.; Leszczynska, M.; Hall, P. 2014. User perspective on geospatial data quality. Case study of the Polish Topographic Database, in *9th International Conference "Environmental Engineering"*, 22–23 May 2014, Vilnius, Lithuania.

CIESIN. 2005. *Gridded population of the world (GPW). V3* [online]. Center for International Earth Science Information Network New York, Columbia University, USA [cited 22 August 2017]. Available from Internet: http://sedac.ciesin.columbia.edu/data/set/gpw-v3-centroids

CSO. 2010. *Basic information about Polish demographic development in the years 2000–2009 (in Polish: Podstawowe informacje o rozwoju demograficznym Polski w latach 2000–2009)* [online], [cited 22 August 2017]. Available from Internet: http://stat.gov.pl/cps/rde/xbcr/gus/lu_podsta_info_o_rozwoju_demograf_polski_2000-2009.pdf

Das, N. 2013. *Anciliary data report. Urban area. Preliminary, v.1.1, SMAP Science Document no. 046* [online], [cited 22 August 2017]. Available from Internet: http://smap.jpl.nasa.gov/system/.../288_046_urban_area_v1.1.pdf

Deichmann, U.; Balk, D.; Yetman, G. 2001. *Transforming population data for interdisciplinary usages: from census to grid* [online], [cited 22 August 2017]. Available from Internet: http://sedac.ciesin.columbia.edu/plue/gpw/GPWdocumentation.pdf

Elvidge, C. D.; Baugh, K. E.; Kihn, E. A.; Kroehl, H. W.; Davis, E. R. 1997. Mapping city lights with nighttime data from the DMSP operational linescan system, *Photogrammetric Engineering and Remote Sensing* 63(6): 727–734.

Elvidge, C. D.; Imhoff, M. L.; Baugh, K. E.; Hobson, V. R.; Nelson, I.; Safran, J.; Dietz, J. B.; Tuttle, B. T. 2001. Nighttime lights of the world: 1994–95, *ISPRS Journal of Photogrammetry and Remote Sensing* 56: 81–99. https://doi.org/10.1016/S0924-2716(01)00040-5

Goodchild, M. F.; Buttenfield, B.; Wood, J. 1994. Introduction to visualizing data quality, Chapter in: Hearshaw, H. M.; Unwin, D. J. (Eds.). *Visualization in Geographical Information Systems.* New York: John Wiley and Sons.

Hall, O.; Stroh, E.; Paya, F. 2012. From census to grids: comparing gridded population of the world with Swedish census records. *The Open Geogr. J.* 5: 1–5. http://sedac.ciesin.columbia.edu/data/set/grump-v1-urban-extents/data-download. https://doi.org/10.2174/1874923201205010001

Heuvelink, G. B. M.; Burrough, P. A.; Stein, A. 1989. Propagation of errors in spatial modelling with GIS, *International Journal of Geographical Information Systems* 3(4): 303–22. https://doi.org/10.1080/02693798908941518

INSPIRE. 2010. D2.8.I.2 INSPIRE Specification on geographical grid systems – guidelines. http://inspire.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_GGS_v3.0.1.pdf

ISO. 1994. ISO 5725-1:1994. *Accuracy (trueness and precision) of measurement methods and results* – Part 1: General principles and definitions.

ISO. 2012. ISO19158:2013. *Geographic information – Quality assurance of data supply*.

ISO. 2013. ISO 19157:2013. *Geographic information – Data quality*.

Linard, C.; Victor, A.; Alegana, V. A.; Noor, A. M.; Robert, W.; Snow, R. W.; Andrew, J.; Tatem, A. J. 2010. A high resolution spatial population database of Somalia for disease risk mapping, *International Journal of Health Geographics* 9: 45 [online], [cited 22 August 2017]. Available from Internet: http://www.ij-healthgeographics.com/content/9/1/45

Nowak Da Costa, J. 2016a. Novel tool for examination of data completeness based on comparative study of VGI data and official building datasets, *Geodetski vestnik* 60(3): 495–508. https://doi.org/10.15292/geodetski-vestnik.2016.03.495-508

Nowak Da Costa, J. 2016b. Towards building data semantic similarity analysis: OpenStreetMap and the Polish Database of Topographic Objects, *Baltic Geodetic Congress (Geomatics) Proceedings*, 2–4 June 2016, Gdańsk, 269–275.

Tobler, W.; Deichmann, U.; Gottsegen, J.; Maloy, K. 1997. World Population in a Grid of Spherical Quadrilaterals, *International Journal of Population Geography* 3: 203–225. https://doi.org/10.1002/(SICI)1099-1220(199709)3:3<203::AID-IJPG68>3.0.CO;2-C