The 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
22-25 September, 2021, Cracow, Poland

Validation of the Global Human Settlement Layer and NASA Population Data for Ukraine

Nataliia Kussul [1,2], Hanna Yailymova [1,2], Sophia Drozd[1], Andrii Shelestov[1,2]
[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 03056 Kyiv, Ukraine;
{sofi.drozd.13@gmail.com, andrii.shelestov@gmail.com}, http://ipt.kpi.ua//
[2] Space Research Institute NASU-SSAU, Glushkov Ave 40, 4/1, 03187, Kyiv 187, Ukraine,
{nataliia.kussul@gmail.com, anna.yailymova@gmail.com}, http://inform.ikd.kiev.ua/

Abstract— **In Ukraine, due to financial difficulties, the planned census is often postponed from year to year. The country is forced to rely on static data, which is sometimes completely untrue. In addition, it is not possible to count the number of inhabitants everywhere. In particular, Ukraine has not known for several years exactly how many people live in the occupied territories. As a result - the wrong distribution of the budget, which entails another financial crisis and a number of other troubles. At the same time there several satellite based products allowing to estimate the population. This study provide validation of satellite based population products delivered by JRC and NASA for the territory of Ukraine. To verify the correctness of the satellite based products, such as Global Human Settlement Layer (GHSL) and NASA population GPWv4 data collection have been compared to official statistics on the number of the largest cities of Ukraine.**

Keywords— population, satellite products, GHSL, corellation, NASA

I. INTRODUCTION

Today, there are many ways to conduct a census in the world. In Ukraine, a mixed type of census is usually used. However, this method requires a large amount of material and human resources. The world practices combined methods of keeping statistics on calculating the number of inhabitants. One of these involves the involvement of satellite imagery [1]. Artificial intelligence uses special algorithms to process the satellite image and returns the approximate number of inhabitants in the area captured in the image. This method allows you to calculate the number of inhabitants even on the territory that is inaccessible to conduct a questionnaire census. Now, Ukraine wants to calculate the number of inhabitants by satellite method in the uncontrolled territories of the eastern part of the country. In this case, come to the aid of global products that are being built around the world. Accordingly, the first question before using this data for the territory of Ukraine is the question of their verification. This article will check the accuracy of the satellite based population products to determine its feasibility in Ukraine.

This work acknowledge the funding received by the National Research Foundation of Ukraine from the state budget[1].

II. DATA AND MATERIALS

To validate the population data, official statistical data on the population for 2015-2020 were used for the 20 largest cities of Ukraine [2]. There are Kharkiv, Odessa, Dnipro, Donetsk, Zaporizhia, Lviv, Mykolaiv, Luhansk, Vinnytsia, Khmelnytsky, Cherginiv, Sumy, Poltava, Chernivtsi, Cherkasy, Rivne, Zhytomyr, Kherson, Ivano-Frankivsk, Lutsk, Ternopil. A separate validation was conducted in the districts of the city of Kyiv, for which statistical information on the population is available for 2007-2021 [3]. Population data were extracted from Global Human Settlement Layer (GHSL) [4] and images of NASA. This spatial raster dataset depicts the distribution of population, expressed as the number of people per cell. Data from GHSL are only available for 2015, and population data from NASA are available for 2015 and 2020. The example of GHSL product for Ukraine is presented on Fig. 1.

The GHSL produces new global spatial information, evidence-based analytics describing the human presence on the planet that is based mainly on two quantitative factors: (i) the spatial distribution (density) of built-up structures and (ii) the spatial distribution (density) of resident people. The GHSL uses global, multi-temporal archives of high-resolution satellite imagery, census data, and volunteered geographic information as input data [5]. The spatial resolution of GHSL data is 250 meters. The full methodology and concepts of population data creation described in [6]. Also, for the territory of Ukraine it is possible to use maps of buildings with 10-meter spatial resolution, which were obtained according to the developed methodology of classification of time series of satellite data at the Space Research Institute NASU and SSAU [7], [8], based on multilayer perceptron. In order to enhance efficiency of classification, it is reasonable to use not an individual perceptron, but an ensemble of them. At the same time, data merging at the decision-making level takes place during classification.

[1] The National Research Foundation of Ukraine project 2020.02/0284 "Geospatial models and information technologies of satellite monitoring of smart city problems".

Also, the GHSL may benefit population disaggregation and mapping. In the work [9] authors test the integration of the GHSL built-up grid and the Gridded Population of the World (GPW) in order to refine the mapping of population distribution in Syria, for the year 2000, greatly improving depiction of population distribution and density. Preliminary results indicate that GHSL is a good proxy for population disaggregation. The Chinese scientists investigated that GHSL tends to underestimate the building density of low-density areas and overestimate the areas with high density [10].

The NASA population GPWv4 data collection is produced using multiple input data sets, including geographic boundaries, census data, United Nations population estimates, and a global water mask with spatial resolution 1.1 km [11].

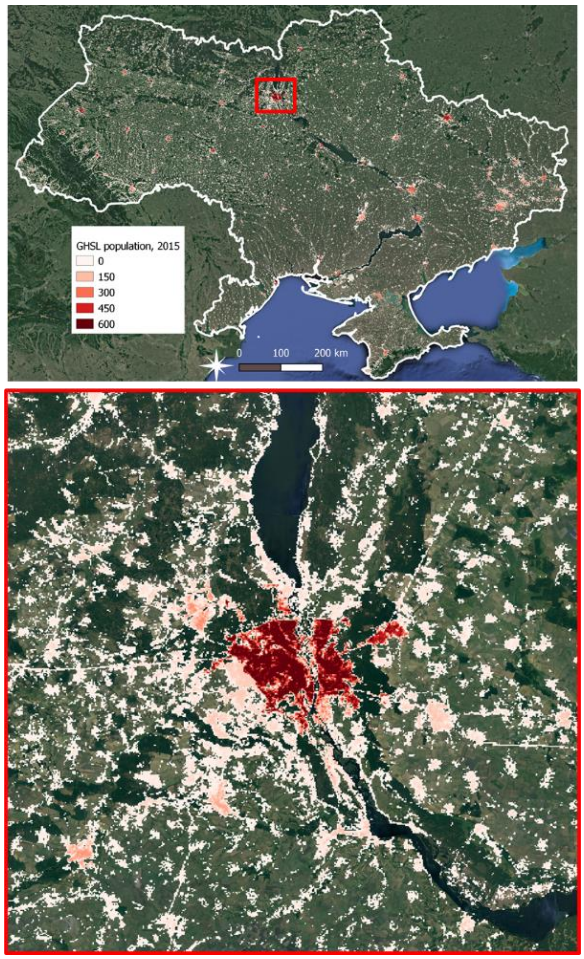

Figure 1. GHSL for Ukraine and Kyiv in 2015

III. METHODS

The vector boundaries of cities [12] and open-source software QGIS [13] with zonal statistics application were used to calculate the population by city based on the GHSL and NASA geospatial data. By counting the values of pixels was determined the population within each city.

Pearson's correlation coefficient was calculated to verify the relationship between satellite population data and statistics. Pearson's correlation coefficient describes the magnitude of the linear relationship between the data.

Pearson correlation coefficient formula:

$$r = \frac{N\sum xy - \sum(x)\sum(y)}{\sqrt{N[\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}},$$

where N is the number of pairs of scores. This article corresponds to the number of cities that were validated (20 of them), as well as the number of districts (10 of them) of the city of Kyiv in the study of the population of the capital. Σx is the sum of x scores. These are data obtained from official population statistics. Σy is the sum of y scores. This is the population data obtained from satellite images. Σxy is the sum of the products of paired scores. Σx^2 is the sum of squared x scores. Σy^2 is the sum of squared y scores.

In addition, to estimate the deviation of satellite data from official statistics, the value of absolute and percentage errors was calculated.

The formula for the absolute error (Δx) is:

$$\Delta x = xi - x,$$

where xi is the measurement, the population data obtained from satellite images, x is the true value, data obtained from official population statistics.

The formula for the percentage error (Y) is:

$$Y = \frac{\Delta x}{x} * 100\%,$$

where Δx is the absolute error, and x is the true value, data obtained from official population statistics.

IV. RESULTS

Pearson's correlation coefficient was calculated to verify the relationship between satellite population data and statistics (Fig. 2). Nationwide, Pearson's correlation coefficient is close to 1 for both Global Human Settlement Layer and NASA satellites. This means that there is a strong linear relationship between the data.

The Global Human Settlement Layer correlation coefficient for Kyiv districts is close to 0.77 (Table 1). This means that there is a high direct linear relationship between GHSL data and statistics. NASA data for the districts of Kyiv show a correlation coefficient close to 0.52, which is characteristic of the direct linear dependence of the average intensity.

During the analysis, the absolute and relative deviations of GHSL and NASA data from statistical data were calculated.

Fig. 3 shows that GHSL data is not significantly reflected in official census data. The maximum relative error according to GHSL is about 33% in the city of Kherson, the minimum error is less than 1% in the city of Mykolaiv. In the main part of cities - the error does not exceed 10%. At the level of Kyiv city districts, the relative deviation mostly fluctuates within 2-10%. Deviations of more than 10% are observed in Desniansky and Obolonsky

districts. A particularly large deviation (about 50%) was recorded in Holosiivskyi district. This is due to the fact that these areas contain a large number of greenery (parks, forests, squares), and the percentage of buildings is much lower compared to such regions of Kyiv as Pechersky, Dniprovsky, Solomyansky (Fig. 4).

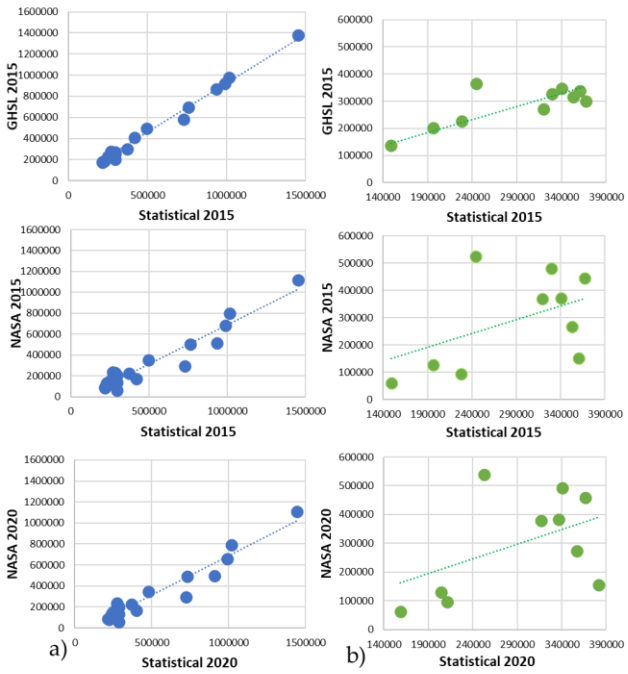

Figure 2. Validation of satellite population data for the largest cities of Ukraine (a) and districts of Kyiv (b)

TABLE I. PEARSON'S CORRELATION COEFFICIENT FOR DIFFERENT PROVIDERS

Population data provider	20 largest cities		districts of Kyiv	
	2015	2020	2015	2020
GHSL	0,995	N/A	0,776	N/A
NASA	0,963	0,963	0,529	0,523

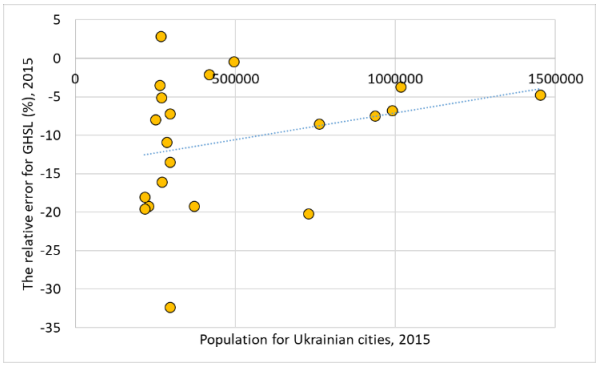

Figure 3. The relative errors of GHSL data for the 20 largest cities in Ukraine in 2015

The largest relative error according NASA population product is recorded in the city of Kherson and is about 80%. For other cities, the error exceeds 40%-60% in 2015 and in 2020 years (Fig. 5, Fig. 6).

At the level of districts of the city of Kyiv - the deviation of data from satellite images from official census data is about 50%. In Holosiivskyi district, deviations reach even more than 100% (Fig. 4). However, when estimating the data not separately by district, but for the whole city, the deviation rates have decreased significantly, in some years they do not reach even 7% for NASA data.

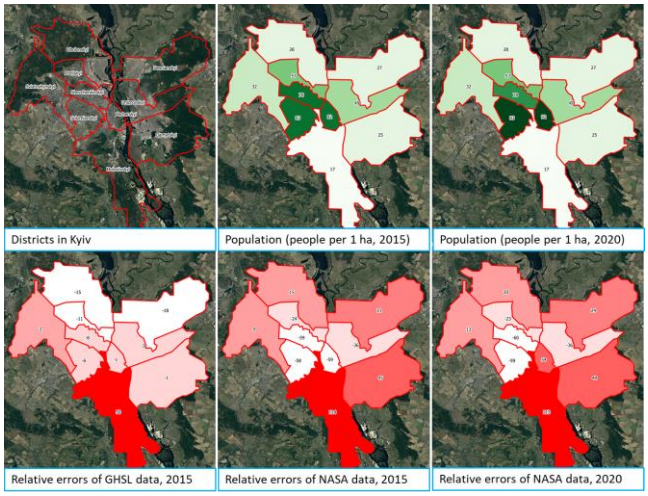

Figure 4. The relative errors of GHSL and NASA data for Kyiv districts

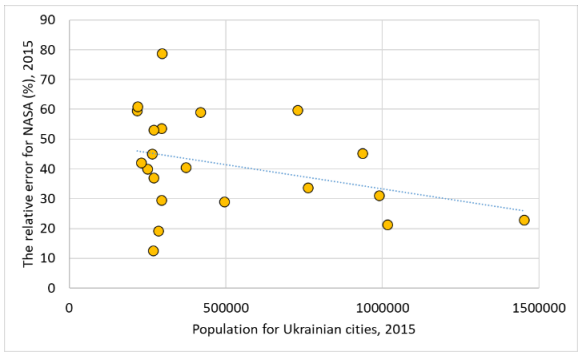

Figure 5. The relative errors of NASA data for the 20 largest cities in Ukraine in 2015

Thus, judging by the map in Fig. 8, it can be seen that GHSL satellite information can be used to calculate the number of inhabitants in eastern Ukraine. The magnitude of the error increases when moving west (the exceptions are the cities of Khmelnytsky and Rivne).

NASA satellites have significant errors in both the eastern and western parts of the country. The error decreases in the north-central part of Ukraine. In the capital, for example, even less than 10%.

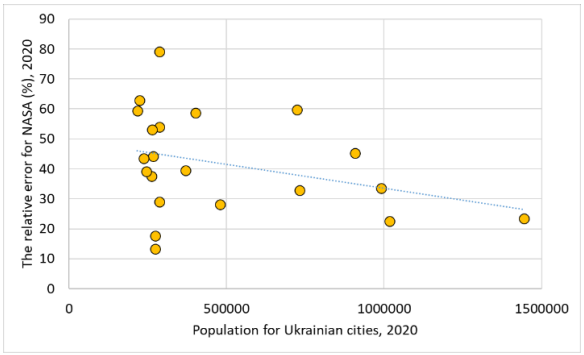

Figure 6. The relative errors of NASA data for the 20 largest cities in Ukraine in 2020

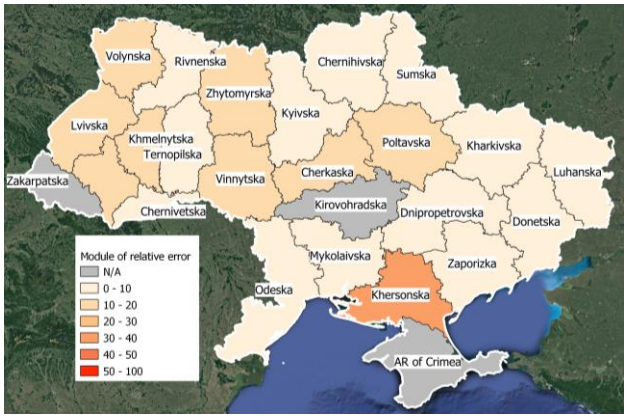

Figure 7. Dependence of GHSL data on the geographical location.

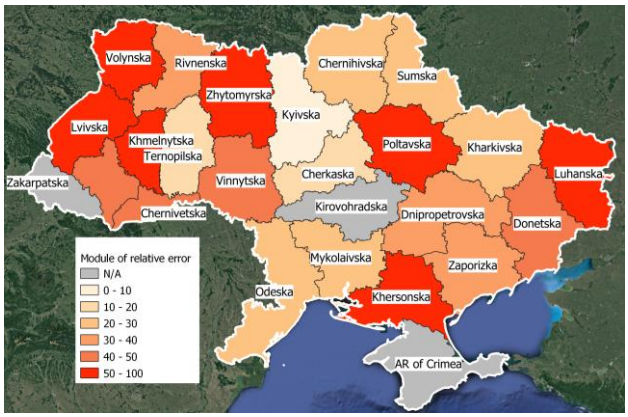

Figure 8. Dependence of NASA data on the geographical location.

V. Conclusions

As a result, we can conclude and state that the satellite census method could be used for different studies, such as SDG indicators estimation for national level. For most cities' deviation does not exceed 10%, but for some cities in Ukraine such as Kherson, Lviv, the error for population estimation is bigger and requires additional study. First of all, errors occur due to differences in spatial resolution of data. However, for most of the territory of Ukraine this method can be used based on the approximate value of the slope. By combining this method with traditional census methods, significant progress can be made in estimating the population of Ukraine.

References

[1] Population Estimation Mining Using Satellite Imagery. Available online: https://link.springer.com/chapter/10.1007/978-3-642-40131-2_25 (accessed on 08 April 2021).

[2] The current population of Ukraine. Available online: https://ukrstat.org/uk/druk/publicat/Arhiv_u/13/Arch_nnas_zb.htm (accessed on 08 April 2021).

[3] Population for districts of Kyiv. Available online: http://www.kiev.ukrstat.gov.ua/p.php3?c=1123%26lang=1 (accessed on 08 April 2021)

[4] Schiavina, Marcello; Freire, Sergio; MacManus, Kytt (2019): GHS population grid multitemporal (1975, 1990, 2000, 2015) R2019A. European Commission, Joint Research Centre (JRC) DOI: 10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218 PID: http://data.europa.eu/89h/0c6b9751-a71f-4062-830b-43c9f432370f

[5] M. Melchiorri, M. Pesaresi, A. Florczyk, Ch. Corbane, T. Kemper. 2019. "Principles and Applications of the Global Human Settlement Layer as Baseline for the Land Use Efficiency Indicator—SDG 11.3.1" ISPRS Int. J. Geo-Inf. 8, no. 2: 96. https://doi.org/10.3390/ijgi8020096

[6] Freire, Sergio; MacManus, Kytt; Pesaresi, Martino; Doxsey-Whitfield, Erin; Mills, Jane (2016): Development of new open and free multi-temporal global population grids at 250 m resolution. Geospatial Data in a Changing World; Association of Geographic Information Laboratories in Europe (AGILE). AGILE 2016.

[7] M. Lavreniuk, N. Kussul, S. Skakun, A. Shelestov and B. Yailymov, "Regional retrospective high resolution land cover for Ukraine: Methodology and results," 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 3965-3968, doi: 10.1109/IGARSS.2015.7326693.

[8] N. Kussul, M. Lavreniuk, A. Shelestov, S. Skakun "Crop inventory at regional scale in Ukraine: developing in season and end of season crop maps with multi-temporal optical and SAR satellite imagery," in European Journal of Remote Sensing, 2018, Vol. 51, pp. 627-636. DOI: 10.1080/22797254.2018.1454265.

[9] S. Freire, T. Kemper, M. Pesaresi, A. Florczyk and V. Syrris, "Combining GHSL and GPW to improve global population mapping," 2015 *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 2541-2543, doi: 10.1109/IGARSS.2015.7326329.

[10] Liu F, Wang S, Xu Y, Ying Q, Yang F, Qin Y (2020) Accuracy assessment of Global Human Settlement Layer (GHSL) built-up products over China. PLoS ONE 15(5): e0233164. https://doi.org/10.1371/journal.pone.0233164

[11] Center for International Earth Science Information Network (CIESIN), Columbia University. 2018. Documentation for the Gridded Population of the World, Version 4 (GPWv4), Revision 11 Data Sets. Palisades NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H45Q4T5F Accessed DAY MONTH YEAR.

[12] Ukraine - Subnational Administrative Divisions. Available online: https://data.humdata.org/dataset/ukraine-administrative-boundaries-as-of-q2-2017 (accessed on 08 April 2021)

[13] QGIS. Free open-source geographic information system. Available online: https://qgis.org/uk/site/ (accessed on 08 April 2021)