IZDELAVA PODROBNE POPULACIJSKE KARTE MESTA NA TEMELJU INFORMACIJ NACIONALNE PODATKOVNE BAZE O STAVBAH

DETAILED MAPPING OF THE DISTRIBUTION OF A CITY POPULATION BASED ON INFORMATION FROM THE NATIONAL DATABASE ON BUILDINGS

Tomasz Pirowski, Karolina Bartos

UDK: 314.1:528.45
Klasifikacija prispevka po COBISS.SI: 1.01
Prispelo: 30. 1. 2018
Sprejeto: 15. 6. 2018

DOI: https://doi.org/10.15292/geodetski-vestnik.2018.03.458-471
SCIENTIFIC ARTICLE
Received: 30. 1. 2018
Accepted: 15. 6. 2018

IZVLEČEK

V prispevku je predstavljena metodologija za dasimetrični razvoj podrobne populacijske karte za mestno okolje. Izračun prebivalstva temelji na povezavi nacionalne podatkovne baze o stavbah (BDOT) z razpoložljivimi statističnimi podatki o gostoti prebivalstva. Praktični primer smo izdelali za območje mesta Krakov, pri tem smo uporabili visokokakovostne demografske podatke, povezane s prostorskimi enotami mesta. Zemljevidi prebivalstva so bili izdelani na več načinov, pri čemer so bile upoštevane značilnosti in lokacije stavb. Faza optimizacije je temeljila na ustrezno prilagojeni površinsko uteženi metodi korelacije na podlagi globalnega števila prebivalcev v mestu, referenčni pa so bili statistični podatki 141 mestnih prostorskih enot. Dobljeni rezultati upravičujejo členitev območij enostanovanjskih in večstanovanjskih stavb. Prvotna opredelitev med številom prebivalcev in stanovanjsko površino stavbe (41 m^2/osebo) se je po optimizaciji spremenila: za območja enostanovanjskih stavb (84 m^2/osebo) in območja večstanovanjskih stavb (37 m^2/osebo). Zaradi tega so se napake MAPE izboljšale z 48 % na 30 % in napake RMSE z 2896 na 2684. Po naknadni členitvi mestnih enot glede na povprečno površino na prebivalca so se parametri zmanjšali na: MAPE 10 %, RMSE 1146.

KLJUČNE BESEDE

demografski podatki, dasimetrično modeliranje, topografski podatki, nacionalna baza stavb

ABSTRACT

The paper proposed a methodology for dasymetric development of a detailed population map for a city environment. The recalculation of population is based on linking the national database on buildings (BDOT) with the available statistical information about population density. The experiment was conducted in the city of Cracow, using demographic data with a high level of detail, related to the urban units of the city. The generation of population maps was performed for several options, dividing the buildings depending on their characteristics and location. The optimisation stage was based on a properly adjusted surface-weight method of correlation, where the global number of people in the city was used, while the statistical data from 141 urban units was considered to be reference data. The obtained results justify the division into single-family and multi-family buildings. The original connection between the function of population and the inhabitable area of a building (41 m^2/person) was differentiated during optimisation: for single-family houses (84 m^2/person) and for multi-family houses (37 m^2/person). Due to this, the MAPE errors were improved from 48% to 30%, and RMSE from 2896 to 2684. Having performed additional segmentation of urban units according to the average inhabitable areas per person recorded for them, the parameters dropped down to: MAPE 10%, RMSE 1146.

KEY WORDS

demographic data, dasymetric modelling, topographic data, national database on buildings

1. INTRODUCTION

Detailed data on the number of people and their distribution in the city are important for the local community and therefore an important element of local Geographic Information Systems (GIS). Numerous methods and proposals for the creation of population maps can be found in the literature. Until the mid-1950s, specific choropleth maps were one of the most popular and most frequently used methods of presentation. However, presenting the number of people in predetermined area units – usually corresponding to the administrative division – leads to the interpretation of information not corresponding to the real field situation. The dasymetric method reflects the spatial variability to a better extent, by introducing reference fields based on additional cartographic information. The first recognised document constituting a basis for this method is the population density map of the Cape Cod peninsula by Wright, 1936 (Mennis, 2003). The computerisation of cartography, the development of Geographic Information Systems and the increasing availability of digital spatial data over recent years have popularised this method of preparing population maps.

The selection of reference units in the dasymetric method is based on the assumption of the existence of areas characterised by an identical intensity of a given phenomenon. Due to the aggregated form, which is a typical feature of source data, it is necessary to use additional details, usually acquired from auxiliary maps, such as the cadaster or land use. For this reason, the dasymetric method is difficult to be automated – it requires the evaluation and association of many additional variables, and the recalculation of statistical data into new units must retain the condition of conformity with respect to basal statistical data (Tobler, 1979).

An important issue related to dasymetric maps is the division of variables into limiting variables and binding variables, introduced by Robinson et al. (1995). The role of limiting variables is to determine the absolute thresholds of values for a presented phenomenon. On the other hand, binding variables include geographic phenomena spatially related to the presented phenomenon. For population density maps, land use types are usually adopted as limiting variables, while land morphology, soil type, physical-geographical conditions and others constitute binding variables (Bielecka et al., 2005). An example of the use of satellite data for the determination of limiting variables is the paper of Harley (2002), however, its derivative products – land use and coverage maps (Eicher and Brewer, 2001; Gallego and Peedell, 2001; Mennis, 2003; Bielecka et al., 2005; Gallego, 2010; Pirowski and Pomietłowska, 2017), including the detection of impermeable surfaces – are used more commonly (Wu and Murray, 2005; Azar et al., 2010, Bajat et al., 2013).

The growing need for better-quality data is reflected by the development of population mapping methods. For example, new approaches to subpixel imperviousness prediction from remote sensing images are being developed to improve the estimation of impervious surfaces and their changes (Drzewiecki, 2016; Wang et al., 2017). Documents providing information on the diversity of population density keep emerging, for example related to daily commuting (Sleeter and Wood, 2006; Smith et al., 2015), including those using information originating from mobile devices (Horanont and Shibasaki, 2010). Important binding variables are sought to enable forecasting changes in the distribution of people (Bajat et al., 2011). The proposed methods allow obtaining higher spatial resolution, like those based on the detection of buildings using photographic interpretation methods (Pirowski and Drzewiecki, 2012), laser scanning (Sridharan and Qiu, 2013), the complementary use of numerical land coverage model and aerial photographs CIR (Ural et al., 2011), using information on the taxation of parcels (Maantay et al., 2007, 2009), relation to the street network (Riebel and Bufalino, 2005; Zandbergen and Ignizio, 2010), the location of address points (Tapp, 2010; Bakillah, 2014), and using databases of buildings (Lwin and Murayama, 2009; Bajat et al., 2013; Całka et al., 2016). The increasing amount of additional data necessitates the development of algorithms that successfully use binding variables and return control-statistical information of the generated products of the dasymetric method. Intelligent Dasymetric Mapping is one example (Mennis and Hultgren, 2006; França et al., 2014).

The paper presents the recalculation of population numbers from 141 urban units (j.u.) of the city of Cracow. Smaller units, so-called urban units, were also introduced taking into account, among others, the old cadastral divisions belonging to the parish, the division into settlements or historical urban units (derived from cadastral municipalities, which in essence meant division of the city for tax purposes). At the same time, these are the best publicly available statistics on Cracow's population. The presented recalculation is based on information about the location of buildings, and their function and size, originating from the Polish domestic Topographic Objects Data Base 1:10 000 (referred to hereafter as BDOT10k). A similar approach, based on the so-called footprint of a building, was previously used, e.g. Lwin and Murayama (2009), and the division of buildings depending on the number of storeys was introduced into population estimation by Bajat et al. (2013). The proposal shown in the present paper is an extension and modification of the method used by Całka et al. (2016) for a scarcely populated area, whose result was a raster map in a 1 km grid. The novelty of the approach presented herein is the implementation of multi-option segmentation of buildings based on their height and location, in order to improve the credibility of estimating the number of people inhabiting them and producing a high-resolution city population map in a 5 m grid. In this approach, the population allocation is not directly related to the surface or size of the building, but it is additionally modified by the factor associated with the segmentation used.

2 METHODOLOGY

Within an urban and industrial agglomeration, the density of buildings and their character, height and functions are very diverse. In population mapping using the dasymetric method, spatial constraints, such as building categories from Corine Land Cover (CLC) and Urban Atlas (UA) are insufficient. Thus, the adopted method is based on an assumption that the number of people is closely related to the inhabitable area of buildings and houses (night population modelling). The research issues are related to answering the question of whether a simple association of any habitable surface area is sufficient, or whether the segmentation of buildings should be performed. How does one take into account shared surface areas in tenements and residential blocks (for example staircases), or service establishments, frequently present especially on ground floors? Is the function of assigning population to a habitable area constant throughout the city, or does it require introducing its local variability using binding variables? If so, how are they to be defined properly?

The simplest case adopted the solution used by Całka et al. (2016), which assumed a simple dependence between the size of a building (a product of the building's surface area and the number of its storeys) and the number of people originating from the statistical data provided by the Central Statistical Office (hereafter referred as CSO, in Polish: GUS) involving the average area of a home and the number of people inhabiting it within the analysed area. Całka et al. (2016) reported high local discrepancies for the individual municipalities between the population calculated in this manner and the actual status (from an underestimation by -10% to an overestimation by + 25%). This is why in the next step the authors iteratively corrected the discrepancies, changing the initially adopted theoretical number of people in one apartment, so that the calculated volume of people could comply with the statistical data in the given municipality.

In order to avoid the abovementioned operations, the adopted solution of calculating the average surface area per one inhabitant of Cracow (in m^2/person) was based on the globally calculated inhabitable area from the BDOT database and the total number of inhabitants resulting from the aggregation of data from 141 j.u. (1).

$$Av(m) = \frac{\sum_{1}^{km}\left(N(b)\times A(b)\right)}{\sum_{i=1}^{141} Pop(ju)} \tag{1}$$

where: *N*(*b*) - the number of storeys in a residential building, *A*(*b*) - the footprint area of a residential building, *Av*(*m*) – the average area per one inhabitant of the city, *Pop*(*ju*) – the number of people in the i-th urban unit, *km* – the number of residential buildings in the city.

The population of the given urban unit may then be estimated (2):

$$Pop_e(ju_i) = \frac{\sum_{1}^{kju(i)}\left(N(b)\times A(b)\right)}{Av(m)} \tag{2}$$

where: $Pop_e\,(ju_i)$ – the estimated sum of people in the i-th urban unit, *kju* – the number of residential buildings in the i-th urban unit.

The values calculated according to formula (1) differ from the real values originating from statistical data. Instead of correcting it iteratively, the authors propose the use of a coefficient (3):

$$K(ju_i) = \frac{Pop(ju_i)}{Pop_e(ju_i)} \tag{3}$$

The number of people per a specified building in the given j.u. then equals (4), and the presented sequence of operations allows fulfilling Tobler's condition (1979) i.e. preserve population totals in census enumeration units.

$$Pop_e(b_i) = \frac{N(b)\times A(b)}{Av(m)} \times K(ju_i) \tag{4}$$

where: $Pop_e\,(b_i)$ – the estimated number of people inhabiting a building in the i-th urban unit.

The calculated values of population for each building are added to a base of attribute data associated with the vector form of buildings from BDOT. In the following step, depending on the needs, it is possible to prepare generalised population maps, including those in the form of a raster model.

The presented course of action (1-5) allows, in a simpler manner compared to what was proposed by Całka et al. (2016), recalculating the quantity of the population from basic statistical data into new spatial units. An additional advantage of the change in the algorithm is the possibility of a preliminary estimation of the product's credibility based on the input data in possession. This is important when there are no reference data allowing an independent evaluation of the process used for the recalculation of people. This is because it can be easily noticed that, based on the formula (4), excluding its correcting segment *K*(*ju*), one can obtain a dasymetric map fulfilling Tobler's condition for the whole city, although not fulfilling it for the individual j.u. Considering the statistical data from j.u. as referential, local discrepancies can be considered as deviations δ (5).

$$\delta_i + Pop_e(ju_i) = Pop(ju_i) \qquad (5)$$

At this stage, not using the *K*(*ju*) coefficient, a decrease in the deviations may be caused by the segmentation of buildings based on their attributes from the BDOT base and/or the introduction of additional binding variables. Statistical values obtained in the process of optimisation could constitute a basis for an evaluation and choice of the best option. In order to accomplish this objective, use was made of the experience based on Land Use / Land Cover (LULC) data from Corine Land Cover (CLC) and Urban Atlas (UA), presented in the papers of Gallego and Peedell (2001), Bielecka (2005), and Pirowski and Pomietłowska (2017), in which surface-weight methods of correlation were used. In these methods, the weight coefficient differentiated the population density among various types of land coverage. In the case of using BDOT, this is related to the adopted inhabitable areas Av (the average usable area per one inhabitant), depending on the type of building and its location. The minimisation of the mean squared error calculated for 141 j.u. was chosen as the condition (6).

$$\sum_{i=1}^{141} (\delta_i)^2 \rightarrow \min \qquad (6)$$

The authors of previous papers (Eicher and Brewer, 2001; Riebel and Bufalino, 2005; Tapp, 2010) also used the mean squared error (RMSE) when analysing the studied methods. The RMSE error allows easy interpretation, assuming the values to be the same units as the mapped variables. In this case, due to the high diversity of urban units with respect to the number of inhabitants, apart from RMSE, the following parameters were used to evaluate the options: R^2 (coefficient of determination), MPE (mean percentage error), and MAPE (mean absolute percentage error).

The demographic data related to Cracow, involving the whole city and divided into districts and urban units, are published on the website of Cracow City Hall supervised by the City Development Department (www.msip2.um.krakow.pl). The data are updated on a yearly basis and compiled based on the publications of the Statistical Office, the National Register of Entities of National Economy REGON database, the Regional and Local Data Bank of GUS, City Status Reports and the population register database supervised by Cracow City Hall.

Information about the location and characteristics of buildings was acquired from the national database of topographic objects BDOT10k of 2011, implemented on the scale of 1:10 000 for the whole country. From the population mapping point of view, this data has a very high level of detail. As reported by Bielecka (2015), the surveying error RMSE involving the location of a building amounts to 1.3 m. A graphical representation of buildings in a vector form is associated with a descriptive database, comprising data important from the standpoint of population calculation, involving features of buildings such as their functions, footprint areas and the numbers of storeys.

In order to differentiate between residential and non-residential buildings, information about the general function of the object was used, followed by proceeding with their selection, using information about the specific function of the building. Ultimately, single-family buildings, multi-family buildings and multi-residence buildings were approved for the analysis.

3 CHARACTERISTICS OF THE STUDY AREA

The study area is the city of Cracow (fFigure 1). Its surface area is 326.80 km^2 and the number of registered inhabitants in 2012 was 758,334 people. Cracow is divided administratively into eighteen districts, which in turn are additionally divided into smaller j.u. They are derived from former cadastral units. The shapes of their boundaries are determined by natural barriers: roads, railways, and watercourses.

In terms of the types of buildings, their functions and population density, Cracow is highly diversified within its administrative borders. There are old houses (tenements, religious buildings), panel building neighbourhoods, low single-family houses, and large industrial areas.

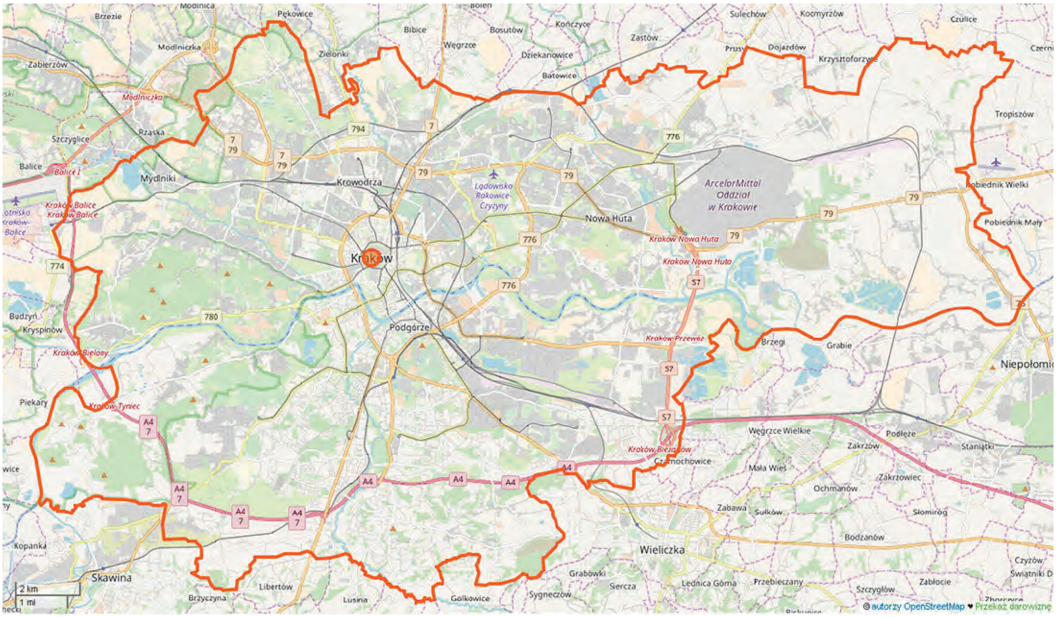

Figure 1: The study area – Cracow (source: © OpenStreetMap contributors, www.openstreetmap.org/copyright).

4 MULTI-OPTION ESTIMATION OF POPULATION DISTRIBUTION

Based on formulas (1-4), and subsequently minimising errors present in the individual j.u. by means of the least square method (5, 6), the theoretical area per one inhabitant was calculated for Cracow. The resulting value $Av(m)$ = 41.14 m^2/person is close to the one produced according to formula (1) (43.6 m^2) and much higher than indicated by the report on the results of the National Population and Housing Census of 2011 (23.8 m^2). This confirms the problems reported by Całka et al. (2016) with adopting the census value directly as a basis for calculations. Such a large discrepancy may be explained by the failure to take into account service and commercial spaces and shared spaces present in multi-family buildings.

Other options no. 2, 3, 4 allow assigning different areas per inhabitant depending on the characteristics of buildings. This results in the different values of *Av*(*m*) for specified groups of buildings. Option 2 assumes that area per inhabitant may differ significantly for single-family and multi-family buildings. This requires expanding the formulas (1-5) into the form (7), so that in the next step the *Av*(*m*) parameters could be calculated using the least squares method (6). Option 3, using the number of people per house, is a modification of option 2 (formula (8)). It is based on the assumption that the relationship between the area of a single-family house and the number of inhabitants is not as correlated as for apartments. This is because owning a house today is becoming a measure of the status of its owners, an intentional choice of lifestyle. Option 4 additionally introduces the division of multi-family houses into lower (I-V storeys) and higher buildings (more than V storeys) (formula (9)). The ultimately adopted segmentation was preceded by numerous tests, which considered other storey intervals and/or introducing specified constant values of the area correcting the surface area of buildings available for division between inhabitants. The ultimately adopted solution is a kind of prosthesis, making it possible, for higher and lower buildings separately, to take into account a different share of non-residential surfaces, such as the commercial and service areas commonly located on the ground floor.

$$\delta_i + \frac{\sum_{1}^{kju\_jm(i)} \left(N(b)\times A(b)\right)}{Av_{jm}(m)} + \frac{\sum_{1}^{kju\_wm(i)} \left(N(b)\times A(b)\right)}{Av_{wm}(m)} = Pop(ju_i) \qquad (7)$$

$$\delta_i + kju\_jm(i)\times Pop_{jm}(m) + \frac{\sum_{1}^{kju\_wm(i)} \left(N(b)\times A(b)\right)}{Av_{wm}(m)} = Pop(ju_i) \qquad (8)$$

$$\delta_i + \frac{\sum_{1}^{kju\_jm(i)} \left(N(b)\times A(b)\right)}{Av_{jm}(m)} + \frac{\sum_{1}^{kju\_wmn(i)} \left(N(b)\times A(b)\right)}{Av_{wmn}(m)} + \frac{\sum_{1}^{kju\_wmw(i)} \left(N(b)\times A(b)\right)}{Av_{wmw}(m)} = Pop(ju_i) \qquad (9)$$

where: *kju_jm* – a number in the i-th urban unit of single-family buildings, *kju_wm* – of multi-family buildings, $Av_{jm}(m)$ – the average area per one inhabitant in single-family buildings, $Av_{wm}(m)$ – in multi-family buildings, $Av_{wmn}(m)$ – in multi-family buildings up to V storeys, $Av_{wmw}(m)$ – in multi-family buildings exceeding V storeys, Pop_{jm} (m) – the average number of inhabitants in one single-family building.

Table 1: Options 1, 2, 3, 4 and *Av* coefficients and errors produced for them, in relation to j.u.

	Option 1	**Option 2**	**Option 3**	**Option 4**
	Av	Av_{jm} **/** Av_{wm}	Pop_{jm} **/** Av_{wm}	Av_{jm} **/**Av_{wmn}**/**Av_{wmw}
Surface area or the number of people Av	41.14	84.21 / 37.08	2.88 / 37.08	81.85/ 56.76/ 22.88
R2 [%]	82.69	84.78	84.98	88.66
MPE [%]	33.81	5.37	10.34	1.98
MAPE [%]	47.67	30.72	28.41	31.43

When searching for additional binding variables, the use of an artificial administrative division into 18 districts was assumed. Options 2a, 3a, 4a (table 2) were developed considering each district separately and independently. Such a spatial segmentation sample is justified by different historical conditions related to the individual districts (the time of the construction of buildings, different social and economic growth associated with a historical context such as the time of partitions or the 1950s), as well as their functions served nowadays (scientific, cultural, residential, industrial), which – if they turn out to be significant – may affect the characteristics of population density.

Table 2: The list of *Av* results in districts - options 2a, 3a, 4a.

Segmentation by districts	**Option 2a**	**Option 3a**	**Option 4a**
	Av_{jm} / Av_{wm}	Pop_{jm} / Av_{wm}	Av_{jm}/Av_{wmn}/Av_{wmw}
Stare Miasto	121.21 / 67.38	1.12 / 67.20	329.76 / 70.38 / 43.25
Grzegórzki	106.00 / 40.34	3.54 / 41.96	57.42 / 47.02 / 35.98
Prądnik Czerwony	33.50 / 41.20	6.82 / 41.23	222.16 / 24.15 / 56.91
Prądnik Biały	82.09 / 35.72	2.61 / 35.50	639.64 / 22.85 / 58.63
Krowodrza	5.98 / 54.63	34.74 / 50.57	5.95 / 66.60 / 40.70
Bronowice	57.51 / 58.47	3.62 / 54.36	48.87 / 468.15 / 34.35
Zwierzyniec	127.18 / 53.10	2.25 / 53.75	89.24 / 149.51 / 7.05
Dębniki	67.02 / 56.22	3.04 / 54.86	70.49 / 98.23 / 27.68
Łagiewniki - Borek Fał.	114.37 / 33.89	1.55 / 32.16	55.43 / 70.54 / 31.39
Swoszowice	72.78 / 54.87	3.34 / 69.06	69.91 / 112.95 / 30.00
Podgórze Duchackie	41.13 / 34.86	5.20 / 34.71	26.13 / 28.41 / 119.82
Bieżanów - Prokocim	78.62 / 25.83	2.88 / 25.84	77.87 / 23.67 / 29.54
Podgórze	53.86 / 46.66	4.39 / 46.70	53.40 / 46.98 / 45.34
Czyżyny	97.75 / 34.71	2.26 / 34.77	142.75/ 14.15/ 191.40
Mistrzejowice	13.94 / 30.96	-159.48 / 11.74	47.12 / 16.79 / 133.34
Bieńczyce	82.78 / 22.39	3.12 / 22.40	297.08 / 16.75 / 33.80
Wzg. Krzesławickie	63.91 / 23.48	2.99 / 22.33	66.56 / 22.47 / 30.44
Nowa Huta	67.68 / 30.21	2.88 / 30.18	69.11 / 29.27 / 32.90
R^2 [%]	98.53	98.57	98,91
MPE [%]	4.58	5.58	2.50
MAPE [%]	18.84	17.17	16.15

Another step was an attempt at discerning areas of similar demographic and social characteristics, which would be better at performing the segmentation of j.u. compared to districts. No information helpful in carrying out such a division for Cracow has been found in the literature related to urbanism. Taking advantage of the capabilities of GIS, the data related to the surface areas of the individual j.u., their footprint areas, and population, the types and numbers of buildings were compared to each other, in order to search for mutual dependences. As it could be expected, only the footprint area and the number of people exhibited a relation. On this basis, j.u. were divided into four groups: with areas below 40 m^2, 40-60 m^2,60-80 m^2 and above 80 m^2. The subsequent actions were analogical to the division into districts, resulting in options 2b, 3b, and 4b (table 3).

Table 3: A list of the results of *Av* in groups of j.u. according to their average area - options 2b, 3b, 4b.

Segmentation by average area in the j.u.	Option 2b Av_{jm} / Av_{wm}	Option 3b Pop_{jm} / Av_{wm}	Option 4b Av_{jm} /Av_{wmn}/Av_{wmw}
below 40 m²	98.06 / 28.21	1.94 / 28.03	228.60 /24.00 / 34.17
40 m² - 60 m²	47.78 / 47.78	4.46 / 46.93	50.78 / 53.16 / 37.89
60 m² - 80 m²	66.75 / 64.00	3.40 / 63.40	68.51 / 72.15 / 47.42
above 80 m²	95.24 / 103.42	2.81 / 100.62	95.28/103.39/101.50
R² [%]	97.23	97.14	97.54
MPE [%]	2.13	6.43	1.51
MAPE [%]	9.90	13.86	10.19

In the final stage, taking into account the *Av* parameters calculated in the individual options, the correction of the population volume took place using the *K*(*ju*) coefficient, followed by supplementing the building attribute database with the estimated numbers of inhabitants. In this manner, 10 city population maps were prepared as vector maps. Due to the limited space of the paper, the six most significant results are illustrated in Figure 2. The other results showed little significant differences or were characterized by unacceptable *Av* parameters (a detailed explanation of the factors of selection is presented in the discussion).

5 DISCUSSION

Considering the primary segmentation of buildings, it can be clearly seen that distinct treatment of single-family buildings in urban areas is justified. Option 1 (Table 1, Figure 2a), not taking into account the division, is characterised by bigger errors compared to the other ones (Figures 2b, 2c, 2d). The values of surface area per 1 person calculated in option 2 are approximately twice as high for single-family houses compared to multi-family buildings. The lack of a significant relation between the surface areas of single-family houses and the number of their inhabitants is confirmed by similar results for options 2 and 3 (Figures 2b, 2c). The introduction of additional division of multi-family buildings in option 4 has clearly distinguished the surface area for "skyscrapers" compared to lower houses and tenements. The lower share of joint spaces, commercial and service areas in the case of high buildings does not explain such large recorded differences. Two facts are probably decisive here: the presence of high block buildings mainly from the 1970s and 1980s, characterised by apartments with small rooms, densely inhabited, additionally lacking service premises on the ground floors, and – at the other end of the spectrum – depopulated tenements in the very centre, with large apartments, and numerous service and commercial premises. This is why the effect of an additional division is visible e.g. in the j.u. of the very centre of the city (Figure 2d), where the concentration of tenements is present. It should be pointed out that choropleth maps of fig. 2 reflect the values of *K*(*ju*) necessary to be used in the individual j.u. at the final stage of the generation of population maps in order to retain Tobler's condition.

Optimisation performed for division into districts (options "a", Table 2) decreased the global errors recorded for all the options. Unfortunately, for several districts (Krowodrza, Mistrzejowice, Czyżyny, Zwierzyniec) the surface areas calculated iteratively are erroneous (even negative), which also challenges the credibility of the calculations for the remaining ones. This indicates that optimisation performed for small samples (for an average of approximately eight j.u. for each district) may produce good results globally (fig. 2e), while locally the result is flawed. In order to correct such a result, additional limitations should be introduced into the iterative process, related to the detection of binding variables characterising the j.u.

The division into four subcategories of urban units (options "b", Table 3) produced more stable results compared to those recorded for districts. Only in a group of j.u. with average small surface areas per person does the surface area calculated in single-family houses change, even twofold depending on the option. Segmentation used in options "b" has merged "block estates" from various parts of Cracow with very high numbers of inhabitants, like Bieżanów, Kurdwanów, Prądnik, and Nowa Huta, which for options "b" resulted in the smallest MAPE errors among all the tested solutions, with the best result amounting to 9.9% (option 2b, Figure 2f).

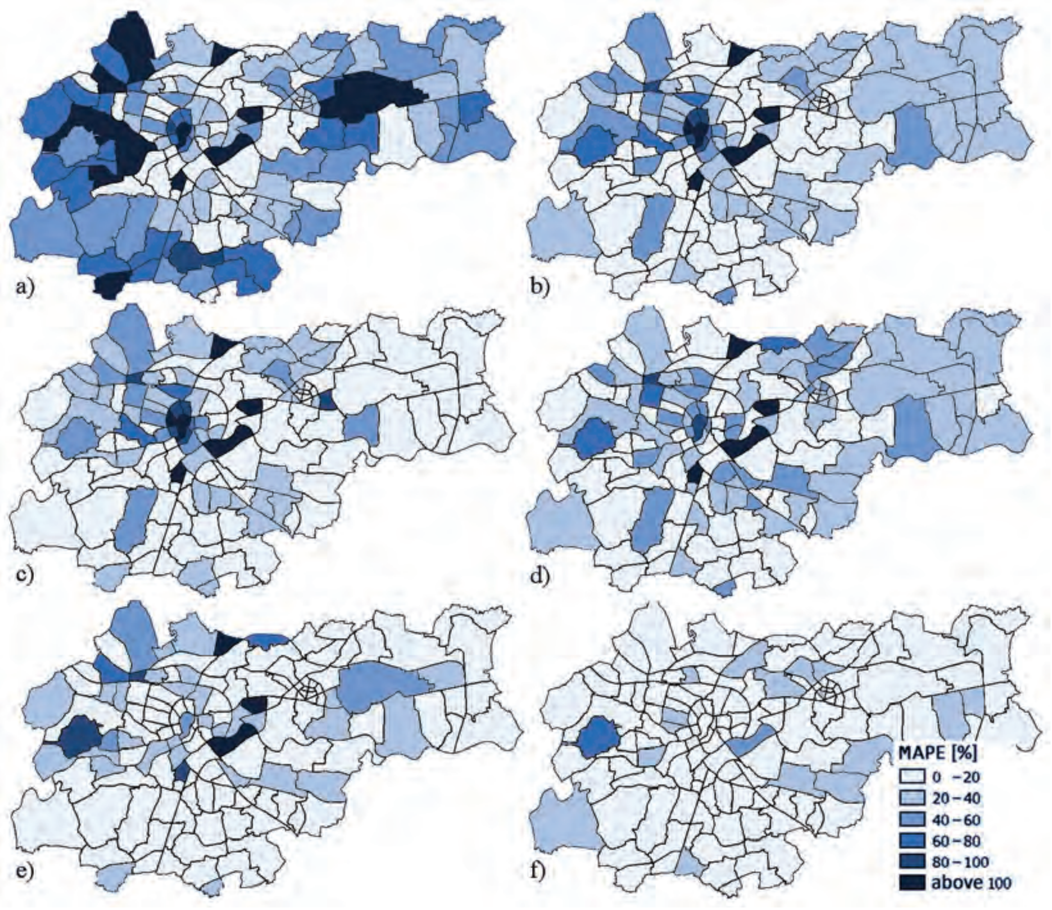

Figure 2: Distribution of the MAPE error for individual j.u. in options: a) 1, b) 2, c) 3, d) 4. e) 2a, f) 2b.

Among the presented statistical parameters, the MAPE error [%] is to be considered useful and easy to interpret. During the stage of selecting weights and the segmentation of data, it allows evaluating the level of estimation which can be reached locally (in j.u.), if the available statistical data involved only the scale of the whole city. The use of ME or MPE parameters mainly provides information on systematic errors.

On the other hand, the RMSE error, directly using the numbers of people, provides a misconception about the major mistakes of the dasymetric method. This results from the fact of the very high diversity of j.u. regarding the number of people living in them. The high squared errors recorded in several j.u. with high numbers of inhabitants determine the RMSE error for the whole city, although percentage-wise these errors do not deviate from the average value for j.u. The parameter R^2 is characterised by low variability (compared to MAPE, its sensitivity is much worse), nor can its value be used to describe errors in the individual j.u.

Based on the obtained results, it can be seen that the applied dasymetric method has limitations. When applying it, it should be taken into account that the input data is of major influence on the quality of the building segmentation process, and thus on population conversion. While the BDOT database is homogeneous across the country, the spatial distribution and character of the buildings may be locally quite different, for example in small-town or rural areas. On the national scale as well, individual census data on the population are very diverse, both in terms of area and volume of the population that lives in it. This is a big problem when trying to optimize the coefficients determining the average living space of the population (Av), differentiated by segmentation per type and location of the building.

Another difficulty that has not been successfully solved by the methodology used is the diversification of non-residential space in multi-dwelling buildings of varying heights. This is important because a significant impact on the actual number of the building's population is to allocate its specific storeys (most often the ground floor) for commercial and service activities. Thus, another meaning is one floor for a low building, another one for a tall one. With automatic population conversion, the only effective solution would be to have this type of information in the building databases. This would also make it easier to model the daily movement of people related to work or the use of commercial and service outlets.

Despite the above reservations and limitations in the application of the method, the segmentation of buildings in the analyzed area has achieved the intended effect, confirmed by reduced errors recorded for census units. The effectiveness of the approach applied to areas of a different nature, for example rural, would require additional research.

6 CONCLUSIONS

The dasymetric method proposed for the recalculation of city population is based on the use of information on spatial location of buildings from BDOT as a limiting variable, and on their characteristics – type, size, and optionally location in a specific part of the city – which information constitutes a basis for introducing proper binding variables. The main result of these procedures was a proper assignment of residential surfaces per inhabitant.

The conducted analyses indicate a high potential of using BDOT to redistribute population at a high level of detail. However, the results produced in the research are ambiguous. Certainly, the division into single-family (single-dwelling) and multi-family (multi-dwelling) buildings affects the improvement of the result, while further particularisations raise doubts. This is why, among the proposed options, the authors considered options 2 and 2b to be the best compromise between the estimated level of erroneousness and the credibility of the final product. The introduction of additional divisions increases the risk of the equifinality phenomenon, as confirmed by the produced incorrect parameters of *Av* in options "a" and partially in "4". Options 3, 3a and 3b produce results comparable to options 2, 2a, and 2b: however, it seems more intelligible to use one optimising measure – the area assigned per person, used by options "2".

Among the applied global statistical parameters, MAPE proved to be the most useful. This is because the use of absolute values allows better error tracing in a set of census units with a very large spread of the population volume. Analysing the spatial distributions of PE provides intelligibility and it can also be one of the empirical methods of searching for further binding variables.

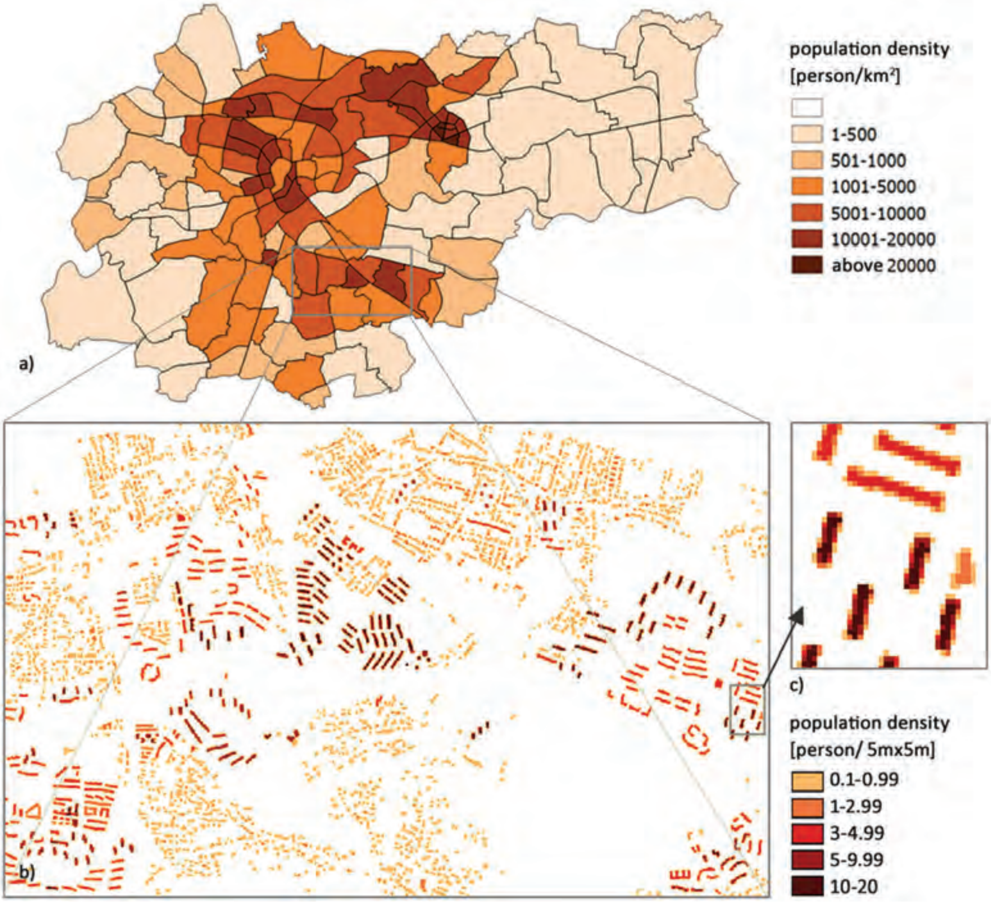

Figure 3: Population maps of Cracow: a) population density in j.u.; b) and c) dasymetric method in a 5 m grid (option 2b, a fragment of the city).

The initial association of habitable area with the number of people (option 1.41 m2/person) as a function, after dividing into single-family (84 m2/person) and multi-family (37 m2/person) buildings, caused the reduction of errors in option 2: MAPE from 48% to 31%, and MPE from 34% to 5%. Further division of multi-family buildings (option 4) simply decreased the MPE error to 2%. On the other hand, the use of segmentation of j.u. based on the average areas (associated with the structure of development in j.u.) in option 2b resulted in an MAPE of 10% and an MPE of 2%. It should be pointed out that these values characterise only the intermediate stage of the developed population maps and are primarily intended to provide credibility of the correctness of weights calculated using the surface-weight correlation method. In the final stage, proper recalculation leads to the compliance of the population volume within each urban unit.

Figure 3 presents the effects of a multi-stage process. Fig. 3a presents the input statistical data in the form of a choropleth map. A dasymetric map of Cracow's population is compiled in option 2b, recalculated into a raster model with a resolution of 5 m (figs 3b, 3c). At this resolution, in dense urban development zones, the network of streets is correctly mapped as unoccupied zones, and scattered development differs from low terraced buildings.

The indicated problem of the equifinality phenomenon can be resolved unambiguously only via the verification of final products based on independent data. Because of the reference data, the presented evaluation of options could be confirmed, since there is no certainty whether options with smaller errors at a j.u. level actually reflect the distribution of people inside the units to a better extent. Such independent evidence would indirectly indicate statistical measures useful in optimising the selection of weights, followed by estimating the credibility of the final products. In order to answer these questions, it is planned to continue research in that area, by preparing and acquiring adequate additional data.

ACKNOWLEDGEMENTS

This article has been prepared within the scope of the AGH UST statutory research no. 11.11.150.949.

Literature and references:

Azar, D., Graesser, J., Engstrom, R., Comenetz, J., Leddy JR, R. M., Schechtman, N. G., Andrews, T. (2010). Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti. International Journal of Remote Sensing, 31 (21), 5635–5655. DOI: https://dou.org/10.1080/01431161.2010.496799

Bajat, B., Krunić N., Samardžić Petrović, M., Kilibarda, M. (2013). Dasymetric modeling of population dynamics in urban areas, Geodetski vestnik, 57 (4), 777–792. DOI: https://dou.org/10.15292/geodetski-vestnik.2013.04.777-792

Bajat, B., Krunić, N., Kilibarda, M., Samardžić Petrović, M. (2011). Spatial Modelling of Population Concentration Using Geographically Weighted Regression Method. Journal of the Geographical Institute "Jovan Cvijic" SASA , 61 (3), 151–167. DOI: https://doi.org/10.2298/IJGI1103151B

Bakillah, M.,. Liang, S., Mobasheri, A., Arsanjani, J. J., Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. International Journal of Geographical Information Science, 28 (9), 1940–1963. DOI: https://dou.org/10.1080/13658816.2014.909045

Barrozo, L. V., Pérez-Machado, R. P., Small, C., Cabral-Miranda, W. (2015). Changing spatial perception: dasymetric mapping to improve analysis of health outcomes in a megacity. Journal of Maps, 11, 1–6. DOI: https://doi.org/10.1080/17445647.2015.1101403

Bielecka, E. (2015). Geographical data sets fitness of use evaluation. Geodetski vestnik, 59 (2), 335–348. DOI: https://doi.org/10.15292/geodetski-vestnik.2015.02.335-348

Bielecka, E., Kuczyk, A., Witkowska, E. (2005). Modelowanie powierzchni statystycznej przedstawiającej gęstość zaludnienia w Polsce przy pomocy metody dazymetrycznej. Polskie Towarzystwo Informacji Przestrzennej, Roczniki Geomatyki, Tom III, Zeszyt 2, 9–16.

Całka, B., Bielecka, E., Zdunkiewicz, K. (2016). Redistribution population data across a regular spatial grid according to buildings characteristics. Geodesy and Cartography, 65 (2), 149—162. DOI: https://doi.org/10.1515/geocart-2016-0011

Drzewiecki, W. (2016). Improving sub-pixel imperviousness change prediction by ensembling heterogeneous non-linear regression models. Geodesy and Cartography, 65 (2), 193-218. DOI: https://doi.org/10.1515/geocart-2016-0016

Eicher, C. L., Brewer, C. A. (2001). Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. Cartography and Geographic Information Science, 28 (2), 125–138. DOI: https://doi.org/10.1559/152304001782173727

França, V. O., Strauch, J. C. M., Ajara, C. (2014). Intelligent Dasymetric Method: an Application in Mesorregião Metropolitana de Belém. Revista Brasileira de Cartografica, 66 (6), 1395–1411.

Gallego, J. (2010). A population density grid of the European Union. Population and Environment, 31 (6), 460–473. DOI: https://doi.org/10.1007/s11111-010-0108-y

Gallego, F. J., Peedell, S. (2001). Using CORINE Land Cover to map population density.

Towards agri-environmental indicators. EEA Topic report 6/2001, pp. 94–105.

Horanont, T., Shibasaki, R. (2010). Estimate ambient population density: discovering the current fl ow of the city. https://www.academia.edu/2004297/estimate_ambient_population_density_discovering_the_current_flow_of_the_city, accessed 29. 5. 2017.

Lwin, K., Murayama, Y. (2009). A GIS Approach to Estimation of Building Population for Micro-spatial Analysis. Transactions in GIS, 13 (4), 401–414. DOI: https://doi.org/10.1111/j.1467-9671.2009.01171.x

Maantay, J. A., Maroko, A. R. (2009). The Cadastral-based Expert Dasymetric System (CEDS) for Mapping Population Distribution and Vulnerability in New York City. Proceedings of the IUSSP International Population Science Conference, Marrakesz, Maroko.

Maantay, J. A., Maroko, A. R., Herrmann, Ch. (2007). Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). Cartography and Geographic Information Science, 34 (2), 77–102. DOI: https://doi.org/10.1559/152304007781002190

Mennis, J. (2003). Generating Surface Models of Population Using Dasymetric Mapping. The Professional Geographer, 55 (1), 92–102. DOI: https://doi.org/10.1080/00330124.2015.1033669

Mennis, J., Hultgren, T. (2006). Intelligent dasymetric mapping and its application to area interpolation. Cartography and Geographic Information Science, 33 (3), 179–194, DOI: https://doi.org/10.1559/152304006779077309

Pirowski, T., Drzewiecki, W. (2012). Mapa gęstości zaludnienia Krakowa, propozycja metodyki opracowania oraz przykładowe zastosowania. Roczniki Geomatyki, 10 (3).

Pirowski, T., Pomietłowska, J. (2017). Modelowanie rozmieszczenia ludności Krakowa metodą dazymetryczną z wykorzystaniem Urban Atlas i ogólnodostępnych danych statystycznych, Geomatics and Environmental Engineering, 11 (4), 83–95. DOI: https://doi.org/10.7494/geom.2017.11.4.83

Reibel, M., Bufalino, M. (2005). Street-weighted interpolation techniques for demographic count estimates in incompatible zone systems. Environment and Planning A, 37 (1), 127–139. DOI: https://doi.org/10.1068/a36202

Robinson, A. H., Morrison, J. L., Muehrcke, P. C., Kimerling, A. J., Guptill, S. C., (1995). Elements of cartography. 6th edition. New York: John Wiley & Sons Inc

Sleeter, R., Wood, N. (2006). Estimating daytime and night time population density for coastal communities in Oregon. Urban and Regional Information Systems Association. Annual Conference, Proceedings, Vancouver, BC, September 26–29.

Smith, A., Newing, A., Quinn, N., Martin, D., Cockings, S., Neal, J. (2015). Assessing the Impact of Seasonal Population Fluctuation on Regional Flood Risk Management. ISPRS International Journal of Geo-Information, 4 (3), 1118–1141, DOI: https://doi.org/10.3390/ijgi4031118

Sridharan, H., Qiu, F. (2013). A Spatially Disaggregated Areal Interpolation Model Using Light Detection and Ranging-Derived Building Volumes. Geographical Analysis, 45 (3), 238–258. DOI: https://doi.org/10.1111/gean.12010

Tapp, A. F. (2010). Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. Cartography and Geographic Information Science, 37 (3), 215–228. DOI: https://doi.org/10.1559/152304010792194976

Tobler, R. T. (1979). Smooth pycnophylactic interpolation for geographic regions. Journal of the American Statistical Association, 74 (367), 519. DOI: https://doi.org/10.2307/2286968

Ural, S., Hussain, E., Shan, J. (2011). Building population mapping with aerial imagery and GIS data. International Journal of Applied Earth Observation and Geoinformation, 13 (6), 841–852. DOI: https://doi.org/10.1016/j.jag.2011.06.004

Wang, J., Wu, Z., Wu, C., Cao, Z., Fan, W., Tarolli P. (2017). Improving impervious surface estimation: an integrated method of classification and regression trees (CART) and linear spectral mixture analysis (LSMA) based on error analysis. GIScience & Remote Sensing, 55 (4), 583–603.DOI: https://doi.org/10.1080/15481603.2017.1417690

Wu, C., Murray, A. T. (2005). A cokriging method for estimating population density in urban areas. Computers, Environment and Urban Systems, 29 (5), 558–579. DOI: https://doi.org/10.1016/j. compenvurbsys.2005.01.006

Zandbergen, P., Ignizio, D. (2010). Comparison of dasymetric mapping techniques for small-area population estimates. Cartography and Geographic Information Science, 37 (3):199–214. DOI: https://doi.org/10.1559/152304010792194985
www.msip2.um.krakow.pl/statkrak/, Web sites StatKrak, accessed 18. 9. 2017.

Pirowski T., Bartos K. (2018). Detailed mapping of the distribution of a city population based on information from the national database on buildings. Geodetski vestnik, 62 (3), 458-471. DOI: https://doi.org/10.15292/geodetski-vestnik.2018.03.458-471

Tomasz Pirowski, Ph.D.
AGH University of Science and Technology,
Faculty of Mining Surveying and Environmental Engineering
Al. Mickiewicza 30, C-4 p.211
PL-30-059 Kraków, Poland,
e-mail: pirowski@agh.edu.pl

Karolina Bartos, M.Sc.
AGH University of Science and Technology,
Faculty of Mining Surveying and Environmental Engineering
Al. Mickiewicza 30, C-4 p.211
PL-30-059 Kraków, Poland,
e-mail: karolinabartos23@gmail.com