Polish Cartographical Review
Vol. 57, 2025, pp. 58–74
DOI: 10.2478/pcr-2025-0004
Received: 03.04.2025
Accepted: 29.05.2025

NATALIA DZIUBA, STANISŁAW SZOMBARA
AGH University of Krakow
Krakow, Poland
orcid.org/0009-0001-0070-1988; ndziuba@agh.edu.pl
orcid.org/0000-0002-0205-7823; szombara@agh.edu.pl

Supporting the editing of dot maps using the spectral clustering algorithm

Abstract. Automation of map production is an important subject of work of many scientists. Particular attention should be paid to dot distribution maps, the editing of which is very time-consuming and complicated due to the lack of support in GIS programs. The aim of this research was to develop a method supporting automation of dot map creation. In the study, equal-size spectral clustering algorithm was used, which was modified with a function equalizing the number of points in clusters. Spatial data on residential buildings and statistical data on the population were integrated to calculate theoretical population distributions. These data were entered into the spectral clustering algorithm based on a predefined dot value. The output clusters were then visualized in ArcGIS Pro, where manual adjustments, such as the definition of the dot size and the dispersion of overlapping markers, completed the map editing process. The results showed that the algorithm successfully created clusters representing the population distribution with an acceptable margin of error of the dot map of less than 5% for the entire county (study area – County of Pszczyna, Poland). The adaptation of the equal-size spectral clustering algorithm for cartographic purposes shows its potential to support automation of the dot distribution map editing process. The study found that reducing the input dot value slightly below the target value improved clustering precision, resulting in more consistent clusters. Despite these successes, the method has limitations, including partial reliance on manual corrections in densely populated areas where overlapping dots could not be fully automatically resolved. These issues underscore the need for further refinement to achieve full automation.

Key words: cartography, dot maps, GIS, clustering, unsupervised learning

1. Introduction

Nowadays, practically everything is subject to automation. This is due to the extremely fast and dynamic development of technology. In every area of life, people try to improve manual processes by automating them in order to save time and energy. A high level of process automation also allows for a significant increase in productivity. This is also an extremely important issue in cartography, because it allows editors to develop maps using a much smaller amount of time. GIS programs are very helpful in editing thematic maps, made using methods such as choropleth map, diagram map or isopleth map. However, the situation is different in the case of editing dot distribution maps – this study raised the problem and attempted to solve it.

The dot method is one of the quantitative methods of presenting data on maps, which allows to show the spatial distribution of a phenomenon (Allen, 2022; Pieniążek & Zych, 2020). Unlike the signature method, in this method, dots are assigned a previously established value of the phenomenon (or several values, if we are dealing with a multi-value dot method, Ratajski, 1989). Using the dot method, the distribution of absolute data is visualized on the map. The number of dots per area symbolizes the accumulation of the occurrence of the studied phenomenon, assuming the appropriate value of one dot (Suchecka, 2014). There are two methods of placing dots on a map – topographic, which is based on the distribution of dots in the places where the phenomenon occurs, and uniform, which involves random or regular placement of an appropriate number of dots in the reference area (Pasławski, 2015; Tennekes, 2018).

During the editing of a dot map, it is important to determine the correct value of the dot. This has a direct impact on the perception of the map. The value of the dot should be selected in such a way that it is as small as possible, but large enough to allow for counting the dots (Pasławski, 2015). Then the condition of countability of the dot map is met. If the cartography selects too high a value, the map may lose the topographic information conveyed because the dots may not correctly show the clusters of the studied phenomenon. An equally important aspect is the size of the dots. It is assumed that the minimum diameter should not be smaller than 0.4 mm, although the minimum range of 0.5–0.7 mm is more suitable for the human eye (Ratajski, 1989). Therefore, the dots cannot be small to be visible in large areas with low density. However, in areas with a high density of dots, care should be taken to ensure that they do not merge with each other. In order to eliminate this problem, multi-value dot method can be used (Pasławski, 2015). It is important to choose the value and size of the dot so that the differentiation into areas of low and high density of the phenomenon is simultaneously clearly visible to the reader. Sometimes reading a specific amount of the phenomenon is relegated to the background and the possibility of reading the spatial patterns of the distribution of the phenomenon is more important (e.g. characteristic differences in rural areas in different regions).

The problem with dot distribution maps is the lack of generally available methods for their automation. Other cartographic presentation methods are supported by computer programs, thanks to which maps using these methods can be created by people who are not specialized in cartography (Hey, 2011). For many years, researchers have been trying to solve the problem related to the size, value and distribution of dots on the map. One possible approach to supporting the editing of dot maps was presented by Mackay (1949). Mackay's nomogram allows for determining the coalescence limit, which means the merging of dots on the map using the size and number of dots per square centimetre. However, this method is not helpful in today's era of digital cartography. Currently, in order to use the nomogram, it should be appropriately transformed, as suggested by Kimerling (2013).

Attempts to automate dot placement have been discussed in the literature many times. One of them is a method based on determining cluster polygons and then placing dots in these areas in a non-randomized manner using a spiral vortex (Hey, 2011; Hey & Bill, 2014). This method maximizes the number of dots that can be placed in a given cluster. The method has been described in more detail in the literature . Another approach to supporting the automation of dot map editing is to divide the studied area using hexagonal tessellation (Szombara, 2017). The entire area of the dot map is divided into hexagons, and the algorithm assigns a number of inhabitants to each hexagon. If number of inhabitants is equal to or greater than the assumed value of the dot, then one or more centroids are created inside the hexagon. If this number is smaller, the hexagons are aggregated to neighbouring areas, and the centroid is created inside the merged area. An attempt was also made to use iterative and clustering algorithms, applied to previously distinguished areas of occurrence of the phenomenon. The research was described in more detail in the literature by de Berg et al. (2004). The graphical principles of designing dot maps were also investigated by Yan and Weibel (2008), who performed a cluster generalization analysis using Voronoi diagrams.

This study focused on an attempt to automate the process of editing dot maps of population distribution. The aim was to find a new, innovative way to automate and speed up the process of creating this type of thematic maps by cartographers. Efforts were made to skip, among others, the hexagonal tessellation stage, but in such a way as to obtain satisfactory results. The study was carried out on a selected test area, for which a spectral clustering algorithm was implemented, modified by the function of equalizing the number of points in clusters (hereinafter, the term “points” refers to points in algorithm clusters, term “dots” refers to dots on a dot map). This allowed for semi-automation of the map editing process, supplemented with manual activities in the ESRI ArcGIS Pro 3.3.3.

The research methodology was presented on an example map of the selected region of interest. This approach allows for a better understanding of the presented algorithm, which is why a case study chapter was not separated.

2. Map examples in the literature context

Nowadays, dot maps are mostly used on small-scale maps, where the arrangement of dots does not require their precise location. Due to the high value of dots, information about the detailed characteristics of the phenomenon becomes insignificant. It is only possible to present its general outline. An example here is the Distribution of rural population in 2016 (2018, p. 25, Figure 1b). One map using the dot method presented the distribution of population in rural areas. The dot value was assumed to be 500 people. On the one hand, such a value allowed for the representation of small towns in a location close to the actual one, but on the other hand, it caused the dots to completely fill the available area in densely populated areas.

The use of a larger dot value makes it possible to read the location of individual dots, but at the cost of their much greater aggregation. This approach was used in the Geographic Atlas (Rozmieszczenie ludności wiejskiej, 2023, p. 201, Figure 1a). In both cases, especially with a large dot size and small scale, it is possible to use administrative divisions as reference fields for dots. An example of a map that uses a relatively small dot size is the Atlas of the Republic of Poland (Pasławski et al., 1993–1997, ark. 62.1, Figure 1c). The map at a scale of 1:1,500,000 uses a dot value of 300 people. This allows the reader to infer the location of even small areas with residential development, but areas of high rural population density are filled with dots covering the entire area.

It is now relatively uncommon to find dot maps at medium and large scales with dots of relatively small value. An example is the map depicting the population distribution of Tirana in 2001. On the 1:75,000 scale map, a dot value of 100 was used (Bërxholi et al., 2003, as cited in Asche, 2023, p. 36).

3. Data sources and research area

The basis for automatic dot generation in this research was spatial data on residential buildings and statistical data about them. Spa-

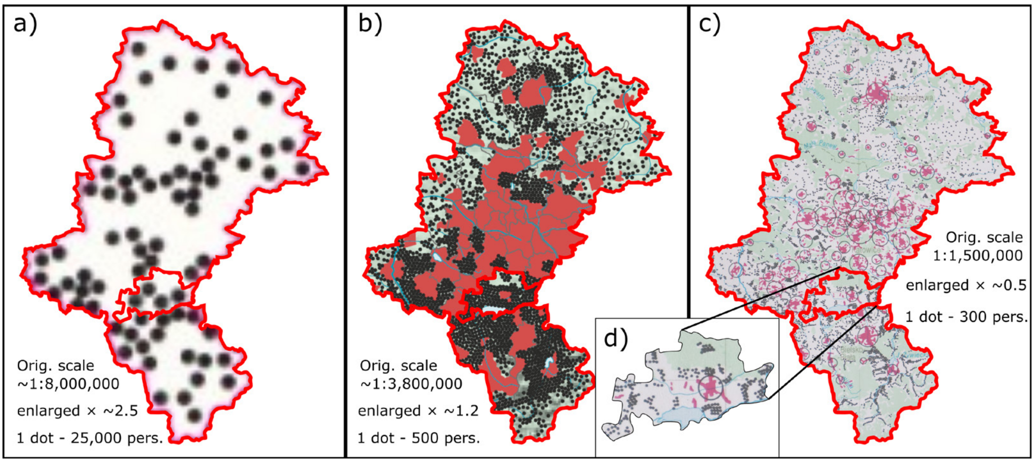

Figure 1. Distribution of rural population in: a) the Geographical Atlas (source: own study, based on: Rozmieszczenie ludności wiejskiej, 2023, p. 201), b) the Statistical Atlas of Poland (source: own study, based on: Distribution of rural population in 2016, 2018, p. 25), c) and d) the Atlas of the Republic of Poland (source: own study, based on: (Pasławski et al., 1993–1997, ark. 62.1). The border of the Silesian voivodship and the Pszczyna county was superimposed on the original maps.

tial data of buildings located in the study area came from the Topographic Object Database (in Polish: Baza Danych Obiektów Topograficznych, BDOT10k), valid as of December 31, 2023. This database is a collection of topographic objects along with their descriptive characteristics and contains the following categories of object classes: water network, communication network, land infrastructure network, land cover, protected areas, territorial division units, buildings, structures and devices, land use complexes and other objects (Rozporządzenie, 2021b) from which only single-family residential buildings with code 0010_317_1 and multi-family residential buildings with code 0010_318_1 were distinguished. In the study area, high-rise residential buildings (code 0010_319) were not distinguished due to their lack of occurrence.

The building class was supplemented with data on the number of residents of each facility. For this purpose, data on the number of inhabitants and housing resources in the communes and the entire county were obtained, valid as of 2022. This information was taken from the Local Data Bank (in Polish: Bank Danych Lokalnych, BDL), which is the largest collection of data on the demographic or socio-economic situation in the country, maintained by the Statistics Poland (in Polish: Główny Urząd Statystyczny, GUS). Information on housing resources included information such as the number of buildings, the number of apartments and the average area of one apartment in the communes.

The region of interest in this research was the Pszczyna county in the Silesian voivodship in Poland. The county consists of five rural communes: Goczałkowice-Zdrój, Kobiór, Miedźna, Pawłowice, Suszec, as well as one urban-rural commune – Pszczyna with the city of the same name. Selecting such a region allowed us to analyze the algorithm's performance in areas with varying population density. In rural regions, inhabited areas are distributed along main streets or occur in small clusters, creating a sparsely built-up area. In the urban area, on the other hand, there are clusters of high density creating a dense and compact built-up area. Knowledge of the population distribution in the analysis area allows us to estimate the resulting distribution of dots on the edited map. The area of the county is 471 km^2 (Poland in numbers, 2024), and the population according to the Statistics Poland as of December 31, 2022 was 110 958 people (Statistics Poland), which gives the population density of the county at the level of 236 people/km^2. The urbanization rate is 23.9% (Poland in numbers, 2024), which is a value much lower than the urbanization percentage rate for the entire Silesian voivodship.

The boundaries of the county, the communes in the county, as well as the precincts in the communes used in the work, come from the State Register of Borders (in Polish: Państwowy Rejestr Granic, PRG). This register contains data on the borders (and areas) for the entire country. It consists of information on voivodeships, counties, communes, units, precincts and special boundaries. Additionally, this register contains information about addresses and their location, and descriptive attributes are attached to all spatial data (Rozporządzenie, 2021a).

4. Methodology

The first step was to prepare data on buildings in the region of interest along with housing and population data. The next step was to implement a modified spectral clustering algorithm to group the population into communes. At this stage, the dot value was defined. In the third stage, the dot size was manually set in ArcGIS Pro. For correct visualization, it was also necessary to use the geoprocessing function – disperse markers, which was supposed to eliminate overlapping dots. In the next stages, the influence of the introduced parameters defining the dot value on the operation of the clustering algorithm was analysed. The obtained results were analysed and the dot map error was calculated. In the last stage, the final maps were edited using the ArcGIS Pro program.

The diagram below on Figure 2 shows the subsequent stages of dot map editing with partial automation used in this paper.

4.1. Data preparation

Before implementing the clustering algorithm for the study area, the data was prepared to reflect the population distribution of the county. For this purpose, a theoretical number

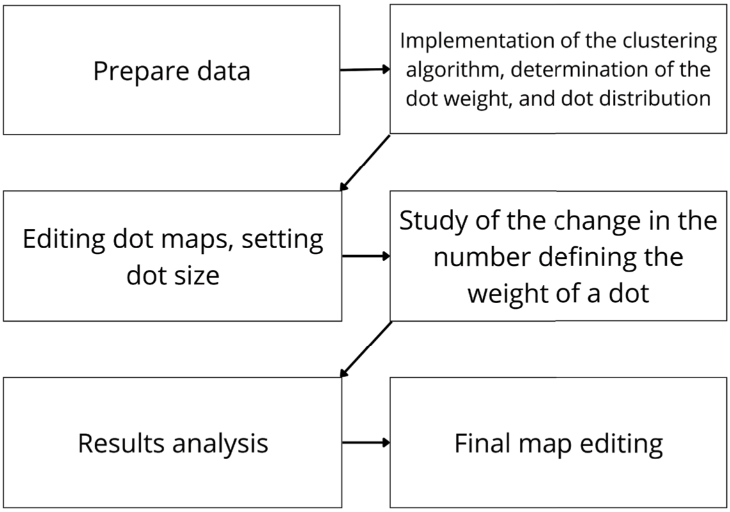

Figure 2. Dot map editing steps performed in this research (source: own study)

of residents was assigned to each building based on data from building attributes (area and number of floors) and statistical data from the Statistics Poland. The approach used is analogous to that described in the literature by Szombara (2017). Data presenting the exact distribution of the population are held by the Statistics Poland, but are subject to statistical confidentiality.

First, data on the average usable area of flats in individual communes of Pszczyna county were used, using data on housing resources of the Statistics Poland. This information was then assigned to each building. The attribute that linked the data were the TERYT codes of the communes. In this way, each object from the class of residential buildings had an assigned value of the average area of the apartment.

In the next stage, the theoretical number of apartments in the buildings was determined. In the case of single-family residential buildings (code 0010_317_1), the number of apartments was set as 1. For multi-family residential buildings (code 0010_318_1), this value was calculated. For this purpose, building attributes such as the area of the building (Shape_Area) and the number of floors in the building (liczbaKondygnacji) were used, as well as the previously assigned average area of an apartment in a given commune. Calculations were performed according to formula (1):

$$\textit{theoretical number of flats} = \frac{\textit{buildings area} \times \textit{number of floors}}{\textit{average flat area}} \quad (1)$$

In order to check the calculation results, the results of the calculation of the number of apartments were compared with the actual number of all apartments in the communes. These results are presented in Table 1. The average relative error, calculated using formula (2), was 2.73%, while the maximum was – 7.76%. This allows us to state that the results are acceptable for cartographic purposes.

$$\textit{relative error}\ [\%] = \frac{\textit{calculated value} - \textit{real value}}{\textit{real value}} \times 100\% \quad (2)$$

The average number of people per apartment was calculated based on data on the number of people in the communes and the number of apartments calculated according to the formula (3):

$$\textit{average number of people in a flat} = \frac{\textit{number of people}}{\textit{number of flats}} \quad (3)$$

Table 1. Comparison of the calculated theoretical number of flats in communes and the real number of flats (source: own study, data from the Statistics Poland)

Commune name	Calculated theoretical number of flats	Real number of flats	Relative error [%]
Goczałkowice-Zdrój	2,189	2,180	0.41
Kobiór	1,803	1,812	–0.5
Miedźna	4,765	5,048	–5.61
Pawłowice	4,899	5,311	–7.76
Pszczyna	17,088	16,837	1.49
Suszec	3,580	3,559	0,59
		Average relative error [%]:	2.73

The last stage of the calculations was to calculate the theoretical number of residents in the building. For this purpose, the theoretical number of apartments in the building was multiplied by the average number of people living in each apartment, which results from the formula (4):

$$\textit{theoretical number of residents} = \textit{theoretical number of flats} \times \textit{average number of people in a flat} \qquad (4)$$

Assigning a theoretical number of residents to each building allowed us to obtain data that presents the theoretical topographic distribution of the population in the study area. Based on this information, centroids were generated for each building, in a number corresponding to the number of residents in the building. The data (points) prepared in this way was used in the next step to perform clustering and create a dot distribution map.

4.2. Spectral clustering algorithm

In this research the equal-size spectral clustering was applied. This algorithm is an extension of standard spectral clustering, designed to create clusters with a specified number of points rather than simply determining the optimal number of clusters (Figure 3). It has been described in more detail in the literature (Fleshman, 2019; Martínez Barbosa, 2023b) and the source code can be found in the GitHub repository (Martínez Barbosa, 2023a). Traditional clustering algorithms focus on parameters like the number of clusters, but they do not ensure an equal distribution of points within those clusters. This modification is particularly useful in various applications, such as editing dot distribution maps, where maintaining a balanced cluster size is crucial due to the predetermined dot value.

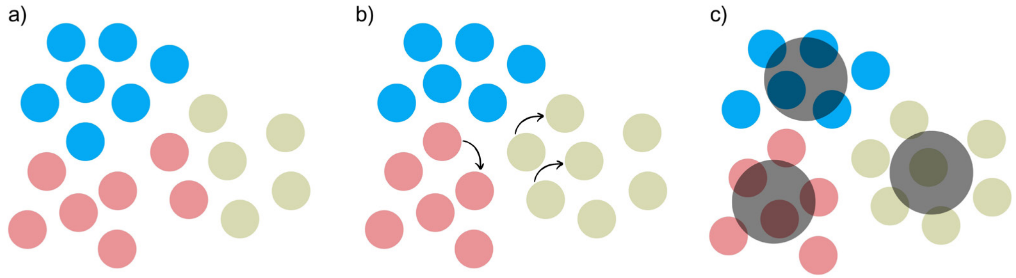

Figure 3. Flow chart of the algorithm: a) spectral clustering, b) equalizing the number of points in clusters, c) creating cluster centroids corresponding to the dots of the dot map (source: own study, based on: Martínez Barbosa, 2023b)

The fundamental aspect of the algorithm is the ability to equalize the number of points across clusters, which is controlled by the *equity_fraction* parameter. This parameter defines the extent to which cluster sizes should be balanced, taking values between 0 and 1. When set closer to 1, the algorithm strongly enforces uniformity, ensuring that clusters are as equal as possible. Conversely, lower values allow for greater variation in cluster sizes. The equalization process is achieved through the redistribution of points, where clusters that contain more points than necessary transfer excess elements to neighbouring clusters with fewer points. This step is performed iteratively, ensuring a gradual balancing of the clusters while maintaining their spatial integrity. The ability to adjust cluster sizes dynamically makes this algorithm highly adaptable to different use cases, particularly those requiring a uniform spatial distribution of data points.

Another crucial parameter of the algorithm is *n_neighbors*, which determines how many neighbouring points are considered during clustering. According to the literature, it is recommended to set this value within 7–15% of all data points (Martínez Barbosa, 2023b). However, in this specific case, it is assumed that the number of neighbours corresponds to the predefined dot value. This parameter plays a vital role in the initial formation of clusters, as it influences how connections between data points are established in the adjacency matrix. A well-chosen *n_neighbors* value ensures that clusters remain spatially coherent and that the redistribution of points during equalization does not lead to distortion of the dataset.

The last step of the algorithm to generate the centroids of each cluster. These centroids were used as dots to edit the dot distribution maps. When entering parameters into the algorithm, the value of the dot was determined (the size was determined in a later manual step). Two dot values were used to analyse and select the appropriate value for the edited map.

4.3. Manual map editing

The implemented algorithm does not provide full automation of dot map editing, so the last part of the work was done manually. At this stage, the dots were given a size. Several dot sizes were used in the work to compare and choose the most suitable for the intended map scale of 1:150,000. This comparative approach ensured that the selected dot size would maintain clarity and readability without overwhelming or underrepresenting the mapped data.

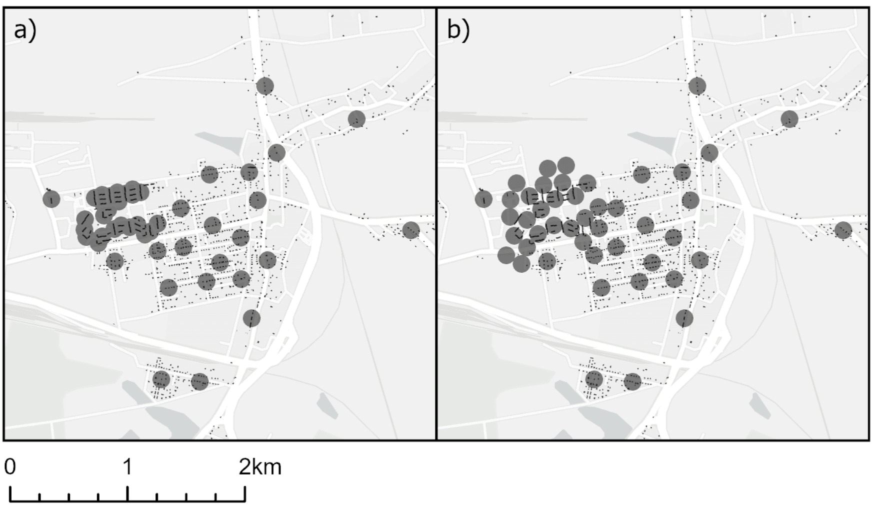

Figure 4. Fragment of the visualization in ArcGIS Pro a) before applying “disperse markers” and b) after applying “disperse markers”, dot value 200, size 1 mm (source: own study)

For correct visualization, it was also necessary to use the geoprocessing function – *disperse markers*, which eliminated overlapping dots and improve spatial distribution (Figure 4). This function was launched with the *expanded* dot spread parameter and at a minimum distance of 0 mm, which means that objects touch each other. Several alternative dispersion methods were explored during the process, yet only the chosen approach produced a natural and visually coherent distribution of dots. As a result, the dots formed clusters in which they did not overlap but were closely adjacent to each other.

5. Results

5.1. Results of the algorithm – calibration parameters for edited maps

After performing clustering, a base for creating dot maps was obtained, but before further analyses, the algorithm result had to be assessed. The distribution of dots with a value of 200 in the region of interest was assessed in relation to the actual population distribution in the Pszczyna county. It was necessary to check whether the dots on the edited map were located in places where buildings inhabited by an appropriate number of residents were located in comparison to the assumed value of the dot. This comparison was carried out by applying a hexagonal tessellation to the study area, assigning each of the hexagons a corresponding number of inhabitants inside (calculated based on housing and building data as before).

The figure below (Figure 5) shows a fragment of a map in sparsely built-up area. Considering that the dots are located in thus areas in places with dense population, it can be stated that in built-up areas this type of algorithm does not cause large errors. However, it should be noted that the dots in areas with lower population density symbolize clusters created from about 200 neighbouring points-residents, which, despite being far apart, were assigned to the same set. This results from the principle of the algorithm, which does not eliminate noise elements, i.e. in this case individual, distant points. This is also consistent with the general principle of generalization in dot maps.

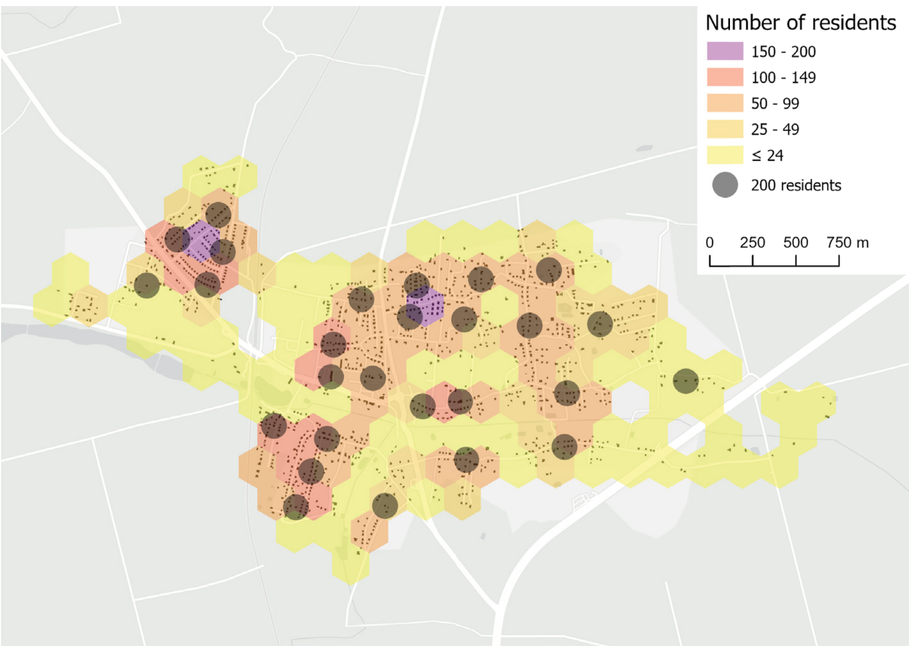

Figure 5. Part of the study area showing sparsely built-up area, dot value 200, size 1 mm, against a background of a hexagonal tessellation grid representing local population density (source: own study)

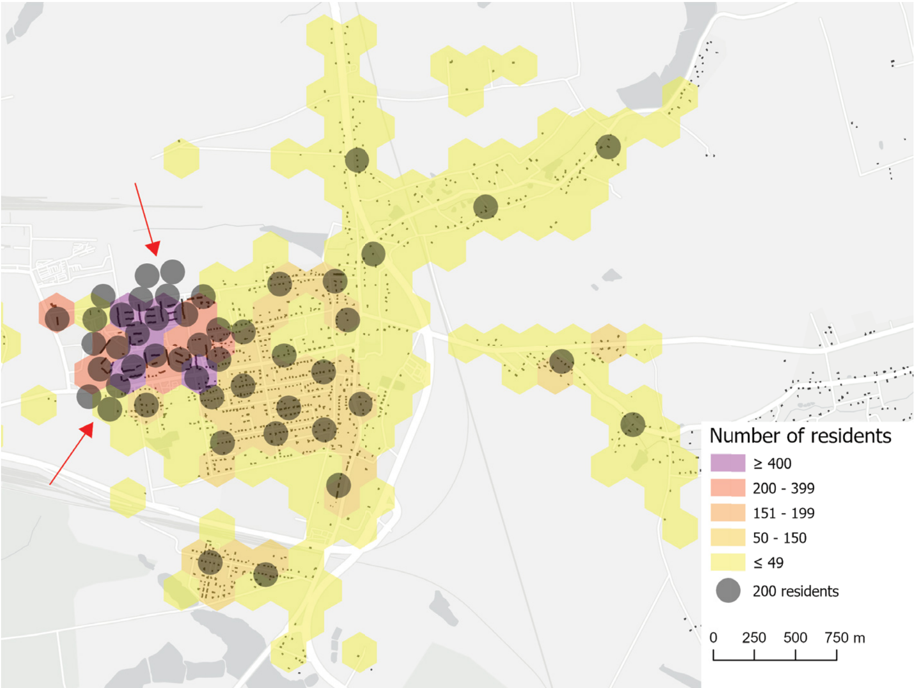

Figure 6. Part of the study area showing compact built-up area, dot value 200, size 1 mm, against a background of a hexagonal tessellation grid representing local population density; red arrows indicate dots moved outside the built-up areas (source: own study)

The visualization is different for dense or compact built-up areas. In this case, the dots are generated by the program in practically the same places, which is caused by the large accumulation of input objects to the clustering algorithm. In order to avoid this problem, it was necessary to spread them so that they do not overlap, but only touch by their edges. It should be noted, however, that after scattering the dots so that they do not overlap, errors occur.

The Figure 6 shows an example fragment of such a compact bult-up area, where arrows indicate dots that have been moved outside the area of buildings and population. However, it was not possible to automatically place the dots so that they are in the right places, because the previously mentioned disperse markers tool did not allow for this. In this case, it would probably be necessary to correct the position of the dots manually. Some of these errors cannot be eliminated, due to the nature of the dot method. They will affect the overall error of the method, but will not interfere (assuming the correct selection of the value and dot size) with the correct reception of the edited map by the user.

While working on automating map editing, statistics of the theoretical number of inhabitants in clusters were calculated to check whether the clustering algorithm actually strives to achieve equinumericity. The results are presented in the Figure 7. The histograms highlight 200 and 150 points as target values, the median and the average number of points in the clusters. The most numerous intervals oscillate around the targets dot value. The algorithm slightly overestimates the number of points in clusters when performing calculations.

A study was conducted to see how the average number of points in clusters and their median changes when the number of dots entered into the algorithm changes. It was assumed that the target dot values were to be 200 and

150, which means that one dot was to symbolize 200 and 150 residents, respectively. The algorithm should therefore strive to obtain such a number of points in clusters. However, the results obtained show that the program overestimates the number of points in clusters.

The Table 2 contain the results of the experiment, which consisted of adopting the following

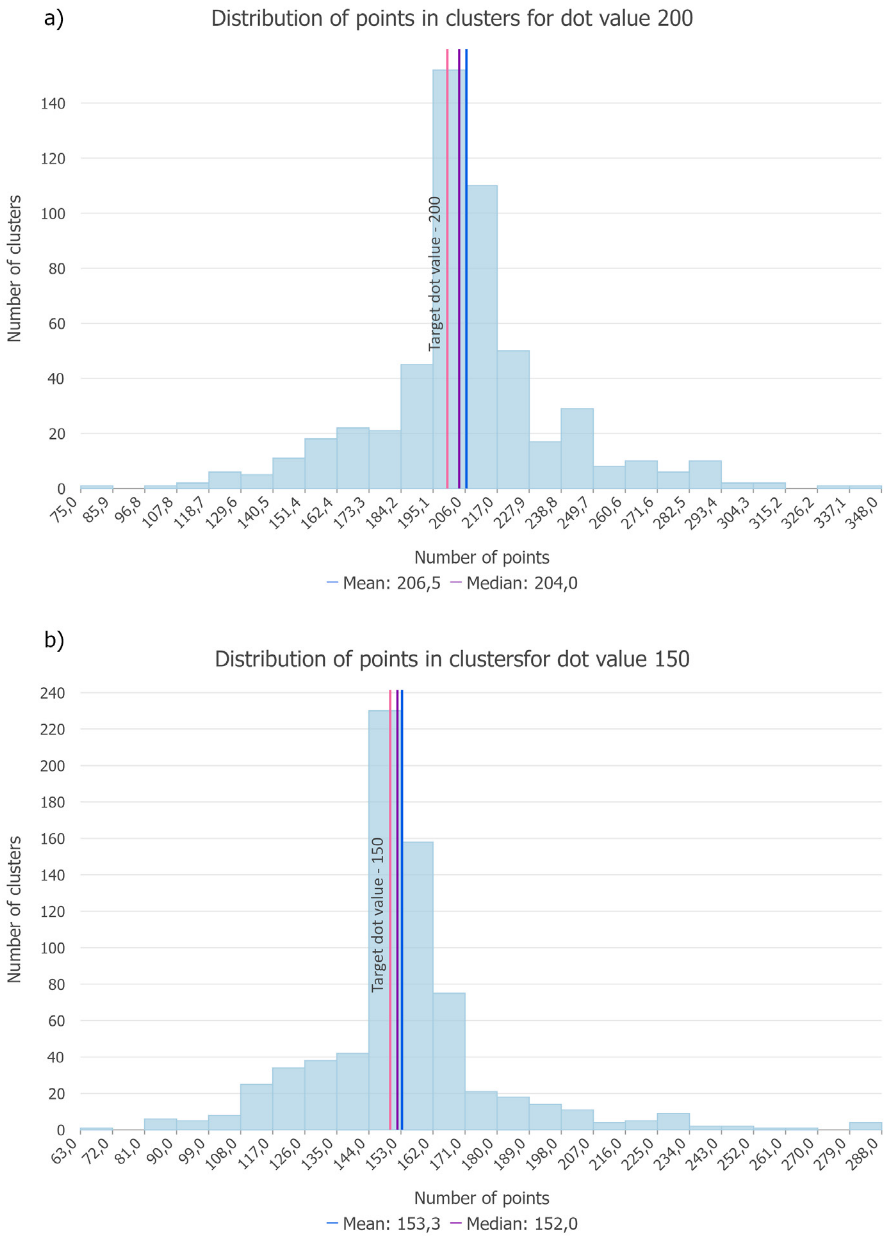

Figure 7. Histograms of the distribution of points in clusters a) for a dot value of 200 and b) for a dot value of 150 (source: own study)

Table 2. Summary of mean values and medians of points in clusters for individual entered dot values (source: own study)

Target dot value	Entered dot value	Average number of points in the cluster	Actual dot value (median)	Relative error	Number of all dots/clusters
	200	206.49	204	3.25	530
	199	205.71	204	3.37	532
	198	204.94	203	3.51	534
200	197	204.57	203	3.84	535
	196	201.92	200	3.02	542
	195	201.54	199	3.35	543
	194	199.71	197	2.94	548
			Average relative error [%]	3.33	
	150	153.28	152	2.19	714
	149	152.63	151	2.44	717
	148	151.58	151	2.42	722
150	147	150.33	149	2.27	728
	146	149.71	148	2.54	731
	145	148.69	147	2.54	736
	144	147.69	146	2.56	741
			Average relative error [%]	2.42	

values for calculations: 200, 199, 198, 197, 196, 195 and 194 (in the case of a target dot value of 200) and 150, 149, 148, 147, 146, 145 and 144 (in the case of a target dot value of 150). In addition to the average and median number of points in clusters (which is the actual value of the dot), the relative error was also calculated using the formula below (5), which allowed us to determine the level at which the algorithm overestimates (or underestimates) the average number of points in clusters in relation to the entered value.

$$\textit{relative error}\ [\%] = \frac{\textit{av. number of points in cluster} - \textit{entered dot value}}{\textit{entered dot value}} \times 100\% \qquad (5)$$

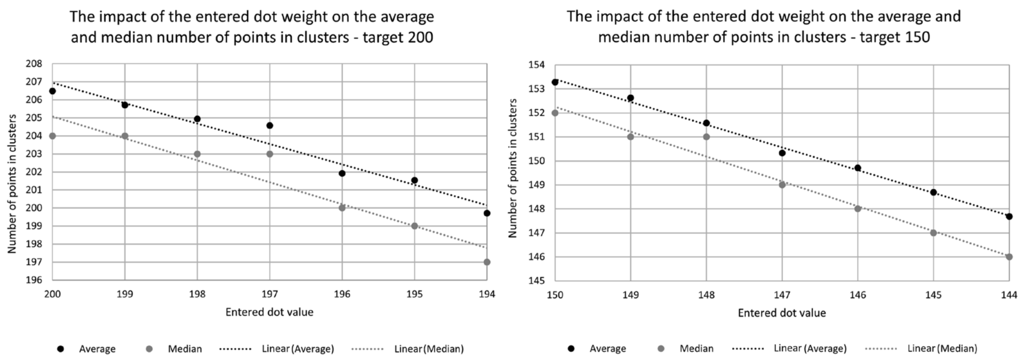

Figure 8. Graphs of the dependence of the dot value entered into the program on the diameter and median number of points in clusters with target dot values a) 200 and b) 150; trend lines were plotted on the graphs (source: own study)

The graphs (Figure 8) show how the number of points in clusters changed during algorithm iterations with subsequent values entered. There is a downward trend in the mean and median number of clustered points. The value closest to the expected one was obtained after entering the number 195 in the program for the target dot value of 200 and 147 – for the dot value of 150. The decrease in the number of points in clusters in both cases adopts a linear trend, in the case of smaller entered values it occurs faster, because the target number of clustered points in the case of the value 150 was reached much faster.

5.2. Dot map error

The focus was also on calculating the error of the dot map, which is a rather complicated problem, because it is not often described in the literature. Following Pasławski (2015), the results indicate that ± 0.5 dot value determines the error of data presentation in individual reference fields. It follows that dots can symbolize sets of objects in a number equal to the value of the dot and ± 50% of this value. Therefore, with a fixed dot value of 200, a dot can represent from 100 to 300 people. The histogram discussed earlier was analyzed again, but this time the focus was on outliers determined by the defined data presentation error. In the histogram (Figure 9), the bars are described by the number of clusters (and thus dots) with a given number of points covered by a given cluster. The arrows indicate the lower and upper thresholds denoting ± 0.5 dot value. The minimum threshold was exceeded by 1 dot, and the maximum by 6, which shows that the algorithm copes worse with too large groups.

The dot map error we calculated is defined as the quotient of the total population in a given reference unit (county or commune) and the number of dots multiplied by the dot value. This made it possible to illustrate the deficiencies and excesses of statistical data. For this purpose, formula (6) was used to calculate the percentage values of errors. Data on the population of the county and communes come from the Local Data Bank of the Statistics Poland.

$$\textit{dot map error}\,[\%] = 1 - \frac{\textit{population}}{\textit{number of dots} \times \textit{dot value}} \times 100\% \quad (6)$$

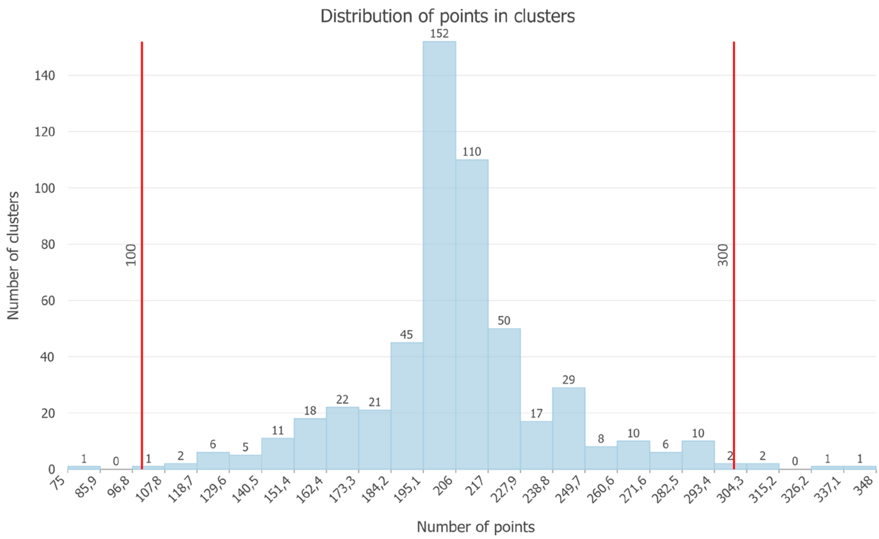

Figure 9. Histogram of the distribution of points in clusters for a dot value of 200 with described frequencies of occurrence of a given value, red lines indicate the limit values of the data presentation error (source: own study)

Table 3. Calculation results of the dot map error for the Pszczyna county and for each of the communes (source: own study)

County name	Dot value	Population (according to GUS)		Number of dots		Dot map error [%]	
		County	Commune	County	Commune	County	Commune
Goczałkowice-Zdrój	200	110,958	6,613	530	33	–4.68	–0.20
Kobiór			5,126		27		5.07
Miedźna			16,098		72		–11.79
Pawłowice			17,642		96		8.11
Pszczyna			52,785		252		–4.73
Suszec			12,694		50		–26.94
Goczałkowice-Zdrój	150	110,958	6,613	714	44	–3.60	–0.20
Kobiór			5,126		36		5.07
Miedźna			16,098		96		–11.79
Pawłowice			17,642		129		8.83
Pszczyna			52,785		340		–3.50
Suszec			12,694		69		–22.65

Table 3 presents the results of the calculation of the dot map error for the entire county, but also for each commune separately. In addition, two dot values were taken into account – 200 and 150. It can be seen that the error for the entire study area is smaller for the dot value of 150, however, both error values for the county are small and acceptable, because they are within a few percent. However, in the case of some communes, the error values are quite large – they reach the order of 20%, which gives an approximate deficiency of 1/5 of the number of people in the commune. This may result from the calculated number of people per building, which is a theoretical value resulting from statistical data. In communes that have multi-family buildings on their areas, these deviations may be greater than for less developed areas with a predominance of single-family buildings. Moreover, in communes with a small population, the dot value does not have a significant impact on the dot map error. The above errors may also result from an incorrect estimation of the number of people using data on buildings from BDOT10k.

Additionally, calculations of the dot map error were performed for different values of dots entered into the algorithm. The results do not indicate a dependence of the dot map error on a decreasing entered value. However, it should be noted that there is a certain dependence between the map error and the relative error given in the table below (Table 4), which was calculated from formula (5). The dot map error increases with the increase of the relative error, which determines the level of deviation of the average number of points in the cluster from the target value, but these errors take opposite signs.

5.3. Final maps

Four dot maps were made using two dot values in order to compare the algorithm's performance with different input values and to compare the correctness of cartographic visualization. The value of the dot and its size had to be selected in such a way that the graphic load of the map was not too large, and the dots correctly illustrated the distribution of the population in the studied area. When analysing the graphic clarity of the composition, it should be noted that maps for which 1 mm diameter dots were used are characterized by a greater graphic load of the analysis area. Considering the value of the dot, using a value corresponding to the number of 150 inhabitants, there are significantly more dots in dense built-up regions,

Table 4. Dot map error calculation results for different entered dot values for target dot values 200 and 150 (source: own study)

Target dot value	Entered dot value	Average number of points in the cluster	Median number of points in the cluster	Number of all dots	Relative error	Dot map error [%]
200	200	206.49	204	530	3.25	–4.68
	199	205.71	204	532	3.37	–4.81
	198	204.94	203	534	3.51	–4.94
	197	204.57	203	535	3.84	–5.28
	196	201.92	200	542	3.02	–4.45
	195	201.54	199	543	3.35	–4.79
	194	199.71	197	548	2.94	–4.37
150	150	153.28	152	714	2.19	–3.60
	149	152.63	151	717	2.44	–3.86
	148	151.58	151	722	2.42	–3.84
	147	150.33	149	728	2.27	–3.68
	146	149.71	148	731	2.54	–3.97
	145	148.69	147	736	2.54	–3.97
	144	147.69	146	741	2.56	–3.99

which makes the map difficult to read and makes it impossible to count the dots. This problem could be solved by introducing multi-valued. In urban areas, a new, higher dot value should be established than in rural areas.

Figure 10 presents a comparison of fragments of maps with the use of different parameters.

6. Discussion

The aim of the work was to find a way to automate the editing dot distribution maps. The limited application of dot maps has been attributed to the complexity of their generation and insufficient implementation in standard GIS tools. It is not possible to automatically generate dot maps with a topographical arrangement of dots in programs such as ArcGIS Pro or QGIS.

The research used a spectral clustering algorithm extended with an additional function that equalized the number of cluster objects. After specifying the target dot value in the program, the number of clusters was calculated, followed by spectral clustering, and then group sizes were equalized. Finally, the program generated cluster centroids, which were to be the dots of the edited dot map.

The visualizations were analysed using different dot values and sizes. The main goal was to ensure that the graphic load of the map was not too high. With a lower value of the dot, too large clusters of dots were created, which made it impossible to count them. This problem could be solved by introducing multi-valued of dots (higher value for urban/densely populated areas). However, that the value would be too high and there would be too few dots, this could lead to an incorrect interpretation of the distribution of the phenomenon on the map. In the case of dot size (i.e. diameter in millimetres) overly large dots darken the research area, hinder readability, and negatively affect the map's perception by users. Such an analysis leads to the conclusion that the best option for a dot map of the population distribution of Pszczyna county at a scale of 1:150,000 can be considered to be the selection of a dot value of 200 inhabitants and a diameter of 1 mm (Figure 11). On the final map it was also possible to see how the algorithm handles the

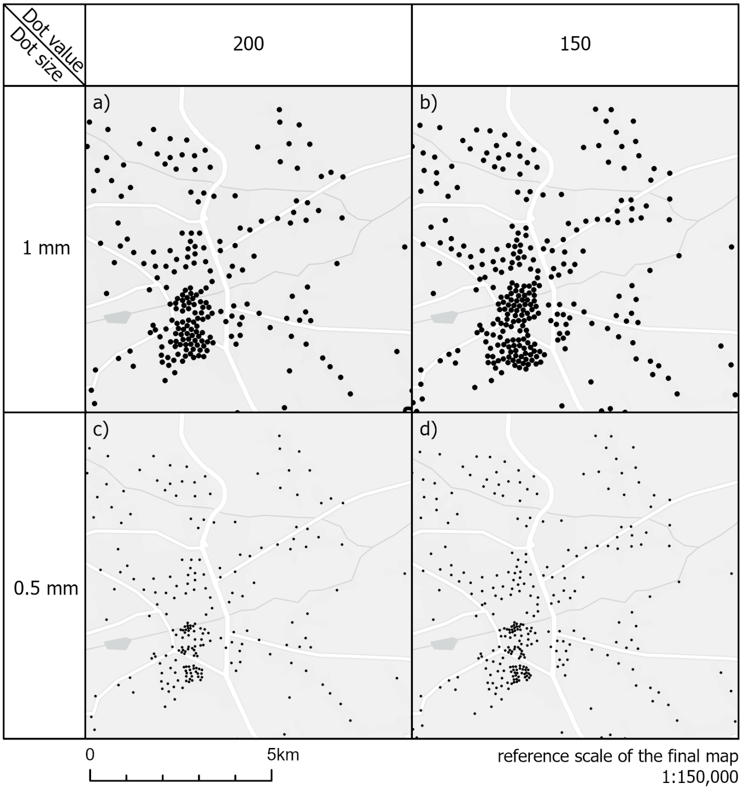

Figure 10. Comparison of map parts with different values and dot sizes (source: own study)

settlement patterns of villages. In some communes, such as Suszec, Miedźna and in the eastern part of the Pszczyna commune, settlements were located along the road, which was well illustrated on the map. In other communes, the population distribution was not so characteristic and was located irregularly or in larger clusters (Kobiór, Pawłowice).

An analysis of the algorithm's operation was also carried out in terms of introducing a lower value of the dot value as a parameter for the clustering algorithm. This resulted from the fact that the algorithm overestimated the average number of points in clusters during calculations. The study showed that after entering a lower value, the program generates clusters with a number of points closer to the target. In this way, a map was developed (Figure 11) which shows the distribution of the population in the Pszczyna county, where the dot value is equal to 200, but the value actually entered into the calculation model was equal to 195.

This analysis was performed for target dot values of 200 and 150. In both cases, a decreasing linear trend was observed after lowering the input value. It is therefore possible that examining analogous relationships for other dot values would allow us to derive a formula for the input value that the algorithm should assume, and which could be calculated by the program after entering the expected dot value. For example, the user could enter a value of 200, while the script before clustering would calculate by how much the entered value should be reduced and enter it into the calculations.

The errors of the dot map were analysed in relation to the population of the entire county, but also to each of the communes. The error for the entire county was –4.68% for a dot value of 200 people and –3.60% for a dot value of

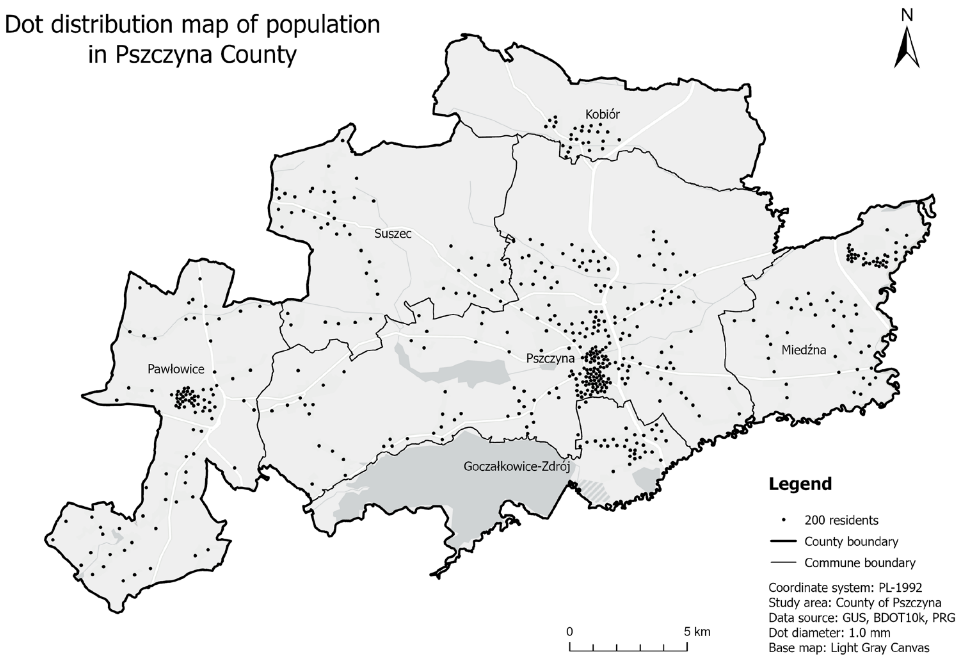

Figure 11. Dot distribution map of population in County of Pszczyna, Poland (source: own study)

150. The error value oscillating around a few percent means that the cartographic visualization developed in this way does not differ significantly from the actual statistical data. The values take on a negative sign, which means that the sum of the values of all dots is lower than the actual number of people living in the county according to the data of the Statistics Poland. The achieved error values should be considered acceptable.

In dot map editing, the dot value is crucial, and the number of dots does not always equal the total population divided by the dot value. For the clustering algorithm used in this work, the problem is therefore the parameter of the *nclusters* function, which requires defining the input number of clusters for calculations and cannot be omitted. This issue could potentially be addressed by modifying the algorithm so that it depends on the entered dot value rather than the number of clusters. An additional option could be to try to use another model, e.g. an agglomerative clustering algorithm. However, this algorithm should be extended, as was the case with spectral clustering – by an element equalizing the number of objects in groups.

7. Conclusion

After the conducted research and analysis of the results, it can be stated that the equal-size spectral clustering algorithm used in the work, which grouped points symbolizing the population into clusters of the most equal sizes, is a good alternative, among others, to the method of automation discussed in the literature by performing the hexagonal tessellation process (Szombara, 2017). It was possible to perform calculations multiple times using different configurations of input parameters. However, the program only helps in part of the process – the diameter of the dots was entered at a later stage in the ArcGIS Pro program. One could consider complete automation of the process, which would also include assigning the desired dot diameter and eliminate dot overlap. An important issue is certainly improving the operation of the algorithm and striving for complete automation.

Acknowledgements. The paper is based on the master's thesis entitled "Supporting the automation of thematic map editing using GIS tools" by the author of the article Natalia Dziuba. The thesis was written under the supervision of Stanisław Szombara defended 15.07.2024 at the Faculty of Geo-Data Science, Geodesy, and Environmental Engineering, AGH University of Krakow. This paper was made at AGH University of Krakow within research subvention no. 16.16.150.545.

References

Allen, J. (2022). Temporal transitions of demographic dot maps. *International Journal of Cartography*, *8*(2), 208–222. https://doi.org/10.1080/23729333.2021.1910184

Asche, H. (2023). *Editorial aspects* – In *The Atlas Cookbook: Ten ingredients how to edit an atlas* (pp. 20–39). ICA Commission on Atlases. https://atlas.icaci.org/wp-content/uploads/Atlas_Cookbook_2023_web.pdf

Bërxholi, A., Doka, D., & Asche, H. (Eds.). (2003). Demographic Atlas of Albania.

De Berg, M., Bose, P., Cheong, O., & Morin, P. (2004). On simplifying dot maps. *Computational Geometry*, *27*(1), 43–62. https://doi.org/10.1016/j.comgeo.2003.07.005

Distribution of rural population in 2016. (2018). In D. Rogalińska (Ed.), *Atlas statystyczny Polski. Statistical Atlas of Poland* (p. 25). Zakład Wydawnictw Statystycznych. https://stat.gov.pl/en/topics/other-studies/other-aggregated-studies/statistical-atlas-of-poland,17,1.html

Fleshman, W. (2019, February 1). *Spectral Clustering. Foundation and Application*. Towards Data Science. https://towardsdatascience.com/spectral-clustering-aba2640c0d5b

Hey, A. (2011). Automated dot mapping – How to generate dot clusters. *Proceedings of the 25th International Cartographic Conference, Paris, France, 3–8 July, CO-476.*

Hey, A., & Bill, R. (2014). Placing dots in dot maps. *International Journal of Geographical Information Science*, *28*(12), 2417–2434. https://doi.org/10.1080/13658816.2014.928822

Kimerling, A. J. (2013). Dotting the dot map, Revisited. *Cartography and Geographic Information Science 36(2)*, 165–182. https://doi.org/10.1559/152304009788188754

Rozmieszczenie ludności wiejskiej [Distribution of the rural population]. (2023). In E. Lodzińska & M. Wieczorek (Eds.), *Atlas Geograficzny: Liceum i Technikum* (p. 201). Demart.

Mackay, J. R. (1949). Dotting the dot map: An analysis of dot size, number, and visual tone density. *Surveying and Mapping*, *9*(1), 3–10.

Martínez Barbosa, C. A. (2023a). *Equal-Size Spectral Clustering*. GitHub repository. https://github.com/anamabo/Equal-Size-Spectral-Clustering

Martínez Barbosa, C. A. (2023b, February 6). *Equal-size spectral clustering. A modification of this popular algorithm that builds clusters balanced in the number of points*. Towards Data Science. https://towardsdatascience.com/equal-size-spectral-clustering-cce65c6f9ba3

Pasławski, J., Lodzińska, E., & Wieczorek, W. (1993–1997). Distribution of population, 1:1500000 (ark. 62.1). In *Atlas of the Republic of Poland*. Główny Geodeta Kraju.

Pasławski, J. (2015). *Jak opracować mapę kropkową*. Uniwersytet Warszawski.

Pieniążek, M., & Zych, M. (2020). *Statistical maps. Data visualisation methods* (Vol. 1). Statistics Poland.

Poland in numbers. (2024). *Polska w liczbach. Powiat pszczyński w liczbach*. https://www.polskawliczbach.pl/powiat_pszczynski

Ratajski, L. (1989). *Metodyka kartografii społeczno-gospodarczej*. PPWK.

Rozporządzenie Rady Ministrów z dnia 16 lipca 2021 r. w sprawie państwowego rejestru granic i powierzchni jednostek podziałów terytorialnych kraju Dz.U. 2021 poz. 1373 (2021a) (Polska). https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210001373

Rozporządzenie Ministra Rozwoju, Pracy i Technologii z dnia 27 lipca 2021 r. w sprawie bazy danych obiektów topograficznych oraz bazy danych obiektów ogólnogeograficznych, a także standardowych opracowań kartograficznych Dz.U. 2021 poz. 1412 (2021b) (Polska). https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210001412

Statistics Poland. https://stat.gov.pl/

Suchecka, J. (2014). *Statystyka przestrzenna. Metody analizy struktur przestrzennych*. Wydawnictwo C.H.Beck.

Szombara, S. (2017). Automation of dot maps production supported by BDOT10k database. *Proceedings of the Baltic Geodetic Congress (BGC Geomatics), Gdansk, Poland, 22–25 June, 100–104.* https://doi.org/10.1109/BGC.Geomatics.2017.34

Tennekes, M. (2018). Tmap: Thematic maps in R. *Journal of Statistical Software*, *84*(6), 1–39. https://doi.org/10.18637/jss.v084.i06

Yan, H., & Weibel, R. (2008). An algorithm for point cluster generalization based on the Voronoi diagram. *Computers & Geosciences*, *34*(8), 939–954. https://doi.org/10.1016/j.cageo.2007.07.008