ADVANCES IN GEODESY AND GEOINFORMATION
(formerly GEODESY AND CARTOGRAPHY)
Vol. 72, no. 1, article no. e35, 2023

https://doi.org/10.24425/agg.2022.141922

Original paper

OpenStreetMap – building data completeness visualization in terms of "Fitness for purpose"

Sylwia Borkowska*, Elzbieta Bielecka, Krzysztof Pokonieczny

Military University of Technology, Warsaw, Poland
e-mail: sylwia.borkowska@wat.edu.pl; ORCID: http://orcid.org/0000-0003-3183-1512
e-mail: elzbieta.bielecka@wat.edu.pl; ORCID: http://orcid.org/0000-0003-3255-1264
e-mail: krzysztof.pokonieczny@wat.edu.pl; ORCID: http://orcid.org/0000-0001-9114-5317

*Corresponding author: Sylwia Borkowska, e-mail: sylwia.borkowska@wat.edu.pl

Received: 2022-09-28 / Accepted: 2022-11-27

Abstract: The purpose of this article was to provide the user with information about the number of buildings in the analyzed OpenStreetMap (OSM) dataset in the form of data completeness indicators, namely the standard OSM building areal completeness index (C Index), the numerical completeness index (COUNT Index) and OSM building location accuracy index (TP Index). The official Polish vector database BDOT10k (Database of Topographic Objects) was designated as the reference dataset. Analyses were carried out for Piaseczno County in Poland, differentiated by land cover structure and urbanization level. The results were presented in the form of a bivariate choropleth map with an individually selected class interval suitable for the statistical distribution of the analyzed data. The results confirm that the completeness of OSM buildings close to 100% was obtained mainly in built-up areas. Areas with a commission of OSM buildings were distinguished in terms of area and number of buildings. Lower values of completeness rates were observed in less urbanized areas. The developed methodology for assessing the quality of OSM building data and visualizing the quality results to assist the user in selecting a dataset is universal and can be applied to any OSM polygon features, as well as for peer review of other spatial datasets of comparable thematic scope and detail.

Keywords: data completeness, data quality, bivariate choropleth map, OSM, VGI

1. Introduction

Buildings are, along with infrastructure, one of the most important physical elements of settlement, determining the morphological, functional and socio-economic structure. The location and number of buildings determine the type of land use, population structure,

energy needs, etc. Currently, there is a high demand for a free dataset describing building facilities, such as its structure, usage characteristics and dynamics. The datasets produced by national mapping agencies often lack the level of detail required to answer all questions in spatial science. Factors that can hinder data collection include scarce financial resources, administrative constraints and, in some cases, strict regulations on privacy and data use. Currently, the largest publicly available spatial database of buildings worldwide is OpenStreetMap. OpenStreetMap data in selected regions is usually evaluated by comparison with commercial or authority data before it is used in a project across application domains (Hecht et al., 2013). Definitions of fitness for purpose and data quality come in two forms: data quality defined by internal characteristics, such as completeness, positional and thematic accuracy, logical consistency, temporal quality (ISO, 2013) and data quality defined in terms of data use (Frank, 2009). According to Barron et al. (2014), the quality of OSM data largely depends on the purpose for which the data are implemented. They referred to this as the "Fitness for Purpose" assessment, previously defined by Veregin as "fitness for use" (Mocnik et al., 2017). Similarly, Chrisman (1984) states that 'Quality information provides the basis for assessing the suitability of spatial data for a given purpose.' Both statements show that data quality and fitness for purpose are commonly understood to be closely related. However, it is important to note that different interpretations of the same data can lead to different information. When the data is interpreted in the right context, the resulting information can be used to solve a given task, and the potential for solving that task depends on both the data and its interpretation. For example, route planning tasks can only be performed if the map is appropriate for the purpose and if the reader knows how to interpret the map in a given context according to their own criteria. An important issue is also appropriate data classification and visualization method. The cartogram method is one of the more commonly used methods of statistical cartography (Korycka-Skorupa and Paslawski, 2017; Korycka-Skorupa and Nowacki, 2019). Since the possibilities of correctly comparing simple cartograms are quite limited, a more synthetic way of presenting the dependencies of phenomena is to use a multi-variate choropleth (Slocum et al., 2009). The multi-variate choropleth is called a variation of the cartogram method, which is formed by superimposing two (bivariate) or more simple cartograms. The essence of bivariate choropleth method is the representation of the values of two phenomena within the boundaries of the spatial units (Leonowicz, 2002a; 2002b; 2006). The criterion for variables selection should be a certain association between them, about which the map informs viewers. Thus, the bivariate choropleth map shows the spatial variation of the structure of the studied phenomenon.

Data fitness for purpose refers to the specific aim and use of the data, but often the data is evaluated without a specific use – for example, to measure the evaluation of semantic compatibility, completeness and consistency, accuracy of location. Data quality does not directly measure, unlike fitness for purpose, how well the data is fit for a particular purpose, but whether it meets our expectations when used for different purposes. The data quality is independent of the specific purpose, as it is evaluated for all possible purposes.

This research aimed to provide the user not only with information on the number of buildings in the OSM data in relation to the reference database, but also to visualise

the quality of the analysed data in terms of its completeness and thus its usability for the fitness for purpose. Three proprietary completeness indicators were used for the analysis, i.e. the *C Index*, *TP Index*, *COUNT Index*, and the two-variable choropleth map presented the results of the quality assessment. Methods for quality assessment and visualisation of results based on cartographic modelling represent a comprehensive, original and innovative approach to assessing the usefulness of spatial data. The research contributes to both the scientific community and practitioners by providing comprehensive, mathematical-statistical based analysis of OSM buildings data completeness and its portrayal in the form of thematic maps.

Related work

OpenStreetMap is a widely used data source in various fields and services, such as environmental monitoring, disaster and emergency management, Sustainable Development Index (SDI) and mapping. OpenStreetMap data in selected regions is generally evaluated by comparing with commercial or authority data (Hayakawa et al., 2012; Demetriou, 2016). In study of Wang et al. (2013) three different quality elements: completeness, thematic accuracy and positional accuracy are presented. The article analyzes and evaluates the quality of OSM data with the 2011 version of the navigation map in Wuhan area of China. The result shows that OSM data on high-level roads and urban traffic are characterized by high positioning accuracy and completeness, so that these OSM data can be used to update the city road network database. The quality of OSM data can be additionally evaluated in the context of sustainable development and it can be a valuable source for monitoring some SDIs (Mobasheri et al., 2018; Borkowska and Pokonieczny, 2022). OSM is currently an important source for building data, despite the existence of potential quality issues. The study assessed the completeness of the OSM building using population data and examined the effectiveness of using population data to create reference data (Zhang et al., 2022). In contrast, Fan et al. (2014) assessed the quality of OSM building outline data for the German city of Munich, whose building outlines and attribute information are used in 3D building developments. In Zacharopoulou et al. (2021) the consistency and attribute completeness of OSM is evaluated and visualized multiscale at the regional, city, and feature levels in six European cities. A number of research projects (MacEachren et al., 1995; Leitner and Buttenfield, 2000; Cliburn et al., 2002; Deitrick, 2007) have shown that high-quality visualization supports decision-making and leads to much better conclusions. Therefore, for OSM data, where quality is often non heterogeneous, adequate visualization is considered essential, as it can strongly influence the viewer's cognitive processing (Keil et al., 2020). Moreover, it provides an exploratory tool that can help users evaluate the suitability of spatial data for a given purpose and interpret it according to established criteria (Mocnik et al., 2018), examine the dependence on external socioeconomic or demographic factors and to study the spatial distribution and heterogeneity of OSM data quality (Ribeiro and Fonte, 2015).

2. Study area and dataset

2.1. Study area

The research area covered the Piaseczno County located in Poland in the central part of the Masovian Voivodeship (Fig. 1).

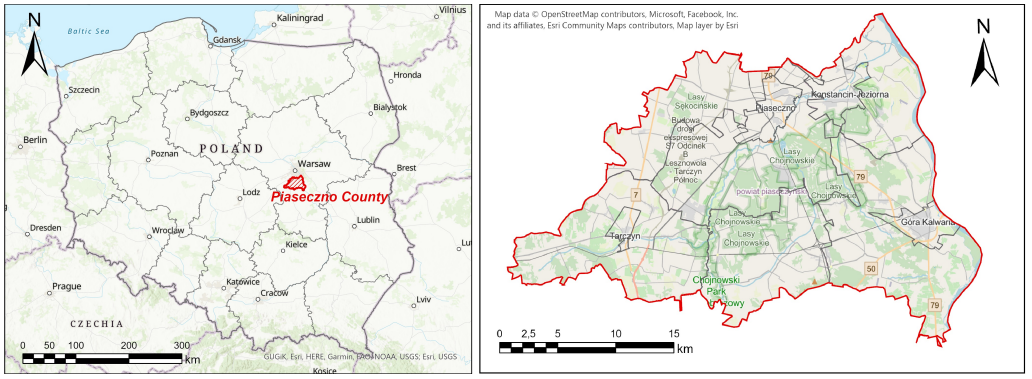


Fig. 1. Location of the analyzed Piaseczno County (source: OpenStreetMap.org, GUGIK, Esri)

Piaseczno County, with an area of 621.04 km^2, consists of six municipalities: four rural-urban municipalities and two rural municipalities. According to 2021 data, the district had a population of 190 606 people (GUS, 2021), which gives a population density of 307 people per km^2. The urbanization rate equals 47.8%. The County is the third richest County in Poland, for which the basic tax revenue per capita in 2021 was 652.70 PLN (PAP, 2022). The district is dominated by coniferous forests and mixed forests, which account for 19.6% of the area. The largest forest complex is the Chojnow Forests. Agricultural land accounts for 49.3% of the total area, and orchards are 10.1%.

2.2. OpenStreetMap dataset

The OpenStreetMap data used for the building completeness analysis was provided by the OSM GEOFABRIK service and its validity is 21 June 2021 (GEOFABRIK, 2021). OSM, building data is not standardized, and there is no clear definition and strict mapping rules (Nowak Da Costa, 2016). As a result, objects and modelling rules are not defined and only recommendations are available. A label, also called a tag, consists of a pair of expressions: “key = value” which can be equated with an attribute. Most features can be described using only a small number of tags. The building tag is used to mark a given object as a building. The most basic use is “building = yes”, but the value may be also used to classify the type of building. For example, a hospital building is labelled “building = hospital”. In the conducted research used all available objects tagged as a

building (building = ∗). Buildings are usually represented by an outline (polygon) – if possible, the outer edge of the building wall should be mapped. Building outlines are captured by different users by different techniques and data acquisition sources. Building data can be captured using handheld GPS devices, vectorization from aerial photographs or satellite imagery, sketch drawing measurements from the street level, or imported from government agency databases or other available spatial data (OSM, 2022). Hence their shape is simplified or very precise. OSM data is characterized by heterogeneous accuracy and level of detail, depending on the method of data collection, the level of generalization of the building outline and the skill and experience of the of the person editing OSM.

The OSM database also lacks unambiguous forms of quality control of entered building data. The author should follow the consensus standards of the OSM community, such as the code of conduct, good practices and general instructions. It is also possible to verify geometric and descriptive data by entering a new measurement by any OSM participant. In addition, a tool in the OSM editor is available for internal checking of the data only, checking its geometric and topological correctness. The OSM data are constantly updated, however, in a heterogeneous approach, depending on user activity and the degree of the area popularity.

2.3. BDOT10k – reference dataset

BDOT10k (Database of Topographic Objects) is a national, spatially continuous, vector database with the thematic scope and a level of detail corresponding to contemporary, civilian topographic maps at a scale of 1:10,000 maintained by the Head Office of Geodesy and Cartography in Poland. The timeliness of the BDOT10k collection used was March 2020. The detailed scope of information collected in BDOT10k, as well as the arrangement, procedure, and technical standards to create, update, verify and share this data is governed by the Regulation of the Minister of Development, Labour and Technology of 27 July 2021 on the topographic objects database and the database of general geographical objects, as well as standard cartographic presentations (RMDLT, 2021). A building in BDOT10k is defined as a construction object, permanently attached to the ground, having foundation, separated from space by building partitions (i.e., walls and covers). The geometrical building representation is the base outline or their maximum extent. All residential buildings and all non-residential isolated buildings are entered into the BDOT10k database. Generalization is not subject to refractions below 4 m on the walls of these buildings. Quality control of the data contained in BDOT10k is ensured by the National Topographic Objects Database Management System (KSZBDOT - Krajowy System Zarzadzania Baza Danych Obiektow Topograficznych) run by the Central Office of Geodesy and Cartography. BDOT10k data quality assessment refers to topology and geometry control, semantic control, syntactic control, attribute control, etc.

2.4. Data preprocessing

OSM buildings shapefile was imported into ESRI's ArcGIS software. All polygon objects that were assigned a value other than NULL for the "building" tag were considered buildings. The analyzed set of OSM building data covered 148,165 objects representing polygons, tagged according to the "building=*" rule. OSM buildings were compared with the objects belonging to the BDOT10k object type called 'buildings, building structures and facilities'. In particular, the following feature objects were mainly involved: buildings (BUBD), sports facilities (BUSP), high technical building structures (BUWT), other technical facilities (BUIT), and several objects from the OIOR class of small building structures of topographical or landmark importance. In order to make a spatial comparison of OSM building polygon with reference BDOT10k building's, it was necessary to harmonize the spatial reference using a common coordinate system. The projected Cartesian Gauss-Kruger coordinate system ETRS 1989 UWPP 1992, which usually serves as a spatial reference for topographic mapping in Poland, was chosen. Other than the preprocessing mentioned above, no other filters were applied to improve the quality of OSM data. In addition, mislabeled polygons could not be identified in the OSM datasets.

3. Methodology

3.1. Main methodological assumptions

The main research problem concerns the cartographic representation of the completeness of spatial data, as an element of data quality, enabling the user to assess the fitness for purpose of the analyzed data and, more specifically, to choose which of the two spatial datasets better suits needs. The methodological approach assumed to explain the research objective was cartographic modelling (as defined by C. Board), covering all stages of the research work from data acquisition, preprocessing and transformation, analysis and visualization (Baranowski et al., 2016). The purpose of cartographic modelling is to create a new cartographic visualization, which is the resultant of the analyses carried out on the spatial database. Thus, the subject of cartographic modelling and the essence of generalization of geographic information are not geometric operations performed on individual features representing topographic objects, but the highlighting of objects and phenomena that are important at a given observational scale, resulting from the understanding of geographic field (Glazewski et al., 2009; French and Li, 2010). It was hypothesized that a holistic approach based on mathematically defined indicators, their statistical analysis with the original cartographic presentation allows for the fitness-for-purpose OSM data assessment in relation to reference data.

The paper is structured as follows. The 1 km^2 hexagonal grid was set up as a minimal mapping unit, described broadly in Subsection 3.2. Three proprietary completeness indicators were used for the analysis, the TP Index, the C Index and COUNT Index, described the OSM data completeness, data quality element, presented in Subsection 3.3.

Finally, the two bivariable choropleth map was introduced, with class ranges based on data statistical distribution (Subection 3.4).

3.2. Hexagonal grid

Data quality index values visualization on choropleth maps makes it possible to identify areas with high or low data completeness levels. Hierarchical administrative units are commonly used as reference regions in thematic mapping, allowing results to be directly compared with official statistics. However, it should be noted that there are some drawbacks to using administrative units to assess the completeness of OSM data due to the Modifiable Areal Unit Problem (MAUP) (Openshaw, 2011). According to MAUP, the resulting aggregate values (e.g., ratios, proportions, densities) are affected by both the shape and scale of the aggregation unit. In addition, the boundaries of administrative units can also be subject to change which means limited comparability over time. Assessing the completeness of OSM data based on single administrative units may also not be detailed enough for small-scale surveys, for which smaller units are better suited.

Geometrically, the analyzed area can be divided into a regular grid of triangles, squares or hexagons. The use of a hexagonal grid provides the superiority over squares and triangles of being closer to the circle, while providing the same complete coverage of the study area (Roick et al., 2011) . In the applied analyses of the completeness of OSM buildings, a hexagonal grid was used, for which the values of completeness indicators were counted in each grid of 1 km^2. Due to the local scale of the study (Piaseczno County), the hexagonal grid in comparison with administrative divisions (e.g. communes) provided a detailed analysis of the spatial distribution of OSM data quality indicators in the studied area.

3.3. Completeness of OSM buildings

The developed indicators are relative, therefore they were calculated for the hexagonal grid in which at least one building from the BDOT10k reference data was located. A graphical interpretation of OSM's completeness indices in relation to BDOT10k is presented in the publication by Borkowska and Pokonieczny (2022). The completeness of OSM buildings was measured by the *C Index* (C – Completeness) which calculates the percentage ratio of OSM buildings to their area in a reference dataset (Tian et al., 2019). The *C Index* was calculated for each of the hexagonal grid with an area of 1 km^2 according to the Eq. (1):

$$C\ Index = \frac{\sum Ft\_Area_{OSM\_match}}{\sum Ft\_Area_{REF\_match}} \cdot 100\%, \tag{1}$$

where: $\sum Ft\_Area_{OSM\_match}$ – area of OSM building matching the building stored in reference BDOT10k database in a given cell of the hexagonal grid, $\sum Ft\_Area_{REF\_match}$ –

area of building from the BDOT10K reference database, corresponding to OSM building in a given hexagonal grid cell.

The *C Index* take values greater or equal 0, where 0 indicates no corresponding buildings in OSM data, 100% – that both datasets include the same buildings, and the value higher than 100 indicates OSM data commission.

Besides comparing aggregate values, the completeness of OSM buildings was analyzed by using object comparison. Hence, the *TP Index* (True Positive Index) was determined, indicating, in addition to completeness, the position of OSM building relative to the BDOT10K building. A graphical interpretation of the *TP Index* is shown in Figure 2.

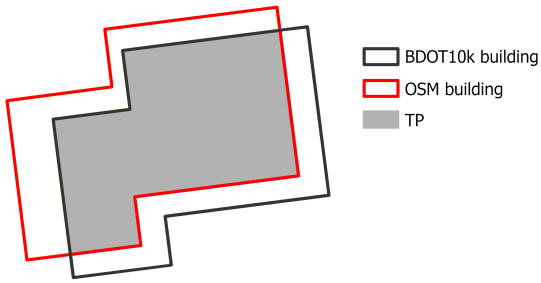


Fig. 2. Presentation of the TP Index

The *TP Index* determines as a percentage the common area of the OSM and BDOT10k buildings. This means that TP defines the degree of coverage of the area of buildings from the OSM relative to the BDOT10k. The *TP Index* was calculated using the Eq. (2):

$$TP\ Index = \frac{BLD_{OSM} \cap BLD_{REF}}{\sum Ft\_Area_{REF\_match}} \cdot 100\%, \tag{2}$$

where: BLD_{OSM} – building from OSM database, BLD_{REF} – building from reference database BDOT10k, $\sum Ft\_Area_{REF\_match}$ – area of building from the BDOT10k reference set in a given hexagonal grid cell.

The TP Index takes values from 0 to 100%. *TP Index* value of 100% is achieved by OSM buildings with exact coverage of BDOT10k dataset. The lower the value of the *TP Index* the less overlapping between OSM and BDOT10k buildings. For a *TP Index* equal to 0%, there is no coverage between OSM and BDOT10k dataset – buildings are disjoint.

The third indicator of OSM building completeness that was used in the study was numerical completeness. For this purpose, the number of OSM buildings located in each cell of the hexagonal grid was calculated and related to the number of buildings of the BDOT10k. The percentage ratio of the number of OSM buildings to the number of

BDOT10k buildings was presented in the form of *COUNT Index* according to Eq. (3):

$$COUNT\ Index = \frac{Ft\_Count_{OSM\_match}}{Ft\_Count_{REF\_match}} \cdot 100\%, \tag{3}$$

where: $Ft\_Count_{OSM\_match}$ – number of OSM buildings matching the reference BDOT10k database buildings in a given cell of the hexagonal grid, $Ft\_Count_{REF\_match}$ – number of buildings from the BDOT10K, corresponding to OSM in a given hexagonal grid cell.

The *COUNT Index* takes values greater than or equal to 0, where 0 means that there are no corresponding buildings in the OSM data, 100% means that both datasets contain the same number of buildings, and a value greater than 100% indicates a numerical commission of buildings in the OSM dataset over BDOT10k.

3.4. Bivariate choropleth classes

A bivariate choropleth map represents data attributes on a single thematic map. Bivariate maps can be useful for visually interpreting spatial patterns, especially for comparing the spatial distribution of two potentially related indicators, as well as for identifying outlier locations that do not follow the expected relationship between indicators (Kraak et al., 2020). Bivariate maps exhibit increased information complexity. Establishing the class ranges is an important step of bivariate choropleth map elaboration. The number of classes is determined by the graphic capabilities and perception of the reader, as well as the complexity of the data presented (Nelson, 2020). With these limitations, however, the map should convey as much information as possible. The number of classes is usually limited to a symmetrical size of nine (3×3) or sixteen (4×4) (Leonowicz, 2002a; 2002b; Calka, 2021).

In order to portrayal the completeness of OSM buildings and the accuracy of their location in relation to BDOT10k, a bivariate choropleth map was applied with the values of class ranges based on the statistical distribution for both variables. Hence, normal probability plots were determined for the studied C, *COUNT Index* and *TP Index*, along with the Shapiro-Wilk test – Figure 3. The normal probability plot identifies substantive departures from normality. In a normal probability plot (also called a "normal plot"), the sorted data is plotted against the values selected so that the resulting image approximates a straight line if the data has an approximately normal distribution. The Shapiro-Wilk test is used to assess whether the collected results of the studied phenomenon have a normal distribution (Hanusz et al., 2016). The null hypothesis for this test assumes that the research sample analyzed comes from a population with a normal distribution. If the Shapiro-Wilk test reaches statistical significance ($\alpha \leq p < 0.05$), this indicates a distribution that deviates from the Gaussian curve.

Deciding which statistical measure is appropriate for determining the bivariate choropleth class ranges was guided primarily by statistical distribution of the *C Index*, *COUNT Index* and *TP Index* values. Data with a normal distribution should be analyzed according to the mean value along with the standard deviation. The lack of a normal

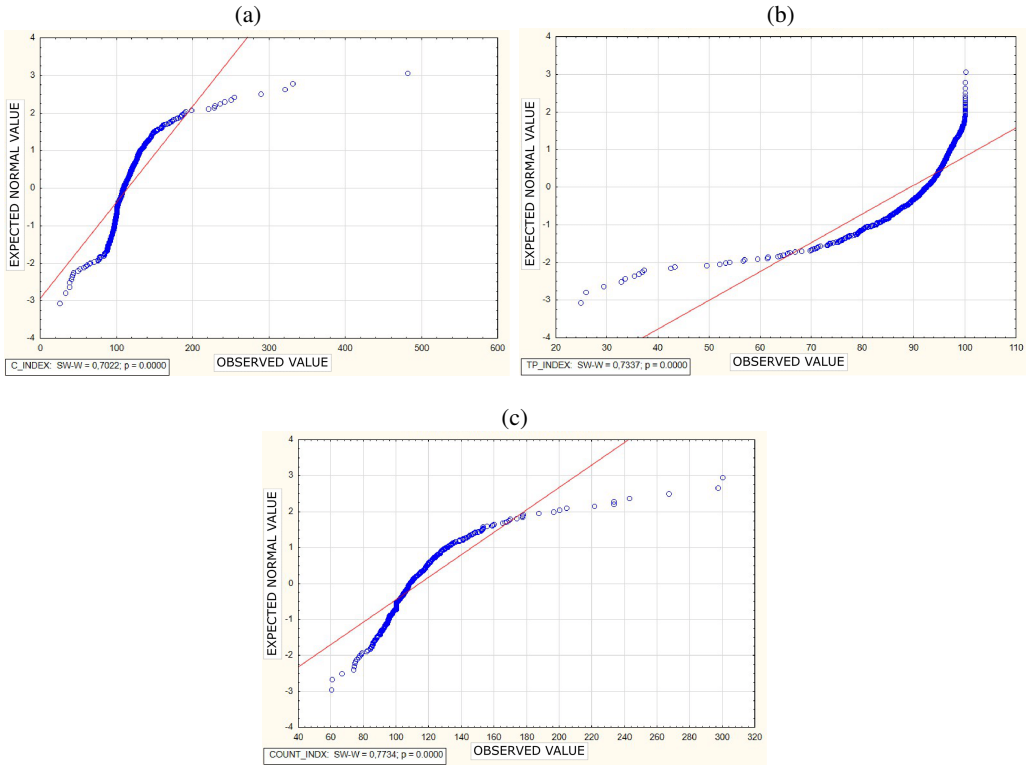


Fig. 3. Normal probability plots with the Shapiro-Wilk test: a) C Index, b) TP Index, c) COUNT Index

distribution implies that there are "outliers" which affect the value of the mean. For the *C Index* and *COUNT Index*, a three-level class range of the two-dimensional choropleth map was used. It was determined on the basis of the range of Indices values and the relationship between the median and the Median of the Absolute Deviation (MAD). MAD is a solid measure of the variability of a one-dimensional sample of quantitative data. The class ranges were determined in accordance with Table 1.

Table 1. Class ranges of bivariate choropleth according to statistical values of C and Count Indicators

Index	Class ranges of bivariate choropleth		
	1	2	3
C / COUNT	0 to M – MAD	from M – MAD to M + MAD	from M + MAD to maximum value

where: M – median value, MAD – value of median absolute deviation.

Furthermore, visualization of OSM buildings completeness based on *C Index* (areal completeness) and *TP Index* (location completeness) Indexes was performed. For the *C Index*, the range up to 100% and above 100% are shown separately. The *C Index*

values up to 100% are shown compared with the *TP Index* in the form of a bivariate choropleth, indicating the relationship between the areal completeness of OSM buildings and their location accuracy. For the *C Index* values above 100%, i.e. over completion (commission) of OSM buildings, a choropleth is used, as there is no reference to the location of BDOT10k data (objects appear only in the OSM database, no corresponding objects in the reference database). The data set for both visualizations adopts a non-normal distribution, hence, values related to the median and median absolute distribution were used to set-up the range of classes. The *C Index* below and above 100%, statistical values were calculated separately according to the distribution of non-normal data. The limits of the intermediate class range were determined for the *C Index* and *TP Index* according to Table 2:

Table 2. Class ranges of bivariate choropleth and choropleth maps according to statistical values of C and TP Indicators

Index	Class ranges of bivariate choropleth		
	1	2	3
C (under 100%) / TP	0 to M	from M to maximum value	–
Choropleth map			
C (above 100%)	0 to M – MAD	from M – MAD to M + MAD	from M + MAD to maximum value

where: M – median value, MAD – value of median absolute deviation.

4. Results

According to the research, two bivariate choropleth maps were developed for visualization the completeness of OSM spatial data in Piaseczno County using three proprietary completeness indices: the *TP Index*, the *C Index* and the *COUNT Index*. The class ranges were based on the statistical distribution of the index data as shown in Table 3:

Table 3. Statistical values of the completeness indicators: C Index, TP Index, COUNT Index

Index	Value of the Index in the Piaseczno County (%)			
	median	median absolute deviation	minimum	maximum
C	109	11	26	481
C ≤ 100%	97	3	26	100
C > 100%	117.5	10	101	481
COUNT	102	9	28	300
TP	92	4	25	100

The bivariate choropleth map shown in Figure 4 is meant to illustrate the relationship between the rate of areal (*C Index*) and numerical completeness (*COUNT Index*) and their spatial distribution. It provides a general purpose and easy-to-understand overview of the completeness of OSM buildings in compered to BDOT10k data in Piaseczno County. The division of classes of the developed bivariate choropleth map (3 × 3) for the analyzed indicators was developed in accordance with their statistics. The data used is highly commission (over 100%) and abnormally distributed, which determined the ranges used.

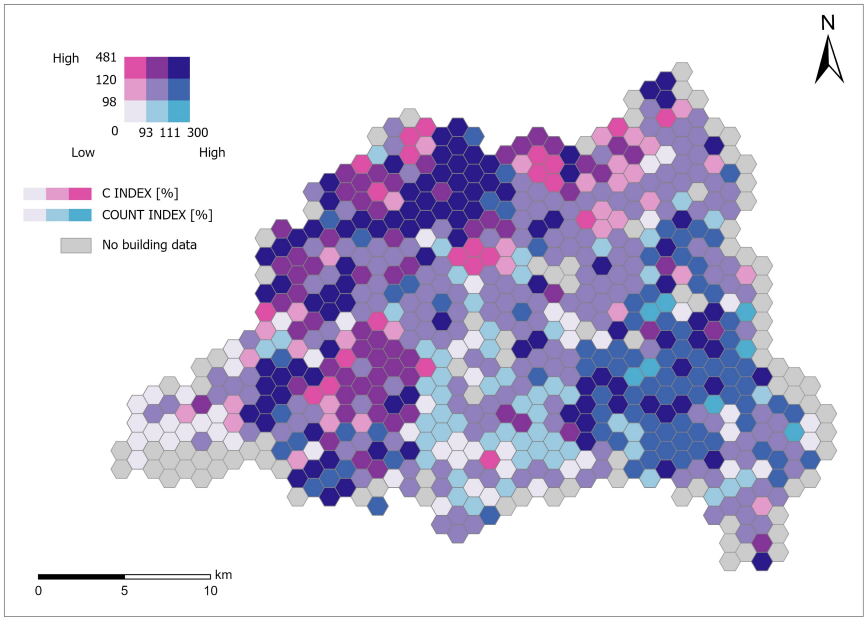


Fig. 4. Bivariate choropleth map of Piaseczno County visualizations C Index and COUNT Index

The first class, for *C Index* ≤ 98% and *COUNT Index* ≤ 93% (shade of white) values, illustrating the deficiency of OSM objects in relation to the BDOT10k database (incompleteness) accounts for 8.8% of the study area. There are regions with a low degree of urbanization, located distant from larger cities. The second class, for *C Index* ≤ 120% and *COUNT Index* ≤ 111% (shade of gray purple), shows the areas with the highest numerical and areal correspondence of OSM buildings to the BDOT10k database (OSM completeness nearest 100%). These regions account for 29.8% of the study area. These are mostly more urbanized zones located near major roads. The third class, with a range of *C Index* > 120% and *COUNT Index* > 111% (shade of dark blue), represents areas with a parallel numerical and areal predominance of OSM buildings in relation to the BDOT10K database (OSM data hypercommission), accounting for 16.6% of the study area. These areas are mainly highly built-up, being parts of larger cities with a developed transportation network. The classes presented in magenta shades refer to region for

which the areal completeness ($C\ Index > 98\%$) of OSM buildings is higher than their numerical completeness ($COUNT\ Index \leq 93\%$) in relation to the BDOT10k base. This means that for these areas the superiority is the areal share rather than the number of OSM buildings. These regions constitute 9.5% of the examined county and concern predominantly urbanized areas. On the other hand, the classes presented in shades of blue concern the region for which the numerical completeness ($COUNT\ Index > 93\%$) of OSM buildings is higher than their areal completeness ($C\ Index \leq 98\%$) compared to the BDOT10k base. This means that for these areas the advantage is taken by the number of OSM buildings, not their area. These regions constitute 10.3% of the Piaseczno County and concern mainly suburban areas.

The bivariate choropleth map combine with choropleth shown in Figure 5 is to present the relationship between the areal completeness ($C\ Index$) and positional completeness ($TP\ Index$) and their spatial distribution. Due to the data scope used, index values up to 100% that were possible for comparison were presented using a bivariate choropleth. The remaining values relating to areal commission (values over 100%), for which there is no $TP\ Index$ reference (only those buildings from the OSM database that have a matching reference database can be compared positionally) are presented separately in the choropleth.

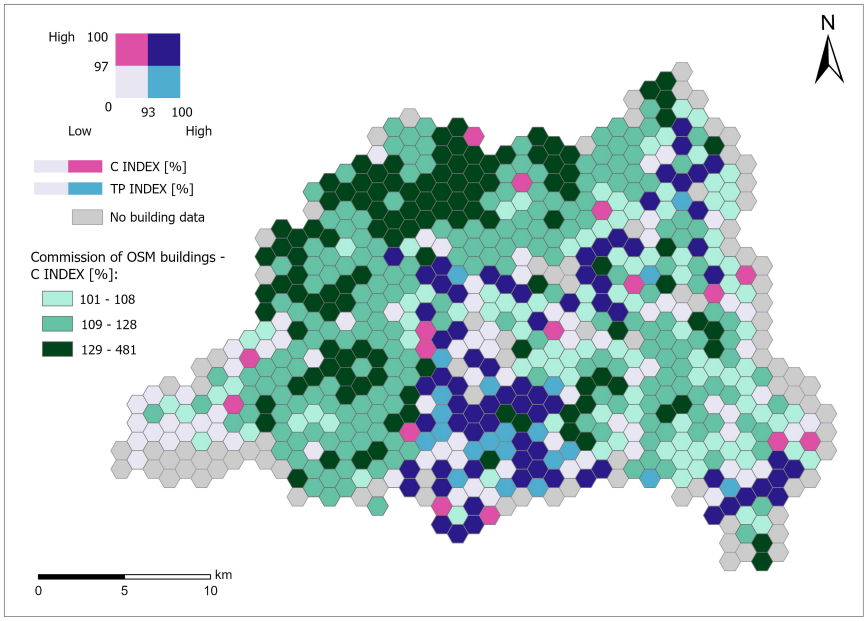


Fig. 5. Bivariate choropleth map of Piaseczno County visualizations C Index and TP Index

To visualize the values of $TP\ Index$ and $C\ Index$ not exceeding 100%, a bivariate choropleth map was used with class division (2×2) consistent with the statistical distribution. The first class, for $C\ Index \leq 97\%$ and $TP\ Index \leq 93\%$ (shade of white)

values, representing both a deficit and lower accuracy of OSM objects in reference to the BDOT10k database, accounts for 8.4% of the study area. These are mostly suburbs and areas with a low level of urbanization. In class two, for index values $C\ Index > 97\%$ and $TP\ Index > 93\%$ (dark blue shade), areas with the highest areal and positional correspondence of OSM buildings with the BDOT10k database are pointed out. These regions account for 13.1% of Piaseczno County and mainly concern areas with a higher degree of urbanization. For areas with magenta saturation ($C\ Index > 97\%$, $TP\ Index \leq 93\%$), the advantage there is in areas with higher areal completeness of OSM buildings than positional accuracy compared to the BDOT10k database. In contrast, for areas with blue saturation ($C\ Index \leq 97\%$, $TP\ Index > 93\%$), there are more OSM buildings with higher positional accuracy than areal completeness. Outlier areas account for 9.7% of the analyzed county.

Three classes of choropleth map were used for areal commission. The smallest areal commission values for the $C\ Index \leq 108\%$ were observed for 16.8% of the study area. The highest values ($C\ Index > 129\%$) were mostly observed in areas with a high degree of urbanization and a well-developed transportation network, accounting for 18.9% of Piaseczno County. The remaining areas with $C\ Index$ commission values between 109% and 128% account for 33.1% of the Piaseczno County.

5. Discussion

Although various indicators and measures have been used so far, OSM quality assessment is still an open research topic. Therefore, visualization of OSM quality is equally important because it acts as an awareness tool for the novice user and an exploration tool for the expert (Zacharopoulou et al., 2021). Completeness of OSM data is an important element of data quality assessment. The present study is concerned with the cartographic representation of the completeness of OSM data, as an element of data quality that allows the user to assess the usefulness of the analyzed data by choosing which of the two spatial datasets better suits his needs.

The proposed methodology deals with the completeness of OSM buildings in a systematic way by comparing OSM features with their counterparts from the official BDOT10k dataset and visualizing the obtained results in the form of bivariate choropleth maps in hexagonal grid of cell size 1 km^2. The results obtained are consistent with other similar studies. Regarding the completeness of OSM buildings in the surveyed district, it was found that some areas are well mapped, especially those with a high degree of development – mainly the completeness of building features is relatively high in city centers, while its value drops sharply further away from city centres. However, In the case of the relationship between the studied indicators of completeness, their spatial distribution is quite diverse as shown in the developed maps (Figs. 4, 5). Some spatial patterns can be observed in relation to the studied completeness indicators.

In the case of the bivariate choropleth map showing the relationship between numerical and areal completeness ($C\ Index$ and $COUNT\ Index$), the areas with the highest

values – OSM buildings significantly exceed those of BDOT10k in terms of both area ratio and their number – are grouped mainly in built-up areas, the outskirts of cities and a developed transportation network. In the case of areas for which the number as well as the area of OSM buildings most closely corresponds to the BDOT10k database, a clear grouping is also evident. These are mainly areas that are city centers and smaller near-by towns. In addition, a clear advantage of post-area commission of OSM buildings against BDOT10k can be seen in the western part of the analyzed district where forest and recreational areas dominate. On the other hand, areas where the advantage is commission of numbers can be seen in the western part of the analyzed district. Areas with a low degree of numerical completeness of buildings but surface completeness similar to BDOT10k are clearly grouped in the southern part of Piaseczno County, where agricultural areas dominate.

Spatial patterns are also evident in the map showing the relationship between the areal completeness and accuracy of OSM buildings. The areal commissions of OSM buildings clearly clusters in areas with a high degree of urbanization reaching the highest values in the northern part of the analyzed district, on the outskirts of the city of Piaseczno and neighboring cities. In visualizing the relationship between the areal completeness and the accuracy of the location of OSM buildings, linear clustering is evident, occurring mainly in areas where the main roads of the analyzed area run.

In addition to visible clustering, there are also outlier cases. The buildings of the OSM and BDOT10k databases were compared with the actual terrain situation as seen on the orthophotos updated to 2020 from the Geoportal service. Examples of buildings identified in OSM and BDOT10k databases against orthophotos in hexagonal grids, along with values of completeness indicators, are shown in Figure 6.

The obtained differences between the completeness of OSM buildings in comparison with the BDOT10k reference database are certainly due to several reasons. In view of the timeliness of the reference data, a common case encountered in the analysis was the presence of buildings in OSM, which is also visible on the orthophotomaps and lacks of its vector in the BDOT10k database – Figure 6a. As a result, there was OSM data commission, eventually reaching a maximum value of 481% for *C Index*. On the other hand, there were also cases in which it was the buildings visible on the orthophotos that had their vectors in the reference database and lacked of their matches in the OSM database, leading to a shortage of objects – Figure 6b. Visible differences between the studied buildings also relate to the displacement of outlines between the databases – Figure 6c. In addition, a significant error in vectorization is the incorrect identification of a building in the OSM database and the inputs of the entire built-up area instead of its outline – Figure 6d. As a result, this leads to a disproportionately high areal commission (*C Index*) with a numerical deficiency of buildings (*COUNT Index*). Other OSM vectorization errors include the identification of terraced buildings. Figure 6e shows the absence of separate outlines of terraced housing in the OSM database, which leads to differences in numerical completeness. On the other hand, Figure 6f shows the opposite case of the lack of separate outlines of terraced buildings in the BDOT10k reference database.

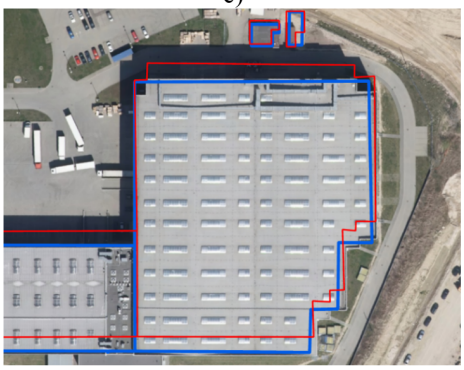
a)

C Index = 135%
TP Index = 89%
COUNT Index = 113%

b)

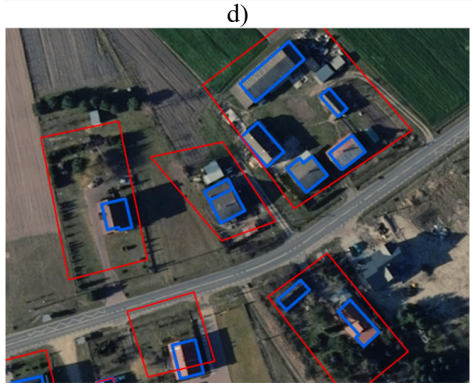
C Index = 75%
TP Index = 81%
COUNT Index = 35%

c)



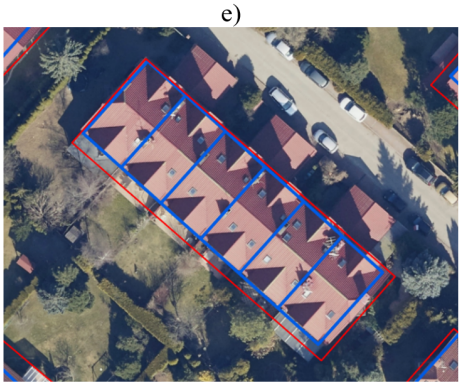
C Index = 101%
TP Index = 70%
COUNT Index = 111%

d)



C Index = 169%
TP Index = 90%
COUNT Index = 85%

e)



C Index = 125%
TP Index = 93%
COUNT Index = 86%

f)

C Index = 110%
TP Index = 96%
COUNT Index = 109%

Fig. 6. Comparison of the analyzed buildings location from the OSM database (red) and BDOT10k (blue) on the orthophotomap, along with the values of the calculated completeness indices: C Index, TP Index, COUNT Index

6. Conclusion

The presented work was aimed at understanding the applicability of OpenStreetMap building data and assessing its quality in the context of official spatial databases such as the Polish Database of Topographic Objects in the study area of Piaseczno County.

The use of completeness indices in the form of $C\ Index$, $TP\ Index$, $COUNT\ Index$ and their visualization in a 1 km^2 hexagonal grid allows a detailed analysis of the quality, structure and spatial patterns of OSM data. The use of bivariate choropleth maps makes it possible to visualize the relationship between the calculated completeness indicators of OSM buildings in comparison with a reference database. Additionally, the resulting two-dimensional map allows these variables to be displayed simultaneously using a single colour scheme. The resulting of bivariate choropleth maps not only provide the user with information on the number of OSM buildings, but also allows to assess the quality of OSM data and provides a basis for evaluating the suitability of spatial data for a given purpose. Thus, the research hypothesis was confirmed.

The results obtained confirm that OSM completeness closest to 100% was obtained mainly in built-up areas. In addition, areas with commission of OSM buildings were distinguished in terms of their area and number of buildings. In less urbanized areas, less "popular" among OSM users, there are gaps in the OSM database and thus lower values of completeness indicators. The elaborated methodology for OSM data quality assessment and visualization the quality results to assist the user in dataset selection is universal and can be applied to any OSM spatial objects, as well as to the peer review (mutual evaluation) of other spatial datasets of comparable thematic scope and detail.

OpenStreetMap constitutes a huge collection of crowdsourced geographic data. It is a widely used data source in various fields and services, such as environmental monitoring, disaster and emergency management, SDI, and mapping. As with any dataset, quality and user needs determine suitability for use. Information regarding not only the quality of the data itself but also the analysis of that data is important from the point of view of the user and its usability. OSM buildings are an important spatial database with a wide range of uses in spatial analysis, emergency management and mapping. Providing the user with information on the number of buildings in the OSM and reference dataset, their quality and structure enables the selection of the most appropriate data according to their purpose. Considering the indicators of OSM data completeness presented in the article, in future research the authors plan to expand the analysis of OSM data quality with the original synthetic data quality indicators.

Author contributions

Conceptualization: S.B., E.B. and K.P; methodology: S.B., E.B.; software development: S.B.; data collection and analyses: S.B.; writing and editing: S.B.; critical revision: E.B. and K.P.; review and editing: S.B; visualization: S.B.

Data availability statement

The data used in the study are available from the corresponding author upon reasonable request.

Acknowledgements

The manuscript does not have external founds.

References

Baranowski, M., Gotlib, D., and Olszewski, R. (2016). Properties of cartographic modelling under contemporary definitions of a map. *Polish Cartographical Review*, 48(3), 91–100. DOI: 10.1515/pcr-2016-0011.

Barron, C., Neis, P. and Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6), 877–895. DOI: 10.1111/tgis.12073.

Borkowska, S. and Pokonieczny, K. (2022) Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability*, 14, 3728. DOI: 10.3390/su14073728.

Calka, B. (2021). Bivariate choropleth map documenting land cover intensity and population growth in Poland 2006–2018. *J. Maps*, 17, 162–168. DOI: 10.1080/17445647.2021.2009925.

Chrisman, N.R. (1984). The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21(2), 79–87.

Cliburn, D.C., Feddema, J.J., Miller, J.R. et al. (2002). Design and evaluation of a decision support system in a water balance application. *Comput. Graph.*, 26, 931–949. DOI: 10.1016/S0097-8493(02)00181-4.

Deitrick, S.A. (2007). Uncertainty visualization and decision making: Does visualizing uncertain information change decisions? In Proceedings of the 23rd International Cartographic Conference, 4-10 August 2007, 4–10. Moscow, Russia.

Demetriou, D. (2016). Uncertainty of OpenStreetMap data for the road network in Cyprus. *Proc. SPIE*, 9688. DOI: 10.1117/12.2239612.

Fan, H., Zipf, A., Fu, Q. et al. (2014). Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.*, 28(4), 700–719. DOI: 10.1080/13658816.2013.867495.

Frank, U.A. (2009). Why is scale an effective descriptor for data quality? The physical and ontological rationale for imprecision and level of detail. In Gerhard Navratil (Ed.) Research trends in geographic information science, pp. 39–61. Springer: Heidelberg.

French, K., and Li, X. (2010). Feature-based cartographic modelling. *Int. J. Geogr. Inf. Sci.*, 24(1), 141–164. DOI: 10.1080/13658810802492462.

GEOFABRIK (2021). Retrieved August 28, 2022 from http://download.geofabrik.de/europe/poland.html.

Glazewski, A., Kowalski, P.J., Olszewski, R. et al. (2009). *New approach to multi scale cartographic modelling of reference and thematic databases in Poland*. Cartography in Central and Eastern Europe, 89–106. Springer: Berlin, Heidelberg.

GUS (2021). Area, population and ranking positions by powiats and cities with powiat status. Retrieved from https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/powierzchnia-i-ludnosc-w-przekroju- terytorialnym-w-2021-roku,7,18.html

Hanusz, Z. Tarasinska, J. and Zielinski, W. (2016). Shapiro–Wilk Test with Known Mean. *Revstat Stat. J.*, 14, 89–100. DOI: 10.57805/revstat.v14i1.180.

Hayakawa, T., Imi, Y. and Ito, T. (2012). Analysis of Quality of Data in OpenStreetMap. 2012 IEEE 14th International Conference on Commerce and Enterprise Computing, 131–134.

Hecht, R., Kunze, C. nd Hahmann, S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS Int. J. Geo-Inf.*, 2, 1066–1091. DOI: 10.3390/ijgi2041066.

ISO (2013). Geographic information-Data Quality. ISO/TC 211 Geographic Information/Geomatics, International Organization for Standardization, Geneva, Switzerland, 2013.

Keil, J., Edler, D., Kuchinke, L. et al.. (2020). Effects of visual map complexity on the attentional processing of landmarks. *PLoS ONE*, 15, e0229575. DOI: 10.1371/journal.pone.0229575.

Korycka-Skorupa, J. and Paslawski, J. (2017). The beginnings of the choropleth presentation. *Polish Cartographical Review*, 49(4), 187–198. DOI: 10.1515/pcr-2017-0012.

Korycka-Skorupa, J. and Nowacki, T. (2019). Cartographic presentation – from simple to complex map. *Miscellanea Geographica*. 23(1), 16–22. DOI: 10.2478/mgrsd-2018-0023.

Kraak, M.J., Roth, R.E., Ricker, B. et al. (2020). *Mapping for a Sustainable World*. The United Nations: New York.

Leitner, M. and Buttenfield, B.P. (2000). Guidelines for the display of attribute certainty. *Cartogr. Geogr. Inf. Sci.*. 27, 3–14.

Leonowicz, A. (2002). Prezentacja zależności zjawisk metodą kartogramu złożonego. *Polski Przegląd Kartograficzny*, 34, 273–85.

Leonowicz, A. (2002). Z problematyki porównywalności kartogramów. *Polski Przegląd Kartograficzny*, 34(1), 22–33.

Leonowicz, A. (2006). Two-variable choropleth maps as a useful tool for visualization of geographical relationship. *Geografija*, 42(1), 33–37.

MacEachren, A.M., Brewer, C. and Pickle, L.W. (1995). Mapping health statistics: Representing data reliability. In Proceedings of the 17th International Cartographic Conference, Barcelona, Spain, 3-9 September 1995, 311-319.

Mobasheri, A., Zipf, A. and Francis, L. (2018). OpenStreetMap data quality enrichment through awareness raising and collective action tools - experiences from a European project. *Geo. Spat. Inf. Sci.*, 21:3, 234–246. DOI: 10.1080/10095020.2018.1493817.

Mocnik, F.B., Fan, H. and Zipf, A. (2017). *Data Quality and Fitness for Purpose. Conference: 20th AGILE Conference on Geographic Information Science*. Wageningen: Netherlands. DOI: 10.13140/RG.2.2.13387.18726.

Mocnik, F.B., Mobasheri, A., Griesbaum, L. et al. (2018). A grounding-based ontology of data quality measures. *J. Spat. Inf. Sci.*, 16(16), 1–25. DOI: 10.5311/JOSIS.2018.16.360.

Nelson, J. (2020). *Multivariate Mapping*. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2020 Edition). DOI: 10.22224/gistbok/2020.1.5.

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

Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Geo Books: Norwick, UK.

OSM (2022). Retrieved August 25, 2022 from: https://wiki.openstreetmap.org/wiki/Pl:Key:building.

PAP (2022). *Najbogatsze i najbiedniejsze powiaty w Polsce część pierwsza (1–99)*. Serwis Samorządowy PAP.

Ribeiro, A. and Fonte, C.C. (2015). A Methodology for Assessing OpenStreetMap Degree of Coverage for Purposes of Land Cover Mapping. ISPRS Annals of Photogrammetry. *Remote Sensing and Spatial Information Sciences*, II3, 297–303. DOI: 10.5194/isprsannals-II-3-W5-297-2015.

RMDLT. (2021). Regulation of the Minister of Development, Labour and Technology of July 27, 2021 on the database of topographic objects and the database of general geographic objects, as well as standard cartographic studies, Dz.U. 2021, nr 30, poz. 1412.

Roick, O., Hagenauer, J. and Zipf, A. (2011). OSMatrix - Grid based analysis and visualization of OpenStreetMap. In Proceedings of the 1st European State of the Map Conference(SOTM-EU), Vienna, Austria.

Slocum, T., McMaster, R.B., Kessler, F.C. et al. (2005). *Thematic cartography and geovisulization, second edition*. Upper Saddle River: Pearson Prentice Hall. ISBN: 9780132298346.

Tian, Y., Zhou, Q. and Fu, X. (2019). An Analysis of the Evolution, Completeness and Spatial Patterns of OpenStreetMap Building Data in China. *ISPRS Int. J. Geo-Inf.*, 8, 35. DOI: 10.3390/ijgi8010035.

Wang, M., Li, Q., Hu, Q. et al. (2013). Quality Analysis of Open Street Map Data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. – ISPRS Arch.*, XL2, 155–158. DOI: 10.5194/isprsarchives-XL-2-W1-155-2013.

Zacharopoulou, D., Skopeliti, A. and Nakos, B. (2021). Assessment and Visualization of OSM Consistency for European Cities. *ISPRS Int. J. Geoinf.*, 10, 361. DOI: 10.3390/ijgi10060361.

Zhang, Y., Zhou, Q., Brovelli, M.A. et al.. (2022). Assessing OSM building completeness using population data. *Int. J. Geogr. Inf. Sci.*. 36(7), 1443–1466. DOI: 10.1080/13658816.2021.2023158.