PREUČEVANJE POPOLNOSTI PODATKOV NA PODLAGI PRIMERJALNE ANALIZE MED PODATKI VGI IN URADNIMI PODATKOVNIMI NIZI O STAVBAH

NOVEL TOOL FOR EXAMINATION OF DATA COMPLETENESS BASED ON A COMPARATIVE STUDY OF VGI DATA AND OFFICIAL BUILDING DATASETS

Joanna Nowak Da Costa

UDK: 551.506:725.1
Klasifikacija prispevka po COBISS.SI: 1.01
Prispelo: 2. 3. 2016
Sprejeto: 9. 9. 2016

DOI: 10.15292/geodetski-vestnik.2016.03.495-508
SCIENTIFIC ARTICLE
Received: 2. 3. 2016
Accepted: 9. 9. 2016

IZVLEČEK

Namen študije je bil prispevati k boljšemu razumevanju kakovosti prostovoljno zbranih geografskih informacij VGI (angl. volunteered geographic information) in njihovih koristi. Pri raziskavi smo se osredotočili na možnost uporabe podatkov o stavbah OpenStreetMap za uradne prostorske podatkovne nize. Z vidika pojavne popolnosti podatkov so ugotovitve raziskave primerljive z rezultati podobnih izvedenih študij. Ugotovili smo, da je popolnost podatkov o stavbah z vidika deleža zajetih stavb relativno visoka v središčih mest, z oddaljenostjo od urbanih središč pa se manjša. Prav tako se je izkazalo, da je popolnost opisnih podatkov o stavbah odvisna od stopnje urbanizacije, dodatno pa še od vrste opisnega podatka. Srednja položajna točnost podatkov o stavbah zbirke OpenStreetMap je za urbana območja ocenjena z 0,6 metra, za podeželje pa z 1,7 metra. Ta ocena je več kot petkrat boljša kot pogosto navedena ocena kakovosti podatkov OpenStreetMap, ki jo je objavil Haklay v letu 2010. V prispevku predstavljamo nov pristop v podporo oceni popolnosti podatkov OpenStreetMap, ki se nanašajo na stavbe. Predlagani kazalnik, ki smo ga poimenovali popolnost ujemanja objekta na podlagi površine (angl. matching feature area-based completeness), omogoča oceno popolnosti podatkov za kakršenkoli ploskovni prostorski podatkovni niz. Kazalnik je tudi prilagodljiv, saj ni pogojen z modeliranjem oboda ploskovnega objekta, niti ne s stopnjo posploševanja. Dodatno je predlagana preprosta metoda za posodabljanje uradnih evidenc o stavbah na podlagi množice podatkov OpenStreetMap.

KLJUČNE BESEDE

kakovost, prostorski podatki, popolnost podatkov, OpenStreetMap, prostovoljno zbrane geografske informacije

ABSTRACT

The goal of this study was a better understanding of the quality of Volunteered Geographic Information (VGI), and by extension its utility. The research focused on the applicability of OpenStreetMap (OSM) building data for official spatial databases. In terms of feature completeness, the achieved results are in-line with other similar studies. The study concluded that in town centres the completeness of OSM data is relatively high but decreases further away from urban centres. It demonstrated that attribute completeness also relies on the level of urbanization as well as the nature of attribute. Furthermore, a very high overall positional accuracy was determined for OSM building data that ranged between 0.6 m in urban areas and 1.7 m in rural areas. This result is more than five times better than the frequently cited OSM accuracy results obtained by Haklay in 2010. In this work, a novel tool is introduced to help assess the completeness of OSM building-tagged features. The proposed index, called the matching feature area-based completeness, estimates the completeness of any areal feature set. This index is also flexible because it is neither affected by discrepancies in the feature outline modelling nor by the degree of abstraction. In addition, the author proposed a simple method to update the official register using the large volume of OSM building data "over-completeness" together with the building data excess indicator.

KEY WORDS

quality, spatial data, data completeness, OpenStreetMap, volunteered geographic information

1 INTRODUCTION

The development and spreading of information and communication technologies along with the growing ability of the public to use them, remained not without impact on geospatial mapping. It appears that virtually everyone, regardless of their education, knowledge and experience, can collect spatial information, for example while walking or cycling with a GPS equipped mobile phone, and produce social network maps. This trend was defined as "neogeography" (Turner, 2006), "crowd sourcing" (in the Web 2.0 setting), or Volunteered Geographic Information (VGI). The latter term is used particularly in relation to spatial data collected voluntarily and free of charge by a large number of volunteers (Goodchild, 2007). The OpenSteetMap (OSM) initiative, the most extensive VGI representative in terms of the number of involved users and the volume of data created, has already gained academic research and commercial interests. However, since the geospatial contributions, skill level and motivations of OSM communities change over time, therefore, monitoring and updated data quality research are necessary to understand the applicability of this important dataset. The data quality is relative to the users' needs and it is neither independent nor absolute (Cooper et al., 2012; Bielecka et al., 2014).

The aim of the study was to understand the applicability of OpenStreetMap building data and to assess its quality specifically with regards to its potential use as complementary or input data for official spatial databases. It focuses in particular on the Polish Database of Topographic Objects for buildings in the Polish county of Siedlce. The study introduces a novel tool that helps to assess the completeness of OSM building-tagged features. The proposed index estimates the completeness of any areal feature set, and it is neither affected by discrepancies in the feature outline modelling nor by the degree of abstraction. Furthermore, the study proposes a simple method to achieve improved updating of official data based on the volume of building data missing from official database, that is OSM "over-completeness".

First, the paper reviews related research in Section 2. Next, it presents the method chosen for the OSM building data quality analysis in Section 3. In Section 4, the paper introduces study area and datasets characteristics. In Sections 5 through 7 the study focuses on thematic and positional accuracy as well as feature completeness. The paper addresses its concluding remarks in Section 8.

2 RELATED WORK

The OpenStreetMap (OSM) project, whose mission is to create a free, digital, open and editable map of the world, and provide a ready-made map or geographic dataset to anyone who wants it, bases on contributions from volunteers. The user-generated geographic information involves many forms of contribution such as online mapping or georeferencing of existing data sources like aerial image, as well as, the collection of data through the user's location-enabled smartphone. The OSM's approach to creating and managing map and geographic dataset was rather intuitive than calculated (Coote and Rackham, 2008) which caused concern among GI experts regarding the quality and usability of such data. As a result, the OSM, and, in general VGI, data credibility and quality are being increasingly studied by researchers (Elwood et al., 2012; Flanagin and Metzger, 2008).

Road network is the most frequently analysed OSM data. In most cases, OSM roads data was compared to official datasets. However, the choice of a reference data for quality control of data collected by non-professional land surveyors is problematic because of its heterogeneity, as noted also by Goodchild and

Li (2012), Haklay (2010), Goodchild and Glennon (2010), and Dorn et al. (2015). Studies of OSM roads completeness concluded that it is heterogeneous and much higher in big cities, lower in towns, and the lowest in rural areas (Haklay, 2010; Girres and Touya, 2010; Esmaili et al., 2013; Zielstra et al., 2013). In terms of positional accuracy of OSM road data, they concluded that some areas are well mapped, however with a tight relation of completeness and urbanization. According to the first ever systematic study, and one of the most cited study, conducted by Haklay in 2010, OpenStreetMap data was, on average, within about 3.2 to 4.8 metres of the position recorded by Ordnance Survey in the centre of London. However, the average in the peripheral districts dropped to 6.8–8.3 meters and the maximum deviations reached 20 meters.

Despite its name, OpenStreetMap is not just a road map; it provides topographic data including buildings. Recently, OSM building data quality has been tested using German and Austrian official data. OSM building completeness was found to be higher in urban areas in comparison with rural ones, but still low (Hecht et al., 2013; Klonner et al., 2014). According to another research on quality assessment of OSM building footprints data in Germany, data was characterised to have a high completeness in terms of area covered, but with limited attributive information, such as building types. While with respect to shape, OSM building footprints have high similarity to those in the German administrative dataset. And there is an offset of about four meters in average in terms of position accuracy (Fan's et al., 2014).

To sum up, many researchers agree that the main advantage of VGI data quality is its good geometric accuracy, while its geographic coverage patchwork and inconsistent semantics are its drawbacks (Goodchild, 2007; Ballatore et al., 2013; Mooney and Corcoran, 2012).

3 METHODOLOGY

The OSM quality analysis focus was on three out of six data quality elements outlined in the current spatial data quality standard, ISO 19157 (2013), namely: completeness, positional and thematic accuracy. The OSM data was compared with the third-party dataset (extrinsic approach), that is the official topographic dataset administered by the Polish Mapping Agency.

The volume of attributive information such as building types and their proper names was calculated for all OSM building features to quantify attribute completeness, a data quality element of thematic accuracy. This automatic procedure was followed by attribute accuracy evaluation based on manual arbitrary comparison of the attributes of corresponding features.

The positional accuracy analysis was based on a manual measurement of the building corner points within OSM dataset and their corresponding points within the reference dataset. The measurements were performed on a fair random sample of OSM buildings evenly distributed within the urban and rural test areas. Spatial accuracy was quantified using the Root Mean Square Error (RMSE).

The resulting high compatibility between the position of building footprints in OSM and the reference set created the basis for automated matching algorithm choice. The feature matching step was a part of the feature completeness investigation. To achieve more reliable results of OSM building completeness analysis, the logical and semantic heterogeneity between two compared datasets were minimised in advance. Moreover, the official data was not considered as the only legitimate reference. Consequently, a novel tool was introduced to help assess the completeness of any dataset of polygon features.

4 STUDY AREA AND DATASETS

The test area, situated in the central-eastern Poland, consists of two sub-areas: Siedlce town (urban district) covering less than 32 sq.km area and a fifty-fold greater area, Siedlce district (rural district) (see Table 1 for their basic characteristics). Siedlce is an average Polish town in terms of both the demographics and economic development. On the other hand, the Siedlce district, surrounding the town, might be described as a poorly urbanized and rather loosely populated area composed of 13 rural communes.

Table 1: The general characteristics of the test sub-areas for the study area Siedleckie County.

District name	District type	Total population	Population density [people per km²]	Total area [km²]	Area after agricultural land and forests deduction [km²]
Siedlce district	rural	81,811	51	1603.3	129.2
Siedlce town	urban	76,603	2,404	31.8	31.8

The test data consists of OSM building-tagged features obtained from the OSM web service, Geofabrik (www.geofabrik.de), in the ESRI shape format. It reflects the state as of May 28, 2015. The examined OSM dataset contains 24,000 objects represented by polygon, most of which lies in Siedlce town (21,434).

Table 2: An overview of the datasets used in this study.

	OSM building data	BDOT10k building data
Definition	Missing.	Unambiguous definition (MSWiA, 2011).
Mapping rule	No strict rules, recommendations only: If possible outer edge of the building wall should be mapped. The outline of building blocks or other complex arrangement of properties allowed.	Building footprint or maximal outline.
Data capture procedure	GPS equipped cell phone or other handheld GPS-device, aerial ortophoto vectorization, sketch drawing from street level, data import from available spatial data sources.	Land and Property Register or other state registers, professional land surveying or ortophoto vectorization.
Accuracy, level of detail	Heterogeneous accuracy and level of detail, depending on the data collection method, generalization level of a building outline, and the contributor's skills and experience.	The level of detail and accuracy equivalent to the scale of 1:10,000.
Quality control	Respect for the OSM consensus norms community, e.g. code of conduct, good practices; Often: geometric and descriptive data verification by introducing a new measurement by any OSM contributor; Potential: intrinsic quality checks (OSM, 2015a, 2015c) using available tools,	Measurement rules and technical supervision over measurements, as well as a system to control data (topology and geometry checks, semantic, syntactic and attribute checks, etc.),
Up-to-datedness	Heterogeneous. Intended to be continuously up-to-date; depends on the contributors' activity,	Homogeneous; kept up-to-date (in practice, updating on a yearly basis),

As a reference data, Polish Database of Topographic Objects (BDOT10k), maintained by the Head Office of Geodesy and Cartography in Poland, were used. BDOT10k is a 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. The ESRI shape format data subset, whose last revision date was August 31,2013, was provided.

OSM buildings were compared with the objects belonging to the BDOT10k group of object classes called 'buildings, building structures and facilities'. In particular, the areal features from the following classes 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. The both studied datasets differ much in data collection and management approaches as can be seen from Table 2 that summarizes their selected characteristics.

The datasets pre-treatment regarded the spatial reference harmonization by 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.

5 THEMATIC ACCURACY

The ISO data quality standard defines thematic accuracy as the *accuracy of quantitative attributes and the correctness of non-quantitative attributes, of the classifications of features and of their relationships* (ISO, 2013). The author of this paper agrees with Koukoletsos (2012) that in the VGI context, thematic accuracy encompasses mostly attribute accuracy along with attribute completeness. The latter needs to be examined here because of possible existence of features lacking their attributes. While classification correctness is barely applicable to OSM quality evaluation due to unlimited range of possible attribute values and their infrequent provision (Al Bakri and Fairbairn, 2010).

Attribute completeness was measured quantitatively as the proportion of the number of OSM buildings that are accompanied with its attribute to the total number of OSM buildings, in percentage. Two attributes were studied, namely building type and building proper name. The results are presented in tab.3.

Table 3: The results of the attribute completeness study based on OSM buildings within the test sub-areas.

	OSM buildings in total	attribute completeness	
		building type	building name
rural district	2,566	32.2%	0.89%
urban district	21,434	76.4%	0.47%

The results of the attribute completeness study confirm the heterogeneous OSM quality across the test site. More developed areas receive more than twice as much contributions as rural ones, as far as the attribute of building type is concerned. It may be associated with a weaker need for knowledge of how buildings are used in rural areas, where generally there are few service buildings and their position is well known to local people (i.e. locals do not need a map to get there). While the attribute that carries information about the proper name of the building is practically not provided; its completeness is below 1% (Table 3).

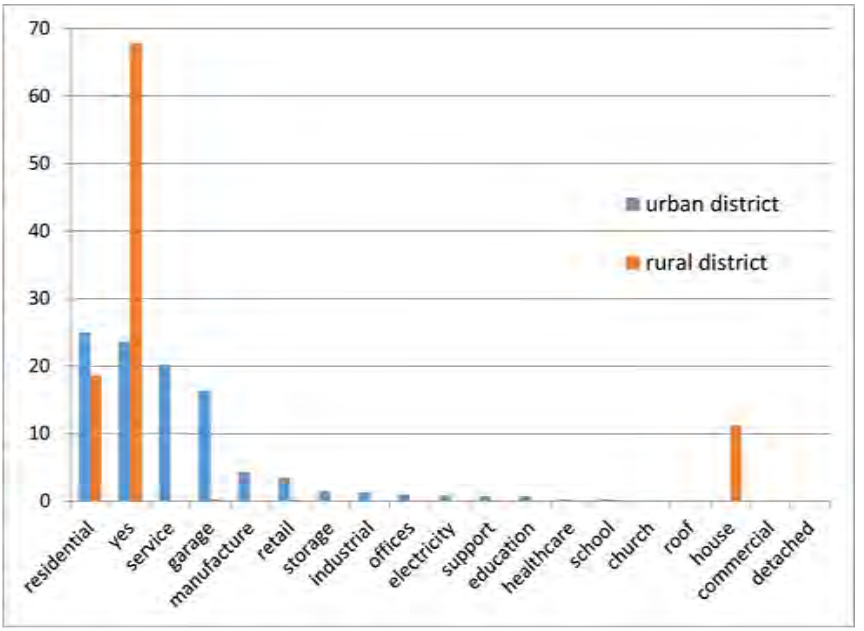

Figure 1: The 'building' features tagged based on their type making up at least 0.2% of the total share as registered in OSM for the urban (blue) and rural district (red).

The list of key values used to tag building type/use includes 10 and 14 items, contributing to at least 0.2% of total share, for the Siedlce district and the town of Siedlce respectively (Figure 1). As many as 67.8% of the buildings located in the rural district and 23.6% in the urban district have the 'yes' tag, which means no information about their use or function.

The attribute deficiencies in buildings featured in the OpenStreetMap database may result from the OSM data collection methods. Often, they cannot be detect based only on satellite or aerial images. Similarly, it is not easy to determine the function of a building observing it form the street level. A high rise can serve as an apartment building, an office building or the seat of a museum of modern arts. Moreover, diverse construction customs - resulting from history or mandated by law in different countries - may distort one's visual assessment, particularly in the case of non-local observers.

In view of the OSM attribute thematic accuracy analysis, the most frequently provided OSM building attribute was chosen. This attribute, referred here to as building type attribute, currently reflects the contributions as for the mapped building typology (i.e. physical nature of the building) or its intended (or original) function or its use (OSM, 2015b). Such a wide range of often contradictory roles prove the previously mentioned problem of the vagueness and ambiguousness of OSM building thematic data. Therefore, its attribute accuracy analysis is not straightforward and it requires OSM semantics better understanding. Consequently, semantic similarity analysis between OSM building data and the Polish Database of Topographic Objects is on-going (the initial findings can be found at (Nowak da Costa, 2016)).

For the purpose of this work, the attribute thematic accuracy analysis was carried out manually and therefore was limited in scope because it required creating time-consuming semantic correspondence

rules for each attribute. The attribute accuracy was evaluated in scrutiny for a small sample of 82 OSM buildings that had their type attribute provided. The accuracy of this attribute was defined as the percentage of the OSM buildings having their attribute equal or very similar to the adequate attribute of the corresponding building feature within the reference dataset. For the studied OSM data sample, the medium level performance of the type attribute accuracy, that is 78%, was obtained.

6 POSITIONAL ACCURACY

Positional accuracy, the component of geometric accuracy, can be defined as a measure of the difference between the position of a distinct object as recorded in the database, and its true location on the ground (Goodchild and Hunter, 1997). Usually, this accuracy is assessed using a reference dataset of higher quality. In the study, the positional accuracy analysis was based on manual measurement of the corner points of building footprints within OSM dataset and their corresponding points within the BDOT10k dataset. The measurements were performed on a fair random sample of OSM buildings, evenly distributed within the urban and rural test areas. On total 782 buildings were measured, where the average number of points measured per building was 5.

Table 4: The accuracy results of the building positions.

	Total number of OSM buildings matched with reference buildings	Number of measured buildings	Minimum/maximum deviation [m]	Mean position deviation [m]	RMSE [m]
rural district	2484	371	0.2/9.4	1.3	1.69
urban district	17917	411	0.1/6.5	0.3	0.59

The set of obtained positon differences is characterised by high discrepancies; the minimum deviation is practically equal to zero, while the maximum reaches almost 10 meters (Table 4). Such heterogeneity is attributed to the variety of methods used by VGI data collectors, their skills and experience.

The positional accuracy was quantified using a traditional statistical measure, the Root Mean Square Error (RMSE). On average, the OpenStreetMap data is within 0.6 and 1.7 meters of the position recorded in the Polish Database of Topographic Objects, for urban and rural test area respectively. The fact that there is practically no positional mismatch between buildings of the two tested data sources, created the basis for the automated matching method choice (see section 7.3). Moreover, the study confirmed that the positional quality of OSM building data is related to urbanization level. In the rural test area, the OSM data quality is, on average, three times worse than in the urban area.

The research reveals surprisingly high positional accuracy of OSM building features, which technically exceeds the accuracy of common handheld GPS receivers or accuracy of available images' amateur vectorization. This may indicate that a part of data was imported in digital form from other spatial databases characterized by high detail level and accuracy.

7 FEATURE COMPLETENESS

Data completeness refers to an indication of whether or not all the data, i.e. features, their attributes and their relationships, are available in the data resource. This chapter focuses on building feature completeness.

7.1 Data pre-treatment

The BDOT10k building data includes all residential and non-residential buildings with the exception of small objects with an area smaller than 40 m^2; however, in the case of small but interconnected structures sharing the same function (e.g. detached garages), they are aggregated and included in the dataset (MSWiA, 2011). In order to ensure comparability across the two studied datasets, the analogous logical constraint was applied to the OSM building data (excluding small features except the detached ones).

Furthermore, as mentioned in section 5, the author investigated the conceptual fuzziness of the OSM features tagged '*building*' and their typologies (*building:type=**). To minimize differences at the semantic level, the reference and the tested datasets have been narrowed down so that they include mostly buildings related to human habitation, educational, healthcare and religious buildings, commercial and main industrial buildings, car garages and sport facilities.

7.2 Choice of measures

The extensive study by Hecht et al. (2013) presents the two significant object-based approaches, the centroid and overlap method, for measuring building completeness using an extrinsic method; and the level of data completeness is determined as a proportion of the corresponding reference buildings to the total set of referenced buildings. These methods are barely sensitive to disparities in object modelling between official and VGI data; however, the official German dataset was considered as the only legitimate reference.

To avoid arbitrary outclassing one of the datasets to be compared for feature completeness, a novel rule of benchmark data lack is introduced here. The assumption of a symmetrical relationship between two datasets, where only the presence or the absence of a specific property is considered, allowed for taking advantage of the *resemblance measures* used unremarkably by mathematicians, e.g. (Batagelj and Bren, 1995), statisticians, e.g. (Czekanowski, 1913; Gower and Legendre, 1986), and environmentalists, e.g. (Legendre and Legendre, 1998).

One of the simplest, and also most frequently selected, coefficients that determine the degree of similarity between two sets is the Jaccard Index (Jaccard, 1901). The Jaccard Index is expressed as a quotient of the cardinality of sets intersection and the cardinality of sets unions. A Polish statistician, Czekanowski (1913), suggested a similar ratio. The Czekanowski's coefficient (also referred to as Bray-Curtis), however, gives more weight (i.e. importance) to the intersection of sets (here: the OSM features that have their counterparts in BDOT10k, and the other way round – based on symmetry assumption). Both coefficients range between 0 and 1, which facilitates comparisons and interpretation of results.

The Jaccard's and Czekanowski's coefficients are defined on the cardinality of a set, which is equivalent to the number of elements in a set. Therefore, their direct adoption for the purpose of feature completeness assessment is greatly affected by the way the number of homologous objects are determined, and may yield confusing results (as depicted in Figure 2).

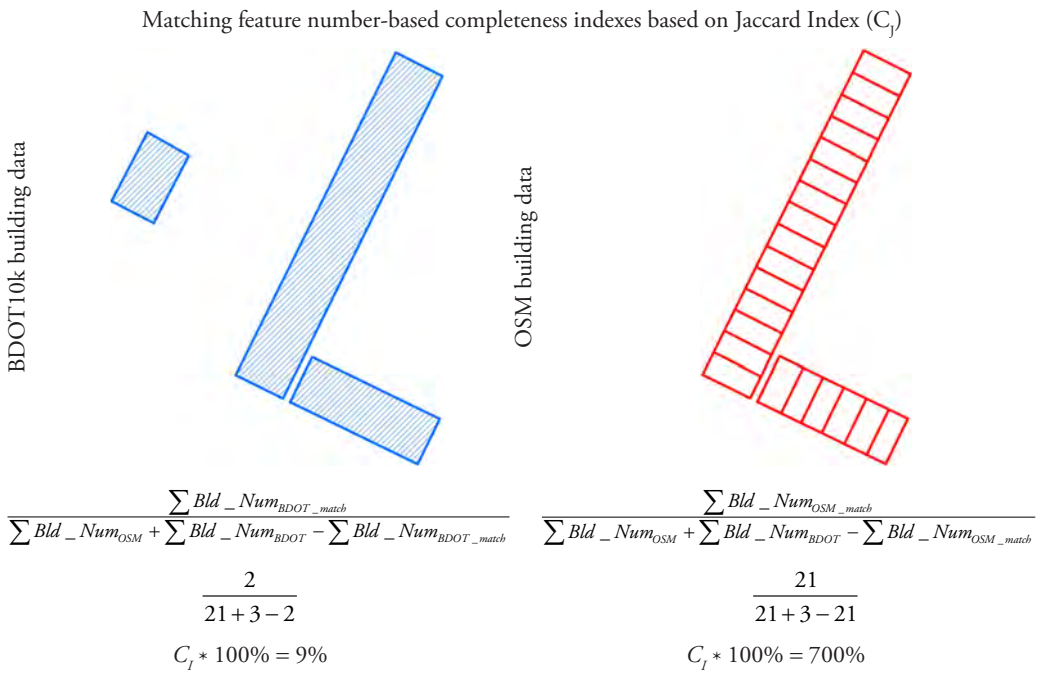

Figure 2: An example of the application of the data completeness evaluation technique based on the Jaccard Index and the spatial objects' number.

The key to the riddle of '700%' are modeling differences between the two sets under examination and, in particular, significantly different levels of data generalization. In the provided example (fig.2), which is based on the real OSM and BDOT10k data from the test area, small adjoining garage building constitute 21 features in the OSM dataset, while in the reference set they are - in compliance with technical guidelines - adequately aggregated and they constitute only two features. If the number of OSM objects is adopted as the power of the homologous objects set, then its value greatly exceeds the overall number of objects in the reference set. This is an example of a failure to satisfy the axiomatic requirements for the application of resemblance measures. If, however, the equal importance is assigned to the area unit (e.g. 1m^2) instead of the feature unit, the requirements are met. Therefore, in the interest of universality, building area substitutes building number determining OSM building completeness (C_{JI}) as follows:

$$C_{JI} = \frac{\sum Bld\_Area_{OSM\_match}}{\sum Bld\_Area_{OSM} + \sum Bld\_Area_{Ref} - \sum Bld\_Area_{OSM\_match}} \quad (1)$$

The respective adaptation of Czekanowski's coefficient for the purposes of determining OSM building completeness (C_{CzI}) is:

$$C_{CzI} = \frac{2 \times \sum Bld\_Area_{OSM\_match}}{\sum Bld\_Area_{OSM} + \sum Bld\_Area_{Ref}} \quad (2)$$

Where: $\Sigma Bld\_Area_{OSM\_match}$ stands for the area of OSM buildings that fulfil the matching criterion, ΣBld_Area_{OSM} - the total area of OSM buildings, and ΣBld_Area_{Ref} is the total area of BDOT10k buildings.

We called C_{JI} and C_{CzI} indexes the **matching feature area-based completeness indexes** based on Jaccard and Czekanowski index, respectively. They are barely sensitive to the building outline modeling and to its degree of abstraction. The completeness indexes values for the example depicted in fig.2 are C_{JI}*100% = 86% and C_{CzI}*100% = 92%. Observably, the coefficients (1) and (2) yield similar results (compare also fig.3). The author is more inclined to the C_{CzI} index based on Czekanowski idea since the calculated C_{CzI} index values are closer to the intuitive (visual) assessment of completeness.

7.3 Matching method

In order to determine homologous objects in both datasets, an automated matching of buildings was carried out. Since the results of the positional accuracy analysis did not reveal significant spatial shifting between manually matched objects, therefore a spatial selection method based on centroid position can be applied as follows. The matching criterion is successfully met if OSM features have their centroid in a reference polygon or a reference feature's centroid lies within an OSM polygon.

The main reason for choosing such algorithm was its simplicity and immediate availability; it uses the common and simple GIS selection method based on spatial location. Moreover, this empiric matching method is proved to be barely sensitive to the discrepancies in the building outline modelling, and to the degree of abstraction, in particular. Its performance was tested on several data samples, extracted randomly from the given OSM building dataset, and compared with manually matched data samples. Only the resting 7% of the building data required manual intervention, mainly the ones characterized by having their geometric centre outside their footprint polygon (non-covex shaped polygon). Also its performance in computational terms is high, although dataset division into smaller sets is recommended.

7.4 Completeness and over-completeness

The OSM completeness was analysed within the administrative borders of 13 communes of Siedlce district, and Siedlce town. The OSM completeness ratio, defined by C_{JI} or C_{Cz} is reported in the form of choropleth map (Figure 3).

The analysis of the obtained OSM building completeness ratio allows to determine the subsequent findings. The OSM building feature completeness is relatively high - that is C_{CzI} of 95% - in the town centres ('M.Siedlce' labelled on fig.3), and its value decreases rapidly as you move away from urban centres. The least completed regions, C_{CzI} of less than 6%, correspond to rural areas with dispersed settlement.

Furthermore, the author investigated the OSM buildings that do not meet the matching criterion. Within the rural test area, those buildings were the subject of visual inspection against both the satellite images shared in the global Internet (e.g. GoogleEarth) and the BDOT10k dataset. The results are summarised in tab.5.

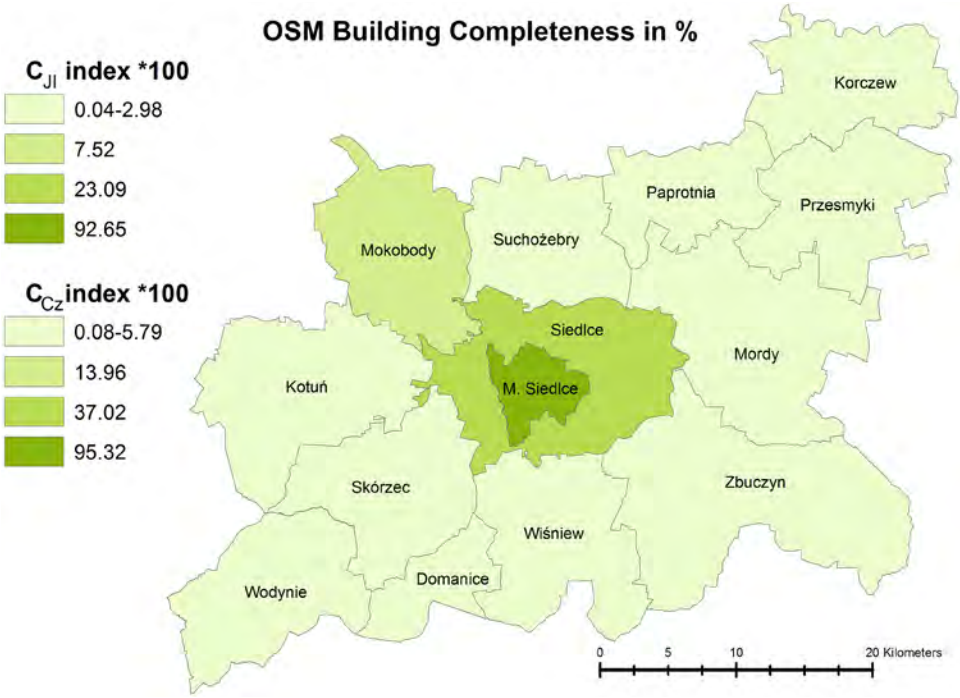

Figure 3: A map of the OSM buildings completeness analysis results for the Siedleckie County as of May, 2015.

Table 5: Typology of OSM buildings in the tested rural district that do not meet the matching criterion.

	Number of instances	Percentage share [%]
OSM buildings non existing in BDOT10k dataset	32	61
OSM buildings belonging to other than the BDOT10k building classes, e.g. roadside shrine, greenhouse	7	13
OSM incorrect building features (measurement error or data entry error)	14	26

It is worth noticing that 61% of objects included in the commission set are correctly registered as building features in the OSM database. They physically exist and they satisfy the technical criteria for BDOT10k building features, nonetheless they are not included in BDOT10k. This proves the OSM building dataset utility on the BDOT10k updating.

To estimate the size of OSM buildings excess (sometimes referred to as VGI over-completeness or data commission), a coefficient analogical to the C_{CzI} index is proposed, as follows:

$$CE_{CzI} \quad \frac{2 \times \sum Bld\_Area_{OSM\_NOmatch}}{\sum Bld\_Area_{OSM} \quad \sum Bld\_Area_{Ref}} \tag{3}$$

Where: $\Sigma Bld\_Area_{OSM\_NOmatch}$ stands for the area of OSM buildings that do not fulfil the matching criterion, ΣBld_Area_{OSM} is the total area of OSM buildings, and ΣBld_Area_{Ref} is the total area of reference buildings.

The CE_{CzI} index to express the OSM building excess, is named the **feature area-based excess indicator**.

7.5 OSM to enhance official spatial datasets

In the countries where official maps are outdated and unequally distributed, OSM is considered a spatial data source until more accurate measurements are available (the Brazilian example can be found at (Camboim et al., 2015)). *On the other hand, where good official data exists and is accessible, crowd-generated data could complement it and provide additional perspectives, without being needed as replacement* (Craglia and Shanley, 2015).

There have already been some attempts of crowd-sourced geodata integration with administrative database, like (Coleman et al., 2010; Siebritz, 2014; Guélat, 2009), proving the idea feasibility. And yet the heterogeneous data quality issue remains. The less time-consuming and non-affecting official data quality alternative is to use the OSM data for detecting changes to the landscape, as also proposed by Sester et al. (2014).

Inspired by harmless 'OSM to change detect' idea, the author proposes a simple method to achieve improved updating of official data based on the volume of building data missing from official database. Instead of time-consuming, targeting country-wide up-to-date coverage, traditional updating cycles, the system prioritises areas requiring topographic map revision and updating. If the feature excess' value, calculated using the proposed feature area-based excess indicator (CE_{Czl}), exceeds the chosen benchmark, it is a premise for a potential update of the examined official reference data subset.

8 CONCLUSIONS AND OUTLOOK

OpenStreetMap data quality research is not a trivial task because of the diversity in which data is collected. All the researchers cited in the paper agree that unrestricted feature conceptualisations, modeling and classifications by volunteer hobbyists and land surveyors alike, generate a melting pot of inconsistent semantics and heterogonous data quality. Since Haklay's and others' studies from 2010, OSM data has changed and it is changing all the time: the contributors, their motivation, interests, skills, and their contributions. Therefore, continuous monitoring of the OSM phenomenon and its data is both necessary and important.

The presented study aimed to understand the applicability of OpenStreetMap building data and assessing its quality in considering its benefits for official spatial databases like the Polish Database of Topographic Objects. The proposed methodology tackles OSM quality in a systematic manner by comparing OSM features with their counterparts from an official dataset of the Polish Mapping Agency. The achieved results are in-line with other similar studies. With regards to OSM building features completeness, the study found that some areas are well mapped especially cities. More concretely, building feature completeness is relatively high in the town centres and its value decreases rapidly further away from urban centres. The study also reveals a very high overall positional accuracy of OSM data: 0.6 m in urban areas and 1.7 m in rural areas. This is more than five times greater than what Haklay noted in 2010. With respect to thematic accuracy, the attribute completeness is low and it relies on the nature of the attribute and the level of urbanization of test area.

In this work, a novel tool is introduced to help assess the completeness of OSM building-tagged features. The proposed index, called the matching feature area-based completeness, is flexible because it is neither

affected by discrepancies in the feature outline modelling nor by the degree of abstraction. Although this novel tool was only applied on building data, it is applicable to any polygon (areal) feature datasets. In addition, the author proposed a method to update the official register using the large volume of OSM building data "over-completeness" together with the building data excess indicator.

In conclusion, the author is of the opinion that in order to fully appreciate the OSM data value, there is a need to understand OSM semantics better. For this reason, the semantic similarity analysis between OSM building data and official data from the Polish Database of Topographic Objects is ongoing. While the procedures of the thematic correspondence, thematic and positional accuracy evaluation are being automated.

Acknowledgments

The author would like to thank Prof. Elżbieta Bielecka, of the Military University of Technology, for being patient and supportive. The author would also like to thank the two anonymous reviewers who have helped improve the manuscript. This research was made possible thanks to a statutory grant from the Institute of Geodesy of the Military University of Technology, no. PBS/933/2016.

Literature and references:

Al Bakri, M., Fairbairn, D. (2010). Assessing the accuracy of 'Crowdsourced' data and its integration with official spatial data sets. In Proceedings of the 9th international symposium on spatial accuracy assessment in natural resources and environmental sciences, Leicester, UK, pp. 317–320.

Ballatore, A., Wilson, D.C., Bertolotto, M. (2013). Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. Knowledge and Information Systems (KAIS), 37 (1), 61–81. DOI: http://dx.doi.org/10.1007/s10115-012-0571-0

Batagelj, V., Bren. M. (1995). Comparing resemblance measures. Journal of Classification, 12, 73–90. DOI: http://dx.doi.org/10.1007/BF01202268

Bielecka, E., Leszczynska, M., Hall, P. (2014). User perspective on geospatial data quality. Case study of the Polish Topographic Database. In The 9th International Conference "ENVIRONMENTAL ENGINEERING" 22–23 May 2014, Vilnius, Lithuania, selected papers. DOI: http://dx.doi.org/10.3846/enviro.2014.193

Camboim, S.P., Meza Bravo, J.V., Robbi Sluter, C. (2015). An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. ISPRS International Journal of Geo-Information, 4 (3), 1366–1388. DOI: http://dx.doi.org/10.3390/ijgi4031366

Coleman, D.J., Sabone, B., Nkhwanana, N.J. (2010). Volunteering Geographic Information to Authoritative Databases: Linking Contributor Motivations to Program Characteristics. Geomatica, 64 (1), 27–40.

Cooper, A., Coetzee, S., Kourie, D., Kaczmarek, I., Iwaniak, A., Kubik, T. (2012). Volunteered geographical information – the challenges. PoPositionIT, Jan/Feb 2012, 34–38.

Coote, A., Rackham, L. (2008). Neogeography data quality—is it an issue? In Holcroft, C. (ed), Proceedings of AGI Geocommunity'08. Association for Geographic Information (AGI), Stratford-Upon-Avon, UK, p. 1.

Craglia, M., Shanley, L. (2015). Data democracy - increased supply of geospatial information and expanded participatory processes in the production of data. International Journal of Digital Earth, 8 (9), 679–693, DOI: http://dx.doi.org/10.1080/17538947.2015.1008214

Czekanowski, J. (1913). Zarys metod statystycznych w zastosowaniu do antropologii. Travaux de la Société des Sciences de Varsovie. III. Classe des sciences mathématiques et naturelles, no. 5. Warsaw: Société des Sciences de Varsovie.

Dorn, H., Törnros, T., Zipf, A. (2015). Quality Evaluation of VGI using Authoritative Data – A Comparison with Land Use Data in Southern Germany. ISPRS International Journal of Geo-Information, 4 (3), 1657–1671, DOI: http://dx.doi.org/10.3390/ijgi4031657

Elwood, S., Goodchild, M.F., Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial data, geographic research, and new social practice. Annals of the Association of American Geographers, 102 (3), 571–590. DOI: http://dx.doi.org/10.1080/00045608.2011.595657.

Esmaili, R., Naseri, F., Esmaili, A. (2013). Quality Assessment of Volunteered Geographic Information, American Journal of Geographic Information System, 2 (2), 19–26. DOI: http://dx.doi.org/10.5923/j.ajgis.20130202.01

Fan, H.; Zipf, A.; Fu, Q.; Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science, 28 (4), 700–719. DOI: http://dx.doi.org/10.1080/13658816.2013.867495

Flanagin, A. J., Metzger, M. J. (2008). The credibility of Volunteered Geographic Information. GeoJournal, 72 (3–4), 137–148. DOI: http://dx.doi.org/10.1007/s10708-008-9188-y

Girres, J. F., Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. Transactions in GIS, 14 (4), 435–459. DOI: http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x

Goodchild, M. F. (2007). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research, 2, 24–32.

Goodchild, M. F., Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth, 3 (3), 231–241. DOI: http://dx.doi.org/10.1080/17538941003759255

Goodchild, M. F., Hunter, G. J. (1997). A simple positional accuracy measure for linear features. International Journal of Geographical Information Science, 11 (3), 299–306.

Goodchild, M. F., Li, L. (2012). Assuring the quality of Volunteered Geographic Information. Spatial statistics, 1, 110–120. DOI: http://dx.doi.org/10.1016/j.spasta.2012.03.002

Gower J. C., Legendre P. (1986). Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3 (1), 5–48. DOI: http://dx.doi.org/10.1007/bf01896809

Guélat, J. C. (2009). Integration of user generated content into national databases - Revision workflow at SwissTopo. 1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases, Wabern, Switzerland.

Haklay, M. (2010). How good is Volunteered Geographic Information? a comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning. B, Planning and Design, 37 (4), 682–703. DOI: http://dx.doi.org/10.1068/b35097

Hecht, R., Kunze, C., Hahmann S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. ISPRS International Journal of Geo-Information, 2, 1066–1091. DOI: http://dx.doi.org/10.3390/ijgi2041066

ISO (2013). ISO 19157:2013 Geographic information – Data quality, International Standard. Geneva: International Organization for Standardization (ISO).

Jaccard, P. (1901). Étude comparative de la distribuition florale dans une portion des Alpes et des. Jura', Bulletin Societe'Vandoise des sciences naturelles, 37, 547–579.

Klonner, C., Barron, C., Neis, P., Höfle, B. (2014). Updating digital elevation models via change detection and fusion of human and remote sensor data in urban environments. International Journal of Digital Earth, 8 (2), 153–171. DOI: http://dx.doi.org/10.1080/17538947.2014.881427

Koukoletsos, T. (2012). A Framework for Quality Evaluation of VGI linear datasets. Doctoral thesis- London: University College London (UCL).

Legendre, P., Legendre, L. (1998). Numerical ecology, 2nd Eng. edition. Amsterdam: Elsevier.

Mooney, P., Corcoran, P. (2012). The annotation process in OpenStreetMap. Transactions in GIS, 16 (4): 561–579. DOI: http://dx.doi.org/10.1111/j.1467-9671.2012.01306.x

MSWiA (2011). Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 17 listopada 2011 r. w sprawie bazy danych obiektów topograficznych oraz bazy danych obiektów ogólnogeograficznych, a także standardowych opracowań kartograficznych (Dz.U. 2011 nr 279 poz. 1642).

Nowak Da Costa, J. (2016). Towards Building Data Semantic Similarity Analysis: OpenStreetMap and the Polish Database of Topographic Objects. BGC Geomatics, pages: 269-275. DOI: http://dx.doi.org/10.1109/BGC.Geomatics.2016.55

OSM (2015a). Good Practice - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Good_practice, accessed 16. 6. 2016.

OSM (2015b). Open Discussion – Key:Building - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Talk:Key:building, accessed 16. 6. 2016.

OSM (2015c). Quality Assurance - OpenStreetMap Wiki. http://wiki.openstreetmap.org/wiki/Quality_assurance, accessed 16. 6. 2016.

Sester, M., Jokar Arsanjani, J., Klammer, R., Burghardt, D., Haunert, J. (2014). Integrating and Generalising Volunteered Geographic Information. In Abstracting Geographic Information in a Data Rich World, Part of the series Lecture Notes in Geoinformation and Cartography, pp. 119–155, Springer. DOI: http://dx.doi.org/10.1007/978-3-319-00203-3_5

Siebritz, L. (2014): Assessing the accuracy of OpenStreetMap data in South Africa for the purpose of integrating it with authoritative data. University of Cape Town.

Turner A. (2006). Introduction to Neogeography, O'Reilly Media Short Cuts Series.

Zielstra, D., Hochmair, H.H., Neis, P. (2013). Assessing the effect of data imports on the completeness of OpenStreetMap—A United States case study. Transactions in GIS, 17 (3), 315–334. DOI: http://dx.doi.org/10.1111/tgis.12037

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 (3): 495-508. DOI: 10.15292/geodetski-vestnik.2016.03.495-508

Asst. Prof. Joanna Nowak Da Costa, PhD.
Military University of Technology
Institute of Geodesy, Gen. S. Kaliskiego 2
01-476 Warsaw, Poland
e-mail: joanna.nowakdc@wat.edu.pl