

WOJSKOWA AKADEMIA TECHNICZNA
im. Jarosława Dąbrowskiego

WYDZIAŁ INŻYNIERII LĄDOWEJ I GEODEZJI



ROZPRAWA DOKTORSKA

**METODYKA WIELOCECHOWEJ OCENY
PORÓWNAWCZEJ PRZYDATNOŚCI
JAKOŚCIOWYCH DANYCH PRZESTRZENNYCH**

Autor: mgr inż. Sylwia BORKOWSKA

Promotor: płk dr hab. inż. Krzysztof POKONIECZNY

Warszawa 2024

Rozprawa doktorska wykonana przez doktoranta: mgr inż. Sylwię Borkowską

dziedzina nauki: nauki inżynieryjno-techniczne

dyscyplina naukowa: inżynieria lądowa, geodezja i transport

METODYKA WIELOCECHOWEJ OCENY PORÓWNAWCZEJ PRZYDATNOŚCI JAKOŚCIOWYCH DANYCH PRZESTRZENNYCH

Dzięki wszechobecności zaawansowanych technologii internetowych oraz urządzeń lokalizacyjnych w ciągu ostatnich dwudziestu lat wszyscy użytkownicy na świecie, niezależnie od ich wiedzy lub doświadczenia, są w stanie tworzyć informacje przestrzenne. Zjawisko to określane jest jako dobrowolna informacja geograficzna VGI (Volunteered Geographic Information). Dane pozyskiwane w ramach VGI wykorzystywane są jako źródło wspierające szeroki zakres usług, tj. monitorowanie środowiska, raportowanie zdarzeń czy zarządzanie sytuacjami kryzysowymi. Należy jednak podkreślić, iż takie heterogeniczne dane przestrzenne dostarczane przez wolontariuszy mają różną jakość. Z tego powodu standardowe wskaźniki jakości danych przestrzennych w kompleksowym podejściu oceny jakości spektrum danych wolontariackich mogą nie być wystarczające z punktu widzenia współczesnego użytkownika.

W niniejszej pracy podjęłam problematykę zewnętrznej oceny jakości wolontariackich danych OpenStreetMap (OSM) w odniesieniu do referencyjnej Bazy Danych Obiektów Topograficznych (BDOT10k) dla głównych sześciu klas pokrycia terenu, wykorzystując do tego standardowe mierniki jakości danych przestrzennych oraz własne autorskie wskaźniki, które przedstawiłam w ramach cyklu składającego się z czterech powiązanych tematycznie artykułów naukowych. Przeprowadzone badania dotyczyły wybranych siedmiu powiatów w Polsce, zróżnicowanych pod kątem środowiska naturalnego jak i antropogenicznego. Pierwszy etap badań poświęcony został opracowaniu metodyki przetworzenia danych przestrzennych OSM w celu dokonania ich zewnętrznej oceny jakości. Ocenę tę przeprowadziłam zgodnie z wskaźnikami ISO, które dodatkowo uzupełniłam o autorskie mierniki, służące do holistycznej oceny jakości oraz pokazania jej zróżnicowania wewnątrz analizowanych zbiorów. Następnie opracowałam metodykę kartograficznej wizualizacji wyników, przedstawiającą prawidłowości w wynikach oceny jakości analizowanych zbiorów danych przestrzennych.

Zgodnie z przeprowadzonym przeglądem literatury i wskazanymi tam głównymi ograniczeniami istniejących wskaźników oceny jakości niejednorodnych danych OSM, określonymi jako niewystarczająco relatywne, w dalszym etapie badań zajęłam się wyznaczeniem złożonej miary oceny jakości danych topograficznych. Wynikiem prac

przedstawionych w niniejszej rozprawie jest opracowanie autorskiego wskaźnika skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych OSM względem BDOT10k w postaci Compound Correspondence Index (CCI), umożliwiającego użytkownikowi ocenę, który z dwóch porównywanych zbiorów jakościowych danych przestrzennych spełnia jego oczekiwania.

METHODOLOGY FOR MULTI-CRITERIA COMPARATIVE ASSESSMENT OF THE USABILITY OF QUALITATIVE SPATIAL DATA

The ubiquity of advanced internet technologies and location-based devices over the past two decades has made it possible for all users in the world, regardless of their knowledge or experience, to produce spatial information. This phenomenon is referred to as Volunteered Geographic Information (VGI). The data produced by VGI is used as a resource to support a wide range of services, i.e. environmental monitoring, incident reporting or emergency management. However, it should be emphasised that such heterogeneous spatial data provided by volunteers have variable quality. Therefore, standard spatial data quality indicators in a comprehensive approach to assessing the quality of a spectrum of volunteer data may not be sufficient from the point of view of a contemporary user.

In the presented study, I examined the external quality assessment of volunteer OpenStreetMap (OSM) data in relation to the reference Topographic Dataset (BDOT10k) for the main six land cover classes, using standard spatial data quality measures and my own original indicators, which I presented in a series of four thematically related scientific articles. The research was carried out on selected seven counties in Poland, diverse in terms of natural as well as anthropogenic environment. The first stage of the research was devoted to the development of a methodology for processing OSM spatial data in order to carry out their external quality assessment. This assessment was conducted in accordance with the ISO indicators, which I additionally supplemented with my authored quality metrics to holistically assess the quality and show its variation within the analysed data sets. Subsequently, I developed a methodology for cartographic visualisation of the results, showing patterns in the quality assessment results of the analysed spatial data sets.

According to the conducted literature review and the main limitations of the existing indicators for assessing the quality of heterogeneous OSM data, identified as insufficiently relative, in a subsequent stage of the research I addressed the determination of a compound measure for assessing the quality of topographic data. The result of the work presented in this dissertation is the development of the author's index of the consolidated analysis of the suitability of OSM spatial data sets in relation to BDOT10k in the form of Compound Correspondence Index (CCI), enabling the user to assess which of the two compared qualitative spatial data fulfils his/her requirements.

SPIS TREŚCI

WYKAZ UŻYTYCH SKRÓTÓW.....	9
1. WPROWADZENIE	10
2. CEL, TEZA, ZAKRES PRACY	13
3. OBSZAR BADAŃ I WYKORZYSTANE DANE	16
3.1. Obszar badań.....	16
3.2. Wykorzystane dane przestrzenne.....	17
4. METODY BADAWCZE	20
4.1. Podstawowe założenia badawcze i ogólny schemat badań	20
4.2. Mierniki jakości danych przestrzennych	23
4.2.1. Mierniki jakości zbiorów danych przestrzennych	23
4.2.2. Współczynnik CCI skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych	28
4.2.3. Autokorelacja przestrzenna.....	31
4.2.4. Wnioskowanie statystyczne	34
5. OPIS WYNIKÓW	35
5.1. Analiza jakości danych OSM zgodnie z miernikami ISO [publikacje A1 oraz A2].....	35
5.2. Kartograficzna wizualizacja wyników oceny kompletności danych przestrzennych dla wybranego przykładu [publikacja A2].....	39
5.3. Opracowanie złożonego indeksu jakości danych przestrzennych [publikacja A3].....	43
5.4. Analiza wrażliwości indeksu CCI [publikacja A4]	47
6. PODSUMOWANIE	53
7. WNIOSKI.....	56
LITERATURA	57
ZAŁĄCZNIKI	62

WYKAZ UŻYTYCH SKRÓTÓW

ATKIS – *ang. Authorative Topographic-Cartographic Information System*, cyfrowy zbiór danych wektorowych o strukturze obiektowej

BDOT10k – Baza Danych Obiektów Topograficznych, odpowiada w ogólności tradycyjnej mapie topograficznej w skali 1:10 000

GRS80 – *ang. Geodetic Reference System '80*, geodezyjny system odniesienia

ISO – *ang. International Organization for Standardization*, Międzynarodowa Organizacja Normalizacyjna

MCDA – *ang. Multi-Criteria Decision Analysis*, wielokryterialna analiza decyzji

MU – *ang. Map Unit*, jednostka mapowa

OSM – *ang. OpenStreetMap*

PUWG 1992 – Państwowy Układ Współrzędnych Geodezyjnych 1992

TOPSIS – *ang. Technique for Order of Preference by Similarity to Ideal Solution*, metoda porządkowania liniowego

VGI – *ang. Volunteered Geographic Information*, dane geograficzne pozyskiwane przez wolontariuszy

WGS84 – *ang. World Geodetic System 1984*, system odniesienia

WLC – *ang. Weighted Linear Combination*, ważona kombinacja liniowa

1. WPROWADZENIE

Dane przestrzenne są podstawą podejmowania większości decyzji, wg Białousza (2004), Bieleckiej i Maja (2009) co najmniej 80% decyzji podejmowanych przez administrację publiczną wymaga odniesienia do wektorowych danych przestrzennych. Od dwóch dekad wielość i stosunkowo łatwa dostępność danych przestrzennych stawiają użytkownika przed dokonaniem trudnego wyboru zbioru spełniającego oczekiwania związane z wykonaniem konkretnego zadania. Urzędowe i komercyjne zbiory danych przestrzennych są najczęściej opisane metadanymi zgodnymi z międzynarodowym standardem ISO 19115:2014 (2014). Stosowane w opisie jakości mierniki mają charakter normatywny i są zdefiniowane w normie ISO 19157-1:2023 (2023).

Zgodnie z normą ISO 19157-1:2023 jakość danych przestrzennych rozumiana jest jako zbiór następujących charakterystyk i atrybutów obiektów zgromadzonych w bazie danych:

- Dokładność geometryczna – opisuje dokładność określania współrzędnych obiektu.
- Dokładność tematyczna – opisuje dokładność lub pewność pozyskania wartości atrybutu.
- Aktualność – opisuje datę, dla której zawartość bazy danych jest zgodna z rzeczywistością.
- Kompletność – określa, jak wyczerpujący jest zbiór obiektów. Może odnosić się do: nadmiaru (nadkompletności), brakujących obiektów, ich atrybutów lub relacji między nimi.
- Spójność logiczna – opisuje spójność relacji zapisanych w strukturze przestrzennej bazy danych (pojęciowej, dziedzinowej i topologicznej).

Jakość danych OpenStreetMap, a w szczególności jej elementy ilościowe, są szeroko interesujące dla potencjalnych użytkowników na całym świecie. Metoda zbierania danych stosowana w OSM uniemożliwia bezpośrednie zastosowanie zasad oceny danych geograficznych zawartych w normie ISO 19157, które odnoszą się do porównania danych ze specyfikacjami technicznymi. Goodchild i Li (2012) wymienili trzy alternatywne podejścia do oceny jakości danych geograficznych pozyskanych w ramach projektów takich jak OpenStreetMap:

- 1) Podejście oparte na crowdsourcingu – uznające założenie, że użytkownicy wykrywają i korygują błędne dane.

- 2) Podejście społecznościowe – zakładające minimalną liczbę kontroli poprawności danych przez administratorów.
- 3) Podejście geograficzne – obejmujące wykorzystanie programów typu GIS do kontroli jakości danych poprzez sprawdzanie poprawności topologii i reguł logicznych.

Zgodnie z wymienionymi standardami jakość jest określana jako zgodność z odpowiednimi specyfikacjami według których zbiory zostały opracowane i nazywana jest „jakością producenta” (Bielewa, 2011). Zbiory społecznościowe, tworzone przez wolontariuszy, nie mają typowych specyfikacji technicznych, a jedynie wskazówki i zalecenia dla osób współtworzących w określony sposób. Ich jakość, a w szczególności kompletność, jest zróżnicowana i zależy od aktywności mapowiczów. Dekadę temu Neis et al. (2012) stwierdzili, że jakość, a konkretnie niekompletność i niejednorodność, to największe wady danych OpenStreetMap znacznie ograniczające ich szerokie wykorzystanie. Dwanaście lat później (sierpień 2024) w bazie Web of Science indeksowanych jest prawie 1000 publikacji naukowych, autorstwa kilku tysięcy osób, poruszających kwestie jakości zasobu OSM. Wśród jednostek naukowych dominują uczelnie wyższe, w szczególności Uniwersytety w Heidelbergu, Wuhan, Londynie, Kalifornii i wiele innych. Wojskowa Akademia Techniczna z dziesięcioma publikacjami znajduje się na 25 miejscu na 1176 instytucji badawczych. Naukowcy najczęściej badają jaka jest jakość budynków (Biljecki et al., 2023; Haklay, 2010; Nowak Da Costa et al., 2016), sieci dróg (Barrington–Leigh & Millard–Ball, 2019; Cichociński, 2012) i lasów (Dorn et al., 2012). Według Fan et al. (2014) budynki w OSM w Monachium cechują się wysoką kompletnością i dokładnością semantyczną, natomiast lokalizacja jest przesunięta średnio około czterech metrów w stosunku do danych z ATKIS.

Drogi i koleje często są analizowane jako elementy pokrycia terenu (Zielstra & Zipf, 2010; Haklay, 2010) lub w aspekcie dostępności do usług (Weiss et al., 2020). Obszarowo badania te dotyczą raczej miast (Neis et al., 2012) niż całych krajów lub regionów. Z globalnej analizy jakości dróg wynika, że w wielu miejscach użytkownicy mogą polegać na kompletności OSM. Barrington–Leigh i Millard–Ball (2019) odkryli, że globalnie drogi w OSM są kompletne w 83%, a dla ponad 40% krajów sieć ulic jest w 100% kompletna. Zauważyli, że w wysoko rozwiniętych krajach z dobrym dostępem do Internetu sieć dróg jest kompletniejsza, niż

w pozostałych regionach, ponadto zarówno słabo zaludnione obszary, jak i gęsto zaludnione miasta są najlepiej zmapowane.

Jakość danych o lasach także oceniana jest jako wysoka. Dorn et al. (2015) wartość kompletności dla regionu Rhine–Neckar w Niemczech szacuje powyżej 90% w stosunku do danych publicznych ATKIS. Podobnie Bielecka i Leszczyńska (2018) oszacowały dokładność kartowania lasów przez wolontariuszy w Polsce jako zróżnicowaną wahającą się w granicach od 76,4% (w województwie lubuskim) do 92,5% w Zachodniopomorskiem.

W geodezji i kartografii jakość danych jest szczególnie istotna, ponieważ produkty (mapy, serwisy, modele) powstałe z ich wykorzystaniem udostępnione są publicznie i służą do podejmowania istotnych społecznie oraz gospodarczo decyzji. Ważna jest świadomość zarówno twórców danych, jak i ich użytkowników, dotycząca jakości danych źródłowych i produktów, do opracowania których zostały one wykorzystane. Nie ma danych nieobarczonych błędami (Su et al., 2007). Opracowana przeze mnie metoda umożliwia określenie stopnia zaufania do danych topograficznych.

2. CEL, TEZA, ZAKRES PRACY

Dylematy użytkownika, stojącego przed wyborem zbioru (lub kilku zbiorów) koniecznego do wykonania konkretnego zadania, stały się główną przesłanką do podjęcia badań dotyczących porównawczej analizy jakości danych przestrzennych. Praca doktorska zatytułowana „**Metodyka wielocechowej oceny porównawczej przydatności jakościowych danych przestrzennych**” stanowi cykl czterech powiązanych tematycznie artykułów naukowych, wyszczególnionych w Tabeli 1.

Tabela 1. Cykl powiązanych tematycznie artykułów.

Numer artykułu	Artykuł w cyklu	Punkty wg. MEiN, IF
A1	Borkowska, S. (90%), & Pokonieczny, K. (10%) (2022). Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. <i>Sustainability</i> , 14, 3728. https://doi.org/10.3390/su14073728	100 IF = 3,3
A2	Borkowska, S. (75%), Bielecka, E. (12,5%), & Pokonieczny, K. (12,5%) (2023). OpenStreetMap - building data completeness visualization in terms of “Fitness for purpose”. <i>Advances in Geodesy and Geoinformation</i> , 72, 1, 1–20. https://doi.org/10.24425/agg.2022.141922	70 IF = 2,1
A3	Borkowska, S. (80%), Bielecka, E. (8%), & Pokonieczny, K. (12%) (2023). Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. <i>Applied Sciences</i> , 13, 7525. https://doi.org/10.3390/app13137525	100 IF = 2,5
A4	Borkowska, S. (75%), Bielecka, E. (10%), & Pokonieczny, K. (15%) (2024). Weights Impact on the Comparative Evaluation of Topographic Data. <i>Geomatics and Environmental Engineering</i> , 18, 4. https://doi.org/10.7494/geom.2024.18.4.97	70

Artykuły cyklu zawierają opis kompletnej metodyki umożliwiającej użytkownikowi ocenę, który z dwóch porównywanych zbiorów spełnia jego

oczekiwania. Metodyka została sprawdzona na przykładzie dwóch zbiorów danych topograficznych, tj. Bazy Danych Obiektów Topograficznych i OpenStreetMap. Pierwszy artykuł [A1] analizuje jakość danych społecznościowych w porównaniu do danych urzędowych przyjętych jako zbiór referencyjny. Do analiz jakości wykorzystałam istniejące wskaźniki oceny położenia i kompletności oraz zmodyfikowane wskaźniki kompletności obiektów topograficznych. W kolejnym artykule [A2] zaproponowałam wizualizację jakości danych modyfikując kartogram złożony (dwuzmienny). Ułatwia to użytkownikowi podjęcie decyzji o przydatności zbioru dla konkretnego miejsca w przestrzeni. W artykule trzecim [A3] zdefiniowałam autorski współczynnik Compound Correspondence Index (CCI) przedstawiający konkretne miejsca w przestrzeni (pola podstawowe), dla których zawartość obu zbiorów danych jest bardzo zbliżona oraz te, dla których jest zdecydowanie różna. Wizualizacja kartograficzna oraz dane tabelaryczne umożliwiają szybki wybór przez użytkownika pożądanego zbioru danych topograficznych. W ostatnim artykule [A4] analizuję wrażliwość wskaźnika CCI na zmianę wag przypisanych do obiektów topograficznych. Wizualizacja kartograficzna ilustruje obszary, w których różne wagi zasadniczo zmieniają wynik wartości wskaźnika.

Głównym celem badawczym jest **opracowanie uwarunkowań kompleksowej oceny porównawczej jakości danych przestrzennych**. Na podstawie powyższego problemu sformułowałam tezę badawczą brzmiącą: „**metoda wielocechowej analizy porównawczej i wizualizacji jakości danych oraz autorskie wskaźniki oceny przydatności danych przestrzennych stanowią podstawy do spójnej ich oceny przez użytkownika**”.

Osiągnięcie celu i udowodnienie tezy wymagało weryfikacji hipotez roboczych poprzez udzielenie odpowiedzi na następujące pytania szczegółowe:

P1. Czy zasadne jest opracowanie nowych mierników jakości danych przestrzennych potrzebnych użytkownikowi do szczegółowej i kompleksowej oceny jakości przestrzennych danych społecznościowych?

H1. Mierniki jakości ISO uzupełnione o autorskie mierniki jakości danych są wystarczające do holistycznej oceny jakości danych przestrzennych z perspektywy użytkownika oraz pokazania zróżnicowania jakości wewnątrz analizowanych zbiorów.

Weryfikacja hipotezy roboczej wymagała zdefiniowania dodatkowych autorskich mierników jakości danych, uwzględniających wymiar geometryczny obiektów

przechowywanych w zbiorze danych przestrzennych. Zostało to przedstawione w artykułach **A1, A2, A3** oraz **A4** cyklu.

P2. Czy autorski współczynnik Compound Correspondence Index (CCI) skonsolidowanej analizy odpowiedniości jest wrażliwy na powiększanie obszaru badań? Innymi słowy, czy metoda analizy odpowiedniości (korespondencji) dwóch zbiorów danych przestrzennych prowadzona dla zbiorów z poszczególnych obszarów badawczych oddzielnie (lokalnie) lub łącznie (regionalnie) daje takie same wyniki?

H2. Autorski współczynnik skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych CCI pokazuje większą zgodność zbiorów danych w ujęciu regionalnym.

Odpowiedź na pytanie drugie i weryfikacja hipotezy drugiej zostały przedstawione w artykułach **A3** i **A4**.

P3. Czy autorskie połączenie kartogramów wizualizujących wyniki analiz jakości danych przestrzennych pokazuje prawidłowości w ocenie jakości danych?

H3. Autorski kartogram strukturalny będący połączeniem kartogramu złożonego i kartogramu prostego podkreśla różnice w zakresie liczby obiektów oraz dokładności ich lokalizacji w porównywanych zbiorach.

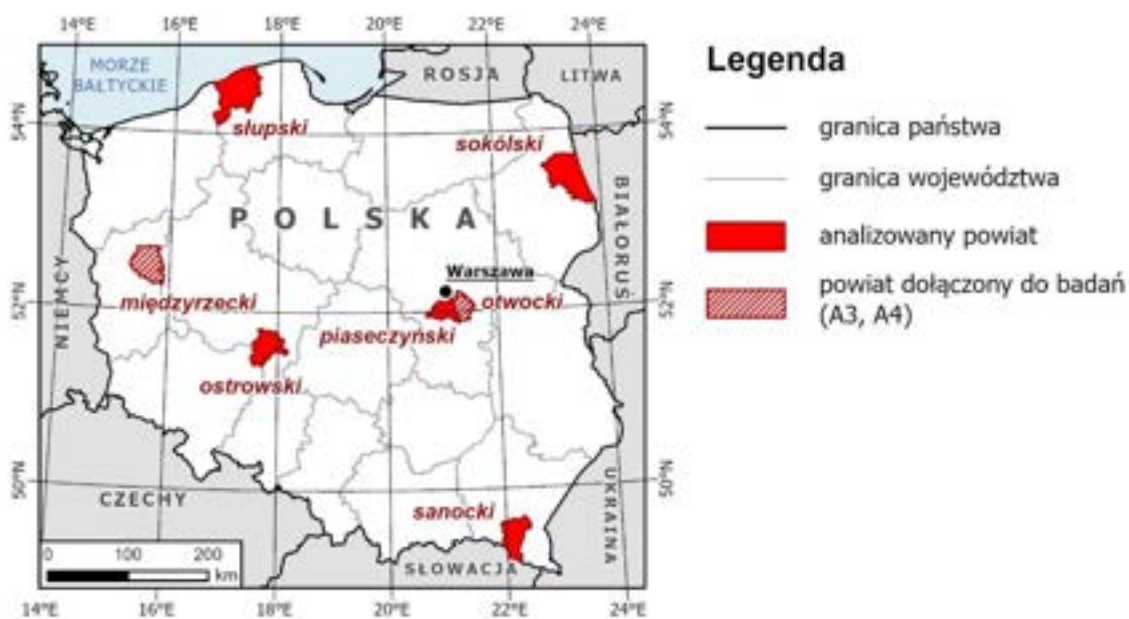
Weryfikacja hipotezy trzeciej została udokumentowana w artykule **A2**.

3. OBSZAR BADAŃ I WYKORZYSTANE DANE

3.1. Obszar badań

Obszarami testowymi było siedem powiatów (Rys. 1):

- piaseczyński, sokólski, sanocki, słupski oraz ostrowski (występujące we wszystkich artykułach);
- otwocki oraz międzyrzecki (włączone do badań w artykułach **A3** oraz **A4**).



Rys. 1. Obszar badań.

Powiaty te znajdują się w różnych mezoregionach fizyczno–geograficznych Polski, których łączny obszar obejmuje 3,1% powierzchni kraju oraz odzwierciedlają różnorodność zarówno środowiska naturalnego, jak i antropogenicznego (Tabela 2), co czyni je obszarami reprezentatywnymi. Dodatkowo ze względu na różne zagrożenia, w tym naruszenie integralności granic państwowych, mają one strategiczne znaczenie dla bezpieczeństwa kraju.

Tabela 2. Ogólna charakterystyka analizowanych powiatów [źródło: A3, Tabela 1].

Cecha	Powiat						
	piaseczyński	sokólski	sanocki	słupski	ostrowski	otwocki	międzyrzecki
Województwo	mazowieckie	podlaskie	podkarpackie	pomorskie	wielkopolskie	mazowieckie	lubuskie
Podprovincia fizyczno–geograficzne	Niziny Środkowo–polskie	Wysoczyzny Podlasko–Białoruskie	Zewnętrzne Karpaty Zachodnie	Pobrzeża Południowo–bałtyckie	Niziny Środkowo–polskie	Niziny Środkowo–polskie	Pojezierza Południowo–bałtyckie

Cecha	Powiat						
	piaseczyński	sokólski	sanocki	ślupski	ostrowski	otwocki	międzyrzecki
Powierzchnia (km ²)	621,12	2054,34	1223,62	2347,59	1159,92	616,46	1387,61
Liczba ludności	190 607	64 902	92 900	98 761	161 581	124 283	57 100
Gęstość zaludnienia (os./km ²)	311	32	81	43	139	202	42
Liczba miast	4	4	2	2	2	3	3
Poziom urbanizacji (%)	47,8	41,7	47,2	20,7	53,7	61,8	52,3
Użytkowanie powierzchni (km²)							
Obszary zabudowane	82,51	72,93	43,65	85,15	55,62	58,91	24,08
Lasy	132,88	547,91	586,67	864,13	347,59	250,15	735,43
Tereny uprawne	387,37	1426,28	512,14	1234,97	728,53	276,51	513,42
Zbiorniki wodne	16,44	6,71	13,60	110,65	13,31	11,14	38,39
Główne obszary chronione	Chojnowski Park Krajobrazowy	Park Krajobrazowy Puszczy Knyszyńskiej, Biebrzański Park Narodowy	Park Krajobrazowy Gór Słonnych, Jaślicki Park Krajobrazowy, Ciśniańsko-Wetliński Park Krajobrazowy	Słowiński Park Narodowy, Dolina Słupi	Park Krajobrazowy Dolina Baryczy, Uroczyska Płyty Krotoszyńskiej	Mazowiecki Park Krajobrazowy, Dolina Środkowego Świdra	Pszczewski Park Krajobrazowy, Puszcza Notecka, Nietoperek

3.2. Wykorzystane dane przestrzenne

W przeprowadzonym badaniu jakości danych wykorzystałam dane OpenStreetMap, stanowiące darmową, otwartą bazę danych geograficznych aktualizowaną i utrzymywaną przez społeczność wolontariuszy za pośrednictwem otwartej współpracy. Dane OSM charakteryzują się heterogeniczną dokładnością i poziomem szczegółowości, w zależności od techniki pozyskiwania, doświadczenia oraz umiejętności edytora bazy. Źródła pozyskiwania danych OSM to przede wszystkim pomiary z przenośnych odbiorników GPS, zdjęcia lotnicze i inne dostępne otwarte źródła danych. Aktualność danych OSM różni się w zależności od aktywności wolontariuszy. OSM ma własną infrastrukturę do przechowywania, udostępniania, wyszukiwania i wizualizacji danych. Dane OSM są przechowywane w relacyjnej bazie danych PostgreSQL, zgodnie z układem odniesienia WGS84.

OSM używa topologicznej struktury danych składającej się z:

- węzłów (nodes) – reprezentują konkretny punkt na powierzchni Ziemi zdefiniowany przez szerokość i długość geograficzną;
- linii (ways) – uporządkowane węzły, reprezentujące linię (drogi, rzeki) lub wielokąt (budynki, lasy);
- relacji (relations) – wielofunkcyjna struktura danych dokumentująca relacje między dwoma lub większą liczbą elementów danych (węzłami, liniami i/lub innymi relacjami);
- tagów (tags) – stosowane do węzłów, linii lub relacji i składają się z pary klucz = wartość. Służą do przechowywania metadanych o obiektach (typ, nazwa, właściwości fizyczne).

Dane OpenStreetMap wykorzystane w analizach pozyskano 24 czerwca 2021 r. z serwisu Geofabrik (2024).

Jako drugi zbiór danych w przeprowadzonych analizach jakości danych wykorzystalam Bazę Danych Obiektów Topograficznych (BDOT10k). Jest to wektorowa baza danych zawierająca przestrzenną lokalizację obiektów topograficznych wraz z ich podstawowymi charakterystykami opisowymi. Zawartość i szczegółowość bazy danych BDOT10k odpowiadają zawartości i szczegółowości tradycyjnej, cywilnej mapy topograficznej w skali 1:10 000. Zasób podstawowy BDOT10k to zbiór obiektów sklasyfikowanych na trzech poziomach szczegółowości i obejmujących swoim zakresem tematycznym 286 rodzajów obiektów zgrupowanych w 57 klasach i 9 kategoriach klas obiektów. Szczegółowy zakres informacji gromadzonych w bazie BDOT10k, ich organizacja, tryb i standardy techniczne tworzenia, aktualizacji, weryfikacji i udostępniania danych określone są w rozporządzeniu (MRPiT, 2021). Baza danych BDOT10k jest dostępna bezpłatnie do dowolnego wykorzystania za pośrednictwem serwisu Geoportal Krajowy (2024). Dane BDOT10k są pozyskiwane poprzez: pomiary geodezyjne, rejestr gruntów i budynków, ortofotomapę lub inne oficjalne rejestry państwowe. W rozporządzeniu określono, że geometrię obiektów BDOT10k pozyskuje się z dokładnością nie mniejszą niż 1,5 m, a w przypadku obiektów, których jednoznaczna identyfikacja w terenie jest utrudniona i zależna od oceny osoby dokonującej identyfikacji – z dokładnością nie mniejszą niż 5 m. Wartości współrzędnych punktów opisujących geometrię obiektów wyraża się w metrach z precyzją zapisu do 0,01 m. Kontrola jakości BDOT10k jest przeprowadzana zgodnie z systemem kontroli danych

przesyłanych do zasobu (sprawdzanie topologii i geometrii, sprawdzanie semantyki, składni i atrybutów itp.) i jest wykonywana w wysokim stopniu szczegółowości. Zestaw danych BDOT10k jest zdefiniowany w prostokątnym układzie współrzędnych PUWG 1992 (Państwowy Układ Współrzędnych Geodezyjnych), który jest układem współrzędnych opartym na mapowaniu Gaussa–Krügera dla elipsoidy GRS80 w jednej dziesięciostopniowej strefie dla Polski. Aktualność dostępnych danych BDOT10k wykorzystanych w tej analizie to marzec 2020 r.

4. METODY BADAWCZE

4.1. Podstawowe założenia badawcze i ogólny schemat badań

W przeprowadzonych badaniach jakości danych przestrzennych zastosowałam jako pole podstawowe siatkę heksagonalną o powierzchni 1 km². Ocena jakości danych przestrzennych w oparciu o pojedyncze powiaty (badania na małą skalę) wymagała zastosowania wystarczająco szczegółowej siatki odniesienia o mniejszych jednostkach niż administracyjne. Siatka sześciokątna jest najbliższa kołu, dając jednocześnie takie samo pełne pokrycie badanego obszaru (Roick et al., 2011). Sześciokąty redukują stroniczość próbkowania z powodu efektów krawędziowych kształtu siatki, co jest związane z niskim stosunkiem obwodu do powierzchni kształtu sześciokąta (Birch et al., 2007). Okrągłość siatki sześciokątnej pozwala na bardziej naturalne odwzorowanie krzywych we wzorcach danych niż siatka kwadratowa.

Opisane analizy wykonałam przy użyciu oprogramowania GIS – ArcGIS Pro oraz Statistica. Ze względu na różne układy współrzędnych analizowanych danych, bazę danych OSM przekształciłam do układu zgodnego z BDOT10k – PUWG 1992, jednolitego dla całego obszaru Polski. Transformację pomiędzy układem geograficznym WGS84 a PUWG 1992 wykonałam zgodnie z narzędziami konwersyjnymi ArcGIS Pro. Proces przeprowadzania transformacji układu współrzędnych nie wpływa na uzyskane wyniki.

W przeprowadzonych badaniach jakości danych przestrzennych OSM szczegółowo przeanalizowałam sześć podstawowych klas pokrycia terenu:

- trzy warstwy powierzchniowe: budynki, lasy i zbiorniki wodne;
- trzy warstwy liniowe: drogi, linie kolejowe i rzeki.

Badane klasy pokrycia terenu odpowiednio wyselekcjonowałam z analizowanych baz BDOT10k oraz OSM, wskazując odpowiadające tematycznie grupy obiektów – Tabela 3. Obiekty BDOT10k przyjąłam jako dane referencyjne w przeprowadzonej analizie jakości, gdyż są to państwowe dane topograficzne o znanej wyższej dokładności względem danych OSM.

Tabela 3. Analizowane dane OSM oraz BDOT10k [źródło: A1, Tabele 1,2].

Lp.	Klasa pokrycia terenu	Reprezentacja geometryczna	Tag OSM	Kod BDOT10k
1	drogi	linia	highway = {motorway, trunk, primary, secondary, tertiary, unclassified, residential}	SKJZ
2	linie kolejowe	linia	railway = rail	SKTR
3	rzeki	linia	waterway = {river, stream, tidal_channel}	SWRS SWKN
4	budynki	wielokąt	building = *	BUBD
5	lasy	wielokąt	landuse = forest, natural = wood	PTLZ
6	zbiorniki wodne	wielokąt	natural = water landuse = reservoir water = reservoir	PTWP

Za wyborem wymienionych danych topograficznych przemawiała złożoność ich relacji między składnikami środowiska geograficznego, związanymi z morfologią, geologią, hydrologią, roślinnością i klimatem. A także ich strategiczne znaczenie w zarządzaniu sytuacją kryzysową (tj. powodzie, pożary, ataki terrorystyczne), dla której ważna jest identyfikacja obszarów zaludnionych, dróg dojazdowych i obszarów niebezpiecznych oraz ich rola w kształtowaniu zrównoważonego rozwoju zgodnie z Agendą 2030.

Zgodnie z wykonanym przeze mnie przeglądem literatury na potrzeby badań oraz pracą Senaratne et al. (2016), analizującą 56 artykułów dotyczących jakości danych VGI zaobserwowałam, iż jednym z głównych ograniczeń jest fakt, że istniejące miary i wskaźniki (w tym te opisane przez ISO) nie są wystarczająco inkluzywne, aby ocenić dane OSM. Dzieje się tak głównie dlatego, że niejednorodny charakter OSM jest zasadniczo odmienny od tego, czym do tej pory zajmowali się eksperci geoprzestrzeni, co w konsekwencji prowadzi do luki badawczej przy określaniu wskaźników jakości i proponowaniu metod ich obliczania. Ponadto, z wykonanego przeglądu literatury wynika, iż przeprowadzono tylko kilka badań mających na celu zbadanie i przeanalizowanie różnic w wymaganiach jakościowych dla różnych dziedzin zastosowań (Senaratne et al., 2016). Prace badawcze nad wymienionymi ograniczeniami, a tym samym ulepszenie istniejących metod, stanowią ważny wkład naukowy

w użyteczność zasobów VGI. Większość metod wykorzystana do oceny jakości danych OSM odnosi się do dokładności geometrycznej, dokładności tematycznej i spójności topologicznej, mniej metod zajmuje się pozostałymi miarami jakości, takimi jak kompletność. Dodatkowo badania te skupiają się na konkretnych obiektach topograficznych, nie dokonując analizy jakości elementów pokrycia terenu w ujęciu kompleksowym. W związku z powyższym opracowany przeze mnie wskaźnik CCI, dotyczący zgodności danych zawartych w dwóch zbiorach, stanowi istotny wkład w obecny stan badań nad jakością danych przestrzennych. Nowość badania polega na złożonym, uniwersalnym podejściu metodycznym, które pozwala na ocenę danych kategoriowych, tj. danych jakościowych pogrupowanych w kategorie, które odnoszą się do formy informacji przechowywanej i identyfikowanej za pomocą nazw lub etykiet (np. las, rzeka, jezioro, miasto) zgodnie z kryteriami zdefiniowanymi przez użytkownika.

Ogólny schemat badań składał się z co najmniej czterech głównych faz: wstępnej, analitycznej, wizualizacyjnej i decyzyjnej (Rys. 2).



Rys. 2. Schemat przeprowadzonych badań.

Faza wstępna obejmowała studia literaturowe, wstępne przetworzenie danych (np. konwersja do wspólnego układu współrzędnych), opracowanie metody badań i sposobu wizualizacji wyników. Faza druga, analityczna, polegała na obliczeniu wartości zaproponowanych wskaźników, ich charakterystyk statystycznych (miar skupienia, rozproszenia, określenia typu rozkładu statystycznego, sprawdzenie zależności statystycznych, np. korelacji). Kolejna faza wizualizacyjna to przedstawienie wyników analiz w postaci map, tabel i wykresów. Natomiast faza decyzyjna obejmowała ocenę wyników, porównanie ich z wynikami innych badaczy i sformułowanie wniosków końcowych.

4.2. Mierniki jakości danych przestrzennych

W kolejnych podrozdziałach opisałam wskaźniki oceny jakości danych OSM w odniesieniu do bazy referencyjnej BDOT10k (Tabela 3). Wyznaczone metodami wielo cechowej analizy porównawczej wskaźniki oraz ich wizualizacja miały na celu dokonanie kompleksowej oceny porównawczej jakości danych przestrzennych. Ze względu na odmienny typ geometryczny analizowanych obiektów pokrycia terenu, szczególnie omówiłam ocenę jakości danych OSM dla obiektów powierzchniowych oraz osobno dla obiektów liniowych. Wszystkie wskaźniki obliczone były dla podstawowego pola odniesienia – heksagonu o powierzchni 1 km².

4.2.1. Mierniki jakości zbiorów danych przestrzennych

W celu uzyskania informacji statystycznej o dokładności geometrycznej obiektów powierzchniowych OSM w porównaniu z referencyjną bazą danych BDOT10k wykorzystałam punkty homologiczne – Rys. 3.



Rys. 3. Półautomatycznie wykryte punkty homologiczne między budynkami OSM (czerwony) a budynkami BDOT10k (niebieski) [źródło: A1, Rys. 5].

Pomiar punktów homologicznych przeprowadziłam półautomatycznie w oprogramowaniu ArcGIS Pro poprzez porównanie współrzędnych odpowiadających sobie narożników (wierzchołków) obiektów w bazach OSM i BDOT10k. Dokładność geometryczną obiektów powierzchniowych OSM przedstawiłam w postaci pierwiastka błędu średniokwadratowego RMSE (równania 1 – 3).

$$RMSE_X = \sqrt{\frac{\sum_i (X_{OSM} - X_{BDOT10k})^2}{N}} \quad (1)$$

$$RMSE_Y = \sqrt{\frac{\sum_i (Y_{OSM} - Y_{BDOT10k})^2}{N}} \quad (2)$$

$$RMSE = \sqrt{RMSE_X^2 + RMSE_Y^2} \quad (3)$$

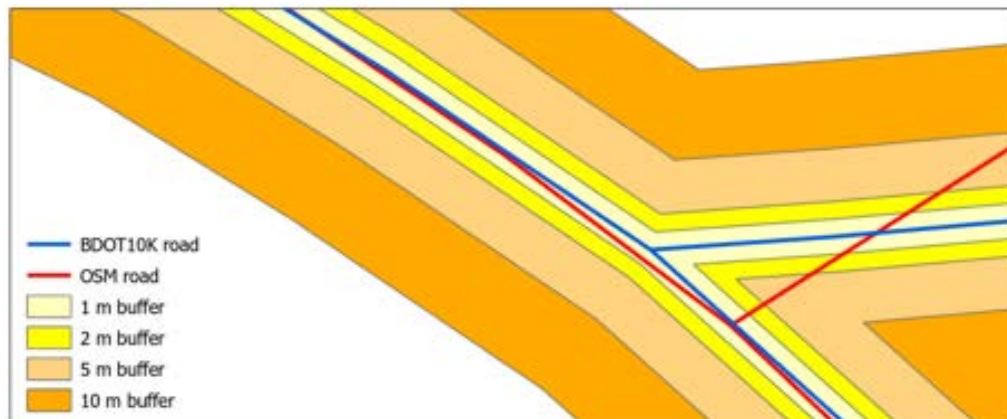
gdzie:

X_{OSM} , Y_{OSM} – współrzędne punktu wierzchołka obiektu bazy danych OSM, odpowiadającemu obiektowi z bazy referencyjnej BDOT10k;

$X_{BDOT10k}$, $Y_{BDOT10k}$ – współrzędne punktu wierzchołka obiektu z bazy referencyjnej BDOT10k;

N – liczba obserwacji (punktów homologicznych).

Do określenia dokładności lokalizacji obiektów liniowych OSM względem bazy BDOT10k wykorzystałam opracowaną w badaniach Goodchild et al. (1997) metodę strefy buforowej. Dla każdej klasy obiektów liniowych (drogi, linie kolejowe i rzeki) wytyczyłam 4 strefy buforowe o szerokościach 1, 2, 5 i 10 m, dobrane optymalnie na podstawie literatury oraz dokładności danych BDOT10k (Rys. 4).



Rys. 4. Utworzone strefy buforowe wokół liniowego obiektu bazy danych BDOT10k [źródło: A1, Rys. 6].

Dokładność lokalizacji obiektu liniowego określiłam, obliczając jaki procent długości poszczególnych odcinków linii danych OSM znajduje się w poszczególnych strefach buforowych (równanie 4).

$$Coverage [\%] = \frac{L_{OSM}}{L_{BDOT10k}} \cdot 100\% \quad (4)$$

gdzie:

L_{OSM} – całkowita długość obiektów OSM testowanych w buforze;

$L_{BDOT10k}$ – całkowita długość obiektów liniowych BDOT10k w buforze.

Jako kolejny wskaźnik jakości danych OSM, obliczyłam kompletność obiektów powierzchniowych OSM, wyrażoną w postaci C Index, za pomocą metody opartej na współczynniku powierzchni, zwanym „area ratio unit” (Tian et al., 2019), który oblicza kompletność jako procentowy stosunek całkowitej powierzchni obiektu OSM do całkowitej powierzchni obiektu bazy danych BDOT10k w obrębie konkretnego pola podstawowego (równanie 5).

$$C\ Index = \frac{\sum A_{OSM}}{\sum A_{BDOT10k}} \cdot 100\% \quad (5)$$

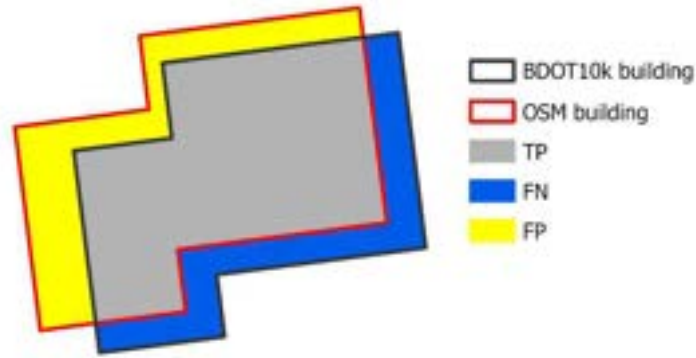
gdzie:

A_{OSM} – powierzchnia obiektu OSM odpowiadająca obiektowi referencyjnej bazy danych BDOT10k w danej komórce siatki heksagonalnej;

$A_{BDOT10k}$ – powierzchnia obiektu w zbiorze referencyjnym BDOT10k w danej komórce siatki heksagonalnej.

Osiągane wartości C Index mogą być większe lub równe 0, gdzie 0 oznacza brak odpowiednich budynków w danych OSM, równe 100% – oba zbiory danych zawierają te same budynki, bądź wyższe niż 100% wskazuje na nadkompletność danych OSM.

Ze względu na nadmiar danych dostępnych w OSM w stosunku do danych BDOT10k, metoda obliczania C Index może wprowadzić zawyżenie kompletności danych – wspomnianą nadkompletność (Brovelli et al., 2016). Z tego powodu obliczyłam trzy dodatkowe wskaźniki: TP Index (True Positive Index), FP Index (False Positive Index) i FN Index (False Negative Index) – Rys. 5.



Rys. 5. Graficzna interpretacja wskaźników TP, FP i FN.

Wskaźnik TP reprezentuje nakładające się obszary obiektów powierzchniowych pomiędzy OSM i BDOT10k, tj. wspólne obszary pomiędzy zbiorami danych (równanie 6). Wskaźnik FP reprezentuje obiekty powierzchniowe OSM, które nie istnieją w zbiorze danych BDOT10k (równanie 7), a wskaźnik FN uwzględnia obiekty BDOT10k, które nie istnieją w zbiorze danych OSM (równanie 8).

$$TP\ Index = \frac{A_{OSM} \cap A_{BDOT10k}}{\sum AREA_{BDOT10k}} \cdot 100\% \quad (6)$$

$$FP\ Index = \frac{A_{OSM} \setminus A_{BDOT10k}}{\sum AREA_{BDOT10k}} \cdot 100\% \quad (7)$$

$$FN\ Index = \frac{A_{BDOT10k} \setminus A_{OSM}}{\sum AREA_{BDOT10k}} \cdot 100\% \quad (8)$$

gdzie:

A_{OSM} – powierzchnia obiektu w zbiorze OSM odpowiadająca obiektowi w bazie danych BDOT10k w danej komórce siatki heksagonalnej;

$A_{BDOT10k}$ – powierzchnia obiektu referencyjnego w zbiorze danych BDOT10k;

$AREA_{BDOT10k}$ – całkowita powierzchnia obiektów w zbiorze BDOT10k w danej komórce siatki heksagonalnej.

Indeks TP przyjmuje wartości od 0 do 100%. Wartość 100% jest osiągnięta przez obiekty OSM z dokładnym pokryciem zbioru danych BDOT10k. Im niższa wartość wskaźnika, tym mniejsze nakładanie się obiektów OSM i BDOT10k. Dla wartości równej 0 nie ma powierzchni wspólnej między OSM a zbiorem danych BDOT10k – obiekty są rozłączne.

Dodatkową autorską miarą oceny jakości danych [wyznaczoną w A2] był wskaźnik informujący o liczbie budynków w analizowanym zbiorze danych OpenStreetMap w postaci numerycznego wskaźnika kompletności COUNT Index. Obliczał udział procentowy jaki stanowi liczba budynków jako obiektów OSM w porównaniu do liczby budynków w danych referencyjnymi BDOT10k. Tak jak w poprzednich badaniach, wskaźniki obliczono w odniesieniu do siatki heksagonalnej o oczku 1 km² (równanie 9).

$$COUNT\ Index = \frac{Count_{OSM}}{Count_{BDOT10k}} \cdot 100\% \quad (9)$$

gdzie:

$Count_{OSM}$ – liczba obiektów (budynków) w zbiorze OSM w danej komórce siatki heksagonalnej;

$Count_{BDOT10k}$ – liczba obiektów (budynków) w zbiorze referencyjnym BDOT10k w danej komórce siatki heksagonalnej.

COUNT Index przyjmuje wartości większe lub równe 0, gdzie 0 oznacza, że nie ma odpowiadających obiektów OSM do danych BDOT10k. Wartość 100% oznacza, że oba zbiory danych zawierają taką samą liczbę obiektów, a wartość większa niż 100% wskazuje na liczbową przewagę obiektów w zbiorze danych OSM nad BDOT10k.

Kompletność dróg, linii kolejowych i rzek dostępnych w bazie OSM została obliczona poprzez porównanie długości danego obiektu liniowego z długością odpowiadającego mu obiektu w zbiorze danych BDOT10k i wyrażona w procentach (równanie 10).

$$Completness\ [\%] = \frac{L_{OSM}}{L_{BDOT10k}} \cdot 100\% \quad (10)$$

gdzie:

L_{OSM} – długość obiektu liniowego w zbiorze danych OSM;

$L_{BDOT10k}$ – długość odpowiadającego obiektu według zbioru danych BDOT10k.

W analizie atrybutów i dokładności semantycznej OSM przedstawiłam ilościowe wyniki dokładności wartości atrybutów bez uwzględnia referencyjnej bazy danych. Analizy ilościowe pokazują, w jakim stopniu wybrany znacznik obiektu OSM jest informowany (zawiera informacje o mapowanym obiekcie) względem całej grupy obiektów danego pokrycia terenu w badanym powiecie. Analizie poddałam liczbę

obiektów, które posiadają uzupełnione tagi dodatkowe do tagu głównego, takie jak nazwa (NAME), a w przypadku budynków także typ (TYPE) – zgodnie z równaniem 11.

$$\text{Attribute accuracy} [\%] = \frac{F_{OSM}}{T_{OSM}} \cdot 100\% \quad (11)$$

gdzie:

F_{OSM} – liczba obiektów z uzupełnionym znacznikiem (tagiem) w zbiorze danych OSM;

T_{OSM} – całkowita liczba obiektów OSM.

4.2.2. Współczynnik CCI skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych

W celu oceny zgodności zbiorów danych opracowałam autorską złożoną miarę, która umożliwiła porównawczą ocenę kompletności dwóch zestawów danych przestrzennych przy użyciu liniowego rankingu autorytatywnego – Compound Correspondence Index (CCI) i statystycznych miar rozproszenia w celu wizualizacji przestrzennej [przedstawione w **A3** oraz **A4**]. W przeciwieństwie do dotychczas przeprowadzonych badań, opisanych w podrozdziale 4.2.1. [**A1**, **A2**], opracowana kompleksowa metodyka opiera się na kompensacyjnej analizie porównawczej z wykorzystaniem metody TOPSIS, wcześniej niestosowanej w ocenie danych kategorycznych.

Podstawowy problem badawczy, do którego odnosi się niniejszy podrozdział, dotyczy definicji Compound Correspondence Index (CCI) w skali lokalnej i regionalnej, zastosowanych wag oraz określenia liczby i optymalnych zakresów klas, które jednoznacznie wskazują przestrzenne położenie różnic w stopniu zgodności dwóch badanych zestawów danych. Metodę porównawczej analizy wielokryterialnej WLC zastosowałam na podstawie takich kryteriów, jak różnice w powierzchni zajmowanej przez budynki, lasy i zbiorniki wodne oraz różnice długości dróg, linii kolejowych i rzek. Minimalna różnica wskazuje na bardzo podobną objętość informacji, podczas gdy maksymalna oznacza duże różnice między dwoma zestawami danych przestrzennych. W celu wyznaczenia CCI zastosowałam standardową metodę TOPSIS, szczegółowo opisaną w wielu pracach, np. Zaltko Pavic i Novoselac (2013) oraz Zavadskas et al. (2015), przedstawioną w sposób ogólny poniżej:

- 1) Wybór zestawów danych BDOT10k lub OSM (dwie alternatywy; $m = 2$) na podstawie minimalnej wartości kryterium k ($k = 6$), jak pokazano w równaniu 12.

$$\begin{aligned} x_1 &= |BT_B - OSM_B|, x_2 = |BT_F - OSM_F|, x_3 = |BT_W - OSM_W|, \\ x_4 &= |BT_{Ro} - OSM_{Ro}|, x_5 = |BT_S - OSM_S|, x_6 = |BT_{Ra} - OSM_{Ra}|, \end{aligned} \quad (12)$$

gdzie:

BT – zbiór danych BDOT10k;

OSM – dane OpenStreetMap;

indeksy dolne B, F, W – całkowity obszar zajmowany przez odpowiednio: budynki, lasy i zbiorniki wodne w komórce siatki heksagonalnej;

indeksy dolne Ro, S, Ra – całkowita długość zajmowana przez odpowiednio: drogi, ciekły wodni i linie kolejowe w komórce siatki heksagonalnej.

2) Normalizacja zmiennych metodą przekształcenia ilorazowego (równanie 13).

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}}, \quad i = 1 \dots m; j = 1 \dots n. \quad (13)$$

gdzie:

x_{ij} – obserwacja j -tej zmiennej dla obiektu i ;

m – liczba alternatyw;

n – liczba komórek siatki heksagonalnej.

3) Obliczenie ważonej znormalizowanej macierzy decyzyjnej (równanie 14).

$$r_{ij} = w_j \times n_{ij}, \quad i = 1 \dots m; j = 1 \dots n. \quad (14)$$

w_j – wartość wagi przypisana j -tej zmiennej.

4) Określenie współrzędnych wzorca (PIS) i antywzorca (NIS) (odpowiednio równania 15 i 16).

$$PIS = \{r_1^-, \dots, r_j^-, \dots, r_n^-\}, \quad \text{gdzie } v_j^+ = \{\max(r_j^i) \text{ if } j \in B; \min(r_j^i) \text{ if } j \in J'\} \quad (15)$$

$$NIS = \{r_1^-, \dots, r_j^-, \dots, r_n^-\}, \quad \text{gdzie } v_j^+ = \{\min(r_j^i) \text{ if } j \in B; \max(r_j^i) \text{ if } j \in J'\} \quad (16)$$

Ponieważ wszystkie użyte kryteria są destymulantami, dodatnie rozwiązanie idealne obliczono jako $r_j^i \min$, podczas gdy ideał negatywny to $r_j^i \max$.

5) Wyznaczenie odległości obiektów od wzorca i antywzorca (równanie 17).

$$S_i^+ = \sqrt{\sum_i^n (r_{ij} - r_j^+)^2}; S_i^- = \sqrt{\sum_i^n (r_{ij} - r_j^-)^2} \quad i = 1, \dots, m \quad (17)$$

6) Obliczenie wartości zmiennej agregatowej (równanie 18).

$$q_i = \frac{S_i^-}{S_i^+ + S_i^-}, \quad \text{gdzie } q_i \in [0; 1] \quad (18)$$

gdzie:

$\max_i\{q_i\}$ – najlepszy obiekt w odniesieniu do kryteriów oceny;

$\min_i\{q_i\}$ – najgorszy obiekt w odniesieniu do kryteriów oceny.

Współczynnik CCI obliczyłam w dwóch podejściach: lokalnym (dalej zwanym CCI_L) oraz regionalnym (dalej zwanym CCI_R). Takie założenie stanowi innowacyjne rozwiązanie, ponieważ pozwala na porównanie dwóch zestawów zawierających dane jakościowe dla obszarów geograficznie rozłącznych, umożliwiając użytkownikowi świadomy i odpowiedzialny wybór jednego z nich. Pokazuje również różnice między klasyfikacjami CCI na poziomie lokalnym i regionalnym, wskazując na wrażliwość wskaźnika CCI [A3]. W przypadku CCI_L wartości wskaźnika zostały obliczone dla każdego z siedmiu analizowanych powiatów osobno.

W odniesieniu do poziomu regionalnego, CCI_R obliczyłam sekwencyjnie, rozszerzając obszar testowy o kolejne powiaty:

- I. W pierwszej sekwencji analizowałam cztery powiaty: piaseczyński, sokólski, sanocki oraz słupecki (obliczając CCI_{R4});
- II. W drugiej sekwencji analizowałam pięć powiatów, dodając do poprzednich czterech powiat ostrowski (obliczając CCI_{R5});
- III. W trzeciej sekwencji analizowałam sześć powiatów, dodając do poprzednich pięciu powiat otwocki (obliczając CCI_{R6});
- IV. W czwartej sekwencji analizowałam siedem powiatów, dodając do poprzednich sześciu powiat międzyrzecki (obliczając CCI_{R7}).

Zgodnie z metodą TOPSIS dla wyznaczanego wskaźnika CCI należało przyjąć wartości wag przypisanych j-tej zmiennej, czyli w tym przypadku badanym klasom pokrycia terenu. Zagadnienie to analizowałam w dwóch wariantach. Pierwszy wariant wagowania (oznaczony jako CCI_{W1}) zakładał, iż przyjęte w metodzie TOPSIS kryterium

wagowania obiektów stanowiła ich rozpoznawalność na zobrazowaniach satelitarnych i zdjęciach lotniczych, z których zostały uzyskane, tj. zdjęcia lotnicze i zobrazowania satelitarne SPOT 5 w strefie przygranicznej Unii Europejskiej. W związku z tym budynkom i lasom nadałam najwyższą wagę 0,25, drogom i liniom kolejowym 0,15, z kolei zbiornikom i ciekom wodnym najniższą 0,10. Wspomniane reguły ważenia wykorzystywane są także w analizach przejezdności terenu oraz zarządzaniu sytuacjami kryzysowymi (Pokonieczny, 2018). Pierwszy wariant wagowania wskaźnika CCI opisałam w **A3**. W drugiej kombinacji wag (CCI_{W2}) przyjąłam, iż wszystkie kryteria są równie ważne, dlatego każdemu z analizowanych obiektów pokrycia terenu przypisałam równą wagę wynoszącą 0,167. Drugi wariant wagowania wskaźnika CCI opisałam w **A4**. Aby wyrazić względną zmianę procentową między dwoma wariantami wag dla lokalnych CCI, zdefiniowałam i wyznaczyłam miarę Relative Change (RC) zgodnie z równaniem 19.

$$RC_{CCI} = \frac{CCI_{W2} - CCI_{W1}}{CCI_{W2}} \cdot 100\% \quad (19)$$

gdzie:

RC_{CCI} – wskaźnik Relative Changes obliczony dla lokalnego CCI;

CCI_{W1} – wartość CCI przy użyciu różnych wag (pierwsza kombinacja);

CCI_{W2} – wartość CCI przy użyciu równych wag (druga kombinacja).

4.2.3. Autokorelacja przestrzenna

Autokorelacja przestrzenna oznacza, iż wartości obiektów bliskich geograficznie są bardziej podobne do siebie niż tych odległych. Zjawisko to powoduje tworzenie się klastrów przestrzennych o podobnych wartościach. Narzędzie Spatial Autocorrelation Global Moran's I (wykorzystane w **A3**) mierzy autokorelację przestrzenną na podstawie lokalizacji komórek siatki heksagonalnej i wartości współczynnika CCI jednocześnie. Biorąc pod uwagę zbiór cech i powiązany atrybut (wartość CCI w danej komórce heksagonalnej) ocenia czy układ jednostek przestrzennych jest skupiony, rozproszony czy losowy. Narzędzie oblicza wartość indeksu Moran's I zarówno wynik „z”, jak i wartość „p”, aby ocenić istotność tego indeksu. Wartości „p” to numeryczne przybliżenia obszaru pod krzywą dla znanego rozkładu, ograniczonego statystyką testową (Getis & Ord, 1992). Matematyczną formułę autokorelacji Global Moran's I przedstawiłam w równaniu 20.

$$I = \frac{n}{s_0} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n c_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})} \quad (20)$$

gdzie:

n – liczba jednostek przestrzennych;

c_{ij} – elementy macierzy sąsiedztwa C ;

s_0 – suma elementów macierzy C ;

x_i – wartość obserwacji dla i -tej jednostki.

Narzędzie autokorelacji Global Moran's I jest statystyką inferencyjną, co oznacza, że wyniki analizy są zawsze interpretowane w kontekście jej hipotezy zerowej. Gdy wartość „p” zwrócona przez to narzędzie jest statystycznie istotna, można odrzucić hipotezę zerową. Podsumowanie interpretacji wyników zamieściłam w Tabeli 4.

Tabela 4. Interpretacja wyników autokorelacji Morana I (Goodchild et al., 1986).

Statystyka	Opis
Wartość „p” nie jest istotna statystycznie	Nie można odrzucić hipotezy zerowej. Najprawdopodobniej rozkład przestrzenny wartości cech jest wynikiem losowych procesów przestrzennych
Wartość „p” jest istotna statystycznie, wynik „z” jest dodatni	Można odrzucić hipotezę zerową. Przestrzenny rozkład wysokich i/lub niskich wartości w zestawie danych jest przestrzennie skupiony
Wartość „p” jest istotna statystycznie, wynik „z” jest ujemny	Można odrzucić hipotezę zerową. Przestrzenny rozkład wysokich i/lub niskich wartości w zestawie danych jest bardziej rozproszony przestrzennie

Globalne statystyki, takie jak narzędzie autokorelacji przestrzennej Global Moran's I, oceniają ogólny wzorec przestrzenny i trend danych. Są najskuteczniejsze, gdy wzorec przestrzenny jest spójny w całym badanym obszarze. Lokalne statystyki (takie jak narzędzie HotSpot Analysis: Getis–Ord G_i^* dostępne w oprogramowaniu ArcGIS Pro) oceniają każdą cechę w kontekście sąsiednich cech i porównują sytuację lokalną z sytuacją globalną.

Analiza skupień punktów (hotspot) została wykorzystana do wskazania relacji przestrzennych i zidentyfikowania przestrzennych skupień wartości RC wskaźnika CCI dla zastosowanych wariantów wag w metodzie TOPSIS. Uzyskane wartości wskazywały, gdzie obiekty o wysokich lub niskich wartościach były zgrupowane przestrzennie

(Getis & Ord, 1992). Gorący punkt można opisać jako obszar o większej koncentracji zdarzeń w porównaniu do oczekiwanej liczby po uwzględnieniu losowego rozkładu. Cecha o wysokiej wartości jest interesująca, ale może nie być statystycznie istotnym punktem zapalnym. Aby obiekt był statystycznie istotnym punktem aktywnym, musi mieć wysoką wartość i być otoczony przez inne obiekty również o wysokich wartościach. Lokalna suma obiektu i jego sąsiadów jest porównywana proporcjonalnie z sumą wszystkich obiektów. Gdy suma lokalna różni się od oczekiwanej sumy lokalnej i gdy różnica jest zbyt duża, aby wynikać z losowego przypadku, uzyskuje się statystycznie istotny wynik „z”, zgodnie z równaniami 21 – 23.

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2}{n-1}}} \quad (21)$$

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \quad (22)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2} \quad (23)$$

gdzie:

x_j – wartość względnej zmiany RC;

$w_{i,j}$ – waga przestrzenna między cechami CCI dla i -tego obiektu j -tej zmiennej;

n – liczba obserwacji.

Statystyka Getis–Ord G_i^* wykorzystywana w analizie skupień, zapewnia wynik „z”, wartość „p” i przedział ufności z interpretacją zgodnie z Tabelą 5.

Tabela 5. Charakterystyka parametrów analizy skupień hotspot (Ord et al., 2010).

Statystyka	Opis	Implementacja
Wartość „p” jest mała, wynik „z” jest dodatni	Klastrowanie typu wysoki–wysoki (im większy wynik „z”, tym większy stopień skupienia)	$CCI_{W1} < CCI_{W2}$ $RC_{CCI} > 0$
Wynik „z” jest bliski 0	Brak wyraźnych klastrów przestrzennych	–
Wartość „p” jest mała, wynik „z” jest ujemny	Klastrowanie typu niski–niski (im mniejszy wynik „z”, tym większy stopień skupienia)	$CCI_{W1} > CCI_{W2}$ $RC_{CCI} < 0$

4.2.4. Wnioskowanie statystyczne

W celu identyfikacji prawidłowości obliczonych mierników jakości danych, ułatwiających ich analizę, wykorzystałam elementy statystyki opisowej, czyli miary położenia (średnia arytmetyczna, mediana, mediana odchylenia bezwzględnego) i rozrzutu (rozstęp, miara zmienności, rozstęp ćwiartkowy, odchylenie standardowe). Wymienione miary statystyczne zastosowałam w artykułach **A2**, **A3** oraz **A4**. Dodatkowo wykorzystując miary statystyki opisowej zdefiniowałam przedziały klas odzwierciedlające zgodność danych OSM względem BDOT10k.

Do oceny czy analizowane zmienne (wyznaczone mierniki jakości) posiadają rozkład normalny wykorzystałam test Shapiro–Wilka [A2]. Normalny wykres prawdopodobieństwa identyfikuje istotne odstępstwa od normalności (Hanusz et al., 2016). Hipoteza zerowa dla tego testu zakłada, że analizowana próba badawcza pochodzi z populacji o rozkładzie normalnym. Jeśli test Shapiro–Wilka osiąga istotność statystyczną ($\alpha \leq p < 0,05$), wskazuje to na rozkład odbiegający od krzywej Gaussa. Wynik testu Shapiro–Wilka warunkował, które z miar statystyki opisowej posłużyły do wyznaczenia metody klasyfikacji.

Celem porównania wartości zmian procentowych wskaźnika RC między przyjętymi wariantami wag CCI dla badanych powiatów, zdefiniowałam cztery przedziały klasowe [A4]. Zostały one utworzone z zakresami wartości, które reprezentowały proporcje odchylenia standardowego. Ujemne wartości RC zostały przeanalizowane w dwóch klasach: dla nich każdy zakres został zdefiniowany zgodnie z przedziałem połowy odchylenia standardowego ($0,5 \sigma$), który został obliczony jako średnia wartość dla analizowanych powiatów. Dodatkowo wartości RC również podzielono na dwie klasy według wartości jednego odchylenia standardowego (σ) jako przedział zakresów.

W celu zbadania związku pomiędzy opracowanymi wskaźnikami CCI oraz RC, a klasami pokrycia terenu wykorzystałam analizę statystyczną w postaci korelacji Pearsona [A4]. Została ona użyta do zapewnienia ogólnego przeglądu wyników na poziomie powiatu oraz określenia, czy istnieje związek liniowy między dwoma zmiennymi – jeśli tak, jaka jest jego siła oraz jaki ma on charakter.

5. OPIS WYNIKÓW

W niniejszych podrozdziałach opisałam szczegółowo uzyskane wyniki badań wchodzące w skład cyklu publikacyjnego, stanowiące niniejszą rozprawę. Artykuły te dołączono w formie załączników.

5.1. Analiza jakości danych OSM zgodnie z miernikami ISO [publikacje A1 oraz A2]

Jak już zostało wspomniane, w przypadku danych zbieranych nieodpłatnie i dobrowolnie przez bardzo dużą liczbę wolontariuszy, zwanych dobrowolnymi informacjami geograficznymi (VGI), do których należą dane OSM, istnieją poważne wątpliwości co do ich jakości, przez co ich wykorzystanie staje się problematyczne. Wynika to z braku szczegółowych specyfikacji technicznych, podających jedynie zasady i wytyczne dotyczące udostępniania danych oraz częstego braku formalnej weryfikacji wszystkich danych wprowadzanych do bazy. Aby móc udostępniać i wykorzystywać zbiór danych w różnych aplikacjach, należy po pierwsze poznać jego jakość, a po drugie opisać go w sposób standardowy i zrozumiały (Goodchild & Li, 2012).

Mając na uwadze powyższe fakty w publikacji A1 oraz częściowo w A2, przedstawiłam kompleksową ocenę jakości danych OSM, wykorzystując mierniki jakości opisane w normie ISO (2023) oraz dodatkowe wskaźniki geometryczne, zwracając także uwagę na aspekt niedoskonałych ustaleń semantycznych i założeń jakościowych. Badane elementy jakości wraz z wyznaczonymi w tym celu miernikami przedstawiono w Tabeli 6 (ich szczegółowy opis zamieściłam w podrozdziale 4.2.1).

Tabela 6. Obliczone wskaźniki jakości danych przestrzennych.

Element jakości	Obliczony wskaźnik jakości	Klasa pokrycia terenu	Numer artykułu
Dokładność geometryczna	1) RMSE, RMSE _X , RMSE _Y	budynki, lasy, wody powierzchniowe	A1
	2) Coverage	drogi, linie kolejowe, rzeki	
Kompletność	3) C Index	budynki, lasy, wody powierzchniowe	A1, A2
	4) COUNT Index		
	5) TP Index		
	6) FN Index		A1
	7) FP Index		

	8) Completeness	drogi, linie kolejowe, rzeki	
Dokładność tematyczna	9) Attribute accuracy	budynki, lasy, wody powierzchniowe, drogi, linie kolejowe, rzeki	A1

Postawiony problem badawczy polegał przede wszystkim na określeniu kompletności, dokładności lokalizacyjnej i atrybutowej głównych klas pokrycia terenu obiektów OSM w stosunku do krajowych danych urzędowych zgromadzonych w bazie obiektów topograficznych – BDOT10k, która stanowiła bazę odniesienia.

Charakterystyka danych przestrzennych, podsumowana w publikacji **A1**, ukazuje podstawowe źródła rozbieżności pomiędzy analizowanymi zbiorami danych, którymi są: odmienny model koncepcyjny, zasady pomiarów i nadzór technologiczny oraz system kontroli danych przesyłanych i przechowywanych w bazach OSM i BDOT10k. Biorąc pod uwagę wspomniane różnice, prace badawcze obejmowały identyfikację obiektów odpowiadających sobie w obu zbiorach oraz analizę dokładności geometrycznej i atrybutowej oraz kompletności obiektów. Uzyskane wyniki kompletności danych OSM wyznaczonych dla obiektów powierzchniowych zwizualizowałam w postaci map tematycznych za pomocą kartogramu prostego (Leonowicz, 2002). Obliczone wskaźniki C, FP, FN oraz TP Index odniosłam do pola przyjętej siatki heksagonalnej o powierzchni 1 km² (zgodnie z opisem w podrozdziale 4.1.). Ze względu na dużą liczbę map, opracowane mapy tematyczne przedstawiłam dla dwóch analizowanych powiatów, które reprezentują największe zróżnicowanie wyników – powiaty piaseczyński oraz sokólski (mapy zamieszono w publikacji **A1**, Rysunki 9 oraz 11, stanowiącej załącznik nr 1).

Analiza dokładności geometrycznej obiektów powierzchniowych wykazała, że najwyższą dokładnością lokalizacji charakteryzowały się budynki – średni błąd RMSE w tej grupie wyniósł 1,92 m. Z kolei najniższe wyniki dokładności lokalizacji obiektów powierzchniowych OSM osiągnięto dla lasów – średni błąd RMSE wyniósł 5,65 m. Pełne wyniki otrzymanej analizy dokładności geometrycznej obiektów powierzchniowych znajdują się w **A1** w Tabeli 3 (załącznik nr 1).

Analizując dokładność geometryczną obiektów liniowych zauważono, że najwyższe dokładności uzyskano dla takich obiektów jak drogi i linie kolejowe. Wraz ze wzrostem szerokości strefy buforowej od 1 do 10 m znacząco wzrastał udział obiektów OSM w danym buforze, gdzie dla szerokości 2 m w większości powiatów znajdowało się

około 50% badanych obiektów sieci transportowych. Najwyższą dokładnością i największym udziałem obiektów charakteryzował się powiat piaseczyński, w którym w OSM zaobserwowano nadkompletność kolei (114%) i kompletność 92% dla dróg w buforze o zasięgu do 5 m. Natomiast najniższą dokładność (83,4% dla kolei i 50% dla dróg w buforze o zasięgu do 10 m) odnotowano w powiecie słupskim o strukturze rolno-leśnej, gdzie sieć komunikacyjna jest słabo rozwinięta, a poziom urbanizacji jest najniższy spośród badanych powiatów i wynosi 20,7%. Dodatkowo najniższe wyniki dokładności geometrycznej uzyskano dla sieci rzecznej: od 6% dla bufora 1 m (powiat ostrowski) do maksymalnie 67% dla bufora 10 m (powiat sokólski). Pełne wyniki otrzymanej analizy dokładności geometrycznej obiektów liniowych znajdują się w **A1** w Tabelach 4 – 8 (załącznik nr 1).

W przypadku analizy wskaźników kompletności danych przestrzennych, uzyskane wyniki były dość zróżnicowane i zależały od rodzaju obiektu i struktury analizowanego powiatu, w tym użytkowania gruntów – wyniki wykonanej analizy znajdują się w **A1** w Tabelach 9 – 10 oraz Rysunkach 9 i 11 (załącznik nr 1). Najwyższe wartości kompletności budynków OSM (w tym nadkompletność, gdzie liczba budynków OSM znacznie przekraczała liczbę budynków BDOT10k) uzyskano na terenach zurbanizowanych badanych powiatów (miasta i obszary gęsto zabudowane). Najniższe wartości kompletności budynków odnotowano na przedmieściach miast i na terenach rolnych powiatów. Obliczony wskaźnik TP, pokazujący stopień nakładania się obiektów OSM i BDOT10k, osiągał najwyższe wartości bliskie 100% dla obszarów, gdzie stopień kompletności C Index wynosił od 100% do 150%. W przypadku znacznej nadkompletności obiektów (C Index > 150%) wskaźnik TP przyjmował niższe wartości (poniżej 70%). Wskaźnik FP, informujący o obiektach powierzchniowych OSM, które nie występują w zbiorze danych BDOT10k, osiągał najwyższe wartości dla obszarów silnie zurbanizowanych, co wynikało bezpośrednio z wysokiej nadkompletności danych OSM. Najwyższe wartości wskaźnika FN, wskazującego na obiekty BDOT10k, które nie istnieją w zbiorze danych OSM, osiągnięto dla obszarów, dla których stopień kompletności danych był najniższy. Analiza stopnia kompletności obiektów liniowych OSM w odniesieniu do bazy referencyjnej BDOT10k dla poszczególnych powiatów wykazała, że sieć transportowa w większości badanych powiatów osiągnęła najwyższe wyniki, w tym nadkompletność dla powiatów o wysokim stopniu urbanizacji i rozwiniętej sieci transportowej (powiaty piaseczyński i ostrowski). W przypadku sieci rzecznej najniższy wskaźnik kompletności (do 75,7%) odnotowano w powiecie sokólskim.

Najwyższe wartości dokładności tematycznej atrybutów badanych obiektów OSM uzyskano dla dróg, rzek i budynków w powiatach rozwiniętych, które cieszyły się również popularnością wśród użytkowników: piaseczyńskim, ostrowskim i nadmorskim słupskim. Mimo to wskaźniki te były stosunkowo niskie na tle pozostałych mierników jakości – dokładność atrybutowa dla rzek wyniosła maksymalnie 58% a dla budynków było to 39,1%. Najniższe wskaźniki (głównie nieprzekraczające 5%) uzyskano dla linii kolejowych, lasów i wód powierzchniowych, Wyjątek w tym zestawieniu stanowi powiat piaseczyński, dla którego linie kolejowe osiągnęły najwyższą wartość 15% dokładności atrybutowej. Wyniki ilościowe pokazują, że główny tag analizowanych obiektów OSM jest w większości uzupełniony, podczas gdy atrybuty drugorzędne są uzupełniane rzadko. Uzyskane wyniki wskazują na potrzebę uzupełnienia informacji o większości obiektów w bazie danych OSM – zgodnie z przeprowadzonymi analizami nie ma informacji o nazwie i typie większości obiektów OSM, jak np. typ budynku, numer drogi, typ kolei, nazwa rzeki. Brak wartości informacyjnej dotyczącej budynków i dróg może być poważną przeszkodą w korzystaniu z bazy danych OSM w wielu analizach przestrzennych. Pełne wyniki otrzymanej analizy dokładności tematycznej obiektów OSM znajdują się w **A1** w Tabeli 11 (załącznik nr 1).

Wyznaczony wskaźnik kompletności COUNT Index [opisany w **A2**], informujący o stosunku liczby budynków OSM do liczby budynków BDOT10k obliczony został dla powiatu piaseczyńskiego, zgodnie z założeniami publikacji **A2**. Wartości COUNT Index powyżej 100%, świadczące o przewadze liczebnej budynków OSM, występują na terenach zabudowanych, stanowiąc nieco ponad połowę badanego obszaru. Im gęstsza zabudowa tym wartość wskaźnika wzrasta do maksymalnie 300%. Wartości COUNT Index poniżej 100%, wskazujące na liczebną przewagę budynków BDOT10k, znajdują się na obrzeżach miast oraz obszarach o niskiej zabudowie, jak tereny rolne lub rolno-leśne.

Dodatkowym krokiem w ocenie jakości danych było wizualne porównanie wzajemnego położenia obiektów OSM i BDOT10k na tle aktualnej ortofotomapy (2020 r.), dostępnej w serwisie Geoportal Krajowy. W tym celu wybrano losowo w każdym powiecie 20 obiektów takich jak budynki, lasy, wody powierzchniowe, rzeki i linie kolejowe. W wyniku analizy stwierdziłam, że największe różnice występowały dla budynków, co wiązało się z wzajemnym przesunięciem obrysu budowli w stosunku do analizowanych baz danych, oraz lasów – zasięg obrysów był różnie interpretowany

zarówno w bazie BDOT10k jak i OSM. Opis oraz wyniki przeprowadzonej oceny znajdują się w **A1**, Rys. 12 (załącznik nr 1).

Przedstawione wyniki analiz przyniosły odpowiedź na pierwsze pytanie szczegółowe P1 i potwierdziły hipotezę H1, iż kompleksowa ocena jakości danych OSM w porównaniu z oficjalną bazą danych referencyjnych BDOT10k zgodnie ze wskaźnikami ISO oraz wzbogacona o dodatkowe mierniki geometryczne, badające kompletność danych, takie jak TP, FP, FN oraz COUNT Index, jednoznacznie wskazują na znacznie szerszy kontekst interpretacyjny, pozwalający na przedstawienie zróżnicowania jakości wewnątrz badanych zbiorów. Holistyczna ocena jakości niejednorodnych danych przestrzennych OSM, wykonana za pomocą zdefiniowanych mierników w odniesieniu do głównych klas pokrycia terenu pięciu reprezentatywnych powiatów na terenie Polski, stanowi istotny wkład w określeniu różnic danych jakościowych potrzebnych użytkownikowi dla różnych dziedzin zastosowań.

5.2. Kartograficzna wizualizacja wyników oceny kompletności danych przestrzennych dla wybranego przykładu [publikacja A2]

Przedstawione w publikacji **A1** wyniki oceny jakości danych OSM umożliwiły porównanie wartości informacyjnej zawartej w dwóch badanych zbiorach danych przestrzennych. Celem niniejszego artykułu było dostarczenie użytkownikowi informacji o liczbie budynków w analizowanym zbiorze danych OpenStreetMap (OSM) w postaci wskaźników kompletności danych. Według Barron et al. (2014) jakość danych i ich możliwość zastosowania do określonego celu są powszechnie rozumiane jako ściśle ze sobą powiązane. Należy jednak pamiętać, że różne interpretacje tych samych danych mogą prowadzić do różnych informacji, toteż jakość danych OSM w dużej mierze zależy od celu, w jakim dane są wdrażane, tzw. "przydatności do celu" (fitness for purpose).

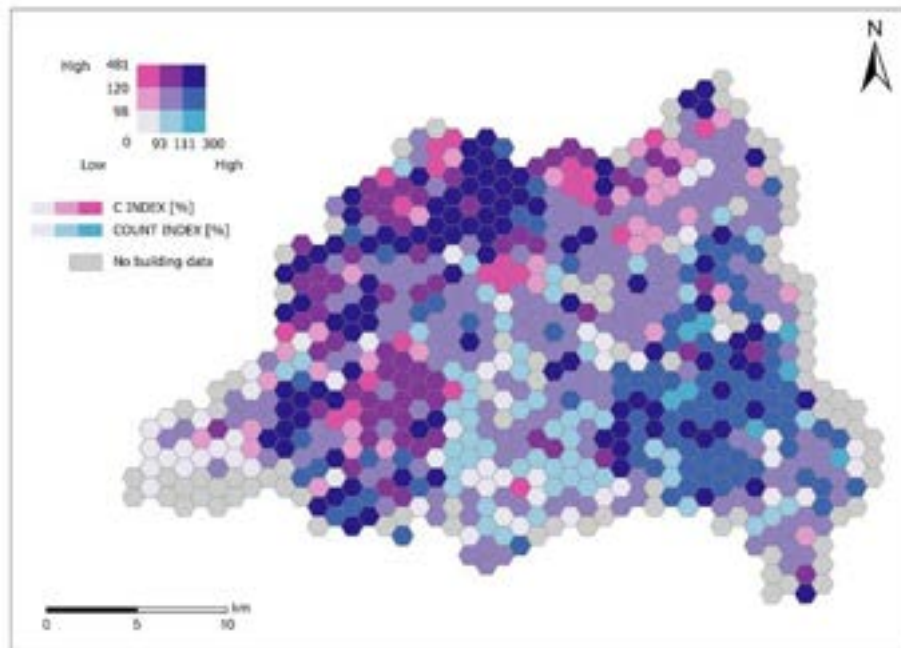
W artykule **A2** główny problem badawczy dotyczył kartograficznej prezentacji kompletności danych przestrzennych, jako elementu jakości danych, umożliwiającej użytkownikowi ocenę ich przydatności do celu (fitness for purpose), a w szczególności wybór, który z dwóch zbiorów danych przestrzennych lepiej odpowiada jego potrzebom. Przyjętym przeze mnie podejściem metodycznym było modelowanie kartograficzne, obejmujące wszystkie etapy pracy badawczej, począwszy od akwizycji, wstępnego

przetwarzania i transformacji, do analizy i ostatecznie wizualizacji danych (Baranowski et al., 2016).

W przeprowadzonych analizach wykorzystałam obliczone w publikacji **A1** wskaźniki kompletności danych, a mianowicie standardowy wskaźnik kompletności powierzchni budynków OSM – C Index oraz wskaźnik lokalizacji budynków OSM – TP Index. Dodatkowym wskaźnikiem wyznaczonym w publikacji **A2** był wskaźnik informujący o liczbie budynków w analizowanym zbiorze danych OpenStreetMap w postaci numerycznego wskaźnika kompletności – COUNT Index. Analizy wykonałam dla powiatu piaseczyńskiego, zróżnicowanego ze względu na strukturę pokrycia terenu i stopień urbanizacji.

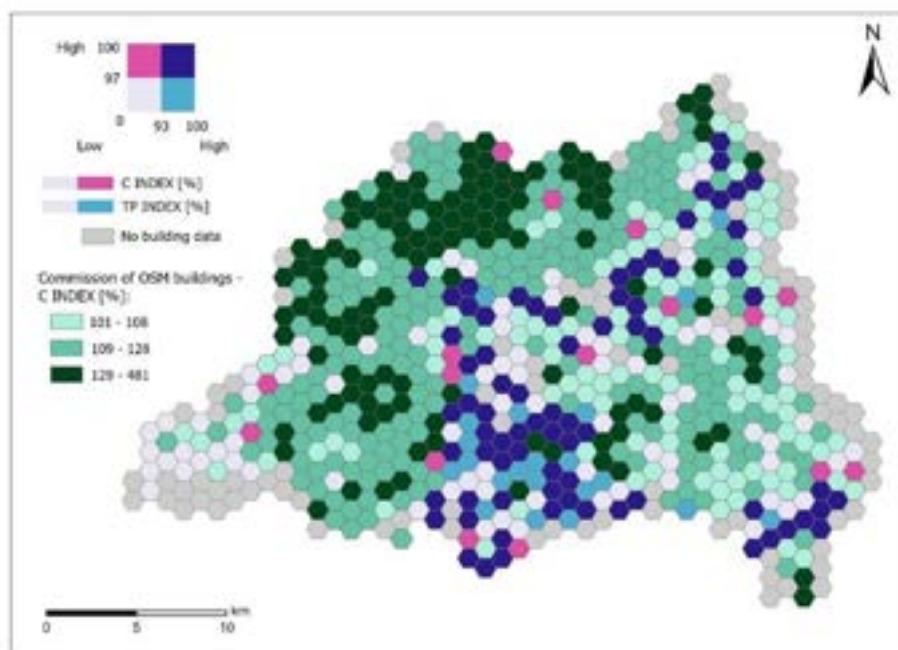
Postawiona przeze mnie hipoteza zakładała, iż holistyczne podejście oparte na matematycznie zdefiniowanych wskaźnikach jakości danych przestrzennych, ich analiza statystyczna z oryginalną prezentacją kartograficzną pozwala na wybór jednego ze zbiorów danych przestrzennych w zależności od potrzeb użytkownika.

W celu wizualizacji kompletności danych przestrzennych OSM w powiecie piaseczyńskim opracowałam dwa kartogramy złożone (Rys. 6 i 7), wykorzystujące trzy wskaźniki kompletności: TP Index, C Index oraz COUNT Index. Wartości zakresów klas wyznaczyłam na podstawie rozkładu statystycznego wartości wskaźników (test Szapiro-Wilka). Z uwagi na to, iż nie przyjmują one rozkładu normalnego, do skonfigurowania zakresu klas użyłam wartości związanych z medianą i medianą odchylenia bezwzględnego, znajdujących się w Tabelach 1–2 publikacji **A2** (załącznik nr 2).



Rys. 6. Kartogram złożony prezentujący kompletność powierzchniową (C Index) oraz liczbową (COUNT Index) danych OSM dla powiatu piaseczyńskiego [źródło: A2, Rys. 3].

Opracowany kartogram złożony (Rys. 6) miał na celu zilustrowanie zależności oraz rozkładu przestrzennego między kompletnością powierzchniową (C Index) a kompletnością liczbową (COUNT Index) budynków w badanych zbiorach. Opracowany kartogram o wymiarze klas 3x3, zapewnia ogólny i łatwy do zrozumienia przegląd kompletności budynków OSM w porównaniu z danymi BDOT10k w powiecie piaseczyńskim. Jednocześnie wskazuje on na obszary, gdzie jednocześnie oba wskaźniki kompletności są mniejsze niż 100% (występuje niedomiar kompletności obiektów OSM), są bliskie lub nieco większe niż 100% oraz znacząco przekraczają 100% – występuje tu znaczna nadkompletność danych OSM. Łatwe do identyfikacji są również obszary gdzie jeden z badanych wskaźników znacząco przewyższa drugi.



Rys. 7. Kartogram złożony prezentujący kompletność powierzchni (C Index) oraz wskaźnik lokalizacji (TP Index) danych OSM dla powiatu piaseczyńskiego [źródło: A2, Rys. 4].

Drugi kartogram złożony w połączeniu z kartogramem prostym (Rys. 7) prezentuje zależności oraz rozkład przestrzenny między kompletnością powierzchni (C Index) a wskaźnikiem lokalizacji (TP Index). Wartości indeksów do 100% zostały przedstawione za pomocą kartogramu złożonego o rozmiarze 2x2. Kolor granatowy oznacza pełną zgodność budynków w obu zbiorach, kolor bardzo jasno szary duże różnice w kompletności. Nadkompletność budynków w OSM (wartości powyżej 100%), opisuje te budynki ze zbioru OSM, które nie mają odpowiedników w zbiorze BDOT10k. Zostało to zilustrowane za pomocą kartogramu prostego, im ciemniejszy kolor tym większa różnica w liczbie budynków.

Uzyskane wyniki są zgodne z innymi podobnymi badaniami i potwierdzają, iż kompletność cech budynków jest stosunkowo wysoka w centrach miast, podczas gdy jej wartość gwałtownie spada w miarę oddalania się od terenów silnie zurbanizowanych, które to są mniej eksplorowane wśród użytkowników OSM, przez to także mniej zmapowane, co w konsekwencji powoduje luki w bazie w porównaniu z danymi referencyjnymi (Hecht et al., 2013). Mimo zaobserwowania pewnych wzorców w przypadku związku między badanymi wskaźnikami kompletności danych OSM ich rozkład przestrzenny jest dość zróżnicowany, jak pokazano na opracowanych mapach (Rys. 6, 7). Dlatego wizualizacja jakości OSM jest równie ważna, ponieważ działa jako

narzędzie świadomości dla początkującego użytkownika i narzędzie eksploracji dla eksperta (Zacharopoulou et al., 2021). Opracowana metoda oceny jakości danych o budynkach OSM i wizualizacji wyników ilościowych w celu pomocy użytkownikowi w wyborze zbioru danych jest systematyczna oraz uniwersalna i może być stosowana do dowolnych obiektów powierzchniowych OSM, a także do wzajemnej oceny innych przestrzennych zbiorów danych o porównywalnym zakresie tematycznym oraz podobnej szczegółowości.

Przeprowadzone badania odpowiedziały na pytania P1 i P3 oraz potwierdziły niniejszym słusność hipotez H1 i H3.

5.3. Opracowanie złożonego indeksu jakości danych przestrzennych [publikacja A3]

Wybór odpowiednich danych jest zwykle wspierany analizą ich jakości lub kompletności, często rozumianej jako pojemność informacyjna. Motywacją do tego badania jest zatem opracowanie złożonej miary, która umożliwi porównawczą ocenę kompletności dwóch zbiorów danych. W odróżnieniu od dotychczas prowadzonych badań, opisanych w podrozdziałach 5.1. oraz 5.2., ocena kompletności odnosi się do podstawowej jednostki analitycznej, umożliwiającej szczegółową ocenę w poszczególnych lokalizacjach geograficznych. Opracowana kompleksowa metodyka opiera się na kompensacyjnej analizie porównawczej TOPSIS, niestosowanej wcześniej w ocenie danych kategorycznych, z wykorzystaniem liniowego rankingu wyrażonego w postaci autorskiego współczynnika skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych – Compound Correspondence Index (CCI) oraz statystycznych miar rozproszenia dla jej wizualizacji przestrzennej.

Podstawowy problem badawczy, do którego odnosi się niniejsze opracowanie, dotyczy zdefiniowania złożonego indeksu (CCI) w ujęciu lokalnym (CCI_L) i regionalnym (CCI_R) oraz określenia liczby optymalnych zakresów klas, które jednoznacznie wskazują przestrzenne położenie różnic w pojemności informacyjnej dwóch badanych zbiorów danych. Do wielokryterialnej analizy porównawczej wykorzystałam metodę WLC, opartą na takich kryteriach, jak różnice w powierzchni zajmowanej przez zabudowę, lasy, zbiorniki wodne oraz długości dróg, linii kolejowych i rzek dla analizowanych zbiorów przestrzennych OSM oraz BDOT10k. Minimalna różnica wskazuje na bardzo podobną

ilość informacji, natomiast maksymalna oznacza duże różnice pomiędzy obydwoma zbiorami – wskaźnik CCI szerzej opisano w podrozdziale 4.2.2.

W publikacji tej postawiłam następujące pytanie szczegółowe [P2]: czy autorski współczynnik CCI (Compound Correspondence Index) skonsolidowanej analizy odpowiedniości jest wrażliwy na powiększanie obszaru badań? Innymi słowy, sprawdziłam czy metoda analizy odpowiedniości (korespondencji) dwóch zbiorów danych przestrzennych prowadzona dla zbiorów z poszczególnych obszarów badawczych oddzielnie (lokalnie) lub łącznie (regionalnie) daje takie same wyniki.

Odpowiedź na postawione pytanie pozwala zweryfikować hipotezę, że CCI_R niedoszacowuje (zaniża) różnice pomiędzy analizowanymi zbiorami danych, wskazując na nieco wyższą zgodność niż CCI_L . Podejście to pokazuje różnice pomiędzy klasyfikacją CCI na poziomie lokalnym i regionalnym, a także pozwala na porównanie dwóch zbiorów zawierających dane jakościowe dla obszarów geograficznie rozłącznych, umożliwiając użytkownikowi świadomy i odpowiedzialny wybór jednego z nich, co czyni je innowacyjnym.

W przeprowadzonych badaniach obliczyłam lokalny wskaźnik CCI osobno dla każdego z analizowanych powiatów. Pięć przedziałów klas, których zakresy wyznaczyłam na podstawie wartości odchylenia standardowego, odzwierciedlają zgodność danych BDOT10k i OSM. W przeciwieństwie do klasycznej skali Likerta (Muhammed, 2023) zastosowano odwrotną kolejność, zgodną z wartościami CCI, przedstawioną w Tabeli 7.

Tabela 7. Klasy zgodności CCI [źródło: A3, Tabela 3].

Klasa	Przedział wartości	Opis klasy
1	$-0.50 \sigma > CCI_L$	pełna zgodność (maximum compliance)
2	$-0.5 \sigma \leq CCI_L \leq 0.5 \sigma$	umiarkowana zgodność (moderate compliance)
3	$0.5 \sigma \leq CCI_L \leq 1.5 \sigma$	pół-zgodność (semi-compliance)
4	$1.5 \sigma \leq CCI_L \leq 2.5 \sigma$	niska zgodność (moderate noncompliance)
5	$CCI_L > 2.5 \sigma$	brak zgodności (maximum noncompliance)

Wyznaczony lokalny CCI różnił się pomiędzy rozpatrywanymi powiatami o czym świadczy przedstawiona statystyka opisowa (Tabela 2 w publikacji A3 - załącznik nr 3), oraz wizualizacja, która opiera się na prezentacji wartości CCI w formie kartogramów – Rys. 8.

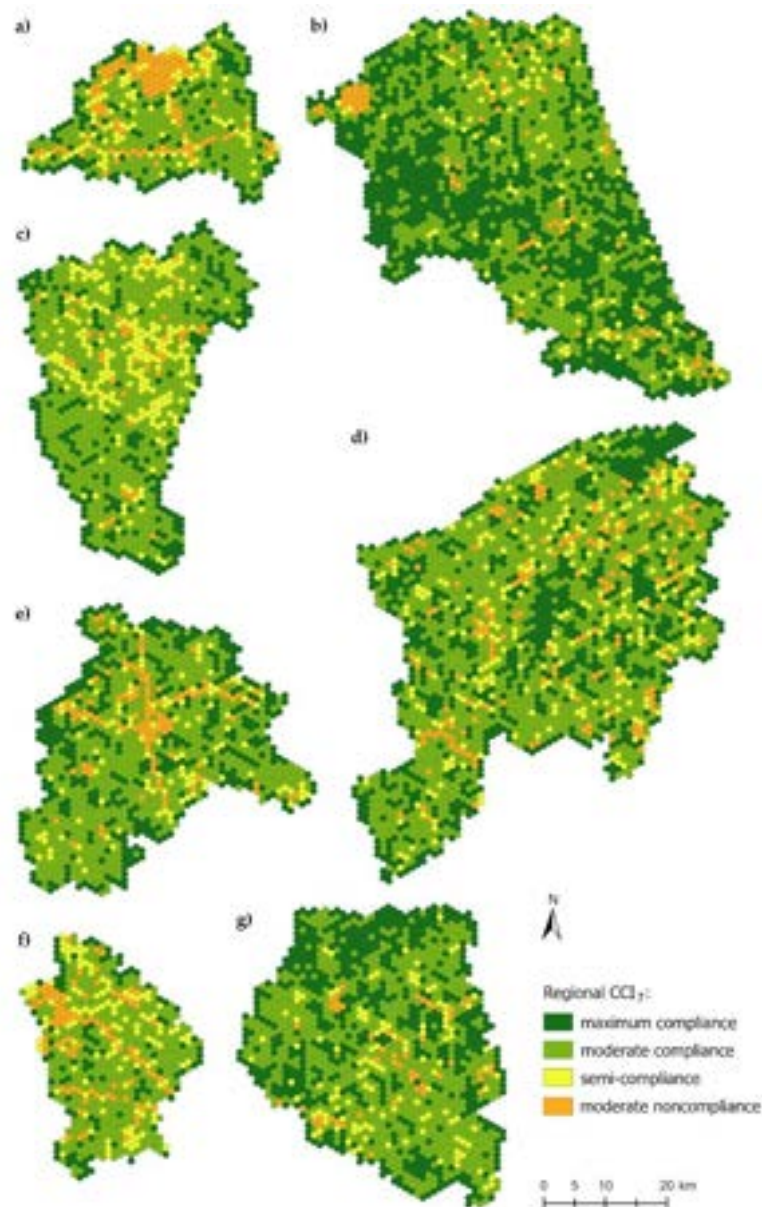


Rys. 8. Wartości CCI_L według przyjętych klas zgodności w analizowanych powiatach: a) piaseczyńskim; b) sokólskim; c) sanockim; d) słupeckim; e) ostrowskim; f) otwockim; g) międzyrzeckim [źródło: A3, Rys. 3].

Wartości CCI_L we wszystkich analizowanych powiatach wykazały klastrowanie, na co wskazuje statystyka autokorelacji przestrzennej Global Moran's I, z zakresami wartości z-score od 21,80 do 14,62 (wartości $p < 0,001$) odpowiednio w powiatach piaseczyńskim i słupeckim.

Aby określić wrażliwość badanego wskaźnika CCI analizę TOPSIS wykonałam w odniesieniu dla czterech podstawowych powiatów, a mianowicie: piaseczyńskiego, sokólskiego, sanockiego i słupeckiego (uzyskując CCI_{R4}), następnie dodając sekwencyjnie po jednym odrębnym powiecie do analizy: ostrowskim (CCI_{R5}), otwockim (CCI_{R6}) oraz

międzyrzeckim (CCI_{R7}). Pozostałe założenia, jak przyjęte wagi oraz klasy zgodności, pozostały bez zmian. W miarę rozszerzania się obszaru badania, zakres regionalnych wartości CCI zwiększa się, a lokalne różnice między danymi stają się mniej znaczne. Miary statystyczne zamieszczone w Tabeli 4 publikacji A3 (załącznik nr 3), jak wariancja, odchylenie standardowe oraz zakres ćwiartkowy maleją, co dowodzi, że wartości CCI_{R7} , których wizualizacja wyników została przedstawiona za pomocą kartogramu na Rys. 9, są mniej zróżnicowane niż CCI_{R6} , CCI_{R5} oraz CCI_{R4} .



Rys. 9. Wartości CCI_{R7} według przyjętych klas zgodności w analizowanych powiatach:
a) piaseczyńskim; b) sokólskim; c) sanockim; d) słupeckim; e) ostrowskim;
f) otwockim; g) międzyrzeckim [źródło: A3, Rys. 4].

Zakładając, że obszar badań składa się z kilku przestrzennie rozłącznych obszarów (np. powiatów, miast), regionalny CCI umożliwia ocenę przydatności zestawów według wspólnej skali opartej na odchyleniu standardowym. Wyniki oceny regionalnej przewyższają wyniki klasyfikacji lokalnej, dając lepsze wyniki, tj. wyższy poziom zgodności danych. Przeszacowanie zgodności regionalnej waha się od 9% do 20% powierzchni powiatu, ze średnią redukcją o 3% w obszarze, na którym dwa zestawy danych (BDOT10k i OSM) mają porównywalny zakres informacji. Obszary o średniej i dużej niezgodności są redukowane średnio o 2,4%. Analiza wrażliwości pokazuje, że ani wielkość regionu, ani położenie przestrzenne powiatów nie miały istotnego wpływu na wartości regionalnego CCI.

Wartości CCI we wszystkich analizowanych powiatach wykazały klastrowanie. Największą zmienność między danymi BDOT10k i OSM zaobserwowano na obszarach o wysokim stopniu urbanizacji (np. miasta Piaseczno oraz Otwock) i w pobliżu przebiegu głównych szlaków transportowych. Należy jednak podkreślić, iż przeprowadzone analizy nie wykazały statystycznie istotnych korelacji między współczynnikami CCI a badanymi elementami pokrycia terenu (budynki, drogi, rzeki, linie kolejowe, lasy i zbiorniki wodne).

Przetwarzanie i porównanie kilku powiatów przy użyciu wspólnych ram analitycznych pozwala na syntetyczną ocenę kompletności analizowanych zbiorów danych geoprzestrzennych oraz identyfikację potencjalnych podobieństw i różnic pomiędzy badanymi rejonami. Metoda proponowana w tym artykule ma kilka ograniczeń, m.in. ważenia zmiennych w metodzie TOPSIS oraz kryterium opierającego się głównie na różnicach geometrycznych w powierzchni i długości obiektu geograficznego analizowanego w zbiorach danych OSM i BDOT10k.

Badania przedstawione w publikacji A3 przyczyniły się do odpowiedzi na pytania szczegółowe P1 oraz P2, udowadniając słuszność hipotez H1 oraz H2.

5.4. Analiza wrażliwości indeksu CCI [publikacja A4]

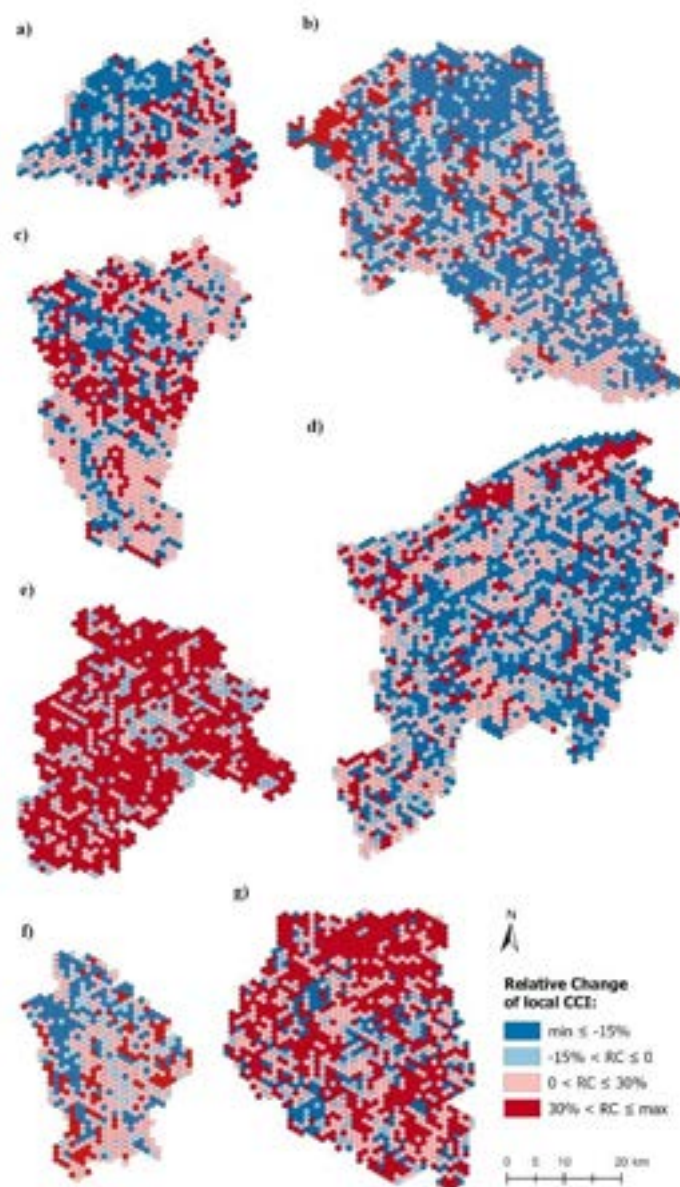
Artykuł porusza problem wagowania kryteriów, który wspiera wybór zbioru danych kategorycznych zgodnie z potrzebami użytkownika. Analiza porównawcza przestrzennych danych jakościowych jest często wykorzystywana do wyboru zestawów danych, które są odpowiednie do celów użytkownika. Zazwyczaj wykorzystuje się w tym ujęciu ocenę wielokryterialną opartą na dostępnych aplikacjach MCDA. Jako narzędzie

wspomagające podejmowanie decyzji, głównym celem MCDA jest pomoc decydom poprzez zapewnienie opcji decyzyjnych zgodnie z przyjętymi kryteriami (Aksoy & San, 2019). W literaturze dotyczącej MCDA przedstawiono wiele reguł ważenia kryteriów, a ich różnorodność prowadzi do następującego pytania: w jaki sposób wybór wag wpływa na ostateczny ranking alternatyw (Odu, 2019; Roszkowska, 2013)? W związku z tym niniejsze badanie ma na celu analizę wpływu wag na ocenę danych topograficznych pod kątem ich przydatności do określonego celu. W publikacji A4 opracowałam autorski wskaźnik CCI obliczony oddzielnie dla każdego powiatu. Z poprzednich badań [publikacja A3] wynika, iż geometria danych jak i waga parametrów obciążały ostateczne wyniki analizy wrażliwości wskaźnika jakości danych CCI, wyznaczonego za pomocą metody TOPSIS. Toteż opisane badania stanowią także wkład w stosunkowo niedawną dyskusję na temat wpływu parametrów początkowych na wyniki analiz wielokryterialnych i wieloatrybutowych (na przykładzie TOPSIS). Wartości CCI bazują na kryteriach takich jak różnice w obszarach pokrytych budynkami, lasami i zbiornikami wodnymi, a także długości dróg, linii kolejowych i rzek, które są przypisane do sześciokątnej siatki o powierzchni 1 km². Syntetyczny wskaźnik CCI, który opisuje różnice między dwoma badanymi zbiorami danych topograficznych (OSM i BDOT10k), został opracowany przy użyciu klasycznej metody TOPSIS w dwóch podejściach. Pierwsza kombinacja wag dla obliczonego wskaźnika jakości CCI_{w1} zakłada zróżnicowane wagi, które mogą zostać wybierane w sposób subiektywny przez użytkownika – opisane w publikacji A3. W drugiej kombinacji wag dla obliczonego wskaźnika jakości CCI_{w2}, zastosowanej w tym badaniu, założyłam iż wszystkie kryteria są równie ważne, dlatego każde z analizowanych obiektów przyjmuje wagę 0,167. Aby wyrazić względną zmianę procentową pomiędzy różnicami wag zastosowanych dla opracowanego współczynnika jakości CCI_L w dwóch wariantach wagowania, zastosowałam wskaźnik Relative Change (RC). Warianty wag dla CCI wraz z zastosowanym wskaźnikiem RC szczegółowo opisałam w podrozdziale 4.2.2. niniejszej rozprawy.

Według przyjętej stopniowej skali zgodności, przedstawionej w Tabelach 4–5 publikacji A4 (załącznik nr 4), wskaźniki CCI w dwóch kombinacjach wag zajmowały podobne udziały powierzchni każdego powiatu. Największe różnice w zajmowanych powierzchniach przypisanych do klas, można było zaobserwować w przypadku powiatu piaseczyńskiego – powierzchnia obszarów maksymalnej i umiarkowanej zgodności zmniejszyły się o 22,8 punktów procentowych udziału w powierzchni powiatu po

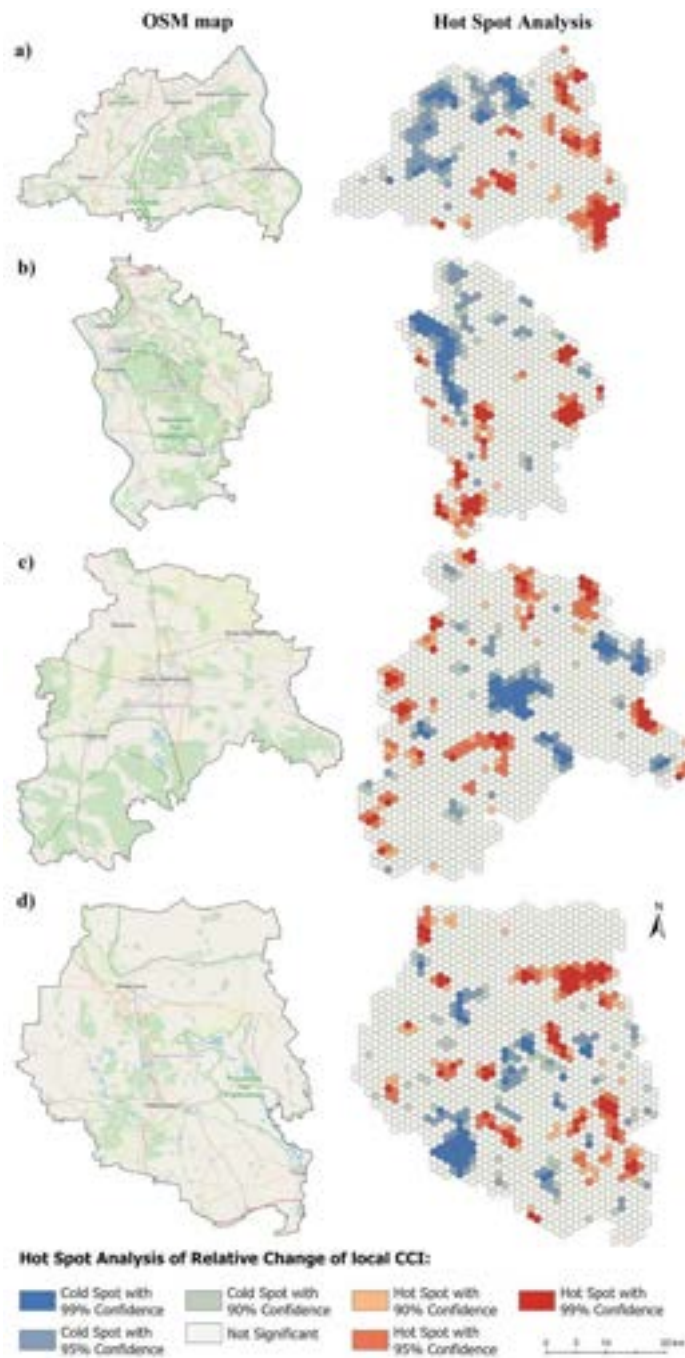
zrównaniu wag CCI (wynosząc 55%). Z kolei obszar zajmowany przez pół-zgodność podwoił się, osiągając 25% udziału w powierzchni powiatu piaseczyńskiego. Dodatkowo udział obszarów ocenionych jako umiarkowanie i maksymalnie niezgodne wzrósł z 9% do 20,2% udziału w powierzchni powiatu po zrównaniu wag. Pozwoliło to stwierdzić, że podobnie jak w poprzednich badaniach, obszary o wysokim stopniu urbanizacji wykazywały największą zmienność między danymi OSM a BDOT10k.

Wartości wskaźnika RC wyznaczonego CCI_L dla analizowanych wariantów wag były dość zróżnicowane – wyniki przedstawiono w Tabeli 7 publikacji **A4** (załącznik nr 4). Kartograficzną prezentację wartości RC przedstawiono na Rys. 10. Ujemne wartości RC ujawniły przewagę wariantu ważenia z różnymi wagami (CCI_{W1}). Zgodnie z wynikami najwyższy udział ujemnych wartości RC wykazano w powiatach sokólskim (58,4%) i piaseczyńskim (51,1%). Z kolei dodatnie wartości RC świadczyły o przewadze wariantu równych wag (CCI_{W2}), było to szczególnie widoczne w powiatach ostrowskim (83,7%), międzyrzeckim (77,9%) i sanockim (69,9%). Powiatem o najbardziej wyrównanych udziałach poszczególnych wariantów wagowych był powiat słupecki (odpowiednio 49,8% i 50,2% dla wartości ujemnych i dodatnich RC).



Rys. 10. Wartości RC dla CCI_L w analizowanych powiatach: a) piaseczyńskim; b) sokólskim; c) sanockim; d) słupeckim; e) ostrowskim; f) otwockim; g) międzyrzeckim [źródło: A4, Rys. 2].

Na poziomie pikseli analiza korelacji Pearsona nie wykazała istotnego związku między typem pokrycia terenu a CCI dla równych wag (CCI_{W2}), podobnie jak w przypadku CCI_{W1} , opisanym w publikacji A3. Nie wykazano również istotnego związku statystycznego we względnych zmianach między zastosowanymi wagami CCI. Z tego powodu przeprowadzono analizę hotspot w celu zidentyfikowania skupisk pikseli o podobnych wartościach. Wynikowe mapy analizy hotspot dla wybranych powiatów zamieszczono na Rys. 11.



Rys. 11. Mapy analizy hotspot wartości RC dla CCI_L wybranych powiatów w zestawieniu do mapy OSM: a) piaseczyński; b) otwocki; c) ostrowski; d) międzyrzecki [źródło: A4, Rys. 4].

Obszary, które zostały zidentyfikowane jako gorące punkty, skupiały się głównie w lasach, na terenach otwartych, obszarach uprawnych, w sąsiedztwie zbiorników wodnych oraz (rzadziej) na obszarach o niskiej gęstości zabudowy. Natomiast obszary, które zostały zidentyfikowane jako zimne punkty, znajdowały się na obszarach gęsto

zabudowanych i wzdłuż głównych linii transportowych. To obszary o największej zgodności i nieznacznym wpływie zmiany wag na wynik preferencji.

Zastosowane różne warianty wag, użyte do obliczenia autorskiego skonsolidowanego wskaźnika odpowiedniości zbiorów danych OSM oraz BDOT10k w postaci CCI, a także wyznaczenie względnej zmiany CCI, dowiodły znacznej wrażliwości badanych zbiorów na przyjęte kombinacje wag na poziomie lokalnym. Świadczy to o dużym wpływie przyjętych kryteriów na otrzymane wyniki oceny jakości danych OSM. Równe wagi w metodzie TOPSIS wpłynęły na liczbę pikseli, a tym samym udział procentowy w powierzchni powiatów przypisanych do określonych klas zgodności CCI.

W obu zastosowanych wariantach wag największy udział procentowy zajmowały obszary o maksymalnej oraz umiarkowanej zgodności, natomiast najmniejszy, obszary o niskiej zgodności bądź jej braku. Największe różnice zauważono w powiecie piaseczyńskim oraz międzyrzeckim, dla których obszary o najwyższej oraz umiarkowanej zgodności po zastosowaniu równych wag zmniejszyły się odpowiednio z 77,8% do 55% oraz z 85,3% do 80,10% na rzecz obszarów o niższej zgodności. Różnice dla pozostałych powiatów na ogół nie przekraczały 3 punktów procentowych, toteż wyznaczenie zmian procentowych różnic wag CCI w postaci wskaźnika RC pozwoliło na zidentyfikowanie ich względnej wielkości oraz poszczególnych miejsc w przestrzeni. Z przeprowadzonych analiz hotspot wynika, iż w badanych powiatach klastry zdefiniowane jako gorące i zimne punkty obejmowały podobne typy pokrycia terenu, co pozwoliło na ich scharakteryzowanie pod względem typu osadnictwa i użytkowania gruntów.

Przedstawione w publikacji A4 wyniki badań przyczyniły się do realizacji pytań szczegółowych P1 oraz P2, potwierdzając słuszność przyjętych hipotez H1 oraz H2.

6. PODSUMOWANIE

Przydatność do celu jest zasadą powszechnie akceptowaną przez analityków jako prawidłowe podejście do uzyskania zestawu danych wysokiej jakości, jednak tylko nieliczni analitycy lub użytkownicy końcowi danych mogą dokładnie określić, jaka jakość danych jest wymagana dla konkretnego zadania. Wybierając zestaw danych przestrzennych oraz użytych do analiz kryteriów, użytkownik powinien być bardzo uważny, ponieważ nie jest możliwa ocena wszystkich mocnych i słabych stron dostępnych danych.

Spójny tematycznie cykl czterech publikacji, pt. „**Metodyka wielocechowej oceny porównawczej przydatności jakościowych danych przestrzennych**”, który stanowi niniejszą rozprawę doktorską, miał na celu przedstawienie oryginalnego rozwiązania polegającego na metodyce oceny zgodności przestrzennych danych jakościowych. Umożliwia to zaproponowany przeze mnie autorski współczynnik skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych Compound Correspondence Index (CCI).

Teza niniejszej rozprawy doktorskiej brzmi: „**metoda wielocechowej analizy porównawczej i wizualizacji jakości danych oraz autorskie wskaźniki oceny przydatności danych przestrzennych stanowią podstawy do spójnej ich oceny przez użytkownika**”. Wynikała ona z faktu, iż istniejące standardowe miary i wskaźniki oceny jakości heterogenicznych danych wolontariackich nie są wystarczająco inkluzywne, aby kompleksowo ocenić dane OSM oraz pokazać ich zróżnicowanie względem analizowanych zbiorów.

Do przeprowadzenia badania jakości danych społecznościowych OSM w odniesieniu do danych urzędowych BDOT10k, przyjętych jako zbiór referencyjny, zastosowałam standardowe wskaźniki jakości danych rozszerzone o dodatkowe wskaźniki kompletności [publikacje A1 oraz A2]. Udowodniłam, iż mierniki jakości ISO uzupełnione o autorskie wskaźniki, umożliwiają bardziej złożoną ocenę jakości przestrzennych danych społecznościowych z perspektywy użytkownika oraz pokazania zróżnicowania jakości wewnątrz analizowanych zbiorów, odpowiadając na pierwsze pytanie szczegółowe [P1] i częściowo potwierdzając tym samym pierwszą hipotezę [H1]. Wyniki tych badań przyczyniły się do opracowania przeze mnie kartograficznej prezentacji wskaźników kompletności OSM w postaci zmodyfikowanego kartogramu złożonego, który poprzez zilustrowanie konkretnego miejsca w przestrzeni uwidoczniał

skupiska wartości wysokich i niskich, tym samym wskazując obszary o podobnej lub różnej jakości danych [publikacja **A2**]. Udowodniłam, iż taka wizualizacja w znaczącym stopniu ułatwia użytkownikowi podjęcie decyzji o przydatności zbioru dla określonych celów, odpowiadając na trzecie pytanie szczegółowe [**P3**] i potwierdzając trzecią hipotezę [**H3**]. W kolejnym kroku [publikacja **A3**] opracowałam autorski współczynnik Compound Correspondence Index (CCI), który wykorzystując przyjęte pola podstawowe, przedstawiał konkretne miejsca w przestrzeni, gdzie zawartość obu zbiorów danych jest najbardziej zgodna oraz te, dla których jest zdecydowanie różna. Wizualizacja kartograficzna w postaci kartogramów oraz dane tabelaryczne umożliwiły szybki wybór przez użytkownika pożądanego zbioru danych topograficznych. Tym samym odpowiedziałam na pierwsze pytanie szczegółowe [**P1**], udowadniając hipotezę pierwszą [**H1**]. Opracowany wskaźnik CCI obliczony został w dwóch podejściach lokalnym oraz regionalnym w celu oceny jego wrażliwości na powiększanie obszaru badań. Udowodniłam, iż autorski współczynnik skonsolidowanej analizy odpowiedniości zbiorów danych przestrzennych CCI pokazuje większą zgodność w ujęciu regionalnym niż lokalnym, odpowiadając tym samym na drugie pytanie szczegółowe [**P2**] oraz potwierdzając drugą hipotezę [**H2**]. W ostatniej publikacji [**A4**] analizowałam wrażliwość wskaźnika CCI na zmianę wag przypisanych do obiektów topograficznych. W tym celu zastosowałam wskaźnik Relative Change (RC), informujący o zmianach procentowych CCI w przyjętych dwóch kryteriach wagowania obiektów przestrzennych. Opracowana wizualizacja kartograficzna oparta na kartogramach oraz analizie hotspot, ilustruje obszary, w których różne wagi zasadniczo zmieniają wynik wartości wskaźnika CCI. Wynik publikacji **A4** w postaci opracowanego skonsolidowanego wskaźnika zgodności danych jakościowych CCI w dwóch wariantach wag w całości odpowiedział na pierwsze pytanie szczegółowe [**P1**], zupełnie udowadniając postawioną hipotezę **H1**, iż zasadne jest opracowanie nowych mierników jakości danych przestrzennych potrzebnych użytkownikowi do szczegółowej i kompleksowej oceny jakości przestrzennych danych społecznościowych. W Tabeli 8 przedstawiłam w sposób schematyczny pytania szczegółowe wraz z postawionymi hipotezami i odpowiadającymi za ich udowodnienia publikacjami w cyklu.

Tabela 8. Schemat systematyki prowadzonych badań.

Pytanie szczegółowe	Hipoteza	Artykuł w cyklu
P1	H1	A1, A2, A3, A4
P2	H2	A3, A4
P3	H3	A2

Ważnym aspektem, który znacząco wpływa na wyniki rankingu TOPSIS, wykorzystanego do opracowania współczynnika CCI, jest wybór obiektów topograficznych i ich priorytetyzacja, co zwykle wiąże się z nadrzędnym celem, tj. odpowiedzią na pytanie o przedmiot analizy. W niniejszym badaniu przyjęto, że cel ten był związany z zarządzaniem sytuacjami kryzysowymi (tj. powodzie, pożary, ataki terrorystyczne), dla którego ważna jest identyfikacja obszarów zaludnionych, dróg dojazdowych, terenów zalesionych i innych miejsc strategicznych. Z tego względu do analiz przyjęto główne elementy pokrycie terenu jak budynki, lasy, sieci transportowe oraz wody powierzchniowe, których znaczenie w wymienionych sytuacjach jest niepodważalne.

Warto zauważyć, iż opracowany wskaźnik CCI można obliczyć dla dowolnej jednostki mapowej (MU), zarówno naturalnej (np. zlewnie, ekotopy), administracyjnej, jak i geometrycznej. Uniwersalność CCI polega również na tym, że w analizach można uwzględnić dowolne obiekty, a jednostki mapowania można uszeregować, biorąc pod uwagę inne nietopograficzne dane kategoryczne. Zaproponowana przeze mnie autorska klasyfikacja zgodności danych wskaźnika CCI (Tabela 7), w sposób optymalny rankinguje wartości w polach podstawowych od najbardziej do najmniej zgodnych, w znaczący sposób ułatwiając użytkownikowi ocenę wykorzystanych danych jakościowych i ich weryfikację pod kątem przydatności do założonego celu. Opracowana skala bazuje na wartości odchylenia standardowego i może być wykorzystana w podobnych badaniach, dotyczących porównania zawartości informacyjnej danych kategorycznych.

7. WNIOSKI

Niniejsza rozprawa wpisuje się w światowe trendy, dotyczące badania jakości danych przestrzennych oraz poszukiwania nowych zastosowań informacji przestrzennej. Stanowi istotny wkład w ocenę porównawczą danych jakościowych.

Na podstawie przeprowadzonych badań sformułowałam następujące wnioski końcowe:

1. Istniejący zestaw wskaźników jakości danych przestrzennych jest niewystarczający do oceny użytkownika (zgodnie z przyjętym założeniem „fitness for purpose”).
2. Autorski wskaźnik Compound Correspondence Index (CCI) umożliwia kompleksową porównawczą analizę o szczegółowości piksela dwóch zbiorów danych jakościowych.
3. Autorskie modyfikacje kartogramu, podkreślające różnice względem badanych zbiorów, ułatwiają wskazanie miejsc szczególnie istotnych z punktu widzenia wyboru zbiorów przestrzennych przez użytkownika.

Przeprowadzone badania potwierdzają stwierdzenia i wnioski wielu autorów o konieczności zdefiniowania dodatkowych wskaźników umożliwiających kompleksową ocenę jakości danych społecznościowych. Udowadniają one zatem prawdziwość postawionej na wstępie tezy badawczej, brzmiącej „**metoda wielocechowej analizy porównawczej i wizualizacji jakości danych oraz autorskie wskaźniki oceny przydatności danych przestrzennych stanowią podstawy do spójnej ich oceny przez użytkownika**”.

LITERATURA

- Aksoy, E., & San, B.T. (2019). Geographical Information Systems (GIS) and Multi-Criteria Decision Analysis (MCDA) integration for sustainable landfill site selection considering dynamic data source. *Bulletin of Engineering Geology and the Environment*, 78(4), 779–791. DOI: 10.1007/s10064-017-1135-z
- Baranowski, M., Gotlib, D., & 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
- Barrington-Leigh, C., & Millard-Ball, A. (2019). Correction: The world's user-generated road map is more than 80% complete. *PLOS ONE* 14(10): e0224742. <https://doi.org/10.1371/journal.pone.0224742>
- Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6), 877-895. DOI: 10.1111/tgis.12073
- Białousz, S., Osińska-Skotak, K., Fiałkowska, A., Pluto-Kossakowska, J., Lady-Drużycka, K., & Różycki, S. (2004). *Systemy Baz Danych Przestrzennych województwa Mazowieckiego*. Oficyna wydawnicza PW, Warszawa.
- Bielawa, A. (2011). Postrzeganie i rozumienie jakości-przeгляд definicji jakości. *Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania*, 21, 144-152.
- Bielecka, E., & Leszczyńska, M. (2018). Usability of the OpenStreetMap forest data. *Sylwan* (6), 460-468. <https://doi.org/10.26202/sylwan.2018040>
- Bielecka, E., & Maj, K. (2009). Systemy informacji przestrzennej: podstawy teoretyczne. *Wojskowa Akademia Techniczna*, Warszawa.
- Biljecki, F., Chow, Y. S., & Lee, K. (2023). Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. *Building and Environment*, 237, 110295.
- Birch, C., Oom, S., & Beecham, J. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*. 206. 347-359. 10.1016/j.ecolmodel.2007.03.041

- Borkowska, S., & Pokonieczny, K. (2022). Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability* 14(7):3728. <https://doi.org/10.3390/su14073728>
- Borkowska, S., Bielecka, E., & Pokonieczny, K. (2023). Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Applied Sciences*, 13, 7525. <https://doi.org/10.3390/app13137525>
- Borkowska, S., Bielecka, E., & Pokonieczny, K. (2024). Weights Impact on the Comparative Evaluation of Topographic Data. *Geomatics and Environmental Engineering*, 18, 4. <https://doi.org/10.7494/geom.2024.18.4.97>
- Borkowska, S., Bielecka, E., & Pokonieczny, K. (2023). OpenStreetMap - building data completeness visualization in terms of “Fitness for purpose”. *Advances in Geodesy and Geoinformation*, 72, 1, 1–20. <https://doi.org/10.24425/agg.2022.141922>
- Brovelli, M., Minghini, M., Molinari, M., & Zamboni, G. (2016). Positional accuracy assessment of the OpenStreetMap buildings layer through automatic homologous pairs detection: The method and a case study. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLI-B2*. 615-620. [10.5194/isprs-archives-XLI-B2-615-2016](https://doi.org/10.5194/isprs-archives-XLI-B2-615-2016)
- Cichociński P. (2012). Ocena przydatności OpenStreetMap jako źródła danych dla analiz sieciowych, *Roczniki Geomatyki 2012, TOM X*, 7(57).
- 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. <https://doi.org/10.3390/ijgi4031657>
- 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. <https://doi.org/10.1080/13658816.2013.867495>
- GEOFABRIK. (2024). Pozyskano z: <http://download.geofabrik.de/europe/poland.html> (dostęp: 22-09-2024).
- Geoportal Krajowy. (2024). Pozyskano z: <https://www.geoportal.gov.pl/pl/aplikacje/geoportal-krajowy/> (dostęp: 22-09-2024).

- Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24, no. 3. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>
- Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information, *Spatial Statistics*, 1, 110-120, ISSN 2211-6753, <https://doi.org/10.1016/j.spasta.2012.03.002>.
- Goodchild, M.F. (1986). Spatial Autocorrelation. *Catmog* 47, Geo Books. ISSN 0306-6142.
- Goodchild, M.F., & Hunter, G.J. (1997). A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.*, 11, 299–306.
- Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design*, 37(4), 682-703. DOI: 10.1068/b35097
- Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro–Wilk Test with Known Mean. *Revstat Stat. J.*, 14, 89–100. DOI: 10.57805/revstat.v14i1.180
- Hecht, R., Kunze, C., & Hahmann, S. (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS Int. J. Geo-Inf.*, 2, 1066-1091. <https://doi.org/10.3390/ijgi2041066>
- International Organization for Standardization (2014). Geographic information - Metadata, ISO Standard No. 19 115:2014, <https://www.iso.org/standard/53798.html> (dostęp: 22-09-2024)
- International Organization for Standardization (2023). Geographic Information - Data Quality, ISO Standard No. 19157-1:2023, <https://www.iso.org/standard/78900.html>, 2023 (dostęp: 22-09-2024)
- Leonowicz, A. (2002). Prezentacja zależności zjawisk metodą kartogramu złożonego. *Polski Przegląd Kartograficzny*, 34, 273–85.
- MacEachren, A. M., & Kraak, M. J. (2013). Research Challenges in Geovisualization. *Cartography and Geographic Information Science*, 28(1), 3-12. <https://doi.org/10.1559/152304001782173970>

- Maruf, M. (2023). Alternative approach to analysing data obtained with Likert scale. *Route Educational and Social Science Journal*, 10(5):96. 10.17121/ressjournal.3439.
- MRPiT. (2021). Rozporządzenie Ministra Rozwoju, Pracy i Technologii z dnia 27 lipca 2021 r. w sprawie bazy danych obiektów topograficznych oraz bazy danych obiektów ogólnogeograficznych, a także standardowych opracowań kartograficznych. Dz.U. 2021, nr 30, poz. 1412. Pozyskano z: <https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210001412> (dostęp: 22-09-2024).
- Neis, P., Zielstra, D., & Zipf, A. (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007-2011. *Future Int.*, 4, 1-21. <https://doi.org/10.3390/fi4010001>
- Nowak Da Costa, J. N., Bielecka, E., & Całka, B. (2016). Jakość danych OpenStreetMap – analiza informacji o budynkach na terenie Siedlecczyzny. *Roczniki Geomatyki-Annals of Geomatics*, 14(2 (72)), 201-211.
- Odu, G.O. (2019). Weighting methods for multi-criteria decision making technique. *Journal of Applied Sciences and Environmental Management*, 23(8), 1449–1457. <https://www.ajol.info/index.php/jasem>.
- Ord, K., & Getis, A. (2010). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27(4), 2010, 286–306. <https://doi.org/10.1111/j.1538-4632.1995.tb00912.x>.
- Pavić, Z., & Novoselac, V. (2013). Notes on TOPSIS Method. *International Journal of Research in Engineering and Science*, 1, 5-12.
- Pokonieczny, K. (2018). Comparison of land passability maps created with use of different spatial data bases. *Geografie*, 123, 317–352. <https://doi.org/10.37040/geografie2018123030317>
- Roick, O., Hagenauer, J., & Zipf, A. (2011). OSMatrix - Grid based analysis and visualization of Open-StreetMap. In *Proceedings of the 1st European State of the Map Conference(SOTM-EU)*, Vienna, Austria.
- Roszkowska, E. (2013). Rank ordering criteria weighting methods – a comparative overview. *Optimum. Studia Ekonomiczne*, 5(65), 14-33. <https://doi.org/10.15290/ose.2013.05.65.02>.

- Senaratne, H., Mobasheri, A., Ali, A., Cristina, C., & Haklay, M. (2016). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1). DOI: 10.1080/13658816.2016.1189556
- Su, Y., Yang, L. & Jin, Z. (2007). Evaluating Spatial Data Quality in GIS Database. *Published in: 2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 5967-5970, Shanghai, China. DOI: 10.1109/WICOM.2007.1463
- Tian, Y., Zhou, Q., & 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. <https://doi.org/10.3390/ijgi8010035>
- Weiss, D.J., Nelson, A., Vargas-Ruiz, C.A. et al. (2020). Global maps of travel time to healthcare facilities. *Nat Med*, 26, 1835–1838. <https://doi.org/10.1038/s41591-020-1059-1>
- Zacharopoulou, D., Skopeliti, A., & Nakos, B. (2021). Assessment and Visualization of OSM Consistency for European Cities. *ISPRS Int. J. Geoinf.*, 10, 361. DOI: 10.3390/ijgi10060361
- Zavadskas, E., Mardani, A., Turskis, Z., Jusoh, A., & Nor, K. (2016). Development of TOPSIS method to solve complicated decision-making problems: An overview on developments from 2000 to 2015. *International Journal of Information Technology & Decision Making*. DOI: 10.1142/S0219622016500176
- Zhang, X., An, J., Zhou, Y., Yang, M., & Zhao, X. (2024). How sustainable is OpenStreetMap? Tracking individual trajectories of editing behavior. *International Journal of Digital Earth*, 17(1). <https://doi.org/10.1080/17538947.2024.2311320>
- Zielstra, D., & Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. *In Proceedings of the 13th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal.

ZAŁĄCZNIKI

Artykuły stanowiące cykl publikacji wraz z oświadczeniami współautorów o procentowym udziale w poszczególnych publikacjach:

1. **Borkowska, S.** (90%), & Pokonieczny, K. (10%) (2022). Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability*, 14, 3728. <https://doi.org/10.3390/su14073728>
2. **Borkowska, S.** (75%), Bielecka, E. (12,5%), & Pokonieczny, K. (12,5%) (2023). OpenStreetMap - building data completeness visualization in terms of “Fitness for purpose”. *Advances in Geodesy and Geoinformation*, 72, 1, 1–20. <https://doi.org/10.24425/agg.2022.141922>
3. **Borkowska, S.** (80%), Bielecka, E. (8%), & Pokonieczny, K. (12%) (2023). Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Applied Sciences*, 13, 7525. <https://doi.org/10.3390/app13137525>
4. **Borkowska, S.** (75%), Bielecka, E. (10%), & Pokonieczny, K. (15%) (2024). Weights Impact on the Comparative Evaluation of Topographic Data. *Geomatics and Environmental Engineering*, 18, 4. <https://doi.org/10.7494/geom.2024.18.4.97>

Article

Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development

Sylwia Borkowska *  and Krzysztof Pokonieczny 

Institute of Geospatial Engineering and Geodesy, Faculty of Civil Engineering and Geodesy, Military University of Technology (WAT), 00-908 Warsaw, Poland; krzysztof.pokonieczny@wat.edu.pl

* Correspondence: sylwia.borkowska@wat.edu.pl

Abstract: One potential source of geospatial open data for monitoring sustainable development goals (SDG) indicators is OpenStreetMap (OSM). The purpose of this paper is to provide a comprehensive evaluation of the spatial data quality elements of OSM against the national official data—the database of topographic objects at a scale of 1:10,000. Such spatial data quality elements as location accuracy, data completeness and attribute compatibility were analysed. In the conducted OpenStreetMap tests, basic land-cover classes such as roads, railroads, river network, buildings, surface waters and forests were analysed. The test area of the study consisted of five counties in Poland, which differ in terms of location, relief, surface area and degree of urbanization. The best results of the quality of OSM spatial data were obtained for highly urbanized areas with developed infrastructure and a high degree of affluence. The highest degree of completeness of OSM linear and area objects in the studied counties was acquired in Piaseczyński County (82%). The lowest degree of completeness of the line and area objects of OSM in the studied counties was obtained in the Ostrowski County (51%). The calculated correlation coefficient between the quality of OSM data and the income per capita in the county was 0.96. The study complements the previous research results in the field of quantitative analysis of the quality of OSM data, and the obtained results confirm their dependence on the geometric type of the analysed objects and characteristics of test areas, i.e., in this case counties in Poland. The obtained results of OSM data quality analysis indicate that OSM data may provide strong support for other spatial data, including official and state data. OSM stores significant amounts of geospatial data with relatively high data quality that can be a valuable source for monitoring some SDG indicators.

Keywords: data quality; OpenStreetMap; open data; VGI; sustainability



Citation: Borkowska, S.; Pokonieczny, K. Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability* **2022**, *14*, 3728. <https://doi.org/10.3390/su14073728>

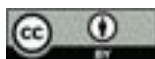
Academic Editors: Haklae Kim, Jangwon Gim and Dongjun Suh

Received: 30 January 2022

Accepted: 18 March 2022

Published: 22 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The issue of spatial data quality has been attracting broad interest for many years, not only of data producers and distributors, but also from users and researchers. The importance of data quality in business and science is well recognized and widely described [1–4]. From the perspective of a data provider or distributor, quality assessment is one of the key elements of production that is always analysed in the context of compliance with technical specifications. The ability to create, collect, store, maintain, transmit, process and present information and data to support business processes in a timely and cost-effective manner requires both an understanding of the characteristics of information and data that determine its quality and the ability to measure, manage and report on it. The issue of geospatial data quality has a crucial role in terms of its ability to be used to determine sustainable development goals indicators. In 2015, the United Nations adopted 17 sustainable development goals (SDGs), which represented a universal call to action to end poverty, protect the planet and ensure peace and prosperity for all by 2030 [5]. To track progress toward sustainable development goals, the UN proposed a set of 231 statistical indicators that range from health outcomes such as infant mortality (Indicator 3.2) to economic indicators such as the percentage of the population living in poverty (Indicator 1.1), environmental indicators

such as air quality (Indicator 11.6) and geospatial data. Indicator-based approaches help ground broad and, in many cases, vague sustainable development goals in more concrete and measurable terms, but obtaining the data needed to monitor indicators on a national or global scale is a significant and fundamental challenge.

Data quality problems are quite widely highlighted by the international standards organization ISO (the International Organization for Standardization). ISO has developed several standards dedicated to the assessment and reporting of geospatial data quality, including: ISO 8000-61:2016—data quality and ISO 19157:2013—geographic information—data quality. ISO 8000-61:2016 specifies the processes required for data quality management. Each process is defined by a purpose, outcomes and the activities that are to be applied for the assurance of data quality [6].

The following are within the scope of this part of ISO 8000: fundamental principles of data quality management, the structure of the data quality management process, definitions of the lower-level processes for data quality management, the relationship between data quality management and data governance and implementation requirements.

The scope of this part of ISO 8000 does not include detailed methods or procedures by which to achieve the outcomes of the defined processes.

ISO 19157:2013 establishes the principles for describing the quality of geographic data. It defines the components for describing data quality, specifies the components and content structure of a register for data quality measures, describes general procedures for evaluating the quality of geographic data and establishes the principles for reporting data quality. ISO 19157:2013 also defines a set of data quality measures for use in evaluating and reporting data quality. It is applicable to data producers who provide quality information to describe and assess how well a data set conforms to its product specification and to data users attempting to determine whether or not specific geographic data are of sufficient quality for their particular application. ISO 19157:2013 does not attempt to define minimum acceptable levels of quality for geographic data.

In ISO normative documents, quality is defined as a comprehensive set of characteristics and features of data sets and services that affect the ability to satisfy current and future user requirements [7]. The characteristics and features mentioned in the standard with respect to spatial data sets are defined by more than a dozen quantitative and qualitative indicators. The most commonly used ones, which are also valid for INSPIRE Directive (Infrastructure for Spatial Information in Europe) data sets, include completeness (lack and excess of objects), logical consistency (conceptual, topological, domain and format), position accuracy, temporal and thematic accuracy (e.g., correctness of classification or correctness of quality attributes) and lineage [8]. All the mentioned quality elements are assessed in terms of compliance with the technical specifications of the data, and the results of the assessment are reported in the metadata. Researchers also pay attention to data availability, which is often a key element of quality, and to the distributor that guarantees better quality, e.g., official data are considered more reliable [3].

As far as data collected voluntarily and free of charge by a very large number of volunteers, referred to as volunteered geographic information (VGI) or crowdsourcing geodata are concerned, the use of the indicators mentioned above becomes problematic. This results from the lack of detailed technical specifications, only giving rules and guidelines for data provision, and the frequent absence of formal verification of all data entered into the database. Volunteers are usually left with a great deal of discretion regarding the accuracy of the data entered and the detail of its descriptive characteristics. Data verification is generally performed by other users, who are potentially more familiar with the area or willing to use community data for specific tasks. The dataset that is the most commonly studied and evaluated for data quality is OpenStreetMap (OSM), which is also a potential source of geospatial data for monitoring SDG indicators [9,10].

1.1. Related Works

The usage of OpenStreetMap has rapidly increased since it was first established in 2004. In line with this increased usage, a number of studies have been conducted to analyse the accuracy and quality of OSM data, but many of them focus mainly on the completeness and accuracy of the location of roads or buildings. In the study [11], the author aimed to analyse the quality of OSM data by comparing it with the Ordnance Survey (OS) data sets in England and selected five areas of London. The geometric accuracy and completeness of road sections were analysed. The analysis shows that OSM information can be quite accurate: on average at a distance of about 6 m from the position recorded by the operating system and with about 80% overlap of highway features between the two data sets. In paper [12], OSM road data was analysed to characterize the behaviour of OSM participants. The study area, Ankara, the capital of Turkey, was evaluated using several network analysis methods such as completeness, degree centrality, proximity, PageRank and a proposed method to measure contributor activation in a limited area from 2007–2017. The results show that the experience level of contributors determines the type of contribution. In general, more experience means more detailed contributions. The paper [13] analyses the spatial pattern, evolution, density and diversity of OSM road networks in Iran between 2008 and 2016 and looks to find casual relations between OSM and census statistics. This is due to the fact that OSM completeness reflects the importance of OSM data in human life. The results show that the road network in Iran considerably increased from 2008 to 2016, with road length increasing to 489,400 km in 2016 from 4300 km in 2008. Road density grew while road diversity and evenness declined. Mapping direction extended from big cities to medium or small-sized ones. Western counties located in mountainous regions are still not very active. The top active counties producing OSM data are mostly populated by urban citizens.

This study [14] aims to provide an analysis of the evolution, completeness and spatial patterns of OSM building data in China from 2012 to 2017 using two quality indicators, OSM building count and OSM building density. The development of OSM numbers from 2012 to 2017 is analysed by province in a regular 1 km² grid. The obtained results showed that the number of OSM buildings increased nearly 20 times from 2012 to 2017, and in most cases, economic (gross domestic product) and OSM road length are two factors that can influence the development of OSM building data in China. Most grid cells in urban areas have no building data, but two typical patterns (dispersion and aggregation) of high-density grid cells are among the prefecture-level divisions. In the article [15], the authors describe the methods of completeness analysis of OSM buildings and their application to various test areas in Germany. The results show that unit-based completeness measurements (e.g., total number or area of buildings) are very sensitive to modelling discrepancies between official data and OSM. The November 2011 analysis in Germany showed a completeness of 25% in the Länder of North Rhine-Westphalia and 15% in Saxony. While further analyses from 2012 confirm that the completeness of the data in Saxony increased to 23%, the pace of new data entry decreased in 2012. In study [16], the quality of OSM land use and land-cover (LULC) data is investigated for an area in southern Germany for two spatial data quality elements: thematic accuracy and completeness. The results show a substantial agreement between OSM and the authoritative dataset. Nonetheless, for this study region, there were clear variations between the LULC classes. Forest covers a large area and shows both a high OSM completeness (97.6%) and correctness (95.1%). In contrast, farmland also covers a large area, but for this class OSM shows a low completeness value (45.9%) due to unmapped areas. Additionally, the results indicate that a high population density, as present in urbanized areas, seems to denote a higher strength of agreement between OSM and the DLM (digital landscape model).

These studies show that OSM information can be quite accurate, but its value depends on the areas for which it was acquired. The aforementioned studies clearly show that the best spatial data quality results were achieved for urban areas and those of interest to OSM

users. Based on the statistical results of the research, the authors have inferred that the experience levels of contributors determine the contribution type and level of object detail.

1.2. Research Purpose

Taking into account the above facts, this paper presents a comprehensive assessment of the quality of OpenStreetMap volunteer data, paying attention to the aspect of imperfect semantic findings and quality assumptions. The research question posed was to determine, first of all, the completeness, location accuracy and attribute compatibility of the main land-cover classes of OSM objects in relation to the national official data collected in the database of topographic objects, which was the reference base in the conducted research. The analyses were performed for five selected counties in Poland, taking into account their diversity in terms of terrain and urbanization level, which allows them to be treated as representative samples. The data to which OSM was referred were official Polish data from BDOT10k (National Database of Topographic Objects). This study complements the previous research results in the field of quantitative and qualitative analysis of OSM data, especially in relation to the Polish territory, taking into account the diversity of land cover and development of the test areas. A novelty in the study is its comprehensive approach to assessing the quality of OSM data for the main classes of land cover in the area of a particular county. Another novelty is the analysis of these elements in relation to the indicators of sustainable development, including economic resources.

2. General Principles of OpenStreetMap

Due to the lack of open access to spatial data in the UK, in 2004 Steve Coast initiated the free, open and editable OpenStreetMap (OSM) project. The main mission of this project is to provide both finished maps and raw geodata to any user by volunteers around the world. Despite its name, OpenStreetMap is not just a map of roads. Roads account for 23.9% of all objects, and buildings are the most represented, accounting for 58.9% of all objects collected in the database [17]. OSM is based on the idea of an open social network and uses wiki technology, which in practice means that anyone can enter or edit any object in the database at any time. In addition, the database stores the history of edits to each object, so the effects of a mistaken vectorization or deliberate vandalism may be retracted.

2.1. OSM Data Structure

OSM has its own infrastructure for storing, sharing, searching and visualizing data that is not compliant with OGC (Open Geospatial Consortium) standards. OSM follows the peer production model that created Wikipedia; its aim is to create a set of map data that is free to use, editable and licensed under new copyright schemes [18]. OSM data are stored in a PostgreSQL relational database, according to the WGS84 datum (World Geodetic System 1984). Basic geometric types are used to represent the geometry, which, when combined with any labelling scheme, allow virtually any geographic object to be described. Elements are the basic components of OpenStreetMap's conceptual data model of the physical world. Elements are of three types (Figure 1):

- Nodes—represent a specific point on the earth's surface defined by its latitude and longitude. Nodes can be used to define standalone point features. For example, a node could represent a park bench or a water well.
- Ways—used to represent linear features such as rivers and roads. Ways can also represent the boundaries of areas (solid polygons) such as buildings or forests. In this case, a way's first and last node is the same. This is called a "closed way".
- Relations—multi-purpose data structure that documents the relationship between two or more data elements (nodes, ways and/or other relations).

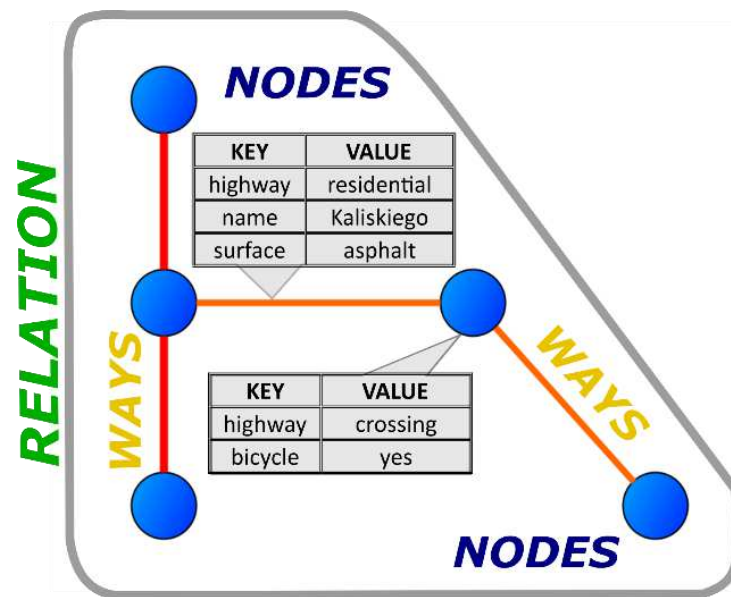


Figure 1. An example of an OSM object data structure.

A label, also called a tag, consists of a pair of expressions: “key = value” which can be equated with an attribute. For example, highway = residential defines the way as a road generally used for local traffic within settlement. The key highway = * is the main key used for identifying any kind of road, street or path. The value of the key helps indicate the importance of the highway within the road network as a whole (Figure 1).

2.2. Methods of Obtaining OSM Data

OSM data users and providers use a variety of techniques and data sources to acquire site information, including vectorization of orthoimagery, measuring a route with a hand-held GPS receiver while walking (or even biking or driving), sketching or measuring from the level of the nearest road, importing public official data, etc. [19,20]. The dependence of geometric accuracy on data acquisition methods and techniques explains the heterogeneous geometric accuracy of the data in the OSM database. In addition, the effect of heterogeneity of object acquisition techniques and devices is compounded by the effect of interpretation. Correct interpretation of the spatial position of an object, e.g., the outline of a building, especially from satellite or aerial images, requires experience, preferably supported by appropriate training.

Goodchild, an ambassador of the idea and term VGI, expressed the opinion that well-distinguishable geographic objects are less demanding in terms of training and experience for volunteer observers placing them [21]. Treating this opinion as a research hypothesis, we confirm that the location of a given spatial object varies depending on its type and the interpretive capabilities of the person editing OSM.

2.3. Methodological Aspects of OSM Data Quality Assessment

According to ISO 19157:2013 [7], spatial data quality is understood as a set of the following characteristics and attributes of the objects collected in a database:

- Geometric accuracy—this parameter describes the accuracy of determining the coordinates of the object. In practice, the preferred method of checking this parameter is to compare it with an independent source of higher accuracy.
- Thematic accuracy—describes the accuracy or certainty of the acquisition of an attribute value. Estimating the accuracy of a quantitative attribute is analogous to location accuracy (comparison with a more accurate data source).
- Currentness—describes the point in time or moment in time when the contents of the database match reality.

- Completeness—determines how exhaustive a set of objects is. It may refer to: excess, missing objects, their attributes or the relationships between them.
- Logical consistency—describes the consistency of the relationships recorded in the spatial database structure (conceptual, domain and topological).

The quality of OpenStreetMap data, and in particular its quantitative elements such as completeness and geometric accuracy, is of broad interest to potential users worldwide. The method of data collection used in OSM makes it impossible to directly apply the principles of geographic data evaluation contained in ISO 19157, which refer to the comparison of data with technical specifications. Goodchild [22] listed three alternative approaches to assessing the quality of geographic data acquired by projects such as OpenStreetMap:

- Crowd-sourcing approach—based on the assumption that redundant data is detected and corrected by users.
- Social approach—assuming minimal data validity checks by administrators.
- Geographical approach—involving the use of GIS-type programs for data quality control by checking the correctness of topologies and logical rules.

Such approaches to assessing the quality of VGI data are gaining popularity, although external assessment is still widely used, requiring access to other, usually more accurate and reliable data. Such an approach was used in this research because it allows for a comprehensive assessment of data quality, which is crucial from the point of view of potential users. However, the selection of reference data is a problematic issue. Concerns about what reference set to choose for quality control of spatial databases fed by non-cartographer volunteers were also expressed in studies [11,23,24]. Given their study, the official state resource was selected for comparison. Considering that most of the previous analyses of the quality and usefulness of OSM concerned large cities or agglomerations [25–27], the research results presented below are novel in that they include an analysis of spatial data covering not only roads and buildings but consider a broad data set consisting of the main land-cover elements for counties in Poland that are diverse in terms of terrain and degree of urbanization (thus, they do not cover only urbanized areas).

3. Materials and Methods

3.1. Source Data

3.1.1. Area of Research

The test area consists of five counties located in the territory of Poland. The counties that were selected are diverse in terms of location, terrain, area, degree of urbanization and land cover, so that they can be representative samples. The location of the analysed test areas is shown in Figure 2.

Piaseczyński County is situated in the central part of Masovian Voivodeship. It is one of the richest and best developed counties in Poland [28]. The county has an area of 621 km² and a population of 182,076 [29]. Within the county there are natural plant communities: meadows, peat bogs and forests. The County of Piaseczyński lies in the belt of the central Polish lowlands. The land use structure of the county is agricultural land: 49.2%, orchards: 10% and forest: 19.6%.

Sanocki County is located in the Bieszczady Mountains in southern Poland, with an area of 1159 km² and a population of 95,035. The county has an agricultural nature, with poorly developed other sectors of the economy [29]. Forestry and related wood industry are dominant here. More than 1/3 of the area is under nature protection.

Sokólski County is a county in the northeastern part of Poland with an area of 2055 km² and population of 66,686 [29]. Sokólski County has rolling and hilly hills ranging in elevation from 110 to 240 m above sea level. The county belongs to the area of the so-called “Green Lungs of Poland”, i.e., ecologically developed areas with great tourist potential. In total, 24% of its area is covered by forests and forest lands. The main forest complex is the Knyszyn Primeval Forest, through which the Supraśl River flows. A significant part of the county’s area is used for agricultural production. The good condition of the natural

environment and traditional farming culture in many farms contribute to the development of agriculture.



Figure 2. Location of analysed counties.

Słupski County is situated in the northwestern part of the Pomeranian Voivodeship, and has an area of 2304 km² (4th largest in Poland) [29]. It is bordered to the north by a 57 km long stretch of the Baltic coast. The relief is varied, with characteristic uplifts of terminal moraines and a specific, coastal landscape in the northern part, with dune areas reaching up to 30 m above sea level. Numerous rivers are an important element of the landscape—with the largest one, Słupia, being a well-known sea trout river. Protected forests cover 83% of the forest area.

Ostrowski County is a county located in the southwestern part of the Greater Poland Province and covers an area of 1160 km². The county of Ostrów Wielkopolski has the second largest population in the the Greater Poland Province (161,581 people) [29] and is situated in the macroregion of South Greater Poland Lowland. The county has an agricultural and industrial nature, with agricultural lands covering 64.9% of the county's area, and 28.3% of forest areas.

3.1.2. OSM Data

In the conducted OpenStreetMap data quality study, 6 main (essential) land-cover classes were used, which included linear objects (roads, railroads and river network) and area objects (buildings, surface water and forests). OpenStreetMap data were obtained from the OSM Geofabrik service [30]. The timeliness of the surveyed OSM data is 24 June 2021. Table 1 provides a description of the data used and Figure 3 shows an example visualization of the OSM data. The analysed OSM objects were selected to match the objects from the BDOT10k reference base as closely as possible.

OSM data are characterized by heterogeneous accuracy and level of detail, depending on the acquisition technique and object contour detailing, which, in turn, depend on the skill and experience of the observer. The means of obtaining OSM data are primarily measurements from handheld GPS receivers, aerial photographs and other available data sources. OSM data timeliness varies depending on volunteer activity.

Table 1. Characteristics of the OSM data analysed with the distinguishing “tag” [17].

No.	Geometric Representation	Tag	Description
1	line	highway = {motorway, trunk, primary, secondary, tertiary, unclassified, residential}	The principal tags for the road network. They range from the most to least important.
2	line	railway = rail	Standard track for passenger or freight trains
3	line	waterway = {river, stream, tidal_channel}	rivers, streams, and other watercourses with water flowing from one place to another
4	polygon	building = *	single building outline.
5	polygon	natural = water landuse = reservoir water = reservoir	Unspecified bodies of water. Typically lakes but can also be larger rivers, harbours, etc.
6	polygon	landuse = forest natural = wood	Forest or woodland. Sometimes considered to have the restricted meaning “managed woodland or tree plantation maintained by humans to obtain forest products”.

**Figure 3.** An example visualization of the OpenStreetMap database.

3.1.3. Reference Data—BDOT10k

In the conducted analyses of OSM data quality, the official spatial data of the National Database of Topographic Objects (BDOT10k) was adopted as the reference dataset. It is a vector database containing the spatial location of topographic objects along with their basic descriptive characteristics. The content and detail of the BDOT10k database correspond to those of a traditional topographic map at the scale of 1:10,000, (Figure 4).

The detailed scope of information collected in BDOT10k, organization, mode and technical standards of creating, updating, verifying and making data available are specified in the legal act [31]. The BDOT10k database is updated on a current basis after obtaining reliable data from feeders. The BDOT10k database is available free of charge for any use. Data may be downloaded free of charge via the GEOPORTAL service [32]. The validity of the available BDOT10k data used in this analysis is March 2020. Objects from the BDOT10k reference database, belonging to the same land-cover classes as the OSM data, were used for the OSM data quality analysis (see Table 2).



Figure 4. An example visualization of the BDOT10k database in Geoportal service.

Table 2. Characteristics of the BDOT10k data analysed [31].

No.	Geometric Representation	Name	Designation BDOT10k	Description
1	line	road	SKJZ	Roadway centreline segments, or portions of the roadway dedicated to vehicular traffic, that have a uniform set of attributes.
2	line	track or set of tracks	SKTR	Segments of track axles or axles of sets of tracks used for the movement of railway vehicles.
3	line	river and stream/channel	SWRS/SWKN	Sections of river and stream axes between hydrographic network nodes.
4	polygon	building	BUBD	Buildings permanently connected to the ground with foundations.
5	polygon	surface water	PTWP	Areas occupied by the waters of rivers, canals and reservoirs.
6	polygon	forest or wooded area	PTLZ	Densely wooded areas: forests, wooded parks and other wooded areas.

The BDOT10k dataset is defined in the PUWG 1992 (Państwowy Układ Współrzędnych Geodezyjnych) rectangular coordinate system, which is a coordinate system based on the Gauss–Krüger mapping for the GRS80 ellipsoid in one ten-degree zone for Poland (EPSG: 2180). The PUWG 1992 is intended for maps with a scale of 1:10,000 and smaller.

BDOT10k data are acquired through: geodetic survey, land and building registry, orthophoto vectorization or other official state registers. BDOT10K quality control is performed in accordance with the control system for data submitted to the BDOT10k resource (topology and geometry checks, semantic, syntactic and attribute checks, etc.) and is carried out in high detail.

3.2. Methodology

This part describes the methods used to assess the quality of OSM data for the test counties (Piaseczyński, Słupski, Ostrowski, Sanocki and Sokólski) in relation to the BDOT10k reference database.

The spatial data characteristics summarized in Section 3.1 show the basic sources of discrepancies between the analysed datasets, namely: different conceptual model, measurement rules and technological supervision and the control system of data transferred and stored in the OSM and BDOT10k databases. Taking into account the mentioned differences in both datasets, the research work included: the identification of corresponding objects in both sets as well as the analysis of geometric accuracy and completeness of objects

and attributes. Sections 3.2.1 and 3.2.2 describe the procedures for assessing the quality of OSM data in terms of geometric accuracy. Due to the different geometric type of the analysed land-cover objects, OSM data quality assessment for area (Section 3.2.1) and linear (Section 3.2.2) objects is detailed. Sections 3.2.3 and 3.2.4 describe the method for assessing the completeness of OSM data, and Section 3.2.5 analyses semantic and attribute accuracy of OSM. These methods take into account the heterogeneous nature of OSM data (linear and area objects). The analyses described were performed using GIS software—ArcGIS Pro and Statistica software. Due to different coordinate systems of the analysed data, the OSM database was transformed to the metric system consistent with BDOT10k—PUWG 1992 uniform for the entire territory of Poland. The use of the metric coordinate system made it possible to perform spatial analyses in GIS software in accordance with the applied methodology. The transformation between the WGS 84 geographic system and PUWG 1992 was performed in accordance with the conversion tools of ArcGIS Pro. The process of performing the transformation of the coordinate system does not affect the results obtained.

3.2.1. Analysis of the Geometric Accuracy of OSM Area Objects

In order to obtain statistical information about the geometric accuracy of OSM surface objects in comparison to the reference database BDOT10k, homologous points were used. Accuracy analysis was based on automatic measurement of corresponding vertices of OSM area objects in relation to BDOT10k—Figure 5.



Figure 5. Homologous points semi-automatically detected between the OSM buildings (red) and the BDOT10k buildings (blue).

The measurement was conducted for all analysed counties, taking into account area objects—buildings, forests and surface waters. Measurement of homologous points was performed by comparing coordinates of corresponding corners (vertices) in OSM and BDOT10k databases.

The geometric accuracy of OSM area objects is presented in terms of root mean square error (RMSE). RMSE is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. root mean square error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how these residuals are spread out. In other words, it tells us how

concentrated the data is around the line of best fit. root mean square error is commonly used in climatology, forecasting and regression analysis to verify experimental results [33].

The value of RMSE error for the determined homologous points in OSM and BDOT10k datasets was calculated according to the Equations (1)–(3):

$$RMSE_X = \sqrt{\frac{\sum_i (X_{OSM} - X_{BDOT10k})^2}{N}} \quad (1)$$

$$RMSE_Y = \sqrt{\frac{\sum_i (Y_{OSM} - Y_{BDOT10k})^2}{N}} \quad (2)$$

$$RMSE = \sqrt{RMSE_X^2 + RMSE_Y^2} \quad (3)$$

where:

X_{OSM}, Y_{OSM} —coordinates of a point from the OSM database,

$X_{BDOT10k}, Y_{BDOT10k}$ —coordinates of a point from the BDOT10k database and N —number of observations (homology points).

3.2.2. Analysis of Geometric Accuracy of OSM Linear Objects

The buffer zone method developed in research [34] was used to determine the accuracy of the location of OSM linear objects relative to the BDOT10k database. Using this method, accuracy is determined by the percentage of an OSM linear object located within the buffer zone of the corresponding linear feature from the BDOT10k reference data. The buffer zones were determined for the BDOT10k data. For each of the linear objects' class (roads, railroads and rivers), 4 buffer zones with widths of 1, 2, 5 and 10 m were delineated—Figure 6.

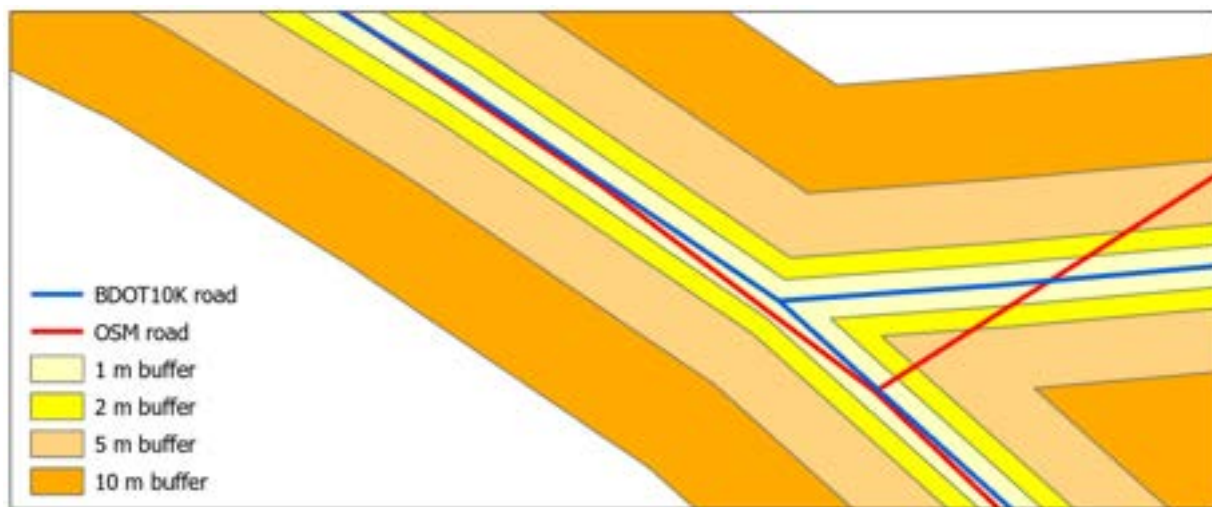


Figure 6. Created buffer zones around a linear object of the BDOT10k database.

The widths of the buffer zones were optimally selected on the basis of the literature and the accuracy of the BDOT10k data. The assumed accuracy of the BDOT10k database corresponds to the accuracy of maps in the scale of 1:10,000, which results in the accuracy of the location of objects in the database amounting to 1.5–5 m, depending on the type of object in accordance with [31]. Other geometric conditions of objects are also specified: the minimum distance between vertices is 2 m, and the accuracy of mapping angles is 1 degree.

The OSM data length was then overlaid on the designated BDOT10k buffer zones to calculate the percentage of overlap using the equation as follows:

$$Coverage [\%] = \frac{A}{B} \cdot 100\% \quad (4)$$

where:

A —length of the OSM dataset tested in the buffer.

B —total length of the tested BDOT10k dataset in the buffer.

3.2.3. Analysis of Completeness of OSM Area Objects

The completeness of OSM area objects is assessed using a method based on area ratio units [14], which calculates completeness (C) as the percentage ratio between the total area of an OSM object and the total area of the corresponding BDOT10k database object within a specific spatial unit (e.g., administrative or geometric). For this purpose, the area of the analysed counties was divided into sub-areas according to a regular grid of hexagons of an area equal to 1 km² each. The choice of this cell shape is suggested by [14], who argued that the hexagonal shape of the basic field has the advantage of approximating a circle, optimally covering the study area. As noted by [35], the area ratio method may introduce an overestimation of C due to the overabundance of data available in OSM relative to BDOT10k data. For this reason, further studies are recommended in which three additional indicators are calculated: true positive (TP), false positive (FP) and false negative (FN) rates. The TP indicator (Figure 7) represents the overlapping areas of surface objects between OSM and BDOT10k, i.e., the common areas between the datasets. The FP indicator represents OSM surface objects that do not exist in the BDOT10k dataset, and the FN indicator considers BDOT10k area objects that do not exist in the OSM dataset.

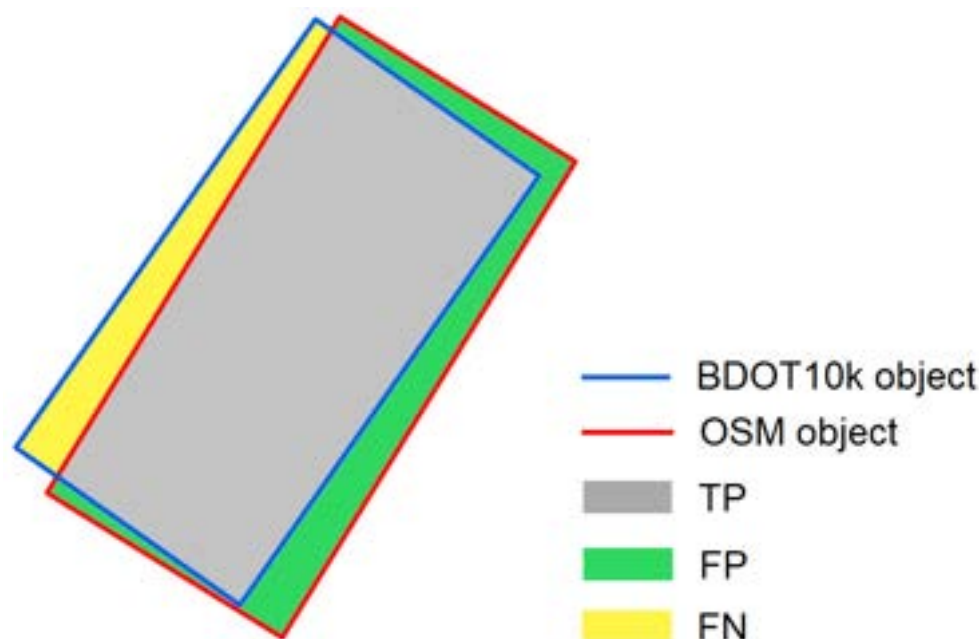


Figure 7. Characteristics of TP, FP and FN indicators.

To assess the completeness of OSM area objects, C , TP, FN and FP values were calculated for each hexagonal grid cell. The obtained results were then related to the total area of objects in the reference base, which is the BDOT10k collection. In the conducted research on OSM data completeness, such analyses were carried out for buildings, surface water and forests.

3.2.4. Analysis of Completeness of OSM Linear Objects

In this study, the completeness of roads, railroads and rivers available in the OSM database was calculated by comparing the length of a given linear feature with the length of the corresponding feature in the BDOT10k dataset (Table 3) using the Equation (5) [36]. The results are presented as percentages:

$$Completeness [\%] = \frac{D}{E} \cdot 100\% \quad (5)$$

where:

D —object length in OSM dataset.

E —total length of the corresponding object according to BDOT10k dataset.

Table 3. Statistics on the homologous pairs detected in the considered test areas.

County	Area Object	Number of Homologous Point Pairs	RMSE _X [m]	RMSE _Y [m]	RMSE [m]
Piaseczyński	buildings	956,277	1.54	1.58	2.21
	forests	32,677	4.22	4.25	5.99
	waters	20,510	2.96	2.89	4.14
Sokólski	buildings	295,224	0.99	0.94	1.36
	forests	109,130	4.20	4.13	5.89
	waters	16,317	2.47	2.51	3.52
Słupski	buildings	288,790	1.14	1.09	1.58
	forests	23,625	3.94	3.90	5.54
	waters	15,185	3.35	3.42	4.79
Ostrowski	buildings	467,641	1.70	1.68	2.40
	forests	12,627	3.85	3.71	5.35
	waters	7204	3.01	2.91	4.19
Sanocki	buildings	285,385	1.43	1.49	2.06
	forests	55,834	3.88	3.87	5.48
	waters	7646	2.73	2.80	3.91

3.2.5. Attribute and Semantic Accuracy Analysis of OSM

The analysis presents quantitative results of the accuracy of OSM database attribute values. The BDOT10k database was not considered in this analysis. The quantitative analyses show the extent to which the selected OSM object tag is informed (contains information of the mapped feature). The analysis includes the number of objects with additional tags to the main tag completed, such as NAME and others describing an additional OSM object type. Attribute tests were performed by analysing the number of objects in a given OSM class and by examining the degree of user-entered information about each object's attributes. The analysis was conducted for linear and area objects. The attribute quality assessment of OSM objects was calculated by comparing the correct number of object names as an Equation (6) [36]:

$$Attribute\ accuracy [\%] = \frac{F}{FG} \cdot 100\% \quad (6)$$

where:

F —number of objects with the informed tag in OSM dataset.

G —total number of OSM objects.

Attribute accuracy assesses the accuracy of attributes captured according to the specifications of the database.

4. Results

4.1. Geometric Accuracy of OSM Area Objects

Using the methods described in Section 3.2.1, the geometric accuracy of area objects was calculated separately for each of the five analysed counties. The obtained results of OSM area object geometric accuracy based on homologous points are presented in Table 3.

According to the results presented in Table 3, it is noticeable that the smallest values of RMSE were obtained in all the counties for building type objects. The smallest error was recorded in Sokólski County—1.36 m (295,224 homologous points). On the

other hand, the lowest value of RMSE was obtained in the Ostrowski County—2.40 m (467,641 homological points).

Objects of the forest and surface water type obtained much higher values of RMSE in the analysed areas. The lowest values on a quite similar level were obtained for objects of forest type (from 5.35 m for Ostrowski to 5.99 m for Piaseczyński County). On the other hand, the values of RMSE for surface waters were slightly more diversified—from 3.52 m for the Sokólski County to 4.79 m for the Słupski County.

4.2. Geometric Accuracy of OSM Linear Objects

According to Section 3.2.2, the accuracy of OSM linear objects' locations was determined by creating buffer zones in relation to BDOT10k objects with widths of 1 m, 2 m, 5 m and 10 m. Then, the percentage of overlap of OSM data in relation to each buffer zone was determined. Linear land-cover objects (roads, railroads and river network) in all analysed counties were included in the analysis. The resulting analysed linear objects by county are shown in Tables 4–8.

Table 4. Geometric accuracy of OSM linear objects in Piaseczyński County in relation to BDOT10k objects in buffer zones.

Buffer Zone Width [m]	Roads		Railway		River Network	
	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]
1	1786.3	48.5	52.3	36.3	79.5	9.9
2	2952.7	80.2	80.5	56.0	147.4	18.3
5	4225.1	114.7	132.3	92.0	267.4	33.2
10	4696.7	127.5	169.8	118.0	336.0	41.7

Table 5. Geometric accuracy of OSM linear objects in Sanocki County in relation to BDOT10k objects in buffer zones.

Buffer Zone Width [m]	Roads		Railway		River Network	
	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]
1	998.6	20.6	93.6	62.8	1092.9	32.5
2	1543.8	31.8	120.9	81.1	1547.5	46.0
5	1989.9	41.0	135.2	90.8	1805.0	53.7
10	2119.2	43.6	139.5	93.7	1926.9	57.3

According to the results presented in Tables 4–8, it can be seen that the geometric accuracy of linear objects varies significantly depending on the reference buffer width used. In case of the road network, the highest increase in the share of OSM data in the analysed buffer zones was recorded in Słupski county—from 14.3% for 1 m buffer width to 50% for 10 m buffer width. On the other hand, in the county of Piaseczyński for 1 m width of the buffer there was a significant share of the OSM data—48.5%. With increasing buffer width, the share of OSM data relative to BDOT10k increased up to 127.5% for 10 m buffer width. Increase in share of OSM data above 100% indicates data overabundance, i.e., the OSM database contains more data than BDOT10k. As far as the remaining counties (Sanocki, Sokólski and Ostrowski) are concerned, the share of OSM data in relation to BDOT10k was at a similar level with the increase in reference buffer zones, but in no case did it exceed 70%.

Table 6. Geometric accuracy of OSM linear objects in Sokólski County in relation to BDOT10k objects in buffer zones.

Buffer Zone Width [m]	Roads		Railway		River Network	
	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]
1	1636.5	23.4	77.3	44.0	92.1	15.9
2	2870.4	41.0	121.0	68.9	173.3	30.3
5	4338.9	61.9	147.4	83.9	317.1	54.8
10	4848.8	69.2	153.5	87.4	385.9	66.7

Table 7. Geometric accuracy of OSM linear objects in Słupski County in relation to BDOT10k objects in buffer zones.

Buffer Zone Width [m]	Roads		Railway		River Network	
	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]
1	1167.1	14.3	77.4	38.5	114.7	8.1
2	2022.4	24.8	117.3	58.5	216.9	15.2
5	3337.6	40.9	152.9	76.2	429.4	30.2
10	4082.0	50.0	167.4	83.4	599.4	42.1

Table 8. Geometric accuracy of OSM linear objects in Ostrowski County in relation to BDOT10k objects in buffer zones.

Buffer Zone Width [m]	Roads		Railway		River Network	
	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]	OSM Data Length [km]	Coverage [%]
1	1216.3	24.9	89.0	35.3	70.1	6.0
2	1960.3	40.1	138.2	54.8	120.6	10.3
5	2733.5	55.9	238.5	94.6	196.4	16.8
10	2996.8	61.3	285.0	113.1	230.5	19.8

The analysis of the railroad network revealed that the situation is slightly different than in the case of the road network. The highest share of OSM data for the reference buffer width of 1 m was recorded in Sanocki county—62%. As the width of the buffer zones increased, the share of OSM data increased slightly—up to 93.7% for a 10 m wide reference buffer. For the Piaseczyński and Ostrowski Counties, the situation is very similar. The share of OSM data in individual buffer zones increases significantly with the increasing width of the reference buffer—from about 30% for 1 m width to over 113% for 10 m width. In the case of the Sokólski and Słupski Counties, the share of OSM data in relation to the intervals of the reference buffer zones is similar and does not exceed 90% in the case of the 10-m-wide buffer.

As for the river network, the share of OSM data in particular buffer zones is the smallest in comparison with the other analysed objects of the road and railroad network. The smallest share of OSM objects of the river network in the analysed buffers was recorded in the Ostrowski County—from 6% for the 1 m buffer to only 19.8% for the 10 m buffer. For Sanocki County, the share of OSM data in relation to BDOT10k data for the width of reference for a buffer of 1 m is the largest share from all analysed counties—32.5%. With the increase in the buffer width, this share increases to 57.3% for 10 m width of the reference buffer. On the other hand, in the Sokólski County, the share of OSM data in relation to BDOT10k reaches the highest value for the reference buffer of 10 m width—66.7%. In the

case of Piaseczyński and Słupski Counties, the obtained values of the accuracy of the OSM data position in relation to the reference buffer zones are similar and amount from about 8% for the buffer zone of 1 m to about 42% for the buffer zone of 10 m width.

4.3. Completeness of OSM Area Objects

The methods described in Section 3.2.3 were used to calculate the completeness of OSM area objects in individual cells of the hexagonal grid. The analysed area was divided into basic fields in the form of a hexagonal grid of 1 km². Finally, 60 thematic maps were developed visualizing the spatial distribution of C, TP, FP and FN indices in the analysed counties. Due to the large number of maps, the obtained thematic maps are presented for the two analysed counties, which represent the greatest diversity of the results. For Piaseczyński County (Figure 8) the developed maps are presented in Figure 9 and for Sokólski County (Figure 10) the obtained maps are shown in Figure 11.

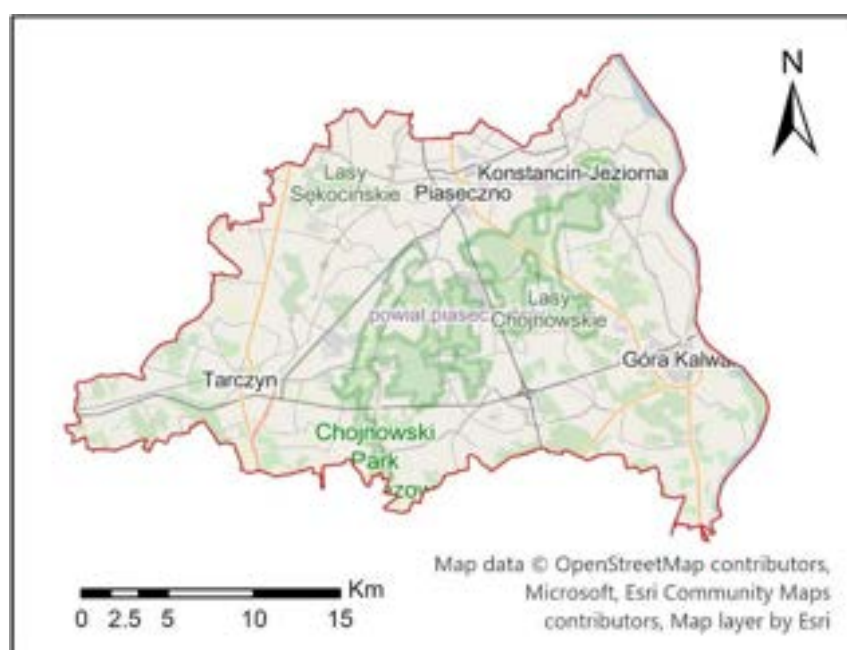
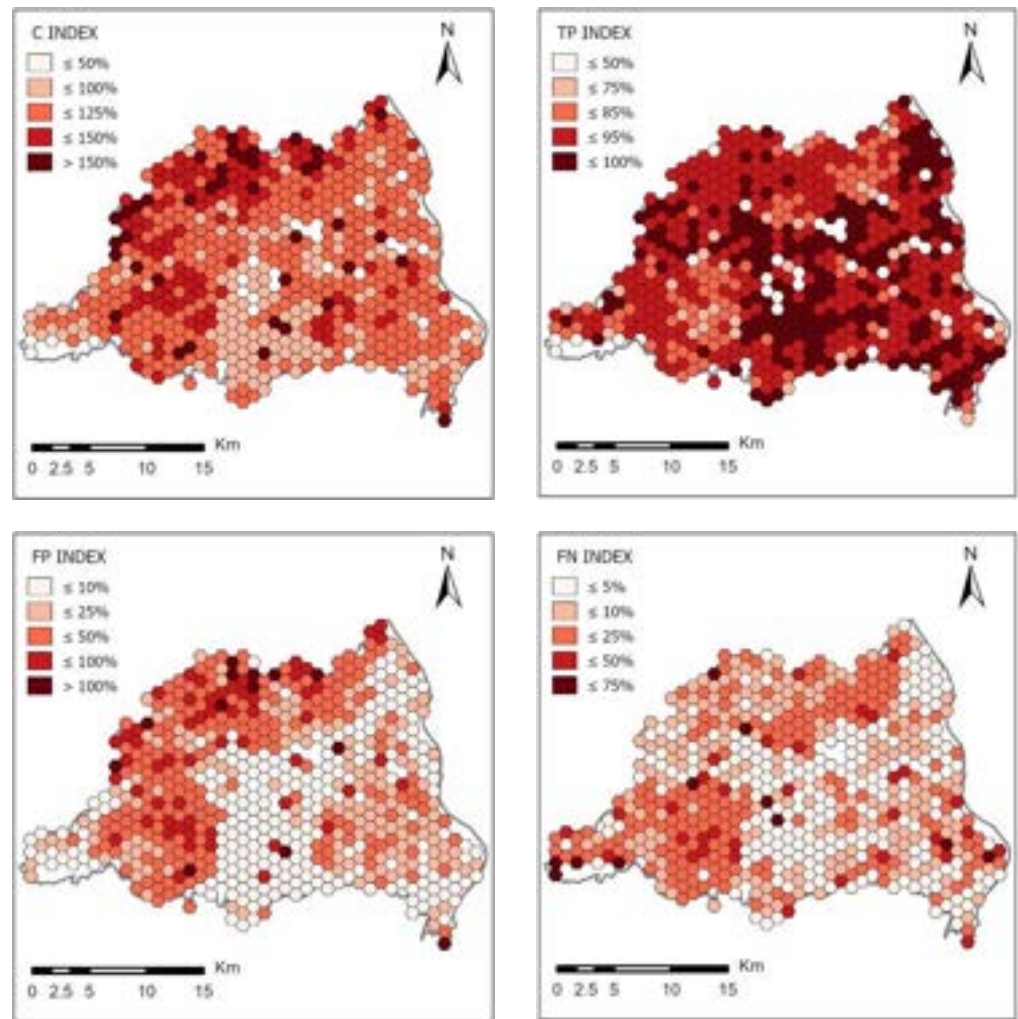


Figure 8. Piaseczyński County—overview map.

The set of maps presented in Figures 9 and 11 shows the spatial distribution of the calculated indices of completeness of OSM area objects in comparison with the BDOT10k reference database in two counties: Piaseczyński and Sokólski. The ranges of values for each of the indicators were determined according to the Natural Breaks algorithm. Jenks Natural Breaks classification (or optimization) is a data classification method designed to optimize the distribution of a set of values into “natural” classes. The range of classes consists of elements with similar characteristics that form a “natural” group in the data set [37].

(a) Buildings



(b) Forests

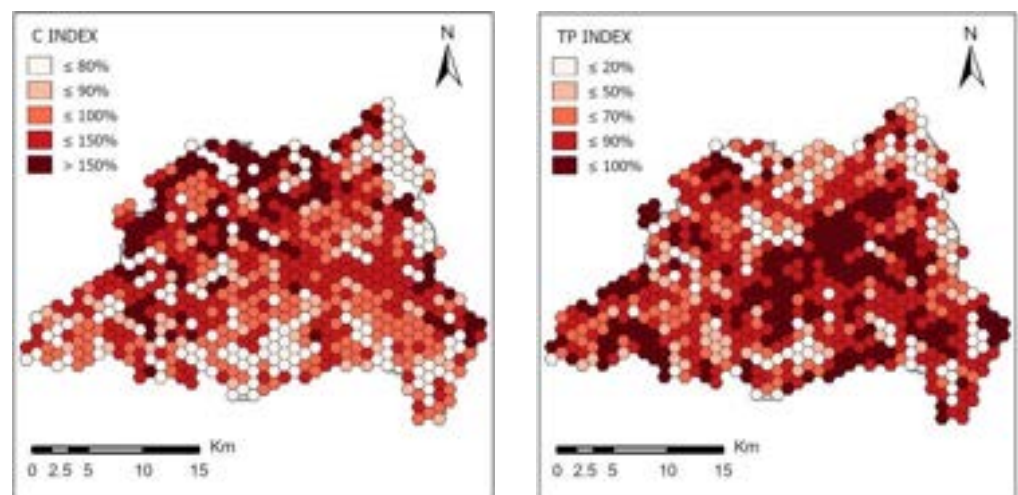


Figure 9. Cont.

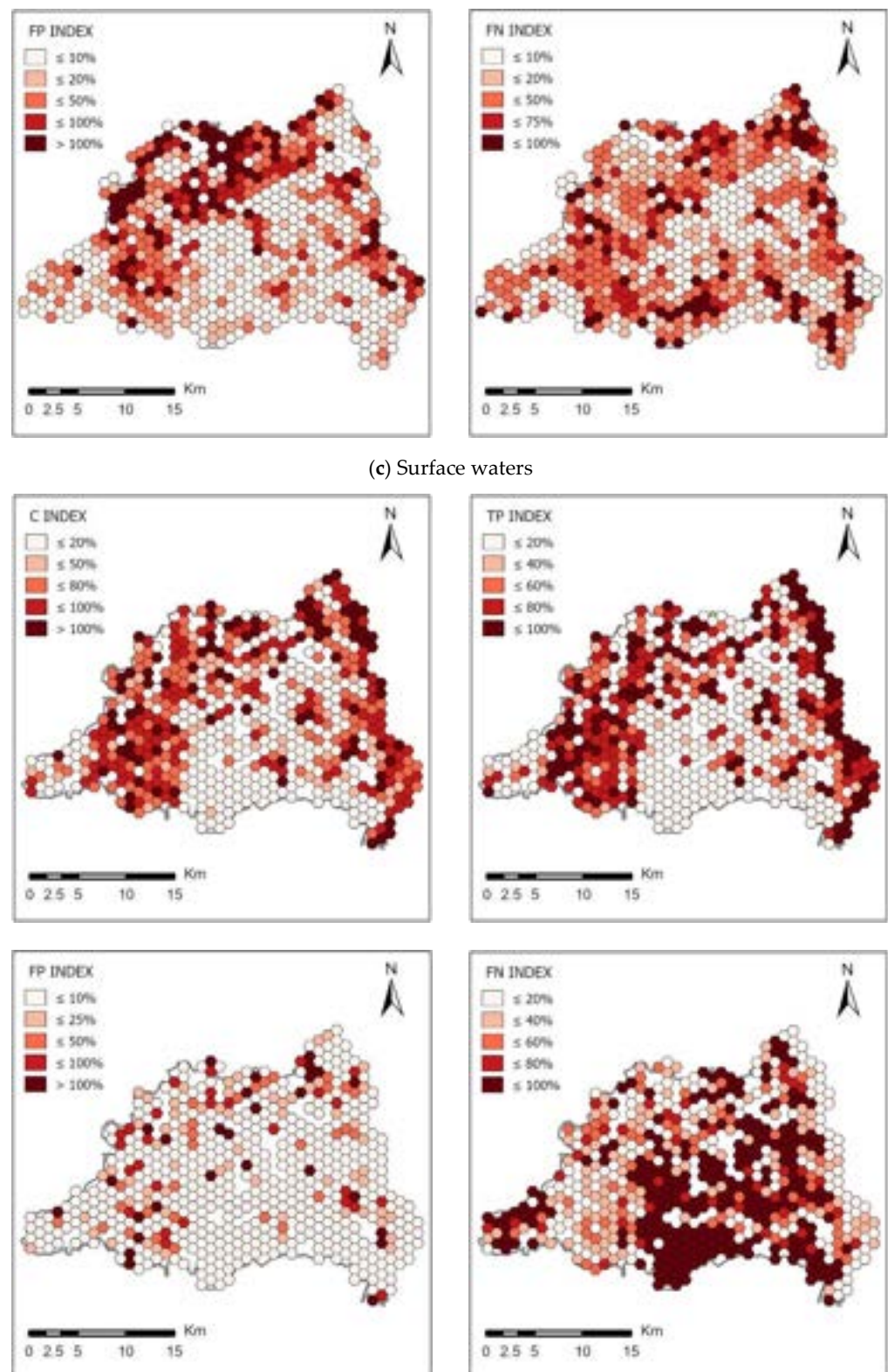


Figure 9. Completeness analysis of OSM area objects in Piaseczyński County for indicators C, TP, FP, and FN: (a) buildings, (b) forests, (c) surface waters.

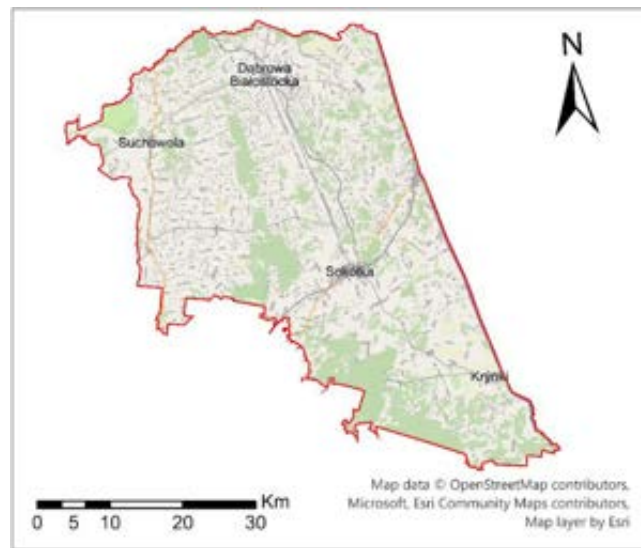


Figure 10. Sokólski County—an overview map.

(a) Buildings

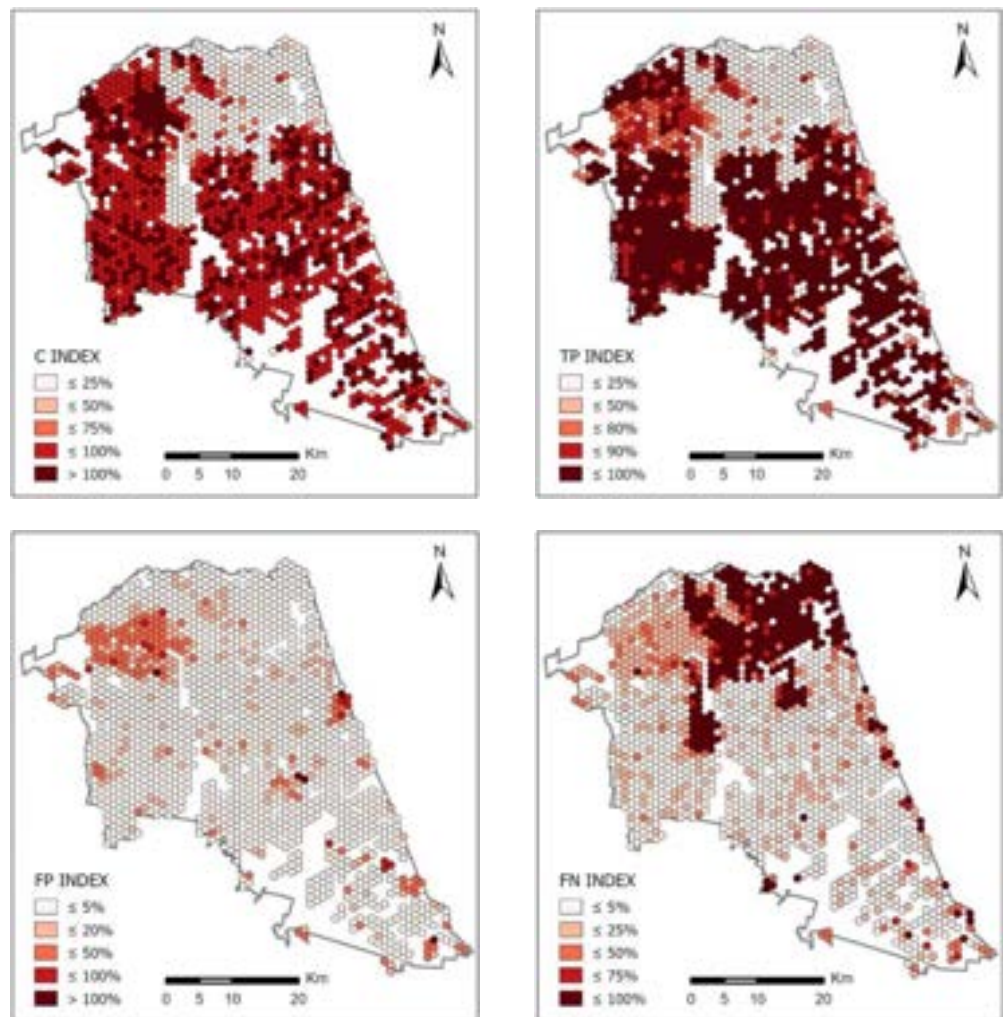
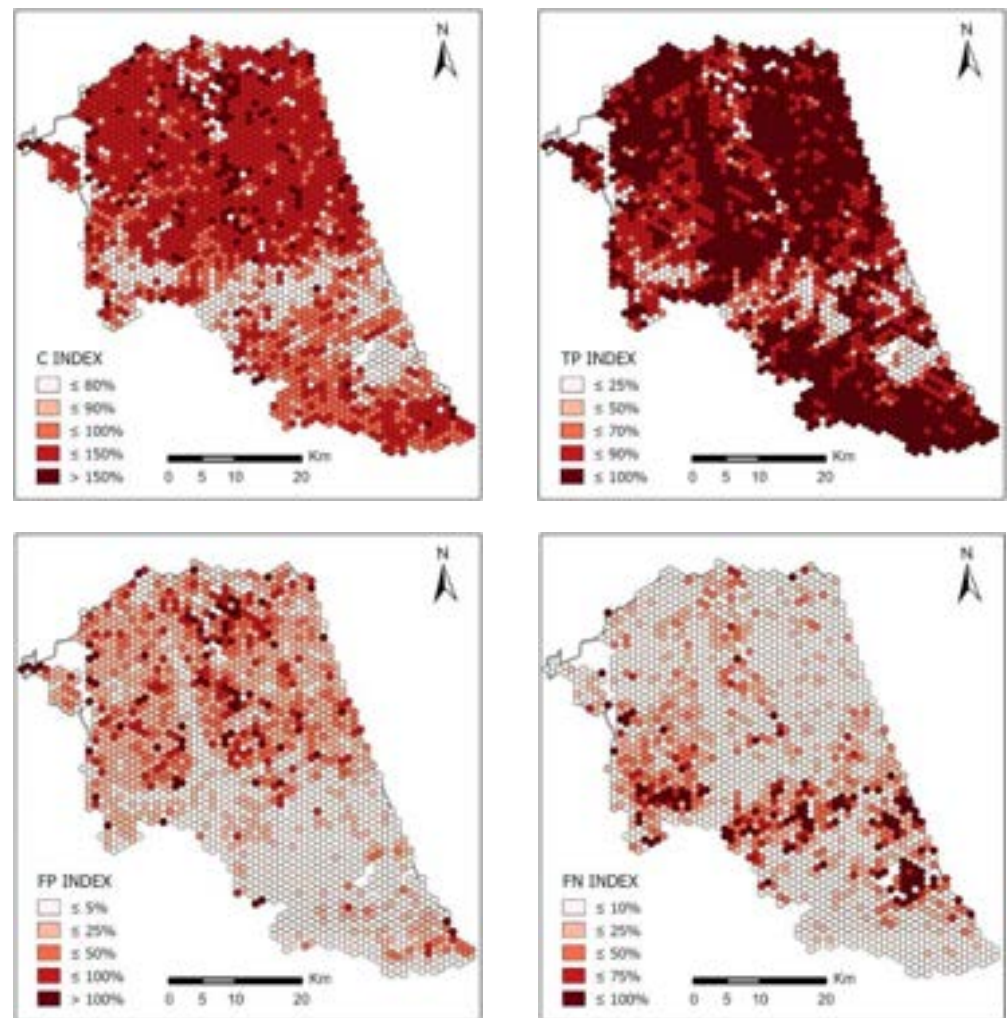


Figure 11. Cont.

(b) Forests



(c) Surface waters

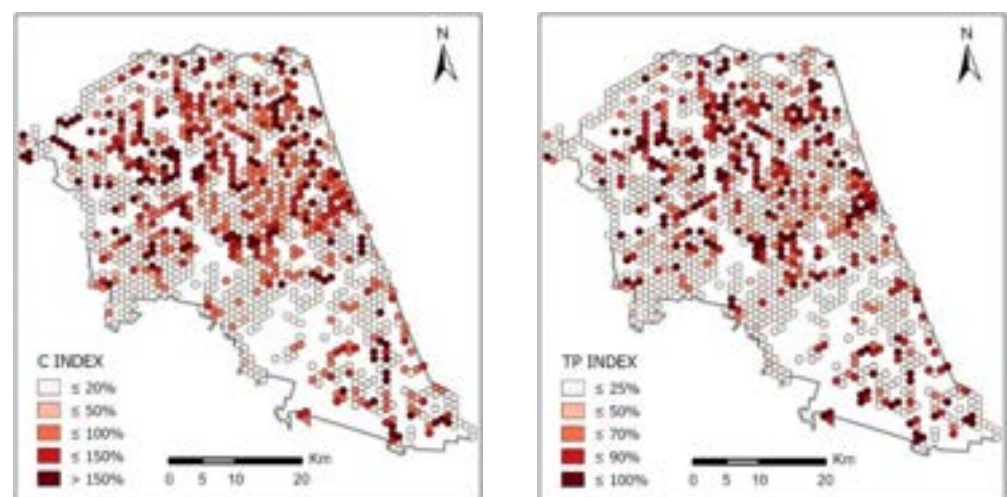


Figure 11. Cont.

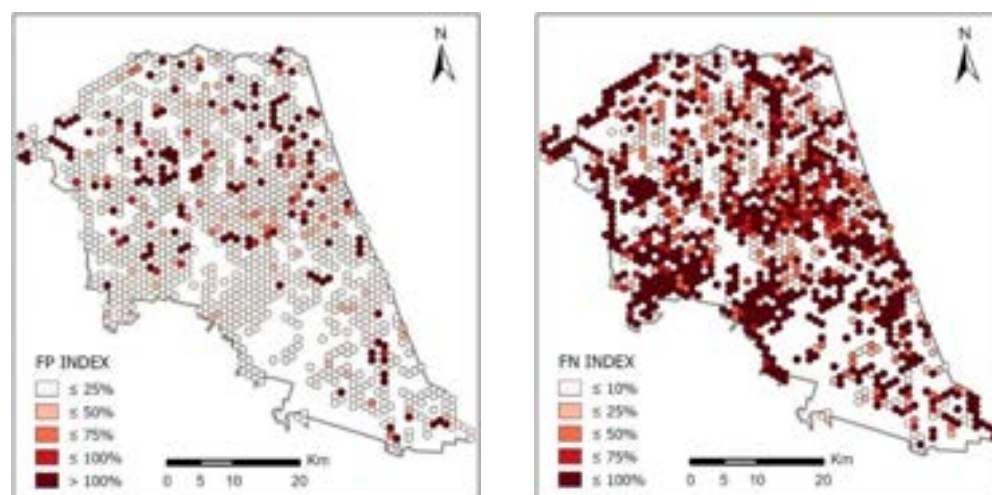


Figure 11. Completeness analysis of OSM area objects in Sokólski County for indicators C, TP, FP and FN: (a) buildings, (b) forests and (c) surface waters.

This classification method seeks to minimize the average deviation from the class mean while maximizing the deviation from the means of the other groups. This method reduces the variance within classes and maximizes the variance between classes.

The mean values of OSM area objects' completeness indices for all counties analysed with respect to the base field of the hexagonal grid are presented below in Table 9.

Table 9. Average values of OSM area objects' completeness indices in the analysed counties.

Class of Area Objects	Index	Average Value of the Index in the County [%]				
		Piaseczyński	Sokólski	Słupski	Ostrowski	Sanocki
buildings	C	115.1	79.5	66.5	83.5	111.2
	TP	89.2	74.3	52.8	67.9	92.3
	FP	25.7	5.3	14.4	15.6	18.9
	FN	10.8	25.4	47.0	32.8	7.3
forests	C	171.2	180.6	80.2	78.4	86.3
	TP	69.5	82.8	72.7	73.3	79.9
	FP	102.3	48.3	7.4	5.6	6.4
	FN	30.4	16.4	27.3	27.1	20.1
waters	C	118.8	154.7	112.6	35.2	41.7
	TP	42.7	32.1	34.6	25.4	29.6
	FP	76.2	122.4	78.3	10.6	11.2
	FN	57.3	67.7	65.7	75.9	70.3

The spatial distribution of the completeness index C for buildings in the analysed counties reaches the highest values in urbanized and densely built-up areas. This index in highly urbanized areas in all cases reaches values above 100%. The highest average values of index C were noted in Piaseczyński County—on average, the completeness index C here equals 115%, and in some grid cells it reaches values above 400%. The lowest value of the C indicator for buildings was observed in Słupski County—about 67%. The total area of the TP index for buildings (i.e., buildings present in both BDOT10k and OSM datasets) is approximately 89% of the total area of BDOT10k buildings on average in Piaseczyński County. The highest value of the TP index for buildings was recorded in Sanocki county (92.3%). In the Słupski County, on the other hand, this index averaged 53%—the lowest value in all the analysed counties. In all cases, the highest TP values were achieved in urban areas, where high values of index C (close to 100%) were obtained. As for the FN indicator (i.e., buildings mapped in BDOT10k but not in the OSM data set), the highest value was obtained in Słupski County—on average 47% of the total BDOT10k area, while the lowest value was obtained in Sanocki County: 7% of the total BDOT10k area. On the other hand, the total FP area (i.e., buildings mapped in OSM but not in the BDOT10k data

set) was on average 26% of the total BDOT10k area in Piaseczyński County (the highest value) and 5% in Sokólski County (the lowest).

In the case of forests, the highest C index was recorded in Sokólski County (180.6%). An equally high value of this indicator was calculated for Piaseczyński County: 171.2%, whereas the lowest C index was found in Ostrowski county: 78.4%. In all analysed counties the TP index was at a similar level, but the highest value was achieved by Sokólski County (82.8%) and the lowest by Piaseczyński County (69.5%). The FP index reached the highest value for the Piaseczyński County—102.3%. Much lower values were achieved in the three analysed counties: Słupski, Sanocki and Ostrowski (7.4%, 6.4% and 5.6%, respectively). The highest values of FN were obtained for three counties—Piaseczyński (30.4%), Słupski (27.3%) and Ostrowski (27.1%). On the other hand, Sokólski County had the lowest FN: 16.4%.

The completeness index C for surface waters obtained the highest value for the Sokólski County: 154.7%. On the other hand, the lowest value was obtained in Sanocki County (41.7%) and Ostrowski County (35.2%). The TP index obtained the highest value for Piaseczyński County (42.7%), while the lowest for Ostrów Wielkopolski County (34.6%). The value of the FP index varied significantly between the analysed counties—the highest value was achieved by Sokólski County (122.4%), while the lowest by Sanocki County (11.2%) and Ostrowski County (10.6%). The FN index for the analysed counties was at a similar level, but the highest value was calculated for Ostrowski County—75.9%, and the lowest for Piaseczyński County: 57.3%.

4.4. Completeness of OSM Linear Objects

As far as linear objects are concerned, OSM data completeness was assessed by comparing the length of OSM data to the length of corresponding objects in the BDOT10k reference database. The results are presented as a percentage in Table 10.

Table 10. Completeness of OSM linear objects in relation to the BDOT10k reference database for each county.

County	Line Object Class	Length in OSM Base [km]	Length in BDOT10k Database [km]	Completeness [%]
Piaseczyński	roads	5730.3	3683.1	155.6
	railways	177.3	143.9	123.2
	rivers	461.5	806.4	57.2
Sokólski	roads	5872.9	7008.0	83.8
	railways	157.1	175.6	89.5
	rivers	578.2	764.3	75.7
Słupski	roads	5030.0	8625.9	58.3
	railways	179.0	200.7	89.2
	rivers	951.062	1423.7	66.8
Ostrowski	roads	3525.1	4886.3	72.1
	railways	292.0	252.0	115.9
	rivers	286.8	1166.2	24.6
Sanocki	roads	2421.7	4857.2	49.9
	railways	144.9	148.9	97.3
	rivers	2147.2	3362.4	63.9

According to the results presented in Table 10, it can be seen that the completeness of OSM linear objects varies considerably by objects type and by the county analysed. As far as roads are concerned, the highest completeness rate was found in Piaseczyński County (155.6%).

The same applies to railroad OSM facilities, for which the highest completeness rate was again found in Piaseczyński County: 123.2%. On the other hand, the lowest index was noted in Słupski County (89.2%). The remaining counties achieved values ranging from 89.5% to 115.9%.

The highest completeness rate for river-type OSM objects was obtained for Sokólski County: 75.7%. On the other hand, the drastically lowest value was obtained for Ostrowski County—only 24.6%. For the remaining counties the obtained index of completeness was quite similar—from 57.2% to 66.8%.

4.5. Semantic and Attribute Accuracy in OSM Database

Tests of attribute accuracy of OSM were performed by analysing the number of objects in OSM database and by checking the degree of information entered by the user concerning particular attributes of a given object. The analysis was performed for linear and area objects in all counties. The attribute to be verified was the NAME key. Additionally, for buildings, the TYPE attribute was also evaluated, denoting information about the type of a given building. All values for buildings other than “yes”, indicating a specific building type and entered values for the amenity key, were included in the analysis (see Table 11).

Table 11. Analysis of the attribute accuracy of OSM data.

County	Object Class	Objects	Fields	Informed	Non-Informed	Ratio %
Piaseczyński	roads	28,924	NAME	11,373	17,551	39.1
	rivers	2495	NAME	429	2066	17.0
	railways	419	NAME	64	355	15.0
	buildings	148,165	NAME	1106	147,059	0.7
			TYPE	15,849	132,316	10.7
	waters	2766	NAME	39	2727	1.4
forests	7716	NAME	95	7621	1.3	
Sokólski	roads	9056	NAME	794	8262	8.7
	rivers	723	NAME	209	514	29.0
	railways	168	NAME	0	160	0
	buildings	49,197	NAME	211	48,986	0.4
			TYPE	938	48,259	2.0
	waters	1460	NAME	1	1459	0
forests	11,631	NAME	15	11,616	0.1	
Słupski	roads	10,575	NAME	2906	7669	27.5
	rivers	1209	NAME	388	821	32.0
	railways	179	NAME	0	179	0
	buildings	40,306	NAME	495	39,811	1.2
			TYPE	11,482	28,824	28
	waters	1078	NAME	54	1024	5.0
forests	1556	NAME	9	1547	0	
Ostrowski	roads	8829	NAME	3018	5811	34.2
	rivers	281	NAME	165	116	58.7
	railways	473	NAME	0	473	0
	buildings	76,572	NAME	522	76,050	0.7
			TYPE	20,488	56,084	26.8
	waters	581	NAME	68	513	11.7
forests	729	NAME	12	717	1.6	
Sanocki	roads	8006	NAME	1542	6464	19.0
	rivers	4754	NAME	509	4245	11.0
	railways	307	NAME	30	277	10.0
	buildings	45,528	NAME	393	45,135	0.8
			TYPE	4562	40,966	10
	waters	346	NAME	12	334	3.5
forests	1209	NAME	14	1195	1.0	

According to the results presented in Table 11, it can be seen that the degree of attribute accuracy of OSM database is very low. The highest degree of compatibility was achieved for rivers in Ostrowski county (58.7%) for the attribute “NAME”. A rate of 0% was recorded

for railroads in the Sokólski, Słupski and Ostrowski counties. As for roads, the highest degree of attribute compatibility of OSM database was achieved in Piaseczyński County (39%), and the lowest in Sokólski County (8.7%). For rivers, the relatively highest results were achieved in Ostrowski County, and the lowest ones were recorded in Sanocki County: 17%. For railroads, the highest rate was recorded in Piaseczyński County: 15%. When analysing the attribute compatibility of buildings in OSM, the information degree of two attributes "NAME" and "TYPE" was examined. For the "NAME" attribute, the highest rate was achieved in Słupski County (1.2%), and the lowest rate was achieved in Sokólski County (0.2%). The situation was similar for the "TYPE" attribute: the highest index was recorded for the Słupski county (28%), and the lowest for Sokólski county (2.0%). In the case of rivers, the highest compatibility index was indicated in Ostrowski County (11.7%), and the lowest in Sokólski County (0%). In the case of forests, the attribute compatibility index remained in all counties at a fairly similar level—about 1%, but it reached the lowest in Sokólski County (0%) and Słupski County (1%).

5. Discussion

In the conducted research, the geometric accuracy, completeness and semantic and attribute accuracy of OSM linear and area objects, representing the main land-cover elements, i.e., buildings, forests, surface water, transportation network and water network, were analysed.

5.1. Geometric Accuracy of OSM Area Objects

The analysis of the geometric accuracy of area objects revealed that the highest accuracy of location was achieved by buildings—the average error of RMSE in this group was 1.92 m. The best results were achieved for counties with a high level of urbanization. In case of the Piaseczyński administrative district the achieved accuracy of RMSE of homological points in comparison with other results was not the highest, but it should be emphasized that the number of investigated pairs of homological points in this county was exceptionally big—even three times bigger than in other counties, which might have influenced the obtained results. The lowest results of accuracy of OSM surface objects' position were achieved for forests: on average the RMSE error was 5.65 m.

5.2. Geometric Accuracy of OSM Linear Objects

Analysing the location of linear objects, it was noted that the highest accuracies were achieved for such objects as roads and railways. As the width of the buffer zone increased from 1 to 10 m, the share of OSM objects in the given buffer also increased significantly. The highest accuracy along with the highest share of objects was recorded in Piaseczyński County, which is characterised by a dense and developed road network, while the lowest accuracy was recorded in Sanocki County, with agricultural and forest structure, where the road network is poorly developed. Quite high accuracy was also noted for railroads: as the buffer zone increased up to 10 m in the analysed counties, the share of railroads oscillated around 100%. The lowest results were obtained for the river network: from 6% for 1 m buffer (Ostrowski County) to maximum 67% for 10 m buffer (Sokólski County).

5.3. Completeness of OSM Area Objects

As far as spatial data completeness analysis is concerned, the results obtained were quite diverse and depended on the type of object and analysed county. For buildings, the highest OSM data completeness values were obtained in urbanized areas of the studied counties (cities and built-up areas). The lowest completeness values were found in the suburbs of the counties and in agricultural areas. For buildings in built-up areas, over-completeness was often recorded, i.e., the number of OSM buildings significantly exceeded the number of BDOT10k buildings. Additionally, the calculated TP index, showing the degree of overlap between OSM and BDOT10k objects, reached the highest values for areas, with the degree of completeness oscillating around 100%. The FP index, which informs

about OSM surface objects that do not exist in the BDOT10k data set, also achieved the highest values for highly urbanized areas, which resulted directly from the high over-completeness of OSM data. Finally, the highest values of the FN index were achieved for areas where the degree of data completeness was the lowest. In these cells, the majority were objects that were not present in the OSM database, although they existed in the BDOT10k database.

5.4. Completeness of OSM Linear Objects

The analysis of the degree of completeness of OSM linear objects in relation to the BDOT10k reference database for individual counties revealed that the transport network in most of the studied counties achieved the highest results, including over-completeness (numerically, the OSM base exceeds the BDOT10k base) for counties with a high degree of urbanization and a developed transport network (Piaseczyński and Ostrowski Counties). In the case of the river network, the lowest completeness index (up to 75.7%) was recorded in the Sokólski County.

5.5. Semantic and Attribute Accuracy in the OSM Database

The average attribute accuracy index obtained for OSM linear and surface objects was only 11.7%. The highest accuracy values were obtained for road network, rivers and buildings in developed counties that were also popular among users: Piaseczno, Ostrowski and the coastal county of Słupsk. The lowest indices were obtained for railroads, forests and surface waters. The quantitative results show that the main tag of each type of analysed OSM objects is mostly informed, while the secondary attributes are rarely informed. The obtained results indicate the need to complete information about most of the objects in the OSM database—according to the analyses performed, there is no information about the name and type of most of the OSM facilities. Lack of information value concerning buildings and roads may be a serious obstacle in using the OSM database for many spatial analyses.

5.6. Comparison of OSM and BDOT10k Data with Orthophotomap

Another element of the OSM data quality assessment was the comparison of objects from the OSM and BDOT10k databases with the actual terrain situation visible on orthophotomap updated to 2020 from Geoportal service. For this purpose, about 20 objects of the analysed geometric types presented on the orthophotomap were selected in random places of each county. These objects were then compared with corresponding objects in the OSM and BDOT10k databases. An example of the objects identified in the OSM and BDOT10k databases on the background of an orthophotomap is presented in Figure 12 below.

As a result of the analysis, it was found that the greatest differences were noted for buildings (outline shift in relation to the analysed databases) and forests (defining the contour of the forest was subject to the interpretation skills of the OSM and BDOT10k database editors). Additionally, it was found that the large over-completion of OSM database objects is mainly due to the entry of a given building in the OSM database and its absence in the BDOT10k database, which is updated relatively less frequently than the OSM database. Such cases occurred in single cells of the analysed hexagonal network, which led directly to a high completeness rate. In addition, it was observed that the location and outline of OSM buildings is more generalized than BDOT10k objects, which retain considerable detail of the building shape. However, as far as forests and surface waters are concerned, the OSM database retains a higher level of contour detail than the BDOT10k database.

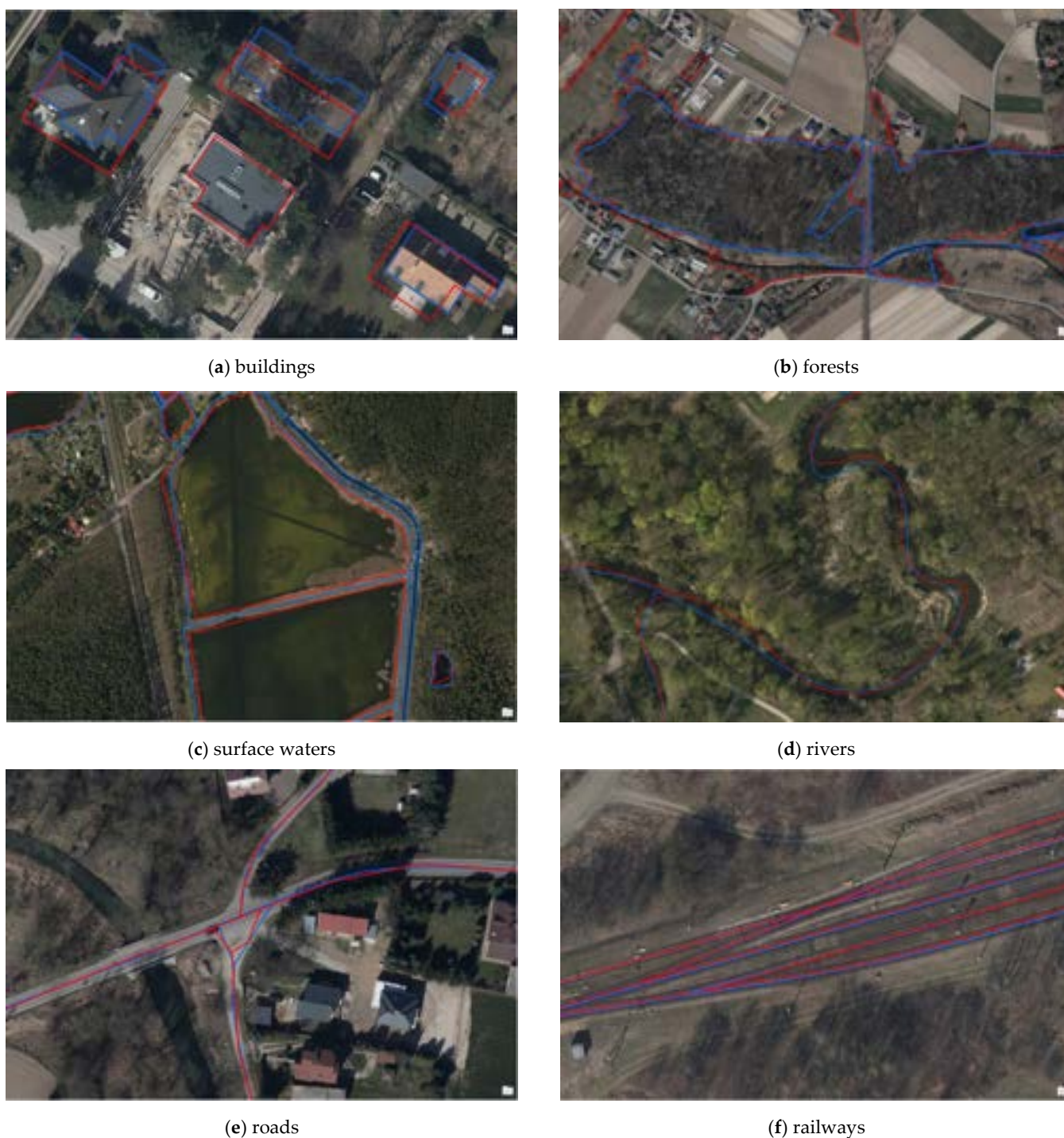


Figure 12. Location of the analysed OSM database objects (red) and BDOT10k (blue) on the background of the current orthophoto: (a) buildings, (b) forests, (c) surface waters, (d) rivers, (e) roads and (f) railways.

The obtained differences between the quality of OSM data in comparison with the BDOT10k reference database are certainly due to several reasons. One of them is the discrepancies in the source images and their shift in relation to reality. Many companies and institutions provide imagery to help create OpenStreetMap. In Poland, it is possible to use the data available on the official Geoport service. They are correctly calibrated for the territory of Poland. These images are used when updating BDOT10k. The available aerial photos or satellite imagery from sources other than Geoport are in a large proportion of

cases shifted compared with reality. In this case, the person editing OSM should suggest GPS traces. The BDOT10k database is regularly updated, while the objects in the OSM database are introduced by users on an ongoing basis. Updating and verification of BDOT10k data sets for a selected area (usually a powiat) is a long-term process subject to strict official regulations. For this reason, there are visible differences in the quality of OSM data, which make it over-completed in relation to the BDOT10k and leads to differences in the mutual spatial position (FN and FP Indices). It should also be emphasized that in the BDOT10k database, the classification of objects is strictly defined by legal regulations corresponding to the detail on a scale of 1:10,000 [31]. There are no such regulations in the assessment of OSM objects, which leads to a fairly large content for users to operate on. An example would be building mapping. In the BDOT10k database, buildings are defined as “construction objects permanently connected with the ground, having foundations separated from the space by means of building partitions (i.e., walls and covers), i.e., enclosed with walls on all sides and covered with a roof, with or without a basement with built-in house connections”. In the case of the OSM base, the building is the outline of a single building created for each complex or “block” that may be associated with one single-family house or more complex buildings. In addition, the outlines can be very simplified outlines, or very closely match the shape of the building. In the case of introducing a building to OSM from satellite imagery, one should try to recognize the geometry of the building next to the ground and not the course of the roof. The OSM mapper community in Poland has a total of over 28,000 members, of which an average of more than 200 members are constantly active on a daily basis (data from December 2021) [38]. Special activity of people editing OSM is visible in the central part of Poland (Piaseczyński county) and the eastern part (Sokólski and Sanocki counties). The remaining parts of Poland, which include the Ostrowski and Słupski counties, show minimal or no activity in recent months [39]. Such a heterogeneous structure of the OSM community and its differentiation between counties affects the quality of OSM data and its level of topicality.

Some people editing the OSM database in Poland independently import objects from selected state registers (e.g., BDOT10k, the Register of Places, Streets and Addresses). According to available data these are mainly address points and outlines of buildings [40]. The highest rate of imported objects to OSM database concerns mainly the central part of Poland. For the analysed areas address points were imported, which is not included in the OSM data quality analysis. In the case of data on buildings and land-cover elements, such an import was not made in the analysed counties [40].

5.7. OSM Data Quality in Relation to Economic Development

From the point of view of economics, one of the main tasks of local government is the equitable distribution of public goods and the creation of conditions for socio-economic development, which promotes the implementation of the objectives of Agenda 2030. Therefore, in the next step, the average indicators of the quality of OSM data in each county were recalculated as the arithmetic mean for the obtained values and compared with the income per capita in the county in 2020 [29]. The obtained results are presented in Table 12.

Table 12. Average indicators of OSM data quality compared with per capita income of budgets in the analysed county.

County	Average OSM Data Quality Index [%]				Income Per Capita (PLN)
	Linear Objects	Area Objects	Attributes	Average	
Piaseczyński	88.3	75.8	12.2	58.8	1638.5
Ostrowski	57.6	44.3	19.1	40.3	1268.3
Sanocki	62.5	47.9	7.9	39.4	1337.2
Sokólski	68.5	74.1	5.7	49.4	1539.2
Słupski	55.8	55.0	13.4	41.4	1381.2

The income of county budgets consists of: (1) own income, (2) subsidies, (3) general subvention and (4) funds for subsidizing tasks. The calculated Pearson correlation coefficient between the average indicator of data quality and income per capita in the county was 0.96. This means that the wealthier the county, the higher the indicator of OSM data quality. A graph of the analysed values is shown in Figure 13.

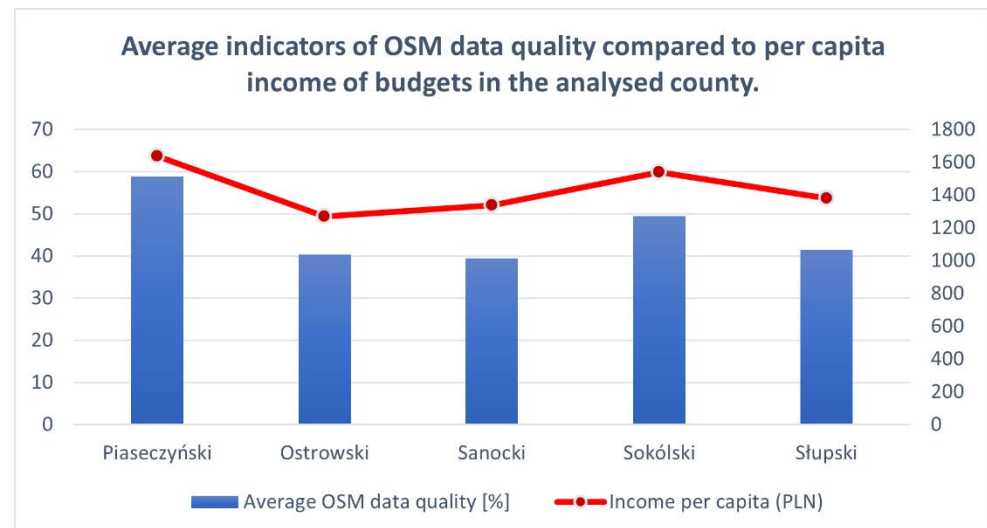


Figure 13. Average indicators of OSM data quality compared with per capita income of budgets in the analysed counties.

Indeed, counties with high per capita income showed the highest quality OSM data. This is probably because there are not only many more OSM objects in relatively developed regions of Poland, but also more users with high incomes and better Internet access.

OpenStreetMap can be a true data source for measuring SDG metrics that require geographic data [41]. Taking into account the demonstrated data quality, OpenStreetMap objects can be used to measure sustainable development goals, which mainly include goal 11 (make cities and human settlements inclusive, safe, resilient and sustainable) and goal 15 (protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification and halt and reverse land degradation and biodiversity loss). For example, the conducted analyses show a relatively good quality of buildings for all analysed poviats, which is taken into account by the SDG Indicator 11.7.1—the average share of built-up area in cities that is open to public use.

6. Conclusions

The presented results of the analysis of the comprehensive assessment of the quality of OSM data in comparison with the official reference database of spatial data BDOT10k clearly demonstrate that the obtained results largely depend on the geometric type of objects analysed and the characteristics of the test counties, including their economic development.

The obtained results confirm that the best quality indicators of OSM data were achieved for objects of quite easily recognizable and interpretable characteristics and location, i.e., buildings and transport network. On the other hand, the lowest values were achieved for objects for which defining the range and type may be a problem for a non-professional user of spatial databases—mainly forests and water network. In addition, the nature of the studied areas influenced the obtained results. The best results of spatial OSM data quality were obtained for highly urbanized areas with developed infrastructure and high per capita income ratio. The degree of coverage of line and area objects of OSM with BDOT10k amounted to 82% in the Piaseczyński county and 71.3% in the Sokólski county, respectively. The worst was in urban outskirts and low urbanized areas with low-income ratio. The obtained results show that the lowest average value of coverage of linear and

surface objects' OSM in relation to the reference database BDOT10k was obtained in the counties: Słupski 55.4%, Sanocki 55.2% and Ostrowski 51%. This is mainly due to the interest of users in a given area and the frequency of introducing new OSM spatial objects. It is also worth noting that in highly urbanized counties there was often an over-completion of data (the number of OSM data significantly exceeded the number of BDOT10k data). In less urbanized areas that are less "popular" among OSM users, there are gaps in the OSM database and "white spots" resulting from the lack of objects introduced there.

The results received from the OSM data quality analysis indicate that OSM data may provide strong support for other spatial data, including official and state data. Additionally, OSM data are mapped by users and appear in the OSM database on an ongoing basis. In the case of the BDOT10k data, the Head Office of Geodesy and Cartography, by virtue of legal provisions, conducts coordination works aimed at maintaining homogeneity, harmonization and consistency of the BDOT10k data in the entire country through cooperation with public administration bodies and voivodship marshals with respect to developing and maintaining up-to-date data. Updating the BDOT10k is a time-consuming process that involves many additional units and institutions. Due to the voluntary nature of the OSM data and the work of the database users, it should be emphasized that the database requires systematic control and supplementation with new objects and information.

Voluntary geographic information (VGI), or geospatial content generated by non-professionals who use mapping systems available on the Internet, provides opportunities for government agencies at all levels to enrich their geospatial databases. Moreover, in some cases, "eyes on the ground" VGIs have an advantage over more expensive accuracy tests conducted by official agencies because the authors have unique local knowledge. OSM's crowdsourced geospatial data helps fill micro-level data gaps and provides insight into SDG progress in a more real-time manner than is possible through annual or biennial surveys and periodic censuses. OpenStreetMap is currently the largest geospatial dataset under an open license. As OSM is increasingly used in various applications, it is important to control the quality of OSM data. OSM service provides its own OSM data quality control tools. Often the tools accomplish this by providing a list of errors in the data that mappers can then fix with editing tools. However, it is an internal tool that may validate data incorrectly. Therefore, external data quality control is important. Taking into account the received discrepancies, it would be recommended to introduce a control system in areas with available reference data of higher accuracy. Consider the OSM data quality indices presented in the article, in future studies the authors plan to extend the OSM data quality analysis with point objects for the entire territory of Poland and to compare these results with the data quality in other areas of Europe.

Author Contributions: Conceptualization, S.B. and K.P.; methodology, S.B. and K.P.; software, S.B.; validation, S.B. and K.P.; formal analysis, S.B.; investigation, S.B. and K.P.; writing—original draft preparation, S.B.; writing—review and editing, S.B. and K.P.; visualization, S.B.; supervision, S.B. and K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Faculty of Civil Engineering and Geodesy, Institute of Geodesy of the Military University of Technology with the frame of statutory research [PBS 933/2017].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study is available on request from the co-author Sylwia Borkowska, e-mail: sylwia.borkowska@wat.edu.pl.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Redman, T.C. *Data Quality for the Information Age*, 1st ed.; Artech House Publishers: Boston, MA, USA, 1997.
2. Loshin, D. Monitoring Data Quality Performance: Using Data Quality Metrics. *White Pap. Inform.* **2010**, 4–11. Available online: https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/informatica_whitepaper_monitoring_dq_using_metrics.pdf (accessed on 28 December 2021).
3. Xia, J.; Myers, R.L.; Wilhiote, S.K. Multiple open access availability and citation impact. *J. Inf. Sci.* **2010**, *37*, 19–28. [CrossRef]
4. Bielecka, E. Geographical data sets fitness of use evaluation. *Geod. Vestn.* **2015**, *59*, 335–348. [CrossRef]
5. Global Indicator Framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development, A/RES/71/313, E/CN.3/2018/2, United Nations. 2017. Available online: https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202020%20review_Eng.pdf (accessed on 28 December 2021).
6. *ISO 8000-61:2016*; Data Quality—Part 61: Data Quality Management: Process Reference Model. ISO/TC 184/SC 4 Industrial Data. International Organization for Standardization: Geneva, Switzerland, 2016.
7. *ISO 19157:2013*; Geographic information—Data Quality. ISO/TC 211 Geographic Information/Geomatics. International Organization for Standardization: Geneva, Switzerland, 2013.
8. Girres, J.F.; Touya, G. Quality Assessment of the French OpenStreetMap Dataset. *Trans. GIS* **2010**, *14*, 435–459. [CrossRef]
9. Bertolotto, M.; McArdle, G.; Schoen-Phelan, B. Volunteered and crowdsourced geographic information: The OpenStreetMap project. *J. Spat. Inf. Sci.* **2020**, *20*, 65–70. [CrossRef]
10. Feldmeyer, D.; Meisch, C.; Sauter, H.; Birkmann, J. Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 498. [CrossRef]
11. Haklay, M. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environ. Plan. B Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
12. Hacar, M.; Kılıç, B.; Şahbaz, K. Analyzing OpenStreetMap Road Data and Characterizing the Behavior of Contributors in Ankara, Turkey. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 400. [CrossRef]
13. Minaei, M. Evolution, density and completeness of OpenStreetMap road networks in developing countries: The case of Iran. *Appl. Geogr.* **2020**, *119*, 102246. [CrossRef]
14. Tian, Y.; Zhou, Q.; Fu, X. An Analysis of the Evolution, Completeness and Spatial Patterns of OpenStreetMap Building Data in China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 35. [CrossRef]
15. Hecht, R.; Kunze, C.; Hahmann, S. Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 1066–1091. [CrossRef]
16. Dorn, H.; Törnros, T.; Zipf, A. Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. [CrossRef]
17. Taginfo OpenStreetMap. Available online: <http://taginfo.openstreetmap.org/tags> (accessed on 28 December 2021).
18. Haklay, M.; Weber, P. OpenStreetMap—User-generated Street Map. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
19. Ramm, F.; Topf, J.; Chilton, S. *OpenStreetMap: Using and Enhancing the Free Map of the World*, 1st ed.; UIT Cambridge: Cambridge, UK, 2010.
20. Arsanjani, J.J.; Zipf, A.; Mooney, P.; Helbich, M. *OpenStreetMap in GIScience: Experiences, Research, and Applications*; Lecture Notes in Geoinformation and Cartography; Springer: Basel, Switzerland, 2015.
21. Goodchild, M.F. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *Int. J. Spat. Data Infrastruct. Res.* **2007**, *2*, 24–32.
22. Goodchild, M.F. Spatial Accuracy 2.0. 2008. Available online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.1444&rep=rep1&type=pdf> (accessed on 20 January 2021).
23. Goodchild, M.F.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241. [CrossRef]
24. Goodchild, M.F.; Li, L. Assuring the quality of Volunteered Geographic Information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
25. Brovelli, M.A.; Zamboni, G. A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 289. [CrossRef]
26. Yagoub, M.M. Assessment of OpenStreetMap (OSM) Data: The Case of Abu Dhabi City, United Arab Emirates. *J. Map Geogr. Libr.* **2017**, *13*, 300–319. [CrossRef]
27. Husen, S.N.R.M.; Idris, N.H.; Ishak, M.H.I. The Quality Of Openstreetmap In Malaysia: A Preliminary Assessment. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *XLII-4/W9*, 291–298. [CrossRef]
28. Report on the state of Piaseczyński Powiat in 2019. Available online: <http://bip.piaseczno.pl/public/getFile?id=511435> (accessed on 28 December 2021).
29. GUS The Central Statistical Office Poland 2021. Available online: <https://stat.gov.pl/obszary-tematyczne/ludnosc/> (accessed on 28 December 2021).
30. GEOFABRIK Service. Available online: <http://download.geofabrik.de/europe/poland.html> (accessed on 28 December 2021).

31. MRPiT 2021: Rozporządzenie Ministra Rozwoju, Pracy i Technologii z Dnia 27 Lipca 2021 r. w Sprawie Bazy Danych Obiektów Topograficznych Oraz Bazy Danych Obiektów Ogólnogeograficznych, a Także Standardowych Opracowań Kartograficznych [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. Available online: <https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210001412> (accessed on 28 December 2021).
32. GEOPORTAL Service. Available online: <https://www.geoportal.gov.pl/dane/baza-danych-objektow-topograficznych-bdot> (accessed on 28 December 2021).
33. Statistics Service. Available online: <https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/> (accessed on 28 December 2021).
34. Goodchild, M.F.; Hunter, G.J. A simple positional accuracy measure for linear features. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 299–306. [[CrossRef](#)]
35. Törnros, T.; Dorn, H.; Hahmann, S.; Zipf, A. Uncertainties of Completeness Measures in Openstreetmap—A Case Study for Buildings in a Medium-Sized German City. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *II-3/W5*, 353–357. [[CrossRef](#)]
36. Kounadi, O. *Assessing the Quality of OpenStreetMap Data*; MSC, Geographical Information Science; University College of London Department of Civil, Environmental and Geomatic Engineering: London, UK, 2009.
37. Jenks, G.F. The Data Model Concept in Statistical Mapping. *Int. Yearb. Cartogr.* **1967**, *7*, 186–190.
38. OSM Stats. Available online: <https://osmstats.neis-one.org> (accessed on 20 February 2022).
39. Overview of the ResultMaps. Available online: <https://resultmaps.neis-one.org/> (accessed on 20 February 2022).
40. OSM Wiki. Available online: https://wiki.openstreetmap.org/wiki/Pl:Importy_oficjalnych_danych_pa%C5%84stwowych (accessed on 9 March 2022).
41. Van Den Hoek, J.; Friedrich, H.K.; Ballasiotes, A.; Peters, L.E.R.; Wrathall, D. Development after Displacement: Evaluating the Utility of OpenStreetMap Data for Monitoring Sustainable Development Goal Progress in Refugee Settlements. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 153. [[CrossRef](#)]

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,



The Author(s). 2023 Open Access. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

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).

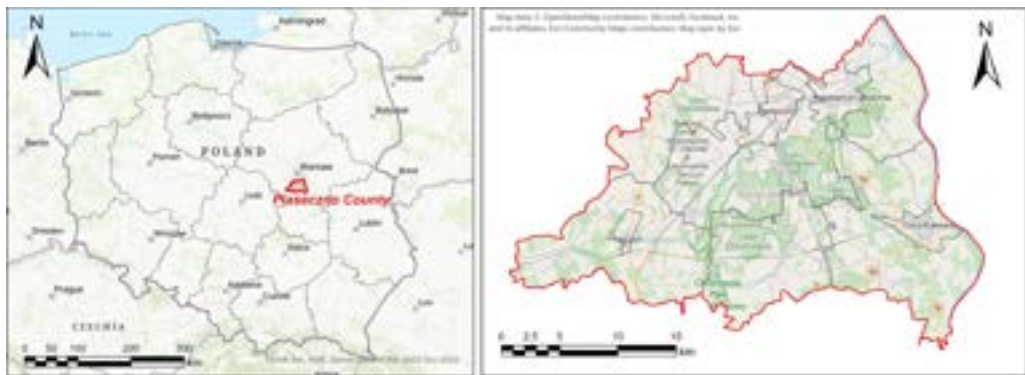


Fig. 1. Location of the analyzed Piaseczno County (source: OpenStreetMap.org, GUGIK, Esri)

Piaseczno County, with an area of 621.04 km², 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 km². 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 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 Zarządzania Baza Danych Obiektów 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 (BUIIT), 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 km² 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 km². 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 km² according to the Eq. (1):

$$C\ Index = \frac{\sum Ft_Area_{OSM_match}}{\sum Ft_Area_{REF_match}} \cdot 100\%, \quad (1)$$

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.

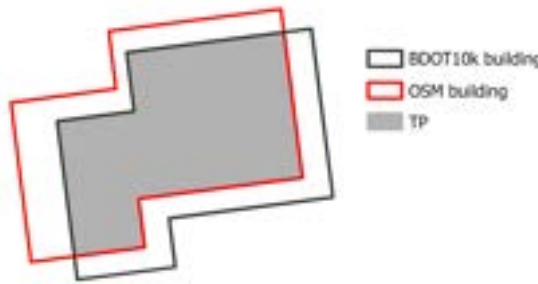


Fig. 2. Presentation of the TP Index

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):

$$TP\ Index = \frac{BLD_{OSM} \cap BLD_{REF}}{\sum Ft_Area_{REF_match}} \cdot 100\%, \quad (2)$$

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):

$$COUNT\ Index = \frac{Ft_Count_{OSM_match}}{Ft_Count_{REF_match}} \cdot 100\%, \quad (3)$$

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

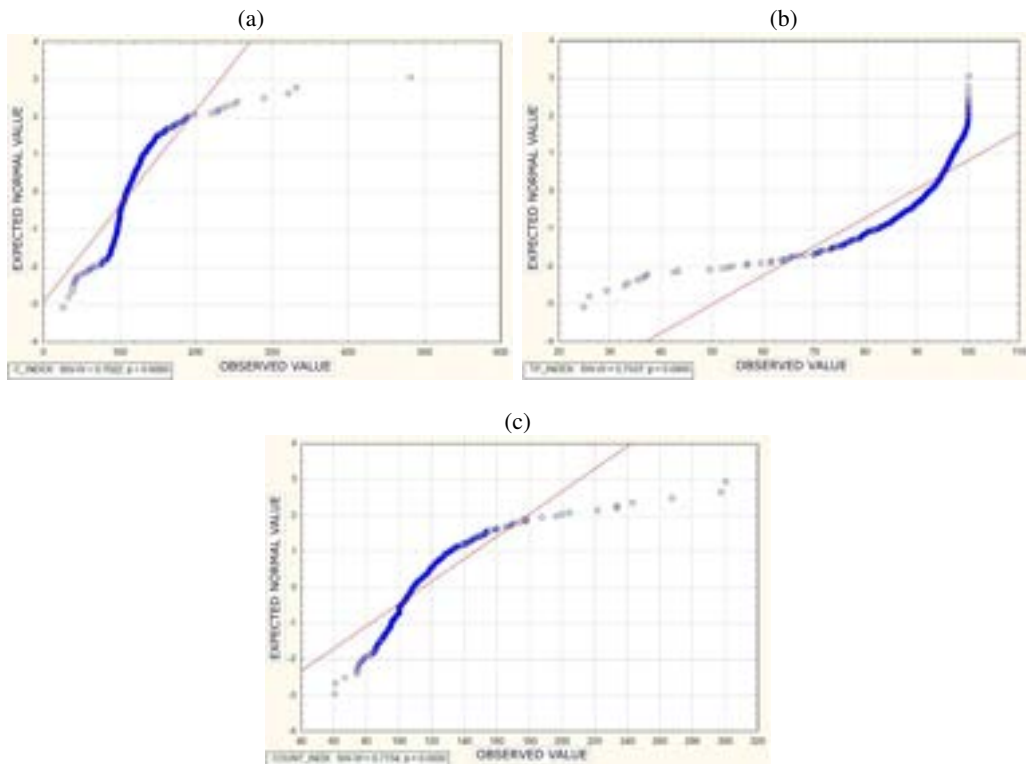


Fig. 3. Normal probability plots with the Shapiro-Wilk test: a) C Index, b) TP Index, c) COUNT Index

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
$C \leq 100\%$	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 compared 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.

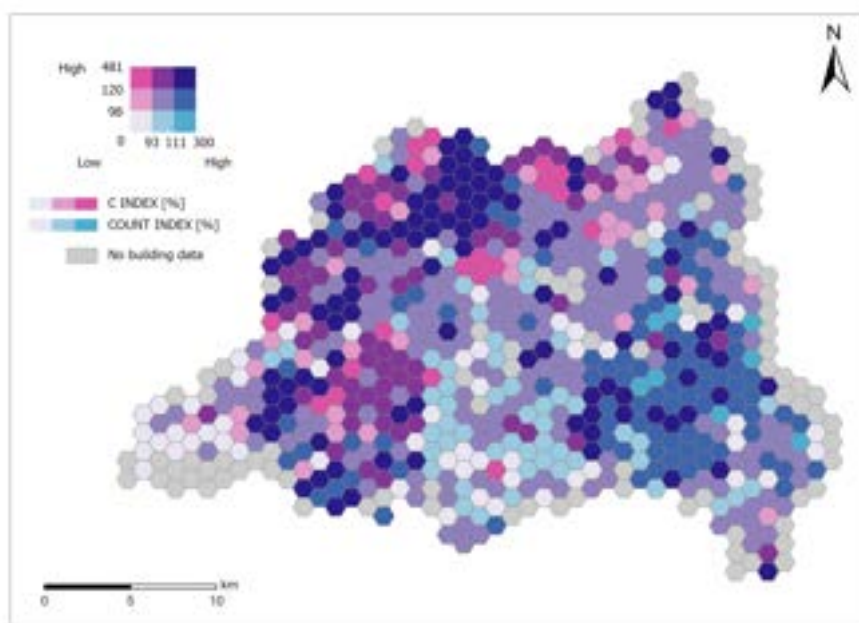


Fig. 4. Bivariate choropleth map of Piaseczno County visualizations C Index and COUNT Index

The first class, for $C Index \leq 98\%$ and $COUNT Index \leq 93\%$ (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 $C Index \leq 120\%$ and $COUNT Index \leq 111\%$ (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.

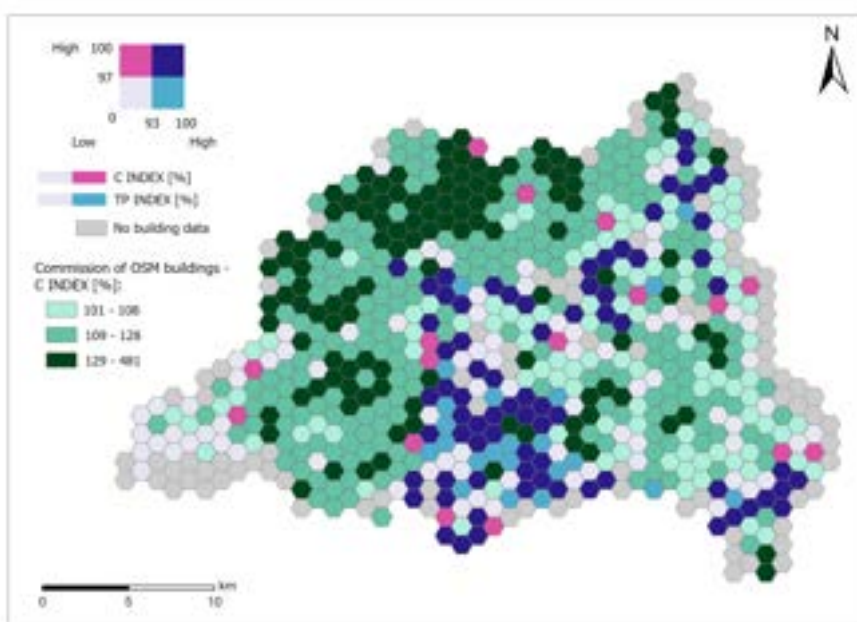


Fig. 5. Bivariate choropleth map of Piaseczno County visualizations C Index and TP Index

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 km². 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.



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 km² 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 funds.

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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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. Tarasinski, J. and Zielinski, W. (2016). Shapiro–Wilk Test with Known Mean. *Revstat Stat. J.*, 14, 89–100. DOI: [10.57805/revstat.v14i1.180](https://doi.org/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. and 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1080/13658816.2021.2023158).

Article

Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data

Sylwia Borkowska ^{*}, Elzbieta Bielecka  and Krzysztof Pokonieczny 

Faculty of Civil Engineering and Geodesy, Military University of Technology, 00-908 Warsaw, Poland; elzbieta.bielecka@wat.edu.pl (E.B.); krzysztof.pokonieczny@wat.edu.pl (K.P.)

* Correspondence: sylwia.borkowska@wat.edu.pl

Abstract: This study aims at a comparative analysis of quantitative data, namely, OSM and BDOT10k. Analyses were conducted in a 1 km² hexagonal grid, in seven test counties located in different regions of Poland, differing in the degree of urbanization, land cover and natural environment. It is assumed that the authors' consolidated regional classification of the Compound Correspondence Index CCI_{Rn} is attributed to the geometric mapping unit based on TOPSIS values, and their statistical measure of dispersion enables the comparison of datasets for individual geographically disjointed areas according to uniform criteria, e.g., the number of topographic features stored in analyzed datasets, both polygonal (buildings, forests, surface water) and linear (roads, watercourses, railroads). The final results of the regional assessment outperform the local classification giving a higher level of data compliance. Overestimation of regional concordance ranges from 9 to 20% of the county area, with an average of 3% reduction in the area where the two datasets (BDOT10k and OSM) have comparable information ranges. Areas of medium and high nonconformity are decreased by an average of 2.4%.

Keywords: OpenStreetMap; topographic data; categorical data; Compound Correspondence Index (CCI); spatial analysis; data agreement; information volume; fit-for-purpose; data quality



Citation: Borkowska, S.; Bielecka, E.; Pokonieczny, K. Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Appl. Sci.* **2023**, *13*, 7525. <https://doi.org/10.3390/app13137525>

Academic Editor: Edoardo Rotigliano

Received: 11 May 2023

Revised: 18 June 2023

Accepted: 23 June 2023

Published: 26 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Information is a key asset in the 21st century, referred to as the information age. The information age is, in turn, inextricably linked to the information society, a society based on information and knowledge. Furthermore, geospatial (geographical) information (GI) plays, as noted by [1,2], a crucial role among available information, as it is the basis for 60–80% of decisions made by public administrations [3]. In the third decade of the 21st century, geospatial information forms a foundation of the information-based society. Geographic information and GI technology are present in almost every sector of science, the economy and industry. Nowadays, in addition to traditional maps and databases, they offer a plethora of web applications that can solve real-world problems. Furthermore, these analyses are presented in a form understandable to the end user and the consumer of information, and add value to the economy, science and society.

The large amount and variety of data not only create new challenges in effective management and analysis, but also create opportunities to explore the potential value of data. Data, including geospatial data, are not error-free, and the results of their analysis are burdened with uncertainty. Data entry errors lead to inaccurate geospatial data, which in turn has wide implications and, as discussed by Bielecka and Burek [4], significantly affects the results of the final analysis, and can lead to wrong decisions and financial losses. Therefore, the GI society is tirelessly striving to improve the reliability of geospatial data, proposing methods for assessing quality and reliability and testing the suitability of data for a specific task (their fitness for purpose). Data quality assessment is particularly important in the context of selecting a dataset that best meets the user's needs. The selection of fit-for-purpose data is usually supported by an analysis of their quality or completeness,

often understood as information capacity. The analysis is carried out mainly by comparison with reference data considered to be very reliable. The motivation for this study is therefore to develop a composite measure that allows a comparative assessment of the completeness of two datasets. Contrary to studies conducted so far, described in Section 2, the assessment of completeness refers to the basic analytical unit, facilitating a detailed assessment of completeness in particular geographical locations. The comprehensive methodology developed is based on a compensatory comparative analysis (TOPSIS), not previously used in categorical data evaluation, using the linear ranking of the authoritative Compound Correspondence Index (CCI) and statistical measures of dispersion for its *in explicite* spatial visualization. Therefore, this study aims to compare the volume of information (capacity) of two open spatial datasets: (1) Database of Topographic Objects (BDOT10k), made available on an open basis by the Head Office of Geodesy and Cartography, and (2) OpenStreetMap, created by volunteers from all over the world. This methodology contributes to both academics and practitioners by helping a user select the best (fit-for-purpose) spatial dataset available for the task at hand. The decision behind the choice of topographic data, both authoritative and voluntary, lies in their wide use in a plethora of applications, e.g., environmental analysis [5,6], protecting and validating landscapes [7,8], and sustainable rural and urban planning [9,10]. The elaborated methodological framework is intended to facilitate cartographic modeling in the analysis of selecting the best geospatial dataset, namely, the data that are most fit for purpose.

2. Related Works

2.1. Categorical Data Comparison—Literature Review

The literature on geospatial data quality is extremely rich, and those on the implementation of machine learning solutions as part of data quality improvement strategies are gaining popularity. This brief review will thus cover only one aspect, namely, the comparative analysis of maps and datasets containing categorical data. Categorical data typically result from mapping, classification, or modeling; hence most of the literature in the field of qualitative data comparison deals with land cover/land use or landscape data [11]. Differences between categorical maps can be characterized and measured in a variety of ways [12], from descriptive or inferential statistics to advanced data mining. However, two approaches dominate among comparative studies of categorical data, the first based on a cross-tabulation matrix to summarize the degree of data association, and the second using spatial and statistical analysis to determine data comparisons in locational and qualitative dimensions.

The Cohen kappa coefficient of agreement derived from a cross-tabulation matrix, i.e., the relative rating of two or more classifications based on the proportion of correctly allocated cases, dominates most of the literature [13,14]. The Kappa coefficient can also be used with missing data, as pointed out by De Raadt et al. [15]. The authors analyzed three kappa coefficients: Gwet's kappa, regular kappa, and listwise deletion kappa. They found that both Gwet's kappa and listwise deletion kappa outperform regular category kappa in terms of bias, and generally have a very small mean squared error. It is even possible to use weighted kappa, introduced by Cohen in 1968 [16], which is of the utmost importance when disagreements between datasets are not equally important. Vanbelle and Albert [17] remarked that, under certain conditions, "the weighted kappa coefficient is equivalent to the product-moment correlation coefficient". In 2000, Pontius [18] introduced some extension of kappa, namely, kappa with random change agreement and kappa for location (Klocation).

The kappa coefficient, however, has some disadvantages. Foody [13] pointed out that the sample used to evaluate maps should be non-dependent. However, in practice, this assumption is almost impossible, since the same sample of ground data sites is often used for each case of map elaboration. In conclusion, in [13], the author also suggests using a measure of the proportion of correctly assigned cases. A similar opinion is shared by Pontius and Millones [18], who summed up more than a decade of research on the

kappa coefficient and concluded that the simple measures of quantitative and allocation disagreement are even more useful in showing differences in categorical data. Moreover, Pontius [19] also suggests that in data agreement analysis, statistical measures of dispersion such as mean deviation, mean absolute deviation and correlation and slope are even more helpful in interpreting data dissimilarity than indices of agreement.

In view of the aforementioned limitations of kappa, researchers use some qualitative–quantitative approaches that deserve attention due to their combination of evaluation and analysis. Multi-criteria analytical techniques predominantly rely on statistical methods or a combination with data mining techniques. Li and Reynolds [20] quantified spatial heterogeneity in categorical maps using ANOVA statistics. Among scientists, a very popular method for comparing categorical data is clustering. However, as noted by Lex et al. [21], the clustering of multi-dimension data can conceal some important relations between object classes, and the final results strongly depend on the algorithm used. Promising results were obtained by Hagen [22], using the fuzzy set theory, especially for ambiguities in determining the “location of the category (fuzziness of location) and in the definition of the category (fuzziness of category)”. The simultaneous analysis of location and quantity was also the subject of research by Pontius and Suedmeyer [23], who developed a new technique of budgeting agreement and disagreement between two categorical maps. Their methodology also includes stratification, hard and soft classification and multiple resolutions to compare maps by quantity and location. Wabiński et al. [24] adapted the structural information measure, introduced by the Russian cartographer Salistchev [25], to compare tactile thematic maps, which extended the use of comparative measures to maps developed for the needs of visually impaired people. Comber et al. [26] recall that different methods of data comparison require many disparate processing steps, and may lead to different results and conclusions. This assessment was also supported by Boots and Csillag [27], based on the results of an expert workshop conducted in March 2004.

2.2. OpenStreetMap Data Quality

With the advent of OSM, due to the lack of control over the VGI data that characterize most user-provided web resources, many questions have been raised about the quality and reliability of the data. However, as summarized in Arsiani et al.’s [28] study on OpenStreetMap, in some regions, OSM geodata are more complete and geometrically and semantically more accurate than the corresponding proprietary datasets. Haklay [29] was the first who compared the quality of OpenStreetMap data with the Meridian2 data maintained by the Ordnance Survey (OS). The research results published in the fourth year of the project indicate the comparable quality of OS and OSM in terms of accuracy and completeness. The most comprehensive assessment of OSM quality was presented by Girres and Touya [30], providing such elements of spatial data quality as geometric, attribute, semantic and temporal accuracy, logical consistency, completeness, lineage and usage. The outcome of their research shows heterogeneous quality across regions and land use elements, indicating relatively high positioning accuracy and very different completeness depending on the density of OSM volunteers. Mondzech and Sester [31] analyzed OpenStreetMap, focusing on determining the optimal routes for pedestrian traffic. Nevertheless, as demonstrated by Ciepluch et al. [32] and Zielstra and Zipf [33], other topographic data with which OSM data are compared are not always more accurate.

Contrary to the extensive analysis of buildings and roads in OSM [31,34], the first attempt at analyzing land use features was carried out by Hagenauer and Helbich [35]. The authors compared the land use polygons and land use patterns of OSM and Urban Atlas data, and found that the location agreements expressed by kappa were 91, 79 and 76% across the three classification levels, while the attributes of both datasets matched at 81, 67 and 65%. Significantly worse completeness and accuracy results were observed by Zhou et al. [36] when comparing OSM land cover with global CCI-LC (Climate Change Initiative Land Cover) data in 168 countries. In 129 countries, completeness did not exceed

40%, and in only 17 European countries was it higher than 60%. Much better results were obtained for accuracy, which was higher than 60% in 149 countries.

3. Materials and Methods

3.1. Study Area and Data Used

The study covers the territory of seven Polish counties (an administrative unit corresponding to the European Territorial Units for Statistics NUTS 4), which are located in different physical–geographical mesoregions and reflect the diversity of both the natural and anthropogenic environment, making them representative (Table 1). Due to various threats, including the violation of state borders' integrity, they are of strategic importance for security. The area of interest (see Figure 1) covers 3.1% of Poland. Piaseczno and Otwocki counties are located in the central part of Poland, adjacent to the capital city of Warsaw, Ostrowski, in Great Poland. The next two study areas are situated along the eastern border of Poland, namely, Sokółski (in the north) along the border with Belarus, and Sanocki (in the south) on the Polish–Ukrainian and Polish–Slovak border. The sixth county Słupski is located in the Baltic Sea coastal zone. Last but not least, Międzyrzeczki county is located in western Poland. Five of the regions of interest have previously been the subject of OSM data quality assessments [37,38].

Table 1. General characteristics of the counties under consideration.

Description	Piaseczno	Sokółski	Sanocki	Słupski	Ostrowski	Otwocki	Międzyrzeczki
Geographical Subprovinces ¹	Central Polish Lowlands	Podlasie-Białystok Upland	Eastern Beskids	South Baltic Coast	Central Polish Lowlands	Central Polish Lowlands	Greater Poland Lake District
Area (km ²) ²	621.12	2054.34	1223.62	2347.59	1159.92	616.46	1387.61
People	190,607	64,902	92,900	98,761	161,581	124,283	57,100
Population density	311	32	81	43	139	202	41.5
Number of cities	4	4	2	2	2	3	3
Urbanization level (%)	47.8	41.7	47.2	20.7	53.7	61.8	52.3
Land use ² (km ²):							
Built-up and artificial	82.51	72.93	43.65	85.15	55.62	58.91	24.08
Forest	132.88	547.91	586.67	864.13	347.59	250.15	735.43
Agriculture	387.37	1426.28	512.14	1234.97	728.53	276.51	513.42
Water bodies	16.44	6.71	13.60	110.65	13.31	11.14	38.39
Protected area	Chojnów Landscape Park, protected landscape areas	Knyszyn Forest, Biebrza National Park	Słonne Mountains Landscape Park, protected landscape areas	Słowiński National Park	Landscape Park Dolina Baryczy, protected landscape area	Masovian Landscape Park, Landscape Park Dolina Śródkowego Świdra	Notecka Forest, Pszczewska Landscape Park

¹ Solon et al. [39]; ² Based on data from the National Register of Boundaries (PRG); Based on Cadastral data 2021 from geoport.gov.pl.

Two topographic datasets were investigated, namely, the National Database of Topographic Objects (BDOT10k), managed by the Head Office of Geodesy and Cartography, and OpenStreetMap, created on a voluntary basis. Six thematic layers were analyzed in detail: three polygon layers, such as buildings, forests and water bodies, and three linear ones—streams and canals, paved roads and railway lines. Topographic data were chosen because they reflect the complex relationships between components of the geographical environment, related to morphology, geology, hydrology, vegetation, and microclimate.



Figure 1. Study area.

3.2. Method Applied

3.2.1. Main Methodological Assumptions and Research Question

The basic research problem to which this study refers concerns the definition of a Compound Correspondence Index (CCI), at local and regional scales, and the determination of the number and optimal class ranges that explicitly indicate the spatial location of differences in the information capacity of two investigated datasets. The WLC (Weighted Linear Combination) method for comparative, multi-criteria analysis was used on the basis of such criteria as differences in the area covered by buildings, forest and water bodies, and the lengths of roads, railways and rivers. The minimum difference indicates a very similar information volume, while maximum implies large differences between the two sets. It is assumed that the consolidated regional classification of the Compound Correspondence Index (regional CCI, hereinafter referred to as CCI_{Rn}), attributed to the 1 km^2 hexagonal grid, and their statistical measure of dispersion, enables the comparison of datasets for individual geographically disjoint areas according to uniform criteria. The decision to use hexagons has some advantages, as it is closer in shape to circles than squares, potentially reducing bias due to edge effects [40].

Therefore, the answering of two research questions is the priority of this study, and the questions are as follows:

Q1—Is there the difference between the CCI value calculated for grid cells of all research areas together (CCI_{Rn}) and the CCI value calculated in grid cells separately for each area (local CCI; hereinafter referred to as CCI_{Ln}), and if so, how big?

Q2—How does the value of regional CCI (CCI_{Rn}) change if we include another research area in the analysis, i.e., how sensitive is the CCI_{Rn} ?

The answers to the questions allow us to verify the hypothesis that CCI_{Rn} underestimates the dissimilarity between analyzed datasets, indicating slightly higher compliance than the local CCI.

This approach is innovative as it allows the comparison of two sets containing qualitative data for geographically disjoint areas, thus enabling the user to choose one of them consciously and responsibly. It also shows the differences between CCI classifications at the local and regional levels.

3.2.2. Research Schema

The research was conducted in four consecutive phases: (1) preparatory; (2) computational; (3) sensitivity analysis; and (4) visualization. The preparatory phase relied on data acquisition, checking and preprocessing, including coordinate transformation, hexagonal

grid creation, as well as assigning appropriate attributes to grid cells, e.g., the area covered by buildings, forests and water reservoirs and the length of roads, streams and railways in both datasets BDOT10k and OSM. This stage is effectively described in a former publication of Borkowska and Pokonieczny [37]. Phase 2—calculation CCI based on TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution). Phase 3—sensitivity assessment, which aimed to analyze the variability of regional CCI values when the research area is extended to other regions (disjoint or adjacent), based on descriptive statistics measures of dispersion and centrality and spatial inferential statistics, Global Moran's I. Phase 4—visualization, which is based on the portrayal of the CCI value in the form of a choropleth map. The five class intervals reflect the conformity of BDOT10k and OSM data. Contrary to the classic Likert scale [41], reverse ordering was used, consistent with the CCI values, and so class 1, starting with the value 0, meant full compliance, and class 5 meant maximum non-compliance. Class ranges are based on one standard deviation interval. All research phases and stages are shown in Figure 2.



Figure 2. The workflow schema.

3.2.3. Compound Correspondence Index Calculation

TOPSIS is perceived as the most widely used approach among multi-criteria decision analysis (MCDA) [42,43]. This classical MCDA method, formerly developed by Hwang and Yoon in 1981 [44], addresses complex decision problems involving conflicting goals, uncertainty and different data formats. TOPSIS has gained popularity in a plethora applications because it evaluates real-world problems based on a variety of criteria, both quantitative and qualitative. Simultaneously, it takes into account the mutual distances to positive and negative ideal solutions, and finally orders the rank of preferences based on their relative proximity (a combination of these two distance measures). This study uses the classical TOPSIS method to portray differences in the thematic scope (understood as the area covered by buildings, forests and water reservoirs and the lengths of roads, rivers and railways) of two topographic datasets (BDOT10k and OSM) in the form of a composite index to compare and rank the obtained alternatives.

The seven standardized steps of the TOPSIS method described in detail by many papers, e.g., Pavic and Novoselac [45], Zavadskas et al. [46], are in general presented below.

1. Problem description

The decision relies on the selection of BDOT10k or OSM sets (two alternatives; $m = 2$) based on the minimal value of the k criteria ($k = 6$), as shown in Equation (1):

$$\begin{aligned} x_1 &= |BT_B - OSM_B|, x_2 = |BT_F - OSM_F|, x_3 = |BT_W - OSM_W|, \\ x_4 &= |BT_{Ro} - OSM_{Ro}|, x_5 = |BT_S - OSM_S|, x_6 = |BT_{Ra} - OSM_{Ra}|, \end{aligned} \quad (1)$$

where BT denotes the BDOT10k dataset; OSM —OpenStreetMap data; subscripts B, F, W —total area covered by buildings, forest and water bodies in the grid cell, and Ro, S, Ra —total length of paved roads, streams, and railways, respectively.

2. Calculation of the normalized decision matrix (n_{ij}) using the quotient method (Equation (2)):

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}}, i = 1 \dots m; j = 1 \dots n. \quad (2)$$

where m is the number of alternatives, and n is the number of hexagonal grid cells.

3. Calculation of the weighted normalized decision matrix—Equation (3):

$$r_{ij} = w_j \times n_{ij} = 1, \quad i = 1 \dots m; j = 1 \dots n. \tag{3}$$

The criterion for weighting objects was their recognizability on satellite and aerial imagery, which is the main source for their acquisition [34,47,48]. Thus, buildings and forests were given a weight of $w_j = 0.25$, paved roads and railways $w_j = 0.15$, while water bodies and streams just $w_j = 0.10$. Furthermore, these weighting rules are also used in accessibility and passability analyses and, as noted by Pokonieczny [49], are extremely important in crisis management.

4. Determine the positive (PIS) and negative (NIS) ideal solutions, as shown by Equations (4) and (5).

$$PIS = \{r_1^-, \dots, r_j^-, \dots, r_n^-\}, \text{ where } v_j^+ = \begin{cases} \max(r_j^i) & \text{if } j \in B; \\ \min(r_j^i) & \text{if } j \in J' \end{cases} \tag{4}$$

$$NIS = \{r_1^-, \dots, r_j^-, \dots, r_n^-\}, \text{ where } v_j^+ = \begin{cases} \min(r_j^i) & \text{if } j \in B; \\ \max(r_j^i) & \text{if } j \in J' \end{cases} \tag{5}$$

As all criteria used are destimulants, the positive ideal solution was computed as r_j^i min, while the negative ideal was r_j^i max.

5. Calculate the separation measure S_i of each alternative (relative closeness to the positive ideal solution) as (Equations (6) and (7)):

$$S_i^+ = \sqrt{\sum_i^n (r_{ij} - r_j^+)^2}; S_i^- = \sqrt{\sum_i^n (r_{ij} - r_j^-)^2} \quad i = 1, \dots, m \tag{6}$$

6. Calculate the closeness coefficient of the alternatives (CC_i) as:

$$CC_i = \frac{S_i^-}{S_i^+ + S_i^-} \tag{7}$$

7. Sort alternatives in descending order, whereby the highest CC_i indicates the best performance in relation to the evaluation criteria.

4. Results

4.1. Local CCI Diversity

The CCI_L differs between regions considered, taking the highest values in Ostrowski and Międzyrżecki (see Table 2), which are both located in Greater Poland. The counties are characterized by a similar urbanization level (53.7 and 52.3, respectively) and land cover structure (see Table 1). Very high statistical dispersion of CCI_L , as measured by the coefficient of variation, is observed in Słupski and Międzyrżecki counties, where population density, afforestation and percentage of area covered by agricultural land are similar. However, both counties differ significantly in terms of area, with Słupski county being twice as large.

The values of the local Compound Correspondence Index in Otwocki and Piaseczno counties are characterized by the highest value of IQR, as well as standard deviation, and thus indicate a high dispersion of local CCI values. Furthermore, mean CCI_L takes a value greater than σ , which indicates that for most of the hexagonal cells, the CCI_L value is lower than the mean value, i.e., the data consistency is relatively high here. The standard deviation of 0.068 and 0.072 in Otwocki and Piaseczno counties is almost twice as high as the lowest value recorded in Sokólski county (0.039). Furthermore, the relatively high diversity of analyzed data is proven by variance, which takes the value of 0.0052, 0.0046 and 0.0039 in Piaseczno, Otwocki, and Sanocki counties, respectively.

Table 2. Descriptive CCI_L statistics.

Statistics	Piaseczno	Sokólski	Sanocki	Słupski	Ostrowski	Otwocki	Międzyrzeczki
Mean	0.0915	0.0390	0.0678	0.0414	0.0427	0.0989	0.0353
Median	0.0754	0.0304	0.0533	0.0313	0.0314	0.0832	0.0271
Minimum	0	0	0	0	0	0	0
Maximum	0.4828	0.5329	0.4995	0.4764	0.5765	0.5069	0.5899
Q1	0.0420	0.0159	0.0315	0.0165	0.0181	0.0542	0.0129
Q3	0.1196	0.0499	0.0790	0.0507	0.0538	0.1246	0.0427
Variance (σ^2)	0.0052	0.0015	0.0039	0.0018	0.0017	0.0046	0.0016
Std. Dev. (σ)	0.0723	0.0386	0.0627	0.0426	0.0406	0.0680	0.0405
Coeff. of variation	78.9755	98.8908	92.3641	103.0427	95.1072	68.7311	114.7669
Interquartile range (IQR)	0.0775	0.0340	0.0475	0.0342	0.0357	0.0704	0.0298

The CCI_L values in all analyzed counties reveal clustering, as indicated by the spatial autocorrelation Global Moran's I statistics, with z-score ranges of 21.80 to 14.62 (p -values < 0.001) in Piaseczno and Słupski counties, respectively.

A predominance of areas with low and very low differentiation between BDOT10k and OSM (CCI_L first and second class) ranging from 83.5% to 85.3% was observed in Słupski and Międzyrzeczki counties. The relative lack of congruence, defined as semi-compliance, with the highest values of 13.1% to 13.3% characterizes Otwocki, Piaseczno and Ostrowski regions. A great diversity in the analyzed datasets (CCI_L fourth and fifth class) is noted in Otwocki (9.3%) and Piaseczno (9%) counties. The remaining districts stand out with relatively small noncompliance, amounting in Ostrowski to 7.1%, in Sanocki to 6.6% and in Słupski to 6.3%, with the least in Sokólski (5.6%) and Międzyrzeczki (5.1%) (see Table 3). The level of compliance between the BDOT10k and OSM datasets is shown in Figure 3.

Table 3. The percentage of a county's area in CCI_L classes.

Class	Description	Range	Percentage of the County's Area (%)						
			Piaseczno	Sokólski	Sanocki	Słupski	Ostrowski	Otwocki	Międzyrzeczki
1	maximum compliance	$-0.50 \sigma < CCI_L$	35.5	32.1	30.6	30.6	33.2	34.1	29.1
2	moderate compliance	$-0.5 \sigma < CCI_L \leq 0.5 \sigma$	42.3	49.6	52.2	52.9	46.5	43.5	56.2
3	semi-compliance	$0.5 \sigma \leq CCI_L \leq 1.5 \sigma$	13.1	12.7	10.6	10.3	13.3	13.1	9.6
4	moderate noncompliance	$1.5 \sigma \leq CCI_L \leq 2.5 \sigma$	5.6	2.7	3.1	3.1	4.6	9.3	2.5
5	maximum noncompliance	$CCI_L > 2.5 \sigma$	3.4	2.9	3.5	3.2	2.5	-	2.6

4.2. Regional CCI Diversity and Sensitivity Analysis

As the region's area expands, the range of regional CCI values increases, and the local differences between the data become blurred. The variance, standard deviation and interquartile range decrease, which proves that CCI_{R7} values are less diverse than CCI_{R6}, CCI_{R5} and CCI_{R4} (Table 4). Regardless of the region's extent, the CCI_{Rn} mean value is greater than the median and slightly lower than the standard deviation, indicating a predominance of values smaller than the mean, i.e., maximum and moderate compliance between BDOT10k and OSM. Nevertheless, CCI_{Rn} is characterized by a large disparity, based on cv, which takes values above 100%.

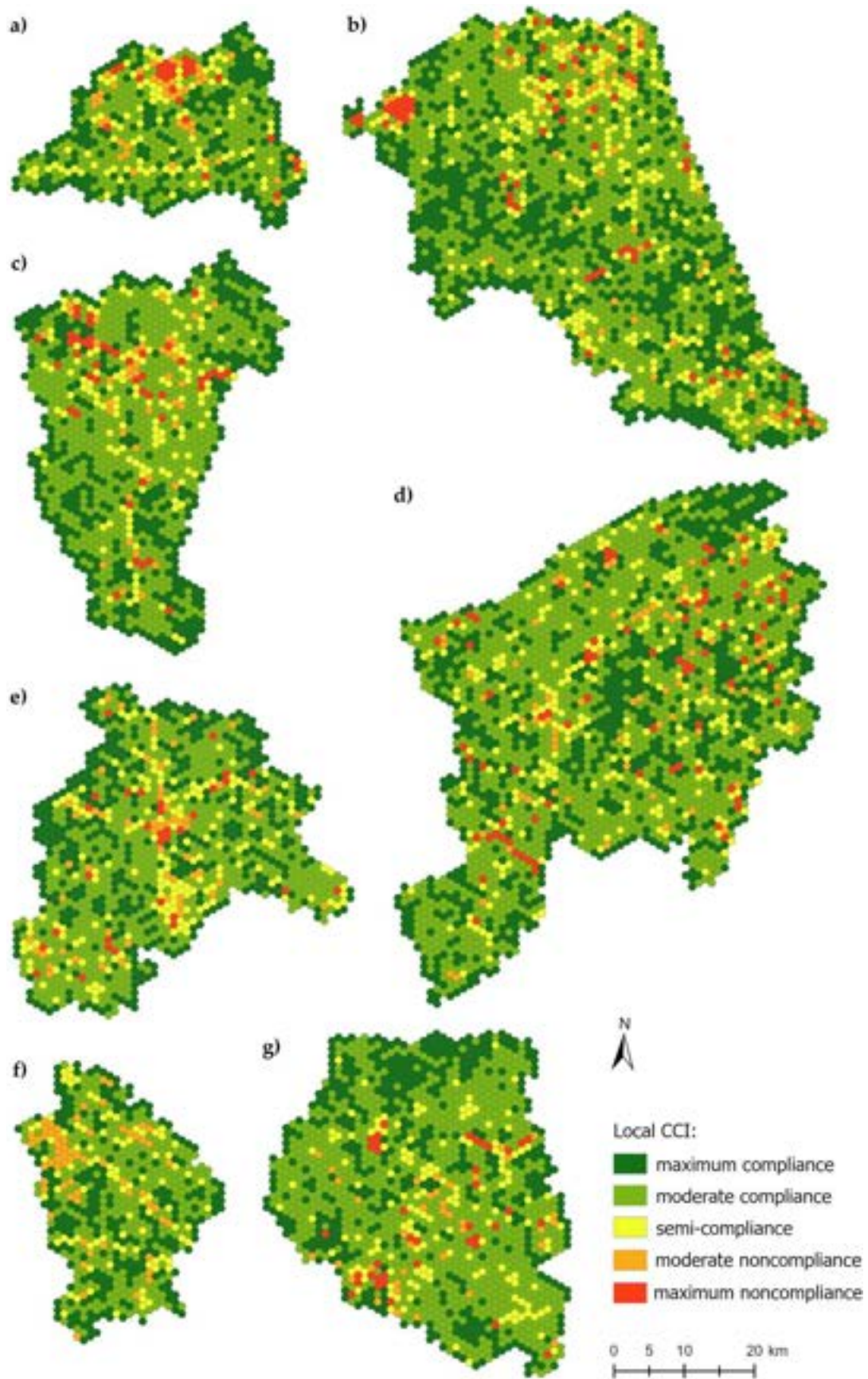


Figure 3. CCI_L in analyzed counties: (a) Piaseczno, (b) Sokółski, (c) Sanocki, (d) Słupski, (e) Ostrowski, (f) Otwocki, (g) Międzyrzecki.

Table 4. Regional CCI descriptive statistics.

Statistics	CCI _{R4} ¹	CCI _{R5}	CCI _{R6}	CCI _{R7}
Mean	0.0327	0.0272	0.0281	0.0263
Median	0.0241	0.0201	0.0208	0.0193
Minimum	0	0	0	0
Maximum	0.4971	0.5289	0.5299	0.5389
Q1	0.0123	0.0104	0.0108	0.01
Q3	0.0410	0.0340	0.0352	0.0327
Variance (σ^2)	0.0012	0.0008	0.0008	0.0008
Std. Dev. (σ)	0.0343	0.0287	0.0289	0.0278
Coefficient of variation (cv)	104.9797	105.3271	102.7925	105.6453
Interquartile range (IQR)	0.0287	0.0236	0.0244	0.0227

¹ CCI_{R4}—denotes the regional CCI computed for four counties, namely, Piaseczno, Sokólski, Sanocki and Słupski; CCI_{R5}—five counties, Ostrowski was added; CCI_{R6}—six counties, additionally Ostrowski; and CCI_{R7}—seven regions, Międzyrzeczki added.

The standard deviation of CCI_{R4} (0.0343) is smaller than that of CCI_L in Sokólski county (CCI_L σ = 0.0386), with the lowest CCI_L value, which value is less than half that of the highest value in Piaseczno county (CCI_L σ = 0.0723). Additionally, in terms of variance, the value of CCI_{R4} (0.0012) is noticeably smaller than that of CCI_L, whereas, in Piaseczno county, the variance is four times higher (CCI_L σ^2 = 0.0052) (see also Table 2). The maximum compliance between BDOT10k and OSM oscillates around 32% of the entire region area. In contrast, moderate compliance, with CCI_{Rn} values in the range of -0.5σ to 0.5σ , represents just over 50% (Table 5). The maximum disagreement between the data is basically negligible. The area where both BDOT10k and OSM data have the lowest correspondence between geospatial objects varies between 3.6 and 6.3%. Figure 4 presents a cartographic visualization of regional CCI₇.

Table 5. The percentage of a county’s area belonging to regional CCI class of compliance.

Class	Description	Interval Size	Percentage of the Counties Area (%)			
			Regional CCI _{R4}	Regional CCI _{R5}	Regional CCI _{R6}	Regional CCI _{R7}
1	maximum compliance	$-0.50 \sigma < CCI_{Rn}$	31.9	31.5	32.3	32.1
2	moderate compliance	$-0.5 \sigma < CCI_{Rn} \leq 0.5 \sigma$	50.9	51.3	50.1	51.2
3	semi-compliance	$0.5 \sigma \leq CCI_{Rn} \leq 1.5 \sigma$	11.1	10.9	11.0	10.4
4	moderate noncompliance	$1.5 \sigma \leq CCI_{Rn} \leq 2.5 \sigma$	6.1	3.6	3.9	6.3
5	maximum noncompliance	$CCI_{Rn} > 2.5 \sigma$	-	2.7	2.7	-

The highest IQR indicates that the spread of the CCI_{Rn} values declines when the region’s area expands, with the highest value in Piaseczno (CCI_{R4}–IQR = 0.045; CCI_{R5-7}–IQR 0.035) and the lowest in Sokólski (CCI_{R4}–IQR = 0.020; CCI_{R5-7}–IQR 0.017). In the remaining counties, the IQR amounted to about 0.025 (see Table 6). In Sanocki, Słupski and Otwocki counties, the standard deviation CCI_{Rn} takes values lower than the mean CCI_{Rn} in more than 50% of the grid cells ($\mu > \sigma$), indicating the predominance of high agreement between BDOT10k and OSM data. In each county, regardless of the spatial extent of the region, the variance does not change, and the standard deviation decreases slightly, as does the achieved maximum value of CCI_{Rn}.

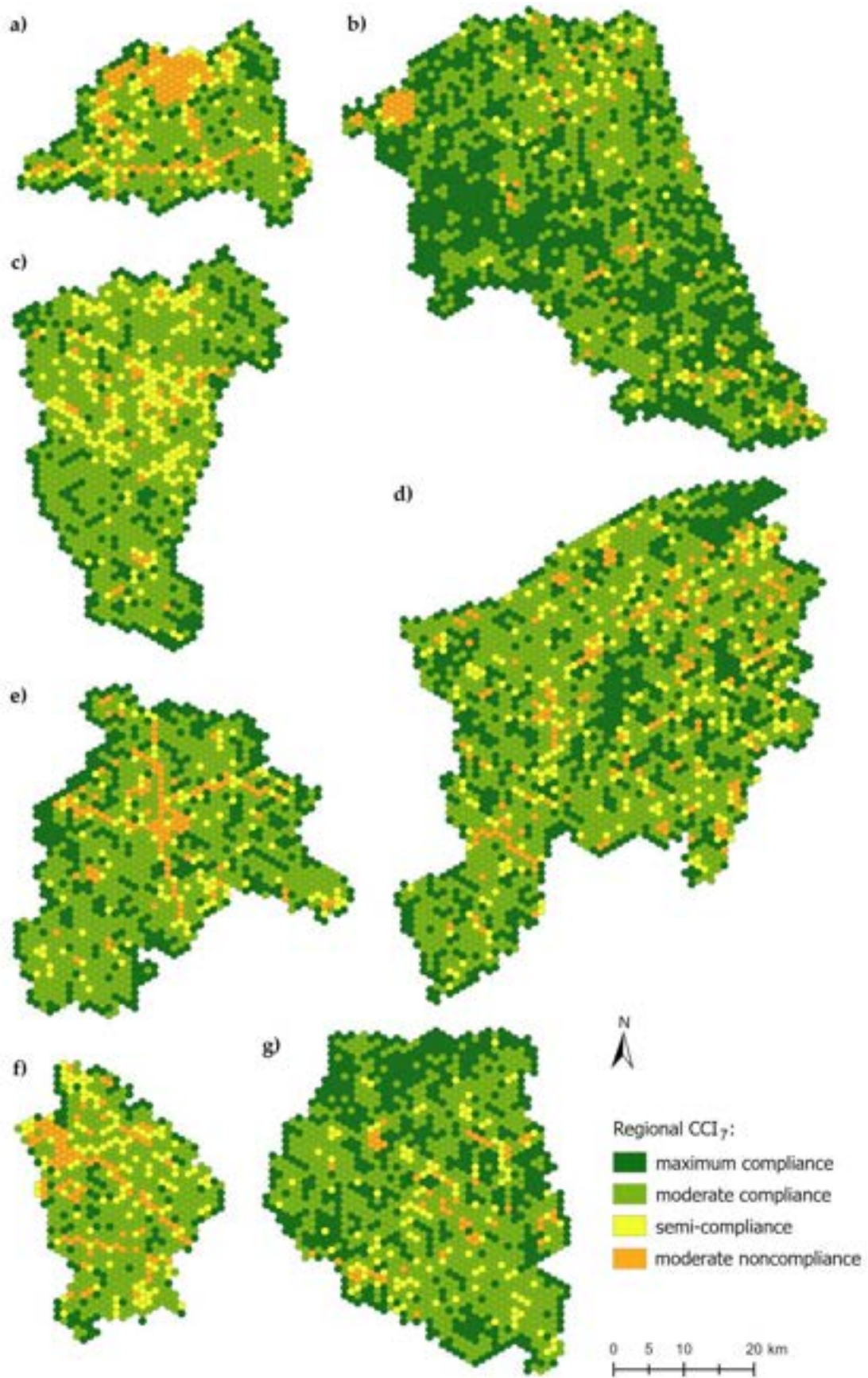


Figure 4. Regional CCI₇ in analyzed counties: (a) Piaseczno, (b) Sokólski, (c) Sanocki, (d) Słupski, (e) Ostrowski, (f) Otwocki, (g) Międzyrzecki.

Table 6. CCI_{Rn} descriptive statistics in each county.

County	CCI _{Rn}	Mean	Median	Min	Max	Q1	Q3	IQR	σ^2	σ
Piaseczno	CCI _{R4}	0.0500	0.0327	0.0000	0.4356	0.0170	0.0620	0.0450	0.0028	0.0527
	CCI _{R5}	0.0403	0.0270	0.0000	0.3778	0.0141	0.0495	0.0354	0.0018	0.0425
	CCI _{R6}	0.0403	0.0271	0.0000	0.3778	0.0142	0.0495	0.0353	0.0018	0.0425
	CCI _{R7}	0.0398	0.0261	0.0000	0.3845	0.0137	0.0491	0.0354	0.0018	0.0429
Sokólski	CCI _{R4}	0.0249	0.0174	0.0000	0.4971	0.0091	0.0300	0.0209	0.0009	0.0304
	CCI _{R5}	0.0208	0.0147	0.0000	0.3949	0.0078	0.0249	0.0172	0.0006	0.0254
	CCI _{R6}	0.0209	0.0147	0.0000	0.3912	0.0077	0.0250	0.0172	0.0007	0.0256
Sanocki	CCI _{R7}	0.0201	0.0138	0.0000	0.3909	0.0072	0.0238	0.0166	0.0006	0.0255
	CCI _{R4}	0.0290	0.0263	0.0000	0.1741	0.0148	0.0382	0.0234	0.0004	0.0204
	CCI _{R5}	0.0348	0.0315	0.0000	0.1975	0.0176	0.0458	0.0282	0.0006	0.0243
	CCI _{R6}	0.0291	0.0264	0.0000	0.1731	0.0149	0.0384	0.0235	0.0004	0.0204
Słupski	CCI _{R7}	0.0279	0.0249	0.0000	0.1619	0.0141	0.0372	0.0231	0.0004	0.0196
	CCI _{R4}	0.0336	0.0257	0.0000	0.3280	0.0133	0.0422	0.0290	0.0011	0.0331
	CCI _{R5}	0.0283	0.0215	0.0000	0.2699	0.0110	0.0355	0.0245	0.0008	0.0280
Ostrowski	CCI _{R6}	0.0283	0.0215	0.0000	0.2684	0.0111	0.0354	0.0243	0.0008	0.0279
	CCI _{R7}	0.0269	0.0203	0.0000	0.2639	0.0105	0.0336	0.0232	0.0007	0.0269
	CCI _{R5}	0.0272	0.0191	0.0000	0.5289	0.0111	0.0328	0.0217	0.0009	0.0298
	CCI _{R6}	0.0272	0.0193	0.0000	0.5299	0.0111	0.0329	0.0218	0.0009	0.0298
Otwocki	CCI _{R7}	0.0268	0.0188	0.0000	0.5389	0.0106	0.0320	0.0213	0.0009	0.0300
	CCI _{R6}	0.0384	0.0310	0.0000	0.2429	0.0193	0.0479	0.0286	0.0009	0.0292
Międzyrzeczki	CCI _{R7}	0.0369	0.0295	0.0000	0.2472	0.0183	0.0460	0.0277	0.0008	0.0289
	CCI _{R7}	0.0217	0.0169	0.0000	0.2373	0.0084	0.0276	0.0192	0.0005	0.0223

Figure 5 presents a cartographic visualization of CCI_{Rn} in two counties: Piaseczno, with the highest spread of CCI, and Sokólski with the lowest.

4.3. Local vs. Regional CCI

Each of the counties analyzed is characterized by an overestimation of the area of maximum compliance by regional CCI compared to local CCI. The largest growth of maximum data compliance takes place in Otwocki and Piaseczno counties, amounting to 50.9% and 40.1%, respectively, and the lowest in Międzyrzeczki (16%) and Słupski counties (14.7%). The overestimation of a very good match between BDOT10k and OSM does not depend on the spatial extent of the region, i.e., it does not change when counties are added and the extent of the region increases. The upsurge in maximum data compliance was at the expense of a moderate compliance class, showing an average of 17.2% (the highest, 29.8%, in Otwocki and the lowest, 6.8%, in Międzyrzeczki counties), and that in comparable data agreement (semi-conformance class) showed an average of 8.2% (the highest, 12.2%, in Otwocki and the lowest, 4.1%, in Słupski); low and very low compliance classes (moderate and maximum non-compliance) increased by 3.5% and 1.9%, respectively. The maximum decline in the area considered non-compliant occurred in Piaseczno and Otwocki counties. However, their extents do not exceed 9%.

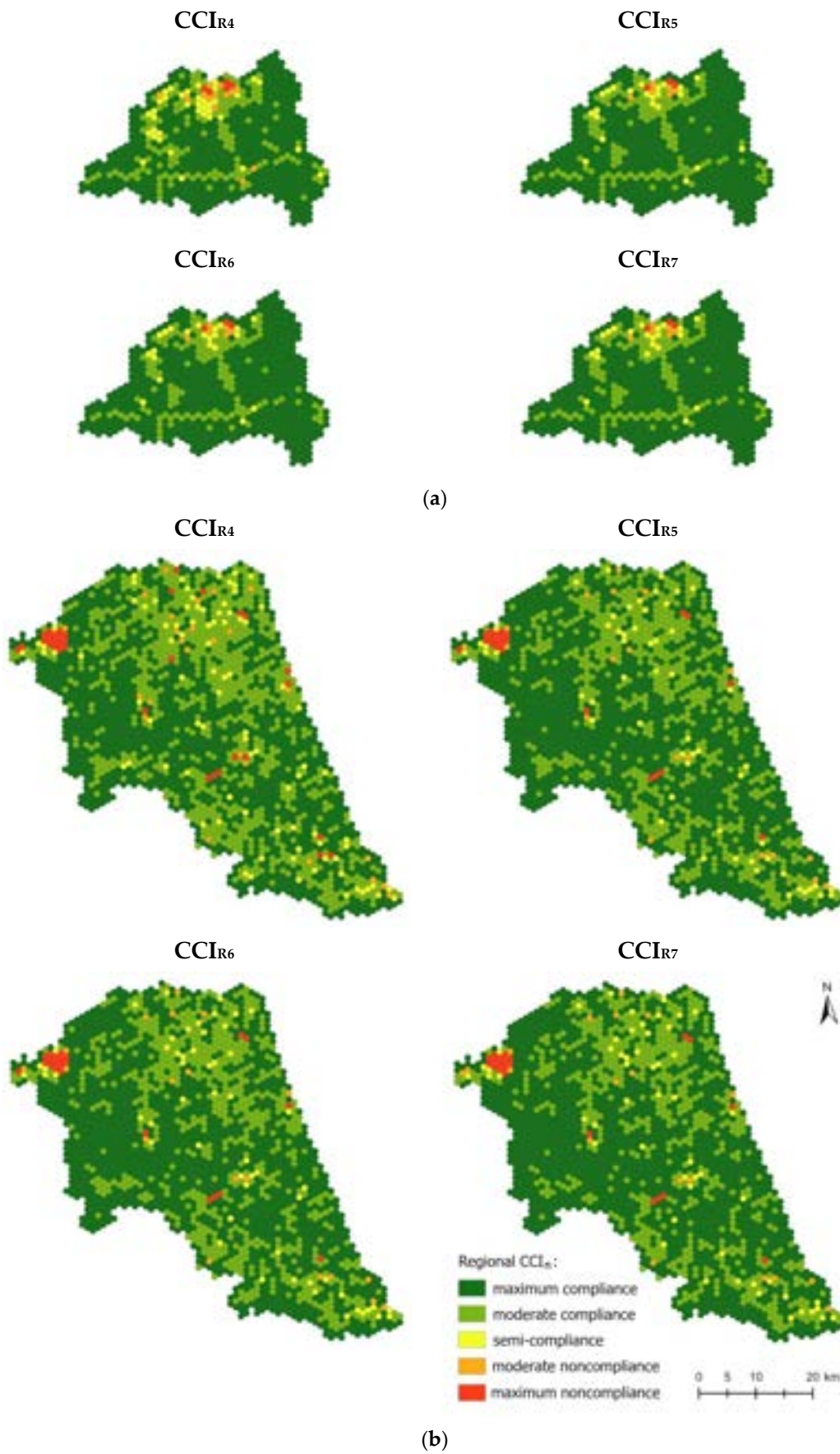


Figure 5. CCI_{Rn} choropleth map: (a) Piaseczno; (b) Sokólski.

5. Discussion

5.1. Semantic Uncertainty

The comparative analysis of data requires semantic similarity in the investigated concepts. As emphasized by many researchers [34,50,51], this assumption is a prerequisite for proper comparative data quality assessment. Semantic similarity in GIScience enables not only data comparison, but also the integration of data from different sources and their further analysis. If the semantic similarity between concepts is ignored, the evaluation or data comparison will be inaccurate [52]. The semantic consistency of concepts in IG can be analyzed at the general level, i.e., only the definition of the object and its geometrical representation are given [53], or at the detailed level, which also comprises definitions of attributes [34,54]. In general, semantic inconsistency results from the classification used, as well as scaling [53]. Scientists use numerous methods to quantify and visualize geodata uncertainties, many of which rely on mathematical models. Semantic uncertainty results from incomplete knowledge of spatial features and phenomena. Traditionally, the boundaries of geographical features are treated as discrete and mutually exclusive, and often, the true location of the boundary is unknown and subject to positional and semantic errors. Zhao et al. [52] introduced a conceptual framework of ontology that includes the definition, semantic relationship, nature, and attributes of geographic features (concept). The authors defined semantic similarity as “calculated on the basis of the feature similarity of ontology concept and the semantic distance of concept to measure the semantic similarity of ontology concept. The smaller the semantic distance between concepts, the higher the similarity between concepts”.

Based on OSMwiki [55] and Ministry Regulation [56], it is seen that in both datasets, a “building” is a man-made structure with a roof, standing (more or less) permanently in one place and geometrically represented by a polygon. Differences in definition only arise when “tag:building use” is considered. During forest surveys, two tags, “natural = wood” and “landuse = forest” are used to map. They represent forests or other areas of trees. Both tags together are compliant with the forest definition in BDOT10k. Water bodies are represented in OSM by the following tags: “natural = water”, “landuse = reservoir” and “water = reservoir”; these are equivalent to “surface water”, i.e., areas occupied by rivers, canals and reservoirs. “Highway = {motorway, trunk, primary, secondary, tertiary, unclassified, residential}” is the principal tag for the road network and corresponds to “road”, while “tag:railway = rail” matches “track or set of tracks” in BDOT10k. Finally, streams, rivers and other watercourses in OSM have the tag “waterway = {river, stream, tidal_channel}”, and in BDOT10k are named “river and stream/channel” [37].

5.2. Validity and Applicability of Data Comparison

Decisions to use fit-for-purpose geospatial datasets are heavily based on data quality, and in particular, information volume, which is understood as the number of geographical features captured. OSM data are rich and heterogeneous, and their quality strongly depends on the degree of urbanization, as mentioned by [34,57]. On the contrary, BDOT10k is perceived as very reliable as authoritative data [58]. It is updated every 2–3 years [56]. Both datasets are used in many applications in Poland by commercial companies and administrations, and in science and education.

Risk assessment and risk management, spatial planning, and environmental protection and monitoring applications often require detailed data when every object is important. In loss assessment (e.g., due to flood or fire), indicating access roads and planning the locations of investments and many other applications, every building, road, railway and object that constitute some kind of barrier, such as water or forest, is important. When the analyzed sets are characterized by high data agreement, the choice of the set is definitely less important than when the two sets vary from each other. It is then necessary to analyze the quality of the datasets based on the indicators described, inter alia, by Borkowska and Pokonieczny [37], and to make a choice supported by the cartographic visualization of data quality [38].

It is worth mentioning that OSM data are made available alongside official data in the form of the WMS (Web Map Service) in the national web portals (Geoportals), providing access to geographic information and spatial data services, e.g., in Poland, Germany, France and Greece [59–62].

6. Conclusions

It was assumed that the selection of a set of geospatial data for a specific application (fit-for-purpose data) could be completed on the basis of the local or regional CCI. Geospatial datasets are assessed by clustering hexagonal grid cells based on ordered CCI similarity. Like any data classification, it is an exploratory procedure because it leads to an understanding of objects and processes. The research is theoretical and methodological in nature, but the close connection to a specific problem situation also gives it a cognitive aspect.

Assuming that the study area consists of several spatially disconnected areas (e.g., counties, cities), the regional CCI makes it possible to assess the suitability of the sets according to a common scale based on the standard deviation. However, the results of the regional assessment outperform the local classification, giving better results, i.e., a higher level of data compliance. The overestimation of regional compliance ranges from 9 to 20% of the county's area, with an average of 3% reduction in the area over which the two datasets (BDOT10k and OSM) have comparable information scope. Areas of medium and large incompatibility are reduced by an average of 2.4%. Sensitivity analysis shows that neither the size of the region nor the spatial location of the counties had a significant impact on the values of the regional CCI.

The CCI values in all analyzed districts revealed clustering. A greatest variation between BDOT10k and OSM data was observed in areas with a high degree of urbanization (e.g., Piaseczno city, Otwock city) and near the course of major transportation routes. The analyses carried out did not prove statistically significant correlations between the CCI coefficients and the land cover elements studied (buildings, roads, rivers, railways, forests and water bodies).

OSM is a valuable source of up-to-date geographic data in emergency mapping, with capacities including, but not limited to, identifying infrastructure at risk of destruction, collapsed buildings, fires and accessibility, which can be important inputs for the orientation of rescuers on the ground. The use of these data in areas with varying degrees of completeness and timeliness of the other official spatial data is of particular importance. Nevertheless, the volunteer type of data may raise issues about quality. The analysis performed of the compliance of the OSM dataset in comparison with official data allows the selection of a set with appropriate characteristics suitable for the intended purpose.

Conducting and comparing several counties using a common analytical framework allows for synthesizing how complete the analyzed geodatasets are, and identifying potential commonalities and differences across places.

The method proposed in this paper has several limitations. Among them are TOPSIS constrains, such as the way that variables are weighted, correlations between variables, and the possibility of an alternative that is close to the ideal point and the nadir point simultaneously. An additional limitation is the CCI designation that is based solely on criteria (variable) such as differences in the area and length of the geographic feature analyzed in the OSM and BDOT10k datasets. In future work, our research will primarily address at least some of the limitations mentioned above. Thus, we will use a different WLC approach to assess the correspondence between OSM and authoritative topographic data. The Hellwig's information capacity method, a the method of optimal predictor selection, is considered for the selection of explanatory variables to be used in a model to evaluate geodatasets. Another research question to deliberate regards the impact of the scope and methods of weighting variables.

Author Contributions: Conceptualization S.B. and E.B.; methodology S.B. and K.P.; formal analysis S.B.; investigation S.B.; resources S.B.; data curation S.B.; writing—original draft preparation S.B.; writing—review and editing S.B., E.B. and K.P.; visualization S.B.; supervision K.P. and E.B.; project administration S.B.; funding acquisition K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Institute of Geospatial Engineering and Geodesy, Faculty of Civil Engineering and Geodesy, Military University of Technology under the statutory research UGB-IG 531-4000-22-816.

Institutional Review Board Statement: Not relevant to this study.

Informed Consent Statement: Not applicable.

Data Availability Statement: The research data are available on the request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodchild, M.F. Geographical information science. *Int. J. Geogr. Inf. Syst.* **1992**, *6*, 31–45. [CrossRef]
2. Sui, D.; Goodchild, M. The convergence of GIS and social media: Challenges for GIScience. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1737–1748. [CrossRef]
3. Pachelski, W. Aktualny stan europejskich i krajowych prac normalizacyjnych w dziedzinie Informacji Geograficznej [Present status of European and National Standardization in Geographic Information]. *Rocz. Geomatyki Ann. Geomat.* **2004**, *2*, 96–105.
4. Bielecka, E.; Burek, E. Spatial data quality and uncertainty publication patterns and trends by bibliometric analysis. *Open Geosci.* **2019**, *11*, 219–235. [CrossRef]
5. Najwer, A.; Jankowski, P.; Niesterowicz, J.; Zwoliński, Z. Geodiversity assessment with global and local spatial multicriteria analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *107*, 102665. [CrossRef]
6. Pluto-Kossakowska, J.; Tulkowska, W.; Władyka, M. GIS technology in green and blue infrastructure analysis. *Ann. Geomat.* **2020**, *XVIII*, 33–50.
7. Bober, A.; Calka, B.; Bielecka, E. Application of state survey and mapping resources for selecting sites suitable for solar farms. In Proceedings of the 16th International Multidisciplinary Scientific Geoconference (SGEM 2016), Albena, Bulgaria, 29 June–5 July 2016; Volume 1, pp. 593–600.
8. Mierzwik, M.; Calka, B. Multi-Criteria Analysis for Solar Farm Location Suitability. *Rep. Geod. Geoinform.* **2017**, *104*, 20–32.
9. Give, S.; Brancia, A.; Satterstrom, F.K.; Linkov, I. Decision Support Systems and Environment: Role of MCDA. In *Decision Support Systems for Risk Based Management of Contaminated Sites*; Marcomini, A., Suter, G.W., Critto, A., Eds.; Springer: New York, NY, USA, 2009.
10. Zykwińska-Rauba, K. Optymalizacja wielokryterialna w procesach decyzyjnych i jej wykorzystanie w zarządzaniu środowiskiem w zrównoważonych miastach. In *Zarządzanie Przedsiębiorstwem Wobec Współczesnych Wyzwań Technologicznych. Społecznych i Środowiskowych*; Walaszczyk, A., Koszewska, M., Eds.; Wydawnictwo Politechniki Łódzkiej: Łódź, Poland, 2021; pp. 174–186.
11. Remmel, T.K.; Fortin, M.-J. Categorical, Class-focused map patterns: Characterization and comparison. *Landsc. Ecol.* **2013**, *28*, 1587–1599. [CrossRef]
12. Boots, B.; Csillag, F. Categorical maps. Comparisons. and confidence. *J. Geogr. Syst.* **2006**, *8*, 109–118. [CrossRef]
13. Foody, G.M. Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633. [CrossRef]
14. Pontius, R.G., Jr. Comparison of categorical maps. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 1011–1016. Available online: http://www2.clarku.edu/~rpontius/pontius_2000_pers.pdf (accessed on 13 January 2023).
15. De Raadt, A.; Warrens, M.J.; Bosker, R.J.; Kiers, H.A. Kappa Coefficients for Missing Data. *Educ. Psychol. Meas.* **2019**, *79*, 558–576. [CrossRef] [PubMed]
16. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [CrossRef] [PubMed]
17. Vanbelle, S.; Albert, A. A note on the linearly weighted kappa coefficient for ordinal scales. *Stat. Methodol.* **2009**, *6*, 157–163. [CrossRef]
18. Pontius, R.G.; Millones, M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* **2011**, *32*, 4407–4429. [CrossRef]
19. Pontius, R.G. Indices of Agreement. In *Metrics that Make a Difference. Advances in Geographic Information Science*; Springer: Cham, Switzerland, 2022. [CrossRef]
20. Li, H.; Reynolds, J.F. A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology* **1994**, *75*, 2446–2455. [CrossRef]
21. Lex, A.; Streit, M.; Partl, C.; Kashofer, K.; Schmalstieg, D. Comparative Analysis of Multidimensional. Quantitative Data. *IEEE Trans. Vis. Comput. Graph.* **2010**, *16*, 1027–1035. [CrossRef]

22. Hagen, A. Fuzzy set approach to assessing similarity of categorical maps. *Int. J. Geogr. Inf. Sci.* **2003**, *17*, 235–249. [[CrossRef](#)]
23. Pontius, R.G., Jr.; Suedmeyer, B. Components of agreement between categorical maps at multiple resolutions. In *Remote Sensing and GIS Accuracy Assessment*; Routledge: Oxfordshire, UK, 2004; Volume 2004, pp. 233–251. Available online: http://www2.clarku.edu/faculty/rpontius/pontius_suedmeyer_2004_rsgisaa.pdf (accessed on 16 January 2023).
24. Wabiński, J.; Mościcka, A.; Kuźma, M. The Information Value of Tactile Maps: A Comparison of Maps Printed with the Use of Different Techniques. *Cartogr. J.* **2021**, *58*, 930. [[CrossRef](#)]
25. Salistchev, K.A. *Kartografia Ogólna*, 2nd ed.; Wydawnictwo Naukowe PWN: Warsaw, Poland, 1998.
26. Comber, A.; Fisher, P.; Wadsworth, R. What is Land Cover? *Environ. Plan. B Plan. Des.* **2005**, *32*, 199–209. [[CrossRef](#)]
27. Csillag, F.; Boots, B. Comparing maps as spatial processes. In *Developments in Spatial Data Handling*; Fisher, P., Ed.; Springer: Berlin/Heidelberg, Germany; New York, NY, USA, 2004; pp. 641–652.
28. Arsanjani, J.J.; Zipf, A.; Mooney, P.; Helbich, M. An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications. In *OpenStreetMap in GIScience: Experiences, Research, and Applications. Lecture Notes in Geoinformation and Cartography*; Springer International Publishing: Cham, Switzerland, 2015; pp. 1–15. [[CrossRef](#)]
29. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Plan. Des.* **2008**, *37*, 682–703. [[CrossRef](#)]
30. Girres, J.F.; Touya, G. Quality Assessment of the French OpenStreetMap Dataset. *Trans. GIS* **2010**, *14*, 435–459. [[CrossRef](#)]
31. Mondzech, J.; Sester, M. Quality Analysis of OpenStreetMap Data Based on Application Needs. *Cartographica* **2011**, *46*, 115–125. [[CrossRef](#)]
32. Ciepluch, B.; Jacob, R.; Mooney, P.; Winstanley, A. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Leicester, UK, 20–23 July 2010.
33. Zielstra, D.; Zipf, A. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, 10–14 May 2010.
34. Nowak Da Costa, J. Novel tool for examination of data completeness based on a comparative study of VGI data and official building datasets. *Geod. Vestn.* **2016**, *60*, 495–508. [[CrossRef](#)]
35. Hagenauer, J.; Helbich, M. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 963–982. [[CrossRef](#)]
36. Zhou, Q.; Wang, S.; Liu, Y. Exploring the accuracy and completeness patterns of global land-cover/land-use data in OpenStreetMap. *Appl. Geogr.* **2022**, *145*, 102742. [[CrossRef](#)]
37. Borkowska, S.; Pokonieczny, K. Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability* **2022**, *14*, 3728. [[CrossRef](#)]
38. Borkowska, S.; Bielecka, E.; Pokonieczny, K. OpenStreetMap—Building data completeness visualization in terms of “Fitness for purpose”. *Adv. Geod. Geoinf.* **2023**, *72*, e35.
39. Solon, J.; Borzyszkowski, J.; Bidłasik, M.; Richling, A.; Badora, K.; Balon, J.; Brzezinska-Wójcik, T.; Chabudzinski, Ł.; Dobrowolski, R.; Grzegorzczak, I.; et al. Physico-geographical mesoregions of Poland: Verification and adjustment of boundaries on the basis of contemporary spatial data. *Geogr. Pol.* **2018**, *91*, 143–170. [[CrossRef](#)]
40. Krebs, C.J. *Ecological Methodology*; Harper Collins: New York, NY, USA, 1989.
41. Song, Z.; Roth, R.E.; Houtman, L.; Prestby, T.; Iverson, A.; Gao, S. Visual storytelling with maps: An empirical study on story map themes and narrative elements. visual storytelling genres and tropes. and individual audience differences. *Cartogr. Perspect.* **2022**, *100*, 10–44. [[CrossRef](#)]
42. Ren, L.; Zhang, Y.; Wang, Y.; Sun, Z. Comparative analysis of a novel M-TOPSIS method and TOPSIS. *Appl. Math. Res. Express* **2007**, *2007*, abm005. [[CrossRef](#)]
43. Zyoud, S.H.; Fuchs-Hanusch, D. A bibliometric-based survey on AHP and TOPSIS techniques. *Expert Syst. Appl.* **2017**, *78*, 158–181. [[CrossRef](#)]
44. Hwang, C.L.; Yoon, K. Methods for multiple attribute decision making. In *Multiple Attribute Decision Making*; Beckmann, M., Kunzi, H.P., Eds.; Springer: Berlin, Germany, 1981; pp. 58–191.
45. Zlatko Pavić, Z.; Novoselac, V. Notes on TOPSIS Method. *Int. J. Res. Eng. Sci.* **2013**, *1*, 5–12.
46. Zavadskas, E.-K.; Mardani, A.; Turskis, A.; Jusoh, A.; Khalil, M.D. Development of TOPSIS Method to Solve Complicated Decision-Making Problems—An Overview on Developments from 2000 to 2015. *Int. J. Inf. Technol. Decis. Mak.* **2016**, *15*, 645–682. [[CrossRef](#)]
47. Goodchild, M.F. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *Int. J. Spat. Data Infrastruct. Res.* **2007**, *2*, 24–32.
48. Cichociński, P. A study on the usability of open spatial data for road network-based analysis—Using OpenStreetMap as an example. *Geoinformatica Pol.* **2021**, *20*, 89–96. [[CrossRef](#)]
49. Pokonieczny, K.J. Comparison of land passability maps created with use of different spatial data bases. *Geografie Prague* **2018**, *123*, 317–352. [[CrossRef](#)]
50. Ballatore, A.; Bertolotto, M.; Wilson, D.C. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl. Inf. Syst.* **2013**, *37*, 61–81. [[CrossRef](#)]

51. Fonte, C.C.; Antoniou, V.; Bastin, L.; Estima, J.; Arsanjani, J.J.; Bayas, J.-C.L.; See, L.; Vatsava, R. Assessing VGI Data Quality. In *Mapping and the Citizen Sensor*; Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V., Eds.; Ubiquity Press: London, UK, 2017; pp. 137–163. [[CrossRef](#)]
52. Zhao, Y.; Wei, X.; Liu, Y.; Liao, Z. A Reputation Model of OSM Contributor Based on Semantic Similarity of Ontology Concepts. *Appl. Sci.* **2022**, *12*, 11363. [[CrossRef](#)]
53. Calka, B.; Orych, A.; Bielecka, E.; Mozuriunaite, S. The Ratio of the Land Consumption Rate to the Population Growth Rate: A Framework for the Achievement of the Spatiotemporal Pattern in Poland and Lithuania. *Remote Sens.* **2022**, *14*, 1074. [[CrossRef](#)]
54. Majic, I.; Winter, S.; Tomko, M. Finding equivalent keys in OpenStreetMap: Semantic similarity computation based on extensional definitions. In Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery GeoAI'17, Los Angeles, CA, USA, 7–10 November 2017; pp. 24–32.
55. OSM. Available online: <https://wiki.openstreetmap.org/wiki/> (accessed on 19 January 2023).
56. 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. Available online: <https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20210001412> (accessed on 26 March 2023).
57. Ribeiro, A.; Fonte, C.C. A Methodology for Assessing Openstreetmap Degree of Coverage for Purposes of Land Cover Mapping. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2015**, *II-3/W5*, 297–303. [[CrossRef](#)]
58. Bielecka, E. Geographical Data Sets Fitness of Use Evaluation. *Geod. Vestn.* **2018**, *59*, 335–348. [[CrossRef](#)]
59. Geoportal Krajowy Service. Available online: https://mapy.geoportal.gov.pl/imap/Imgp_2.html?gpmap=gp0 (accessed on 26 March 2023).
60. Geoportal. de Service. Available online: <https://www.geoportal.de/map.html> (accessed on 26 March 2023).
61. Geoportail Service. Available online: <https://www.geoportail.gouv.fr/donnees/openstreetmap-monde> (accessed on 26 March 2023).
62. Geodata.gr Service. Available online: <http://geodata.gov.gr/maps/?locale=en> (accessed on 26 March 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Sylwia Borkowska¹, Elżbieta Bielecka², Krzysztof Pokonieczny³


Weight Impact on Comparative Evaluation of Topographic Data


Abstract: The paper addresses the problem of weighting in an analysis that supports the selection of a categorical data set according to user needs. Using the Relative Change (RC) of the Compound Correspondence Index (CCI), it is shown that weights have a significant impact on user choice – reaching extreme values in both urbanized and forested areas. Decreasing the weights from 0.25 to 0.17 in forested and built-up areas resulted in the maximum variations that were seen in the hot spot maps, with cold areas generally corresponding to built-up regions and hot areas to forested areas. The analysis covers seven counties that are located in different regions of Poland: Pomerania, Podlasie, Mazovia, Greater Poland and the Beskidy Mountains.


Keywords: quantitative data, OSM, sensitivity analysis, MCDA, TOPSIS

Received: May 6, 2024; accepted: July 22, 2024

© 2024 Author(s). This is an open-access publication, which can be used, distributed, and reproduced in any medium according to the Creative Commons CC-BY 4.0 License.

¹ Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geospatial Engineering and Geodesy, email: sylwia.borkowska@wat.edu.pl (corresponding author),  <https://orcid.org/0000-0003-3183-1512>

² Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geospatial Engineering and Geodesy, email: elzbieta.bielecka@wat.edu.pl,  <https://orcid.org/0000-0003-3255-1264>

³ Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geospatial Engineering and Geodesy, email: krzysztof.pokonieczny@wat.edu.pl,  <https://orcid.org/0000-0001-9114-5317>

1. Introduction

The comparative analysis of spatial quantitative data is often used to select data sets that are suitable for a user's purpose. This generally uses multi-criteria evaluation [1] based on available MCDA (multi-criteria decision analysis) applications. As a decision-support tool, the main objective of MCDA is to assist decision-makers by providing decision options according to accepted criteria. As noticed by [2], however, the criteria should be rational, transparent and non-overlapping. Despite their high diversity, multi-criteria decision applications share some characteristics: (1) a determinate number of comparable alternatives; (2) many criteria against which the alternatives are compared; (3) measurable values that define the quality of the alternative with respect to each of its criteria; and (4) weights for each of the criteria that determine the importance of each of them. Researchers [e.g., 3, 4] have claimed that weights and the choice of how to measure the distances between given criteria are, in general, fundamental and predominantly influence the results.

Many criteria-weighting rules have been presented in the MCDA literature [5, 6]. Their variety leads to the following question: how does the choice of weights affect the final ranking of decision alternatives? Hence, this study aims to analyze the weight impact in a fit-for-purpose assessment of topographical data. It uses the TOPSIS (technique for order of preference by similarity to ideal solution) methodology as well as the Comparative Compound Index (CCI) that was previously introduced in [1]. The CCI was calculated separately for each county in our study; hence, it was demarcated as local. The presented research used and summarized the results of the suitability analysis of the topographical data that was published in [1, 7, 8]. Therefore, the sensitivity analysis of the TOPSIS ordering was carried out on the same seven counties and two topographical data sets; namely, official data that is maintained by the Head Office of Geodesy and Cartography – Database of Topographic Objects (BDOT10k) as well as volunteer data – OpenStreetMap (OSM). This work is part of the discourse on the importance of attribute weights in final TOPSIS ratings. The study confirms the significant influence of the adopted weights on the usability evaluation of the data and the final decision that is made. The novelty of the research lies in the complex universal methodological approach that allows for an evaluation of categorical data; i.e., qualitative data grouped into categories [1] rather than measured data that refers to a form of information that is stored and identified by names or labels (e.g., forest, river, lake, city) according to user-defined criteria. To the best of our knowledge, this research concerns a problem that has not yet been addressed by researchers regarding changes in final TOPSIS rankings as related to changes in attribute weights at the pixel level as well as the relationship between changes in TOPSIS rankings and land use.

The paper is structured as follows: Section 2 describes selected publications on MCDA sensitivity analysis, focusing on the use of TOPSIS and weighting methods.

Section 3 describes the materials and methods that were used, Section 4 presents the results of the sensitivity analysis, and Section 5 is a scientific discussion of the obtained results. Finally, Section 6 concludes the paper.

2. Literature Review

TOPSIS is one of the most popular multi-criteria decision techniques [7, 9–11]. Based on a thorough literature review of TOPSIS applications, Behzadian et al. [9] found that the TOPSIS model had been used mainly in technical and socio-economic research but still needed a broader focus on environmental issues. A similar opinion was shared by Zyoud and Fuchs-Hanusch [10], who found that TOPSIS was mostly used in supply chain management and sustainability research, while analytic hierarchy process (AHP) was predominant in risk modeling and analysis in Geographic Information Systems [10]. The traditional TOPSIS model suffers from correlations between criteria [11] because it uses Euclidean distance, which does not take correlation into account; therefore, its results are affected by overlapping information. To overcome this, the correlation of the attributes should be checked a priori [12]. Furthermore, Li et al. [4] observed that TOPSIS studies generally assumed that parameter weights were invariant and mostly subjectively determined by experts. Yet, only a few studies have included TOPSIS sensitivity analyses based on weight changes [e.g., 4, 13–18], although the results of previous analyses are difficult to generalize today. Criteria weights have various interpretations and implications that are misunderstood and neglected – not only by decision makers, but also by academics. Kobryń and Prystrom [17] found that rating alternatives in TOPSIS strongly depended on the nature of the accepted criteria and the version of TOPSIS (classical, interval, or fuzzy). Choo et al. [15] identified several plausible interpretations of criteria weights and their appropriate roles in decision models, such as scale validity, commensurability, criteria importance, and rank consistency. They also insisted on defining the concept of criteria importance, noting that the “proper interpretation and application of criteria weights would improve the quality of results obtained by using the variety of MCDM models” [15]. Based on investigations of some MCDA applications and available weighting methods on the objectivity of the resulting rankings, Bączkiewicz et al. [16] observed that (1) a proper method for the problem to be adequately solved was essential, (2) a comparative analysis of the results was strongly recommended, and (3) a selection of criteria weights that reflected the preferences of the decision-maker were essential parts of MCDM.

Więckowski and Zwiech [19] used TOPSIS and entropy for selecting energy-efficient materials. The results of analyzing the correlations between weighting and MCDA methods came to the conclusion that, although there were similarities between the rankings, they were not so significant that the weighting methods could be applied equally without changes in the final rankings. Chen et al. [20] used

sensitivity analysis to examine the dependence of model results on input parameters, identify criteria that are particularly vulnerable to weight changes, and show the impact of changing the criteria weights on the model results in the spatial dimension as well as their relative impact on the final evaluation results. The study was carried out using the example of assessing the suitability of irrigated farmland in Australia. The authors of [20] altered and examined the original weights for the five different criteria over a range of 40 simulations using a method of deviating the weights from a base range, defined as a limited set of discrete percentage changes ($\pm 20\%$) in which the weight of each criterion was varied by 1%. A similar study of parameter-sensitivity analysis for determining the variability in the results caused by different input weights for four criteria (climate, soil, slope and erosion) was conducted for land suitability for sorghum cultivation in the Republic of Yemen by [21]. Sixteen weighting schemes were constructed and related to the layers of the criteria map. The results showed that slope and soil were highly sensitive elements in the suitability classification, while climate and erosion were moderately sensitive. Liern and Pérez-Gladish [22] proposed a new TOPSIS approach in which the weights were not determined a priori in an exact way. Weights were considered to be decision variables in a set of optimization problems whose goal was to maximize the relative closeness of each alternative to the ideal solution. The result was a new index of relative proximity that was a function that depended on the values of the weights. The method [22] can be useful in such decision-making situations where it is difficult to determine precise subjective weights.

Undoubtedly, the data and parameter weight burdened the final results of the analysis; hence, wide-ranging and thoughtful TOPSIS sensitivity is still challenging. Our research contributes to the relatively recent discussion of the influence of initial parameters on the results of multi-criteria and multi-attribute analyses (exemplified by TOPSIS).

3. Material and Methods

The study focuses on a sensitivity analysis of TOPSIS weighting in ranking the local Compound Correspondence Index (CCI) that was previously described in detail in [1] expressed as the fitness-for-purpose of six types of topographical objects that are stored in OpenStreetMap (OSM) and the National Database of Topographic Objects (BDOT10k); namely, buildings, forests, water bodies, roads, railroads and rivers. The rationale behind the choice of topographical objects is their clear and unambiguous definition in both databases and their importance in analyses of sustainable development and crisis management. They are also consistent with the authors' previous research on assessing the usability of topographical data. The research was carried out via the following three steps: (1) ranking the local CCI using the TOPSIS method and equal weighting; (2) comparing the CCI ranking results of

expert subjective and equal weighting by the Relative Change (RC_{CCI}); and (3) spatial and statistical analyses of RC_{CCI} . The priority of this study is to answer the following research questions:

1. Does the combination of weights that are used in the local CCI calculation affect the final hexagonal pixel ranking? If so, by how much?
2. Do changes in the local CCI values that are expressed as relative change RC_{CCI} cluster spatially?
3. Are high and low RC_{CCI} values related to land cover types?

3.1. Study Area

Studies were conducted in seven Polish counties; these were characterized in Borkowska et al. [1] and are shown in Figure 1. Słupski County, the largest of the counties (2,300 km²), is situated in the northern part of Poland (along the Baltic Sea coastline), and Sokółski County (along the Polish-Belarusian border). In the central part of Poland (and belonging to the Warsaw agglomeration) are located Otwocki and Piaseczno (Piaseczyński) Counties (each with an area of more than 600 km²). Ostrowski and Międzyrzeczki Counties (with areas of more than 1,100 km² each) are located in the western part of Poland. With an area that is comparable to each of the previous two, Sanocki County is situated in southern Poland (near the border with Slovakia).



Fig. 1. Study area: locations of analyzed counties

Source: [1]

The geographical and geopolitical locations of the counties, their sizes, different use structures and levels of urbanization determined their representativeness in the conducted analyses.

3.2. CCI Rating by TOPSIS

The TOPSIS technique aims at gaining an order preference that is similar to an ideal solution; i.e., a hypothetical solution with maximum benefits and minimum costs of the criteria that are used (attributes or alternatives). The best alternative is that which is nearest to the positive ideal solution and furthest from the negative one [11]. The similarity (or difference) is described by the Euclidean or Mahalanobis geometric distance; in this study the Euclidean distance was applied. The ideal solution and the negative one are examined based on the maximum (or minimum) values of the distances. As mentioned by [18], the TOPSIS method allows for trade-offs between criteria, as it allows a poor performance on one criterion to be ignored in favor of a good performance on another. When choosing the best alternative, the TOPSIS technique is comprised of the following main steps – normalizing the decision matrix, calculating the weighted normalized matrix, calculating the ideal positive and negative solutions, calculating the separation measure and calculating the relative closeness and alternative rankings. The steps that are mentioned above are followed by establishing non-overlapping criteria and their weights [11, 18].

The following subsection provides a comparison of CCI ratings. The CCI values are based on criteria such as differences in the areas that are covered by buildings, forests, and water bodies as well as the lengths of roads, railroads and rivers that are assigned to a 1 km² hexagonal grid [1, 8]. The CCI synthetic indicator, which describes the differences between the two studied topographical data sets (OSM and BDOT10k), was developed by using the classical TOPSIS method in two approaches. The first assumes varied weights, while the second assumes equal weights of the studied topographical objects. For each 1 km² hexagonal grid, the differences for the lengths or areas of the OSM topographical objects that were studied against the BDOT10k objects were calculated in ArcGIS Pro environment according to Equation (1):

$$x_i = |BT_i - OSM_i| \quad (1)$$

where:

x_i , $i = 6$ – difference value of topographical objects (buildings, forests, water bodies [area] or roads, railroads, rivers [length]) that is assigned to the hexagonal grid attribute,

BT_i – BDOT10k object,

OSM_i – OSM object.

The adopted weighting assumes that all of the criteria are equally important; hence, each criterion takes on a weight amount of 0.167 (the second combination in Equation (2)). In order to express the relative percentage change, the Relative Change (RC) between the local CCI is subsequently determined with two variants of weights according to Equation (2).

$$RC_{CCI} = \frac{CCI_{w_2} - CCI_{w_1}}{CCI_{w_2}} \cdot 100\% \quad (2)$$

where:

RC_{CCI} – RC of CCI values,

CCI_{w_1} – value of CCI using various weights (first combination),

CCI_{w_2} – value of CCI using equal weights (second combination).

The first combination (CCI_{w_1}) utilizes object weighting due to the objects' recognizability in the satellite and aerial images from which they were obtained; i.e., aerial orthoimages (10 m pixel) and SPOT 5 orthoimages in the EU border zone. Thus, buildings and forests were each given a weight of 0.25, paved roads and railroads – 0.15, water bodies and streams – 0.10. These weighting rules are also used in accessibility analyses and are extremely important in emergency management [23]. In the second combination (CCI_{w_2}), the weights were equal and amount to 0.167.

In order to compare the RC values that were obtained for the studied counties, four class divisions were defined; these were created with ranges of values that represented the proportions of the standard deviation. The negative RC values were analyzed in two classes; for these, each range was defined according to the interval of half of the standard deviation (0.5σ) that was calculated as the average value for the analyzed counties. The positive RC values were also divided into two classes according to the value of one standard deviation (σ) as the interval of the ranges.

3.3. Hot Spot and Statistical Analysis

Hot spot analysis was used to indicate the spatial relationships and identify the spatial clustering of the RC values. The resulting values showed where objects with high or low values were spatially clustered [24]. A hot spot can be described as an area with a higher concentration of events as compared to an expected number after considering the random distribution of events. A feature with a high value is interesting but may not be a statistically significant hot spot. For an object to be a statistically significant active point, the object will have a high value and be surrounded by other objects with high values as well. The local sum of an object and its neighbors is compared proportionally with the sum of all of the objects. When the local sum is different from the expected local sum and when the difference is too large to be due to random chance, a statistically significant “z” result is obtained [25] according to Equations (3)–(5):

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{\sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}} \quad (3)$$

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \tag{4}$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2} \tag{5}$$

where:

- x_j – RC value of CCI feature,
- w_{ij} – spatial weight between CCI features i and j ,
- n – total number of features.

The Getis-Ord G_i^* statistic provides a z-score, p -value and confidence interval with an interpretation according to Table 1.

Table 1. Hot spot analysis parameter interpretation

Statistics	Description	Implication
$z > 0$ and p -value is small	high-high spatial cluster (the larger the z-score, the greater the clustering degree)	$CCI_{w1} < CCI_{w2}$ $RC_{CCI} > 0$
z is closer to 0	no obvious spatial clustering	–
$z < 0$ and p -value is small	low-low spatial cluster (the smaller the z-score, the greater the clustering degree)	$CCI_{w1} > CCI_{w2}$ $RC_{CCI} < 0$

Source: own elaboration based on [25]

A statistical analysis that was based on descriptive statistics and Pearson correlations was used to provide an overall overview of county-level results.

4. Results

4.1. CCI with Equal Weights Overview

Otwocki and Piaseczno Counties were characterized by the highest inter quantile range (IQR) values as well as the highest standard deviations this indicated the high dispersion of their local CCI values (Table 2). The standard deviation took values that were lower than the mean of the CCI values, which indicated that the CCI values were more concentrated; i.e., the consistency of the data was relatively high in these instances. Otwocki and Piaseczno Counties were characterized by

standard deviations of 0.065 and 0.062, respectively; these were nearly double the lowest value that was recorded in Sokólski County (0.036).

Table 2. Descriptive CCI statistics for TOPSIS analysis with equal weights

Statistics	Mean	Median	Minimum	Maximum	Q1	Q3	σ	IQR
Międzyrzeczki	0.0426	0.0328	0.0000	0.4607	0.0152	0.0553	0.0425	0.0401
Ostrowski	0.0565	0.0470	0.0000	0.5170	0.0277	0.0720	0.0455	0.0442
Otwocki	0.0989	0.0887	0.0000	0.4172	0.0555	0.1268	0.0619	0.0713
Piaseczno	0.0864	0.0756	0.0000	0.4954	0.0425	0.1153	0.0645	0.0728
Sanocki	0.0714	0.0615	0.0000	0.5102	0.0347	0.0946	0.0572	0.0600
Sokólski	0.0355	0.0281	0.0000	0.5855	0.0140	0.0479	0.0356	0.0339
Słupski	0.0389	0.0314	0.0000	0.4572	0.0157	0.0482	0.0382	0.0325

The descriptive statistics of the CCI with the different weights are presented below in Table 3 for comparison purposes (a detailed analysis is described in [1]).

Table 3. Descriptive CCI statistics for TOPSIS analysis with various weights

Statistics	Mean	Median	Minimum	Maximum	Q1	Q3	σ	IQR
Międzyrzeczki	0.0353	0.0271	0.0000	0.5899	0.0129	0.0427	0.0016	0.0405
Ostrowski	0.0427	0.0314	0.0000	0.5765	0.0181	0.0538	0.0017	0.0406
Otwocki	0.0989	0.0832	0.0000	0.5069	0.0542	0.1246	0.0046	0.0680
Piaseczno	0.0915	0.0754	0.0000	0.4828	0.0420	0.1196	0.0052	0.0723
Sanocki	0.0678	0.0533	0.0000	0.4995	0.0315	0.0790	0.0039	0.0627
Słupski	0.0414	0.0313	0.0000	0.4764	0.0165	0.0507	0.0018	0.0426
Sokólski	0.0390	0.0304	0.0000	0.5329	0.0159	0.0499	0.0015	0.0386

Source: own elaboration based on [1]

According to the five data-compliance ranges that were defined by Borkowska et al. [1], the percentages of the local CCI classes that were calculated with equal weights are presented in Table 4. The significant predominance of areas with low and very low differentiations between BDOT10k and OSM (the first and second classes of the CC compliance) could be observed in almost all of the analyzed counties – from 81.7% in Słupski County to 76.6% in Otwocki County. The exception was Piaseczno County; such areas accounted for slightly more than half of the county's size (55%). In this county, the highest diversity (defined as a semi-compliance

[the third class]) could be noted, with a value of 24.7%; the noncompliance (the fourth and the fifth classes) amounted to as high as 20.2%. The other counties in the semi- and non-compliance classes ranged from approximately 13% to 15% (semi-compliance) and 6% to 8% (noncompliance).

Table 4. County area percentages in CCI_L classes for equal weights

Class	Description	County area percentage [%]						
		Międzyrzeczki	Ostrowski	Otwocki	Piaseczno	Sanocki	Słupski	Sokółski
1	maximum compliance	33.2	32.1	33.5	24.0	32.4	31.5	32.3
2	moderate compliance	46.9	47.5	43.1	31.0	46.4	50.2	48.8
3	semi-compliance	14.0	12.9	15.1	24.7	15.2	13.1	14.0
4	moderate noncompliance	5.9	5.2	6.7	16.4	6.0	2.3	4.9
5	maximum noncompliance	–	2.4	1.6	3.8	–	2.9	–

Table 5 below shows the percentages of the local CCI classes calculated with various weights (an analysis is widely described in [1]).

Table 5. County area percentages in CCI_L classes for various weights

Class	Description	County area percentage [%]						
		Międzyrzeczki	Ostrowski	Otwocki	Piaseczno	Sanocki	Słupski	Sokółski
1	maximum compliance	29.1	33.2	34.1	35.5	30.6	30.6	32.1
2	moderate compliance	56.2	46.5	43.5	42.3	52.2	52.9	49.6
3	semi-compliance	9.6	13.3	13.1	13.1	10.6	10.3	12.7
4	moderate noncompliance	2.5	4.6	9.3	5.6	3.1	3.1	2.7
5	maximum noncompliance	2.6	2.5	–	3.4	3.5	3.2	2.9

Source: own elaboration based on [1]

4.2. Relative Changes of CCI

The results of the RC between the local CCIs for the variant of differentiated and equal weights were quite diverse (Tables 6, 7, Fig. 2). A similar range of the minimum (from -34.4% to -30.5%) and maximum (from 64.6% to 74.6%) RC values could be observed in Piaseczno, Sokólski, Sanocki and Otwocki Counties. However, the high values that were obtained in Ostrowski and Międzyrzecki Counties (where the minimum values of the percentage changes in the local CCIs were -12.7% and -23.1%, respectively, and the maximum values were 120.6% and 98.2%, respectively) significantly exceeding the obtained maximum results for the other counties. Ostrowski County also had the highest median (44.0%) and variance (1,429.1%) among the studied counties. The standard deviation values in the analyzed counties ranged from 25.6% (Sokólski County) to 29.6% (Międzyrzecki County), with relatively low means (from 5.5% to 1.1%); these indicated greater variability. Ostrowski County achieved the highest σ value (37.8%) with a mean value of 45.4%.

Table 6. Descriptive statistics of RC of CCI

Statistics	Piaseczno	Sokólski	Sanocki	Słupski	Ostrowski	Otwocki	Międzyrzecki
Mean	1.1	-5.5	13.1	0.1	45.4	5.8	25.8
Median	-1.4	-8.1	15.3	0.1	44.0	6.3	29.6
Minimum	-34.5	-34.2	-30.5	-34.4	-12.7	-31.9	-23.1
Maximum	65.0	64.6	74.6	64.7	120.6	73.5	98.2
Q1	-25.1	-29.5	-8.2	-23.5	15.1	-17.7	6.6
Q3	16.9	9.1	28.9	9.8	72.7	17.7	32.1
Variance (σ^2)	761.7	657.8	691.8	712.4	1429.1	717.1	877.0
Std. dev. (σ)	27.6	25.6	26.3	26.7	37.8	26.8	29.6

Table 7. Percentages of county areas for RC of CCI ranges

Class	Range of RC	Percentage of the county's area [%]						
		Piaseczno	Sokólski	Sanocki	Słupski	Ostrowski	Otwocki	Międzyrzecki
1	$\min \leq -15\%$	36.2	44.1	19.3	34.3	-	27.6	11.8
2	$-15\% < RC \leq 0$	14.9	14.3	11.1	15.4	16.3	16.1	10.3
3	$0 < RC \leq 30\%$	32.2	31.0	45.5	36.5	17.4	39.5	29.7
4	$30\% < RC \leq \max$	16.7	10.6	24.1	13.7	66.4	16.8	48.2

a)

b)

c)

d)

e)

g)

f)

Fig. 2. RCs of local CCIs in analyzed counties:
a) Piaseczno; b) Sokólski; c) Sanocki; d) Słupski; e) Ostrowski; f) Otwocki; g) Międzyrzeczki

Piaseczno, Sokólski, Słupski, and Otwocki Counties (Table 7) achieved similar proportions of negative RC values (accounting for about half of each county), with a clear predominance of values of up to -15% RC (the maximum being 44.1% of the area of Sokólski County). Values from -15% to 0% RC for these districts represented from 14.3% of the area in Sokólski County to 16.1% of the area of Otwocki County. In Ostrowski, Międzyrzecki and Sanocki Counties, negative RC values account for 16.3%, 22.1%, and 30.4%, respectively. Positive values of up to 30% of RC predominated in Sanocki (45.5%) and Otwocki (39.5%) Counties. However, the shares of Piaseczno, Sokólski, and Międzyrzecki Counties were similar, amounting to about one-third of the analyzed set. The largest shares of a county’s area (within a range of more than 30% RC) were represented by Ostrowski (66.4%) and Międzyrzecki (48.2%) Counties, and the smallest shares were those of Słupski (13.7%) and Sokólski (10.6%) Counties.

The land use of the analyzed areas was dominated by agricultural land (53% on average) and forests (36% on average). The relative sizes of the built-up areas varied from less than 2% in Międzyrzecki County to 13.3% in Piaseczno County (Fig. 3).

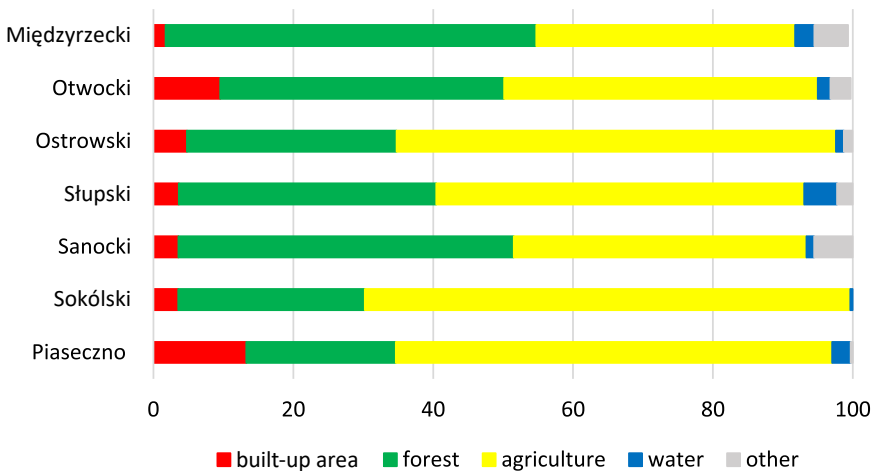


Fig. 3. Land use in study area

Source: own elaboration based on official cadastral statistics from 2021

Pearson linear correlations (r) provide insight into the associations of land use and RC. At a significance level of $p < 0.0500$, the Pearson correlation varies depending on the range of the RC levels. A moderate negative correlation (-0.52) can be observed between a forest and an RC level that is less than 0, while a strong positive correlation (0.76) can be observed between a forest and an RC level that is greater than 30%. A strong negative correlation (-0.82) was recorded between agriculture and an RC level that was greater 30%. Built-up areas are moderately negatively (-0.61) correlated with an RC range that is between 0% and 15%.

4.3. Variants of Local CCI Weights – Hot Spot Analysis

A Getis-Ord G_i^* analysis identified the statistically significant hot and cold spots that are shown in Figure 4.

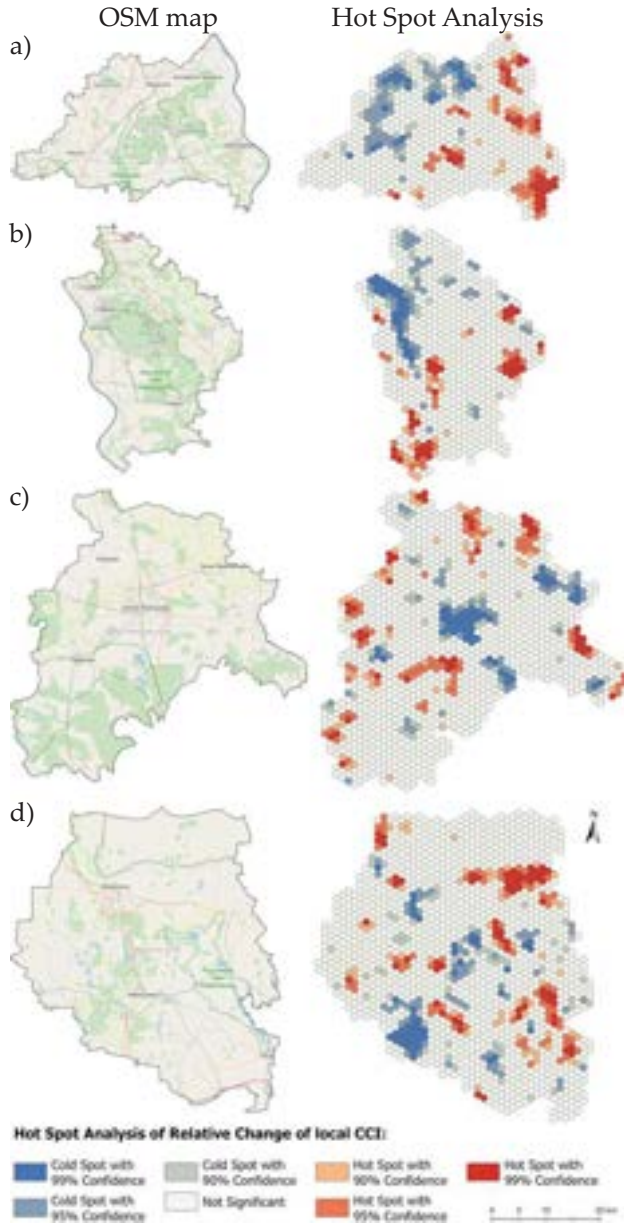


Fig. 4. Hot spot analysis of RC of local CCI in analyzed counties: a) Piaseczno; b) Otwocki; c) Ostrowski; d) Międzyrzeczki (compared to OSM map)

Based on the hot spot maps (Fig. 2), a visual analysis of the relationship between the landscape and the hot and cold clusters was performed. In Piaseczno and Otwocki Counties, those areas that were identified as hot spots were clustered mainly in open areas (meadows, farmlands), in the vicinity of water bodies (ponds in Żabieniec) and (less frequently) in areas of dispersed settlements (mainly rural areas) and forested areas (Chojnowskie Forests, Masovian Landscape Park). However, those areas that were identified as cold spots occurred in urban areas (the cities of Piaseczno [along with its neighboring towns south to the city of Tarczyn] and Otwock [with its neighbors Józefów and Karczew]) as well as along major transportation lines (Krakowska Avenue). In Ostrowski and Międzyrzecki Counties, the hot spots were similarly concentrated in open areas (the northeastern part of the county), large water bodies, and forests (Pszczewski Landscape Park, Barycz Valley Landscape Park). Cold spots also occurred in areas of compact development – the cities of Skierzyna, Międzyrzecz and Ostrów Wielkopolski, the towns of Odolanów and Nowe Skalmierzyce as well as in the forests in the northern part of Międzyrzecki County (Nietopek Nature Reserve).

5. Discussion

Fitness for purpose is a principle that is widely accepted among analysts as the correct approach for obtaining a quality data set [26, 27]. However, only a few analysts or end users of data can accurately determine what data quality is required for a specific task. When selecting a particular spatial data set, the user should be very attentive, as it is impossible to evaluate all of the strengths and weaknesses of available data. One aspect that is difficult to assess is the up-to-dateness, which is given for an entire data set, while its parts could be characterized by a different topicality [28]. Topographical data are updated periodically according to the rules in force (which vary from country to country). In Poland, this used to be a ten-year period [29, 30]; however, it was recently changed to a two-year period. The OSM data is updated by users (mappers), so the data up-to-dateness depends on their activities. The BDOT10k data that was used in this research was from March 2020, while it was not possible to determine the year of the OSM data update for the analyzed areas. A literature research [31, 32] showed that the most up-to-date data was on roads and buildings.

Nevertheless, an important aspect that significantly influences the TOPSIS ranking results is the selection of topographical objects and their prioritization, which is usually associated to the overarching objective; i.e., answering the question about the purpose of an analysis. In the present study, it was assumed that this objective was related to crisis management (i.e., floods, fires, terrorist attacks) for which the identification of populated areas, access routes and hazard areas is important. The second application area was sustainable development according to

Agenda 2030. The greatest weights were assigned to buildings and forests, whose importance in both applications was indisputable [1]. The research did not consider object attributes due to the relatively small number of objects that were described with attributes in OSM [1, 31].

It is worth noting that a CCI can be calculated for any map unit (MU), whether natural (e.g., catchments, ecotopes), administrative, or geometric. The universality of a CCI also lies in the facts that any objects can be included in analyses and mapping units can be ranked by considering other (non-topographical) categorical data (even MUs within one data set).

In the presented research, two variants of the weights of the CCI, a comparative measure of OSM, and the BDOT10k quantitative data were analyzed. In the first approach (in accordance with [1]), differentiated weights were adopted, which corresponded to the relevance of the objects under study that were adopted by the authors; i.e., buildings, forests (with the greatest value of the weights), communication networks (a moderate value of the weights), and watercourses/water bodies (the lowest value of the weights). In contrast, all of the analyzed objects were considered to be equally important in the presented variant, and their weights were assumed to be equal. The CCI values with different and equal weights differed, as was previously mentioned in [18, 19]. The local CCI values showed clustering in all of the analyzed counties. According to the adopted gradual scale of compliance, the CCI in two combinations of weights occupied similar shares of the area of each county. The greatest differences in the occupied areas could be seen in the case of Piaseczno County – the sizes of the maximum and moderate compliance areas decreased by a 22.8% share of the county's area after equalizing the CCI weights (amounting to 55%). However, the area that was occupied by semi-compliance doubled to a 25% share of Piaseczno County's area. Similarly, the share of the areas that were assessed as being of moderate and maximum noncompliance increased from 9% to 20.2% of the share of the county's area after equalizing the weights. This allowed us to conclude that, as in the previous studies, those areas with high degrees of urbanization showed the greatest variability between the BDOT10k data and the OSM data.

At the pixel level, the Pearson correlation analysis did not show a significant relationship between the land cover type and the CCI for the equal weights (CCI_{w_2}); this was similar to the CCI that was analyzed in the previous article (CCI_{w_1}) [1]. Also, no significant statistical relationship was shown in the Relative Changes between the CCI weights that were used (significance level $p < 0.0500$). For this reason, a hot spot analysis was performed in order to identify clusters of spatial phenomena. The hot spot detection evolved from studying the distribution of the points or the spatial distribution of the points in space in order to comprehend the spatial patterns [33]. A visual dependency analysis revealed observations that, in the counties that were studied, the clusters that were defined as hot spots and cold spots included similar land cover types, thus allowing them to be characterized in terms of settlement type and land use.

6. Conclusions

The Relative Changes of the CCI showed the effect of the weights on the obtained results. The negative RC values revealed the predominance of the variant of the weighting with different weights (CCI_{w1}). According to the results, the highest share of negative RC values was shown in Sokólski (58.4%) and Piaseczno (51.1%) Counties. However, positive RC values proved the prevalence of the equal-weight variant (CCI_{w2}); this was especially evident in Ostrowski (83.7%), Międzyrzeczki (77.9%) and Sanocki (69.9%) Counties. The county with the most equal shares of the different weighting variants was Slupski (49.8% and 50.2%, respectively). The equal weights in the TOPSIS method influenced the number and, thus, the area, while both of the topographical data sets (BDOT10k and OSM) had the highest and moderate compliances. These differences varied from county to county, taking 4.1% (Międzyrzeczki) and a minimum of 0.2% (Sokólski) in the maximum-compliance and from 9.3 and 0.8% for the same counties in the moderate-compliance CCI classes. For 56% of the total area, the change in the weights altered the ranking by half a standard deviation. Relatively large changes of more than 2.5 standard deviations could be observed in 4% of the analyzed area. The demonstrated analyses prove that the studied data sets of OSM and BDOT10k were quite sensitive to the adopted weighting combinations.

A hot spot analysis of the CCI's Relative Changes indicated spatial relationships between the studied data sets despite the absence of a statistically significant Pearson correlation. Those areas that were identified as hot spots were mainly clustered in forests, open areas, cultivated areas, neighborhoods of water bodies, and (less frequently) areas with low building density. However, those areas that were identified as cold spots were found in the areas of urban-rural development and along major transportation lines.

Funding

The paper preparation was founded by the statutory research at the Military University of Technology, Faculty of Civil Engineering and Geodesy, Institute of Geospatial Engineering and Geodesy, Grant Number USG 531-4000-22-705.

CRedit Author Contribution

S. B.: conceptualization, methodology, formal analysis, data curation, writing – original draft preparation, writing – review and editing.

E. B.: conceptualization and result discussion, writing – review and editing.

K. P.: conceptualization, validation, writing – review and editing, supervision, funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Public Data:

OpenStreetMap data: <https://download.geofabrik.de>.

BDOT10k data: <https://mapy.geoportal.gov.pl>.

Use of Generative AI and AI-assisted Technologies

No generative AI or AI-assisted technologies were employed in the preparation of this manuscript.

References

- [1] Borkowska S., Bielecka E., Pokonieczny K.: *Comparison of land cover categorical data stored in OSM and authoritative topographic data*. Applied Sciences, vol. 13(13), 2023, 7525. <https://doi.org/10.3390/app13137525>.
- [2] Aksoy E., San B.T.: *Geographical information systems (GIS) and multi-criteria decision analysis (MCDA) integration for sustainable landfill site selection considering dynamic data source*. Bulletin of Engineering Geology and the Environment, vol. 78(4), 2019, pp. 779–791. <https://doi.org/10.1007/s10064-017-1135-z>.
- [3] Anthoff D., Tol R.S.J.: *On international equity weights and national decision making on climate change*. Journal of Environmental Economics and Management, vol. 60(1), 2010, pp. 14–20. <https://doi.org/10.1016/j.jeem.2010.04.002>.
- [4] Li P., Qian H., Wu J., Chen J.: *Sensitivity analysis of TOPSIS method in water quality assessment. Sensitivity to the parameter weights*. Environmental Monitoring and Assessment, vol. 185(3), 2013, pp. 2453–2461. <https://doi.org/10.1007/s10661-012-2723-9>.
- [5] Odu G.O.: *Weighting methods for multi-criteria decision making technique*. Journal of Applied Sciences and Environmental Management, vol. 23(8), 2019, pp. 1449–1457. <https://www.ajol.info/index.php/jasem>.
- [6] Roszkowska E.: *Rank ordering criteria weighting methods – a comparative overview*. Optimum. Studia Ekonomiczne, vol. 5(65), 2013, pp. 14–33. <https://doi.org/10.15290/ose.2013.05.65.02>.
- [7] Borkowska S., Bielecka E., Pokonieczny K.: *OpenStreetMap – building data completeness visualization in terms of ‘Fitness for purpose’*. Advances in Geodesy and Geoinformation, vol. 72(1), 2023, pp. 2–20. <https://doi.org/10.24425/agg.2022.141922>.
- [8] Borkowska S., Pokonieczny K.: *Analysis of OpenStreetMap data quality for selected counties in Poland in terms of sustainable development*. Sustainability, vol. 14(7), 2022, 3728. <https://doi.org/10.3390/su14073728>.
- [9] Behzadian M., Khanmohammadi Otaghsara S., Yazdani M., Ignatius J.: *A state-of-the-art survey of TOPSIS applications*. Expert Systems with Applications, vol. 39(17), 2012, pp. 13051–13069. <https://doi.org/10.1016/j.eswa.2012.05.056>.

-
- [10] Zyoud S.H., Fuchs-Hanusch D.: *A bibliometric-based survey on AHP and TOPSIS techniques*. Expert Systems with Applications, vol. 78, 2017, pp. 158–181. <https://doi.org/10.1016/j.eswa.2017.02.016>.
- [11] Çelikkbilek Y., Tüysüz F.: *An in-depth review of theory of the TOPSIS method: An experimental analysis*. Journal of Management Analytics, vol. 7(2), 2020, pp. 281–300. <https://doi.org/10.1080/23270012.2020.1748528>.
- [12] Wang Z.X., Wang Y.Y.: *Evaluation of the provincial competitiveness of the Chinese high-tech industry using an improved TOPSIS method*. Expert Systems with Applications, vol. 41(6), 2014, pp. 2824–2831. <https://doi.org/10.1016/j.eswa.2013.10.015>.
- [13] Kusumadewi S., Hartati S.: *Sensitivity analysis of multi-attribute decision making methods in Clinical Group Decision Support System*. [in:] *2007 International Conference on Intelligent and Advanced Systems: Kuala Lumpur, Malaysia: 25–28 November 2007*, IEEE, pp. 301–304. <https://doi.org/10.1109/ICIAS.2007.4658395>.
- [14] Dalalah D., Hayajneh M., Batieha F.: *A fuzzy multi-criteria decision making model for supplier selection*. Expert Systems with Applications. 2011, vol. 38(7), pp. 8384–8391. <https://doi.org/10.1016/j.eswa.2011.01.031>.
- [15] Choo E.U., Schoner B., Wedley W.C.: *Interpretation of criteria weights in multi-criteria decision making*. Computers & Industrial Engineering, vol. 37(3), 1999, pp. 527–541. [https://doi.org/10.1016/S0360-8352\(00\)00019-X](https://doi.org/10.1016/S0360-8352(00)00019-X).
- [16] Bączkiewicz A., Wątróbski J., Kizielewicz B., Sałabun W.: *Towards objectification of multi-criteria assessments: A comparative study on MCDA methods*. [in:] Ganzha M., Maciaszek L., Paprzycki M., Ślęzak D. (eds.), *Proceedings of the 16th Conference on Computer Science and Intelligence Systems: September 2–5, 2021*, Annals of Computer and Information Systems, vol. 25, IEEE, pp. 417–425. <https://doi.org/10.15439/2021F61>.
- [17] Kobryń A., Prystrom J.: *A data pre-processing model for the TOPSIS method*. Folia Oeconomica Stetinensia, vol. 16(2), 2016, pp. 219–235. <https://doi.org/10.1515/fofi-2016-0036>.
- [18] Pavić Z., Novoselac V.: *Notes on TOPSIS Method*. International Journal of Engineering Research and General Science, vol. 1(2), 2013, pp. 5–12.
- [19] Więckowski J., Zwiech P.: *Can weighting methods provide similar results in MCDA problems? Selection of energetic materials study case*. Procedia Computer Science, vol. 192, 2021, pp. 4592–4601. <https://doi.org/10.1016/j.procs.2021.09.237>.
- [20] Chen Y., Yu J., Khan S.: *Spatial sensitivity analysis of multi-criteria weights in GIS-based land suitability evaluation*. Environmental Modelling & Software, vol. 25, 2010, pp. 1582–1591. <https://doi.org/10.1016/j.envsoft.2010.06.001>.
- [21] Al-Mashreki M. H., Akhir J.B.M., Abd Rahim S., Tukimat L., Haider A.R.: *GIS-based sensitivity analysis of multi-criteria weights for land suitability evaluation of sorghum crop in the Ibb Governorate, Republic of Yemen*. Journal of Basic and Applied Scientific Research, vol. 1(9), 2011, pp. 1102–1111.

- [22] Liern V., Pérez-Gladish B.: *Multiple criteria ranking method based on functional proximity index: Un-weighted TOPSIS*. *Annals of Operations Research*, vol. 311, 2022, pp. 1099–1121. <https://doi.org/10.1007/s10479-020-03718-1>.
- [23] Dawid W., Pokonieczny K., Wyszzyński M.: *The methodology of determining optimum access routes to remote areas for the purposes of crisis management*. *International Journal of Digital Earth*, vol. 15(1), 2022, pp. 1905–1928. <https://doi.org/10.1080/17538947.2022.2134936>
- [24] Getis A., Ord K.: *The analysis of spatial association by use of distance statistics*. *Geographical Analysis*, vol. 24, 1992, pp. 189–206. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>.
- [25] Ord K., Getis A.: *Local spatial autocorrelation statistics: distributional issues and an application*. *Geographical Analysis*, vol. 27(4), 2010, pp. 286–306. <https://doi.org/10.1111/j.1538-4632.1995.tb00912.x>.
- [26] Jimenez J.: *Fitness for purpose in relation to specification limits*. *Accreditation and Quality Assurance*, vol. 17(1), 2012, pp. 27–34. <https://doi.org/10.1007/s00769-011-0825-7>.
- [27] Sheng J., Wilson J.P., Chen N., Deviny J.S., Sayre J.M.: *Evaluating the Quality of the National Hydrography Dataset for Watershed Assessments in Metropolitan Regions*. *GIScience & Remote Sensing*, vol. 44(3), 2007, pp. 283–304. <https://doi.org/10.2747/1548-1603.44.3.283>.
- [28] Bielecka E., Jenerowicz A.: *Intellectual structure of CORINE land cover research applications in web of science: A Europe-wide review*. *Remote Sensing*, vol. 11(17), 2019, 2017. <https://doi.org/10.3390/rs11172017>.
- [29] Bielecka E.: *Geographical data sets fitness of use evaluation*. *Geodetski Vestnik*, vol. 59(2), 2016, pp. 335–348. <https://doi.org/10.15292/geodetski-vestnik.2015.02.335-348>.
- [30] Bac-Bronowicz J., Dygaszewicz J., Grzempowski P., Nowak R.: *Bazy danych referencyjnych jako źródła zasilania i aktualizacji warstw dotyczących budynków w Wielorozdzielczej Topograficznej Bazie Danych*. *Roczniki Geomatyki*, t. 8, z. 5(41), 2010, pp. 7–22.
- [31] Biljecki F., Chow Y. S., Lee K.: *Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes*. *Building and Environment*, vol. 237, 2023, 110295. <https://doi.org/10.1016/j.buildenv.2023.110295>.
- [32] Marczak S.: *Ocena zaangażowania społeczeństwa w tworzenie danych przestrzennych w Polsce na przykładzie projektu OpenStreetMap*. *Roczniki Geomatyki*, t. 13, z. 3(69), 2015, pp. 239–253.
- [33] Chakravorty S.: *Identifying crime clusters: The spatial principles*. *Middle States Geographer*, vol. 28, 1995, pp. 53–58.

OŚWIADCZENIE

Niniejszym potwierdzam, że w ramach artykułów, stanowiących cykl publikacji:

- 1) **Borkowska, S., & Pokonieczny, K. (2022).** Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability*, 14, 3728. <https://doi.org/10.3390/su14073728>

Byłam odpowiedzialna za zaprojektowanie badania, opracowanie metodyki, pozyskanie oraz przetworzenie danych, wykonanie analiz oraz wizualizację i ocenę wyników. Opracowałam manuskrypt oraz brałam udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 90% całości artykułu.

- 2) **Borkowska, S., Bielecka, E., & Pokonieczny, K. (2023).** OpenStreetMap - building data completeness visualization in terms of "Fitness for purpose". *Advances in Geodesy and Geoinformation*, 72, 1, 1–20. <https://doi.org/10.24425/agg.2022.141922>

Byłam odpowiedzialna za zaprojektowanie badania, opracowanie metodyki, wykonanie analiz wraz z ich oceną statystyczną, wizualizację i ocenę wyników, sformułowanie wniosków. Opracowałam manuskrypt oraz brałam udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 75% całości artykułu.

- 3) **Borkowska, S., Bielecka, E., & Pokonieczny, K. (2023).** Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Applied Sciences*, 13, 7525. <https://doi.org/10.3390/app13137525>

Byłam odpowiedzialna za zaprojektowanie badania, opracowanie metodyki, wykonanie analiz, obliczenie wartości zaproponowanych wskaźników wraz z ich opisem statystycznym oraz wizualizację i ocenę wyników. Opracowałam manuskrypt oraz brałam udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 80% całości artykułu.

- 4) **Borkowska, S., Bielecka, E., & Pokonieczny, K. (2024).** Weights Impact on the Comparative Evaluation of Topographic Data. *Geomatics and Environmental Engineering*, 18, 4. <https://doi.org/10.7494/geom.2024.18.4.97>

Byłam odpowiedzialna za zaprojektowanie badania, opracowanie metodyki, wykonanie analiz i obliczenie wartości zaproponowanych wskaźników wraz z ich charakterystyką statystyczną, wizualizację i ocenę wyników oraz sformułowanie wniosków końcowych. Opracowałam manuskrypt oraz brałam udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 75% całości artykułu.



(podpis)

Elżbieta Bielecka
Wydział Inżynierii Lądowej i Geodezji
Wojskowa Akademia Techniczna im. J. Dąbrowskiego
Warszawa, Polska

Warszawa, 25.09.2024

OŚWIADCZENIE

Niniejszym potwierdzam, że w ramach artykułów, stanowiących cykl publikacji:

- 1) Borkowska, S., **Bielecka, E.**, & Pokonieczny, K. (2023). OpenStreetMap - building data completeness visualization in terms of "Fitness for purpose". *Advances in Geodesy and Geoinformation*, 72, 1, 1–20. <https://doi.org/10.24425/agg.2022.141922>

Uczestniczyłam w konsultacjach na etapie koncepcji badań, wykonałam korektę manuskryptu oraz brałam udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 12,5% całości artykułu.

- 2) Borkowska, S., **Bielecka, E.**, & Pokonieczny, K. (2023). Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Applied Sciences*, 13, 7525. <https://doi.org/10.3390/app13137525>

Byłam współodpowiedzialna za przegląd literatury oraz przygotowanie odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 8% całości artykułu.

- 3) Borkowska, S., **Bielecka, E.**, & Pokonieczny, K. (2024). Weights Impact on the Comparative Evaluation of Topographic Data. *Geomatics and Environmental Engineering*, 18, 4. <https://doi.org/10.7494/geom.2024.18.4.97>

Brałam udział w opracowaniu końcowej wersji manuskryptu oraz przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 10% całości artykułu.



(podpis)

OŚWIADCZENIE

Niniejszym potwierdzam, że w ramach artykułów, stanowiących cykl publikacji:

- 1) Borkowska, S., & **Pokonieczny, K.** (2022). Analysis of OpenStreetMap Data Quality for Selected Counties in Poland in Terms of Sustainable Development. *Sustainability*, 14, 3728. <https://doi.org/10.3390/su14073728>

Uczestniczyłem w zaprojektowaniu badań i opracowaniu metodyki. Brałem udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 10% całości artykułu.

- 2) Borkowska, S., Bielecka, E., & **Pokonieczny, K.** (2023). OpenStreetMap - building data completeness visualization in terms of "Fitness for purpose". *Advances in Geodesy and Geoinformation*, 72, 1, 1–20. <https://doi.org/10.24425/agg.2022.141922>

Byłem współodpowiedzialny za opracowanie koncepcji badań oraz korektę manuskryptu. Brałem udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 12,5% całości artykułu.

- 3) Borkowska, S., Bielecka, E., & **Pokonieczny, K.** (2023). Comparison of Land Cover Categorical Data Stored in OSM and Authoritative Topographic Data. *Applied Sciences*, 13, 7525. <https://doi.org/10.3390/app13137525>

Uczestniczyłem w opracowaniu metodyki badań oraz brałem udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 12% całości artykułu.

- 4) Borkowska, S., Bielecka, E., & **Pokonieczny, K.** (2024). Weights Impact on the Comparative Evaluation of Topographic Data. *Geomatics and Environmental Engineering*, 18, 4. <https://doi.org/10.7494/geom.2024.18.4.97>

Byłem współodpowiedzialny za opracowanie koncepcji badań. Wykonałem korektę manuskryptu oraz brałem udział w przygotowaniu odpowiedzi na uwagi recenzentów.

Mój wkład oceniam na 15% całości artykułu.

.....

(podpis)