POLSKIE TOWARZYSTWO INFORMACJI PRZESTRZENNEJ
ROCZNIKI GEOMATYKI 2017 ❍ Tom XV ❍ Zeszyt 2(77): 233–248

Wyznaczenie cech społeczeństwa wpływających na zaangażowanie w tworzenie VGI w Polsce

Determination of socioeconomic features of a society influencing the involvement in VGI creation in Poland

Sylwia Marczak

Politechnika Warszawska, Wydział Geodezji i Kartografii,
Zakład Fotogrametrii, Teledetekcji i Systemów Informacji Przestrzennej

Słowa kluczowe: OpenStreetMap, społecznościowe dane przestrzenne, korelacja, regresja liniowa wieloraka, regresja ważona geograficznie
Keywords: OpenStreetMap, volunteered geographic information, correlation, linear regression analysis, geographically weighted regression

Wstęp

Od momentu powstania terminu *voluntereed geographic information* (VGI), sformułowanego przez Goodchild'a w 2007 roku można zaobserwować znaczny wzrost znaczenia danych przestrzennych zbieranych na zasadzie wolontariatu, przez użytkowników Internetu niebędących profesjonalistami w tym zakresie. Wzrost ten dotyczy zarówno rozwoju serwisów umożliwiających tworzenie VGI, jak i ciągle zwiększającej się liczby użytkowników – wolontariuszy, a także zainteresowania naukowców z całego świata zagadnieniami między innymi jakości i możliwości zastosowań tego rodzaju danych.

Najważniejszym czynnikiem wpływającym na tworzenie danych w oparciu o *crowdsourcing* (ang. *crowd* – tłum, ang. *sourcing* – czerpanie) była technologia Web 2.0, która umożliwia użytkownikom Internetu edytowanie istniejących lub tworzenie nowych treści (Neis, Zielstra, 2014). Spowodowało to powstanie i szybki rozwój serwisów społecznościowych, między innymi Facebooka, Twittera oraz innych, działających w oparciu o dane tworzone przez użytkowników między innymi YouTube lub Wikipedia. Czynnikiem mającym bezpośredni wpływ na tworzenie przez użytkowników Internetu danych przestrzennych było upowszechnienie korzystania z sygnału z systemu GPS (*Global Positioning System)*, który obecnie dostępny jest nie tylko w telefonach komórkowych, ale również w zegarkach, bądź tak zwanych inteligentnych ubraniach. Przyczyniło się to do powstania serwisów umożliwiających tworzenie i korzystanie z VGI, z których najbardziej popularny to OpenStreetMap (OSM), ale są to również WikiMapia lub Google Map Maker. Należy w tym miejscu zwrócić

uwagę na fakt, że bez względu na serwis mamy do czynienia z pewną społecznością twórców danych. W przypadku polskiej wersji językowej serwisu OSM – openstreetmap.org.pl termin ten użyty został do opisu strony *Portal polskiej społeczności OpenStreetMap.* W ramach tej inicjatywy organizowane są również spotkania pod nazwą *mapping party,* w czasie których sympatycy OSM tworzą dane przestrzenne dla okolicy, w której odbywa się impreza. Nie ma zatem wątpliwości, że serwisy umożliwiające tworzenie VGI znacząco przyczyniają się do popularyzacji wiedzy geoprzestrzennej wśród społeczeństwa. Na profilu Facebook OpenStreetMap Polska można przeczytać – *Stowarzyszenie OpenStreetMap Polska to organizacja non-profit, którego celem jest promocja i wspieranie projektu OpenStreetMap na terenie Polski, ale także popularyzowanie idei wolnej kartografii oraz wykorzystania jej dla ogólnospołecznych celów takich jak popularyzacja wiedzy z zakresu geodezji i kartografii czy wspieranie tworzenia, gromadzenia i rozpowszechniania ogólnodostępnych danych geograficznych.* W związku z powyższym zaproponowany przez Marczak (2015) polski odpowiednik terminu *voluntereed geographic information* – „społecznościowe dane przestrzenne" wydaje się być słuszny.

W literaturze światowej istnieje wiele pozycji opisujących badania nad zjawiskiem *crowdsourcingu* w różnych jego aspektach począwszy od jakości danych (m.in. Haklay, 2010; Girres, Touya, 2010; Marczak, 2015; Nowak Da Costa i in., 2016), przez analizy użytkowników tworzących i korzystających z danych (m.in. Neis, Zipf, 2012; Budhathoki, 2010), aż po możliwości zastosowań społecznościowych danych przestrzennych (m.in. Arsanjani, Vaz, 2015; Cichociński, 2012; Cichociński, Dębińska, 2012). Wszystkie przytoczone powyżej prace dotyczyły danych pobranych z serwisu OSM, a ich wspólny wniosek to zjawisko wysokiej heterogeniczności zarówno w odniesieniu do danych OSM, których zdecydowanie większa ilość znajduje się na obszarach miejskich, jak i społeczności je tworzącej, której zróżnicowanie dotyczy zarówno liczby użytkowników, jak i poziomu ich zaangażowania.

Ze względu na fakt, że w OSM nie są zbierane dane dotyczące użytkowników, takie jak wiek bądź miejsce zamieszkania, kilka pozycji w literaturze zagranicznej dotyczyło próby scharakteryzowania przeciętnego użytkownika projektu. Można tu wyróżnić dwa zasadnicze podejścia. Pierwsze z nich to przeprowadzenie ankiety wśród użytkowników OSM i statystyczne opracowanie wyników (m.in. Haklay, Budhathoki, 2010; Stephens, Rondinone, 2012; Steinmann i in., 2013; Schmidt, Klettner, 2013). Z badań tych wynika, że przeciętny użytkownik tworzący VGI to mężczyzna z wyższym wykształceniem w wieku od 20 do 50 lat. Ponadto w badaniu przeprowadzonym przez Haklay i Budhathoki (2010) wśród 426 ankietowanych, aż 51% posiadało wiedzę geoprzestrzenną. Natomiast wśród 389 mężczyzn z badania przeprowadzonego przez Schmidt i Klettner (2013) 56% zadeklarowała korzystanie z serwisu OSM (w tym tworzenie danych) w celach prywatnych, podczas gdy wśród 122 kobiet było to 33,8%, a najczęściej wskazywanymi przyczynami korzystania z OSM była praca (61,5%).

Drugie podejście dotyczące próby scharakteryzowania społeczności OSM polega na znalezieniu korelacji między cechami demograficznymi społeczeństwa danego obszaru, a ilością tworzonych danych VGI na tym obszarze. Wymaga to założenia, że wśród społeczeństwa danego obszaru są użytkownicy OSM tworzący dane na tym obszarze. Dotychczasowe doświadczenia pokazują, że założenie to jest zasadne, gdyż zdecydowana większość wolontariuszy OSM to użytkownicy lokalni (*local mappers*) (Neis i in., 2013). W pracy Mashhadi i in. (2013) autorzy badali zależność między liczbą punktów POI, stworzonych przez użytkowników OSM w poszczególnych dzielnicach Londynu a gęstością zaludnienia, liczbą lud-

ności przypadającą na 1 punkt POI, liczbą osób bezdomnych, liczbą wizyt (wyrażoną liczbą zameldowań w serwisie Foursquare) oraz odległością do najbliższego obszaru metropolitarnego (*poly-centre*). W badaniu wykorzystano regresję liniową jednokrotną i wielokrotną do stwierdzenia czy istnieje wpływ cech społeczeństwa na ilość danych OSM. Otrzymany skorygowany współczynnik determinacji na poziomie 0,17 dla Londynu i 0,16 dla tak zwanego Londynu Wewnętrznego wskazał jednak na niski stopień objaśniania, co nie pozwoliło na wskazanie cech społeczeństwa mających szczególnie duży wpływ na ilość tworzonych danych VGI. W pracy Arsanjani i Bakillah (2015) zastosowano inne podejście, które zakładało w pierwszym etapie wyznaczenie obszarów o szczególnie dużej liczbie danych OSM (*hot spots*), następnie pozyskanie danych demograficznych na poziomie powiatów w niemieckim landzie Badenia-Wirtembergia i ułożenie modeli regresji logistycznej. Wynikiem badania jest stwierdzenie, iż wysoki wpływ na pozyskiwanie szczególnie dużej liczby danych OSM mają takie cechy społeczeństwa jak: gęstość zaludnienia, poziom wykształcenia, średnie wynagrodzenie, turystyka (wyrażona liczbą pobytów czasowych na co najmniej jedną noc), wiek, liczba cudzoziemców i bliskość obszarów zabudowanych.

Celem niniejszego artykułu jest uzupełnienie tych prac o badanie dotyczące obszaru Polski i wyznaczenie takich cech społeczeństwa, które mają szczególnie duży wpływ na liczbę tworzonych danych OSM w powiatach, przy założeniu, że może on być różny w zależności od typu geometrycznego danych przestrzennych. Należy spodziewać się, iż można wyznaczyć te cechy przy jednocześnie wysokim stopniu objaśniania zmiennej zależnej, co pozwoli na prognozowanie rozwoju serwisu OSM w przyszłości.

Zmienne objaśniane – wykorzystane dane z projektu OpenStreetMap

OpenStreetMap jest bez wątpienia najpopularniejszym serwisem umożliwiającym tworzenie społecznościowych danych przestrzennych i korzystanie z nich. Liczba zarejestrowanych użytkowników projektu wzrasta nieprzerwanie od 2004 roku, kiedy projekt został zainicjowany w Wielkiej Brytanii przez Steve'a Coasta (Zielstra, Zipf, 2010). Obecnie (stan na 27 października 2016 roku) w serwisie zarejestrowanych jest 3 185 114 użytkowników, którzy tworzą dane przestrzenne dwoma głównymi sposobami – wektoryzując zdjęcia satelitarne lub lotnicze, bądź wgrywając ścieżki lub punkty z odbiorników GPS.

Struktura bazy danych OSM jest różna od powszechnie stosowanej w systemach informacji geograficznej struktury relacyjnej. Składają się na nią trzy rodzaje obiektów – węzły (*nodes*), linie (*ways*) i relacje (*relations*). Za pomocą węzłów tworzone są obiekty punktowe, natomiast linii – liniowe i poligonowe. Powiązania między obiektami reprezentowane są za pomocą relacji. Cechy obiektów zapisywane są za pomocą tagów przyjmujących postać klucz-wartość (Cichociński, 2012). Różnice w modelu OSM i relacyjnym sprawiają trudności w korzystaniu z danych w oprogramowaniu typu GIS. Pewnym rozwiązaniem tego problemu jest skorzystanie z danych w postaci plików .shp, które udostępniane są przez firmę Geofabrik. Niestety darmowo można skorzystać tylko z 8 klas obiektów, co znacząco zmniejsza użyteczność tego produktu. Firma ta udostępnia również dane w formacie .osm xml, które zawierają pełną strukturę OSM, a w przypadku Polski podzielone są na województwa. Warto tutaj zaznaczyć, że objętość jednego zbioru w zależności od wielkości województwa waha się od 600 MB do 2,4 GB i rośnie z każdą aktualizacją, co znacznie wpływa na czas przetwarzania i analizowania danych zapisanych w tym formacie.

W niniejszej pracy dane OSM zostały wykorzystane jako zmienne objaśniane w modelach regresji, a także jako zmienne do określenia stopnia korelacji między cechami demograficznymi społeczeństwa a ilością danych VGI. W tym celu w pierwszym kroku dane pobrane w formacie .osm xml (o aktualności na dzień 19.07.2016 roku) zostały zaimportowane do geobazy plikowej za pomocą zestawu narzędzi „ArcGIS Editor for OSM" stworzonego przez firmę Esri na potrzeby korzystania z danych OSM. Wynikiem importu dla każdego województwa były trzy klasy obiektów – dane punktowe, liniowe i poligonowe. Następnie w oparciu o atrybut *osmtimestamp* z każdego zbioru zostały wybrane obiekty, które powstały w okresie od 1 stycznia 2013 do 31 grudnia 2015 roku. Kolejnym krokiem było zliczenie liczby punktów, długości linii i powierzchni poligonów wybranych obiektów w powiatach w Polsce. Na koniec wartości te zostały podzielone przez powierzchnię każdego z powiatów. W ten sposób powstały trzy zmienne objaśniane – liczba punktów OSM na 1 km^2, długość linii OSM na 1 km^2 i procentowe pokrycie danymi poligonowymi OSM powiatów w Polsce powstałymi w okresie trzech lat – od 1.01.2013 do 31.12.2015 rok.

Zmienne objaśniające – wybór cech demograficznych społeczeństwa

Biorąc pod uwagę cel niniejszego artykułu, którym jest wyznaczenie cech społeczeństwa mających istotny wpływ na pozyskiwanie danych z OSM, na podstawie przeglądu literatury i dostępności danych na poziomie powiatowym wybrano 15 zmiennych, które zostały w późniejszych etapach wykorzystane do analiz korelacji i regresji. Wszystkie dane zostały pozyskane z Banku Danych Lokalnych, prowadzonego przez Główny Urząd Statystyczny. Większość cech została wybrana w oparciu o cechy wskazane w badaniach Mashhadi i in. (2013) oraz Arsanjani i Bakillah (2015). Ponadto biorąc pod uwagę dostępność danych i analizując ich możliwy wpływ na pozyskiwanie danych VGI zbiór ten uzupełniono o takie cechy jak:

- małżeństwa zawarte na 1000 ludności – zakładając, że im jest ich więcej tym liczba pozyskiwanych danych jest mniejsza ze względu na mniejszą ilość czasu osób w związkach małżeńskich będących jednocześnie użytkownikami OSM;
- liczba fundacji, stowarzyszeń i organizacji społecznych na 10 tys. ludności – zakładając, że ich liczba pozytywnie wpływa na ilość pozyskiwanych danych, gdyż osoby działające w organizacjach *non-profit* mają większe predyspozycje do działań na rzecz społeczeństwa, za które można uznać tworzenie danych VGI;
- frekwencja wyborcza w wyborach samorządowych w 2014 roku – zakładając, że wyższa frekwencja przekłada się na większą liczbę pozyskiwanych danych, ze względu na większe zainteresowanie sprawami lokalnej społeczności ludności, co może mieć wyraz również w tworzeniu VGI;
- turyści zagraniczni z Niemiec – według literatury (Neis, Zipf, 2012) z Niemiec pochodzi najwięcej aktywnych użytkowników OSM, a co za tym idzie tworzona jest największa ilość danych, w związku z tym założono, że liczba niemieckich turystów pozytywnie wpływa na ilość tworzonych danych w polskich powiatach, gdyż mogą się wśród nich znajdować amatorzy tworzenia danych VGI.

W miarę dostępności cechy zostały pobrane dla trzech lat – 2013, 2014, 2015. Następnie dla każdej z nich obliczono wartości średnie, które zostały wykorzystane do dalszych analiz jako zmienne objaśniające. Spis wybranych zmiennych przedstawia tabela 1.

Tabela 1. Wybrane cechy demograficzne społeczeństwa przyjęte jako zmienne objaśniające

Lp.	Skrót	Nazwa	Jednostka	Lata
1	GZ	Gęstość zaludnienia	os/km^2	2013, 2014, 2015
2	W1	Procent liczby ludności w wieku do 20 lat	%	2013, 2014, 2015
3	W2	Procent liczby ludności w wieku 20-30 lat	%	2013, 2014, 2015
4	W3	Procent liczby ludności w wieku 30-40 lat	%	2013, 2014, 2015
5	W4	Procent liczby ludności w wieku 40-50 lat	%	2013, 2014, 2015
6	W5	Procent liczby ludności w wieku powyżej 50 lat	%	2013, 2014, 2015
7	M	Małżeństwa zawarte na 1000 ludności	liczba	2013, 2014, 2015
8	PS	Liczba osób w gospodarstwach domowych otrzymująca pomoc społeczną na 1000 ludności	osoba	2013, 2014
9	SB	Stopa bezrobocia rejestrowanego	%	2013, 2014, 2015
10	WF	Współczynnik feminizacji	–	2013, 2014, 2015
11	SW	Średnie wynagrodzenie brutto	zł	2013, 2014, 2015
12	WW	Procent ludności w wieku 13 lat i więcej z wykształceniem wyższym	%	2011
13	FS	Liczba fundacji, stowarzyszeń i organizacji społecznych na 10 tys. mieszkańców	liczba	2013, 2014, 2015
14	FW	Frekwencja wyborcza w wyborach samorządowych w 2014 roku	%	2014
15	TZ	Turyści zagraniczni (nierezydenci) na 1000 mieszkańców	osoba	2013, 2014, 2015
16	TN	Turyści zagraniczni będący obywatelami Niemiec (nierezydenci) na 1000 mieszkańców	osoba	2013, 2014, 2015

Przyjęta metodyka

Wyznaczenie cech społeczeństwa mających szczególnie istotny wpływ na liczbę pozyskiwanych danych OSM podzielono na trzy zasadnicze etapy. Pierwszy z nich zakładał wyznaczenie współczynników korelacji między zmiennymi objaśnianymi i wszystkimi zmiennymi objaśniającymi. Następnie te zmienne, dla których współczynnik ten był istotny statycznie, zostały wykorzystane do ułożenia modeli regresji liniowej wielorakiej. W kolejnym etapie zbadano autokorelację przestrzenną reszt regresji liniowej, chcąc określić zasadność zastosowania regresji ważonej geograficznie. Ostatnim etapem było porównanie wyników z regresji liniowej i regresji ważonej geograficznie i zidentyfikowanie cech społeczeństwa mających szczególnie istotny wpływ na pozyskiwanie społecznościowych danych przestrzennych. Do wyznaczenia siły zależności między zmiennymi objaśniającymi i objaśnianymi wykorzystano współczynnik korelacji liniowej Pearsona, który wyraża się następującym wzorem (Koop, 2011):

$$r = \frac{\Sigma^{N}_{i=1} (Y_i - \overline{Y}) (X_i - \overline{X})}{\sqrt{(\Sigma^{N}_{i=1} (Y_i - \overline{Y})^2} \sqrt{(\Sigma^{N}_{i=1} (X_i - \overline{X})^2}}$$

gdzie: X_i – kolejne obserwacje zmiennej X, Y_i – kolejne obserwacje zmiennej Y, $\overline{X}$, $\overline{Y}$ – średnie wartości zmiennych X, Y.

Współczynnik ten przyjmuje wartości z przedziału <-1;1>, gdzie wartości bliskie 1 oznaczają silną dodatnią korelację, bliskie -1 ujemną korelację, a o braku korelacji świadczą wartości współczynnika bliskie zeru (Koop, 2011). Oprócz wyznaczenia samego współczynnika korelacji przeprowadzono również test istotności na poziomie α=0,001, który wykazał, że dla liczby obserwacji wynoszącej 380 (liczba powiatów w Polsce) wartości współczynnika *r* większe od 0,17 są istotne statystycznie. Statystyka testowa ma rozkład t-Studenta i przyjmuje następującą postać:

$$T_{n-2} = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

gdzie: *n* – liczba obserwacji, *r* – wartość współczynnika korelacji Pearsona. Za jej pomocą weryfikowana jest hipoteza zerowa $H_0 : \rho = 0$ – brak zależności liniowej między badanymi zmiennymi, przeciw alternatywnej $H_0 : \rho \neq 0$.

Oprócz korelacji Pearsona obliczono również współczynniki korelacji Spearmana i Kendalla, które są ogólniejsze od współczynnika korelacji Pearsona – są miernikami monotonicznych zależności (nie tylko liniowych) między zmiennymi, a także są odporniejsze na obserwacje odstające. Wartości tych współczynników mieszczą się w przedziale <-1;1>, gdzie wartości bliskie 1 świadczą o silnie dodatniej monotonicznej zależności między zmiennymi, a bliskie -1 o silnie ujemnej zależności.

Wszystkie współczynniki korelacji zostały obliczone dla trzech zmiennych objaśnianych i wszystkich zmiennych objaśniających dla powiatów w Polsce.

Należy pamiętać, że korelacja pozwala na stwierdzenie czy zależność między zmiennymi istnieje i jaka jest jej siła, nie pozwala natomiast na określenie czy zachodzi objaśnianie zmiennej zależnej zmiennymi niezależnymi. Do określenia tego rodzaju związków służy regresja. W niniejszej pracy zostało wykorzystane modelowanie w oparciu o regresję liniową wieloraką, która wyraża się następującym wzorem (Koop, 2011):

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \quad i=1, 2, \ldots, n$$

gdzie: *i* – numery pojedynczych obserwacji, $\alpha_1, \beta_1 \ldots, \beta_k$ – nieznane parametry modelu, ε_i – błąd (reszta) modelu.

Oszacowania nieznanych parametrów modelu dokonuje się z wykorzystaniem metody najmniejszych kwadratów, natomiast oceny dopasowania modelu za pomocą współczynnika determinacji R^2 i skorygowanego współczynnika determinacji R^2_{adj}, który służy do porównywania jakości modeli o różnej liczbie zmiennych. Wartości R^2 zawierają się w przedziale <0;1>, im wartość współczynnika bliższa 1 tym lepsze dopasowanie modelu. Oprócz współczynnika determinacji do wyboru najlepszego modelu stosowane jest również kryterium informacyjne Akaikego AIC (*Akaike Information Criterion*), im mniejsza jego wartość tym model jest lepszy. Proces modelowania był iteracyjny i wykorzystywał jedną z metod doboru zmiennych objaśniających do modeli ekonometrycznych – regresję krokową wsteczną. W pierwszym kroku jako zmienne objaśniające zostały wykorzystane wszystkie zmienne istotnie skorelowane ze zmiennymi objaśnianymi, następnie z modelu usuwana była ta z nich, dla której bezwzględna minimalna wartość statystyki t-Studenta była mniejsza od wartości krytycznej $t_{n-(k+1),\,1-\alpha/2}$, gdzie *n* – liczba obserwacji, *k* – liczba zmiennych objaśniających w modelu, wyznaczonej dla poziomu istotności 0,05. Następnie obliczano nowy model i powtarzano powyższą procedurę, aż do osiągnięcia wartości statystyk t-Studenta dla wszystkich zmiennych większych od wartości krytycznej. Istotną kwestią na etapie budowania

modelu jest określenie stopnia skorelowania (współliniowości) zmiennych objaśniających między sobą, co ma znaczący wpływ na otrzymywane wyniki. Stopień skorelowania można ocenić obliczając czynnik inflacji wariancji (VIF), który mówi o tym ile razy wyznaczona wariancja estymatora jest większa od wariancji prawdziwej – niezakłóconej współliniowością (Gruszczyński i in., 2003). Przy braku współliniowości VIF=1, uważa się, że VIF>10 (według innych źródeł VIF>5) świadczy o znaczących zakłóceniach wywołanych współliniowością (Gruszczyński i in., 2003). Ostatecznie wyznaczono 3 modele regresji liniowej wielorakiej, w których za zmienne objaśniane przyjęto: (1) liczbę punktów OSM na 1 km^2, (2) długość linii OSM na 1 km^2 i (3) procentowe pokrycie danymi poligonowymi OSM. Kolejnym krokiem było testowanie normalności reszt z regresji co jest warunkiem poprawności przeprowadzonego modelowania. Istnieje wiele testów sprawdzających normalność rozkładu, w niniejszej pracy wykonano wykresy kwantylowe dla reszt (QQ plot) oraz wykonano test Shapiro-Wilka. Pierwszą metodę można zaliczyć do metod wizualnych, gdyż polega ona na ocenie czy punkty wykresu leżą wzdłuż prostej. Warunek ten w przybliżeniu był spełniony dla wszystkich zbudowanych modeli, co potwierdził test Shapiro-Wilka. Hipotezą zerową w tym teście jest stwierdzenie, iż badany rozkład jest normalny, a alternatywną, że nie można mówić o normalności rozkładu. Jeśli obliczona wartość *p-value* jest większa od α należy przyjąć hipotezę zerową. Dla przyjętego poziomu istotności α=0,001 otrzymano następujące wartości *p-value* – 0,06 dla modelu 1, 0,2 dla modelu 2 i 0,1 dla modelu 3. Następnie wyznaczono reszty z regresji i zbadano ich rozkład przestrzenny, co było niezbędne do określenia zasadności modelowania z wykorzystaniem regresji ważonej geograficznie. W tym celu zastosowano globalną statystykę Morana I, której wartości mieszczą się w przedziale <-1,1>, a jej interpretacja jest następująca (Marczak, Pluto-Kossakowska, 2015):

- I>0 – zachodzi dodatnia korelacja przestrzenna,
- I≈0 – brak autokorelacji,
- I<0 – zachodzi ujemna korelacja przestrzenna.

Dla tych modeli, dla których stwierdzono występowanie autokorelacji przestrzennej reszt z regresji liniowej zastosowano regresję ważoną geograficznie. Zakłada ona, że badane zjawisko charakteryzuje się niestacjonarnością, czyli różnym stopniem oddziaływania czynników sprawczych w zależności od położenia jednostki odniesienia zmiennych w przestrzeni geograficznej. Regresja ważona geograficznie (*Geographically Weighted Regression,* GWR) umożliwia oszacowanie parametrów modelu w każdej jednostce odniesienia, dla której znane są wartości zmiennych zależnych i niezależnych, co potwierdza poniższa postać modelu (Cellmer, 2010):

$$Y_i = \beta_0 (x_i, y_i) + \beta_1 (x_i, y_i) X_{1i} + \beta_2 (x_i, y_i) X_{2i} + ...+ \beta_k (x_i, y_i) X_{ki} + \varepsilon_i \text{ dla } i = 1,2,...n$$

gdzie parametry β_k są związane z lokalizacją, wyrażoną współrzędnymi x_i, y_i.

Oceny dopasowania modelu dokonano analogicznie jak w regresji liniowej. Ostatecznie wyznaczono te cechy społeczeństwa, które mają istotny wpływ na pozyskiwanie danych OSM, co biorąc pod uwagę zastosowane modele regresyjne umożliwia prognozowanie przyrostu społecznościowych danych przestrzennych w powiatach Polski.

Wyniki analizy korelacji

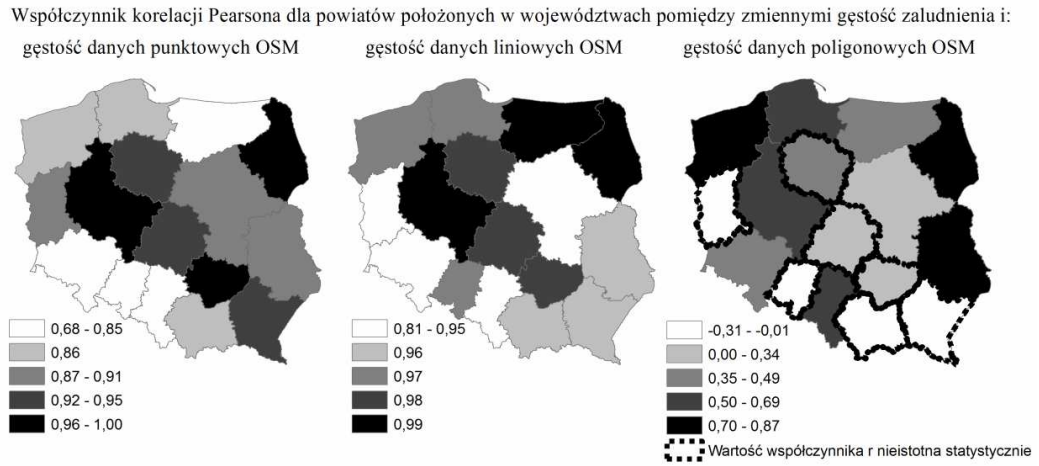
W pierwszym etapie – analizie korelacji – wyznaczono w sumie 144 współczynniki korelacji Pearsona, Spearmana i Kendalla z czego 112 było istotnych statystycznie. Wyniki dla wszystkich zmiennych objaśnianych i objaśniających przedstawiono w tabeli 2. Należy zwrócić uwagę na znaczne różnice w sile skorelowania w zależności od typu geometrycznego pozyskiwanych danych OSM. Największe bezwzględne wartości wszystkich współczynników korelacji otrzymano dla danych liniowych, a najmniejsze dla danych poligonowych. Należy zauważyć, że dla danych tych brak istotności współczynników korelacji Spearmana i Kendalla występuje dla tych samych zmiennych, dla których nieistotna jest korelacja Pearsona. Dla danych punktowych i liniowych występuje inna zależność – brak istotności korelacji Spearmana i Kendalla występuje dla większej liczby zmiennych niż w przypadku braku istotności korelacji Pearsona. Oznacza to, że dla danych poligonowych dla największej liczby zmiennych nie można mówić ani o zależności monotonicznej ani tym bardziej liniowej. Natomiast dla danych liniowych i punktowych można wskazać zmienne objaśniające, dla których zachodzi tylko zależność liniowa, gdyż współczynniki korelacji Spearmana i Kendalla są nieistotne. Największe wartości wszystkich rodzajów korelacji uzyskano dla zmiennej gę-

Tabela 2. Wartości współczynników korelacji Pearsona (r) dla zmiennych objaśnianych i objaśniających; kolorem szarym zaznaczono wartości nieistotne statystycznie

Skrót zmiennych objaśniających	Zmienne objaśniane								
	liczba punktów OSM/1 km²			długość linii OSM/1 km²			procentowe pokrycie danymi poligonowymi OSM powiatów		
	współczynniki korelacji			współczynniki korelacji			współczynniki korelacji		
	Pearsona	Spearmana	Kendalla	Pearsona	Spearmana	Kendalla	Pearsona	Spearmana	Kendalla
GZ	0,75	0,79	0,61	0,92	0,87	0,70	0,48	0,34	0,23
W1	-0,37	-0,31	-0,20	-0,50	-0,36	-0,23	-0,27	-0,25	-0,17
W2	-0,40	-0,45	-0,29	-0,50	-0,50	-0,33	-0,33	-0,31	-0,21
W3	0,30	0,15	0,10	0,34	0,41	0,28	0,18	0,21	0,14
W4	-0,21	-0,11	-0,07	-0,21	0,02	0,03	-0,08	-0,10	-0,07
W5	0,30	0,32	0,21	0,40	0,26	0,16	0,23	0,21	0,14
M	-0,23	-0,17	-0,11	-0,30	-0,32	-0,22	-0,12	-0,15	-0,10
PS	-0,40	-0,61	-0,43	-0,49	-0,67	-0,47	-0,33	-0,31	-0,21
SB	-0,32	-0,51	-0,35	-0,36	-0,42	-0,29	-0,15	-0,12	-0,08
WF	0,63	0,58	0,41	0,77	0,68	0,49	0,31	0,29	0,19
SW	0,31	0,35	0,23	0,44	0,44	0,30	0,33	0,28	0,19
WW	0,67	0,62	0,44	0,77	0,67	0,48	0,29	0,35	0,24
FS	0,22	-0,12	-0,10	0,25	-0,03	-0,03	-0,08	-0,08	-0,05
FW	-0,42	-0,23	-0,15	-0,56	-0,49	-0,34	-0,33	-0,25	-0,17
TZ	0,14	0,16	0,11	0,18	0,37	0,27	0,19	0,22	0,15
TN	0,00	-0,01	0,00	0,01	0,30	0,21	0,13	0,23	0,15

stość zaludnienia, przy czym w zależności od długości danych liniowych OSM wynosiła ona aż 0,92 (korelacja Pearsona), podczas gdy dla danych poligonowych było to zaledwie 0,23 (korelacja Kendalla). Wysokie wartości dodatniej korelacji uzyskano również dla zmiennych – procent ludności z wykształceniem wyższym i współczynnik feminizacji, o ile w przypadku pierwszej z nich należało się spodziewać takiego wyniku, o tyle w przypadku drugiej jest on zaskakujący, gdyż według literatury zdecydowanie więcej danych OSM tworzonych jest przez mężczyzn. Największe wartości ujemnej korelacji, świadczące o wzroście jednej zmiennej przy jednoczesnym spadku drugiej, uzyskano dla zmiennej dotyczącej frekwencji wyborczej, co wskazuje iż przyjęte założenie o pozytywnym wpływie tej zmiennej na pozyskiwanie danych OSM było błędne. Zaskakujące wyniki uzyskano również dla procentowego udziału ludności w wieku do 20 i od 20 do 30 lat. Należało się spodziewać, że zmienne te są dodatnio skorelowane ze zmiennymi objaśnianymi, podczas gdy korelacja ta jest ujemna. Dodatnie wartości współczynników korelacji uzyskano dla grup wiekowych 30-40 lat i powyżej 50 lat, o ile pierwsza z nich nie powinna dziwić, o tyle druga jest dosyć zaskakująca, chociaż siła związku jest na średnim poziomie. Wartości wszystkich współczynników korelacji dla zmiennych związanych z turystyką są zbliżone do zera i w większości nieistotne statystycznie.

Oprócz wyznaczenia współczynników korelacji dla wszystkich powiatów w Polsce, obliczeń dokonano również dla powiatów każdego z województw z osobna. Pozwala to stwierdzić czy siła związków zachodzących między zmiennymi objaśnianymi i objaśniającymi jest różna w zależności od regionu Polski (rys. 1).



Rysunek 1. Współczynnik korelacji Pearsona obliczony dla powiatów położonych w poszczególnych województwach Polski

Należy przy tym pamiętać, że w związku z mniejszą liczbą obserwacji (liczba powiatów w poszczególnych województwach) za istotne należy uznać zdecydowanie większe wartości współczynnika *r*. Na postawie analizy współczynnika *r* w województwach można stwierdzić, że występuje stosunkowo niska korelacja między zmienną gęstość zaludnienia a zmiennymi objaśnianymi w regionie Śląska i Dolnego Śląska, co jest dosyć zaskakujące biorąc pod uwagę fakt, że to jedne z liczniej zaludnionych obszarów kraju. Wynika z tego iż w regionach

tych inne cechy społeczeństwa są powiązane zależnościami z ilością pozyskiwanych danych VGI. Podobne analizy przeprowadzono również dla innych zmiennych objaśniających – wnioski z nich przedstawiono w rozdziale „Podsumowanie i wnioski".

Wyniki modelowania za pomocą regresji liniowej i GWR

Na podstawie analizy korelacji wyznaczono modele regresji liniowej zawierające wszystkie zmienne objaśniane o istotnym statystycznie współczynniku *r*. Następnie modele były optymalizowane metodą eliminacji, której etapy nie zostaną przedstawiona w niniejszym artykule, ze względu na ograniczenia w jego długości. Ostatecznie wyznaczono trzy modele regresji dla każdej ze zmiennych objaśnianych (tab. 3).

Tabela 3. Parametry modeli regresji liniowej o najlepszym dopasowaniu

Zmienna objaśniająca	Wartość parametru	Błąd standardowy	Poziom istotności *(p-value)*	Wartość statystyki t-Studenta	VIF
Model 1 Zmienna zależna: liczba punktów OSM/1 km^2					
Stała	-47,344	27,410	0,085	-1,727	–
GZ	0,053	0,006	0,000	8,801	2,14
M	6,953	3,464	0,045	2,007	1,21
SW	-0,007	0,004	0,048	-1,983	1,39
WW	3,929	1,184	0,001	3,318	2,81
R^2 = 0,6006, R^2_{adj} = 0,596, AIC= 3913,8					
Model 2 Zmienna zależna: długość linii OSM/1 km^2					
Stała	5749,773	1420,150	0,000	4,049	–
GZ	6,953	0,662	0,000	10,500	2,49
PS	-12,764	4,126	0,002	-3,093	1,51
WW	187,699	69,860	0,008	2,687	2,56
FW	-70,806	22,090	0,001	-3,205	1,48
R^2 = 0,8789, R^2_{adj} = 0,8774, AIC= 6930,9					
Model 3 Zmienna zależna: procentowe pokrycie danymi poligonowymi OSM					
Stała	1851,733	315,322	0,000	5,872	–
GZ	0,103	0,010	0,000	9,862	2,52
W2	-26,126	6,546	0,000	-3,991	2,62
PS	-0,859	0,205	0,000	-4,195	3,08
SB	3,888	1,019	0,000	3,815	2,21
WF	-12,914	2,472	0,000	-5,222	4,05
SW	0,036	0,009	0,000	4,116	1,36
FS	-2,073	0,793	0,009	-2,615	1,24
R^2 = 0,3773, R^2_{adj} = 0,3657, AIC= 4464,97					

Istotność statystyczną poszczególnych zmiennych objaśnianych w modelach o największym dopasowaniu określa wartość *p* (*p-value*) – im jest ona niższa tym istotniejsza jest dana zmienna w modelu.

Największy stopień objaśniania wyrażony za pomocą współczynnika determinacji (R^2) uzyskano dla zmiennej długość linii OSM/1 km^2, natomiast najmniejszy dla zmiennej procentowe pokrycie danymi poligonowymi OSM, czego należało się spodziewać po uprzednio przeprowadzonej analizie korelacji. W przypadku gęstości danych liniowych OSM interpretacja otrzymanego wyniku jest następująca – 88% zmienności gęstości danych liniowych OSM jest objaśniane przez zmienność: gęstości zaludnienia, liczby osób w gospodarstwach domowych otrzymujących pomoc społeczną, procentu osób z wykształceniem wyższym, liczby fundacji, stowarzyszeń i organizacji społecznych oraz frekwencji wyborczej. Należy zauważyć, że z tych zmiennych tylko wzrost gęstości zaludnienia i osób z wykształceniem wyższym wpłynie na wzrost pozyskiwanych danych liniowych VGI. Dla pozostałych zmiennych znak oszacowanego parametru jest ujemny, zatem należy się spodziewać, że wzrost zmiennej zależnej spowoduje spadek wartości zmiennej objaśnianej.

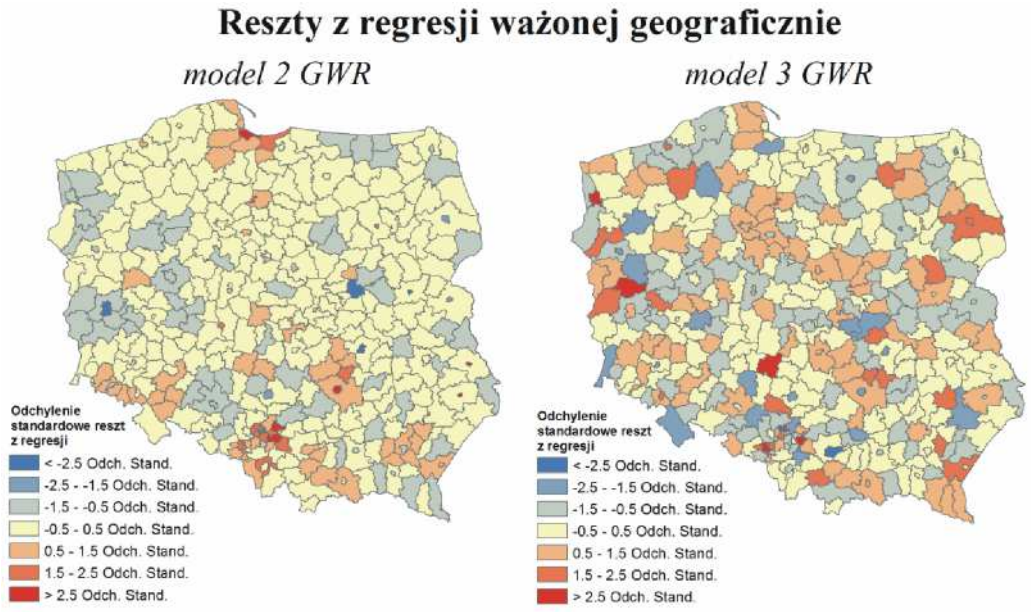
Po przeprowadzeniu modelowania za pomocą regresji liniowej obliczono wartości reszt dla każdego z powiatów, a następnie zbadano czy zachodzi autokorelacja przestrzenna, która wskazuje na tworzenie klastrów przestrzennych przez powiaty o niskim bądź wysokim dopasowaniu modelu. W przypadku jej stwierdzenia istnieją przesłanki do zastosowania modelu regresji ważonej geograficznie. Dla modelu 1 wartość globalnej statystyki Morana I wynosiła 0,02 (na poziomie istotności 0,001), co świadczy o braku autokorelacji przestrzennej. Dla modeli 2 i 3 było to odpowiednio 0,21 i 0,42 (na poziomie istotności 0,001), co z kolei wskazuje na istnienie dodatniej autokorelacji i potrzebę zastosowania modelu regresji ważonej geograficznie GWR (tab. 4).

Tabela 4. Wyniki modelowania regresją ważoną geograficznie

Model 2 GWR Zmienna zależna: długość linii OSM/1 km^2				Model 3 GWR Zmienna zależna: procentowe pokrycie danymi poligonowymi OSM		
zmienna objaśniająca	średnia wartość parametru	średni błąd standardowy		zmienna objaśniająca	średnia wartość parametru	średni błąd standardowy
Stała	5779,760	1587,911		Stała	231,887	84,324
GZ	6,690	0,328		GZ	0,039	0,038
WW	231,978	38,416		PS	-0,124	1,089
FW	-103,952	29,129		SB	0,180	5,343
R^2 = 0,8902, R^2_{adj} = 0,8868, AIC= 6902,67, średni lokalny R^2=0,8837				R^2 = 0,8900, R^2_{adj} = 0,7653, AIC= 4346,86, średni lokalny R^2=0,4091		

GWR zastosowano wykorzystując zestaw narzędzi Spatial Statistics Tools z oprogramowania ArcGIS. W narzędziu tym budowane są lokalne modele regresji oparte na macierzy sąsiedztwa z wykorzystaniem funkcji wagowej. Wagi zmieniają się wraz z oddalaniem się od punktu, w którym obliczany jest model lokalny. W budowanych modelach zastosowano zmienną macierz sąsiedztwa, w której liczba sąsiadów do budowy modelu lokalnego była dobierana w oparciu o maksymalizację kryterium informacyjnego Akaikego obliczanego dla danego modelu lokalnego.

Analiza reszt z regresji dla tych dwóch modeli wskazuje, iż są one zdecydowanie większe dla modelu objaśniającego pozyskiwanie danych poligonowych OSM (model 3 GWR). W przypadku modelu 2 GWR można zauważyć, że skrajnie wysokie bezwzględne wartości reszt występują dla powiatów grodzkich, co prawdopodobnie związane jest z wpływem zmiennej gęstość zaludnienia na wyniki modelowania (rys. 2). Dla powiatów ziemskich reszty przyjmują niskie bądź umiarkowane wartości. W przypadku modelu 3 GWR nie można stwierdzić występowania podobnej zależności – wartości reszt i ich rozłożenie przestrzenne wydaje się być losowe.



Rysunek 2. Odchylenia standardowe reszty z regresji dla modelu 2 GWR i modelu 3 GWR

Podsumowanie i wnioski

W artykule podjęto próbę wyznaczenia cech społeczeństwa mających szczególny wpływ na liczbę pozyskiwanych społecznościowych danych przestrzennych w powiatach Polski, w zależności od ich typu geometrycznego. Dokonano tego analizując wartości współczynnika korelacji Pearsona między liczbą punktów OSM na 1 km^2, długością linii OSM na 1 km^2 i procentowym pokryciem danymi poligonowymi OSM powiatów a cechami demograficznymi społeczeństwa. Zastosowano również modele regresji liniowej i regresji ważonej geo-

graficznie w celu wyznaczenia stopnia objaśniania zmiennych zależnych (liczba pozyskiwanych danych OSM) przez zmienne niezależne (cechy społeczeństwa).

Analiza korelacji wykazała, że istnieją znaczne różnice w wartościach współczynników korelacji Pearsona, Spearmana i Kendalla w zależności od typu geometrycznego danych, co potwierdza słuszność przyjętego założenia dotyczącego tego zjawiska. Dla większości zmiennych objaśnianych otrzymano wyniki zgodne z przyjętymi założeniami. Najwyższe dodatnie wartości współczynników otrzymano dla zmiennej gęstość zaludnienia, czego należało się spodziewać biorąc pod uwagę, że to właśnie ludzie tworzą dane OSM. Zmienne dotyczące podziału ludności na pięć grup wiekowych wykazały, iż istnieje korelacja między każdą z nich a ilością pozyskiwanych danych OSM z tym, że jest ona ujemna dla grup do 20 lat, 20-30 lat i 40-50 lat. O ile pierwsza grupa wiekowa zawierająca dzieci i młodzież nie powinna dziwić, o tyle druga jest zaskakująca gdyż badania literaturowe wskazują, że to właśnie z tej grupy wiekowej pochodzi najwięcej osób tworzących VGI. Ponadto pewnym zaskoczeniem są również dodatnie współczynniki korelacji dla grupy wiekowej powyżej 50 lat. Dlatego też optymalnym rozwiązaniem byłoby zastąpienie tych pięciu zmiennych jedną, która byłaby współczynnikiem w sposób kompleksowy odnoszącym się do wieku ludności wyznaczonym w oparciu o grupy wiekowe, na przykład z zastosowaniem wyższych wag dla tych, dla których według literatury tworzonych jest więcej danych OSM. W przypadku zmiennych małżeństwa zawarte na 1000 ludności, liczba osób otrzymujących pomoc społeczną i stopy bezrobocia otrzymano ujemne wartości współczynników korelacji, co było zgodne z przyjętymi założeniami. Dosyć zaskakujący wynik otrzymano dla zmiennej współczynnik feminizacji, dla którego uzyskano silne dodatnie korelacje, podczas gdy w literaturze jako głównych twórców VGI wskazuje się mężczyzn. Należałoby zatem przypuszczać, że ich większa liczba, związana jest z większą ilością pozyskiwanych danych, w takim wypadku znak współczynnika korelacji dla zmiennej współczynnik feminizacji powinien być ujemny. Być może również w tym przypadku należałoby zastosować inną zmienną, na przykład procent liczby mężczyzn w liczbie ludności. Równie wysokie wartości korelacji uzyskano dla zmiennej wykształcenie wyższe, co pokrywa się z przyjętymi założeniami na podstawie przeglądu literatury. Korelacje dodatnie, jednak na średnim poziomie, uzyskano dla zmiennej średnie wynagrodzenie brutto, co wskazuje iż zmienna ta nie jest silnie związana z ilością pozyskiwanych danych OSM. Dla zmiennej frekwencja wyborcza uzyskano ujemne wartości korelacji – przeciwnie do przyjętych założeń, co wskazuje iż udziału w wyborach samorządowych i tworzenia VGI nie można uznać za dwa przejawy pewnego rodzaju lokalnego patriotyzmu. Niskie wartości korelacji dodatniej uzyskano dla zmiennej liczba fundacji, stowarzyszeń i organizacji społecznych, co świadczy o mniejszej zależności między działaniem na rzecz społeczeństwa w ramach organizacji *non-profit* i tworzeniem danych społecznościowych. Dla zmiennych związanych z turystyką uzyskane wartości korelacji są bardzo niskie, co świadczy o jej braku i nieistotności statystycznej.

Oprócz globalnej analizy korelacji wykonano również analizy regionalne, poprzez obliczenie współczynników *r* dla każdego z województw oddzielnie. Pozwoliło to na dostrzeżenie pewnych regionalnych zależności, co nie było możliwe w modelu globalnym. Dla zmiennej liczba ludności z wykształceniem wyższym w zależności od ilości pozyskiwanych danych punktowych OSM otrzymano współczynniki *r* w przedziale (0,31-0,94), w odniesieniu do danych liniowych był to przedział (0,39-0,96), a poligonowych (-0,23-0,87). Wartości te wskazują na istnienie dużych różnic regionalnych, co skłania autorkę do wniosku, że ciekawych wyników należałoby się spodziewać przeprowadzając lokalną analizę korelacji, polega-

jącą na próbkowaniu pełnego zbioru danych i wyznaczaniu lokalnych współczynników *r,* co może stanowić propozycję przyszłych badań.

Na podstawie analizy regresji można stwierdzić, że występują znaczne różnice w stopniu objaśniania zmiennych zależnych przez zmienne niezależne. W przypadku regresji liniowej było to: 60% w przypadku danych punktowych, 88% w przypadku danych liniowych oraz 37% w przypadku danych poligonowych. Po zastosowaniu regresji ważonej geograficznie wyniki te udało się poprawić uzyskując współczynnik $R^2_{adj} = 0{,}89$ dla danych liniowych i $R^2_{adj} = 0{,}76$ dla danych poligonowych. Wskazuje to, iż w szczególności w odniesieniu do danych powierzchniowych należy wykorzystywać model GWR, przy czym wskazanie zmiennych objaśniających powinno być poprzedzone analizą korelacji lokalnej. Analiza za pomocą modeli regresji pozwoliła na wyznaczenie ostatecznej listy cech społeczeństwa, mających wpływ na pozyskiwanie danych OSM w zależności od ich typu geometrycznego (tab. 5).

Tabela 5. Ostatecznie wyznaczone cechy społeczeństwa wpływające na ilość pozyskiwanych danych OSM

Cechy społeczeństwa wpływające na ilość pozyskiwanych danych punktowych OSM	– Gęstość zaludnienia – Procent ludności z wykształceniem wyższym – Średnie wynagrodzenie brutto – Małżeństwa zawarte na 1000 ludności
Cechy społeczeństwa wpływające na ilość pozyskiwanych danych liniowych OSM	– Gęstość zaludnienia – Procent ludności z wykształceniem wyższym – Frekwencja wyborcza w wyborach samorządowych
Cechy społeczeństwa wpływające na ilość pozyskiwanych danych poligonowych OSM	– Gęstość zaludnienia – Liczba osób w gospodarstwach domowych otrzymująca pomoc społeczną – Stopa bezrobocia rejestrowanego

Należy przy tym zauważyć, że dobór zmiennych był co prawda poprzedzony badaniami literaturowymi, jednak charakteryzuje się pewną subiektywnością. Dlatego przedstawione badania należy rozszerzyć o inne zmienne dotyczące cech społeczeństwa, a także przeprowadzenie analiz na najniższym poziomie podziału administracyjnego kraju. Ostatecznie uzyskane wyniki, wskazujące na wysoki stopień objaśniania pozyskiwania VGI w zależności od cech społeczeństwa, pozwalają na prognozowanie rozwoju projektu OSM w Polsce. Należy przy tym pamiętać, że mimo tego, iż wpływ zmiennych zależnych na zmienną niezależną jest istotny statystycznie nie przesądza o istnieniu przyczynowości, co wynika z faktu, że na każde społeczeństwo składają się jednostki, którymi zawsze kierują indywidualne wybory, których nie można modelować statystycznie, a co za tym idzie wykazana zależność nie musi istnieć w rzeczywistości.

Literatura

Arsanjani J.J., Bakillah M., 2015: Understanding the potential relationship between the socio-economic variables and contributions to OpenStreetMap. *International Journal of Digital Earth* 8(11): 861-876.

Arsanjani J.J., Vaz E., 2015: An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. *International Journal of Applied Earth Observation and Geoinformation* 35: 329-337.

Budhathoki N., 2010: Participants' Motivations to Contribute to Geographic Information in an Online Community. University of Illinois, USA.

Cellmer R., 2010: Analiza przestrzenna dynamiki zmian cen nieruchomości lokalowych z wykorzystaniem regresji ważonej geograficznie. *Acta Scientiarum Polonorum. Administratio Locorum* t. 9, nr 3: 5-14.

Cichociński P., 2012: Ocena przydatności OpenStreetMap jako źródła danych dla analiz sieciowych. *Roczniki Geomatyki* t. 10, z. 7(57)*:* 15-24, PTIP, Warszawa.

Cichociński P., Dębińska E., 2012: Badanie dostępności komunikacyjnej wybranej lokalizacji z wykorzystaniem funkcji analiz sieciowych. *Roczniki Geomatyki* t. 10, z. 4(54): 41-48, PTIP, Warszawa.

Girres J.F., Touya G., 2010: Quality assessment of the French OpenStreetMap dataset. *Transaction in GIS* vol. 14 iss. 4: 435-459.

Goodchild M.F., 2007: Citizens as sensors: the Word of volunteered geography. *GeoJournal* vol. 69.

Gruszczyński M., Kluza S., Winek D., 2003: *Ekonometria*. Wyższa Szkoła Handlu i Finansów Międzynarodowych: Elipsa, Warszawa.

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* vol. 37: 682-703.

Haklay M., Budhathoki N., 2010: OpenStreetMap: Overview and Motivational Factors. Referat prezentowany na Horizon Infrastructure Challenge Theme Day, University of Nottingham, marzec 2010.

Koop G., 2011: Wprowadzenie do ekonometrii. Wolters Kluwer Polska Sp. z o.o.: 30-31.

Kulczycki M., Ligas M., 2007: Regresja ważona geograficznie jako narzędzie analizy rynku nieruchomości. *Geomatics and Environmental Engineering* 1: 59-68.

Marczak S., 2015: Ocena zaangażowania społeczeństwa w tworzenie danych przestrzennych w Polsce na przykładzie projektu OpenStreetMap. *Roczniki Geomatyki* t. 13, z. 3(69): 239-253, PTIP, Warszawa.

Marczak S., Pluto-Kossakowska J., 2015: Zastosowanie statystyki przestrzennej do analizy wykorzystywania funduszy europejskich w Polsce. *Roczniki Geomatyki* t. 13, z. 1(67): 105-116, PTIP, Warszawa.

Mashhadi A., Quattrone G., Capra L., 2013: Putting ubiquitous crowd-sourcing into context. Proceedings of the 2013 conference on Computer supported cooperative work: 611-622.

Neis P., Zielstra D., 2014: Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet* 6: 76-106.

Neis P., Zielstra D., Zipf A., 2013: Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet* 5(2): 282-300.

Neis P., Zipf A., 2012: Analyzing the contributor activity of a Volunteered Geographic Information project – The case of OpenStreetMap. *ISPRS International Journal of Geo-Information* 1(2): 146-165.

Nowak Da Costa J., Bielecka E., Całka B., 2016: Jakość danych OpenStreetMap – analiza informacji o budynkach na terenie Siedlecczyzny. *Roczniki Geomatyki* t. 14, z. 2 (72): 201-211, PTIP, Warszawa.

OpenStreetMap Polska. Dostęp 26.10.2016 r. https://www.facebook.com/osmpolska/about/

Schmidt M., Klettner S., 2013: Gender and experience-related motivators for contributing to openstreetmap. International workshop on action and interaction in volunteered geographic information (ACTIVITY), Leuven: 13-18.

Steinmann R., Häusler E., Klettner S., Schmidt M., Lin Y., 2013: Gender Dimensions in UGC and VGI: A Desk-Based Study. Jekel/Car/Griesebner (Eds.): GI_Forum 2013 Creating the GISociety, Niemcy.

Stephens M., Rondinone A., 2012: Gendering the GeoWeb. Prezentacja na Annual Meeting. Nowy Jork. http://www.scoop.it/t/opensource-geo/p/1452578643/gendering-the-geoweb-analysingdemographic-difference-in-usvgi

Zielstra D., Zipf A., 2010: A comparative study of proprietary geodata and volunteered geographic information for Germany. 13th AGILE International Conference on Geographic Information Science 2010, Portugal.

Streszczenie

W ostatnich latach tworzenie obywatelskich (społecznościowych) danych przestrzennych przez użytkowników Internetu, niebędących profesjonalistami w tym zakresie, jest coraz bardziej popularne. Świadczy o tym również rosnąca liczba inicjatyw opartych o dane zbierane na zasadzie crowdsourcingu (ang. crowd – tłum, ang. sourcing – czerpanie). Przyczynia się to do wzrostu świadomości społecznej dotyczącej danych geoprzestrzennych. Celem artykułu było zbadanie jakie cechy społeczeństwa wpływają na zaangażowanie obywateli w tworzenie VGI (ang. volunteered geographic information) w Polsce. Do jego realizacji wykorzystano dane z projektu OpenStreetMap oraz dane charakteryzujące społeczeństwo pozyskane z Głównego Urzędu Statystycznego. Były to między innymi: poziom wykształcenia, miesięczne wynagrodzenie, współczynnik feminizacji. Pierwszym etapem było określenie stopnia korelacji między danymi opisującymi społeczeństwo a danymi pozyskanymi w projekcie OpenStreetMap w podziale na powiaty. Następnie dla najbardziej skorelowanych zmiennych ułożono modele regresji wielorakiej i regresji ważonej geograficznie (GWR), co pozwoliło na wyznaczenie tych cech społeczeństwa, które miały istotny wpływ na pozyskiwanie VGI w Polsce.

Abstract

In recent years, the creation of volunteered geographic information (VGI) by Internet users, who are not professionals in this area is becoming increasingly popular. There is also a growing number of initiatives based on the data collected on the basis of crowdsourcing. This contributes to increase of the public awareness of geospatial data. The aim of the paper was to examine what features of society affect the involvement of citizens in creating VGI in Poland. To achieve this objective, data from the OpenStreetMap project and society data obtained from the Central Statistical Office (this included level of education, monthly salary, the feminisation rate) were used. The first stage was to determine the degree of correlation between the data describing the society, and the OpenStreetMap data divided into districts. Then, for the most correlated variables multiple regression and geographically weighted regression (GWR) models were arranged, which allowed the determination of the characteristics of a society that had a significant effect on the acquisition of VGI in Poland.

mgr inż. Sylwia Marczak
sylwia.marczak1@gmail.com