Metodyka pozyskiwania danych big data z telefonii komórkowej i możliwości ich wykorzystania w modelowaniu podróży[1]

KRYSTIAN BIRR
dr inż., Politechnika Gdańska, Wydział Inżynierii Lądowej i Środowiska, Katedra Inżynierii Drogowej i Transportowej, ul. Narutowicza 11/12, 80-233 Gdańsk, tel. 58 3472731, e-mail: krystian.birr@pg.edu.pl

Streszczenie: Celem niniejszego artykułu jest przedstawienie przykładowych możliwych źródeł danych o wielkim wolumenie (big data) ze szczególnym uwzględnieniem metody pozyskania danych z telefonii komórkowej – kart SIM oraz potencjału ich wykorzystania w modelowaniu podróży na poziomie makroskopowym. Na podstawie doświadczeń zdobytych podczas zakupu danych typu big data od kilku dostawców dla jednostek samorządowych w województwie pomorskim opisano najważniejsze zagadnienia metodyczne związane problematyką pozyskania i weryfikacji danych big data o rozmieszczeniu i przemieszczeniach ludności.
Słowa kluczowe: ruch drogowy, big data, modelowanie podróży.

Wprowadzenie

Rozwój technologii w ostatnich dziesięcioleciach oraz konieczność cyfryzacji i przetwarzania ogromnej ilości danych stworzyły nowe możliwości pozyskiwania informacji o funkcjonowaniu systemów wykorzystywanych w codziennym życiu. Większość organizacji, bez względu na wielkość, profil działalności i zasoby finansowe, podejmuje działania w zakresie pozyskiwania informacji, ich gromadzenia i przetwarzania. W zakresie planistycznym związanym z rozwojem przestrzeni oraz systemu transportowego, nowe technologie umożliwiają gromadzenie użytecznych informacji, które mogą stać się fundamentem kreowania polityki przestrzennej i szerzej – polityki rozwoju. Współcześnie podejmowanie właściwych decyzji planistycznych w projektowaniu zagospodarowania przestrzennego czy też szerzej podejmowanych działań w sferze publicznej wymaga stosowania właściwych narzędzi rozpoznania i diagnozy stanu przestrzeni, a także przy ich użyciu dokonywania prognoz zmian. Jednym z takich narzędzi są wielopoziomowe modele transportowe przetwarzające i dostarczające dane wspomagające decyzje dotyczące rozwoju transportowego analizowanego obszaru. Do właściwego działania tych narzędzi niezbędne jest jednak pozyskanie wiarygodnego i reprezentatywnego zbioru danych. Obecnie podstawą ich pozyskiwania są kompleksowe badania ruchu, zlecane regularnie w największych miastach, a ostatnimi laty także na poziomie metropolii, regionu, częściowo także kraju.

Postępujący rozwój technologiczny stworzył nowe możliwości zdobywania danych, które w zasadzie mogą być pozyskiwane bez jakiegokolwiek zaangażowania człowieka. Dotyczy to przykładowo danych z systemów kart płatniczych, sieci telefonii komórkowej czy też samych pojazdów. Źródła te, wykorzystywane przez administrację publiczną, mogą być podstawą lub elementem szeregu wniosków istotnych dla bieżącego i strategicznego zarządzania procesami rozwojowymi w dynamicznie zmieniającej się rzeczywistości.

Celem artykułu jest przedstawienie przykładowych możliwych źródeł danych o wielkim wolumenie (big data), ze szczególnym uwzględnieniem metody pozyskania danych z telefonii komórkowej – kart SIM oraz potencjału ich wykorzystania w modelowaniu podróży na poziomie makroskopowym.

Stan badań w wykorzystywania kart SIM w modelowaniu podróży

W literaturze światowej zagadnienie praktycznego wykorzystania kart SIM w modelowaniu podróży nie jest szeroko rozwijane, choć oczywiście również prowadzone są badania w tym zakresie. W jednej z pierwszych publikacji [1] związanych z tym zagadnieniem dokonano opisu metody generowania macierzy źródło-miejsce docelowe (O-D) z wykorzystaniem „ruchomych" danych telefonicznych. Autorzy skoncentrowali się na modelowaniu rozkładu dobowego podróży oraz identyfikacji rozkładu przestrzennego. Przedstawiono także możliwość wykorzystania zbudowanego modelu podróży do prognozowania ruchu na odcinkach sieci transportowej z wykorzystaniem oprogramowania PTV VISUM. Dane z telefonii komórkowej wykorzystano zatem do kalibracji i weryfikacji modelu.

Nieco inny aspekt wykorzystania danych o wielkim wolumenie poruszono między innymi w publikacjach [2, 3] odnoszących się do danych z sondowania pojazdów (*floating car data* – FCD), czyli danych pozyskiwanych z urządzeń zainstalowanych w pojazdach, do których można zaliczyć także aplikacje nawigacyjne. Autorzy wskazują na możliwość wykorzystania tego rodzaju danych zarówno do szacowania rozkładu przestrzennego ruchu samochodowego, a także do analiz prędkości przejazdu na odcinkach sieci transportowej.

Najnowsze publikacje wskazują na możliwość wykorzystania danych z telefonii komórkowej w badaniach przepływu ludności w trakcie pandemii COVID-19 [4]. W pracy zaproponowano wykorzystanie danych roamingowych do modelowania podróży międzynarodowych. Wykorzystano modele grawitacyjne i radiacyjne do uchwycenia przepływów przed i podczas wprowadzania ograniczeń w przemieszczaniu się. Ponieważ tradycyjne modele podróży mają pewne ograniczenia w zakresie modelowania podróży, dla takich stanów w artykule autorzy przedstawili propozycję COVID Gravity Model (CGM), czyli rozszerzenie tradycyjnego modelu grawitacyjnego, który jest dostosowany do scenariusza pandemii, bazując

[1]

na danych z kart SIM z okresu trwania pandemii. Pomijając kwestię wyzwania związanego z prognozowaniem ruchu, powyższe podejście również ukazuje potencjał wykorzystania tego rodzaju danych.

Wśród publikacji polskich autorów swoje szerokie doświadczenie w tym zakresie przedstawiali A. Brzeziński, T. Dybicz oraz Ł. Szymański, którzy są pionierami w zakresie wykorzystywaniach danych z kart SIM w modelowaniu podróży. Na podstawie doświadczeń sporządzili publikacje opisujące możliwość wykorzystania danych big data w budowie modeli oraz kalibracji macierzy podróży [5, 6, 7]. W swoich publikacjach opisali analizy podróży na podstawie kart SIM w obszarze aglomeracji warszawskiej oraz wykorzystaniu tych danych na etapie korekty generacji ruchu oraz wspomagania modelowania rozkładu przestrzennego, wykazując ostatecznie wysoki stopień zgodności liczby podróży w poszczególnych relacjach względem danych z kompleksowych (warszawskich) badań ruchu. W podobnym zakresie również R. Kucharski z zespołem wykorzystali dane z kart SIM do aktualizacji modelu dla aglomeracji krakowskiej [8]. W zakresie projektu badawczego INMOP 3 Brzeziński z zespołem podjęli próbę analizy możliwości wykorzystania big data, w tym danych z kart SIM operatora telefonii komórkowej, oraz dane z sondowania pojazdów jako źródła danych do przeprowadzania analiz ruchu i modelowania podróży wszystkimi środkami transportu w Polsce z wykorzystaniem budowanego w ramach tego projektu krajowego modelu podróży, wykorzystywanego obecnie przez Generalną Dyrekcję Dróg Krajowych i Autostrad [9]. Uzyskane wyniki streszczono między innymi w publikacji [10], ukazując potencjał wykorzystania tego rodzaju danych w modelowaniu podróży.

Źródła pozyskiwania danych

Jak napisano we wstępie, rozwój technologii stwarza nowe możliwości w zakresie gromadzenia i przetwarzania danych o wielkim wolumenie, możliwych do wykorzystania w narzędziach do wspomagania decyzji w zakresie planowania rozwoju i zarządzania transportem. Na rynku polskim możliwe jest również pozyskanie tego rodzaju danych, jednak istotnym aspektem jest zdefiniowanie celu i pozyskanie oraz możliwości przetworzenia danych. Big data nie muszą być wykorzystywane do modelowania podróży. Ich potencjał jest znacznie szerszy i nawet dane historyczne mogą być wartościowym wkładem ukazującym stan funkcjonowania systemu transportowego, a także zagospodarowania przestrzennego. Dane te nie muszą również dotyczyć przemieszczeń, ale też dostarczają informację o rozmieszczeniu ludności w przestrzeni w różnych stanach i okresach.

Spośród obecnie możliwych źródeł danych o rozmieszczeniu i przemieszczeniach ludności można wyróżnić:

- **dane z kart SIM** – dane dostarczane przez operatorów telefonów komórkowych lub, w niektórych przypadkach, pośredników. Dane te bazują na punktach logowania się użytkowników sieci komórkowej, lokalizowanych za pomocą infrastruktury operatora, na którą składają się maszty telefonii komórkowej i umieszczone na nich urządzenia nadawcze i odbiorcze. Użytkownik lokalizowany jest w obrębie środka ciężkości obszarów (tzw. celki) obsługiwanych przez dany maszt, skierowanych w trzech kierunkach od masztu. Z obszaru celek użytkownicy sumowani są do zadanych obszarów, np. gmin i dzielnic. W zależności od zakresu zamówienia dane te mogą być pozyskane jedynie od jednego operatora lub kilku, co przekłada się na zwiększenie próby. Przykładowo, na rynku polskim występuje 4 głównych operatorów, których udział w rynku jest zbliżony, co przekłada się na informację o około 25% użytkownikach w przypadku pozyskania danych od jednego operatora. Uwzględniając dodatkowo konieczność anonimizacji danych oraz udziału zgód użytkowników na przetwarzanie tych danych, wielkość próby w populacji może spaść do 15–20%;
- **dane z aplikacji wykorzystujących lokalizację GPS telefonów komórkowych** – dane pozyskane w wyniku zapytań reklamowych lub od wydawców aplikacji. Dane te dostarczane są przez firmy analityczne i marketingowe specjalizujące się w reklamie mobilnej. Użytkownik lokalizowany jest z wysoką dokładnością (dokładność GPS) w przypadku wyświetlenia reklamy na urządzeniu mobilnym (np. Google AdSense) lub poprzez dane zbierane przez różne aplikacje mobilne, monitorujące położenie użytkownika (aplikacje sklepów, restauracji, sportowe, wynajmu pojazdów i inne). Doświadczenie autora wykazało teoretyczny dostęp do próby około 80% użytkowników telefonów komórkowych. To samo doświadczenie wykazało jednak niedoskonałość w przetworzeniu tych danych przez dostawców, a tym samym niską jakość otrzymywanych wyników. Źródło to charakteryzuje się wysokim potencjałem, jednak konieczne jest posiadanie narzędzi do weryfikacji jakości otrzymanych danych oraz kontrola metodyki ich pozyskiwania i przetwarzania przez dostawcę;
- **dane z GPS pojazdów oraz aplikacji nawigacyjnych** – w podstawowym zakresie zbieranych danych możliwe jest pozyskanie danych o lokalizacji i czasie pojawienia się pojazdu, jego kierunku poruszania się, prędkości chwilowej. Wraz z rozwojem technologii pojawiają się także dodatkowe dane związane z warunkami ruchu pojazdów. Możliwa jest identyfikacja rodzaju pojazdu, wyróżniając pojazdy ciężkie i lekkie. Na podstawie tych danych możliwe jest poznanie informacji o relacjach przejazdu źródło-cel pojazdów, a także o prędkościach przejazdu na poszczególnych odcinkach sieci drogowej. Umożliwia to identyfikację miejsc występowania zatorów drogowych, prowadzenie badań związanych ze zdarzeniami drogowymi. Ponadto, z punktu widzenia modelowania ruchu, w połączeniu z danymi o natężeniu ruchu możliwe jest opracowanie funkcji oporu odcinka oraz oszacowanie wartości prędkości w ruchu swobodnym na poszczególnych typach odcinków;

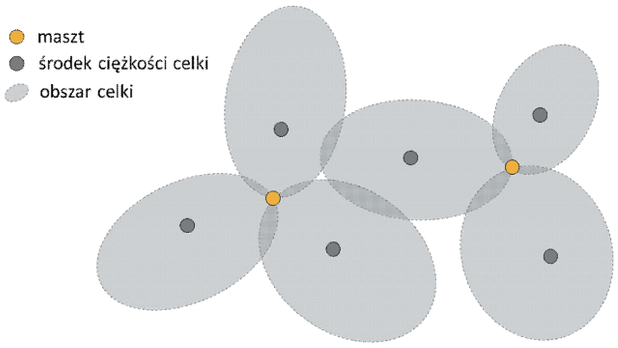
- **dane z systemów sterowania ruchem** – dane gromadzone przez zarządców i operatorów systemów sterowania ruchem. Dane te, w zależności od budowy systemu, umożliwiają pozyskanie danych o natężeniu ruchu na przekrojach i skrzyżowaniach, ale także o czasie przejazdu pojazdów pomiędzy wybranymi punktami z wykorzystaniem łączności bluetooth lub identyfikacji pojazdów po numerach rejestracyjnych. Dodatkowo, z uwzględnieniem systemu priorytetyzacji pojazdów transportu zbiorowego oraz informacji pasażerskiej, możliwe jest także pozyskanie danych o czasie przejazdu odcinków międzyprzystankowych lub między skrzyżowaniami lub poszczególnymi punktami meldunkowymi [11]. Łącząc powyższe z systemem automatycznego zliczania liczby pasażerów wsiadających i wysiadających, możliwe jest także zdobycie danych o wielkości ruchu pasażerskiego. Wszystkie z powyższych danych stanowić mogą istotny element w kalibracji modeli podróży, ale także być podstawą do wprowadzania usprawnień i rozwoju lokalnego systemu transportowego;
- **dane z kart płatniczych** – dane umożliwiające identyfikację czasu i miejsc wykonywania podróży z informacją o miejscu zamieszkania użytkownika (na podstawie regularności i rodzaju wykonywanych transakcji), celu podróży (transakcje w miejscu poza miejscem zamieszkania), w ograniczonym stopniu wykorzystanego środka transportu (transakcja na stacji paliw, w kasie biletowej lub na lotnisku) oraz majętności użytkownika (wielkość transakcji w hotelu, restauracji). Tego rodzaju dane wykazują potencjał przede wszystkim w przypadku podróży ponadlokalnych, w szczególności o charakterze turystycznym i biznesowym. Mogą być dobrym uzupełnieniem danych z telefonii komórkowej. To innowacyjne źródło danych, będzie przedmiotem kolejnych badań autora.

Metodyka pozyskiwania danych z telefonii komórkowej

Jak przedstawiono w poprzednim rozdziale, istnieje kilka źródeł danych o wielkim wolumenie możliwych do wykorzystania w planowaniu i zarządzaniu rozwojem systemu transportowego. W niniejszym rozdziale skoncentrowano się na danych pozyskiwanych z telefonii komórkowej (kart SIM lub GPS). Bazując na doświadczeniu autora zdobytym podczas zakupu tego rodzaju danych od kilku operatorów dla jednostek samorządowych w województwie pomorskim, wyróżniono kilka istotnych elementów, na które należy, zdaniem autora, zwrócić szczególną uwagę, decydując się na zakup tego rodzaju danych.

Podział obszaru na rejony transportowe

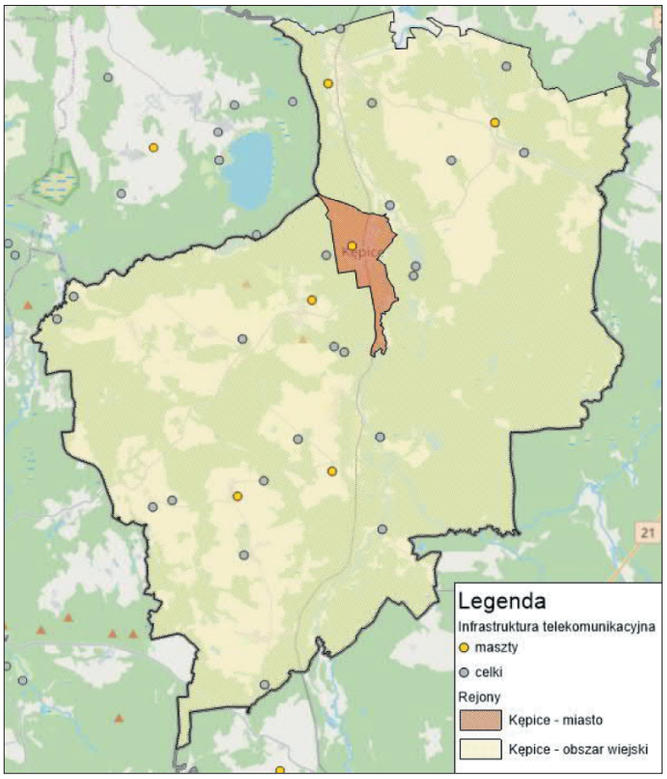
Podstawowym elementem procesu przygotowawczego do pozyskania danych big data jest zdefiniowanie obszaru analizy oraz jego delimitacja. Aspekt ten jest istotny zarówno w przypadku zakupu danych z kart SIM, jak i GPS, ponieważ nawet w przypadku danych GPS, w tym FCD, występuje ryzyko błędnej identyfikacji lokalizacji ze względu na dokładność GPS w danym obszarze, wynikające z uwarunkowań zewnętrznych, np. wysokości budynków. W przypadku kart SIM istotne jest rozpoznanie rozmieszczenia masztów stacji bazowych (BTS) w analizowanym obszarze. Lokalizacje BTS warunkują możliwą skalę delimitacji analizowanego obszaru – na dany rejon powinien przypadać co najmniej jeden BTS, zlokalizowany możliwie blisko geograficznego środka ciężkości danego rejonu. Stacje zlokalizowane przy granicy rejonu obarczone są większym ryzykiem błędu. Oczywiście im większa liczba stacji, tym bardziej ograniczony zostaje błąd lokalizacji. Ma to szczególne znaczenie ze względu na zasięg danego BTS oraz ryzyko logowania się telefonu do BTS przypisanego do sąsiedniego rejonu. W praktyce, zazwyczaj na jednym maszcie BTS operuje się tzw. celkami umieszczonymi w trzech kierunkach (rys. 1). Dla celek kalkulowane są środki ciężkości. Operator telekomunikacyjny nie jest w stanie określić, w którym miejscu wewnątrz celki znajduje się użytkownik. Dlatego jako jego położenie przyjmuje się środek ciężkości celki, z której on aktualnie korzysta.



Rys. 1. Schemat pokrycia obszaru celkami
Źródło: opracowanie własne

Badania i analizy autora dla województwa pomorskiego wykazały, że w przypadku dużych miast optymalnym podziałem jest podział na dzielnice, a w obszarze pozamiejskim na gminy. W przypadku większej gęstości stacji bazowych, szczególnie w obszarach podmiejskich, zasadne może być dokonanie podziału gminy na rejony. Należy przy tym pamiętać, że z uwagi na ograniczenia techniczne, polegające na ograniczonej liczbie celek przypadających na dany rejon, a także z uwagi na automatyczne przełączanie się komórek między celkami (tzw. szum, który może być częściowo filtrowany), zastosowana metoda ma niską dokładność w zakresie opracowania danych dla przemieszczeń wewnątrzrejonowych.

Dokonując podziału obszaru na rejony transportowe, należy także uwzględnić ich kształt. Utrudnione może być pozyskanie danych dla rejonów podłużnych lub w kształcie litery L, U, O. Rejony podłużne mogą wynikać na przykład z podziału administracyjnego. W takim przypadku, jeśli rejony są małe, zalecane jest ich zagregowanie z którymś z sąsiadujących rejonów. Przypadki z kształtami L i U występują, kiedy punktem wyjścia jest podział administracyjny,



Rys. 2. Przykład problematycznego podziału na rejony ze względu na kształt oraz lokalizację celek
Źródło: opracowanie własne

na przykład gminny. Często mniejsze gminy miejskie są otoczone przez gminę wiejską. W takim przypadku również należy albo zagregować rejony, albo dokonać dodatkowego podziału rejonu o tego rodzaju kształcie na kilka części (zazwyczaj dwie), jeśli liczba BTS na to pozwala. Na rysunku powyżej przedstawiono przykładową problematyczną lokalizację, dla której uzyskano zaburzone wyniki bez agregacji rejonów (rys. 2) z uwagi na geometrię rejonu miasta Kępice oraz rozmieszczenie masztów i przypisanych do nich celek.

Definicje

Kolejnym etapem procesu przygotowawczego do zakupu danych jest ustalenie zakresu danych wynikowych. W zależności od zakresu możliwe jest ustalenie tzw. definicji, czyli kryteriów określających przyjęte w selekcji dane oraz interpretację otrzymanych wyników. Dane z telefonii komórkowej pozwalają na pozyskanie informacji zarówno o przemieszczeniach ludności (tzw. przepływy), jak i ich rozmieszczeniu w przestrzeni (tzw. migawki). W zakresie migawek można podjąć próbę identyfikacji liczby mieszkańców, codziennych użytkowników, turystów, odwiedzających lub inne. Przypisanie cech np. mieszkańca do danego użytkownika umożliwia w kolejnym etapie identyfikację jego przemieszczeń, oczywiście w przypadku spełnienia wymogów anonimizacji danych.

Powyższe podejście zastosowano w województwie pomorskim w zamówieniu realizowanym na zlecenie Pomorskiego Biura Planowania Regionalnego, Obszaru Metropolitalnego Gdańsk-Gdynia-Sopot, Urzędu Miasta Gdańska oraz Urzędu Miasta Gdyni [12]. Przystępując do zadania, zdefiniowano kilka wariantów definicji mieszkańców, użytkowników itd., a następnie, korzystając z uprzejmości Wykonawcy dostarczającego dane, firmy T-Mobile, przeprowadzono iteracje testujące poszczególne definicje. Z uwagi na obszerność zagadnienia w niniejszym artykule posłużono się przykładem definicji określających liczbę mieszkańców z pominięciem pozostałych grup. Poniżej przedstawiono przykładowe definicje liczby mieszkańców obszaru:

- definicja nr 1: liczba mieszkańców – liczba osób przebywających w jednym miejscu tzw. nieruchome karty SIM w godzinach nocnych tj. 3.00–4.00, przez 15 lub więcej dni w miesiącu;
- definicja nr 2: liczba mieszkańców – liczba osób przebywających najdłużej w danym rejonie względem pozostałych rejonów w godzinach 19.00–7.00 przez 15 lub więcej dni w miesiącu;
- definicja nr 3: liczba mieszkańców – liczba osób przebywających najdłużej w danym rejonie względem pozostałych rejonów w godzinach 19.00–7.00. Warunek: osoba spędziła min. 15 nocy w województwie pomorskim.

Dla każdej z powyższych definicji otrzymano różne wartości wynikowe, co świadczy o istocie zastosowanego podejścia i wyboru właściwej do danego zadania definicji (tab. 1).

Tabela 1

Zestawienie przykładowych danych porównawczych dla różnych definicji liczby mieszkańców

Liczba mieszkańców	Miesiąc: październik 2019			Miesiąc: sierpień 2019		
	Definicja 1	Definicja 2	Definicja 3	Definicja 1	Definicja 2	Definicja 3
Gdańsk	478 671	521 055	582 734	400 329	434 431	516 695
Gdynia	231 266	249 419	274 466	199 303	214 625	249 346
woj. pomorskie	2 087 421	2 237 091	2 421 767	1 970 426	2 103 115	2 373 854

Źródło: opracowanie własne na podstawie [12]

Ostatecznie na podstawie iteracyjnych analiz uzyskanych wyników, ich logiki oraz korelacji z innymi źródłami danych będących w dyspozycji jednostek samorządowych oraz danych Głównego Urzędu Statystycznego, do dalszych analiz wybrano jedną definicję. W przypadku liczby mieszkańców była to powyższa definicja nr 2. W zakresie pozostałych definicji przyjęto zgodnie z tabelą 2. Uzyskane dane zobrazowano w postaci graficznej oraz porównano do bazy GUS, wykazując między innymi przybliżoną zgodność w zakresie ogólnej liczby mieszkańców województwa pomorskiego, większą o około 10% liczbę osób mieszkających w Gdańsku oraz mniejszą liczbę mieszkańców większości gmin pozamiejskich.

Spośród przedstawionych definicji dodatkowego komentarza wymaga definicja liczby turystów, którą należy interpretować dosłownie, jak opisano to w polu interpretacji. Nie są to zatem tylko „typowi" turyści, którzy spędzają urlop w danym miejscu, ale także osoby spędzające noc poza miejscem zamieszkania i regularnego użytkowania (podróże biznesowe, spotkania towarzyskie). W celu pozy-

Tabela 2

Zestawienie definicji wykorzystanych w analizach dla województwa pomorskiego

Grupa	Interpretacja	Definicja
liczba mieszkańców	liczba stałych mieszkańców w danym rejonie – średnia w danym miesiącu	liczba osób przebywających najdłużej w danym rejonie względem pozostałych rejonów w godzinach 19.00–7.00 przez 15 lub więcej dni w miesiącu
liczba użytkowników	liczba osób przebywających regularnie w danym obszarze w ciągu dnia (np. pracujących, uczących się), niebędącym miejscem zamieszkania	liczba użytkowników – na podstawie miejsca najdłuższego przebywania w godzinach 7.00-19.00 przez 11 lub więcej dni roboczych w miesiącu
liczba odwiedzających	liczba osób odwiedzających dany rejon w ciągu dnia w celu zaspokojenia potrzeb (np. biznesowych, osobistych), który nie jest ani miejscem ich miejscem zamieszkania, ani regularnego użytkowania	liczba osób, która spędziła w danym rejonie co najmniej 4 godziny w godz. 23.00-7.00 przez nie więcej niż 14 dni w miesiącu. Osoby nieuznane za mieszkańców ani użytkowników danej gminy
liczba turystów	liczba osób, która spędziła od 1 do 14 nocy w danym rejonie transportowym, niebędącym miejscem zamieszkania ani regularnego użytkowania	liczba osób przebywających w danym rejonie transportowym w godz. 7.00-23.00 przez minimum 2 godziny. Osoby, których miejscem zamieszkania lub użytkowania nie jest dana gmina

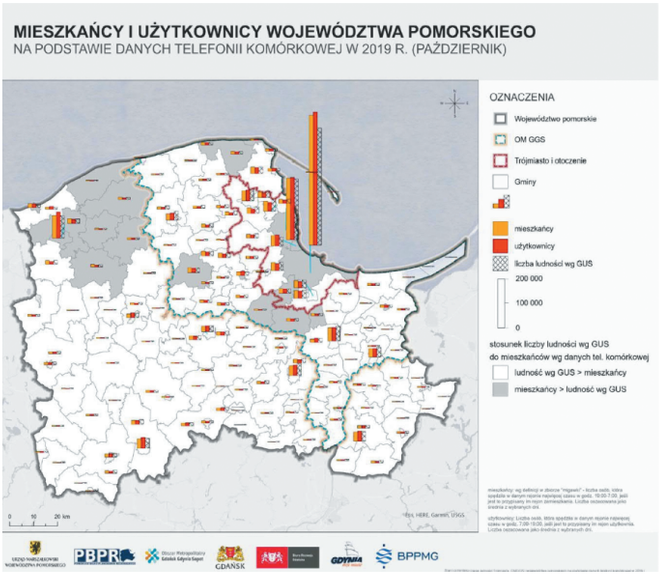
Źródło: opracowanie własne na podstawie [12]

skania informacji o „typowych" turystach powyższą definicję można rozszerzyć poprzez doprecyzowanie na przykład zakresu liczby dni spędzonych w danym miejscu. Kwestię tę pominięto w analizach dla województwa pomorskiego z przyczyn organizacyjnych.

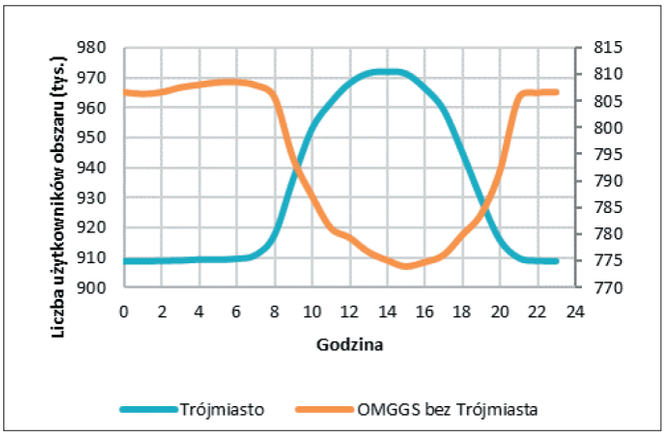
Dodatkowo, w przypadku odwiedzających i turystów, dokonano podziału ze względu na miejsce zamieszkania, uzyskując w ten sposób informację o liczbie odwiedzających i turystów mieszkających w OMGGS, województwie pomorskim, pozostałej część Polski, za granicą.

Przedstawione powyżej definicje nie należy traktować jako jedynych właściwych, jednak wykonane analizy wskazują na istotę przypisania uwagi do przyjętej definicji w celu właściwej interpretacji uzyskanych wyników i celu ich dalszego wykorzystywania. Spośród alternatywnych definicji równie zasadne są definicje odnoszące się do czasu przebywania w danym rejonie w ciągu roku – na przykład w okresie nocnym, co umożliwia uzyskanie danych o liczbie mieszkańców w odniesieniu do roku, a nie miesiąca. Powyższe wymaga jednak od dostawcy pracy na danych rocznych. Z wykorzystaniem powyższych cech przypisanych do użytkowników możliwe było opracowanie ich więźb przemieszczeń. Zidentyfikowano tym samym na przykład relacje podróży turystycznych, a także podróże pomiędzy rejonem zamieszkania a rejonem regularnego użytkowania, odwzorowujących potencjalne podróże obligatoryjne. Poszerzone wyniki analiz dla województwa pomorskiego przedstawiono w raporcie [12].

Oprócz danych opracowanych na podstawie definicji możliwe jest także pozyskanie rozmieszczenia kart SIM w danej chwili na podstawie ich ostatniego miejsca logowania. Zestawiając ze sobą takie dane w przedziale godzinnym, możliwe jest zidentyfikowanie migracji ludności w ciągu doby. Na rysunku 4 przedstawiono przykładową zmianę liczby osób przebywających w Trójmieście i pozostałym obszarze OMGGS.



Rys. 3. Liczba mieszkańców i użytkowników w rejonach – porównanie do GUS
Źródło: raport [12]



Rys. 4. Godzinowe rozmieszczenie ludności w Trójmieście i pozostałej części OMGGS
Źródło: raport [12]

Problematyka rozszerzenia próby

Wykonane badania i analizy wykazały, że kluczowym aspektem w przetwarzaniu danych wynikowych z telefonii komórkowych jest problematyka rozszerzenia próby na populację. Najprostszym i najczęstszym dotychczasowym podejściem jest stosowanie przez dostawców współczynnika rozszerzającego próbę na populację, obliczonego na podstawie liczby kart SIM, dla których użytkownicy wyrazili zgodę na przetwarzanie danych osobowych oraz liczby mieszkańców Polski na podstawie danych GUS. Podejście to, jakkolwiek logiczne, w skali globalnej doprowadza do rozbieżności w przypadku występowania różnych udziałów operatorów w danym obszarze.

Zjawisko to występuje szczególnie w obszarach pozamiejskich, gdzie z uwagi na zasięg sieci komórkowej mieszkańcy preferują wybór jednego lub dwóch operatorów. Sytuacja ta doprowadza do powstawania błędów w skali szacowanej liczby mieszkańców oraz podróży związanych z danym rejonem. Problem ten dotyczy przede wszystkim przypadku podziału obszaru na gminy.

Przeprowadzone badania dla obszaru Trójmiasta nie wykazały istotności tego problemu. Podobnie w przypadku skali powiatowej, gdzie dane dla podróży międzypowiatowych oraz liczby mieszkańców były skorelowane z innymi bazami danych. W przypadku gmin rozwiązaniem minimalizującym powyższy problem jest zastosowanie wskaźnika korygującego, obliczonego na podstawie wyników badań udziału danego operatora w obsłudze mieszkańców. Dane takie wymagają jednak przeprowadzenia badania ankietowego wśród reprezentatywnej próby mieszkańców, co wiąże się z dodatkowymi kosztami, lecz jest zalecane do uzyskania wiarygodnych wyników.

Kordony

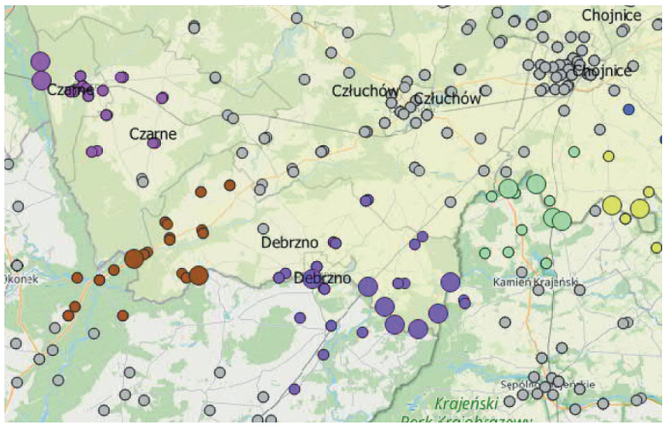
W przypadku ograniczonych możliwości pozyskiwania danych o ruchu zewnętrznym i tranzytowym z innych źródeł, w tym danych z FCD wykorzystujących GPS, możliwe jest skorzystanie z danych z kart SIM. Podobnie jak w przypadku poprzednich zagadnień, tutaj również wysoce zalecana jest weryfikacja zastosowanej metody oraz uzyskanych wyników. W przypadku analiz dla województwa pomorskiego, dla określenia ruchu na kordonach opracowano dwie reguły:

1. **Reguła zewnętrznej granicy**. Podczas przekraczania granic województwa oraz OMGGS podróżujący mogą w niektórych sytuacjach pojawić się w celkach przypisanych więcej niż jednemu kordonowi. Wtedy przypisanie do kordonu jest uzależnione od tego, która celka jest bardziej na zewnątrz. Dlatego przypisujemy kordony wg poniższej logiki:
   a) przemieszczenia źródłowe: czyli sytuacje, gdy osoba wyjeżdża z obszaru analizy – o przypisaniu do kordonu decyduje ostatnia celka sparowana z kordonem, w której pojawił się podróżujący;
   b) przemieszczenia docelowe: o przypisaniu do kordonu decyduje pierwsza celka sparowana z kordonem, w której pojawił się podróżujący;
   c) przemieszczenia tranzytowe: obydwie powyższe reguły są stosowane, zależnie czy jest to wlot, czy wylot z obszaru objętego badaniem.

2. **Reguła pojawienia się**. W tej regule chodzi o uchwycenie osób przyjeżdżających z zagranicy, dla których nie jest możliwe ustalenie początku lub końca podróży (miały on miejsce poza siecią operatora). Dotyczy ona portów oraz lotniska. Dlatego:
   a) jeśli osoba nie była widoczna przez minimum 6 godzin w sieci, wtedy sprawdzamy, czy osoba ta najpierw pojawiła się w którymś z portów lub na lotnisku;
   b) jeśli osoba przestaje być widoczna przez minimum 6 godzin w sieci, wtedy sprawdzamy, czy osoba ta pojawiła się w ostatnim miejscu w którymś z portów lub na lotnisku.

Zastosowanie powyższych reguł wymusza wykonanie działania poprzedzającego polegającego na identyfikacji celek przypisanych do danego kordonu, uwzględniających przemieszczających się użytkowników, którzy niekoniecznie muszą zalogować się do danej celki. Dlatego też grupa tych celek musi obejmować odpowiedni obszar wzdłuż danej drogi. Problem pojawia się jednak w przypadku występowania kilku kordonów obok siebie lub przebiegania drogi w niedalekiej odległości od granicy analizowanego obszaru. W takich przypadkach konieczne jest zastosowanie dodatkowych reguł, weryfikacja grup celek lub agregacja kordonów. W badaniach dla województwa pomorskiego zidentyfikowano dwie problematyczne lokalizacje na kordonach:

a) kordon trasy S7, przebiegającej wzdłuż wschodniej granicy województwa pomorskiego – z uwagi na bliskość trasy, część podróży pojawiała się w obszarze analizy podwójnie: w okolicy Elbląga oraz w okolicy Żuławki Sztumskiej;
b) kordony na południowo-zachodniej granicy województwa – z uwagi na dużą liczbę punktów kordonowych (rys. 5).



Rys. 5. Przykład przypisania grup celek w problematycznych obszarach kordonowych
Źródło: opracowanie własne

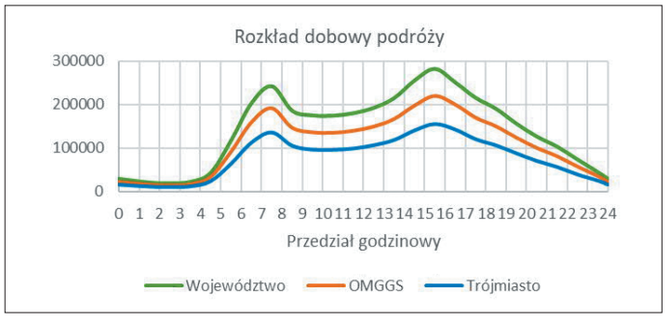
Dla każdego z powyższych przypadków zastosowano pogłębioną analizę doboru celek do kordonów oraz identyfikację logowania się do celek poza obszarem analizy.

Zastosowanie pozyskanych danych w modelowaniu

Zgodnie z doświadczeniami wskazanymi w literaturze, przywołanej w rozdziale 2, pozyskane dane z telefonii komórkowej można wykorzystać do budowy i kalibracji makroskopowych modeli podróży. W kolejnych podrozdziałach wyszczególniono oraz pokrótce opisano możliwości wykorzystania tych danych. Szersze informacje dotyczące tych zagadnień można odnaleźć we wskazanej literaturze.

Rozkład dobowy ruchu

Podstawową informacją możliwą do opracowania na podstawie danych z telefonii komórkowej jest charakterystyka popytu w zakresie rozkładu dobowego, określającego liczbę podróży rozpoczynanych w danym obszarze w poszczególnych przedziałach godzinowych. Na rysunku 6 zaprezentowano wyniki rozkładu dobowego podróży z podziałem na obszar Trójmiasta, OMGGS oraz województwa pomorskiego.



Rys. 6. Rozkład dobowy podróży na podstawie danych z telefonii komórkowej
Źródło: raport [12]

Dane o liczbie podróży

Liczba zidentyfikowanych podróży w poszczególnych relacjach może zostać odniesiona do grup podróży międzygminnych, międzypowiatowych, wewnętrznych (wewnątrz analizowanego obszaru) i zewnętrznych (w tym tranzytowych) (tab. 3). Na tej podstawie możliwe jest oszacowanie ruchliwości mieszkańców w zakresie rodzaju podróży z podziałem dla każdego rejonu (np. gminy, dzielnicy) z osobna. Podobnie jak w przypadku badań zespołu A. Brzezińskiego [5] możliwe jest także określenie udziału podróży z danej gminy, na przykład do stolicy województwa, miasta powiatowego itd. Rozwiązanie to z powodzeniem zastosowano również w autorskim modelu dla województwa pomorskiego zwiększając jego stopień dokładności.

Tabela 3

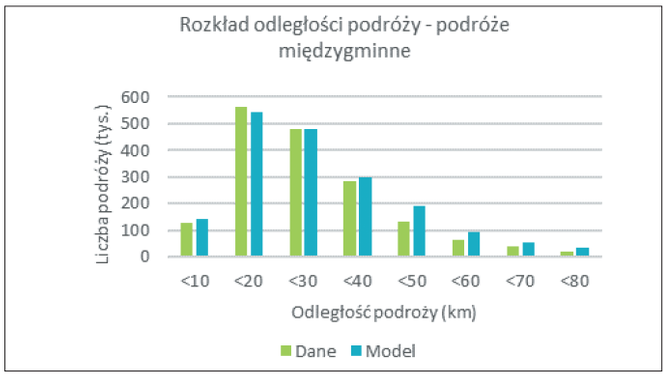
Przykładowe zestawienie danych o liczbie podróży z telefonii komórkowej

Dane z poziomu województwa	Październik		Sierpień	
	Dzień powszedni	Dzień weekendu	Dzień powszedni	Dzień 18.08.2019
Liczba wszystkich podróży	7 288 522	6 127 660	7 554 377	6 413 500
Liczba podróży wewnętrznych	6 967 861	5 807 283	7 134 645	5 799 795
Liczba podróży wew. międzygminnych	2 143 773	1 812 583	2 299 682	1 860 076
Liczba podróży wew. międzypowiatowych	1 211 902	995 497	1 242 680	937 308

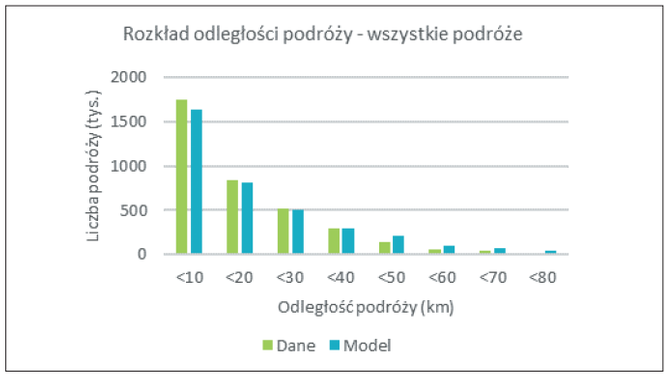
Źródło: raport [12]

Modelowanie rozkładu przestrzennego

Dokonując zestawienia liczby podróży w poszczególnych relacjach z odległościami podróży obliczonymi na przykład w modelach makroskopowych, możliwe jest opracowanie funkcji oporu przestrzeni. Możliwy zakres analiz w tym zakresie jest ograniczony ze względu na niewielką możliwość w identyfikacji motywacji podróży. Dane te jednak wydają się mieć potencjał do wykorzystania, na przykład do szacowania funkcji oporu dla podróży turystycznych, co będzie przedmiotem kolejnych badań autora. Z uwagi na powyższe w standardowych modelach pozyskane w ten sposób dane mogą zostać wykorzystane do kalibracji i walidacji rozkładu przestrzennego dla wszystkich podróży, co również dokonano dla modelu województwa pomorskiego (rys. 7, 8).



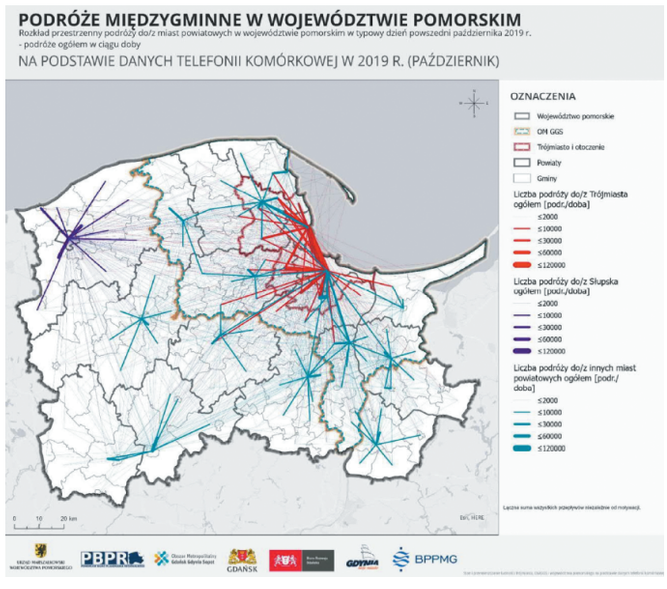
Rys. 7. Rozkład odległości podróży dla podróży międzygminnych – porównanie danych z wartościami z modelu podróży dla województwa pomorskiego
Źródło: raport [12]



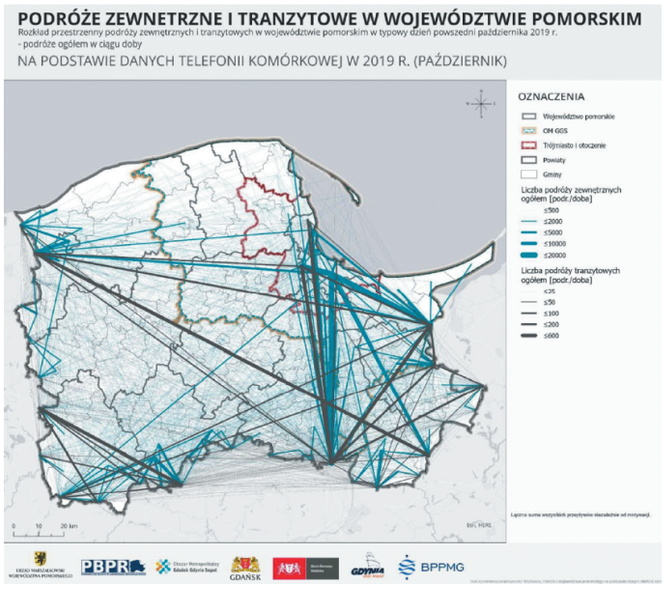
Rys. 8. Rozkład odległości podróży dla wszystkich podróży – porównanie danych z wartościami z modelu podróży dla województwa pomorskiego
Źródło: raport [12]

Ruch zewnętrzny i tranzytowy

Z uwagi na trudność techniczno-prawną w wykonywaniu badań ankietowych na kordonach, dane big data z telefonii komórkowej są obecnie jedną z ciekawszych alternatyw w pozyskiwaniu danych o ruchu zewnętrznym i tranzytowym. Dane te, zweryfikowane przez pomiary natężenia ruchu i potoków pasażerskich w transporcie zbiorowym, można wykorzystać do kalibracji rozkładu przestrzennego (rys. 9, 10).



Rys. 9. Rozkład podróży wewnętrznych na podstawie danych z telefonii komórkowej
Źródło: raport [12]



Rys. 10. Rozkład podróży zewnętrznych i tranzytowych na podstawie danych z telefonii komórkowej
Źródło: raport [12]

Wnioski i rekomendacje

Wykonane badania i analizy wykazują wysoki potencjał danych o wielkim wolumenie w zakresie ich wykorzystania do identyfikacji rozmieszczenia, migracji, przemieszczeń osób w celu analizy funkcjonowania badanego obszaru. W zakresie analiz transportowych dostępne w Polsce źródła danych nie ograniczają się jedynie do danych z telefonii komórkowych. Możliwe jest także pozyskiwanie danych z innych źródeł big data na podstawie danych z sondowania pojazdów oraz z systemów sterowania ruchem i zarządzania transportem zbiorowym.

Z uwagi na kilka możliwości pozyskania tego rodzaju danych oraz ze względu na różnych dostawców, promujących zróżnicowane podejścia metodyczne o różnej jakości, istotne jest odpowiednie rozpoznanie zagadnienia, określenie celu zakupu danych, opracowanie zakresu potrzebnych danych, wybór właściwej metodyki przetwarzania danych, a następnie ich weryfikacja oraz właściwa interpretacja. W niniejszym artykule przedstawiono podstawowe zagadnienia dotyczące ważnych aspektów metodycznych w zakresie możliwości i ograniczeń technicznych w pozyskiwaniu tego rodzaju danych, a także przykładowych obszarów weryfikacji danych.

Wskazane w artykule niektóre z ograniczeń dotyczących dokładności lokalizacji kart SIM, z uwagi na rozmieszczenie BTS i celek, będą zmniejszane wraz z rozwojem sieci 5G i budową nowych stacji bazowych. Ponadto operatorzy sieci komórkowej mają możliwość pozyskiwania podstawowych danych na podstawie transferowanych danych, wskazujących czas korzystania z danego rodzaju aplikacji mobilnych, czy też ogólny sposób korzystania z telefonu. Odpowiednia analiza tego rodzaju danych powinna umożliwić identyfikację przedziału wiekowego użytkownika, jego płeć i inne dane. Potencjał tego źródła danych jest znacznie szerszy i wymagać będzie dalszych badań i analiz.

Dane z sieci komórkowych oraz innych wskazanych w niniejszym artykule źródeł big data, mogą być z powodzeniem wykorzystywane do budowy i kalibracji makroskopowych modeli podróży. Dane o przemieszczeniach ludności na tym poziomie dokładności nie można jednak traktować jako danych wystarczających. Z uwagi na brak informacji o charakterystyce podróży, w szczególności motywacji oraz wybranego środka transportu, niezbędne jest bazowanie na tradycyjnych źródłach danych, w szczególności na wywiadach w gospodarstwach domowych z wykorzystaniem dzienniczków podróży. Big data mają jednak potencjał do wykorzystania w zakresie modelowania podróży o charakterze turystycznym oraz w rozbudowie modeli na inne okresy prognostyczne niż typowy dzień powszedni.

Pełny raport z danych zakupionych dla województwa pomorskiego, z którego wykorzystano dane w tym artykule dostępny jest na stronach wskazanych jednostek samorządowych oraz w postaci streszczenia na stronie bigdata.birr.pl (alias przekierowujący).

Literatura

1. Friedrich M., Immisch K., Jehlicka P., Otterstätter T., Schlaich J., *Generating origin-destination matrices from mobile phone trajectories*, "Transportation Research Record: Journal of the Transportation Research Board", 2010, no 2196.
2. Bowman C.N., Miller J.A., *Modeling traffic flow using simulation and Big Data analytics*, 2016 Winter Simulation Conference (WSC), 2016.
3. Dabbas H., Fourati W., Friedrich B., *Floating Car Data for Traffic Demand Estimation – Field and Simulation Studies*, 2020 IEEE 23rd International Conference on Intelligent Transportation Systems,
4. Luca M., Lepri B., Frias-Martinez E., Lutu A. *Modeling international mobility using roaming cell phone traces during COVID-19 pandemic*. EPJ Data Science, 4/2022.
5. Brzeziński A., Dybicz T., Szymański Ł., *Demand model in the agglomeration using SIM cards*, "Archives of Civil Engineering", 1/2019.
6. Brzeziński A., Dybicz T., *Possibility of Big Data application for OD-matrix calibration in transport demand models*,"Archives of Civil Engineering", 1/2021.
7. Brzeziński A., Dybicz T., Szymański Ł., *Doświadczenia z budowy modelu ruchu dla obszaru metropolitarnego Warszawy z wykorzystaniem innowacyjnych źródeł danych*, „Annały inżynierii ruchu i planowania transportu", t. III, Planowanie ruchu a wyzwania globalne, SITK, 2009.
8. Kucharski R., Mielczarek J., Drabicki A., Szarata A., *Metoda aktualizacji modelu podróży z wykorzystaniem macierzy przemieszczeń telefonów komórkowych*, „Transport Miejski i Regionalny", 2018, nr 5.
9. Brzeziński A., Projekt INMOP3, *Intermodalny Krajowy Model Ruchu*, Konferencja Naukowo-Techniczna, Innowacyjne Metody Prognozowania Ruchu Krajowego – Regionalnego – Lokalnego, Warszawa, 28 maja 2019.
10. Suchorzewski W., Brzeziński A., Waltz A., *Modelowanie i prognozowanie ruchu – od liczydła do Big Data*, „Transport Miejski i Regionalny", 2020, nr 12.
11. Helbin M., Wyszomirski O., *Możliwości wykorzystania Big Data w badaniach popytu i podaży w transporcie miejskim*, „Transport Miejski i Regionalny", 2019, nr 2.
12. Raport: Analiza aktywności i potencjału ludnościowego województwa pomorskiego, obszaru metropolitalnego i Trójmiasta w oparciu o zachowania użytkowników sieci telefonii komórkowych w 2019 r., Praca zbiorowa (Birr K. i inni), Gdańsk-Gdynia 2021.