Techniki analizy i modelowanie więźby ruchu miejskiego[1]

MARIUSZ KAMOLA
dr inż., Politechnika Warszawska, Instytut Automatyki i Informatyki Stosowanej, e-mail: M.Kamola@ia.pw.edu.pl
JAKUB WESOŁOWSKI
mgr, Neptis SA – operator i właściciel systemów Yanosik i Flotis, e-mail: jakub.wesolowski@neptis.pl

Streszczenie: Artykuł prezentuje metody matematyczne zmierzające do uchwycenia zależności więźby ruchu od prędkości średniej w obszarach aglomeracji oraz warunków pogodowych. Uwzględniono również zależności autoregresyjne. Uzyskane wyniki wskazują na dominację zjawisk cyklicznych i sezonowości nad wymienionymi czynnikami. Niemniej jednak zastosowane metody analityczne odsłaniają inne, nieprzewidziane zależności i mogą stanowić materiał do bezpośredniej interpretacji przez specjalistów z dziedziny, jak również punkt wyjścia do budowy bardziej złożonych i adekwatnych modeli predykcyjnych. Artykuł prezentuje również analizę i modelowanie zachowań kierowców na autostradzie w obliczu utrudnień wynikających z planowych prac remontowych.
Słowa kluczowe: więźba ruchu, modelowanie ruchu, modelowanie podróży.

Wprowadzenie

Więźba ruchu, czyli liczba podróży pomiędzy rejonami transportowymi, jest obecnie pojęciem kluczowym dla różnorodnych zadań analitycznych i inżynierskich w transporcie. Precyzyjne informacje o podróżach pozwalają na efektywne bieżące zarządzanie ruchem oraz warunkują racjonalne prowadzenie inwestycji infrastrukturalnych. Te same dane umożliwiają tworzenie poprawnych modeli ruchu: makroekonomicznych, dekomponujących ruch według przyjętych założeń oraz mikroekonomicznych, agregujących decyzje transportowe gospodarstw domowych i przedsiębiorstw [1].

Celem artykułu jest zaprezentowanie technik analizy i obrazowania danych o ruchu, umożliwiających wykrycie nieznanych dotąd zależności i ostatecznie prowadzących do konstrukcji statystycznych modeli ruchu.

Punktem wyjścia do rozważań jest więźba ruchu (*origin-destination matrix*). Pojęcie to w piśmiennictwie naukowym równie często odnosi się do ruchu drogowego, jak do ruchu w sieci komputerowej. Do niedawna w obu dziedzinach stosowano również podobne instrumentarium w celu wyznaczenia więźby ruchu, wykorzystując model grawitacyjny. W modelu tym przyjmuje się, że całkowity ruch wychodzący z obszaru rozkłada się na obszary docelowe proporcjonalnie do udziału ruchu przychodzącego w tych obszarach do ogółu ruchu przychodzącego [2,3]. Model był adekwatny do dostępnych metod pomiarowych, z reguły ograniczających się do obserwacji łącznego ruchu na granicach obszarów sieci transportowej lub komputerowej.

Z czasem środki techniczne dostępne w obu dziedzinach uległy zróżnicowaniu, adekwatnie do postępu technologii. Estymacja macierzy ruchu komputerowego wciąż opiera się na zbiorczych statystykach wolumenu danych obsłużonych przez routery brzegowe (odpowiedniki ulic wylotowych z rejonów transportowych). Z uwagi na wykładniczy trend wzrostowy ruchu, niemożliwa jest rejestracja źródła i przeznaczenia wszystkich pakietów; niekiedy zbiera się niewielką ich próbkę (poniżej 1%). Dodatkowo, o ile pewnym ułatwieniem w dalszej analizie są znane, deterministyczne trasy ruchu, o tyle istotnym utrudnieniem pozostaje natura przepływów traktowanych jako szeregi czasowe, wykazujące silne samopodobieństwo i opisywanych tzw. ciężkoogonowymi rozkładami prawdopodobieństwa [4,5].

W przypadku ruchu drogowego wydaje się, że dostępność nowych technologii zbierania danych rosła szybciej niż samo natężenie ruchu, skutkując powiększaniem się możliwości stosowania nowych technik akwizycji i analizy danych. Wiodącą rolę odegrał rozwój sieci komórkowych, umożliwiających najpierw zgrubną lokalizację pojazdów z wykorzystaniem techniki trilateracji, a następnie, w połączeniu z GPS i transmisją pakietową, pełny i dokładny monitoring przejazdów drogowych środków lokomocji. I choć, podobnie jak dla sieci komputerowych, zbierane dane dotyczą tylko pewnego podzbioru ogólnej liczby przejazdów, zostały ogólnie uznane za wystarczająco reprezentatywne [6].

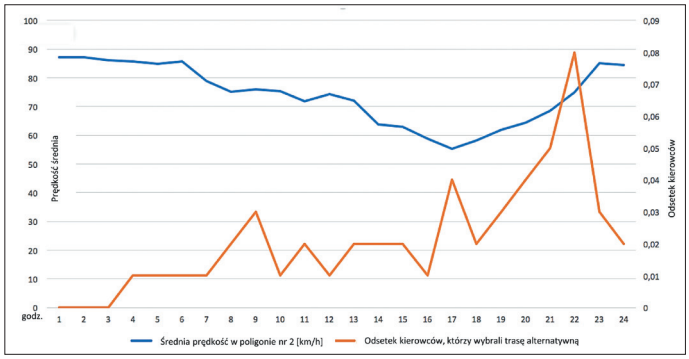
Model wyboru trasy alternatywnej w ruchu dalekobieżnym

Dzięki powszechnemu wykorzystywaniu technologii mobilnych oraz bardzo dużej liczbie użytkowników systemu Yanosik w Polsce, narzędzia, których przeznaczeniem z natury jest wymiana informacji drogowej pomiędzy kierowcami, stają się również bogatym źródłem danych przydatnych do modelowania i zarządzania ruchem, np. w takich procesach jak planowania remontów dróg i prognozowanie rozkładu ruchu na drogach alternatywnych. Przykładem takiego wykorzystania danych pochodzących z urządzeń mobilnych jest analiza ruchu na autostradzie A4 w czasie przeprowadzanego na niej remontu pomiędzy węzłami Opole Południe i Krapkowice. Analiza miała na celu znalezienie odpowiedzi na pytanie: „Jakie warunki ruchu na remontowanym odcinku skłaniają kierowców do przejazdu trasą alternatywną?”. Analizie poddane zostały próbki GPS generowane przez użytkowników systemu Yanosik od poniedziałku do piątku w dniach 14–18 maja 2018 roku. Do zbadania prędkości wyznaczone zostały trzy poligony pomiarowe: pierwszy przed węzłem Opole Południe, drugi to odcinek prowadzonych prac w okolicach węzła Krapkowice, a trzeci to odcinek przed Gliwicami, por. rysunek 1.

[1] ©Transport Miejski i Regionalny, 2019. Wkład autorów w publikację: M. Kamola 50%, J. Wesołowski 50%.

Rys. 1. Poligony pomiarowe na autostradzie A4

Na rysunku 2 przedstawiono wykres średnich prędkości godzinowych oraz procentowy udział kierowców, którzy zdecydowali się opuścić autostradę A4 i przejechali do poligonu trzeciego drogą alternatywną. Innymi słowy jest to odsetek wszystkich kierowców korzystających z systemu Yanosik, którzy w badanych dniach znaleźli się w poligonie numer jeden i trzy, a nie odnotowano ich obecności w poligonie numer dwa.

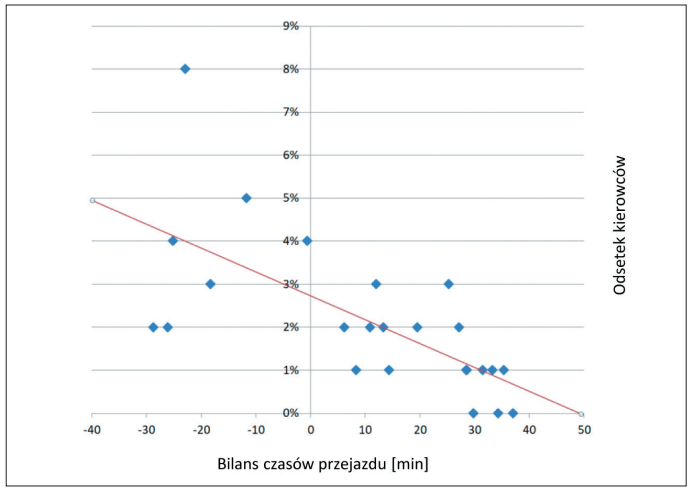


Rys. 2. Warunki drogowe i decyzje kierowców w ujęciu godzinowym

W związku z powyższym wnioski z przeprowadzonej analizy próbek GPS są następujące:

a. pierwszy zauważalny spadek średniej prędkości przejazdowej (o 10 km/h) przez remontowany odcinek A4 pojawia się wraz z początkiem porannych szczytów komunikacyjnych w godzinach od 7:00 do 9:00;
b. kolejny wzrost zainteresowania trasami alternatywnymi pojawia się o godz. 17:00, kiedy średnia prędkość spada do najniższego poziomu, czyli 55 km/h;
c. od godziny 18:00 prędkość zaczyna wzrastać do prawie 90 km/h, jednak prawdopodobna obawa kierowców przed poruszaniem się z prędkością 50 km/h, jak to miało miejsce po godz. 15:00, powoduje, że zainteresowanie trasami alternatywnymi trwa aż do godz. 22:00.

Można modelować decyzje kierowców w tej sytuacji za pomocą prostego modelu regresyjnego. Na rysunku 3 przedstawiono te same odsetki kierowców jadących drogą alternatywną, ale w funkcji bieżącego bilansu czasu przejazdu autostradą i drogami alternatywnymi. Rysuje się oczywista zależność, przybliżona na rysunku prostą regresji.



Rys. 3. Model regresyjny decyzji o ominięciu korka na autostradzie

Analiza więźby ruchu w aglomeracji

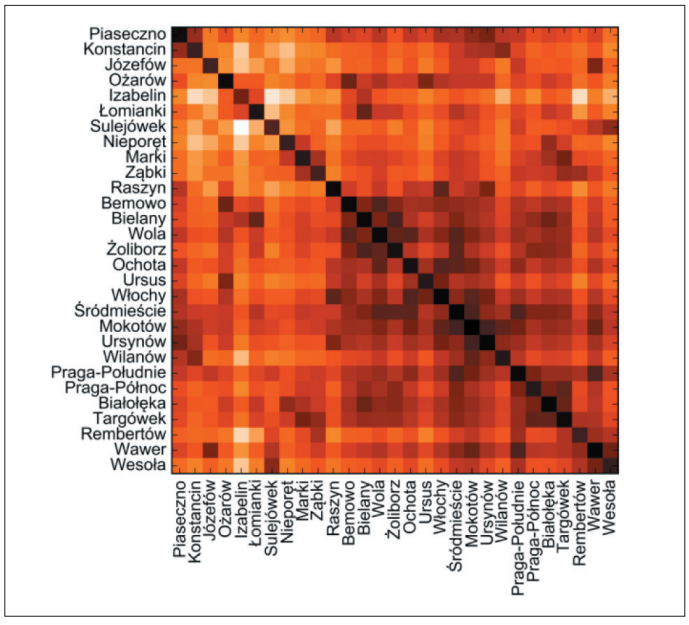
W analizie więźby ruchu miejskiego posługujemy się zagregowanymi danymi o przejazdach, pochodzącymi z systemu Yanosik oraz systemu monitoringu floty dostarczanego przez Neptis SA. Dane obejmują przejazdy pomiędzy wybranymi arbitralnie 29 rejonami transportowymi, tj. 18 dzielnicami Warszawy oraz 11 okolicznymi gminami, por. rysunek 4. Uwzględnione gminy podwarszawskie wykazują silne zróżnicowanie pod względem gęstości zaludnienia, odległości od centrum, cen nieruchomości, dominującej funkcji użytkowej (mieszkalna, przemysłowa, handlowa, rekreacyjna). Aby zapewnić odpowiednią liczność danych w każdej godzinie, ograniczono się do przejazdów rozpoczętych w godzinach od 6:00 do północy. 0dtąd, używając terminu „godzina przejazdu", będziemy odnosili się do najpóźniejszej pełnej godziny, po której rozpoczął się prze-

Rys. 4. Rejony transportowe uwzględnione w analizie więźby ruchu

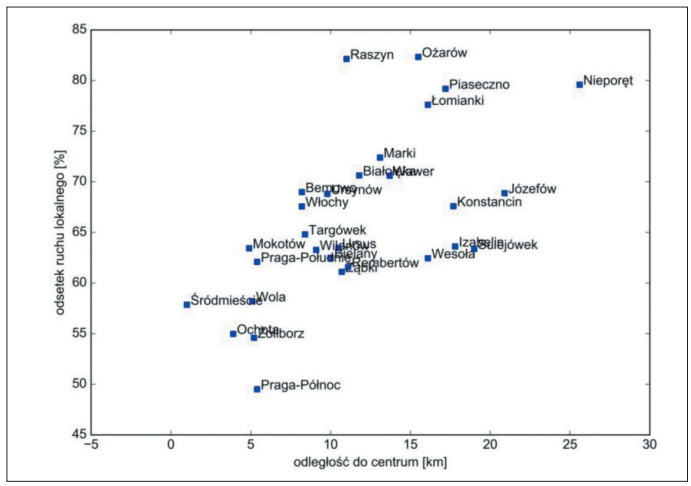
jazd. Rozpatrywane dane obejmują okres od 1 lutego do 30 kwietnia 2018 roku. W tym okresie nie wystąpiły poważne zmiany w stołecznej sieci transportowej ani inne szczególne wydarzenia, które mogłyby generować anomalie ruchu. Rozpatrywane dane obejmują wszystkie środki lokomocji, których kierowcy lub pasażerowie korzystali z aplikacji Yanosik. W naszej analizie odstępujemy więc nieco od klasycznego, czterostadiowego modelowania [3], wymagającego podziału podróży na kategorie środków transportu. Podział taki jest możliwy za pomocą algorytmów firmy Neptis.

Na rysunku 5 przedstawiono poglądowo więźbę ruchu dla całości danych o podróżach w postaci kwadratowej mozaiki. Kolor elementu w kolumnie i oraz wierszu j odpowiada logarytmowi z ogólnej liczby odnotowanych przejazdów z rejonu transportowego i do j. Bardziej nasycone kolory odpowiadają większemu natężeniu ruchu. Nawet pobieżna ocena wykresu pozwala stwierdzić co najmniej, że: 1) natężenie ruchu jest bardzo zróżnicowane, 2) więźba ruchu ma praktycznie postać macierzy symetrycznej, 3) elementy na przekątnej (tj. ruch lokalny) dominują nad ruchem wychodzącym w różnym stopniu. Pomijając dwa pierwsze spostrzeżenia, dość oczywiste dla ruchu drogowego, możemy sprawdzić, jak ma się udział ruchu lokalnego do wybranego innego czynnika, np. szacunkowej odległości rejonu od centrum aglomeracji. Wyniki przedstawiono w postaci wykresu punktowego na rysunku 6.

Również tutaj ocena jakościowa wykresu niesie kolejne istotne spostrzeżenia. Zasadniczo udział ruchu lokalnego w rejonach rośnie wraz ze wzrostem odległości od centrum, co dotyczy nie tylko gmin leżących na peryferiach analizowanego obszaru, lecz również tych wewnątrz (np. Białołęka, Bemowo, Ursynów). Zależność ta nie wynika zatem z niekompletności modelu na brzegach, ale prawdopodobnie z malejącej motywacji do podróży. Motywacja ta jest zresztą zdeterminowana innymi, nieuwidocznionymi czynnikami, wyraziście dzielącymi rejony w odległości 15–20 km na



Rys. 5. Więźba ruchu dla całości danych



Rys. 6. Odsetek ruchu lokalnego w zależności od odległości rejonów od centrum Warszawy

mocno (Ożarów, Piaseczno, Łomianki) i słabo autonomiczne (Józefów, Konstancin, Izabelin, Wesoła, Sulejówek). Typowanie tych czynników i stawianie ogólniejszych hipotez na podstawie analizy jedynie tej aglomeracji byłoby ryzykowne; niemniej jednak przedstawione, proste metody wizualizacji uświadamiają takie zależności.

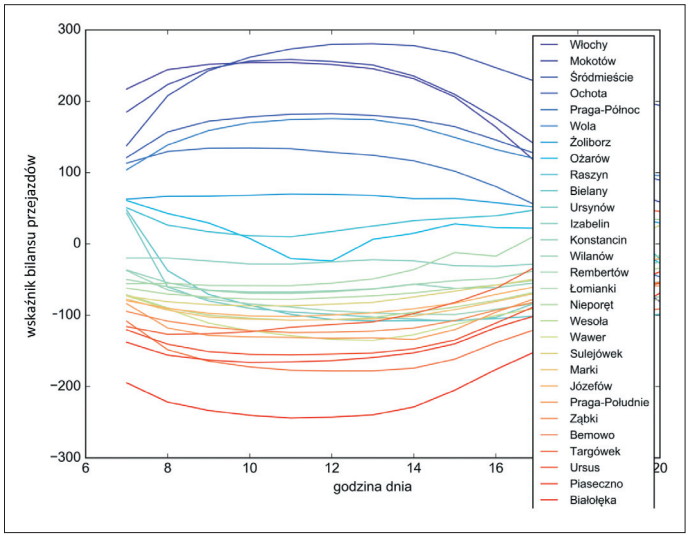
Bilans ruchu

Wróćmy teraz do spostrzeżenia o symetryczności więźby ruchu. Wynika ona w sposób naturalny z faktu, że podróżni, a więc i środki podróżowania, ostatecznie powracają do swoich pierwotnych lokalizacji. (W przypadku ruchu w sieci komputerowej taka sytuacja z reguły nie ma miejsca). Więźba ruchu z rysunku 5 przedstawia zatem ruch praktycznie zupełnie zbilansowany. Wiemy jednak skądinąd, że poszczególne rejony mają zróżnicowany charakter, który będzie manifestował się w więźbie ruchu sporządzonej dla pewnego podzbioru przejazdów. Proponujemy, aby do wyznaczenia przedziału godzin przejazdów, których użyjemy do sporządzenia specyficznej więźby ruchu, wykorzystać następujący *wskaźnik bilansu przejazdów* dla rejonu w godzinie:

$$b_i(t) = \sum_j a_{ji}^{-} - \sum_j a_{ij}^{-} - \sum_j a_{ji}^{+} + \sum_j a_{ij}^{+} \quad ,$$

gdzie: a_{ji}^{-} to liczba przejazdów z rejonu i do j wykonanych w godzinach do t włącznie, a a_{ji}^{+} to analogiczna liczba przejazdów wykonanych w godzinach następujących po t. Dwie pierwsze sumy w powyższym wzorze składają się więc na bilans ruchu do godziny t, a dwie pozostałe na bilans ruchu w pozostałym czasie. W godzinach porannych i wieczornych bilans ten jest bliski zeru, co wynika z symetrii więźby ruchu. Natomiast w ciągu dnia jest równy saldu przejazdów, tj. różnicy pomiędzy przyjazdami do a wyjazdami z rejonu i.

Na rysunku 7 zestawiono przebieg bilansów dla wszystkich rejonów i godzin przejazdów, począwszy od 7:00. (W celu lepszej wizualizacji wykreślono pierwiastki kwadratowe z bilansów, z zachowaniem pierwotnego znaku). Rejony z ujemnym dziennym bilansem szczytowym to typowe „sypialnie” (Białołęka, Piaseczno, Bemowo); rejony



Rys. 7. Wizualizacja wskaźnika bilansu przejazdów (oś rzędnych w skali nieliniowej) w kolejnych godzinach dnia. Legenda uporządkowana wg bilansów z godz. 7:00

z bilansem silnie dodatnim pełnią funkcje ośrodków pracy, kultury, wypoczynku (Śródmieście, Mokotów, Włochy). W środku zestawienia można znaleźć stosunkowo ludne dzielnice, których bilans oscyluje wokół zera (Bielany, Ursynów). Może to świadczyć o zrównoważonym występowaniu funkcji mieszkaniowych i biznesowych, co jednakże nie musi przekładać się na duży odsetek ruchu lokalnego, por. rysunek 6, gdzie jest widoczna różnica blisko 10% na korzyść Ursynowa względem Bielan.

Analiza wykresu pozwala na szacunkowe ustalenie pory dnia, w której, w większości rejonów obserwujemy maksimum wartości bilansu przejazdów. Jest to godzina 11:00 – dla tej godziny maksimum bilansu obserwujemy aż w 12 rejonach transportowych. W siedmiu rejonach godzina maksimum bilansu wypada wcześniej (Ożarów, Raszyn, Targówek, Sulejówek, Konstancin, Piaseczno, Praga-Płn.), a w dziesięciu – później (Izabelin, Ursynów, Wilanów, Praga-Płd., Śródmieście, Wawer, Bielany, Wola, Ochota, Bemowo). W dalszych rozważaniach uwzględnimy więc tylko przejazdy zaobserwowane do południa, gdyż w tym okresie dobrze artykułują się różnice w charakterze poszczególnych rejonów.

Segmentacja dat

Zasadnicze zagadnienie badawcze polega na wskazaniu przyczyn kształtujących więźbę ruchu, czyli na opracowaniu modelu matematycznego zdolnego, na podstawie tych przyczyn, przewidzieć liczbę przejazdów pomiędzy rejonami w nieodległej przyszłości. Pierwsza trudność związana z modelowaniem polega na wytypowaniu mierzalnych, hipotetycznych czynników determinujących więźbę ruchu. Kolejna to dobór klasy modeli, np. linowych, neuronowych, i ich ewentualnego połączenia. Pokusa stworzenia modelu uwzględniającego wszelkie dostępne informacje podlega naturalnym ograniczeniom wynikającym z liczby dostępnych danych ruchowych oraz z groźby stworzenia modelu *przeuczonego*, tj. dobrze dopasowanego do danych historycznych, ale pozbawionego cech generalizacji i wskutek tego niezdolnego do wykonywania sensownych predykcji. Dlatego też postanowiliśmy ograniczyć rozważania do modeli prostych, kładąc nacisk na należytą selekcję użytych danych i ich wizualizację.

Pierwszy etap selekcji polegał na ograniczeniu zbioru danych do przejazdów porannych. Następny wynika z ogólnej wiedzy o zróżnicowaniu motywacji do podróży w zależności od dnia tygodnia. W związku z tym proponujemy następujący podział dat na klasy:

1. niedziele i święta wolne od handlu – podróże o charakterze towarzyskim, edukacyjnym i rekreacyjnym;
2. niedziele handlowe oraz soboty – jw. plus domniemany udział podróży w celu dokonania zakupów lub oględzin;
3. dni bezpośrednio po świątecznych oraz poniedziałki – dni wzmożonych przyjazdów do Warszawy w celach zawodowych;
4. dni robocze przed świętem oraz piątki – dni wzmożonych wyjazdów weekendowych;
5. dni pozostałe – pospolite dni robocze.

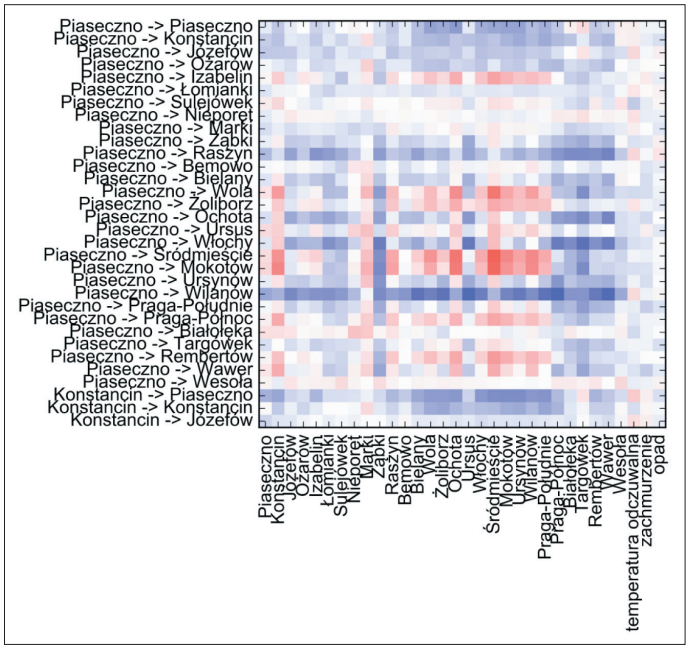
Zasadność takiego podziału można sprawdzić zgrubnie poprzez porównanie współczynników zmienności (tj. ilorazu odchylenia standardowego i średniej liczby przejazdów w godzinie) dla ruchu niepodzielonego oraz dla podzielonego na klasy. Współczynnik ten dla ruchu niepodzielonego wynosi około 0,79 i maleje do około 0,77 w klasach 3 i 4. Wynika stąd w szczególności, że przejazdy w dni robocze charakteryzują się nieco mniejszą zmiennością na tle ogółu i dlatego kolejne rozważania będziemy prowadzić dla nich.

Zmienne objaśniające

W zadaniu budowy modelu więźby ruchu będziemy mieli do dyspozycji następujące zmienne objaśniające, tj. pomiary czynników potencjalnie determinujących liczbę przejazdów w kolejnych godzinach:

- średnie prędkości pojazdów w poszczególnych rejonach transportowych i godzinach;
- historyczne dane pogodowe ze stacji meteorologicznej Okęcie: temperaturę odczuwalną, stopień zachmurzenia oraz wielkość opadu – również w poszczególnych godzinach.

W celu sprawdzenia stopnia zależności elementów więźby ruchu od ww. czynników wyznaczono macierz współczynników korelacji Pearsona, której fragment prezentuje rysunek 8. Pojedyncza próbka danych dotyczy liczby podróży lub wartości zmiennych objaśniających w konkretnej godzinie; statystyki wyznaczono dla 216 próbek. Przedstawiona macierz korelacji obejmuje głównie kierunki ruchu wychodzącego z Piaseczna, które, zachowując charakter typowego źródła ruchu porannego (por. rys. 7), wykazuje jednocześnie zaskakująco duży udział ruchu lokalnego (por. rys. 6). Rysunek 8 uwidacznia przede wszystkim słabą i niespójną zależność generowanego ruchu od czynników pogodowych. Zauważmy, że na najistotniejszych kierunkach dośrodkowych (na Mokotów, Ursynów i do Śródmieścia) obserwujemy dość silną pozytywną korelację natężenia ruchu z prędkościami średnimi w dzielnicach centralnych i pobliskich Piasecznu.



Rys. 8. Macierz korelacji Pearsona wszystkich zmiennych objaśniających z liczbą przejazdów na wybranych relacjach. Odcienie czerwieni odpowiadają dodatnim współczynnikom korelacji, odcienie granatu – ujemnym. Natężenie barwy odpowiada modułowi współczynnika korelacji

Na kierunkach obwodowych (z Piaseczna do Konstancina, Raszyna i Włoch) analogiczne korelacje mają wartości zdecydowanie ujemne lub bliskie zeru. Korelacja danych nie świadczy o przyczynowości zmiennych objaśniających. Nie odpowiada w szczególności na pytanie: czy dobre warunki ruchu w centrum okazują się dodatnio skorelowane z ruchem dośrodkowym z Piaseczna, ponieważ stymulują popyt – a zatem wpływają na *podjęcie* decyzji – czy raczej umożliwiają wjazd większej liczby pojazdów – a zatem determinują sposób *wykonania* decyzji? Odpowiedź na to pytanie wymaga głębszej analizy; ponieważ słabe skorelowanie więźby z czynnikami pogodowymi znamionuje nieelastyczność ruchu, to obserwowane korelacje z prędkościami średnimi są zapewne tylko artefaktem wprowadzonym przez układ transportowy miasta.

Składowe główne

Analiza pełnej macierzy korelacji, jak na rysunku 8, pozwala na wychwycenie istotnych powiązań elementu więźby z bardzo wieloma czynnikami zewnętrznymi i pozostaje niewątpliwie cennym narzędziem w eksperckiej ocenie sytuacji. Jednak ta sama wielość zmiennych objaśniających utrudnia budowę modeli matematycznych. W takiej sytuacji można zastosować redukcję przestrzeni czynników zewnętrznych poprzez *analizę głównych składowych* (PCA – *principal component analysis*), która polega na liniowym przekształceniu pierwotnej przestrzeni zmiennych objaśniających w nową przestrzeń. Zmienne objaśniające po transformacji do nowej przestrzeni stają się głównymi składowymi. Główne składowe są nieskorelowane – nie dublują informacji i dlatego wystarczy wykorzystać kilka najważniejszych, aby dobrze opisać oddziaływanie czynników zewnętrznych. Składowe główne powstają jako liniowe kombinacje oryginalnych czynników dobrane tak, aby maksymalnie wykorzystać informację w nich zawartą.

Tabela 1

Współczynniki przekształcenia zmiennych objaśniających w trzy najważniejsze składowe główne

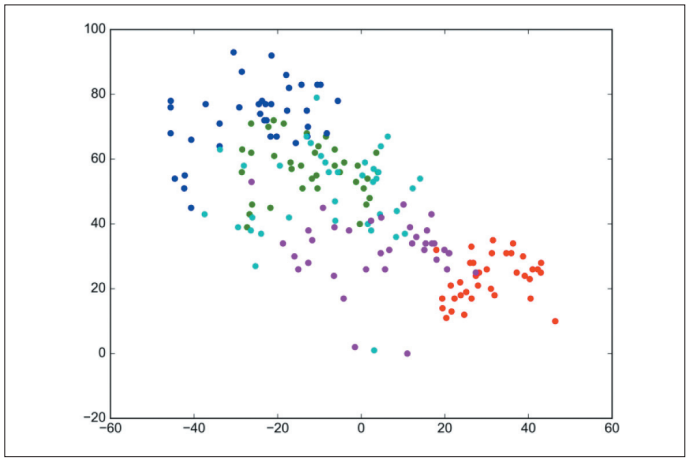
1. składowa główna	2. składowa główna	3. składowa główna
0.235 Bielany	0.326 Śródmieście	-0.40 temperatura odczuwalna
0.230 Ursynów	-0.32 Ząbki	0.330 Piaseczno
0.228 Praga-Południe	-0.30 Targówek	0.319 Konstancin
0.225 Włochy	0.293 Ochota	0.298 Nieporęt
0.221 Mokotów	-0.26 Białołęka	-0.28 Praga-Północ
0.219 Wawer	0.237 Mokotów	-0.26 Ursus
0.213 Wola	0.218 Wilanów	0.261 Marki
0.205 Żoliborz	-0.21 Sulejówek	-0.20 Ożarów
0.203 Wilanów	-0.20 Łomianki	-0.19 Żoliborz
0.203 Rembertów	0.196 Wola	-0.19 Targówek
0.201 Praga-Północ	-0.19 Wesoła	0.169 Raszyn
0.190 Ursus	0.180 Konstancin	0.167 Sulejówek
0.189 Ochota	-0.16 Rembertów	-0.15 Wola
0.188 Józefów	-0.16 Józefów	-0.14 Praga-Południe
0.186 Raszyn	0.159 Ursynów	0.121 Śródmieście
0.180 Śródmieście	0.158 Praga-Południe	0.120 Józefów
0.175 Izabelin	-0.14 Nieporęt	-0.11 zachmurzenie
0.172 Bemowo	-0.13 Wawer	-0.09 Włochy
0.167 Piaseczno	-0.12 Bemowo	-0.08 Wawer
0.166 Konstancin	0.128 Włochy	0.075 Wilanów
0.165 Łomianki	0.117 temperatura odczuwalna	0.074 Izabelin
0.158 Ożarów	0.112 Raszyn	-0.07 Bielany
0.149 Nieporęt	0.083 Żoliborz	0.070 Wesoła
0.144 Białołęka	-0.08 Piaseczno	-0.06 Ząbki
0.143 Ząbki	-0.07 Ursus	0.056 Białołęka
0.141 Wesoła	-0.07 Praga-Północ	-0.04 Mokotów
0.122 Sulejówek	-0.05 zachmurzenie	-0.04 Bemowo
0.120 Targówek	-0.05 Ożarów	-0.03 Ursynów
0.103 Marki	0.027 opad	0.034 Ochota
-0.10 temperatura odczuwalna	-0.02 Izabelin	-0.03 opad
0.026 zachmurzenie	-0.01 Bielany	-0.02 Łomianki
-0.01 opad	-0.00 Marki	0.000 Rembertów

Skuteczność redukcji liczby zmiennych przez PCA ocenia się na podstawie odsetka wariancji wszystkich wyznaczonych składowych, jaki stanowi wariancja pierwszych kilku składowych głównych. W naszym przypadku 1 składowa główna „objaśnia" 47% ogólnej wariancji – można powiedzieć, że zawiera w sobie prawie połowę użytecznej informacji w modelowaniu. Wariancja składowych drugiej i trzeciej to odpowiednio 13% i 7% wariancji ogólnej. Zatem użycie tylko trzech składowych głównych zamiast 32 czynników zewnętrznych wykorzystuje około 2/3 dostępnej informacji. Stopień tej kompresji zależy wyłącznie od wzajemnego skorelowania czynników zewnętrznych; zdarzają się przypadki, w których 1 składowa główna zawiera ponad 80% ogólnej wariancji.

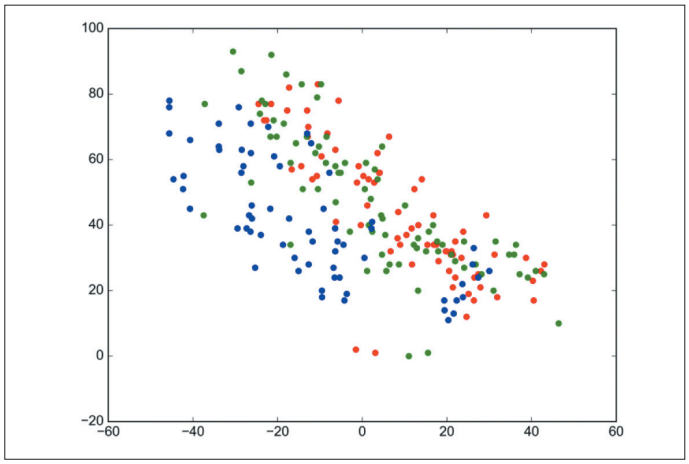
Stosowanie PCA uwalnia nas od nadmiaru współzależnych czynników zewnętrznych – z drugiej strony utrudnia interpretację uzyskanego modelu. O ile interpretacja mieszanki stanowiącej 1. składową główną (por. tab. 1) nie nastręcza problemów – jest ona jakimś agregatem prędkości w rejonach (z dominacją rejonów bliskich centrum aglomeracji) oraz dość nieistotnych przy nich danych pogodowych – o tyle skład kolejnych mieszanek jest trudny do oceny jakościowej. Do 2. składowej wchodzą ze znakiem dodatnim prędkości w rejonach centralnych, a z ujemnym prędkości w rejonach peryferyjnych; można ją interpretować jako szczególny rodzaj bilansu. W 3. składowej widzimy nagłą wysoką pozycję temperatury oraz trudną do zinterpretowania mieszankę pozostałych prędkości. Trudności w interpretacji wynikają stąd, że kolejne składowe główne tworzone są tak, aby wykorzystać informację resztkową z poprzednich

etapów PCA, niewykorzystaną przez składowe o wyższej randze. Dlatego PCA, podobnie jak wiele innych współczesnych technik modelowania (np. sieci neuronowe), za swoją skuteczność każe płacić hermetycznością modelu zależności.

Wprowadzając istotne ograniczenie liczby zmiennych zewnętrznych, PCA ułatwia również prezentację danych w wygodny sposób. Na rysunkach 9 i 10 przedstawiono liczby przejazdów z Pragi-Południe do Śródmieścia w funkcji 1. składowej głównej. Układ chmury punktów wyraźnie świadczy o zależności natężenia ruchu pomiędzy tymi dzielnicami i wartości 1. składowej głównej. Ewidentna korelacja nie oznacza jednak, że natężenie ruchu jest determinowane zagregowaną prędkością w aglomeracji. Pokolorowanie punktów odpowiadające poszczególnym godzinom szczytu porannego na rysunku 9 wskazuje, że tutaj czynnikiem podstawowym jest dzienna okresowość ruchu, a obserwowane wstęgowe ułożenie punktów wynika raczej ze specyfiki układu transportowego. Podobnie objawia się w tych samych danych sezonowość roczna, por. rysunek 10. Przy takim stopniu zdeterminowania więźby ruchu czynnikami sezonowymi modelowanie wpływu dodatkowych czynników (np. pogodowych) musiałoby zostać przeprowadzone dla wąskiego podzbioru danych, co w sposób naturalny ogranicza sensowność takiego przedsięwzięcia z powodu skąpej liczby próbek uczących.



Rys. 9. Liczba przejazdów z Pragi-Płd. do Śródmieścia w funkcji 1. składowej głównej. Kolorami oznaczono godziny dnia: 6:00, 7:00, 8:00, 9:00, 10:00



Rys. 10. Liczba przejazdów z Pragi-Płd. do Śródmieścia w funkcji 1. składowej głównej. Kolorami oznaczono miesiące: luty, marzec, kwiecień

Parametry modeli liniowych

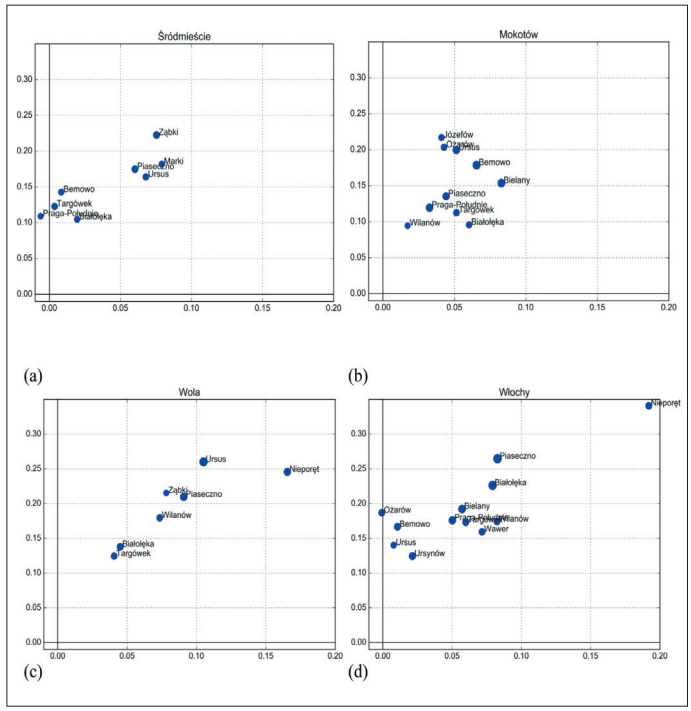
Nawet jeśli obserwowane czynniki zewnętrzne, takie jak składowe główne, nie stanowią pierwotnej przyczyny decyzji o przejazdach, możemy wykorzystywać je wtórnie do budowy modeli prognostycznych. Zdecydowaliśmy się wykorzystać proste modele liniowe w postaci

$$y = c_0 + c_1 x_1 + c_2 x_2 + \dots \ ,$$

gdzie y jest prognozowaną wartością (wyjściem modelu), $x_1, x_2, \dots$ to zmienne objaśniające (wejścia modelu), zsumowane po pomnożeniu przez współczynniki $c_1, c_2, \dots$ – odpowiednio. Model posiada też wyraz wolny, c_0. Moglibyśmy też wykorzystać wszechstronne modele nieliniowe, jak i modele specjalizowane — por. np. nieliniowe modele zależności czasu podróży pomiędzy miejscami zatrzymania [7]. Nie mamy jednak podstaw merytorycznych, aby stosować którykolwiek z modeli specjalizowanych, ani dostatecznej liczby próbek, aby stosować wieloparametrowe modele nieliniowe. W naszej analizie ograniczymy się do modeli o dwu wejściach.

Na rysunku 11 przedstawiono parametry modeli regresyjnych liczby przejazdów do wybranych czterech warszawskich dzielnic o dominującym charakterze biurowym. Wyjściem modeli jest liczba przejazdów w konkretnej relacji t w godzinie w zależności od wejść – wartości dwóch pierwszych składowych głównych zaobserwowanych w godzinie $t - 1$. Można dostrzec następujące zjawiska:

- praktycznie wszystkie modele położone są w I ćw. układu współrzędnych, czyli wykazują ten sam rodzaj zależności od składowych głównych;



Rys. 11. Położenie modeli ruchu do wybranych czterech dzielnic Warszawy na płaszczyźnie parametrów modelu regresji linowej. Wielkość znaczników odpowiada jakości modelu mierzonej statystyką R^2; uwzględniono tylko modele dla których $R^2 > 0{,}3$. (a) Dzielnica Śródmieście, (b) Dzielnica Mokotów, (c) Dzielnica Wola, (d) Dzielnica Włochy

- modele dostatecznej jakości dotyczą z reguły ruchu z rejonów sąsiadujących z dzielnicą docelową;
- siła uzależnienia od składowych głównych, tj. wartość parametrów modelu, rośnie wraz z ze wzrostem odległości pomiędzy rejonem początkowym a dzielnicą docelową (por. np. Piaseczno, Bemowo, Praga-Płd., Wilanów);
- istnieją wyjątki od powyższych obserwacji, np. peryferyjne Białołęka i Targówek pozostają dość niezależne od składowych głównych.

Powyższą analizę można wykorzystać jakościowo do dalszego poszukiwania przyczyn zasygnalizowanych prawidłowości. Można również wykorzystać ją ilościowo do prognozowania więźby ruchu w kolejnej godzinie na podstawie bieżących obserwowanych prędkości w rejonach, bez wnikania w istotę modelu. To ostatnie podejście zyskało dużą popularność w różnych dziedzinach; zauważmy jednak, że modele nauczone maszynowo pozbawiają użytkowników wglądu w przyczyny występowania określonego zjawiska. Jest to źródłem dyskomfortu decydenta pragnącego wniknąć w istotę zjawiska – aby wpływać na zjawisko, monitorować jego ewolucję, kształtować jego postrzeganie przez innych itp.

Podsumowanie

Zaprezentowane wyniki dowodzą, że dane o przejazdach dotyczących wyłącznie użytkowników systemu Yanosik są wystarczająco reprezentatywne do tworzenia wartościowych, ogólniejszych modeli matematycznych. Wykorzystane przez nas niskowymiarowe liniowe modele regresyjne są adekwatne do liczby posiadanych danych i wiedzy o naturze obiektu. Podjęte próby modelowania więźby w weekendy nie doprowadziły do wytworzenia sensownych modeli, gdyż ruch powstały w kierunku rejonów o dużej liczbie centrów handlowych okazał się stosunkowo nieduży. Zaobserwowana w [8] reguła 50% wzrostu ruchu w takich centrach w weekendy oraz okresowość dzienna nie objawiła się w posiadanych danych ani w sensie ilościowym, ani jakościowym. Również analiza ruchu okołopołudniowego do dzielnic o charakterze rekreacyjnym (Śródmieście, Wilanów, Praga-Północ) w święta wolne od handlu nie doprowadziła do spójnych modeli zależnych od jakichkolwiek zmiennych objaśniających.

Arsenał środków technicznych monitorowania przejazdów rośnie i tanieje, obejmując – oprócz świadomie użytkowanych aplikacji mobilnych – w pełni pasywne metody śledzenia, oparte na monitoringu wizyjnym, systemach parkowania, a w przyszłości zapewne dane dostarczone przez producentów samochodów i wypożyczalnie. Obfitość danych ułatwi modelowanie oparte o uczenie maszynowe. Tymczasem interesującą perspektywą badawczą wydaje się być modelowanie z uwzględnieniem aktualnego bilansu pojazdów w poszczególnych rejonach, jako *stanów* systemu transportowego postrzeganego jako układ dynamiczny. Jednym ze sposobów opisu matematycznego takich układów są *modele autoregresyjne z wejściem zewnętrznym* (ARX) w ogólnej postaci:

$$y(t) + d_1 y(t-1) + d_2 y(t-2) + \ldots =$$
$$= c_0 x(t) + c_1 x(t-1) + c_2 x(t-2) + \ldots .$$

Na wartość wyjścia modelu w chwili t mogą mieć wpływ zarówno aktualne, jak i historyczne wartości wejść i wyjść. (W odniesieniu do opisywanego tutaj scenariusza wyjściem modelu jest liczba przejazdów na pojedynczej relacji, a wejściami – wartości zmiennych objaśniających).

Warto zauważyć, że ubocznym skutkiem upowszechniającego się monitoringu pojedynczych przejazdów może być naruszenie prywatności podróżnych, gdyż poruszają się oni po rutynowych, unikatowych trajektoriach. Możliwe jest więc opracowanie indywidualnych profili aktywności, a nawet, łącząc dane ruchowe z informacjami z różnych rejestrów i portali społecznościowych, zidentyfikowanie tożsamości podróżnych. Aby temu zapobiec, stosowane są różne techniki anonimizacji przejazdów [9]. Prace własne autorów doprowadziły do zaproponowania metody anonimizacji siatki ulic, przy zachowaniu istotnych informacji o stylu jazdy kierowców. Wykorzystano w tym celu zaobserwowaną prawidłowość, że najbardziej reprezentatywnym etapem dla dynamiki jazdy pomiędzy zatrzymaniami jest etap końcowy obejmujący około 100 m [10, s. 233–235].

Literatura

1. http://www.slownikkbr.pl
2. Erlander S., Stewart N. F., *The gravity model in transportation analysis: theory and extensions* (Vol. 3), Vsp., 1990.
3. Dybicz T., *Modelowanie i symulacje ruchu, rys historyczny i aktualnie stosowane oprogramowanie*, „ Zeszyty Naukowo-Techniczne SiTK RP Oddział Kraków", 2009, z. 90.
4. Vardi Y., *Network tomography: Estimating source-destination trafic intensities from link data*, „Journal of the American Statistical Association", 1996, vol. 91, no. 433.
5. Kamola M., *Estimation of correlated flows from link measurements,* Methods and Models in Automation and Robotics (MMAR), 20th International Conference, IEEE, 2015.
6. Calabrese F., Di Lorenzo G., Liu L., Ratti C., *Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area,* IEEE Pervasive Computing, 2011, 10(4) .
7. Birr K., Jamroz K., Kustra W., *Analiza czynników wpływających na prędkość pojazdów transportu zbiorowego na przykładzie Gdańska,* „Prace Naukowe Politechniki Warszawskiej – Transport", 2013, nr 96.
8. Romanowska A., Jamroz K., *Wielkopowierzchniowe obiekty handlowe – zwykłe generatory ruchu czy źródła problemów transportowych?*, „Transport Miejski i Regionalny", 2015, nr 2.
9. Zan B., Sun Z., Gruteser M., Ban X., *Linking anonymous location traces through driving characteristics,* Proceedings of the third ACM conference on Data and application security and privacy (pp. 293-300), ACM, 2013.
10. Kamola M., Arabas P., *Sieci społeczne i technologiczne. Jak zrozumieć, jak wykorzystać,* PWN, Warszawa 2018.