

Szacowanie inflacji usług taksówkarskich w Polsce w oparciu o dane skrapowane

V Kongres Statystyki Polskiej

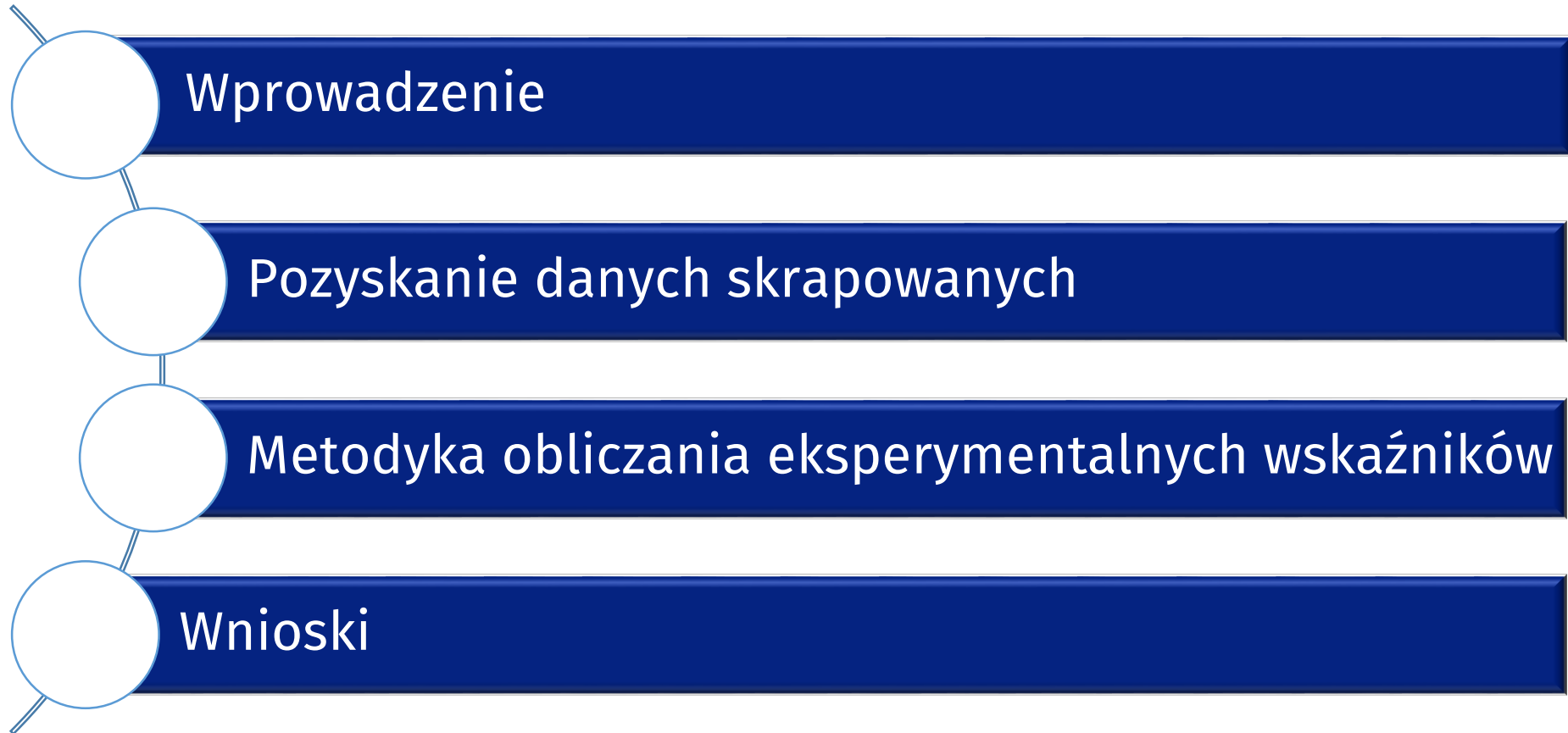
Jacek Białek

Beata Cebula

Agnieszka Matulska-Bachura

Departament Handlu i Usług

Plan prezentacji



Wprowadzenie

- Prace realizowane w ramach umowy o dotację KE *Terminowa, bardziej odpowiednia i lepiej zintegrowana europejska statystyka przedsiębiorstw; obszar Ulepszenie procesów tworzenia europejskich statystyk*; (nr 101143415 – 2023-PL-EBS);
- Cel główny - rozpoznanie możliwości wykorzystania informacji gromadzonych na stronach internetowych przewoźników na potrzeby usprawnienia procesu zbierania i przetwarzania danych w zakresie statystyki cen producentów usług;
- Cele szczegółowe, tj.:
 - ✓ pozyskanie danych na temat cen wybranych usług transportu lądowego pasażerskiego poprzez web-scraping;
 - ✓ rozwój metodyki obliczania wskaźników cen producentów usług w zakresie transportu lądowego pasażerskiego uwzględniających wykorzystanie danych pochodzących z różnych źródeł;
- Stan obecny – wskaźniki cen usług taksówkarskich:
 - ✓ statystyka cen producenta – dane pochodzą z cenników firm i korporacji taksówkarskich;
 - ✓ statystyka cen detalicznych – dane zbierane przez ankietatorów;



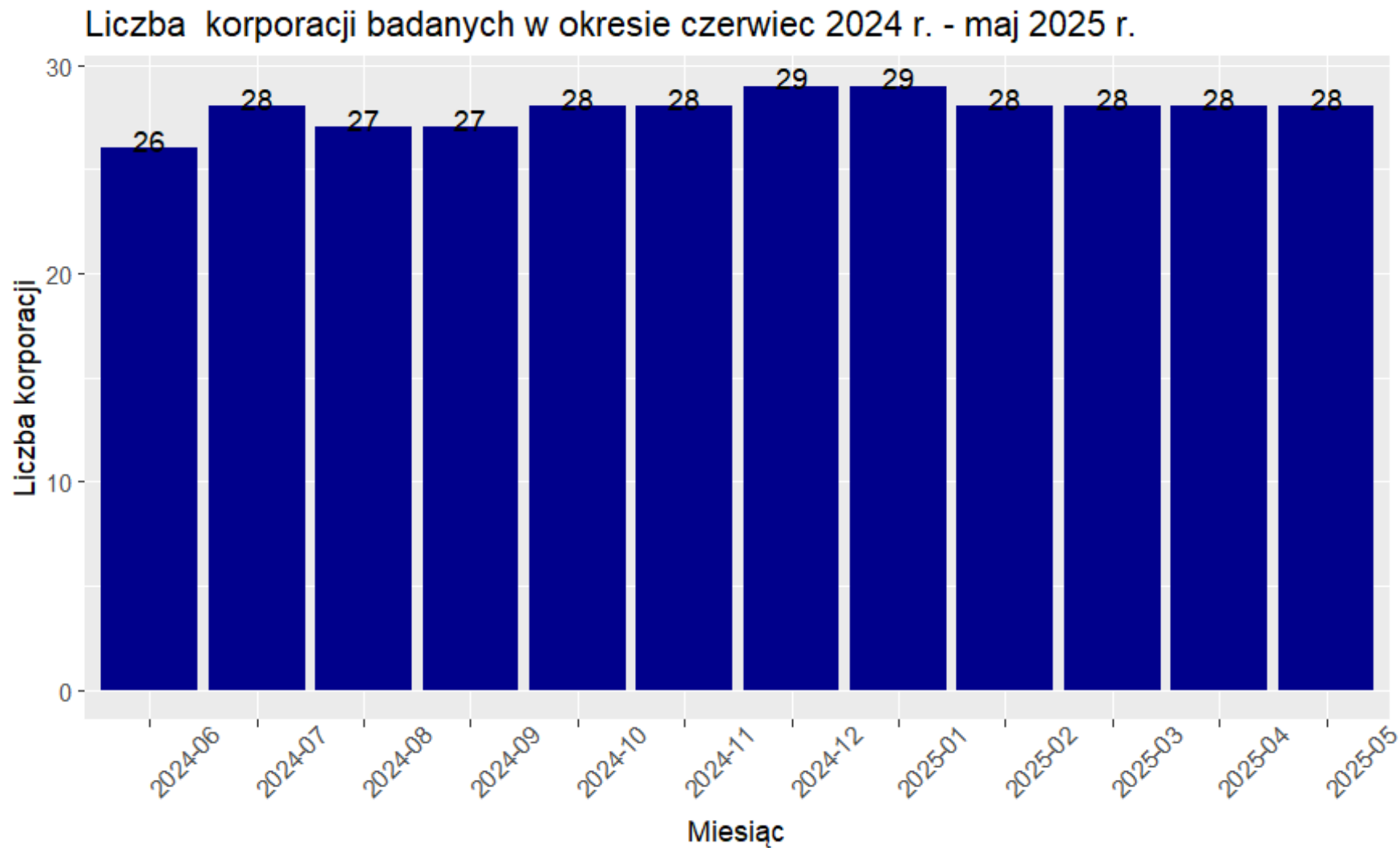
Pozyskanie danych skrapowanych

- Rozpoznanie pozastatystycznych źródeł danych gromadzących informacje na temat działalności taksówek osobowych;
- Ustalenie schematów struktury stron internetowych zawierających informacje o cenach świadczonych usług;
- Opracowanie narzędzia umożliwiającego automatyczne pozyskiwanie danych ze stron internetowych;
- Pozyskanie danych pochodzących ze stron internetowych zgodnie z opracowanymi schematami (web-scraping):
 - ✓zakres zbieranych informacji: miasto, nazwa korporacji, adres strony, na której zamieszczany jest cennik, data notowania, cena za opłatę początkową i cena za km w taryfie 1;
 - ✓częstotliwość – dzienna;
 - ✓okres analiz – czerwiec 2024 r. – maj 2025 r.;

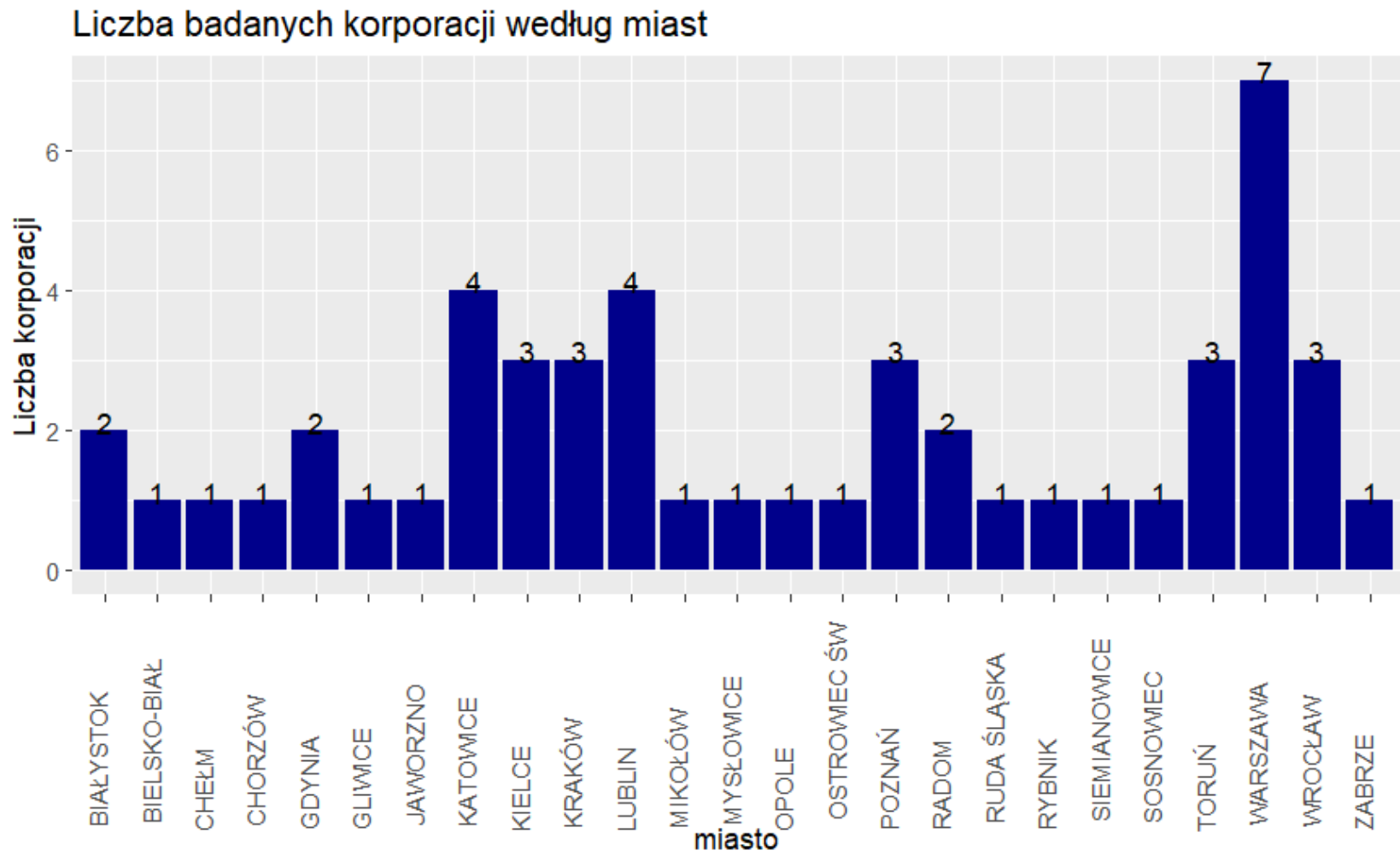
Miasta objęte skrapowaniem



Pozyskanie danych skrapowanych



Pozyskanie danych skrapowanych



Metodyka obliczania eksperymentalnych wskaźników

- Ustalenie reprezentanta usług – przejazd 10 km w ramach taryfy 1 (średnia długość przeciętnego kursu przejazdu taksówką w mieście);

$$P_m = P0_m + 10\text{km} \times Pkm_m,$$

gdzie:

P_m - cena za usługę w danym miesiącu;

$P0_m$ - opłata początkowa w danym miesiącu;

Pkm_m - cena za 1 km przewozy w taryfie 1 w danym miesiącu;

- Ustalenie ceny za miesiąc dla każdej korporacji jako średnia geometryczna z dziennych notowań;
- 3 podejścia do badania zmian cen:
 - ① **tzw. unit value indeks** – uwzględniane są wszystkie obserwacje w okresie badanym (licznik) i wszystkie obserwacje w okresie bazowym (mianownik);
 - ② **bilateralny indeks Jevonsa** - dla obserwacji występujących jednocześnie w okresie badanym i bazowym;
 - ③ **łańcuchowy indeks Jevonsa** - dla każdego badanego okresu obliczane są wskaźniki zmian cen w odniesieniu do okresu poprzedniego dla obserwacji wspólnych, tj. w ramach danego miast dla tych samych korporacji – rekomendowane!!!

Metodyka obliczania eksperymentalnych wskaźników

- dwuetapowa procedura obliczania wskaźników:

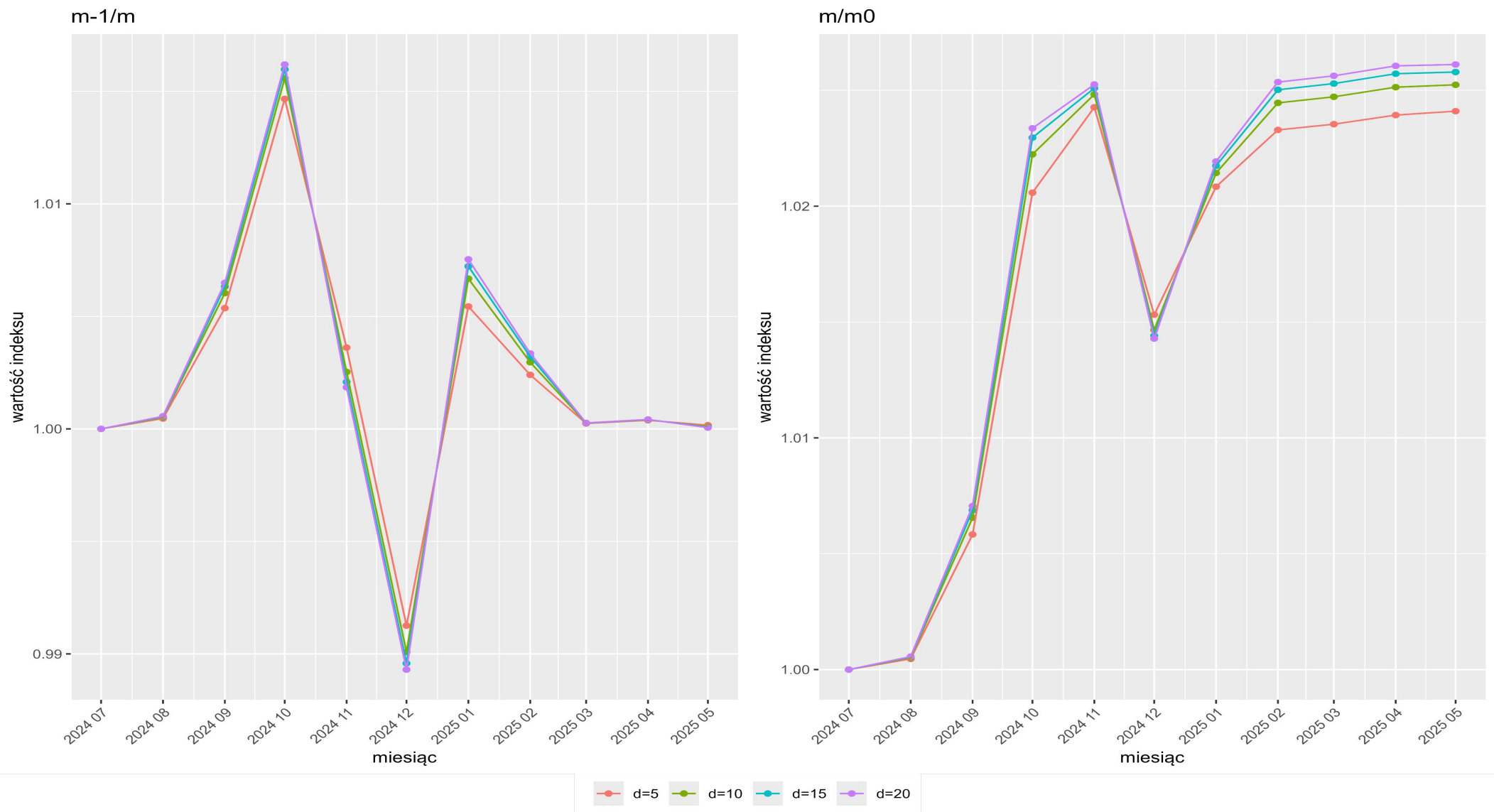
① cena dla miasta jako **średnie geometryczne**;

② cena dla kraju jako **średnia geometryczna ważona** z cen dla miast,

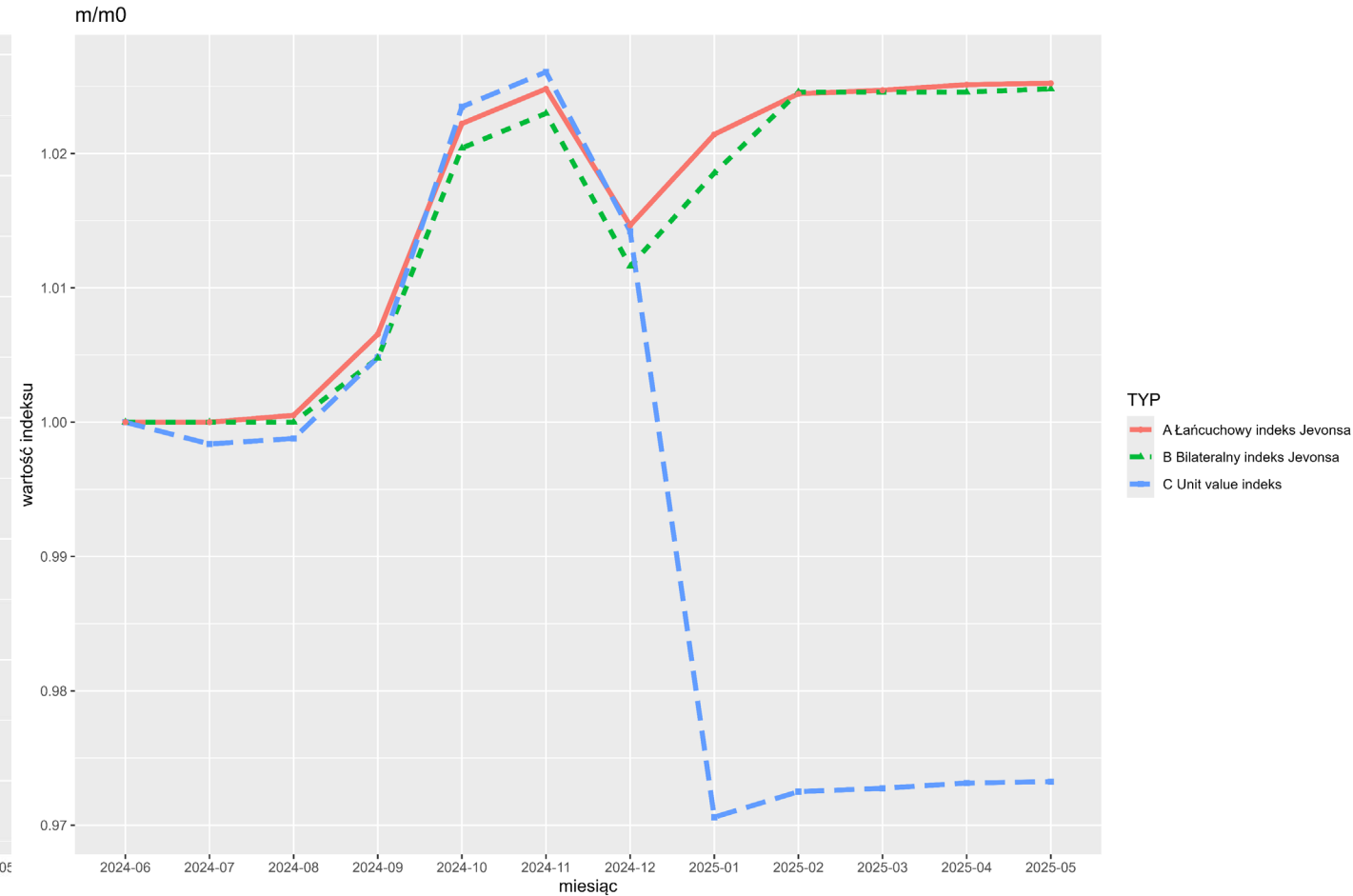
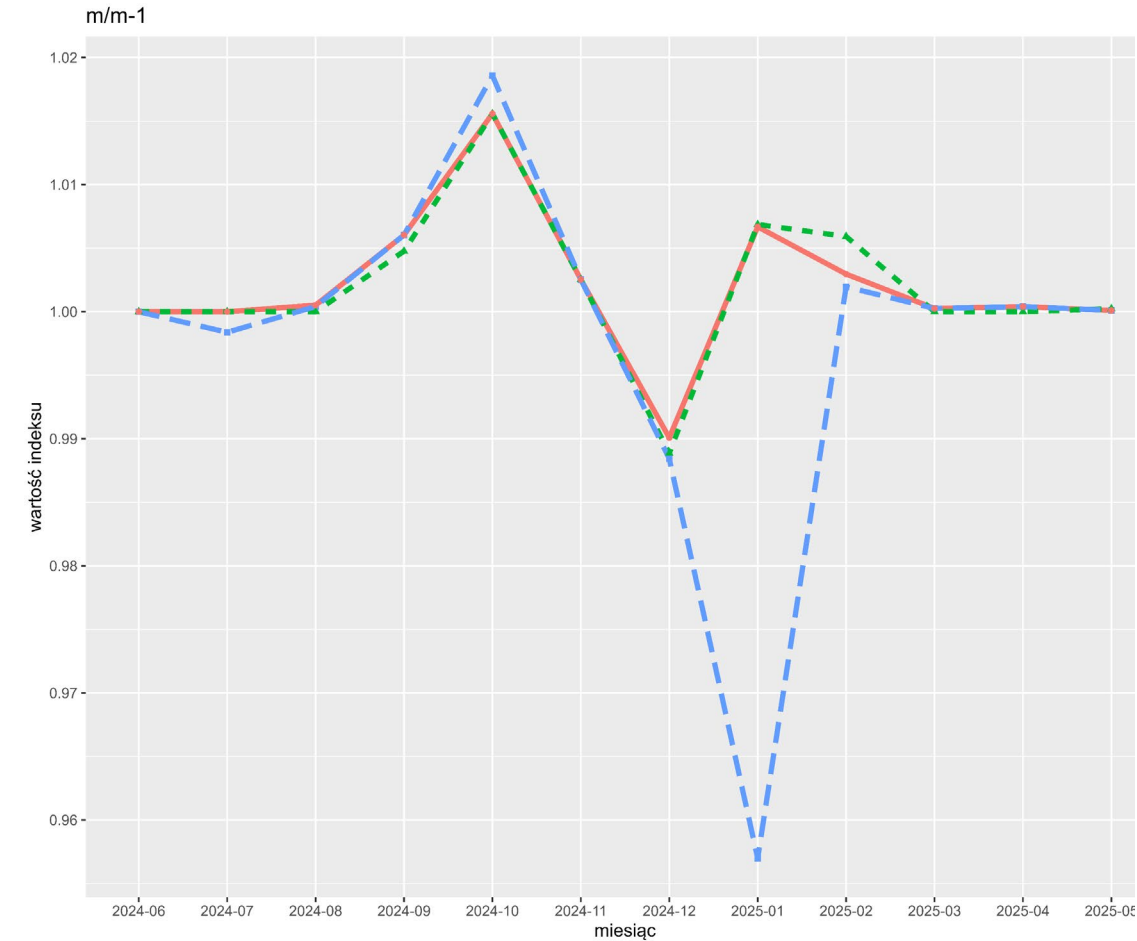
przy czym rozważano metody ważenia wskaźnika cen:

- ✓ wagi według liczby skrapowanych w danym mieście firm taksówkarskich,
- ✓ wagi według liczby ludności miast,
- ✓ wagi według wielkości miast w km²,
- ✓ wagi według liczby ludności i rozchodów na 1 mieszkańca,
- ✓ wagi według obrotów firm taksówkarskich prowadzących działalność w danym mieście,
- ✓ **brak wag** – nieważony wskaźnik cen - **średnia geometryczna cen dla miast**;

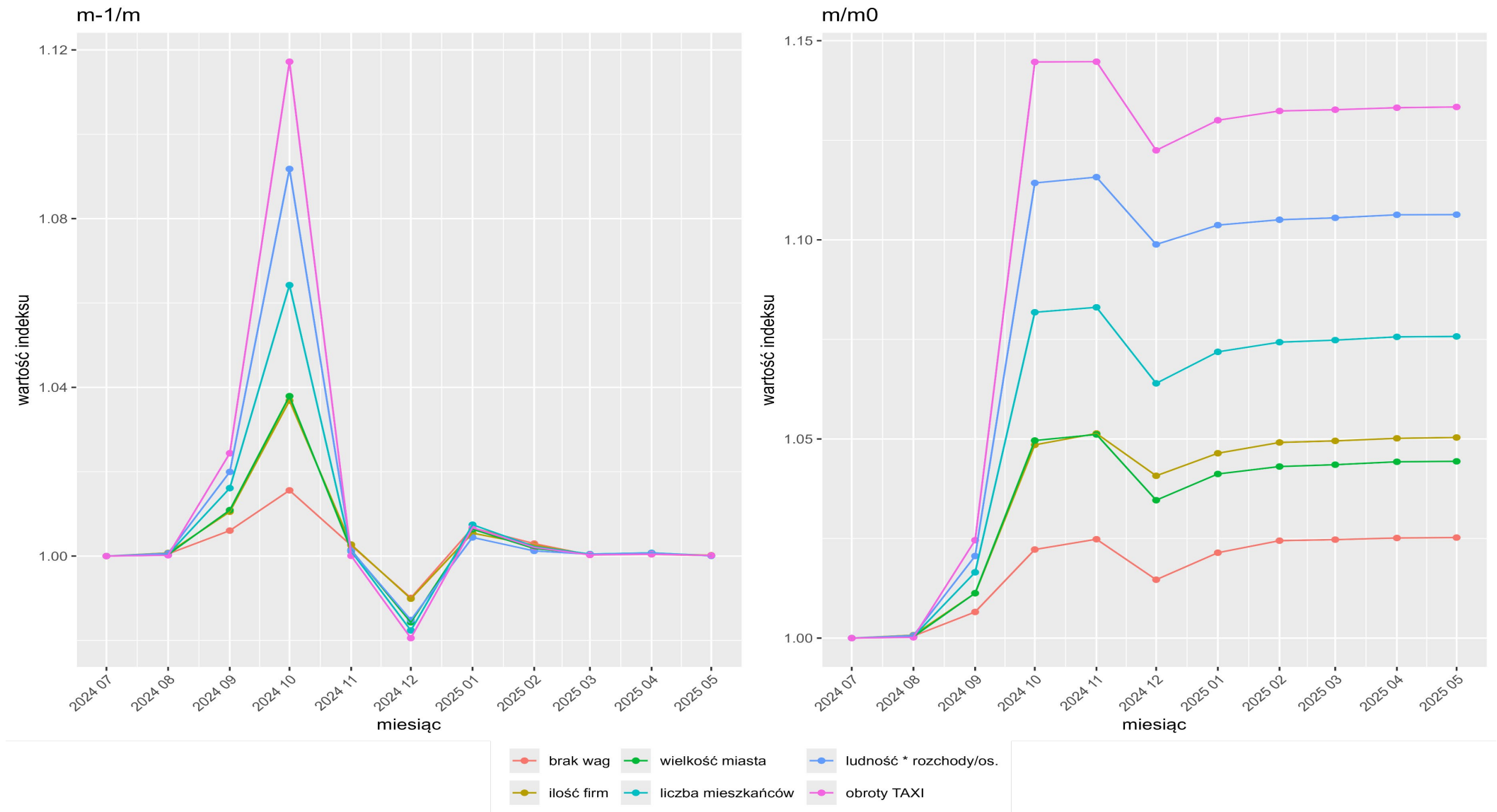
Metodyka obliczania eksperymentalnych wskaźników – długość trasy



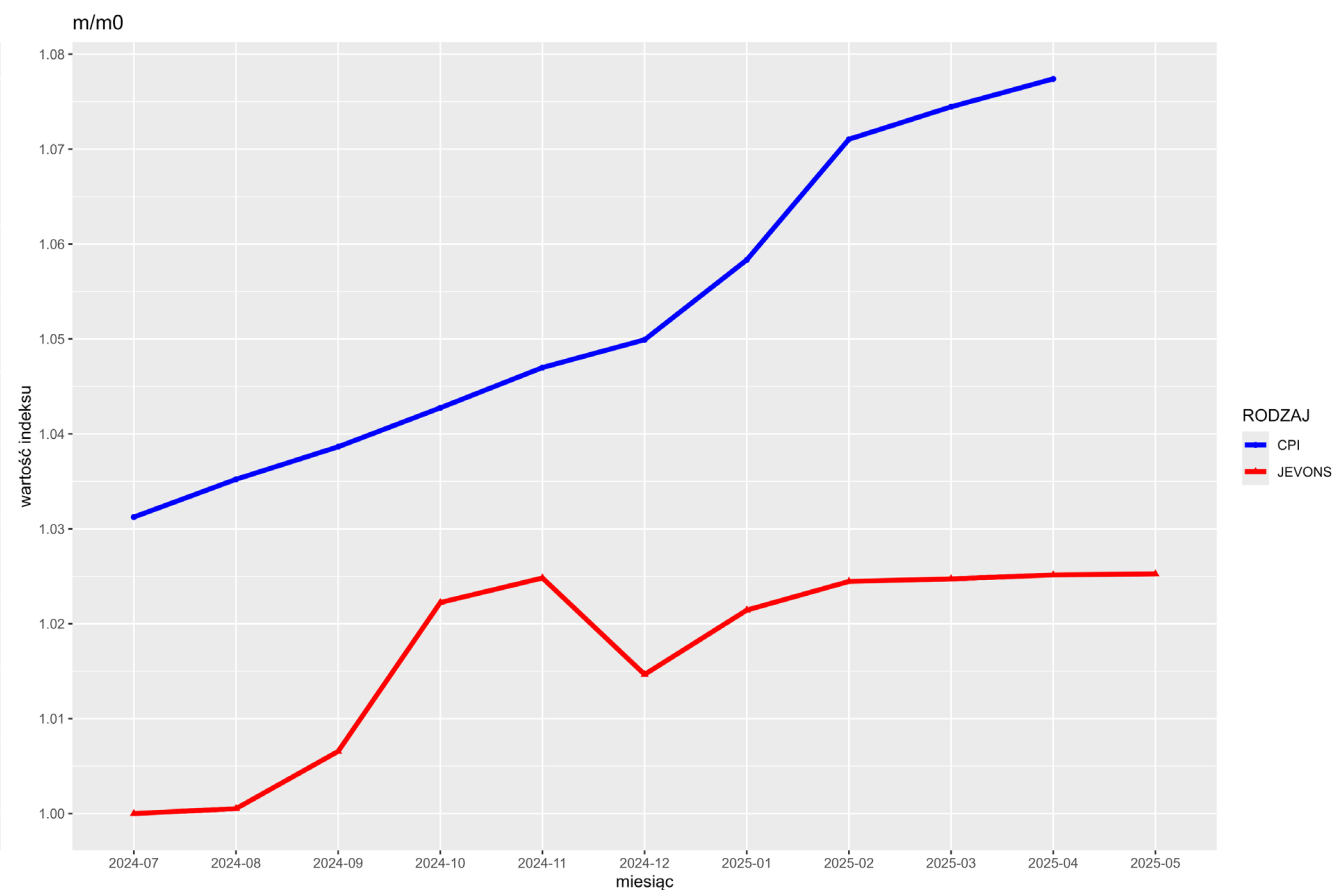
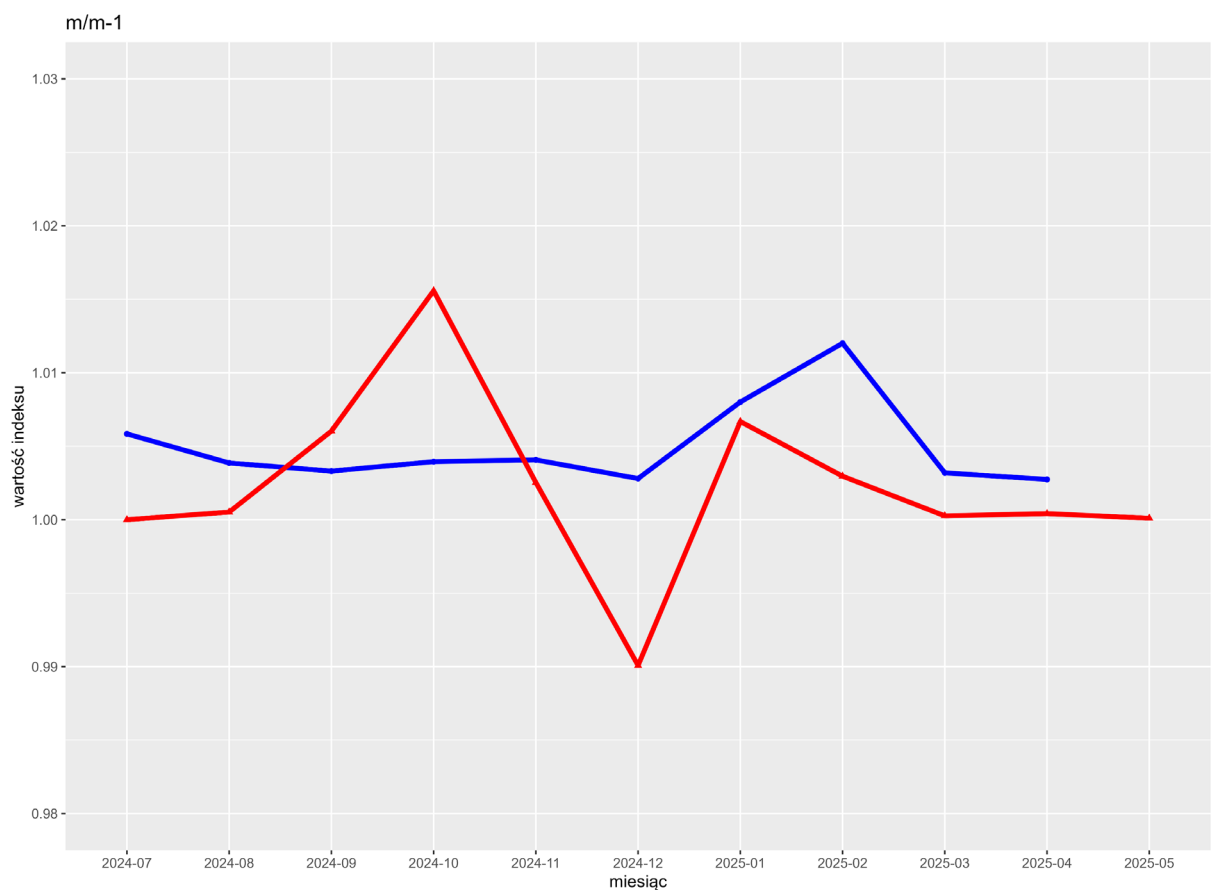
Metodyka obliczania eksperymentalnych wskaźników – typ indeksu



Metodyka obliczania eksperymentalnych wskaźników – system wag



Porównanie eksperymentalnych wskaźników cen z wynikami badania cen konsumpcyjnych



Wnioski

- Skrapowanie danych bardziej wydajne i szybsze w porównaniu z tradycyjnym zbieraniem danych;
- Dostęp do większej ilości informacji – w przypadku statystyki cen – większej liczby reprezentantów;
- Konieczność opracowania narzędzia do automatycznego zbierania danych ze stron internetowych i jego aktualizacji (zapewnienie administratora bazy danych);
- Potrzeba archiwizacji danych i zapewnienia miejsca na serwerze;
- Wdrożenie narzędzi informatycznych umożliwiających obliczanie wskaźników cen na podstawie danych skrapowanych, np. środowisko R;
- Niezbędne dostosowanie metodyki obliczania wskaźników cen do źródła danych – duża liczba notowań, brak informacji o wielkości obrotów, duża dynamika zmian w badanej populacji;



- W zależności od podejścia do konstrukcji indeksu uzyskano znaczne różnice w wartościach wskaźników cen – rekomendowane podejście tzw. dynamiczne na bazie łańcuchowego indeksu Jevonsa;
- Wybór długości trasy nie wydaje się tak istotny w porównaniu z wyborem systemu wag;
- Wyniki ze skrapingu w porównaniu z tradycyjnym źródłem danych (mimo, że odbiegają od siebie) wskazują na ten sam kierunek zmian;
- Aby móc generalizować przedstawione wnioski konieczne jest prowadzenie dalszych analiz na dłuższych szeregach czasowych;

V Kongres Statystyki Polskiej

Dziękujemy za uwagę 😊

Jacek Białek (J.Bialek@stat.gov.pl)
Beata Cebula (B.Cebula@stat.gov.pl)
Agnieszka Matulska-Bachura (A.Matulska@stat.gov.pl)
Departament Handlu i Usług