Urząd Statystyczny
w Opolu

Grupa homogeniczna w danych scrapowanych jako źródło informacji o dynamice cen

Tomasz Markuszewski, Katarzyna Widera

Urząd Statystyczny w Opolu
Ośrodek Statystyki Cen
Dział Inżynierii Danych

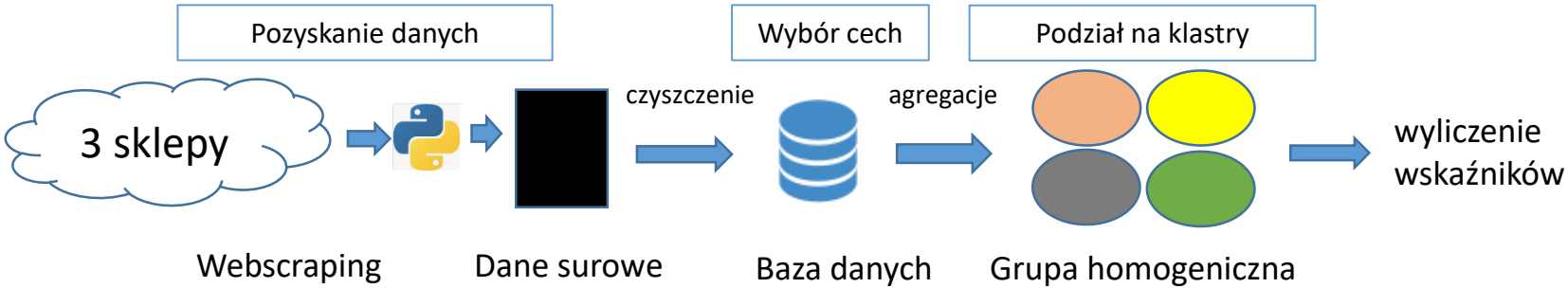
Warszawa 2-4.07.2024 r.

Plan prezentacji

- Pozyskanie i walidacja danych scrapowanych
- Grupa homogeniczna jako reprezentant
- Metody obliczania wskaźnika cen
- Eksperyment, wyniki i podsumowanie

Pozyskanie i walidacja danych scrapowanych

Badanie ma charakter eksperymentalny i składa się z następujących zadań:



Metodyka eksperymentu:

Zadanie 1. Pozyskanie danych przez okres wielu miesięcy oraz zapis do bazy danych
Zadanie 2. Agregacja w czasie, uśrednienie cen do miesięcznych oraz agregacja produktów (klucz: „model produktu")
Zadanie 3. Podział na grupy homogeniczne, musi uwzględniać cechy dostępne na stronach internetowych, jak i cechy mające wpływ na cenę i najbardziej różnicujące
Zadanie 4. Wyliczanie wskaźników

Grupa homogeniczna jako reprezentant

Teoria **C**lustering **L**arge datasets **I**nto **P**rice indices (**CLIP**) (Wielka Brytania)
zakłada ze klient chce kupić pewne typy produktów niż konkretny produkt
oczywiście niektórzy klienci zawsze kupują to samo (lojalność względem marki),
inni kupują to co jest akurat dostępne albo produkt, który oferuje lepszą jakość do ceny,
na podstawie dostępnej oferty mając na uwagę określony budżet.

Wybór cech do budowy grupy

Podział manualny na grupy homogeniczne

Podział na grupy za pomocą algorytmu (ML)

- **https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip**

Cechy opisujące reprezentanta i ich charakterystyki liczbowe

Sklep	Liczba obserwacji		Średnia	Mediana	Minimum	Maksimum	Kwartyl dolny	Kwartyl górny	Odch.std
1	739	Cena	1777,00	1649	739	3299	1397	2099	534,78
2	2420		1799,13	1609,86	549	6249	1334,36	2107,57	682,5
3	1685		1824,42	1686,5	761,04	5999	1396,14	2079	680,97
4	2491		1885,70	1712,5	686,66	16829	1386,33	2240	803,39

Rodzaj Piekarnika	Liczba obserwacji		Średnia	Mediana	Minimum	Maksimum	Kwartyl dolny	Kwartyl górny	Odch.std
Gazowy	516	Cena	1260,99	1208,09	549	3209	1032,81	1489,25	316,93
Elektryczny	6819		1875,33	1720,71	699	16829	1407,15	2199	717,02

TypPłyty	Liczba obserwacji		Średnia	Mediana	Minimum	Maksimum	Kwartyl dolny	Kwartyl górny	Odch.std
Gazowa	4410	Cena	1623,51	1520,17	549	16829	1261	1794	663,77
Inne	178		1848,52	1420,38	995,4	5999	1134	1645,15	1294,07
Ceramiczna	913		1930,88	1899	963,57	3729,99	1644,99	2199	434,97
Indukcyjna	1834		2282,95	2299	899	6249	1799	2699	640,87

Wymiar	Liczba obserwacji		Średnia	Mediana	Minimum	Maksimum	Kwartyl dolny	Kwartyl górny	Odch.std
50	4747	Cena	1606,14	1511,09	549	3299,99	1260,11	1863,55	480,55
59	44		3444,13	2879	1999	5999	2460,43	4941,5	1332,16
59,6	138		2608,92	2490,67	1479,99	3619	2346,14	2941,86	409,35
60	2077		2089,01	1899	959	6249	1649	2434,42	697,73
85	90		1907,61	1884,19	1311	2963	1704	2154,5	322,19
90	215		3404,79	3356,71	1699,99	15074	3219	3505	888,77
110	1		16829	16829	16829	16829	16829	16829	

Wybór cech w oparciu o narzędzia statystyczne

Podsumowanie regresji zmiennej zależnej: Price (kuchenki regresja)
R= ,62364808 R^2= ,38893693 Popraw. R2= ,38860347
F(4,7330)=1166,4 p<0,0000 Błąd std. estymacji: 558,23

N=7335	b*	Bł. std. z b*	b	Bł. std. z b	t(7330)	p
W. wolny			-1273,89	63,49210	-20,0637	0,000000
Wymiar	0,461870	0,009203	37,40	0,74522	50,1860	0,000000
TypPłyty	0,380160	0,009399	206,61	5,10811	40,4478	0,000000
RodzajPiekarnika	0,094327	0,009400	263,31	26,23994	10,0348	0,000000
ShopName	0,065490	0,009201	45,94	6,45386	7,1179	0,000000

Bez zmiennej – rodzaj sklepu:

Podsumowanie regresji zmiennej zależnej: Price (kuchenki regresja)
R= ,62025261 R^2= ,38471330 Popraw. R2= ,38446151
F(3,7331)=1527,9 p<0,0000 Błąd std. estymacji: 560,12

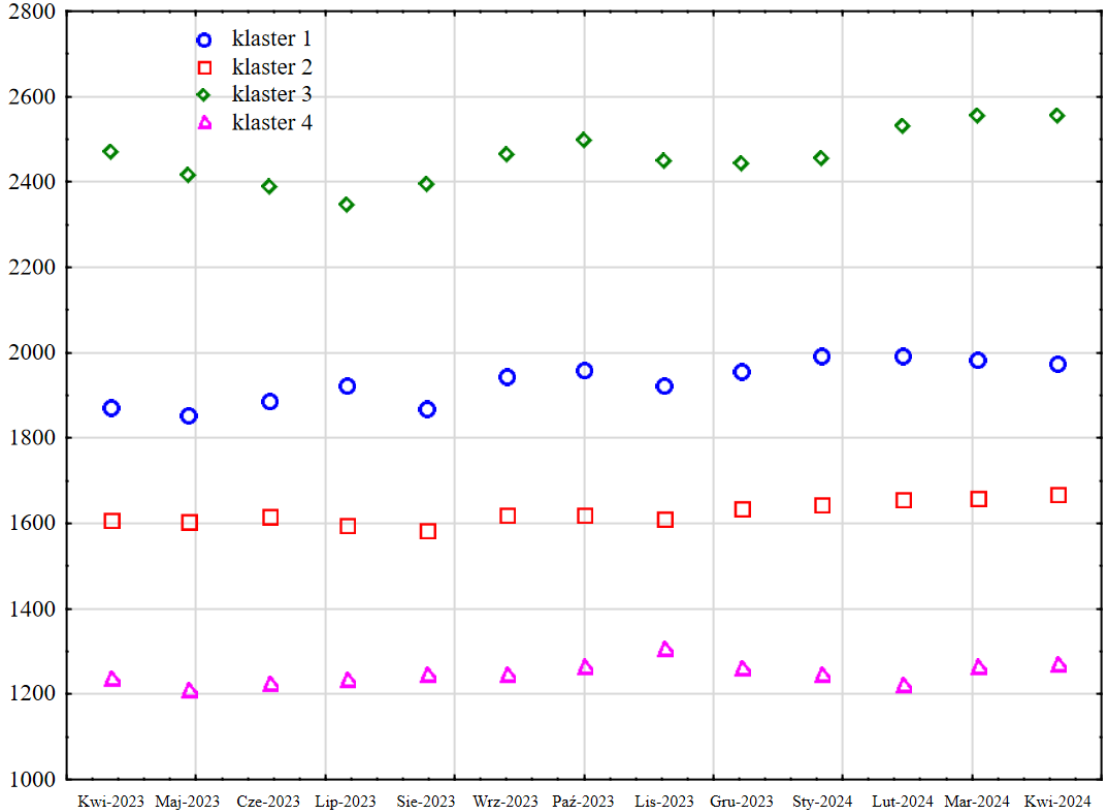
N=7335	b*	Bł. std. z b*	b	Bł. std. z b	t(7331)	p
W. wolny			-1153,55	61,40707	-18,7854	0,000000
Wymiar	0,466317	0,009213	37,76	0,74601	50,6153	0,000000
TypPłyty	0,373614	0,009385	203,05	5,10079	39,8084	0,000000
RodzajPiekarnika	0,093607	0,009431	261,30	26,32715	9,9252	0,000000

Bez zmiennej – wymiar kuchenki:

Podsumowanie regresji zmiennej zależnej: Price (kuchenki regresja)
R= ,41193910 R^2= ,16969382 Popraw. R2= ,16946734
F(2,7332)=749,24 p<0,0000 Błąd std. estymacji: 650,62

N=7335	b*	Bł. std. z b*	b	Bł. std. z b	t(7332)	p
W. wolny			666,1666	57,82639	11,52011	0,000000
TypPłyty	0,356507	0,010895	193,7571	5,92116	32,72284	0,000000
RodzajPiekarnika	0,143674	0,010895	401,0630	30,41254	13,18742	0,000000

Średnie ceny grupy homogenicznej

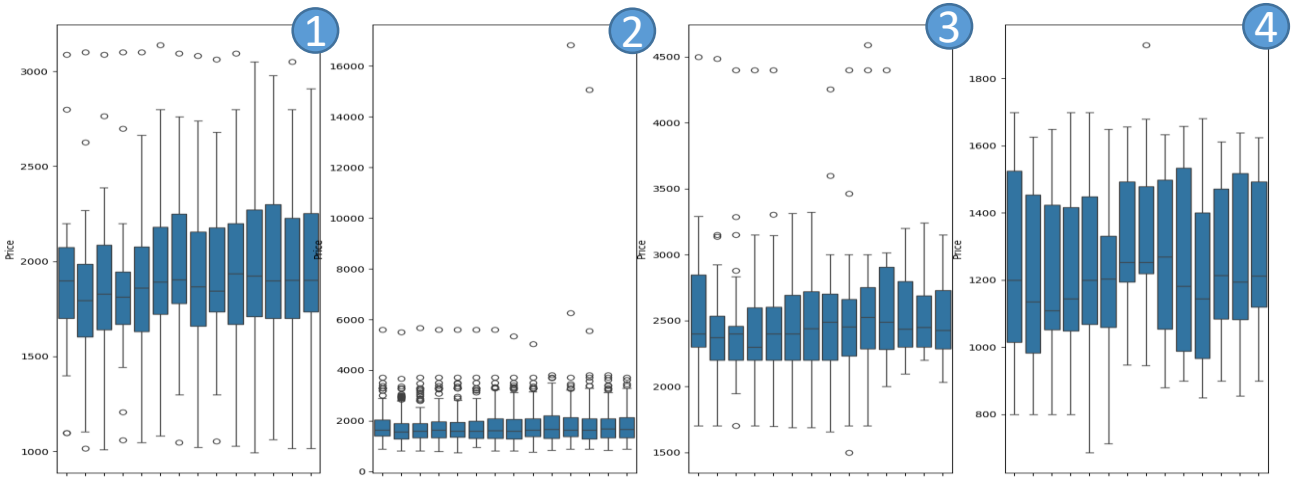


Można pozyskać informację jak zmienia się cena konkretnej grupy homogenicznej.

Wykresy pudełkowe grup homogenicznych

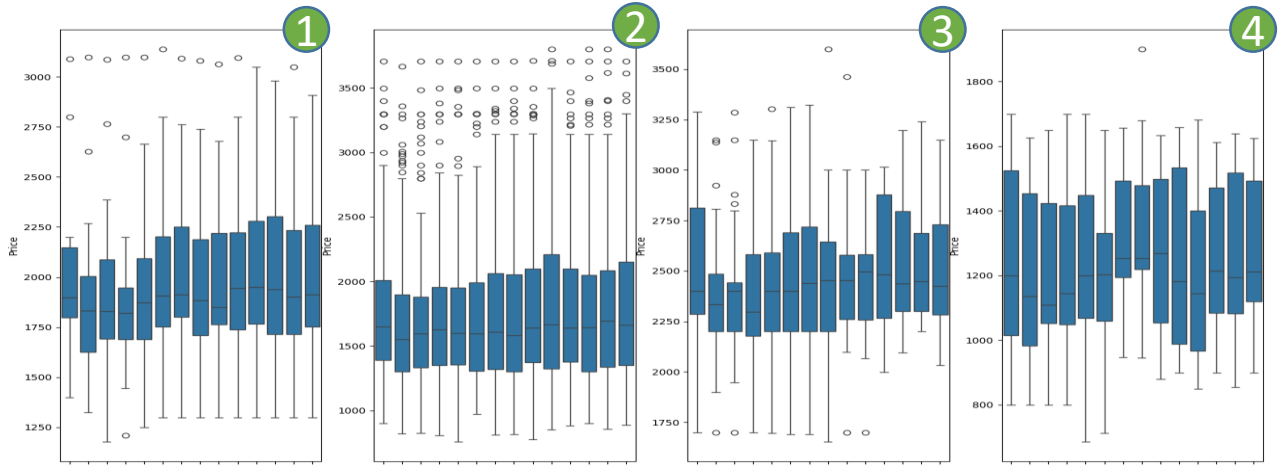
Grupy homogeniczne kuchenki

Przed czyszczeniem



Po usunięciu wartości odstających

Q1-2.5•Qx< cena <Q3+2.5•Qx



Metody wyliczania wskaźników

Brak informacji transakcyjnych o ilości sprzedanych towarów ogranicza liczbę dostępnych sposobów wyliczania wskaźników dla danych scrapowanych. Oto wybrane metody nie wymagające użycia tych wielkości:

- regresja hedoniczna

Wskaźniki bilateralne:

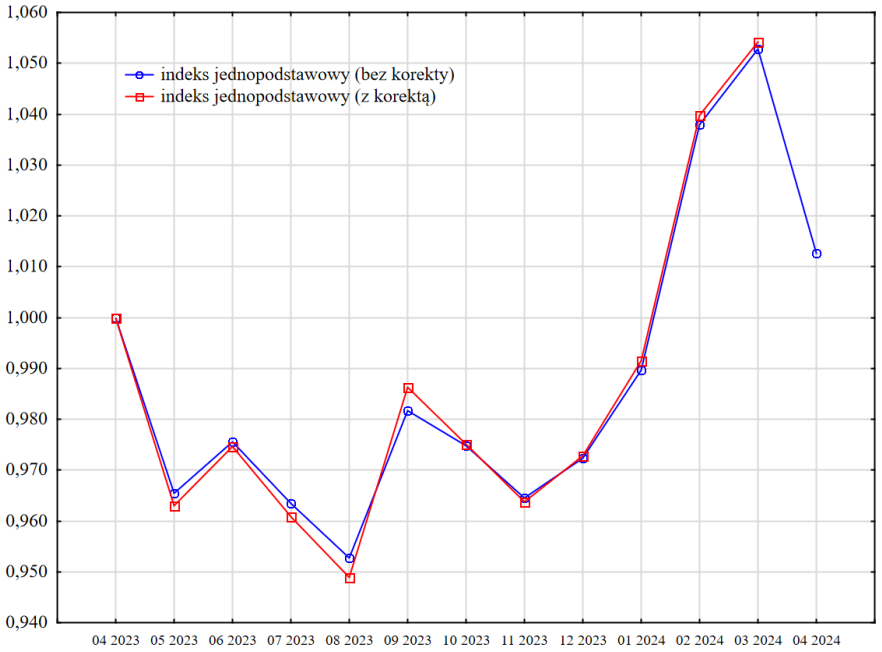
- Jevons (ur. 1835 – 1882) - iloraz średnich geometrycznych z cen
- Dutot (ur. 1684 – 1741*)* - iloraz średnich arytmetycznych z cen

Wskaźniki multilateralne:

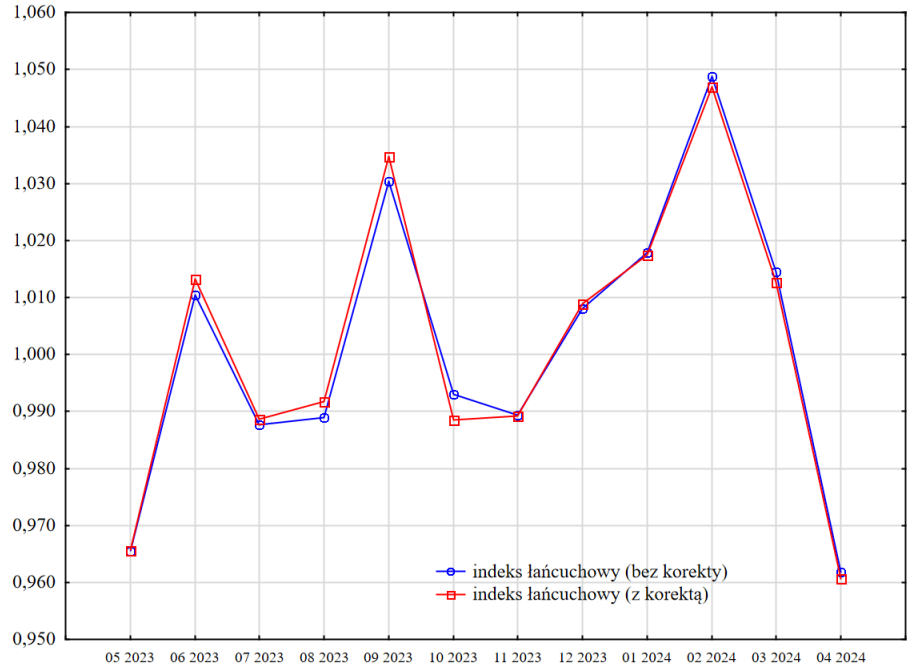
- liczone w GEKS (Gini-Eltetö-Köves-Szulc) uśredniając geometrycznie bilateralne indeksy w oknie czasowym T=13 miesięcy)

Indeks na podstawie regresji hedonicznej

$$I_n = \frac{p_n^{t+1}}{\hat{p}_n^t}$$



rys. 1



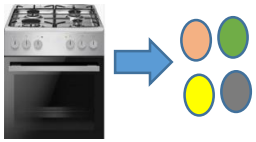
rys. 2

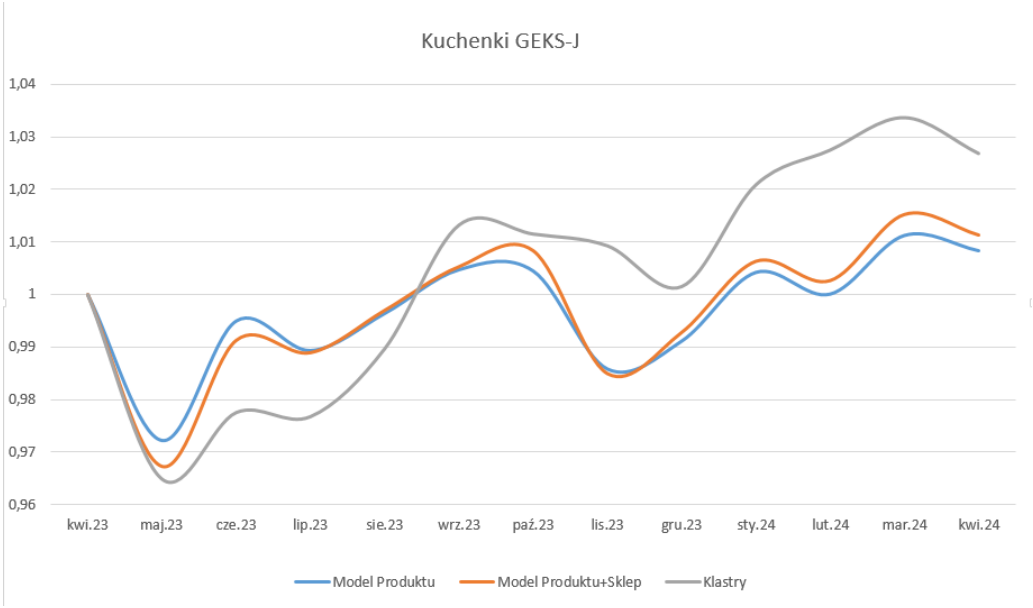
Eksperyment 1 - GEKS-J dla kuchenek

Kuch. elektryczno-gazowa Producent
58ME2.35HZpMs(W) kolor - … | SklepA

Reprezentant		
Model Produktu	Model Produktu+Sklep	Grupa homogeniczna (4 grupy)

manualny





Model Produktu
58ME2.35HZpMs(W)

Model Produktu+Sklep
58ME2.35HZpMs(W)**SklepA**

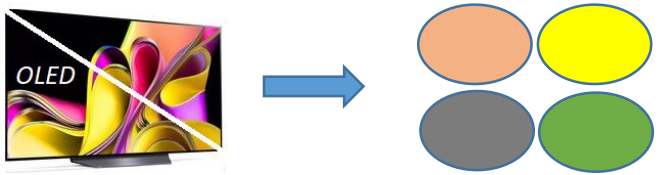
Palnik/Piekarnik (4 grupy)
Klaster 1
Klaster 2
Klaster 3
Klaster 4

Eksperyment 2 dla TV

Jeśli wybrane cechy produktów mają charakter ilościowy, po umieszczeniu w kartezjańskim układzie współrzędnych, podobne będą leżeć blisko siebie.

Budowanie grupy homogenicznej produktów w sposób automatyczny zakłada podział na grupy z wykorzystaniem algorytmu np. *MeanShift*

Algorytm np. *MeanShift*



Grupa homogeniczna

1.Wybór cech X1,X2 dla telewizorów:
X1- przekątna ekranu (ilościowa w calach),
X2- typ matrycy (jakościowa konwertowana na liczby => 1:LED,2:QLED lub 3:OLED)

2.Dla cech X1, X2 wykonujemy algorytm *MeanShift (*bandwith=5).

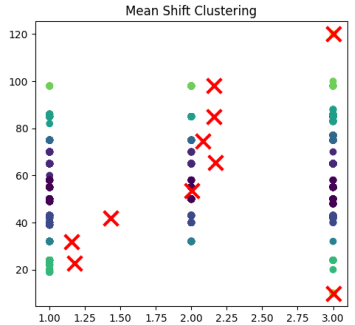
3.Klastrowanie - przypisanie kolumny z numerem klastra do danych.

4. Rozstęp między kwartylowy Q_x i oznaczane są wartości odstające wg niespełnionej zasady $Q1-2.5 \bullet Q_x <= cena <= Q3+2.5 \bullet Q_x$, nie są one brane do dalszych wyliczeń.

5. Ponowne wykonanie algorytmu *MeanShift* oraz podział na klastry.

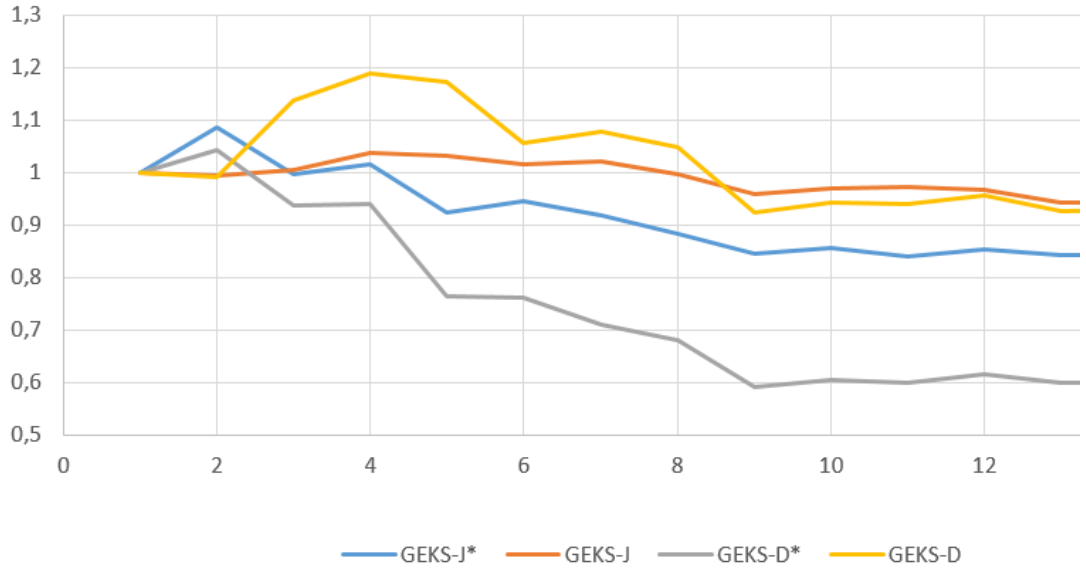
6. Wyliczenie indeksów cen *GEKS-J, GEKS-D* - reprezentant jako średnia arytmetyczna ceny dla klastra.

Telewizory zostały podzielone na 10 klastrów *MeanShift* (bandwith=5)



Wyniki eksperymentu 2 - TV (przekątna, matryca)

GEKS-J vs GEKS-D (* - bez usunięcia)



GEKS-J po odrzuceniu wartości odstających(seria czerwona)

GEKS-D (seria żółta)

Międzykwartylowy rozstęp Q_x

Q_1-2.5•Q_x< cena <Q_3+2.5•Q_x

Podsumowanie

W celu wyliczenia wskaźnika cen dla grupy homogenicznych produktów dla danych scrapowanych potrzebne jest:

- pozyskanie dużej ilości danych z wielu miesięcy (bez obciążenia ankieterów statystycznych- zmniejszenie kosztu badań statystycznych),
- walidacja pozyskanego zbioru oraz agregacje danych, analiza i wybór cech, podział na grupy:
  - „manualnie” poprzez wyrażenia regularne i słowniki,
  - „automatycznie” przy użyciu algorytmów grupujących.

Grupa homogeniczna (klaster) traktowana jako reprezentant pozwala na:

- obliczenie wskaźnika cen towarów w przypadku utraty ciągłości notowań dla pojedynczego produktu,
- wykorzystanie większej ilości danych (poszczególne produkty nie muszą być dopasowywane w czasie), nawet jeśli produkt pojawi się tylko w jednym okresie, to nadal będzie uwzględniony w indeksie.

Literatura

- Guide on Multilateral Methods in the Harmonised Index of Consumer Prices (2022) Manuals and Guidelines, Eurostat
- Metcalfe E., Flower T., Lewis T., Mayhew M., Rowland E. (2016): *Research indices using web scraped price data: clustering large datasets into price indices (CLIP)* Office for National Statistic of UK https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip

  Mayhew M. (2017): *ONS methodology working paper series number 12 – a comparison of index number methodology used on UK web scraped price data* Office for National Statistics https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber12acomparisonofindexnumbermethodologyusedonukwebscrapedpricedata

  Ayoubkhani D., Thomas H. (2022): Estimating Weights for Web-Scraped Data in Consumer Price Indices Journal of Official Statistics, Vol. 38, No. 1, 2022, pp. 5–21, http://dx.doi.org/10.2478/JOS-2022-0002
- Białek J., Kłopotek M., Panek T. (2022): *Nowoczesne technologie i nowe źródła danych w pomiarze inflacji*. GUS, Biblioteka Wiadomości Statystycznych
- https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip
- https://eitca.org/artificial-intelligence/eitc-ai-mlp-machine-learning-with-python/clustering-k-means-and-mean-shift/mean-shift-from-scratch/examination-review-mean-shift-from-scratch/what-are-the-basic-steps-involved-in-the-mean-shift-algorithm/
- Mean Shift: A Robust Approach toward Feature Space Analysis

Urząd Statystyczny
w Opolu

Dziękujemy za uwagę

t.markuszewski@stat.gov.pl
k.widera@stat.gov.pl