ROZDZIAŁ 1
Projekt „Budowa zintegrowanego systemu statystyki cen detalicznych – INSTATCENY" jako wsparcie statystyki publicznej w badaniu zjawisk inflacyjnych

1.1. Główne przesłanki podjęcia badań naukowych w ramach projektu

Rolą statystyki publicznej jest obiektywne i systematyczne informowanie społeczeństwa, organów państwa i administracji publicznej oraz podmiotów gospodarki narodowej o sytuacji ekonomicznej, demograficznej, społecznej oraz środowiska naturalnego. Badania prowadzone przez statystykę publiczną, realizowane z zachowaniem rygorów metodologicznych i transparentności, stanowią wiarygodne źródło informacji i podstawę kształtowania polityki publicznej. Dynamiczne zmiany relacji i procesów społeczno-gospodarczych, będące wynikiem szybko postępującej digitalizacji, stawiają statystykę publiczną przed koniecznością ciągłego udoskonalania i dostosowywania systemu gromadzenia, przetwarzania i analizy danych.

Ważną część badań statystycznych stanowią wskaźniki makroekonomiczne, wykorzystywane w analizach oraz bieżącym monitorowaniu zjawisk ekonomicznych. Jedną z najważniejszych informacji statystycznych jest wskaźnik cen konsumpcyjnych, uznawany za miarę inflacji. Dane statystyczne w tym zakresie publikowane przez Główny Urząd Statystyczny odgrywają kluczową rolę w realizowanej przez Narodowy Bank Polski strategii polityki pieniężnej – strategii bezpośredniego celu inflacyjnego.

Szczególne wyzwanie dla prowadzenia wiarygodnych i dokładnych badań cen konsumpcyjnych stanowią dynamiczne przeobrażenia rynku detalicznego i zachowań konsumentów dotyczące ich preferencji i konsumpcji. Digitalizacja rynku detalicznego oraz zwiększenie dostępu do elektronicznej informacji o cenach powodują, że konieczne jest nie tylko modernizowanie sposobów gromadzenia danych i metod zarządzania pozyskiwanymi informacjami, lecz także stałe doskonalenie metodyki

obserwacji zmian cen i obliczania wskaźników. W tym kontekście wyzwaniem staje się również kwestia rotacji i zmian jakości produktów. Jedną z istotnych przesłanek podjęcia prac w tym projekcie było zwiększenie zakresu danych (w tym informacji dodatkowych o produktach) i ich wykorzystania w badaniu w celu poprawy dokładności i wiarygodności obliczeń.

Tradycyjnie stosowane metody pozyskiwania danych o cenach bazują głównie na zbieraniu danych przez ankieterów statystycznych bezpośrednio w punktach sprzedaży i nie uwzględniają nowych źródeł danych, prowadząc do uzyskiwania coraz mniej reprezentatywnej próby, ograniczającej m.in. możliwości opisania zachodzących na rynku detalicznym procesów i prognozowania zjawisk inflacyjnych.

Modyfikowanie technik zbierania, przetwarzania i analizowania danych musi uwzględniać przede wszystkim konieczność przygotowywania informacji o cenach detalicznych i ich zmianach w wymaganych terminach oraz w stopniu gwarantującym wysoką jakość tych danych. Rozwiązania wypracowane w ramach projektu ukierunkowane były na wdrożenie do badań cen konsumpcyjnych nowatorskich technik automatycznego pozyskiwania dużej ilości danych, zaawansowanych metod ich standaryzacji i przetwarzania oraz na znaczną poprawę efektywności procesu analizy danych poprzez budowę platformy przetwarzania i analizy danych, z wykorzystaniem nowoczesnych narzędzi informatycznych. Przy budowie platformy poddano ocenie i ewentualnemu wykorzystaniu narzędzia z obszaru big data i data science. Założeniem projektu było również zwiększenie – poprzez automatyzację procesów produkcyjnych – efektywności kosztowej prowadzenia wymienionych badań oraz przyspieszenie analiz i przygotowywania nowego zakresu informacji statystycznych, m.in. w ujęciu terytorialnym, a także raportów dla gestorów danych (firm prywatnych, regulatorów rynku itp.) przy zachowaniu wysokich standardów jakościowych.

Intensywne przeobrażenia poszczególnych segmentów rynku detalicznego (m.in. pod wpływem cyfryzacji, rozwoju komunikacji wielokanałowej), jak również zmiany struktury demograficznej i społecznej ludności, a także zachowań zakupowych konsumentów determinują kierunki prac rozwojowych podejmowanych w obszarze statystyki cen detalicznych. Niezmiernie istotnym wyzwaniem dla statystyki cen jest konieczność adaptacji obecnie stosowanych praktyk i metod w zakresie obliczania wskaźników cen detalicznych do nowych warunków, jakie stwarza dostęp do alternatywnych źródeł danych oraz innowacyjnych technologii. Doświadczenie GUS, jak również innych krajowych urzędów statystycznych w zakresie prowadzenia badania cen detalicznych wskazuje, że w celu zwiększenia jego efektywności metodyka obliczania wskaźników cen konsumpcyjnych powinna zostać znacznie wzmocniona, szczególnie w zakresie pozyskania i wieloaspektowego wykorzystania nowych źródeł danych.

1.2. Cele i zadania projektu w kontekście wsparcia statystyki publicznej w zakresie gromadzenia i przetwarzania danych na potrzeby obliczania wskaźników cen detalicznych

Celem projektu „Budowa zintegrowanego systemu statystyki cen detalicznych – INSTATCENY" było unowocześnienie procesu pomiaru zmian cen detalicznych z wykorzystaniem nowych źródeł danych i narzędzi big data oraz jego wdrożenie w GUS.

Główne cele praktyczne projektu to:

- zwiększenie liczby i zróżnicowanie źródeł danych adekwatnie do specyfiki rynku detalicznego w Polsce poprzez włączenie do badania nowych źródeł i metod ich pozyskiwania;
- rozwinięcie metodyki integracji heterogenicznych danych w procesie produkcji wskaźników cen detalicznych;
- zmodernizowanie systemu statystyki cen detalicznych poprzez rozbudowę zakresu wykorzystywanych danych i usprawnienie sposobów ich pozyskiwania.

Wobec konieczności i dużej wagi procesu modernizacji standardów pozyskiwania i przetwarzania danych GUS podjął prace związane z negocjowaniem porozumień z gestorami danych transakcyjnych, którzy mogliby przekazywać dane o poziomach cen i podstawowych parametrach jakościowych produktów. Poszerzenie badania o nowe źródła danych poprawia jego jakość poprzez znaczne zwiększenie dotychczasowej liczebności próby towarów i usług, a także częstotliwości notowań ich cen, co pozytywnie wpływa na precyzję opracowywanych wskaźników cen. Jednocześnie ogranicza ryzyko wystąpienia błędów, które wynikają z ręcznego gromadzenia danych o poziomach cen przez ankieterów statystycznych w wybranych segmentach rynku detalicznego, poprzez automatyzację procesu pozyskania dużych zbiorów danych, ich analizowania i przetwarzania oraz wsparcie pracy ankietera zmodernizowanymi metodami rejestracji informacji.

Projekt obejmował zasadniczą zmianę podejścia do sposobu pozyskania i walidacji danych w celu zapewnienia precyzji i wiarygodności wyników badań zmian cen. Pozyskiwanie danych o cenach detalicznych towarów i usług z nowych źródeł to również korzyści związane ze wzrostem jakości danych i redukcją kosztów badań, ale jednocześnie – nowe wyzwania w zakresie metodycznym i organizacyjnym. Możliwość wykorzystania nowych źródeł danych wymagała bardzo wnikliwego i szczegółowego przetestowania. Wdrażanie kolejnych potencjalnych źródeł informacji o cenach detalicznych produktów musi być bowiem poprzedzone analizą ich wpływu na jakość badania, w tym na dokładność uzyskiwanych wyników. Istotnym czynnikiem pozostaje zachowanie proporcjonalności nakładów w relacji do możliwych do osiągnięcia korzyści (m.in. poprawa jakości danych).

W ramach projektu rozpoczęto systematyczne wdrażanie do praktyki badań następujących danych o cenach detalicznych:

- danych z systemów informatycznych sieci handlowych (dane skanowane);
- danych uzyskiwanych bezpośrednio od ich gestorów (firmy prywatne, regulatorzy danego segmentu rynku);
- danych o cenach detalicznych z internetu (m.in. web scraping);
- danych administracyjnych.

Równocześnie rozpoczęto prace w obszarze adaptacji do zastosowania nowatorskich narzędzi wspierających pracę ankieterów statystycznych rejestrujących dane bezpośrednio w placówkach handlowych.

Rezultatem tych prac jest rozpoczęcie automatyzacji procesu pozyskiwania i przetwarzania danych z zakresu badania cen detalicznych oraz doskonalenia ich analizy poprzez stworzenie innowacyjnego systemu informatycznego służącego do integracji danych o cenach detalicznych towarów i usług pozyskiwanych z różnych źródeł, stanowiącego narzędzie wsparcia procesu analizy danych z badania cen detalicznych przez statystyków GUS i generowania danych dla użytkowników zewnętrznych.

ROZDZIAŁ 2
Metody gromadzenia i przetwarzania danych na potrzeby badania cen detalicznych

2.1. Metodyka konstrukcji koszyka towarów i usług

Do obliczenia wskaźników cen towarów i usług konsumpcyjnych niezbędne są dane na temat:

- poziomu cen produktów, które notowane są przez ankieterów urzędów statystycznych bezpośrednio w punktach sprzedaży detalicznej, zbierane przez internet lub pozyskane od gestorów danych z ich systemów informatycznych;
- wydatków gospodarstw domowych na zakup towarów i usług konsumpcyjnych, które są zbierane podczas badania budżetów gospodarstw domowych i stanowią podstawę do budowy systemów wag.

Statystyka publiczna równolegle oblicza dwa wskaźniki cen konsumpcyjnych: wskaźnik cen towarów i usług konsumpcyjnych (ang. Consumer Price Index – CPI) na potrzeby krajowe oraz zharmonizowany wskaźnik cen konsumpcyjnych (ang. Harmonised Index of Consumer Prices – HICP) na potrzeby porównań międzynarodowych, głównie w ramach Unii Europejskiej. Zasady obliczania CPI nie mają, tak jak w przypadku HICP, podstawy prawnej. Metodyka obliczeń tworzona jest na podstawie wytycznych i rekomendacji międzynarodowych oraz potrzeb użytkowników krajowych. Zasady badania cen konsumpcyjnych doskonalone są w ścisłej współpracy z biurem statystycznym Unii Europejskiej – Eurostatem w ramach opracowywania HICP. Wytyczne międzynarodowe w pewnych obszarach wymagają dostosowania do uwarunkowań krajowych. Przykładowo konieczne jest uwzględnianie czynników, które wpływają na kształt wykorzystywanej w badaniu próby, m.in. zmian zachowań konsumentów i struktury sprzedaży detalicznej oraz dostępności nowych źródeł danych, konieczności redukcji obciążeń ankieterów, zaleceń Eurostatu czy nowelizacji aktów prawnych.

Wskaźniki cen konsumpcyjnych CPI za dany miesiąc publikowane są dwukrotnie – pod koniec miesiąca jako szybki szacunek CPI FE (ang. *CPI flash estimate*), a następnie w połowie następnego miesiąca jako wskaźnik ostateczny. Wskaźniki obliczane są w ujęciu miesięcznym i rocznym z podziałem na szczegółowe kategorie konsumpcyjne.

Istnieje szereg założeń koncepcyjnych przyjętych w zdefiniowaniu zakresu CPI i jego konstrukcji – zarówno uwzględnianych cen, jak i wag. Decyzje dotyczące tego zakresu powinny wynikać z podstawowego zastosowania wskaźników. Praktyka wskazuje jednak, że obliczany i publikowany wskaźnik zaspokaja różne potrzeby.

Ceny konsumpcyjne, będące przedmiotem badania, oznaczają ceny zakupu zapłacone przez gospodarstwa domowe w celu nabycia poszczególnych produktów w drodze transakcji pieniężnych, a *produkty* oznaczają towary i usługi w rozumieniu statystyki Rachunków Narodowych. Podstawą badania zmian cen konsumpcyjnych są obserwacje ich poziomów dla wytypowanego zestawu towarów i usług. Obserwacje te są prowadzone co miesiąc i polegają na zebraniu danych o cenach dokładnie tych porównywalnych towarów lub usług, które zostały wybrane do badania i zamieszczone na liście reprezentantów. Pomiar zmian cen nie odzwierciedla ruchów cen związanych ze zmianą atrybutów towarów takich jak jakość i użyteczność, a jedynie te, które wynikają z czystej zmiany ceny. Informacje gromadzone o cenach dotyczą jedynie tych towarów i usług, które są w sprzedaży na terenie kraju. Przykładowo ceny produktów sprzedawanych przez zagraniczne sklepy internetowe, w których konsument dokonuje zapłaty obcą walutą, nie są uwzględniane w badaniu cen konsumpcyjnych. Metody pozyskiwania informacji o cenach są dostosowane do rodzaju produktów będących przedmiotem badania, jak również do zwyczajów i upodobań konsumentów. Wybór sposobu gromadzenia danych o cenach nie może wpływać na wskaźnik, powinien natomiast zapewniać takie możliwości obserwacji zmian cen, aby obliczony na ich podstawie wskaźnik odzwierciedlał rzeczywiste tendencje ich zmian.

Podstawowe założenia koncepcyjne budowy systemu wag do obliczania CPI, na podstawie zaleceń międzynarodowych, są następujące:

Założenia	Opis
Zakres geograficzny	Koncepcja narodowa
Populacja	Prywatne gospodarstwa domowe
Zakres przedmiotowy	Wydatki konsumpcyjne z wyłączeniem: • związanych z biznesem; • przeznaczanych na aktywa, takie jak dzieła sztuki, inwestycje finansowe (w odróżnieniu od usług finansowych), płatności składek na ubezpieczenie społeczne, grzywny lub podatki dochodowe, płatności odsetek lub spłat długów. Transakcje pieniężne realizowane: • w gotówce; • czekiem; • kartą kredytową; • jako inne zobowiązanie finansowe do zapłaty w zamian za nabycie towaru lub usługi. Z wyłączeniem transakcji niepieniężnych w zakresie: • wydatków na żywność i usług wyprodukowanych na własny rachunek; • wynagrodzeń rzeczowych; • żywności i usług dostarczanych bezpłatnie lub dotowanych przez rząd i instytucje non profit.
Koncepcja	Wagi plutokratyczne
Struktura wag	Wagi regionalne – w każdym województwie wytypowane rejony badania cen (liczba zależy od wielkości województwa). Punkty sprzedaży odpowiadają najpowszechniejszym lokalom handlowym i odzwierciedlają różnorodność sieci dystrybucji w rejonie badania cen. Obecnie przyjęte założenia w zakresie konwencjonalnych źródeł danych nie przewidują stosowania wag według rodzaju sklepów. Ceny towarów i usług konsumpcyjnych zbierane są w ok. 170 punktach sprzedaży detalicznej na badanym rejonie. W 2008 r. włączony e-commerce. Udziały zakupów internetowych oparte na danych z badania budżetów gospodarstw domowych dotyczących zakupów internetowych.

Źródło: opracowanie własne.

CPI jest wynikiem agregacji ok. 350 tys. zebranych w ciągu badanego miesiąca cen detalicznych (CPI FE z założenia jest obliczany na podstawie mniej kompletnych danych). Zebrane dane są podstawą obliczenia dynamik cen dla ok. 340 grup elementarnych. Grupy te są następnie agregowane zgodnie z hierarchiczną klasyfikacją stosowaną w badaniu aż do wskaźnika cen ogółem.

Metodyka obliczania CPI w Polsce oparta jest na formule Laspeyresa. Jako system wag przyjmuje się strukturę faktycznie ponoszonych wydatków przez gospodarstwa domowe w okresie jednego roku (tak aby ująć wszystkie rodzaje wydatków, także ponoszonych sezonowo). Wybór takiej formuły jest uzasadniony łatwością obliczeń i komunikacji. Poza wskaźnikami obliczanymi dla kategorii zgodnych z odpowiednią

klasyfikacją wskaźniki obliczane są także dla kilkudziesięciu specjalnych agregatów, zgodnie z zapotrzebowaniem użytkowników danych.

Podział towarów i usług stosowany w przypadku CPI wykorzystuje Klasyfikację Spożycia Indywidualnego według Celu – COICOP (ang. Classification of Individual Consumption by Purpose). Jest to klasyfikacja opracowana i zalecana przez ONZ. COICOP klasyfikuje towary i usługi nabywane przez gospodarstwa domowe według celu przeznaczenia.

Wdrożona do obliczeń CPI w Polsce w 2014 r. szczegółowa Europejska Klasyfikacja Spożycia Indywidualnego według Celu – ECOICOP (ang. European Classification of Individual Consumption According to Purpose) stanowi rozszerzenie o podklasy klasyfikacji ONZ i składa się z pięciu poziomów:

- dwucyfrowego – 12 działów;
- trzycyfrowego – 44 grup;
- czterocyfrowego –110 klas;
- pięciocyfrowego – 296 podklas;
- sześciocyfrowego – ok. 340 grup elementarnych (poziom krajowy).

Główny podział wydatków w tej klasyfikacji przebiega według 12 działów zgodnie z głównym celem spożycia indywidualnego (zaspokojenie określonych potrzeb):

01 – Żywność i napoje bezalkoholowe;
02 – Napoje alkoholowe i wyroby tytoniowe;
03 – Odzież i obuwie;
04 – Użytkowanie mieszkania lub domu i nośniki energii;
05 – Wyposażenie mieszkania i prowadzenie gospodarstwa domowego;
06 – Zdrowie;
07 – Transport;
08 – Łączność;
09 – Rekreacja i kultura;
10 – Edukacja;
11 – Restauracje i hotele;
12 – Inne towary i usługi.

Klasyfikacja ma strukturę hierarchiczną składającą się z kilku poziomów, w ramach których wzrasta stopień szczegółowości. Na najniższym poziomie agregacji wyróżnia się grupy elementarne, reprezentujące wąskie grupy wydatków konsumpcyjnych gospodarstw domowych. Wyodrębnione kategorie charakteryzują się bardzo różnym stopniem jednorodności. Od tego zależy m.in. liczba wydzielonych grup elementarnych. Przykładowo dział „Żywność i napoje bezalkoholowe" dzieli się na ponad 85 grup elementarnych, np. ryż, pieczywo, twarogi, masło, jabłka, pomidory, kawa, herbata, a dział „Restauracje i hotele" obejmuje 7 jednorodnych grup, np. restauracje, kawiarnie, stołówki. Grupy elementarne są najniższym poziomem w systemie wag służącym do obliczania wskaźników cen w wyższych agregacjach.

ECOICOP, przeznaczona do badania cen detalicznych na potrzeby HICP, mimo dokonywanych rewizji wydaje się nie zaspokajać różnorodnych potrzeb informacyjnych w zakresie kształtowania się cen konsumpcyjnych. Zastosowane w badaniu formuły i zasady obliczeń dają jednak możliwość tworzenia dodatkowych (specjalnych) agregatów, niezwykle istotnych tak ze względów praktycznych (np. możliwość deflowania innych statystyk), jak i analitycznych (np. ocena i prognozowanie zjawisk inflacyjnych).

Poza koszykiem inflacyjnym dla przeciętnego gospodarstwa domowego GUS opracowuje koszyki dla siedmiu grup społeczno-ekonomicznych. Wyznaczane są one według kryterium głównego źródła utrzymania. Badaniem objęte są gospodarstwa domowe: pracowników, rolników, pracujących na własny rachunek, emerytów i rencistów, a także osób utrzymujących się ze źródeł niezarobkowych.

2.2. Konwencjonalne metody gromadzenia danych o cenach detalicznych

Jak wspomniano, CPI jest wynikiem agregacji ok. 350 tys. cen w badanym miesiącu (bez uwzględnienia danych skanowanych). Dane te stanowią podstawę obliczenia zmian cen dla ok. 340 grup elementarnych, a następnie wyższych poziomów agregacji aż do wskaźnika ogółem CPI. Liczba grup elementarnych w ramach poszczególnych działów COICOP oraz przeciętna liczba cen zbieranych w ciągu miesiąca w 2021 r. były następujące:

Dział COICOP	Liczba grup elementarnych	Przeciętna liczba cen w miesiącu[a]	Przykłady towarów i usług
Żywność i napoje bezalkoholowe	86	71 000	Różne rodzaje pieczywa, mięsa, mleko, sery, jaja, owoce, warzywa, wyroby cukiernicze, kawa, herbata, woda, soki
Napoje alkoholowe i wyroby tytoniowe	13	6600	Różne rodzaje napojów spirytusowych, wina, piwa, papierosów
Odzież i obuwie	18	20 200	Materiały odzieżowe, odzież damska, męska i dziecięca, wyroby pasmanteryjne, usługi odzieżowe, obuwie i usługi obuwnicze
Użytkowanie mieszkania lub domu i nośniki energii	25	18 200	Opłaty na rzecz właścicieli, usługi związane z konserwacją i naprawami, utrzymanie porządku, monitoring, konserwacja wind, zaopatrywanie w wodę, wywóz śmieci, usługi kanalizacyjne, nośniki energii
Zdrowie	15	16 400	Wyroby farmaceutyczne (w tym refundowane), sprzęt terapeutyczny, usługi lekarskie, stomatologiczne, sanatoryjne

a Bez danych skanowanych.

(dok.)

Dział COICOP	Liczba grup elementarnych	Przeciętna liczba cen w miesiącu	Przykłady towarów i usług
Transport	28	131 800	Samochody osobowe nowe i używane, części zamienne, paliwa do prywatnych środków transportu, opłaty za parkowanie, transport powietrzny, kolejowy, drogowy, wodny
Łączność	11	1100	Usługi pocztowe, sprzęt telekomunikacyjny, usługi telefonii stacjonarnej i komórkowej, usługi internetowe, usługi w pakiecie
Rekreacja i kultura	53	20 000	Sprzęt audiowizualny, fotograficzny i informatyczny, gry, zabawki, sprzęt sportowy, artykuły ogrodnicze, artykuły dla zwierząt domowych, opłaty za telewizję, książki, gazety, czasopisma, artykuły papiernicze, usługi turystyki zorganizowanej
Edukacja	11	2300	Opłaty za pobyt dziecka w przedszkolu, opłaty związane z nauką na różnych poziomach nauczania, kursy dla osób uczących się w systemie pozaszkolnym
Restauracje i hotele	7	9000	Usługi gastronomiczne świadczone przez restauracje, kawiarnie, bary, stołówki, usługi zakwaterowania świadczone np. przez hotele i schroniska, zakwaterowanie dla studentów
Inne towary i usługi	31	19 300	Usługi fryzjerskie i kosmetyczne, środki kosmetyczne i higieniczne, biżuteria, artykuły podróżne, opłaty za pobyt dziecka w żłobku, ubezpieczenia, usługi finansowe, prawnicze, administracyjne, usługi pogrzebowe i cmentarne

a Bez danych skanowanych.
Źródło: opracowanie własne.

W ciągu ostatnich kilku lat zintensyfikowano prace nad dywersyfikacją źródeł danych o cenach.

Ceny większości produktów gromadzone są w okresach miesięcznych. Obowiązują różne wytyczne, ze względu na specyfikę procesu nabycia w zakresie momentu włączenia do obliczeń cen towarów i usług. Zgodnie z ramami prawnymi ceny towarów powinny być włączone do obliczeń w miesiącu, w którym są obserwowane, a usług – w miesiącu, w którym następuje rozpoczęcie konsumpcji danej usługi po obserwowanej cenie. W przypadku usług występuje często duża różnica w czasie pomiędzy okresem nabycia, płatności, dostawy i konsumpcji. Z praktycznego punktu widzenia obserwacja cen usług jest znacznie bardziej pracochłonna. Przykładowo

w zakresie usług transportowych notowanie cen w wytypowanym terminie odbywa się z odpowiednim wyprzedzeniem odpowiadającym zachowaniom zakupowym gospodarstw domowych. Uwzględnienie rzeczywistej ceny transakcyjnej danej usługi jest zatem w pewnym stopniu uzależnione od dostępnych dodatkowych informacji na temat zachowań rynkowych konsumentów.

W obliczeniach wskaźników cen konsumpcyjnych wykorzystuje się coraz więcej danych jednostkowych, pozyskiwanych z wielu różnorodnych źródeł. W przypadku niektórych grup towarów nie wystarczy pojedynczy pomiar ceny w ustalonym okresie badania (5.–22. dzień miesiąca). Są produkty charakteryzujące się bardziej dynamicznymi zmianami cen (np. warzywa i owoce), których ceny notowane są dwa razy w miesiącu (między 5.–13. a 14.–22. dniem miesiąca). Uwzględniana jest przy tym różnorodność regionalnych rynków. Coraz więcej usług nabywanych jest z wyprzedzeniem, wówczas ceny takich usług obliczane są jako średnie z kilku obserwacji poprzedzających termin ich realizacji (np. dotyczy to usług transportowych, takich jak loty czy przejazdy koleją). W uzasadnionych przypadkach notowania prowadzone są rzadziej niż raz w miesiącu. Dotyczy to np. cen leków refundowanych, których wykaz aktualizowany jest przez Ministerstwo Zdrowia raz na kilka miesięcy..

Źródłem informacji o cenach detalicznych towarów i usług są:

- notowania cen dokonywane przez ankieterów;
- cenniki, zarządzenia i decyzje w zakresie cen jednolitych obowiązujących na terenie całego kraju lub jego części, wydawane przez organy administracji rządowej i organy jednostek samorządu terytorialnego, jak również przez podmioty prowadzące działalność gospodarczą;
- notowania cen towarów i usług kupowanych przez internet;
- bazy danych sieci handlowych i zakładów ubezpieczeń.

W zakresie notowań cen gromadzonych przez ankieterów od 2019 r. badanie cen prowadzone jest w 207 rejonach badania na terenie kraju. Rejonem badania jest miasto lub – w przypadku dużych miast – jego część. Na przykład Warszawa jest podzielona na sześć rejonów, a Łódź – na dwa. Liczba miast, w których są prowadzone notowania cen, wynosi 194. Dane gromadzone są przez sieć ankieterów zatrudnionych w urzędach statystycznych w 16 województwach. Ankieterzy gromadzący ceny prowadzą również inne badania niezwiązane z cenami, m.in. badania warunków życia i BAEL. Notowania cen w rejonach badania realizowane są w punktach sprzedaży detalicznej (zarówno dużych, jak supermarkety, hipermarkety i domy towarowe, jak i w małych sklepikach i na straganach), placówkach gastronomicznych oraz jednostkach świadczących usługi.

Centralne (lub quasi-centralne, z udziałem urzędów wojewódzkich) gromadzenie danych o cenach obejmuje: jednolite ceny detaliczne w kraju, ceny wybranych towarów i usług kupowanych przez internet, samochody używane, samochody nowe,

usługi finansowe świadczone przez banki, opłaty za energię elektryczną i gaz. W ramach badania jednolitych cen detalicznych w kraju obserwowane są m.in. takie kategorie, jak: leki (z refundacją) oraz usługi sanatoryjne, transportowe, pocztowe, telekomunikacyjne, a także opłaty radiowe i telewizyjne, gazety, czasopisma, opłaty administracyjne i usługi prawne.

W badaniu cen detalicznych towarów i usług kupowanych przez internet notowane są ceny artykułów, które odzwierciedlają najpopularniejsze grupy produktów nabywane w sklepach internetowych. Produkty te należą m.in. do następujących działów i grup COICOP: żywność i napoje bezalkoholowe, odzież, obuwie męskie, meble, elementy wyposażenia mieszkania, farmaceutyki, urządzenia i sprzęt terapeutyczny, części zamienne i akcesoria do prywatnych środków transportu, sprzęt telekomunikacyjny, sprzęt do odbioru, nagrywania i odtwarzania dźwięku i obrazu, sprzęt do przetwarzania informacji, sprzęt do rekreacji, gry, zabawki, artykuły dla zwierząt domowych, usługi fotograficzne, książki, wakacje zorganizowane, usługi zakwaterowania, inne urządzenia, towary i produkty higieny osobistej, zegarki i zegary.

W obliczeniach wskaźników cen od kilku lat wykorzystuje się także informacje uzyskane bezpośrednio od gestorów danych, m.in. firm prywatnych oraz regulatorów danego segmentu rynku (sektor ubezpieczeń i paliwa do prywatnych środków transportu). Przykładowo od 2018 r. wskaźnik cen dla grupy „Ubezpieczenie pojazdów silnikowych" jest obliczany na podstawie danych o dokonanych transakcjach, przesyłanych przez towarzystwa ubezpieczeniowe. Dane te zastąpiły notowania prowadzone przez ankieterów, które ze względu na specyfikę rynku stały się niemożliwe do pozyskania. Od 2018 r. dane otrzymane od gestora służą także udoskonalaniu sposobu ważenia poszczególnych reprezentantów w ramach grupy elementarnej „Usługi telefonii komórkowej" poprzez nadanie wag wewnętrznych opłatom post-paid i pre-paid.

Jednym z kluczowych kryteriów jakości badania cen konsumpcyjnych jest zachowanie reprezentatywności próby, czyli dostosowanie badania do zmieniających się uwarunkowań zewnętrznych i wewnętrznych. Działaniem, które ma zapewnić odpowiednią reprezentatywność zmiennych podlegających badaniu, jest coroczna weryfikacja próby.

Proces weryfikacji listy reprezentantów jest wieloetapowy i wielowymiarowy. Weryfikacja próby produktów odbywa się m.in. na podstawie wyników obserwacji ankieterów, danych z innych źródeł, w tym danych administracyjnych, różnych analiz i ekspertyz, a także zapisów w książeczkach budżetowych i notatek respondentów uczestniczących w badaniu budżetów gospodarstw domowych. W wyniku weryfikacji z reguły kilka procent próby ulega zmianom polegającym na:

- aktualizacji opisów;

- usuwaniu pozycji, które straciły cechy reprezentatywności;
- dodaniu nowych reprezentantów osiągających znaczący udział w wydatkach konsumpcyjnych gospodarstw domowych.

Specyficzną kategorią zmian są zmiany spowodowane czynnikami innymi niż bezpośrednio wynikające z uwarunkowań rynkowych czy preferencji konsumentów, np.:

- zmiana klasyfikacji stosowanej w badaniu (konieczność zapewnienia reprezentatywności dla innego zakresu grup produktów);
- konieczność ograniczenia próby ze względu na skrócenie okresu notowania;
- dostosowywanie badania do nowych standardów międzynarodowych (np. w zakresie jakości próby);
- potrzeby publikacyjne.

Lista reprezentantów obejmuje informacje niezbędne do pracy ankieterów w celu identyfikacji produktów objętych badaniem w roku sprawozdawczym oraz do pozyskania informacji o cenach z innych źródeł. Poza nazwą towaru lub usługi zawiera opis podstawowych cech jakościowych produktów, np. gramaturę, wymagany skład surowcowy, markę, typ, model czy producenta. Wytypowane do badania cen towary i usługi są przyporządkowywane do najniższego poziomu COICOP, tak aby zapewnić reprezentatywność dla danej grupy elementarnej. Liczba wybranych reprezentantów nie jest taka sama w każdej grupie elementarnej i zależy od stopnia jej zróżnicowania oraz wielkości wydatków na towary lub usługi, które są ponoszone przez gospodarstwa domowe.

Lista reprezentantów zawiera produkty:

- opisane bardzo dokładnie, z podaniem konkretnych parametrów ściśle określających ich cechy;
- stanowiące wąskie grupy asortymentowe, reprezentujące, np. ubiory, bieliznę, obuwie itp.

Reprezentantami objętymi badaniem cen konsumpcyjnych są towary nabywane często, np. pieczywo, mięso, mleko, sery, warzywa, owoce, gazety, środki do mycia, prania i czyszczenia, ale także – dobra trwałego użytkowania, kupowane rzadziej, w tym meble, sprzęt gospodarstwa domowego, sprzęt radiowy, telewizyjny czy komputerowy. Badanie cen uwzględnia również opłaty regularnie ponoszone przez gospodarstwa domowe, np. związane z użytkowaniem mieszkania, korzystaniem z nośników energii, komunikacji miejskiej, a także obejmuje usługi kupowane rzadziej, np. turystyczne, w zakresie konserwacji mieszkania i inne. Lista reprezentantów do badania cen konsumpcyjnych nie obejmuje (i nigdy nie obejmowała) produktów, które nie są kupowane przez gospodarstwa do celów bieżącego spożycia, czyli dóbr niekonsumpcyjnych, np. lokomotyw, szyn, tankowców. Są one ujęte w innych badaniach statystycznych, jak badanie cen producentów.

Liczba danych jednostkowych na potrzeby obliczeń wskaźników cen konsumpcyjnych wzrasta z roku na rok. Liczba notowanych cen wzrasta zarówno ze względu na zwiększającą się różnorodność oferty towarów i usług na rynku, jak i z uwagi na większe techniczne możliwości pozyskiwania informacji o cenach, m.in. przez internet oraz otrzymywanie danych bezpośrednio z sieci handlowych (dane skanowane).

Notowania dokonywane na potrzeby obliczania wskaźników cen to nie tylko gromadzenie cen produktów. Oprócz samej ceny zbierane są także obszerne dodatkowe informacje charakteryzujące wytypowane towary i usługi i wpływające na ich cenę (np. gramatura, skład surowcowy, model, zakres usługi), a także informacje o naliczonych podatkach.

Powyższe sposoby zbierania danych o zmianach cen mają swoje zalety, ale też ograniczenia. Decyzję o wyborze właściwej metody podejmuje GUS we współpracy z urzędami statystycznymi w poszczególnych województwach. Prowadzone są jednak intensywne prace o charakterze badawczo-rozwojowym, np. dotyczące analizy możliwości zastosowania nowej metodyki dla alternatywnych formuł indeksów cen, które w połączeniu z aktualizacją wytycznych międzynarodowych mogą przełożyć się na zmiany w sposobie prowadzenia badań cen w GUS.

2.3. Konwencjonalne źródła danych o konsumpcji gospodarstw domowych

Wskaźnik cen konsumpcyjnych mierzy zmiany cen reprezentatywnego koszyka towarów i usług. Obliczenia wskaźnika prowadzone są w kilku etapach. Pierwszy polega na obliczeniu zmiany cen dla danego reprezentanta w konkretnym rejonie notowań cen. Kolejnym krokiem jest wyznaczanie ogólnopolskich wskaźników cen dla poszczególnych reprezentantów. Ogólnopolskie wskaźniki cen wszystkich reprezentantów objętych notowaniami obliczane są jako średnie geometryczne wskaźników cen reprezentantów ze wszystkich rejonów. Następnie, stosując średnią geometryczną, opracowuje się wskaźniki cen dla grupy elementarnej na najniższym poziomie agregacji systemu wag (w przypadku wskaźnika cen towarów i usług konsumpcyjnych jest to ok. 340 grup elementarnych). Na ostatnim etapie wskaźniki cen grup towarów i usług wykorzystuje się (przy zastosowaniu systemu wag) do obliczenia wskaźników wyższych poziomów agregacji aż do uzyskania wskaźnika cen towarów i usług konsumpcyjnych ogółem. O znaczeniu zmian cen produktów zaliczonych do określonej klasyfikacją grupy elementarnej w kształtowaniu wskaźnika cen decyduje zatem waga danej grupy, czyli jej udział w wydatkach ponoszonych przez konsumentów. Struktura wydatków konsumpcyjnych zmienia się z roku na rok, jednak od lat dominujący udział stanowią wydatki na żywność i napoje bezalkoholowe oraz użytkowanie mieszkania lub domu i nośniki energii. Strukturę procentową systemu wag w poszczególnych działach COICOP w latach 2018–2022 przedstawiono w tabl. 2.1.

Tabl. 2.1. System wag stosowany w obliczeniach wskaźników cen towarów i usług konsumpcyjnych w latach 2018–2022

Wyszczególnienie	2018	2019	2020	2021	2022
Żywność i napoje bezalkoholowe	24,36	24,89	25,24	27,77	26,59
Napoje alkoholowe i wyroby tytoniowe	6,19	6,37	6,25	6,91	6,32
Odzież i obuwie	5,37	4,94	4,94	4,21	4,47
Użytkowanie mieszkania lub domu i nośniki energii	20,35	19,17	18,44	19,14	19,33
Wyposażenie mieszkania i prowadzenie gospodarstwa domowego	5,25	5,70	5,80	5,83	5,71
Zdrowie	5,69	5,12	5,29	5,39	5,69
Transport	8,74	10,34	9,89	8,88	9,54
Łączność	4,87	4,18	4,54	5,00	4,90
Rekreacja i kultura	6,92	6,44	6,62	5,78	6,07
Edukacja	1,00	1,07	1,15	1,02	1,16
Restauracje i hotele	5,71	6,20	6,12	4,56	4,77
Inne towary i usługi	5,55	5,58	5,72	5,51	5,45

Źródło: opracowanie własne.

Wagi dla CPI w Polsce, podobnie jak w większości krajów, obliczane są na podstawie badania budżetów gospodarstw domowych. Do obliczeń wskaźników cen konsumpcyjnych (zarówno w przekroju ogółem, jak i dla grup społeczno-ekonomicznych) stosowane są dane dotyczące rozchodów (wydatków) pieniężnych gospodarstw domowych pochodzące z badania budżetów gospodarstw domowych. Badanie dostarcza informacji o przeciętnych wydatkach konsumpcyjnych według COICOP. W przypadku gdy badanie budżetów gospodarstw domowych jest podstawowym źródłem danych do wyprowadzania wag, z uwagi na możliwość niedoszacowania lub zawyżania pewnych kategorii wydatków, zaleca się ich porównanie z danymi z innych źródeł. Najczęściej korekty dokonywane są w przypadku wydatków na alkohol, wyroby tytoniowe i usługi gastronomiczne. Klasyfikacja ma charakter uniwersalny i dlatego nie wszystkie rodzaje wydatków w niej ujęte znajdują odzwierciedlenie badaniu budżetów gospodarstw domowych w każdym kraju. W Polsce odnotowuje się zwykle kilkanaście kategorii COICOP, dla których w ciągu roku nie zostały zarejestrowane wydatki lub odnotowano roczne wydatki na poziomie kilkunastu groszy. Niektóre rodzaje wydatków są wyłączone z obliczeń z uwagi na praktyczne ograniczenia pomiaru cen (np. w zakresie narkotyków).

W obliczeniach CPI wagi aktualizowane są co roku. Wykorzystywane w tym celu źródło danych umożliwia aktualizację wag na najniższym poziomie klasyfikacji stosowanej w badaniu, zatem w przypadku CPI nie stosuje się praktyki częściowej aktu-

alizacji wag. Jest to szczególnie istotne ze względu na to, że wagi dla najniższych poziomów agregacji są narażone na szybszą dezaktualizację. Ogólną zasadą opisaną w rekomendacjach organizacji międzynarodowych jest minimalizacja opóźnienia we wdrażaniu nowych wag, tak aby wagi stosowane w CPI były jak najbardziej aktualne. W Polsce opóźnienie to zostało zniwelowane poprzez stosowanie w roku t wstępnych danych o wydatkach pochodzących z badania budżetów gospodarstw domowych z roku $t-1$.

Dostępne dane o wydatkach gospodarstw domowych pozwalają określić udział grupy produktów, a nie poszczególnych produktów, w wydatkach ogółem. Obserwuje się jednak, że coraz więcej grup produktów jest zróżnicowanych pod względem obserwowanych trendów cenowych. Dlatego też, na podstawie dodatkowych danych, konstruowane są wagi wewnętrzne. Wagi wewnętrzne stosowane są np. w przypadku zakupów internetowych. Informacje takie pochodzą również z dodatkowego pytania zawartego w badaniu budżetów gospodarstw domowych. Zmiany cen produktów nabywanych online wpływają na wskaźnik danej grupy elementarnej w stopniu odpowiadającym przypisanej wadze wewnętrznej. Takie wagi wewnętrzne (czyli udziały wydatków w tej grupie na towary nabywane online) są bardzo zróżnicowane – od poniżej 1% w przypadku towarów żywnościowych do kilkudziesięciu procent w przypadku towarów w zakresie rekreacji i kultury.

Dane o wydatkach konsumpcyjnych są podstawowym źródłem informacji do budowy systemu wag, służącego do obliczeń wskaźników cen konsumpcyjnych na wyższych poziomach agregacji. Bez tych informacji nie byłoby możliwe obliczenie wiarygodnego wskaźnika cen. Jakość zastosowanych wag jest istotnym atrybutem procesu obliczeń wskaźników cen i w dużej mierze decyduje o dokładności i wiarygodności wyników tych obliczeń. Niezbędne jest zastosowanie takich wag, które odzwierciedlają możliwie najbardziej aktualny wzorzec konsumpcji gospodarstw domowych. Wagi przypisane poszczególnym grupom elementarnym są bardzo zróżnicowane. Wśród ok. 340 badanych grup elementarnych są grupy, których udział w wydatkach gospodarstw domowych może wynosić zaledwie 0,01% (np. mięso cielęce czy owoce morza). Istnieją również takie grupy, na które gospodarstwa domowe mogą wydawać ok. 3–4% swoich dochodów (np. energia elektryczna, wyroby farmaceutyczne czy benzyna). Wagi przypisane do każdej grupy elementarnej klasyfikacji decydują także o tym, jaki wpływ na wskaźnik ogółem w badanym okresie ma zmiana cen obliczona dla tej grupy. Wielu użytkowników danych jest bowiem zainteresowanych informacjami, które grupy w największym stopniu wpłynęły na wskaźnik cen ogółem w danym okresie (co powoduje obniżenie, a co podwyższenie wskaźnika) czy też w których grupach zmiany cen nie podążają za ogólnymi trendami cenowymi. Z wykorzystaniem wag obliczane są też wskaźniki cen specjalnych agregatów (np. towary, towary nieżywnościowe trwałego użytku czy usługi).

2.4. Alternatywne źródła danych w pomiarze inflacji

Tradycyjny sposób zbierania danych przez ankieterów, poza ograniczeniami omówionymi w podrozdziałach 2.2 i 2.3, jest kosztochłonny i czasochłonny. Wraz z dynamicznym rozwojem branży e-commerce oraz coraz szerszym dostępem do danych transakcyjnych sieci handlowych, od ponad 20 lat obserwujemy dynamiczny wzrost zainteresowania urzędów statystycznych alternatywnymi źródłami danych w pomiarze inflacji. Należy tu wymienić przede wszystkim dane skrapowane, które – najogólniej rzecz biorąc – pochodzą z witryn internetowych, oraz dane skanowane, które są gromadzone i archiwizowane dzięki elektronicznym terminalom sieci handlowych. Poniżej przedstawiona została charakterystyka obu wymienionych źródeł danych z uwzględnieniem głównych problemów i wyzwań, jakie wiążą się z ich praktycznym wykorzystaniem w pomiarze inflacji.

2.4.1. Dane skrapowane

2.4.1.1. Definicja danych skrapowanych

Web scraping, nazywany też *zdrapywaniem* lub *skrobaniem stron*, to zautomatyzowany proces pozyskiwania danych ze stron internetowych, najczęściej przy wykorzystaniu specjalnego programu czy skryptu (zwanego *skraperem*) naśladującego zachowanie człowieka przeglądającego stronę internetową.

Wyodrębnianie danych jest stosowane w celu podejmowania wyważonych decyzji na podstawie ogromnej ilości dostępnych danych internetowych. Jednym z istotnych zastosowań tego procesu jest zbieranie cen. Dostarcza ono informacji o cenach i ofercie produktów, jak również pozwala na ocenę trendów rynkowych, a w efekcie umożliwia uzyskanie ogólnego obrazu rynku. Dzięki temu można np. przekonać się o trafności przyjętej strategii cenowej, co jest szczególnie przydatne dla marek i firm e-commerce. Innym interesującym zastosowaniem skrapowania jest *lead generation* – działanie polegające na pozyskiwaniu danych o klientach potencjalnie zainteresowanych oferowanymi produktami bądź usługami.

2.4.1.2. Metodyka skrapowania danych

Zasadniczo wyróżnia się dwa rodzaje web scrapingu. Pierwszy z nich określany jest mianem specyficznego (ang. *specific*), w którym zarówno struktura, jak i zawartość stron internetowych, które mają być pobrane, są z góry znane, a algorytm musi jedynie replikować zachowanie człowieka odwiedzającego stronę internetową i zbierającego interesujące go informacje. Typowym obszarem zastosowań tego typu metody jest zbieranie danych do wyznaczania wskaźników cen konsumpcyjnych. Przedmiotem zainteresowania jest tutaj zazwyczaj gromadzenie bardzo konkretnych informacji zawartych na stronie internetowej (np. wybrana wartość w wierszu). Drugi wa-

riant web scrapingu – ogólny (ang. *generic*) – zakłada, że użytkownik nie ma wcześniejszej wiedzy na temat treści i budowy strony internetowej, z której korzysta. Pobierana jest wtedy zawartość całej witryny, która następnie jest przetwarzana w celu wyodrębnienia potrzebnych informacji. Przykładem zastosowań ogólnego web scrapingu może być badanie dotyczące wykorzystania ICT w przedsiębiorstwach prowadzone we włoskim urzędzie statystycznym.

Koncepcyjnie wydobywanie danych ze stron internetowych jest prostym procesem, który wymaga współpracy dwóch robotów: robota indeksującego i skrobaka internetowego (webscrapera). Robot indeksujący prowadzi skrobaka przez internet, a ten wydobywa żądane dane (o pająkach internetowych – zob. dalej). Roboty zajmujące się zbieraniem danych, czyli skrobaki internetowe, to specjalistyczne narzędzia zaprojektowane do dokładnego i szybkiego wyodrębniania danych z konkretnej strony internetowej. Różnią się one pod względem konstrukcji i złożoności, w zależności od stawianych im wymagań. Ważną częścią każdego skrobaka są lokalizatory (selektory) danych, służące do wyszukiwania danych, które użytkownik chce wyodrębnić z pliku HTML (strony internetowej). Zazwyczaj korzysta się tu z XPath[1], czyli narzędzia umożliwiającego nawigację między elementami i atrybutami dokumentów XML, selektorów CSS[2], wzorców opisujących łańcuchy symboli (wyrażenia regularne – regex[3]) lub ich kombinacji. Narzędzie do scrapingu zazwyczaj wysyła żądania HTTP do docelowej witryny i wyodrębnia dane ze strony. Zwykle analizuje zawartość, która jest publicznie dostępna i widoczna dla użytkowników oraz renderowana przez serwer jako HTML. Czasami skrobak wysyła również żądania do wewnętrznych interfejsów programowania aplikacji (API) o niektóre powiązane dane – takie jak ceny produktów lub dane kontaktowe – które są przechowywane w bazie danych i dostarczane do przeglądarki za pośrednictwem żądań HTTP. Istnieją różne rodzaje narzędzi, których możliwości można dostosować do różnych projektów ekstrakcji. Przykładowo może być potrzebne narzędzie do scrapingu, które rozpoznaje unikalne struktury witryny HTML lub wyodrębnia, zmienia formatowanie i przechowuje dane z interfejsów API. Narzędzia do skrobania mogą być dużymi strukturami zaprojektowanymi do wszelkiego rodzaju typowych zadań związanych ze skrobaniem, ale można również używać bibliotek programistycznych ogólnego przeznaczenia i łączyć je w celu utworzenia skrobaka. Na przykład można użyć biblioteki żądań HTTP – takiej jak Python-Requests library[4] – i połączyć ją z pythonową biblioteką BeautifulSoup[5], aby pobrać dane z określonej strony. Można też użyć specjalnego frameworka, który łączy klienta http z biblioteką analizującą

[1] https://www.w3schools.com/xml/xpath_intro.asp.
[2] https://developer.mozilla.org/pl/docs/Web/CSS/CSS_Selectors.
[3] https://docs.microsoft.com/pl-pl/dotnet/api/system.text.regularexpressions.regex?view=net-6.0.
[4] https://realpython.com/python-requests.
[5] https://beautiful-soup-4.readthedocs.io/en/latest/.

HTML. Jednym z popularnych przykładów jest Scrapy[6], biblioteka typu open source stworzona z myślą o zaawansowanych potrzebach scrapingu.

2.4.1.3. Doświadczenia międzynarodowe dotyczące danych skrapowanych

Jednym z pierwszych projektów prowadzonych na szeroką skalę, który potwierdził, że dane pozyskiwane ze stron internetowych mogą być przydatne do obliczania wskaźników cen konsumenta, był *Billion Price Project* realizowany przez MIT. Przez kilka lat od 2007 r. dla wybranych krajów pobierane były dane ze sklepów internetowych, a następnie na ich podstawie wyznaczane wskaźniki cen towarów i usług konsumpcyjnych. W przypadku Brazylii, Chile i Kolumbii zebrane dane dość dobrze przybliżały trend obecny we wskaźniku inflacji publikowanym przez urzędy statystyczne tych krajów. Natomiast w przypadku Argentyny zebrane dane nie były spójne z oficjalnymi publikacjami na temat inflacji, nawet w dłuższym horyzoncie czasowym. W pracy Cavallo i in. (2008) podobieństwo wyników uzyskiwanych dzięki skrapowaniu cen i wskaźników opartych na cenach zebranych w sposób tradycyjny ustalono na poziomie 72%. W kolejnych latach projekt rozszerzono także na kolejne kraje.

Podobny projekt rozpoczęła firma Google. Jego celem była budowa wskaźnika cen na bazie danych pobieranych ze stron internetowych. Jednak wskaźnik ten nie został opublikowany oraz nie udostępniono szczegółowych informacji na temat samego projektu. Można podejrzewać, że w przypadku tego projektu problemem była nie tyle mała liczba obserwacji, ile metodyka konstrukcji wskaźnika i sposób wyboru stron, produktów oraz zakresu danych.

Obecnie wiele urzędów statystycznych na całym świecie prowadzi już badania nad możliwością wykorzystania danych pobieranych z sieci internetowej do obliczania oficjalnych wskaźników cen. Wstępne wyniki tych badań wskazują, że tego typu dane mogą przyczynić się do zmniejszenia obciążenia pracą respondentów i ankieterów, a także do usprawnienia i przyspieszenia obliczeń oraz obniżenia kosztów procesu gromadzenia danych i opracowywania wskaźników. Prowadzone są też prace nad wykorzystaniem tych danych do opracowywania nowych wskaźników gospodarczych. Doświadczenie w zakresie zbierania danych za pomocą metod web scrapingu mają m.in. urzędy statystyczne z Holandii, Wielkiej Brytanii, Niemiec, Norwegii, Austrii, Belgii oraz Włoch. Większość urzędów wprowadzała projekty stopniowo, poszerzając zarówno zakres przedmiotowy badanych cen, jak i liczbę witryn sklepowych, z których pobierane są informacje, co miało na celu zwiększenie reprezentatywności badania. Holenderski urząd statystyczny (Statistics Netherlands – CBS) zapoczątkował prace nad web scrapingiem od ściągania z witryn internetowych danych o cenach paliw oraz informacji o cenach przewozów lotniczych. Z kolei bry-

[6] https://docs.scrapy.org/en/latest/intro/tutorial.html.

tyjski urząd statystyczny (Office for National Statistics – ONS) rozpoczął pilotażowy program analizy witryn internetowych od trzech wybranych dużych sieci handlowych (Tesco, Sainsbury's oraz Waitrose & Partners) i stopniowo zwiększał zakres badanych cen z 33 kategorii produktów do pełnego składu koszyka CPI. Web scraping wykorzystuje również norweski urząd statystyczny (Statistics Norway – SSB). Dotychczas otrzymywał on dane z czterech głównych sieci handlowych dla 60 różnych kategorii dóbr konsumpcyjnych. Stosunkowo najbardziej zaawansowane prace związane z wykorzystaniem web scrapingu do pozyskiwania cen towarów i usług konsumpcyjnych realizuje obecnie niemiecki urząd statystyczny (Federal Statistical Office of Germany – Destatis). Część cen pozyskiwania tradycyjnymi metodami została już zastąpiona cenami zebranymi w internecie. W szczególności w ten sposób gromadzone są ceny wynajmu samochodów, części leków, biletów na autobusy dalekobieżne i kolejowych, a także ceny energii elektrycznej i gazu. GUS skrapuje ceny m.in. leków, a także, dzięki realizacji projektu INSTATCENY, ceny wybranych produktów z segmentu spożywczego. W tym ostatnim przypadku skrapowane ceny pobierane są z witryn internetowych wybranych sieci handlowych.

2.4.1.4. Zawartość danych skrapowanych

Zakres kategorii towarów i usług, których ceny gromadzone są przez urzędy statystyczne przy wykorzystaniu metod web scrapingu, jest dla większości krajów zbliżony. Zazwyczaj zbierane są ceny dotyczące tych kategorii koszyka CPI, które są w większości kupowane lub zamawiane za pośrednictwem internetu, czyli bilety lotnicze, kolejowe, usługi hotelowe, wycieczki zorganizowane, elektronika, gry, AGD oraz odzież i obuwie.

Zakres informacji o produkcie, które z reguły pobiera skraper, to poza ceną produktu jego opis (etykieta), tj. unikatowy kod produktu zastosowany przez sklep lub sieć handlową, a także, o ile jest dostępny na stronie internetowej, kod zewnętrzny produktu (EAN, GTIN, SKU itp.). Jeśli sieć handlowa umieszcza dodatkowe informacje o produkcie (np. o jego gramaturze, jednostce sprzedaży czy wielkości udzielonego rabatu), to również te informacje są skrapowane lub ewentualnie ekstraktowane z opisu produktu.

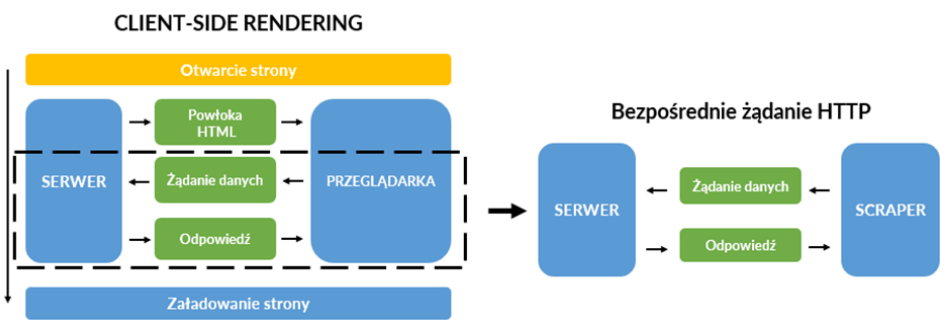
2.4.1.5. Zbieranie danych skrapowanych

Proces zbierania danych skrapowanych w dużej mierze zależy od wykorzystanej metody skrapowania, jednak możemy wyróżnić kilka najczęściej powtarzających się etapów. Pierwszym z nich jest pobranie wszystkich stron internetowych danego sklepu, które zawierają informacje o produktach. Dane o produktach są nieustrukturyzowane i osadzone w kodzie źródłowym strony, dlatego w kolejnym kroku analizowany jest uzyskany kod i wyodrębniane są informacje o cenach i charakterysty-

kach produktów. Istnieje również możliwość sprawdzenia poprawności przeprowadzonego procesu skrapowania, np. czy skraper nie pominął nazw produktów, zbierając same ich ceny. Na ostatnim etapie dane są ujednolicane, a produkty skrapowane selekcjonuje się i klasyfikuje. Tak przygotowane ramki danych (ang. *data frames*) poddawane są dalszej analizie.

W przypadku statystycznych stron internetowych scraping można sprowadzić do analizy DOM (ang. Document Object Model). DOM jest to sposób prezentacji dokumentów HTML lub XML w postaci struktury drzewa, w której każdy węzeł (znacznik) jest obiektem reprezentującym część dokumentu. Definiuje on sposób dostępu oraz określa dalsze prace nad dokumentem. Coraz popularniejsze staje się jednak tworzenie dynamicznych stron internetowych, pozwalających użytkownikowi na interakcje z witryną. Z pomocą renderingu (w wersji *client-side*) możemy tworzyć aplikacje zawierające tylko jedną stronę (ang. *single-page applications*), która zmienia się dynamicznie w zależności od potrzeb użytkownika. Liczba stron w takich aplikacjach ulega ciągłym zmianom i nie jest na stałe określona, tak jak w przypadku statycznych serwisów. Właśnie z myślą o dynamicznych serwisach internetowych powstała kolejna metoda skrapowania danych. Jest nią wysyłanie żądań HTTP do punktów końcowych (ang. *endpoints*) będących punktami połączenia, w których dane, pliki HTML lub aktywne strony serwerów są ujawnione (schemat 2.1).

Schemat 2.1. Bezpośrednie żądanie HTTP jako wycinek renderingu *client-side*



Źródło: opracowanie własne na podstawie: Christianis (2020).

W odpowiedzi na żądanie HTTP uzyskujemy ustrukturyzowane dane, które możemy natychmiast pobrać do pliku. Pozwala to zaoszczędzić dużo czasu w porównaniu z dopasowywaniem i mapowaniem selektorów w HTML DOM. Żądania HTTP całkowicie pomijają warstwę powłoki HTML, dzięki czemu strona, z której chcemy zebrać dane, nie musi być w pełni załadowana, a co za tym idzie skrapowanie tą metodą jest o wiele szybsze. Ponadto uzyskamy więcej danych, ponieważ ściągamy nie tylko dane widoczne na stronie internetowej, lecz także wszystkie zbiory z serwera.

Doświadczenia urzędów statystycznych wskazują, że zbieranie danych ze stron internetowych może być albo zlecane zewnętrznemu podmiotowi, albo prowadzone w ramach urzędu. Do pobierania danych ze stron internetowych można wykorzystać dostępne środowiska programistyczne i programy, takie jak R, Phyton, iMactos, SAS czy import.io, a także platformy zorientowane na duże zbiory danych (big data), takie jak Hadoop czy Spark. Język R oferuje m.in. pakiety przeznaczone do scrapingu, np. rvest oraz Rcrawler. Natomiast w Pythonie możemy wykorzystać inne pakiety, jak: bs4 (Beautiful Soup), requests, lxml, selenium czy też skrapy. Przykładowo urząd statystyczny w Wielkiej Brytanii opiera scraping na Pythonie, środowisko R wykorzystują Holendrzy, Niemcy i Włosi korzystają z iMactos, natomiast austriacki urząd wykorzystuje do skrapowania cen import.io.

Dane mogą być pobierane z dowolnie określoną częstotliwością – najczęściej urzędy statystyczne stosują częstotliwość dzienną, ale w przypadku badań niemieckiego urzędu statystycznego ceny skrapowano co godzinę. Należy jednak pamiętać, że zwiększenie częstotliwości prowadzi do wzrostu objętości danych oraz wydłuża czas ich przetwarzania.

2.4.1.6. Przetwarzanie danych skrapowanych

Dane pobrane z witryn internetowych są często nieujednolicone, nieuporządkowane, niewystandaryzowane (np. jednostką sprzedaży danego produktu może być w tym samym zbiorze danych gram lub kilogram) i niekompletne. Na pierwszym etapie należy więc ustrukturyzować ramkę danych do wymaganej postaci, usunąć rekordy z brakami danych oraz duplikaty (powtarzające się rekordy). Należy wyczyścić (usunięcie niepotrzebnych znaków) i sformatować według przyjętej konwencji etykiety produktów. Na drugim etapie produkty należy zaklasyfikować do odpowiednich homogenicznych grup produktów (najczęściej zdefiniowanych na poziomie COICOP 5 lub bardziej szczegółowym – COICOP 6). Następnie poklasyfikowane dane można wystandaryzować, dzięki czemu możliwa jest późniejsza opcjonalna filtracja danych (usunięcie z bazy danych produktów z ekstremalną zmianą ceny). W kolejnym kroku należy dopasować produkty w czasie (matchingu), czyli sparować ze sobą produkty obserwowane w różnych momentach. Ramka danych z dopasowanymi produktami jest punktem wyjścia do wyznaczenia wskaźników cen.

2.4.1.7. Wyzwania związane ze stosowaniem danych skrapowanych

Wykorzystanie danych skrapowanych w pomiarze inflacji wiąże się z wieloma wyzwaniami i problemami, jakie stoją przed urzędami statystycznymi. Należą do nich:

a) problemy związane z reprezentatywnością danych skrapowanych:

- dane pobierane ze stron internetowych nie są reprezentatywne dla koszyka typowego konsumenta, ponieważ nie uwzględniają informacji o cenach z małych

sklepów, targowisk itp. Dane są ograniczone tylko do tych sprzedawców, którzy mają swoje strony internetowe oraz umieszczają na nich cenniki;

- pobrane dane nie są reprezentatywne dla przeprowadzanych transakcji. Oznacza to, że zbierane są ceny o wszystkich dostępnych produktach niezależnie od tego, czy i w jakich ilościach zostały zakupione. W przypadku tradycyjnego gromadzenia danych o cenach ankieterzy wybierają tylko te produkty, które są reprezentatywne dla struktury zakupów konsumentów, a produkty, na które jest niski popyt, zwykle nie są brane pod uwagę;
- ceny produktów w internecie mogą nie być reprezentatywne dla całej sieci i różnić się od cen w sprzedaży tradycyjnej w zależności od lokalizacji sklepu. Ceny w sklepie tradycyjnym oraz w sklepie internetowym mogą się znacząco różnić, co wynika z różnej polityki cenowej prowadzonej przez każdy sklep w zakresie rabatów, promocji czy wyprzedaży. Należy też pamiętać, że ceny skrapowane są cenami ofertowymi, a nie transakcyjnymi;
- w przypadku wyboru produktów należy też ograniczyć kategorie i rodzaj towarów do tych, które są częścią tradycyjnych koszyków, na bazie których budowane są później wskaźniki cen. Warto też zachować ostrożność w przypadku kategorii, które są szczególnie nadreprezentowane w internecie, np. płyty CD i DVD czy książki. W takim przypadku może okazać się zasadne pobieranie informacji o cenach tylko w zakresie kategorii „najlepiej się sprzedające" (tak robi urząd statystyczny Norwegii);
- w przypadku danych zbieranych w internecie istnieje duża rotacja towarów. Oznacza to, że dostępność danego towaru może być bardzo krótkotrwała;

b) problemy techniczne:

- potencjalna zmiana kodu strony internetowej jest jednym z najważniejszych problemów dla programistów. W przypadku dużych zmian program skrapujący może nie zadziałać. Powoduje to przerwę w procesie zbierania danych i konieczność przeprogramowania skryptu oraz luki w szeregach czasowych;
- zmiany sposobu prezentacji cen – cena w kodzie strony internetowej nie zawsze musi być prezentowana jako ciąg znaków. W niektórych przypadkach właściciel strony internetowej może zapisywać cenę w formie obrazu lub tekstu;
- należy zwrócić również uwagę na takie zagrożenia jak potencjalne pobranie i wykonanie zawirusowanego oprogramowania. W związku z tym wskazane jest umiejscowienie programów pobierających dane ze stron internetowych w ramach samodzielnego systemu. W ten sposób wirusy i złośliwe oprogramowanie, które zostały pobrane, nie mogą przeniknąć do wewnętrznego środowiska IT urzędu statystycznego;
- braki danych mogą się pojawiać ze względu na przerwy w dostępie do internetu lub inne błędy systemu IT;

c) problemy prawne – prawne uregulowania w zakresie pobierania danych ze stron internetowych należy analizować oddzielnie na poziomie każdego kraju, ponieważ nie ma jednolitych międzynarodowych rozwiązań w tym zakresie. Założenia organizacyjne i metodyczne do skrapowania danych na potrzeby badań cen muszą być zgodne z obowiązującymi w danym kraju uregulowaniami prawnymi w zakresie statystyki, przepisami dotyczącymi ochrony baz danych, praw własności intelektualnej, ochrony prywatności i zasadami netykiety. W szczególności należy (rozwiązania stosowane m.in. przez urząd statystyczny Wielkiej Brytanii czy Holandii) zapewnić możliwość identyfikacji podmiotu zbierającego dane za pomocą web scrapingu, stosować okres bezczynności między żądaniami odsłony kolejnych stron, uruchamiać skrypty w nocy (w okresie najmniejszego ruchu w sieci), a także przestrzegać protokołu wykluczania robotów (robots.txt) oraz informować właścicieli strony internetowej o wykorzystaniu ich danych, o ile jest to wykonalne;

d) zasoby ludzkie – proces skrapowania cen wymaga zwiększenia zatrudnienia wysoko wykwalifikowanych osób, które przygotują platformę do pobierania danych oraz zapewnią kontrolę poprawności ich pobierania, wstępnego oczyszczania i ustrukturyzowania danych na potrzeby obliczeń CPI. Dodatkowo działające programy wymagają ciągłego nadzoru m.in. ze względu na zmiany w strukturze stron internetowych;

e) problemy dotyczące agregacji źródeł danych:
   - integracja nowych źródeł danych z istniejącymi procesami kompilacji indeksu cen wymaga zmian metodologicznych i obliczeniowych. Szczególną uwagę należy zwrócić na to, że pełne wykorzystanie danych ze źródeł internetowych często prowadzi do ich niespójności z istniejącymi standardami jakości, paradygmatami metodycznymi, a także wewnętrznymi procesami przetwarzania danych w ramach urzędu statystycznego;
   - ceny ze sklepów internetowych są zbierane codziennie, a nie raz w miesiącu, jak w przypadku CPI obliczanego metodami tradycyjnymi. Z jednej strony duża ilość zebranych danych oferuje możliwość zwiększenia liczby reprezentantów, z drugiej strony ogranicza to zakres możliwości porównywania wskaźnika obliczonego metodą tradycyjną i wskaźnika liczonego na bazie danych internetowych, a co za tym idzie utrudnia włączanie tych danych do obliczania CPI;
   - pobrane dane ze względu na swoją charakterystykę wymagają stosowania innych metod obliczania wskaźników cen (np. metod multilateralnych).

2.4.2. Dane skanowane

2.4.2.1. Definicja danych skanowanych

Zgodnie z definicją podaną w podręczniku na temat metod obliczania wskaźników cen konsumpcyjnych – *Consumer Price Index [CPI] Manual* (International Labour Organization [ILO] i in., 2004; International Monetary Fund [IMF] i in., 2020) przez dane skanowane (ang. *scanner data*) rozumiemy szczegółowe dane o dobrach konsumpcyjnych uzyskane dzięki skanowaniu ich kodów kreskowych w punktach sprzedaży.

2.4.2.2. Doświadczenia międzynarodowe dotyczące danych skanowanych

Technologia użytkowania kodów kreskowych produktów pojawiła się w latach 70. XX w., a ich wykorzystanie do analizy dynamiki cen i poprawy szacunków CPI nabrało szczególnego przyspieszenia w ciągu ostatnich 20 lat. W tym czasie nie tylko ewoluowały zakres i techniki zbierania danych skanowanych, lecz także poszerzyły się możliwości ich pozyskiwania i dalszego doskonalenia metodologii w zakresie konstrukcji indeksów cen (nowe formuły i nowe sposoby aktualizacji wag). Kraje europejskie nadal w zdecydowanej mniejszości wykorzystują dane skanowane, choć proces ten zaczyna obejmować coraz większą grupę państw. Przykładowo do 2015 r. w Europie tylko Holandia, Norwegia, Szwecja i Szwajcaria wykorzystywały dane skanowane do obliczeń indeksów cen detalicznych, a zaledwie rok później dołączyły do nich kolejne kraje: Belgia, Dania i Islandia. W 2022 r. ok. 1/3 krajów UE korzysta z danych skanowanych do szacowania inflacji. Luksemburg, Portugalia i Francja także eksperymentują z danymi skanowanymi dla wybranych podgrup koszyka CPI. GUS w 2022 r. współpracuje z trzema sieciami handlowymi (z kolejnymi dwiema sieciami prowadzone są rozmowy zmierzające do podpisania porozumienia w sprawie transferu danych). Pozyskiwane przez GUS dane skanowane służą na razie do oceny reprezentatywności produktów oraz imputacji braków danych wywołanych pandemią COVID-19, niemniej jednak oficjalne wdrożenie tego źródła danych do pomiaru inflacji jest kwestią najbliższych miesięcy. Generalnie Europa nie odbiega od trendów światowych w tym zakresie. Pionierami w stosowaniu danych skanowanych są Stany Zjednoczone, a niedościgniony poziom zaawansowania metodologicznego reprezentują Australia i Japonia. Japonia jest bardzo ciekawym przykładem kraju, który w pełni wykorzystuje swój potencjał technologiczny oraz współpracę potężnych instytucji (np. Bank of Japan, Nikkei Digital Media), aby jak najdokładniej szacować inflację, bazując na możliwie najszerszej gamie danych skanowanych. Japonia od 1998 r. gromadzi dane skanowane pochodzące od ponad 300 sieci supermarketów z całego kraju, przy czym jest to celowa próba dostawców (w przeciwieństwie np. do Stanów Zjednoczonych, gdzie pobiera się próbę losową). W kraju tym realizowany jest właśnie projekt rządowy funkcjonujący pod nazwą *UTokyo*

Daily Price Index, którego wymiernym efektem, jak wskazuje nazwa, jest publikacja rocznej stopy inflacji każdego dnia (w Japonii oficjalny poziom CPI publikowany jest raz na miesiąc).

2.4.2.3. Zawartość danych skanowanych

Elektroniczne terminale w punktach sprzedaży obsługują najczęściej następujące kody kreskowe: GTIN (ang. Global Trade Item Number) lub jego europejską wersję EAN (ang. European Article Number), PLU (ang. Price Look-Up) lub SKU (ang. Stock Keeping Unit). Najbardziej rozpowszechniony jest GTIN (EAN), choć na świecie funkcjonują też szczególne jego przypadki, np. UPC (ang. Universal Product Code) czy lokalny APN (ang. Australian Product Number). Przykładowo GTIN składa się z 8, 12, 13 lub 14 cyfr. Najbardziej popularna jest pełna wersja 13- i 14-cyfrowa, obejmująca: 1 cyfrę wskazującą poziom pakowania, 3-cyfrowy kod organizacji krajowej GS1 (potocznie: kod kraju, np. 590 – Polska), od 4 do 7 cyfr numeru jednostki kodującej GS1, od 2 do 5 cyfr kodu produktu i 1 cyfrę kontrolną. PLU jest oszczędniejszy w dostarczanych informacjach od GTIN, ponieważ jest krótszy, z kolei SKU jest bardziej ogólny niż GTIN, który dostarcza więcej detalicznych informacji. Paradoksalnie to właśnie zbyt szczegółowy poziom informacji o produkcie, jakiego dostarcza GTIN, sprawia, że posługując się nim, trudno wyłonić homogeniczne grupy produktów (np. ten sam produkt, ale w innym opakowaniu, może mieć dwa różne kody GTIN). Niektóre kraje korzystające z danych skanowanych przy obliczaniu CPI używają więc bardziej ogólnego SKU.

Poza omówionym kodem kreskowym produktu dane skanowane zawierają jeszcze wiele innych cennych informacji. Zasób tych informacji jest różny w poszczególnych krajach, m.in. w zależności od dostawcy (sieci supermarketowej) czy rodzaju produktu. W idealnym przypadku dane skanowane zawierają: kod sprzedawcy (określa grupę towarową według indywidualnej klasyfikacji danej sieci), kod identyfikujący punkt sprzedaży w obrębie danej sieci, etykietę produktu (dodatkowy opis produktu i jego charakterystyki), jednostkę sprzedaży (optymalnie według ujednoliconego formatu, np. „szt.", „kg", „paczka", „500 g", „1 litr" itd.), wartość sprzedaży, liczbę sprzedanych jednostek produktu, flagę (np. oznacza się produkty z przecen i promocji) i informację o VAT.

2.4.2.4. Pozyskiwanie danych skanowanych

Wyróżniamy kilka podstawowych źródeł danych skanowanych, z którymi wiążą się pewne korzyści, ale również pewne ograniczenia i problemy. Najcenniejszym źródłem tego typu danych wydają się bezpośredni dostawcy, a więc punkty sprzedaży, ze szczególnym uwzględnieniem sieci supermarketów. Supermarkety to potężni potencjalni dostawcy danych skanowanych – typowy supermarket posiada bazę 10 000–25 000 kodów kreskowych sprzedawanych produktów, z których większość

stanowią żywność i napoje, ale są reprezentowane również takie kategorie produktów, jak odzież i obuwie, drobna elektronika czy kosmetyki. Teoretycznie podobnymi dostawcami danych skanowanych mogą być również mniejsze markety, drobni sprzedawcy, apteki, biura turystyczne lub nawet sklepy internetowe, o ile tylko archiwizują dane o sprzedaży, uwzględniając kodowanie produktów. Drugim, alternatywnym źródłem danych skanowanych mogą być firmy wyspecjalizowane w badaniu rynku. Niektóre kraje korzystają z danych skanowanych dostarczanych np. przez firmę Nielsen lub GfK i włączają je do szacunków krajowego CPI (Krsinich, 2014). Korzystanie z takiego rozwiązania jest jednak kosztowne i przez to mniej popularne.

2.4.2.5. Przetwarzanie danych skanowanych

Dane skanowane dostarczane są od dostawców najczęściej w formacie CSV lub rzadziej, ze względu na rozmiar, w formacie XLS (GUS pobiera dane w formacie CSV). Niektóre kraje stosują też transfer danych przez serwisy sieciowe (format XML) lub stosują swój wewnętrzny format zapisu. Niekiedy, tuż po dostarczeniu danych, niezbędne jest ich przeformatowanie do takiej ramki danych, jakiej wymaga środowisko IT, w którym dokonywane są dalsze analizy, lub konkretny skrypt danego środowiska. Może nim być środowisko R, Phyton, SAS czy inne, współpracujące w wymienionymi formatami zapisu danych. W zależności od kraju częstotliwość pozyskiwanych danych skanowanych może być dzienna, tygodniowa lub miesięczna. Dane skanowane najczęściej poddawane są już na wstępie procesom oczyszczania (usunięcie duplikatów, rekordów podejrzanych lub z niekompletnymi danymi) oraz agregowania. W zależności od potrzeb kraju dane skanowane są agregowane do dnia, tygodnia lub, najczęściej, do miesiąca. W tym ostatnim przypadku z reguły uwzględnia się transakcje z trzech środkowych tygodni każdego miesiąca. Następnie, mając tak przygotowane dane, dokonuje się klasyfikacji produktów do grup COICOP (minimalnym wymogiem jest COICOP 5, ale najczęstszą praktyką jest poziom COICOP 6). Do celów klasyfikacyjnych wykorzystuje się kody produktów (EAN, kod dostawcy), a w przypadku ich braku lub niejednoznaczności – dodatkowo etykiety produktów, sięgając po metody uczenia maszynowego (ang. *machine learning*), a także metody analizy tekstu (ang. *text mining*). Po etapie klasyfikacji następuje etap dopasowywania produktów z porównywanych momentów czasowych (matching). W przypadku dopasowywania produktów chodzi o to, aby obserwować w czasie ceny tego samego homogenicznego produktu, nawet jeśli zmieni on kolor opakowania i wagę, a co za tym idzie zostanie mu nadany inny kod, np. EAN. Dysponując kodami kreskowymi (GTIN, EAN, SKU itp.), kodami wewnętrznymi sprzedawców oraz dostarczanymi przez nich charakterystykami ilościowymi i jakościowymi, można ustalić zbiór dopasowanych homogenicznych produktów dla porów-

nywanych ze sobą jednostkowych okresów, np. kolejnych miesięcy. Dalej, posiadając już zbiory poprawnie sklasyfikowanych i dopasowanych produktów, można przystąpić do filtrowania danych. Etap ten najczęściej uwzględnia udział sprzedaży danego produktu w łącznej sprzedaży w przynależnej mu grupie produktów w porównywanych miesiącach czy też zmianę ceny produktu z miesiąca na miesiąc. W przypadku zbyt małych udziałów w rynku danego produktu (np. w porównywanych miesiącach) najczęściej jest on usuwany z próby, a w przypadku ekstremalnie małej lub dużej zmiany ceny takie produkty są oznaczane specjalną notką (flagowane) i poddawane dalszej analizie (niektóre kraje również usuwają je z próby). Dalsza analiza również ma wiele wariantów w zależności od kraju. Niekiedy taką ekstremalną cenę traktuje się jako brakującą informację i dokonuje jej imputacji. Często także sprawdza się, czy np. nietypowo niskiej cenie towarzyszy nietypowy spadek ilości sprzedaży. Jeśli tak jest, to analizowany produkt uznaje się za wycofywany z oferty sieci, a cenę za „zrzucaną" (ang. *dump price*). W konsekwencji taki produkt usuwany jest z próby. Dopiero przygotowane w ten sposób dane nadają się do tego, aby na ich podstawie wyznaczyć indeks cenowy.

2.4.2.6. Wyzwania związane ze stosowaniem danych skanowanych

Głównymi problemami i wyzwaniami, jakie wiążą się z wdrożeniem danych skanowanych do pomiaru CPI, są:

- wybór dostawcy danych i współpraca z nim. Doświadczenia Polski oraz innych krajów wskazują, że niełatwo pozyskać nowego dostawcę danych skanowanych. Sieci marketów są niekiedy niechętne do współpracy z urzędami statystycznymi, nie widząc w takim działaniu korzyści finansowych oraz obawiając się osłabienia w ten sposób swojej konkurencyjności (obawa przed niepełną poufnością danych). Od momentu nakłonienia właścicieli supermarketu do współpracy do sfinalizowania umowy legalizującej jej realizację upływa najczęściej ok. pół roku;
- dobór próby do analizy. Polityka krajów stosujących dane skanowane w zakresie doboru próby jest różna, począwszy od tego, że część z nich stosuje próby losowe, a część – celowe. Zależy to głównie od tego, z iloma sieciami handlowymi podpisano umowy o współpracy oraz od zróżnicowania geograficznego i demograficznego danego kraju. Gdy dana sieć marketów różnicuje swoją ofertę oraz ceny w zależności od regionu, wówczas niezbędne jest uwzględnienie rejonizacji przy doborze próby. Podobnie jeśli w danym rejonie różne punkty notowań należące do tej samej sieci stosują np. różne godziny otwarcia czy też promocje, dobór próby powinien również to uwzględniać i traktować takie punkty notowań indywidualnie (agregacja do całej sieci jest tu niewskazana). Wybór momentu dla poboru próby także może być problematyczny. Rozkład cen w ciągu dni tygodnia nie jest równomierny, np. przed weekendem ceny często są zawyżane. Podobnie agregacja

danych do tygodnia może rodzić wątpliwości, które tygodnie (i ile) z danego miesiąca są najbardziej reprezentatywne;

- reprezentatywność danych skanowanych. Należy podkreślić, że nie każdy produkt dostępny w ogólnokrajowej sprzedaży i będący jednocześnie w koszyku CPI jest dostępny w supermarketach. Dodatkowo istnieje znaczna grupa ludzi, która nie robi zakupów w sieciach handlowych lub czyni to bardzo sporadycznie. Z tego względu dane skanowane, przynajmniej na razie, traktuje się jako wsparcie w pracy ankieterów w terenie, a nie ich zastąpienie;
- budowa środowiska IT. Środowisko IT przeznaczone do analizy danych skanowanych musi korzystać z zaawansowanych metod i technik uczenia maszynowego (klasyfikacja, rozpoznawanie tekstu). Dynamiczny charakter zbiorów danych skanowanych w połączeniu z ich ogromnym wolumenem i różnorodnością sprawia, że omówione wcześniej niezbędne procesy (filtrowanie danych, dopasowywanie produktów, klasyfikacja do homogenicznych grup, obliczenia indeksów cen) są dość czasochłonne i wymagają nie tylko szybkich algorytmów, lecz także często uruchomienia dodatkowych serwerów. Odrębnego oprogramowania wymaga również moduł odpowiedzialny za wyznaczanie indeksów cenowych na podstawie danych skanowanych. W środowisku R można zainstalować pakiety, które temu służą (np. IndexNumR czy PriceIndices);
- problemy techniczne. Nagła zmiana sposobu kodowania i etykietowania produktów przez sieć handlową współpracującą z urzędem statystycznym powoduje trudności z klasyfikowaniem produktów (należy przebudować modele klasyfikujące, co wiąże się z przygotowaniem nowych zbiorów uczących dla metod uczenia maszynowego) oraz ich dopasowaniem w czasie;
- opracowanie metodologii przygotowania i analizy danych skanowanych. Mimo bogatych doświadczeń wielu urzędów statystycznych w zakresie stosowania danych skanowanych nadal istnieją pewne problemy metodologiczne. Chodzi m.in. o filtrowanie danych. Kraje dzielą się na jego zwolenników i przeciwników (ci drudzy uważają, że filtrując, nadmiernie redukujemy próbę i tracimy dynamiczny charakter zbioru danych na rzecz bardziej statycznego). Kraje, które zdecydowały się na filtrowanie i imputację danych, stoją przed wyborem parametrów progowych dla tych filtrów (np. jaką zmianę ceny uznać za już nietypową albo jaki udział w rynku uznać za wystarczająco duży, aby włączyć produkt do próby) lub też samej techniki imputacji (w tym zakresie są rekomendacje Eurostatu). Metodologia powinna również określać, za pomocą jakich technik uczenia maszynowego dopasowuje się produkty w porównywanych miesiącach oraz ich klasyfikacji do grup/podgrup COICOP (właściwe mapowanie produktów nadal jest dużym wyzwaniem dla większości krajów). Kolejnym wyzwaniem jest uwzględnienie dóbr sezonowych. Niektóre kraje wręcz wykluczają takie produkty z próby

(zwłaszcza jeśli chodzi o mocną sezonowość, polegającą na tym, że w pewnych okresach roku produkty są całkowicie niedostępne). Choć metodologia konstruowania indeksów cen bazuje na pojęciu homogenicznej grupy produktów, to kwestia właściwego zdefiniowania tego pojęcia oraz stworzenia technik do selekcji takich grup (np. metoda MARS; por. Chessa, 2021) jest nadal szeroko dyskutowana w Eurostacie oraz krajach członkowskich UE i pozostaje otwarta. Ogromnym metodologicznym wyzwaniem jest wybór formuły indeksu cenowego, który czyniłby zadość pewnym formalnym i praktycznym wymogom (np. powinien mieć dobre własności aksjomatyczne i redukować zjawisko łańcuchowego dryfu – ang. *chain drift*; por. Chessa, 2015). Wreszcie finalnym problemem, jaki stoi przed każdym urzędem statystycznym chcącym korzystać z danych skanowanych, jest wybór systemu wag, który służy przechodzeniu do wyższych poziomów agregacji i włączeniu wyliczonych indeksów cen do publikowanej informacji o poziomie inflacji.

2.5. Metodyka konstrukcji indeksów cen w monitorowaniu rozmiarów inflacji

W każdym miesiącu rejestrowanych jest ok. 350 tys. danych jednostkowych dotyczących poziomów cen towarów i usług. Na podstawie przygotowanego zbioru notowań cen obliczane są średnie miesięczne ceny wszystkich reprezentantów jako średnie arytmetyczne z uwzględnieniem liczby notowań. Na tym etapie GUS dysponuje więc pełną wiedzą o kształtowaniu się cen wewnątrz grup elementarnych. Następnie wyznacza się wskaźniki cen reprezentantów pochodzących z poszczególnych grup elementarnych. Wskaźnik cen reprezentanta w rejonie wynika z odniesienia jego średniej miesięcznej ceny z okresu bieżącego do średniej ceny z okresu bazowego. Ogólnopolskie wskaźniki cen wszystkich reprezentantów objętych notowaniami obliczane są jako średnie geometryczne wskaźników cen reprezentantów ze wszystkich rejonów. Na ich podstawie, stosując ponownie średnią geometryczną ze wskaźników wyznaczonych wewnątrz grup elementarnych, oblicza się wskaźniki cen grup towarów i usług konsumpcyjnych na najniższym poziomie agregacji systemu wag. Poziom ten nazywamy właśnie *poziomem elementarnym* (ang. *elementary level*).

Szczegóły konstrukcji wybranych formuł indeksów cen omówiono w podrozdz. 4.5, w tym miejscu jednak przedstawione zostaną najważniejsze aspekty metodyki konstrukcji indeksów cen, w szczególności odnoszące się do alternatywnych źródeł danych. Jak widać z powyższego opisu, w przypadku tradycyjnej metody gromadzenia danych o cenach w GUS (i większości krajów świata) podstawową formułą wyznaczania indeksu cenowego na elementarnym poziomie agregacji danych jest formuła Jevonsa (1865). Ponieważ formuła ta uwzględnia w swojej konstrukcji tylko ceny dóbr, a informacja o poziomie konsumpcji jest całkowicie pomijana, podobnie

jak każda inna formuła tego rodzaju nazywana jest *indeksem nieważonym*. Indeks Jevonsa, stanowiący średnią geometryczną z cząstkowych indeksów cen wyznaczonych dla uwzględnianych dóbr, ma bardzo dobre własności aksjomatyczne i stochastyczne, w związku z tym jest to rekomendowana formuła indeksowa dla poziomu elementarnego (ILO i in., 2004; IMFi in., 2004, 2020). W literaturze przedmiotu ugruntowały się również inne (nieważone) formuły, które mogą być stosowane na elementarnym poziomie obliczeń: indeks Dutota (1738) i indeks Carliego (1804). Pierwszy z nich stanowi iloraz średnich arytmetycznych z cen dóbr obserwowanych odpowiednio w okresie bieżącym i w okresie badanym, a drugi – średnią arytmetyczną z cząstkowych wskaźników cen. W literaturze można znaleźć wiele prac porównujących własności tych indeksów (Białek, 2020, 2022c; Diewert i Silver, 2008, Levell; 2015; Silver i Heravi, 2007). Dowiedziono m.in., że indeks Carliego prowadzi do przeszacowania danych o inflacji i generuje wartości najczęściej wyższe niż indeks Jevonsa. Jest to jeden z powodów, dla których żaden z krajów UE nie stosuje indeksu Carliego na najniższym poziomie agregacji. Rekomendacje dotyczące obliczeń CPI i HICP dają pewną dowolność w stosowaniu indeksu Dutota, ale z zastrzeżeniem, że formuła ta dotyczy homogenicznego zbioru towarów lub usług. Zwróćmy zatem uwagę, że w przypadku tradycyjnej metody gromadzenia danych, polegającej na notowaniu przez ankietera jedynie informacji o cenach dóbr, spektrum potencjalnych formuł jest na tym etapie mocno ograniczone. Co więcej, pewne wątpliwości może budzić to, że każdy z reprezentantów tworzących daną grupę elementarną otrzymuje tę samą wagę przy konstrukcji indeksu cenowego (brak systemu ważenia oznacza równe wagi). Jest to niedogodność, którą można wyeliminować, jeśli zastosuje się dane skanowane. Dane tego rodzaju zawierają kompletną informację o sprzedawanych produktach zawartą w kodzie kreskowym, a zatem dysponujemy tu wiedzą o wielkości sprzedaży produktów już na poziomie reprezentantów (por. pkt 2.4.2).

Wróćmy jednak do danych zbieranych tradycyjną metodą. Po wyznaczeniu indeksów cen dla grup elementarnych wykorzystuje się je następnie, przy zastosowaniu systemu wag, do obliczenia wskaźników na wyższych poziomach agregacji aż do wskaźnika cen towarów i usług ogółem. Na wyższych poziomach agregacji urząd statystyczny dysponuje już nie tylko wskaźnikami cen dla grup towarów i usług, lecz także informacjami o konsumpcji charakteryzującej te grupy, dlatego możliwe staje się na tym etapie zastosowanie ważonych indeksów cen. W przypadku krajów, które do konstrukcji systemu wag stosują strukturę wydatków konsumpcyjnych z badania budżetów domowych z roku poprzedzającego rok badany, a następnie ważoną średnią arytmetyczną ze wskaźników cząstkowych, możemy mówić o zastosowaniu indeksu Laspeyresa (1871). Indeks ten, w większości krajów (w tym w Polsce), obliczany jest z częstotliwością miesięczną.

Indeks Laspeyresa (por. pkt 4.4.2) reprezentuje rodzinę indeksów typu Laspeyresa (ILO i in., 2004; IMF i in., 2020). Upraszczając, można mówić o indeksie, który na podstawie porównania okresu badanego do bazowego ustala ilości konsumowanych dóbr lub ich udział w konsumpcji na poziomie z okresu wcześniejszego niż okres bazowy (indeksy Lowe'a i Younga; por. Białek, 2017; ILO i in., 2004; IMF i in., 2020). Działanie takie jest zasadne, gdy ze względu na ograniczanie kosztów badania budżetów gospodarstw domowych jest ono realizowane rzadziej niż raz na rok. W tym miejscu poczynimy jednak uwagę analogiczną do tej, która dotyczyła indeksów nieważonych i dobrodziejstw danych skanowanych. Otóż kompletność danych skanowanych nie tylko umożliwia zastosowanie indeksów ważonych na najniższym poziomie agregacji danych, lecz także zwiększa możliwość wyboru indeksu ważonego ze względu na dostęp do pełnych danych również dla okresu bieżącego. Innymi słowy, przy obliczaniu wskaźników cen z wykorzystaniem danych skanowanych nie jesteśmy ograniczeni do indeksów typu Laspeyresa. Nie ma najmniejszych przeszkód, aby stosować tu indeksy trzeciej generacji, powstające przez krzyżowanie wag (np. indeks Walsha (1901), Marshalla (1887) czy Geary'ego-Khamisa (1958 i 1970)) lub przez krzyżowanie formuł (indeks Drobischa, 1871) jako średnia arytmetyczna z indeksów Laspeyresa i Paaschego czy też indeks Fishera (1922) – jako ich średnia geometryczna. Ostatni z wymienionych indeksów uważany jest za najlepszą ważoną formułę bilateralną, ponieważ spełnia najwięcej wymaganych aksjomatów (testów; Balk, 1995; von der Lippe, 2007), oraz najlepsze możliwe przybliżenie indeksu kosztów utrzymania (ang. Cost of Living Index – COLI).

Podsumowując dotychczasowe rozważania, dochodzimy do wniosku, że w przypadku tradycyjnej metody zbierania danych wybór indeksu cenowego na poszczególnych poziomach agregacji podyktowany jest dostępnością danych i właściwie jest przesądzony (Jevons, Laspeyres). W przypadku nowych źródeł danych wybór formuły indeksu jest kwestią otwartą i wydaje się, że wciąż nierozstrzygniętą. Wynika to z faktu, że skanując lub skrapując informacje o produktach, napotykamy zjawisko określane w literaturze przedmiotu jako *product churn*, a więc dużą rotację produktów, np. wśród produktów sprzedawanych przez sieci handlowe spory odsetek stanowią dobra sezonowe, produkty znikające lub zupełnie nowe. Dynamika zmian w ciągu roku jest tutaj bardzo duża, dlatego zwykły indeks bilateralny, który nie bierze pod uwagę zmian cen w analizowanym przedziale czasowym, wydaje się zbytnio obciążać pomiar. Z kolei zastosowanie indeksów łańcuchowych (ILO, i in., 2004; IMF i in., 2020), które stanowią iloczyn wszystkich indeksów cen wyznaczonych dla sąsiadujących ze sobą okresów (miesięcy), jest albo niepełne w przypadku ich nieważonych wersji (ponieważ nie uwzględniamy wiedzy o poziomie konsumpcji), albo niewłaściwe w przypadku ważonych indeksów łańcuchowych. W tej drugiej sytuacji problem polega na tym, że nawet najlepsze ważone indeksy łańcuchowe (np.

łańcuchowy indeks Fishera) generują efekt łańcuchowego dryfu, co znacznie obciąża wynik
pomiaru dynamiki cen. Zjawisko łańcuchowego dryfu pojawia się wtedy, gdy mimo powrotu cen i ilości konsumowanych produktów do swoich wyjściowych wartości indeks cenowy wbrew oczekiwaniom nie przyjmuje wartości jednostkowej. Dlatego właśnie w przypadku dynamicznych danych skanowanych rekomenduje się zastosowanie indeksów multilateralnych (Białek i Bobel, 2019), które działają w zadanym oknie czasowym (z reguły 13-miesięcznym), skutecznie eliminując obciążenie pomiaru wynikające z efektu łańcuchowego dryfu. Wśród nich najpopularniejszymi metodami są indeksy GEKS, CCDI, Geary'ego-Khamisa, TPD czy SPQ (szersze omówienie tych formuł zamieszczono w pkt 4.5.4).

2.6. Wyzwania statystyki publicznej w kontekście badań cen detalicznych

CPI mierzy zmiany w czasie poziomu cen towarów i usług, które gospodarstwa domowe nabywają w celu konsumpcji. W wielu krajach wskaźniki CPI zostały pierwotnie wprowadzone w celu zapewnienia pomiaru zmian kosztów życia pracowników, tak aby wzrost płac mógł być powiązany ze zmieniającymi się poziomami cen. Jednak z biegiem lat zakres CPI ulegał rozszerzeniu i obecnie wskaźniki CPI są szeroko stosowane jako makroekonomiczne wskaźniki inflacji. Ogromną wagę przywiązuje się zatem do jakości i dokładności krajowych wskaźników CPI, ale także ich porównywalności międzynarodowej.

Sposób konstrukcji wskaźników powinien, w odpowiedzi na oczekiwania użytkowników danych, pozwalać na dostosowanie do szerokiego zakresu szczególnych zastosowań, np. do obliczania wskaźników inflacji dla poszczególnych grup społecznych. Zakres produktów objętych badaniem cen powinien pozwolić także na prezentację wskaźników dla określonych segmentów rynku, jak np. nośniki energii, CPI z wyłączeniem pewnych grup produktów (np. wyrobów tytoniowych i alkoholu) czy umożliwiać oszacowanie wpływu zmian stawek podatkowych lub cen regulowanych. Wpływ czynników, takich jak zmiany zachowań konsumentów, uwarunkowania rynku, dostępności nowych źródeł danych, a także nowelizacje zaleceń międzynarodowych, powoduje, że badanie cen konsumpcyjnych wymaga udoskonaleń i modyfikacji i w rezultacie stale się zmienia. W *CPI Manual* (ILO i in., 2004; IMF i in., 2020) sprecyzowano cztery wyzwania dla urzędów statystycznych:

1) identyfikacja potrzeb użytkowników;
2) konceptualizacja potrzeb użytkowników w odniesieniu do koncepcji ekonomicznych;
3) przełożenie przyjętej koncepcji na warunki pomiaru statystycznego zgodnie z podstawowymi zasadami pomiaru zmian cen;

4) obliczanie zdefiniowanych wskaźników oraz ocena ich zgodności z założonym celem.

Konceptualizację potrzeb należy rozpatrywać w kontekście m.in.:

- przemian społeczno-ekonomicznych (np. starzenie się społeczeństwa, źródła utrzymania, zróżnicowanie dochodów, zróżnicowanie regionalne);
- możliwości pomiaru zmian cen w zakresie nowych grup towarów i usług;
- uwzględniania grup dotychczas nieuwzględnionych w CPI, np. cen związanych z kupnem domu lub mieszkania i kosztów ponoszonych przez ich właścicieli (ang. owner-occupied housing – OOH);
- obecnych założeń badania;
- dostępnych zasobów kadrowych i kosztów badania;
- kompromisu pomiędzy jakością danych wynikowych a terminem publikacji (np. korzystanie jedynie z danych dostępnych w momencie opracowywania wskaźnika).

Opracowany przez ekspertów międzynarodowy program rozwoju badań cen konsumpcyjnych obejmuje zagadnienia koncepcyjne i metodologiczne, a także aspekty praktyczne i przyszłe wyzwania. Niezbędne są dalsze prace analityczne i wymiana doświadczeń międzynarodowych w definiowaniu i rozwijaniu najlepszych praktyk, w tym w obszarach, w których nie uzyskano konsensusu co do stosowanych metod. Takim obszarem są przykładowo zagadnienia związane z implementacją nowych źródeł danych, wyzwania związane z pomiarem cen usług, kwestie koncepcyjne i pomiarowe ujmowania we wskaźnikach cen konsumpcyjnych kosztów ponoszonych przez właścicieli mieszkań czy zagadnienia obejmujące doskonalenie wag wykorzystywanych w obliczeniach. Dużym wyzwaniem stojącym przed statystyką cen będzie wdrożenie do procesu obliczeń klasyfikacji COICOP2018, przyjętej jako nowy standard przez Komisję Statystyczną ONZ i wprowadzającej wiele zmian strukturalnych (w tym podklasę – poziom 5-cyfrowy, strukturę podziału wydatków na 13 działów).

Niezbędne jest unowocześnienie procesu pomiaru zmian cen detalicznych z wykorzystaniem nowych źródeł danych i narzędzi big data, co będzie się wiązać z dostosowaniem systemu przetwarzania danych do:

- znacznie większej ilości i szerszego zakresu pozyskiwanych danych;
- nowych technologii przetwarzania danych;
- nowej formuły obliczania wskaźnika;
- zmiany metody agregacji danych, pozwalającej na łączenie danych zebranych tradycyjną metodą z danymi pochodzącymi z alternatywnych źródeł.

Często ograniczenia rozwoju metodologii wynikają z problemów praktycznych, jak brak odpowiednich źródeł danych i odpowiednich metod ich przetwarzania czy niewystarczające zasoby urzędu statystycznego.

Pojawienie się kryzysu wywołanego pandemią COVID-19 spowodowało kolejne, bezprecedensowe wyzwanie w obliczaniu wskaźników cen konsumpcyjnych. Urzędy statystyczne musiały zbadać nowe źródła danych i wdrożyć metody uzupełniania brakujących obserwacji cen, a także dokonać zmian w sposobie przetwarzania danych, tak aby zapewnić publikację wiarygodnego CPI. Zaistniała konieczność opracowania dodatkowych obszernych wytycznych i rekomendacji w zakresie gromadzenia danych, metod ich przetwarzania i obliczania wskaźników cen, stosowanych systemów wag, a także komunikacji z użytkownikami danych na temat dokładności i wiarygodności prowadzonych obliczeń.

ROZDZIAŁ 3
Wsparcie statystyki publicznej w metodyce i narzędziach pomiaru inflacji – założenia projektu

3.1. Uwarunkowania i wyzwania związane z gromadzeniem danych o cenach w terenie

Obecnie stosowane metody gromadzenia danych o cenach bazują w dużej mierze na ręcznym zbieraniu danych przez sieć ankieterów statystycznych. Do zadań ankieterów zatrudnionych w 16 wojewódzkich urzędach statystycznych należy – poza zanotowaniem cen wytypowanych reprezentantów towarów i usług – wybór placówek do badania cen detalicznych, nawiązywanie kontaktów z tymi placówkami w celu podjęcia stałej współpracy, dobór towarów i usług do badania cen, notowanie informacji dodatkowych zgodnie z zasadami opracowanymi przez GUS, dokonywanie podmian obserwowanych produktów, a także udział w weryfikacji próby.

Stałą kontrolę i nadzór nad ankieterami sprawują pracownicy wojewódzkich urzędów statystycznych. Do ich obowiązków należy m.in. wyrywkowe sprawdzanie cen i cech jakościowych produktów, poza tym kontroler weryfikuje, czy obserwowany produkt jest dostępny na półce sklepowej, czy jego jakość jest taka sama jak podana w formularzu, a także czy prawidłowo zarejestrowano jego cenę. Gromadzenie danych przez ankieterów jest wspomagane komputerowo – korzystają oni z urządzeń mobilnych z zainstalowaną aplikacją umożliwiającą również porównanie ceny zarejestrowanej w badanym miesiącu z ceną i innymi podstawowymi informacjami z poprzedniego okresu notowania. Oprogramowanie pozwala na automatyczną kontrolę kodów, jednostek sprzedaży i przedziału cenowego. Jeśli rozbieżność między ceną z ostatniego miesiąca a ceną z bieżącego miesiąca przekroczy wskazany zakres, to oprogramowanie wymaga przeprowadzenia dalszych weryfikacji, a uwagi zostają przekazane na kolejne etapy kontroli. Wszelkie zmiany obserwowanych reprezentantów są wyjaśniane i potwierdzane. Gdy zanotowanie ceny danego produktu nie jest możliwe, program wymusza wprowadzenie specjalnego kodu objaśniającego. Użycie tego kodu wskazuje, że brak danych nie wynika z błędu ankietera, a dany

produkt jest niedostępny w wybranym punkcie sprzedaży. Jeśli jest to chwilowy brak notowań, to szacunku ceny dokonuje się na późniejszym etapie. Jeśli zarejestrowana cena jest poprawna, czyli dotyczy produktu porównywalnego jakościowo z produktem notowanym w poprzednim miesiącu, również stosuje się specjalny kod wyjaśniający zmianę cen. Jeżeli jednak pozycja różni się pod względem jakości od podanej w poprzednim miesiącu, to ankieter wprowadza oznaczenie wskazujące, że na późniejszym etapie analizy należy dokonać korekty ceny po wyeliminowaniu zmian dotyczących jakości produktu. Podmiana produktu następuje w sytuacji, gdy dotychczas obserwowany produkt zostanie wycofany ze sprzedaży lub stanie się mniej popularny wśród konsumentów. Dla danych zgromadzonych w terenie przez sieć ankieterów stosowane są następujące metody korekty jakości (ich zastosowanie i dalsze postępowanie z nimi w znacznej mierze uzależnia się od oceny ankietera na etapie notowania cen):

- porównanie bezpośrednie – gdy nowy produkt uznawany jest za równoważny jakościowo z poprzednim;
- korekta związana ze zmianami rozmiarów opakowań – wielkość opakowania jest jedyną istotną zmianą jakości;
- zakładka – dopuszczalna w przypadku zmiany punktów sprzedaży; obserwowany jest jednocześnie stary i nowy produkt (wykonywana jest wycena dwutorowa);
- zakładka pomostowa – jedyną możliwością oszacowania zmiany jakości jest uznanie, że zmiana ceny pozycji zastępczej w danym rejonie badania jest równoważna ze średnią zmianą ceny w innych lub wybranych (np. tych, w których notowania są realizowane w placówkach handlowych tego samego typu) rejonach badania cen.

Wymogi stawiane badaniu cen oraz uwarunkowania rynku detalicznego powodują, że notowanie cen bezpośrednio w terenie wymaga dużych nakładów pracy ankieterów i osób nadzorujących. Metody oparte głównie na ręcznym zbieraniu danych przez ankieterów statystycznych mogą prowadzić do uzyskiwania coraz mniej reprezentatywnej próby (zwłaszcza w niektórych segmentach konsumpcyjnych), ograniczającej m.in. możliwości opisania procesów zachodzących na rynku detalicznym i prognozowania zjawisk inflacyjnych przy wysokim ryzyku błędu, wynikającym np. z konieczności rozstrzygania w sytuacjach problemowych. Wymagane jest wypracowanie kompromisu między maksymalizacją wielkości próby a minimalizacją kosztów prac terenowych i dalszego przetwarzania.

W pewnych sytuacjach notowanie cen przez sieć ankieterów w niektórych segmentach rynku staje się niemożliwe – za przykład może posłużyć gromadzenie cen usług ubezpieczeniowych. Z uwagi na wymóg podawania danych osobowych w celu obliczenia składki, brak czasu agentów ubezpieczeniowych na to, aby udzielić ankieterowi niezbędnych informacji, a także coraz większą złożoność kalkulacji ceny

usługi konieczna stała się zmiana źródła danych i tym samym stopniowe wyłączanie sieci ankieterów z notowania cen dla tego sektora usług.

W 2020 r., z powodu wprowadzenia obostrzeń epidemicznych, praca ankieterów statystycznych została wstrzymana na wiele miesięcy. Informacje o cenach detalicznych, zamiast bezpośrednio w terenie, zbierano za pośrednictwem Internetu i telefonicznie, wprowadzono również wiele niestandardowych rozwiązań dotyczących gromadzenia i przetwarzania danych. Trudna sytuacja gospodarcza na rynku detalicznym często powoduje likwidację placówek oraz niechęć do przekazywania ankieterom informacji na temat cen i dodatkowych cech produktów. Konieczne stało się opracowanie dodatkowych rekomendacji i wytycznych. Utrzymanie efektywnej sieci ankieterów wymaga także stałego wsparcia w dostarczaniu im odpowiedniego oprogramowania i ergonomicznego, dostosowanego do wielogodzinnej pracy w terenie, sprzętu.

3.2. Badanie budżetów gospodarstw domowych jako źródło danych o strukturze konsumpcji

Niezbędnym elementem procesu obliczania wskaźników cen towarów i usług konsumpcyjnych są przeciętne roczne wydatki gospodarstw domowych przeznaczane na zakup towarów i usług konsumpcyjnych. Głównym statystycznym źródłem danych o wydatkach konsumpcyjnych jest badanie budżetów gospodarstw domowych. Jego wyniki stanowią podstawę m.in. budowy systemów wag na potrzeby badań cen detalicznych. Z uwagi na zmieniający się model i strukturę konsumpcji (pojawianie się na rynku nowych produktów i marek, zmiana stylu życia, trendów w modzie, coraz większa świadomość ekologiczna) aktualizacja systemu wag wykorzystywanego do obliczania wskaźników cen następuje co roku i wymaga zastosowania najbardziej aktualnych i szczegółowych informacji o wydatkach konsumpcyjnych gospodarstw domowych. Badanie budżetów gospodarstw domowych prowadzone jest metodą reprezentacyjną, opartą na próbie losowej, która daje możliwość uogólnienia uzyskanych wyników na wszystkie gospodarstwa domowe w kraju. Dane są dostępne w wielu przekrojach, w tym według grup społeczno-ekonomicznych, niezbędnych do budowy systemów wag do wskaźników cen towarów i usług konsumpcyjnych w tym zakresie. Wyniki badania budżetów gospodarstw domowych zasilają również system rachunków narodowych. Na ich podstawie powstaje m.in. kategoria finalne konsumpcyjne pieniężne wydatki gospodarstw domowych (ang. *household final monetary consumption expenditure* – HFMCE), która stanowi źródło danych przy opracowywaniu systemu wag do zharmonizowanego wskaźnika cen konsumpcyjnych.

3.2.1. Metodyka badania budżetów gospodarstw domowych

Przedmiotem badania budżetów gospodarstw domowych jest przede wszystkim wielkość przychodów i rozchodów wszystkich członków badanego prywatnego gospodarstwa domowego, a także ilościowe spożycie wybranych produktów. To badanie prowadzone jest na próbie gospodarstw domowych wybranej metodą doboru losowego, przy zachowaniu wszystkich rygorów i wymagań metody reprezentacyjnej, w dążeniu do zapewnienia wysokiej jakości wyników badania. Stosowany jest terytorialny, warstwowy, dwustopniowy schemat losowania z różnym prawdopodobieństwem wyboru i warstwowaniem na pierwszym stopniu. Jednostką losowania pierwszego stopnia (JPS) jest rejon statystyczny, względnie zespół połączonych sąsiednich rejonów statystycznych liczących minimum 584 mieszkania w miastach i 204 na wsi, a drugiego stopnia – mieszkania. Losowanie JPS odbywa się w każdej warstwie osobno, z prawdopodobieństwami wyboru proporcjonalnymi do liczby mieszkań, z tym że w celu uzyskania lepszej precyzji ocen dla gospodarstw domowych rolników oraz gospodarstw z gmin bardziej zamożnych liczba JPS wylosowanych z warstw uznanych za rolnicze lub zamożniejsze jest większa, niż wynikałoby to z alokacji proporcjonalnej. Operat losowania JPS stanowią wykazy rejonów statystycznych opracowane na potrzeby narodowego spisu powszechnego. Jednostki są powarstwowane według podregionów, a w ramach każdego podregionu – według wielkości miejscowości. Ogółem losowane są 783 JPS. Operat losowania drugiego stopnia stanowią wykazy mieszkań w wylosowanych JPS. W każdej JPS losowane są 24 mieszkania (po dwa na każdy miesiąc) oraz próba rezerwowa licząca od 120 do 500 mieszkań. Badaniem objęte są wszystkie gospodarstwa domowe pod danym adresem. W wylosowanym mieszkaniu badanie przeprowadzane jest w danym miesiącu w dwóch kolejnych latach.

W badaniu budżetów gospodarstw domowych stosuje się metodę rotacji całkowitej miesięcznej o cyklu kwartalnym. Miesięczna rotacja gospodarstw domowych oznacza, że w każdym miesiącu w ciągu roku podejmują badanie inne gospodarstwa. Przy rotacji całkowitej wymianie podlegają wszystkie gospodarstwa domowe uczestniczące w badaniu w danym okresie. Cykl kwartalny oznacza natomiast, że po zakończeniu kwartału z badanymi gospodarstwami domowymi przeprowadzany jest dodatkowy wywiad, w czasie którego zbierane są informacje m.in. o rozchodach rzadkich i wyposażeniu w przedmioty trwałego użytkowania.

W przypadku nieprzystąpienia do badania żadnego gospodarstwa domowego z wylosowanego mieszkania dobierane jest inne mieszkanie ze specjalnie przygotowanej w tym celu rezerwowej listy mieszkań. Taką listę losuje się oddzielnie dla każdego mieszkania. Mieszkanie z próby rezerwowej dobiera się zgodnie z zasadą losowo ustalonego porządku mieszkań tak długo, aż co najmniej jedno gospodarstwo

domowe, które zamieszkuje pod wylosowanym adresem, wyrazi zgodę na uczestnictwo w badaniu.

Badanie budżetów gospodarstw domowych prowadzone jest przez ankieterów (pracowników urzędów statystycznych w poszczególnych województwach), którzy organizują badanie w terenie, utrzymują bezpośredni kontakt z badanymi gospodarstwami domowymi, udzielają niezbędnych wskazówek i wyjaśnień badanym, a następnie analizują uzyskane informacje.

Tym badaniem są objęte wszystkie społeczno-ekonomiczne grupy gospodarstw domowych. Grupa społeczno-ekonomiczna ustalana jest na podstawie przeważających dochodów osób wchodzących w skład danego gospodarstwa domowego. Wyróżnia się następujące grupy społeczno-ekonomiczne gospodarstw domowych: gospodarstwa pracowników, rolników, pracujących na własny rachunek, emerytów i rencistów oraz osób utrzymujących się z niezarobkowych źródeł.

W każdym miesiącu badanych jest ponad 1,3 tys. gospodarstw pracowników (w skali roku ok. 16,3 tys.), ponad 110 gospodarstw rolników oraz ok. 210 gospodarstw pracujących na własny rachunek (w skali roku odpowiednio ok. 1,3 tys. i 2,6 tys.), ok. 1,1 tys. gospodarstw emerytów i rencistów (w ciągu roku ok. 12,4 tys.), a także ok. 80 gospodarstw utrzymujących się z niezarobkowych źródeł (rocznie prawie 1,0 tys.). Łącznie w ciągu roku badaniem objętych jest ponad 33,5 tys. gospodarstw domowych, co stanowi ok. 0,3% gospodarstw domowych w kraju.

W badaniu budżetów gospodarstw domowych nie są ujmowane osoby lub rodziny zamieszkujące obiekty zbiorowego zakwaterowania (np. internaty czy klasztory) oraz gospodarstwa domowe, w skład których wchodzą obywatele obcego państwa pracujący w przedstawicielstwach dyplomatycznych i urzędach konsularnych. Natomiast biorą udział w badaniu gospodarstwa domowe obywateli obcych państw, którzy zamieszkują w Polsce na stałe (lub przez dłuższy okres) i posługują się językiem polskim.

Badanie budżetów gospodarstw domowych dostarcza m.in. informacji dotyczących przeciętnych miesięcznych wydatków, dochodów i spożycia na jedną osobę w gospodarstwie domowym. Strukturę poszczególnych grup wydatków oblicza się w stosunku do wydatków ogółem. Oblicza się również udział przeciętnych miesięcznych wydatków w dochodzie rozporządzalnym na osobę w gospodarstwie domowym. Wyniki tego badania okazują się przydatne m.in. w analizie poziomu i zróżnicowania warunków życia (w tym potrzeb) pomiędzy grupami społeczno-ekonomicznymi gospodarstw domowych.

Cechą istotną dla innych statystyk pod kątem dalszego wykorzystania wyników badania budżetów gospodarstw domowych jest stosowana klasyfikacja, oparta na COICOP, czyli międzynarodowej klasyfikacji służącej do agregacji danych zarówno w tym badaniu, jak i w systemie rachunków narodowych oraz opracowywaniu

wskaźników cen konsumpcyjnych. Dla celów badania budżetów gospodarstw domowych tę klasyfikację uszczegółowiono na szczeblu 6-, a nawet 7-cyfrowym (w oryginalnej COICOP najniższym szczeblem agregacji jest poziom 5-cyfrowy) i oznaczono skrótem COICOP/HBS (Classification of Individual Consumption by Purpose for Houesehold Budget Surveys – klasyfikacja spożycia indywidualnego według celu do badania budżetów gospodarstw domowych). Poszczególne poziomy COICOP/HBS odpowiadają kategoriom ECOICOP, klasyfikacji stosowanej zarówno do obliczania wskaźników cen towarów i usług konsumpcyjnych (CPI) na potrzeby krajowe, jak i zharmonizowanych wskaźników cen konsumpcyjnych (HICP) wykorzystywanych do porównań międzynarodowych w UE.

Głównym źródłem danych o wydatkach każdego badanego gospodarstwa domowego są książeczka budżetowa prowadzona w formie papierowej lub elektronicznej przez cały miesiąc oraz paragony dostarczane ankieterowi przez gospodarstwa domowe wypełniające książeczkę. Książeczkę budżetową w formie papierowej wypełnia ok. 13% gospodarstw domowych, w formie elektronicznej – ok. 4%, a ok. 83% gospodarstw przekazuje informacje o swoich wydatkach na konkretne towary i usługi konsumpcyjne w formie paragonów. Ankieter przenosi zapisy w formie papierowej i informacje z paragonów na nośnik elektroniczny oraz przeprowadza analizę uzyskanych informacji za pośrednictwem specjalnych aplikacji. Następnie dokonuje symbolizacji podanych wydatków, przyporządkowując je do grup elementarnych według COICOP/HBS. W książeczce gospodarstwo domowe zapisuje wszystkie wydatki pieniężne związane z zakupem towarów i usług konsumpcyjnych, podając ich ilość i wartość pieniężną oraz formę płatności (gotówka, karta płatnicza, przelew bankowy lub karta kredytowa), i niepieniężne (otrzymane bezpłatnie) oraz wskazuje dodatkowo, czy zakupu dokonano przez Internet.

Dodatkowe źródło informacji o wydatkach gospodarstwa domowego stanowi kwestionariusz *Informacje uzupełniające o gospodarstwie domowym w … kwartale*, w którym notowane są m.in. poniesione w danym kwartale rzadkie wydatki gospodarstwa domowego, np. na zakup mebli, pralki, lodówki, samochodu, a także związane z uprawianiem turystyki.

3.2.2. Znaczenie badania budżetów gospodarstw domowych dla badania cen detalicznych

Pozyskane z badania budżetów informacje o wydatkach gospodarstw domowych pogrupowane według COICOP/HBS, zgodną z klasyfikacją ECOICOP, stanowią podstawę opracowania systemu wag wykorzystywanego do obliczeń wskaźników cen konsumpcyjnych, które wyliczane są na wyższych poziomach agregacji jako średnie arytmetyczne ważone. Dzięki zastosowaniu systemu wag wskaźniki obliczone dla najniższego poziomu klasyfikacji mogą być agregowane na wyższe poziomy, aż do

wskaźnika cen ogółem. Ponadto na podstawie struktury wag można określić, jaki wpływ na wskaźnik ogółem w badanym okresie miała zmiana cen produktów w danej grupie elementarnej oraz które grupy w największym, a które w najmniejszym stopniu wpłynęły na wskaźnik w danym czasie. Zastosowanie wag umożliwia także obliczanie wskaźników cen dla kategorii innych niż wymienione w przyjętej do obliczeń klasyfikacji ECOICOP. Są to wskaźniki dla specjalnych agregatów (np. towary, usługi i mieszkanie), które wykorzystuje się m.in. jako deflatory do opracowywania innych statystyk w porównywalnych warunkach cenowych, a także do oceny aktualnego przebiegu i prognozowania zjawisk inflacyjnych. Uzyskane z badania budżetów informacje dotyczące zakupów przez Internet są wykorzystywane do ustalenia wag dla grup elementarnych według miejsca dokonania zakupu – w sklepach i stacjonarnych punktach usługowych oraz przez Internet.

System wag stanowią wydatki konsumpcyjne podzielone według ok. 340 grup asortymentowych pochodzących z badań budżetów gospodarstw domowych. W systemie wag przyjętym do obliczenia wskaźników cen towarów i usług konsumpcyjnych, zgodnie z międzynarodowymi zaleceniami, nie uwzględnia się takich wydatków, jak:

- podatki i opłaty celne;
- wydatki na nieruchomości i ziemię;
- wpłaty na cele społeczne, np. dobroczynne lub związane z praktykami religijnymi (ofiara na tacę), składki na różnego rodzaju fundusze;
- zakup papierów wartościowych i waluty obcej;
- kary pieniężne;
- wpłaty kaucji;
- dopłaty do delegacji służbowych;
- alimenty;
- dary ofiarowane osobom prywatnym i udzielone im pożyczki;
- koszty utrzymania młodzieży uczącej się i mieszkającej poza domem (z wyjątkiem zakwaterowania);
- spłaty pożyczek i kredytów.

Struktura spożycia w gospodarstwach domowych przyjęta jako wagi do opracowania wskaźników cen konsumpcyjnych decyduje ostatecznie o kształtowaniu się wskaźnika cen towarów i usług konsumpcyjnych ogółem. Analiza poszczególnych kategorii wydatków pozwala określić ich wpływ na wskaźnik ogółem. Przykładowo jeśli z badania budżetów wynika, że gospodarstwa domowe wydają znacznie więcej na opłaty za energię elektryczną niż na zakup sprzętu gospodarstwa domowego, wówczas zmiany cen energii mają większy wpływ na wskaźnik cen ogółem niż zmiany cen AGD. Opracowanie systemu wag z wykorzystaniem danych z badania budżetów gospodarstw domowych na potrzeby wskaźników cen, w tym w podziale regionalnym, a także według grup społeczno-ekonomicznych, jest bardzo ważnym eta-

pem przygotowywania badania cen i odbywa się według zasad i rekomendacji instytucji międzynarodowych.

Zgodnie z metodologią badań cen wagi powinny uwzględniać możliwie najbardziej aktualny wzorzec konsumpcji gospodarstw domowych, dlatego na początku każdego roku dokonuje się ich aktualizacji na podstawie danych z badania budżetów z poprzedniego roku. Z punktu widzenia badań cen istotna jest więc ciągłość badania budżetów rodzinnych i opracowanie wyników w jak najkrótszym czasie przy zapewnieniu ich odpowiedniej jakości.

3.2.3. Jakość danych oraz wyzwania statystyki publicznej w kontekście badania budżetów gospodarstw domowych

Jakość wyników badania budżetów gospodarstw domowych, które są wykorzystywane do opracowywania systemów wag, wywiera istotny wpływ na kształtowanie wskaźnika cen towarów i usług konsumpcyjnych i w dużej mierze decyduje o jego dokładności i wiarygodności. Zastosowanie metody reprezentacyjnej w badaniu budżetów pozwala na uogólnienie wyników badania, określenie ich precyzji oraz w pewnym stopniu zredukowanie ewentualnych błędów.

Jednym z największych wyzwań stojących przed badaniem budżetów gospodarstw domowych jest zapobieganie brakom odpowiedzi, które mogą wynikać zarówno z przyczyn obiektywnych (brak kontaktu z gospodarstwem, dłuższa nieobecność, choroba itd.), jak i subiektywnych (odmowa udziału w badaniu). Mimo wysiłków ankieterów zmierzających do tego, aby przekonać rodzinę do podjęcia się wypełniania książeczki lub gromadzenia i przekazywania paragonów, zdarzają się przypadki nieprzystąpienia do badania lub rezygnacji z dalszego uczestnictwa w nim. Udziału w badaniach odmawiają najczęściej mieszkańcy dużych miast. Wynika to z ogólnej niechęci do badań statystycznych i brak czasu. Szczególnie trudno dotrzeć do mieszkańców osiedli zamkniętych, ze zorganizowaną ochroną. Pracownicy ochrony często nie chcą wpuszczać na osiedle ankieterów, tym bardziej że znają oni tylko adresy i nie są w stanie powiedzieć, z kim konkretnie chcą się spotkać. Aby temu przeciwdziałać, urzędy statystyczne wysyłają pisma do spółdzielni i wspólnot mieszkaniowych z informacją o planowanym badaniu i prośbą o umożliwienie ankieterom wejścia na osiedle, jednak równie często spotykają się z odmową.

Na uczestnictwo w badaniu często nie godzą się seniorzy, osoby o wysokim statusie materialnym oraz zapracowane. Osoby starsze nie chcą wpuszczać do swoich mieszkań nieznajomych, ponieważ obawiają się o m.in. swoje bezpieczeństwo. A nawet jeśli się zgodzą na badanie, to w wielu przypadkach następnego dnia ktoś z ich rodziny wycofuje zgodę na badanie. Ludzie zamożni z kolei nie chcą ujawniać swoich dochodów i wydatków. Wyzwanie stanowią nie tylko odmowy. Z części adresów trzeba zrezygnować, ponieważ mieszkania, do których ankieterzy udają się wie-

lokrotnie, są zamknięte, gdyż wiele osób pracuje do bardzo późnych godzin. W takich sytuacjach, aby utrzymać założoną wielkość próby, dokonuje się zamiany mieszkania poprzez pobranie innego adresu z listy rezerwowej, co z jednej strony poprawia precyzję badania, ale z drugiej może prowadzić do pewnego zniekształcenia wyników ze względu na to, że nowe gospodarstwo może się charakteryzować innymi cechami demograficznymi i społeczno-ekonomicznymi.

Mimo tych przeszkód ankieterom udaje się dotrzeć do ok. 33,5 tys. wylosowanych i dobranych z próby rezerwowej mieszkań. Wprawdzie struktura próby zbadanych gospodarstw domowych może się różnić od struktury próby wylosowanej, ale przy opracowywaniu wyników badania te różnice koryguje się odpowiednio za pomocą wag, tak aby wyniki stały się reprezentatywne dla poszczególnych grup gospodarstw domowych.

Badanie budżetów gospodarstw domowych może być obciążone błędami losowymi i nielosowymi. Wielkość błędów losowych jest uzależniona od zastosowanego schematu losowania, wielkości próby, metod estymacji parametrów oraz częstości występowania badanych cech. Ich pomiar daje informację o tym, na ile uogólnione wyniki z próby mogą różnić się od wyników, które można byłoby uzyskać z pełnego badania. Błędy nielosowe natomiast, wynikające głównie z przekazania przez respondentów nieprawidłowych informacji czy niewłaściwego przygotowania ankieterów, mogą w większym stopniu niż błędy losowe wpływać na prawidłowość wyników badania. Przeciwdziała się im poprzez organizowanie efektywnych szkoleń ankieterów oraz wnikliwe analizowanie informacji przekazanych przez gospodarstwa biorące udział w badaniu. Uzyskanie wysokiej jakości (dokładności) wyników wymaga ciągłej kontroli na wszystkich etapach badania. Zarejestrowane w książeczce budżetowej dane o rozchodach i przychodach gospodarstwa, jak również ilościowym spożyciu artykułów żywnościowych analizowane są z uwzględnieniem cech demograficznych i społeczno-ekonomicznych członków gospodarstwa oraz warunków mieszkaniowych. Dużego znaczenia nabiera dobra współpraca ankietera z respondentami oraz jego zdolność kojarzenia i interpretacji faktów. Wszelkie wątpliwości dotyczące zarejestrowanych danych ankieter wyjaśnia z badaną rodziną. Może się okazać, że wykryte nieprawidłowości spowodowane są tym, że członkowie gospodarstwa domowego zapomnieli dokonać odpowiednich zapisów lub nie zrozumieli niektórych zasad wypełniania książeczki budżetowej.

Następny aspekt kontroli dotyczy prawidłowej symbolizacji rozchodów i przychodów gospodarstw domowych, zgodnej ze stosowaną w badaniu klasyfikacją. Ankieter dokonuje rejestracji danych zawartych w kwestionariuszach za pomocą specjalnie zaprojektowanych aplikacji, które dają możliwość przeanalizowania danych oraz ułatwiają symbolizację przychodów i rozchodów. W przypadku błędu aplikacja wyświetla odpowiedni komunikat. Rejestracja danego kwestionariusza nie może

zostać zakończona bez poprawienia wszystkich wykrytych błędów, akceptacji przez ankietera sytuacji wątpliwych oraz tzw. superakceptacji, której dokonuje osoba nadzorująca pracę ankietera.

Powszechnie występującym błędem nielosowym w badaniu budżetów jest zaniżanie przez gospodarstwa domowe wydatków na niektóre grupy produktów, szczególnie na alkohol, wyroby tytoniowe, a także w restauracjach. To zjawisko występuje nie tylko w Polsce. Dużym wyzwaniem dla statystyki jest odpowiednie skorygowanie tych nieścisłości. Z tego względu w przyjmowanym do obliczeń wskaźnika cen systemie wag dokonuje się korekty danych uzyskanych z badanych gospodarstw na podstawie danych makroekonomicznych oraz bieżących informacji ze statystyki handlu detalicznego. Korekta polega na podwyższeniu poziomu wydatków na te grupy towarów. Ogólna wartość wydatków z badania budżetów gospodarstw domowych, przyjmowana do obliczeń wskaźników, również korygowana jest o różnicę między oszacowanymi wydatkami ponoszonymi w restauracjach oraz na napoje alkoholowe i wyroby tytoniowe a kwotą wydatków na te artykuły wynikającą z badania budżetów.

Niespotykanym wcześniej na taką skalę wyzwaniem było prowadzenie omawianego badania w czasie zamrożenia gospodarki i ograniczenia kontaktów międzyludzkich w związku z wybuchem pandemii COVID-19. W tym czasie podstawową formą kontaktu z wytypowanymi do badania gospodarstwami domowymi było prowadzenie wywiadu telefonicznego oraz korespondencji mailowej.

Badania budżetów zapewniają dane spełniające kryteria wymagane w badaniu cen detalicznych, przy zastosowaniu formuły Laspeyresa, ale nie dostarczają żadnych informacji dotyczących bieżących zmian w strukturze konsumpcji. Przy zrównoważonym rynku i niewystępowaniu szoków podażowych i popytowych wyniki badania budżetów są całkowicie wystarczające do prowadzenia badania cen detalicznych. Niedawne doświadczenia związane z pandemią COVID-19, która spowodowała konieczność częściowego, a w przypadku niektórych branż całkowitego, zamknięcia działalności, wywołały ogólnoświatową dyskusję nad adekwatnością stosowania struktury wydatków konsumpcyjnych z roku poprzedzającego rok badany w badaniach cen konsumpcyjnych. W celu zapewnienia porównywalności danych oraz utrzymania ciągłości metodyki rekomendacje międzynarodowych organizacji i instytucji statystycznych zawierały wytyczne dotyczące stosowania ostatnich dostępnych danych z badania budżetów do opracowywania systemów wag na potrzeby badania cen, jednak zalecały ich dostosowanie (często poprzez imputację) na podstawie innych badań statystycznych do bieżącej struktury spożycia. GUS stosował się do zaleceń Eurostatu w tym zakresie.

Wspomniane doświadczenia spowodowały także znaczne przyspieszenie podjętych wiele lat wcześniej działań w kierunku rozeznania możliwości wykorzystania

danych skanowanych i skrapowanych w badaniu cen detalicznych. Na większą skalę zaczęto również badać możliwości zastosowania danych pochodzących ze źródeł administracyjnych, a także transmitowanych przez gestorów danych oraz pobieranych bezpośrednio z serwerów sieci handlowych na podstawie podpisanych z tymi sieciami porozumień.

3.3. Potencjalne obciążenia pomiaru inflacji

3.3.1. Substytucja dóbr i punktów sprzedaży

Wskaźnik cen towarów i usług konsumpcyjnych uznawany jest umownie za miarę inflacji. Aby dokonywać pomiaru w tym zakresie, należy przyjąć odpowiednie założenia, które zapewnią zarówno pożądaną jakość badania, jak i jego spójność w aspekcie tworzenia szeregów czasowych oraz możliwości dokonywania porównań. Należy jednak pamiętać, że występują czynniki, które utrudniają dokonywanie tego pomiaru.

Podstawową zasadą przyjętą w badaniu jest prowadzenie go na podstawie dobranego według ustalonych metod stałego koszyka towarów i usług – reprezentantów, jednak w praktyce nie można przewidzieć, jakie towary i usługi w danym okresie będą nabywać konsumenci. Wynika to z zachowań konsumentów, którzy uzależniają swoje decyzje zakupowe od aktualnej sytuacji na rynku, a dokładnie – od oferty punktów sprzedaży. Należy założyć, że w zależności od oferty rynkowej konsumenci wybiorą korzystniejsze dla siebie warunki zakupu danego produktu lub zamiennik danego towaru. W takiej sytuacji występuje efekt substytucji dóbr (ang. *substitution bias*). Częsta aktualizacja systemu wag, zgodna z rekomendacjami międzynarodowymi, jest jednym z narzędzi stosowanych w celu minimalizacji wpływu tego efektu na wyniki badania cen konsumpcyjnych.

Zgodnie z założeniami badania obserwacja wybranych towarów i usług powinna być prowadzona w tych samych punktach sprzedaży detalicznej. Z uwagi na możliwość wyboru korzystniejszej oferty konsumenci mogą jednak w praktyce odwiedzać inne sklepy niż te, które zostały dobrane do badania. W takim przypadku mamy do czynienia z efektem substytucji punktów sprzedaży (ang. *outlet substitution bias*). Częsta weryfikacja wytypowanych do badania cen punktów sprzedaży pozwala zmniejszyć wpływ tego efektu na wskaźnik cen.

3.3.2. Zmiany jakości produktów

Kolejnym problemem jest pojawianie się na rynku nowych towarów i usług w wyniku rozwoju technicznego i nowych technologii. Zadaniem indeksu cen konsumpcyjnych jest oddanie czystej zmiany cen, nieobciążonej wpływem innych czynników. Nieunikniona w praktyce wymiana towarów reprezentantów często wiąże się rów-

nież ze zmianą ich jakości z punktu widzenia nabywcy, więc uwzględnienie tych zmian (poprzez odjęcie ich od łącznej zmiany ceny) jest jednym z warunków poprawności indeksu. W tym celu powinny zostać uwzględnione zarówno obiektywne (techniczne) charakterystyki produktu składające się na jego jakość, jak i niektóre subiektywne, mające wpływ na zadowolenie konsumenta. W celu minimalizacji obciążenia pomiaru CPI z tytułu pojawiających się nowych dóbr – efektu nowych dóbr (ang. *new good bias*), niezbędna jest podmiana produktów na liście reprezentantów. GUS, zgodnie z zaleceniami międzynarodowymi, poza przeprowadzaniem cyklicznych badań (pomagających ocenić skalę efektu substytucji), podejmuje następujące działania w tym zakresie:

- corocznie aktualizuje system wag stosowany do obliczeń wskaźnika cen tak, aby jego struktura zapewniała właściwe odzwierciedlenie bieżących wydatków gospodarstw domowych (czyli ujmowała także zakupy towarów i usług nabywanych po niższych cenach, w nowych punktach sprzedaży oraz nowych technologicznie produktów). Polska statystyka wykorzystuje w tym celu wyniki badania budżetów gospodarstw domowych. To badanie jest przeprowadzane raz do roku, co nie zdarza się często w statystyce światowej z powodu kosztów oraz stopnia trudności badania. Niemniej jednak prowadzone w profesjonalny sposób dostarcza danych niezbędnych m.in. w badaniu cen. Jakość danych, a w efekcie także wyników badania cen, zależy w dużej mierze od rzetelności gospodarstw domowych biorących udział w badaniu (przede wszystkim od skrupulatnego wypełniania książeczek budżetowych);
- raz w roku weryfikuje listy reprezentantów towarów i usług, m.in. na podstawie obserwacji rynku przez ankieterów oraz zgłaszanych przez nich uwag i propozycji dotyczących podmiany reprezentanta w tym samym punkcie notowań lub zaproponowania notowania cen występującego na liście reprezentantów produktu w innym punkcie sprzedaży. Do weryfikacji listy towarów i usług reprezentantów wykorzystuje się także notatki gospodarstw domowych w książeczkach budżetowych. Cennych informacji dostarczyć może również statystyka handlu. Weryfikacja uwzględnia też wprowadzanie do badania nowych produktów, których ceny nie były dotychczas notowane.

W sytuacji gdy występują produkty o innych parametrach jakościowych, stosuje się bezpośrednie lub pośrednie metody szacunku jakości produktów. W metodach pierwszego typu w celu wyceny zmiany jakości produktów wykorzystuje się charakterystyki produktu, a w metodach drugiego typu – różnice lub ilorazy cen produktów uznanych za ekwiwalentne i zakłada, że różnice te wynikają z rozbieżności jakościowych. W obydwu przypadkach cena nominalna jest dzielona (jeżeli pochodzi z okresu bieżącego – najczęstszy przypadek) lub mnożona (jeżeli pochodzi z okresu bazowego) przez współczynnik uwzględniający zmiany jakościowe. Zabieg ten stosu-

je się zarówno do pojedynczych produktów, jak i ich podstawowych agregatów. W przypadku wszystkich szczegółowych metod (omówionych poniżej) można stosować ich kombinacje. Różnicuje się je również ze względu na specyfikę produktów czy agregatów na niższych poziomach.

Do metod bezpośrednich wyceny zmian jakości produktów należą:

- wycena zmiany wielkości (zawartości) opakowania – nie jest ona zmianą jakości sensu stricto, lecz postępowanie w tym przypadku jest identyczne. Metoda ta bazuje na założeniu, że użyteczność konsumenta zmienia się proporcjonalnie do wielkości produktu, co można uznać za uzasadnione w przypadku, gdy zmiana jest niewielka. Algorytm wyceny sprowadza się do podzielenia lub pomnożenia nominalnej ceny przez współczynnik równy wzrostowi lub spadkowi wielkości;
- wycena dodatkowych opcji (ang. *option pricing*) – jeżeli zmiana jakości polega na pojawieniu się w nowym produkcie różnych dodatków wcześniej niewystępujących (np. dodatkowa poduszka w samochodzie czy podzespół w komputerze), to jej wycenę można przeprowadzić, zwiększając wartość produktu o łączną wartość dodatków. Do tego celu wykorzystuje się ceny rynkowe a otrzymaną w ten sposób wielkość mnoży się przez arbitralnie ustalony współczynnik mniejszy od jedności (najczęściej równy 0,5);
- oceny eksperckie z wykorzystaniem dodatkowych informacji (ang. *supported judgmental quality adjustment*) – podobnie jak w przypadku poprzedniej metody należy wycenić dodatkowe opcje, lecz zestaw potencjalnych metod wyceny jest szerszy, jeżeli umożliwiają to dodatkowe informacje uzyskane podczas zbierania danych (można np. porównać ceny dwóch produktów różniących się tylko jedną charakterystyką – jej wycena będzie równa różnicy cen);
- regresja hedoniczna (ang. *hedonic regression*) –polega na oszacowaniu modelu regresji, w którym zmienną objaśnianą jest cena produktu (w praktycznych zastosowaniach jej logarytm naturalny), a zmiennymi objaśniającymi jego charakterystyki mające wpływ na użyteczność konsumenta. W przypadku urządzeń elektronicznych mogą to być parametry techniczne, a w przypadku żywności – zawartość substancji zarówno niekorzystnych (np. konserwantów), jak i korzystnych (np. cenne składniki naturalne) z punktu widzenia konsumenta. Niektóre urzędy statystyczne stosują regresję hedoniczną, lecz uzyskane za jej pomocą wyceny najczęściej mają charakter eksperymentalny. Eksperci Eurostatu oceniają tę metodę wyceny zmian jakościowych jako najbardziej obiektywną (z uwagi na to, że ma czysto empiryczny charakter), choć jednocześnie podkreślają teoretyczne trudności związane z jej stosowaniem.

Wśród metod pośrednich wyceny zmian jakości produktów można wskazać:

- mostkowanie (ang. *bridged overlap*) – w przypadku zniknięcia produktu z rynku lub znacznego zmniejszenia jego sprzedaży i braku możliwości znalezienia za-

miennika jako jego indywidualny indeks cen przyjmuje się indeks obliczony dla wszystkich pozostałych składników agregatu na najniższym poziomie;
- miesięczne indeksy łańcuchowe z uzupełnianiem (ang. *monthly chaining and replenishment*) – produkty reprezentanty losuje się co miesiąc ze wszystkich produktów aktualnie dostępnych. Metoda ta wymaga przyjęcia założenia, że ceny podobnych, z konsumenckiego punktu widzenia, dóbr zmieniają się proporcjonalnie. Oznacza to, że zmiana cen w dostępnych połączeniach (produkt jest dostępny w obu okresach) to czysta zmiana ceny a pozostały wzrost wynika ze zmiany jakości;
- imputacja ceny bazowej (ang. *backcasting*) – to wariant metody mostkowania; cena pochodzi z okresu bazowego całego okresu sprawozdawczego (najczęściej grudzień poprzedniego roku) a cena bieżąca dotyczy dowolnego miesiąca. Metoda ta charakteryzuje się niższą precyzją oszacowań niż w przypadku porównywania dwóch sąsiednich miesięcy.

3.3.3. Obecnie stosowane formuły obliczeń wskaźników cen

Zasady obliczania indeksów cen detalicznych towarów i usług różnią się w zależności od źródeł danych o poziomach cen produktów reprezentantów. Gdy dane uzyskiwane są na podstawie notowań cen towarów i usług reprezentantów przez ankieterów, przez Internet oraz na bazie centralnych notowań cen z cenników, zarządzeń i decyzji w zakresie cen jednolitych, punktem wyjścia do obliczeń indeksów cen są średnie miesięczne ceny produktów reprezentantów. Przeciętne ceny produktów reprezentantów obliczane są na podstawie następujących algorytmów:
- gdy informacje o cenach produktów reprezentantów uzyskiwane są wyłącznie na podstawie notowań ankieterów w rejonach notowań, przeciętny miesięczny poziom ceny produktu reprezentanta w rejonie obliczany jest jako średnia arytmetyczna z uwzględnieniem liczby notowań. W praktyce tylko ceny reprezentantów warzyw i owoców notowane są dwa razy w miesiącu i ich przeciętne miesięczne ceny w rejonie są obliczane według wspomnianej formuły średniej arytmetycznej. Cena pozostałych produktów reprezentantów jest notowana w danym rejonie tylko raz i stanowi przeciętną miesięczną cenę tego reprezentanta w rejonie;
- gdy informacje o cenach produktów reprezentantów są uzyskiwane na podstawie notowań ankieterów i notowań centralnych przez Internet, to do obliczeń ich indeksów wykorzystuje się zarówno przeciętne ceny tych produktów obliczane dla każdego punktu notowań, jak i uzyskiwane z notowań internetowych;
- w przypadku cen jednolitych oraz taryf przeciętne miesięczne ceny detaliczne produktów reprezentantów obliczane są jako średnie arytmetyczne ważone liczbą dni obowiązywania poszczególnych cen.

Gdy dane do obliczeń indeksów cen detalicznych produktów z danej grupy elementarnej pochodzą od gestorów danych (np. stacje paliw czy zakłady ubezpieczeń), to ze względu na ich niejednolity charakter (dotyczą zarówno produktów reprezentantów, jak i grup asortymentowych w ramach danej grupy elementarnej) nie są obliczane dla tych grup przeciętne ceny ich produktów reprezentantów, lecz indeksy cen na poziomie całego kraju. Krajowe indeksy cen dla pozostałych grup elementarnych obliczane są na bazie wcześniej wyliczonych przeciętnych miesięcznych cen produktów reprezentantów z użyciem następujących algorytmów:

- w przypadku wykorzystywania informacji pochodzących wyłącznie z notowań cen produktów reprezentantów dokonywanych przez ankieterów w pierwszym kroku obliczany jest indeks cen reprezentanta w rejonie (jako stosunek średniej ceny w danym miesiącu w rejonie do jego średniej ceny z okresu bazowego). Ogólnopolskie indeksy cen produktów reprezentantów obliczane są jako średnie geometryczne wskaźników cen produktów reprezentantów ze wszystkich rejonów;
- w przypadku wykorzystywania informacji pochodzących zarówno z notowań cen produktów reprezentantów dokonywanych przez ankieterów, jak i z notowań prowadzonych centralnie za pośrednictwem Internetu w pierwszym kroku obliczana jest przeciętna miesięczna cena danego produktu reprezentanta jako średnia arytmetyczna z przeciętnych cen danego reprezentanta na poziomie kraju wyliczanych na podstawie dwóch wspomnianych źródeł informacji. Wagami dla tych dwóch przeciętnych cen są udziały wydatków gospodarstw domowych na danego reprezentanta według dwóch źródeł zakupu (poza sklepami internetowymi oraz w takich właśnie sklepach) w ich wydatkach na danego reprezentanta ogółem, uzyskiwanych z książeczek budżetowych. W kolejnym kroku oblicza się ogólnokrajowe indeksy cen tych produktów reprezentantów.

Indeksy cen detalicznych towarów i usług konsumpcyjnych na kolejnych wyższych poziomach agregacji obliczane są jako średnia arytmetyczna ważona z indeksów cen obliczanych dla wszystkich wskaźników cen na bezpośrednio niższym poziomie agregacji. System wag bazuje na rocznej strukturze wydatków gospodarstw domowych przeznaczonych na zakup towarów i usług z roku poprzedzającego rok badany co oznacza, że w obliczeniach stosuje się formułę Laspeyresa.

Miesięczne indeksy cen detalicznych obliczane przy w różnych podstawach: grudzień poprzedniego roku, poprzedni miesiąc, analogiczny miesiąc poprzedniego roku, analogiczny okres narastający poprzedniego roku.

Mocne i słabe cechy poszczególnych formuł obliczania wskaźników cen, zarówno na poziomie grupy elementarnej, jak i wyższych poziomach agregacji, a także ich potencjalnego obciążenia dla pomiaru inflacji, omówione są w podrozdz. 2.5. i 4.5. Szczegółowy opis zagadnienia wyboru formuły obliczeń można znaleźć w wielu innych źródłach (m.in. Białek, 2014).

3.3.4. Obecnie wykorzystywane aplikacje do zbierania i przetwarzania danych

W ostatnich dwóch pandemicznych latach zmieniły się metody uzyskiwania danych o cenach na potrzeby badania. Przed 2019 r. podstawowe źródło informacji o cenach detalicznych towarów i usług konsumpcyjnych stanowiły notowania cen dokonywane przez ankieterów. Dodatkowo pracownicy urzędów statystycznych prowadzili notowania centralne na podstawie cenników, zarządzeń i decyzji dotyczących cen jednolitych obowiązujących na terenie całego kraju lub jego części wydawanych przez organy administracji rządowej i organy jednostek samorządu terytorialnego oraz podmioty prowadzące działalność gospodarczą. W mniejszym stopniu, jedynie dla wybranych grup produktów, wykorzystywane były ceny uzyskiwane przez pracowników urzędów statystycznych dla towarów i usług nabywanych przez Internet oraz dane z systemów informatycznych gestorów danych. W ten sposób miesięcznie gromadzonych było ok. 350 tys. danych dotyczących cen, w tym ok. 8 tys. notowanych centralnie (pozostałe trzy źródła informacji).

Po wybuchu pandemii oraz wprowadzeniu czasowych ograniczeń pracy ankieterów w terenie stopniowo zaczęto rozszerzać zakres danych o cenach pozyskiwanych z innych źródeł. Do badania zostały włączone m.in. rozpoznane wcześniej podczas prac eksperymentalnych informacje, które są regularnie przekazywane na potrzeby badań statystycznych z sieci handlowych (dane skanowane), oraz informacje automatycznie zbierane z wybranych stron internetowych (dane skrapowane).

Pomimo wdrażania nowych źródeł monitorowanie cen przez ankieterów bezpośrednio w sklepach pozostaje ważną częścią badania. Wynika to zarówno z konieczności kontrolowania danych pozyskiwanych z nowych źródeł, jak i z przeprowadzanych analiz dotyczących udziału wydatków ponoszonych przez konsumentów w różnych punktach sprzedaży (sklepy stacjonarne i Internet). Dostępne informacje wskazują, że nadal dużej części zakupów nie dokonuje się w sieciach handlowych ani przez Internet. Biorąc pod uwagę, że udział zakupów stacjonarnych (poza dużymi sieciami handlowymi) jest nadal znaczący, utrzymanie tradycyjnych notowań cen dla określonych grup towarów i usług wydaje się niezbędne. Poza tym uzyskanie danych o aktualnej cenie niektórych produktów jest możliwe jedynie dzięki kontaktowi ankietera ze sprzedawcą. Warto też zwrócić uwagę, że utrzymanie tradycyjnego sposobu zbierania danych o cenach umożliwi ocenę różnic pomiędzy informacjami pozyskiwanymi z poszczególnych kanałów dystrybucji i miejsc dokonywania zakupów. Niewątpliwie jednak dostęp do nowych źródeł danych ma wiele korzyści, przede wszystkim niski koszt ich uzyskania. Wiąże się to z możliwością ograniczenia pracy ankieterów i przekierowania ich do notowań cen w mniejszych sklepach. Nie będzie to jednak korzystne dla dużych sieci handlowych, na które nakładane są kolejne obowiązki.

3.4. Analiza możliwości zwiększenia liczby i zróżnicowania źródeł danych adekwatnie do specyfiki rynku detalicznego w Polsce

Główne elementy determinujące włączanie do obserwacji ruchów cen alternatywnych źródeł danych są następujące:

- Włączenie danych uzyskanych z nowych źródeł wymaga stworzenia odpowiedniego środowiska informatycznego do ich przetwarzania, zapewniającego właściwe narzędzia do analizy danych skanowanych z wykorzystaniem zaawansowanych metod i technik. Ogromny wolumen i różnorodność danych pochodzących ze źródeł alternatywnych powoduje, że ich przygotowanie do dalszych etapów obliczeń, czyli filtrowanie, dopasowywanie produktów i klasyfikowanie, wymaga opracowania niezawodnych algorytmów oraz zapewnienia odpowiedniej mocy serwerów wykorzystywanych do tych obliczeń.
- Automatyczne pozyskiwanie i przetwarzanie danych z nowych źródeł daje możliwość obniżenia kosztów badania. W procesie wdrażania informacji z sieci dąży się do maksymalnej automatyzacji procesu przetwarzania i analizowania danych skanowanych, a ich obsługa manualna powinna zostać sprowadzona do niezbędnego minimum (najczęściej dotyczy bardzo kłopotliwej klasyfikacji produktu lub problemów ze znalezieniem zamiennika). W związku z tym, że wizyta ankietera w punkcie notowań staje się w tym przypadku praktycznie zbędna (np. jego obecność w supermarkecie może się jedynie sprowadzać do sporadycznych wizyt kontrolnych czy związanych z dokumentacją i legislacją procesu przysyłania danych skanowanych), to koszty samego badania ulegają znacznemu obniżeniu.
- Rozmiar zbiorów danych pozyskiwanych z alternatywnych źródeł jest nieporównywalnie większy od rozmiaru baz danych tworzonych w wyniku tradycyjnego pomiaru CPI i sięga dziesiątek tysięcy transakcji rocznie oraz wielomilionowych wartości sprzedaży. Wolumen transakcji sprzedaży w przypadku danych skanowanych jest wielokrotnie większy niż ten, który uwzględniamy w przypadku tradycyjnego sposobu zbierania danych, i co ważniejsze, otrzymujemy dane o ilości sprzedaży na najniższych poziomach agregacji.
- Dane skanowane, poza kodami kreskowymi umożliwiającymi identyfikację produktu, zawierają często jego drobiazgowy opis i charakterystykę (np. kolor, rozmiar, masę czy zawartość tłuszczu). Umożliwia to zaklasyfikowanie produktów do homogenicznych podgrup, które znajdują się poniżej dostępnego obecnie najniższego poziomu agregacji danych (COICOP 5 czy nawet COICOP 6). Zejście do tak niskich poziomów agregacji, przy jednocześnie dostępnej informacji o wolumenie sprzedaży, musi prowadzić do dokładniejszego pomiaru dynamiki cen produktów reprezentantów i grup elementarnych.
- W przypadku danych skanowanych, nawet na najniższych poziomach agregacji (niższych niż COICOP 6), dysponujemy informacją zarówno o cenach sprzeda-

wanych dóbr, jak i ilościach i/lub wartościach ich sprzedaży. Pozwala to na zastosowanie ważonych formuł indeksów cen, które mają znacznie lepsze umocowanie ekonomiczne w pomiarze inflacji niż indeksy nieważone. Tymczasem w tradycyjnym pomiarze CPI, na najniższych poziomach agregacji, gdzie dysponujemy jedynie informacjami o cenach dóbr, jedynymi możliwymi do zastosowania formułami indeksów są formuły nieważone. Dane skanowane dają więc znacznie więcej możliwości w zakresie dostępnych formuł indeksów i tym samym sposobów ważenia dostępnych informacji.

3.5. Analiza możliwości włączenia do badania nowych źródeł danych

W związku ze zmieniającymi się uwarunkowaniami rynku, zwiększoną dostępnością nowych źródeł danych, a także koniecznością adaptacji do nowych zaleceń międzynarodowych badanie cen konsumpcyjnych wymaga wprowadzania udoskonaleń i modyfikacji. Ogromnym wyzwaniem jest m.in. dywersyfikacja źródeł danych i automatyzacja ich przetwarzania. Podjęcie się tego zadania skutkuje wieloma zmianami w procesie realizacji badania cen, w tym następujących zagadnień:

- Zastosowanie nowych technologii gromadzenia i przetwarzania informacji pozwala na niemal nieograniczone zwiększanie dostępności do poziomów cen i ilości zakupionych przez konsumentów towarów i usług. W przyszłości mogą to być olbrzymie wolumeny i duża częstotliwość dostarczania danych.
- Dane z sieci handlowych (dane skanowane) stwarzają nowe możliwości pomiarowe. Obecnie do monitorowania zmian cen stosowany jest koszyk stały w ciągu badanego roku, natomiast dane z sieci handlowych dotyczą bieżących zakupów (ceny i ilości).
- Możliwość zastosowania metod uczenia maszynowego, w szczególności metod rozpoznawania i analizy tekstu. Mogą być one również wykorzystywane w procesie automatycznego dopasowywania produktów polegającym na odnajdywaniu w zbiorze danych odpowiadających sobie produktów obserwowanych w różnych okresach.
- Prace eksperymentalne dotyczące obliczania wskaźnika cen z wykorzystaniem danych skanowanych obejmują także prace nad nowymi, alternatywnymi, formułami indeksów cen i ich testowanie.
- Ustalenie schematu i systemu wag, który umożliwi przechodzenie do wyższych poziomów agregacji i łączenie wskaźników oszacowanych na bazie alternatywnych źródeł danych do finalnie publikowanej informacji o poziomie inflacji.

Podsumowując, zdobyte doświadczenia wskazują na możliwość poszerzenia badania cen konsumpcyjnych o dane z sieci. Niezbędne wydaje się jednak podjęcie dalszych intensywnych prac, które by umożliwiły (lub zagwarantowały) wykorzystanie danych skanowanych w regularnej produkcji:

- nawiązanie stałej współpracy z sieciami w celu uzyskania gwarancji otrzymania danych w każdym miesiącu, maksymalnej redukcji błędów w zbiorach, poprawy terminowości transmisji danych, a także dostosowania do rekomendacji (np. w zakresie otrzymywania danych tygodniowych);
- zwiększanie wykorzystania danych skanowanych zarówno poprzez włączanie danych z kolejnych sieci, jak i poszerzenie zakresu grup produktów (żywnościowych i nieżywnościowych);
- opracowanie informacji zwrotnej skierowanej do sieci handlowych;
- kontynuacja prac dotyczących eksperymentalnych obliczeń subindeksów cen i integracji danych z różnych źródeł przy założeniu utrzymania notowań ankieterów w wybranych obszarach cen konsumpcyjnych (np. żywność);
- zapewnienie wsparcia IT, utworzenie maksymalnie zautomatyzowanego i skalowalnego systemu przetwarzania danych skanowanych, co stanowi duże wyzwanie w kontekście zróżnicowanego zakresu danych;
- monitorowanie rynku sprzedaży detalicznej w celu jak najlepszego odzwierciedlenia zmian zachodzących w badaniu cen (np. informacje na temat pogarszającej się sytuacji w ciągłości dostaw lub zamknięcie sieci detalicznej);
- zwiększenie możliwości wieloaspektowego wykorzystania danych z sieci, w tym w celu redukcji obowiązków sprawozdawczych respondentów;
- doprecyzowanie krajowych wymogów prawnych wzorem rozwiązań zastosowanych w innych krajach.

3.6. Kierunki działań w celu modernizacji systemu statystyki cen detalicznych

Prowadzenie obliczeń wskaźników CPI wymaga stałego przeglądu i doskonalenia wszystkich etapów badania, w tym identyfikacji potencjalnych nowych źródeł danych dla poszczególnych kategorii klasyfikacji stosowanej w badaniu i metod ich przetwarzania. Wszelkie udoskonalenia powinny być prowadzone w ramach określonych standardów jakości obowiązujących w statystyce publicznej. Z punktu widzenia oczekiwań użytkowników danych wynikowych z zakresu statystyki cen szczególną uwagę w ocenie przydatności źródeł danych zwraca się na terminowość i częstotliwość danych. Planując jakiekolwiek udoskonalenia, należy mieć na uwadze bardzo rygorystyczne wymogi dotyczące harmonogramu publikacji danych. Wstępne dane wynikowe za badany miesiąc są publikowane już pod koniec badanego miesiąca (lub w pierwszych dniach kolejnego). Udostępniane są wprawdzie dane zagregowane, natomiast zgodnie z przyjętą metodyką szybkie szacunki obliczane są w pełnym układzie, czyli dla wszystkich grup klasyfikacyjnych: od grup elementarnych aż do wskaźnika ogółem. Z tego powodu wszystkie potencjalne źródła danych należy oceniać z punktu widzenia ich dostępności w czasie umożliwiającym ich prze-

tworzenie i włączenie również do obliczeń tzw. szybkich szacunków. Harmonogram publikacji danych determinuje zatem prace dotyczące np. rozbudowy wykorzystania danych.

Dodatkowym rygorystycznym wymogiem odnoszącym się do danych wynikowych z zakresu CPI jest brak możliwości ich rewizji. Jedyny przypadek, w którym dane CPI są rewidowane, to dane za styczeń każdego roku. Taka praktyka jest podyktowana dostępnością wag pochodzących z badania budżetów gospodarstw domowych, to rewizja planowa i komunikowana z wyprzedzeniem użytkownikom danych. W podstawach prawnych HICP przewidziana jest możliwość korekty błędów i rewizji opublikowanych danych. W procesie produkcji wskaźników CPI nie przewiduje się takiej możliwości, więc przetworzenie wszelkich danych wejściowych musi zapewnić eliminację ryzyka wystąpienia błędu.

Dane ostateczne w zakresie CPI za badany miesiąc są publikowane w połowie kolejnego miesiąca. Taki cykl publikacji danych znacznie ogranicza zwiększenie wykorzystania źródeł administracyjnych na etapie bieżącego opracowania wskaźnika cen. Dane administracyjne nie są jeszcze wykorzystywane w satysfakcjonującym zakresie jako bezpośrednie źródło danych cenowych dla wskaźników cen konsumpcyjnych.

Oczekiwanym kierunkiem działań modernizacyjnych jest stopniowe włączanie dostępnych różnorodnych źródeł danych na kolejnym etapie produkcji wskaźników CPI. Przykładowo nowe alternatywne źródła danych mogą stać się cennym źródłem informacji dla doskonalenia notowań cen przeprowadzanych obecnie metodą tradycyjną. Dane z wybranych jednostek handlowych wykorzystuje się do weryfikacji poprawności notowań prowadzonych przez ankieterów, np. w zakresie oceny właściwego doboru reprezentantów, dodatkowych opisów, ceny, gramatury, występowania i rodzajów promocji, zmian w obowiązujących stawkach VAT.

Dane skanowane są również wykorzystywane podczas corocznie dokonywanej we współpracy z urzędami statystycznymi weryfikacji próby do badania cen na kolejny rok. Propozycje zgłaszane przez urzędy są konfrontowane z danymi otrzymywanymi z sieci m.in. pod kątem reprezentatywności, możliwości doprecyzowania opisu i jednostki miary. Z zastosowaniem danych skanowanych prowadzone są także pewne obserwacje dotyczące porównywania polityki cenowej sieci detalicznych prowadzonej w różnych kanałach sprzedaży – w sklepach stacjonarnych oraz online.

Główne działania zmierzające do coraz szerszego wykorzystania danych ze źródeł alternatywnych (tu: danych skanowanych) powinny być skoncentrowane m.in. na:

- zintensyfikowaniu współpracy z sieciami w celu maksymalnej redukcji błędów w otrzymywanych zbiorach, poprawy terminowości transmisji danych, a także dostosowania do rekomendacji;

- kontynuacji prac w zakresie opracowywania założeń doboru próby i prowadzeniu eksperymentalnych obliczeń subindeksów cen;
- monitorowaniu rynku sprzedaży detalicznej w celu jak najlepszego odzwierciedlenia zachodzących na nim zmian w próbie CPI (np. informacje na temat pogarszającej się sytuacji lub zamknięcia sieci detalicznej). Z jednej strony obecne warunki ekonomiczne utrudniają proces nawiązywania i utrzymywania pozytywnych relacji z sieciami handlowymi, a z drugiej – powodują niechęć do współpracy z ankieterem i ograniczenie możliwości pozyskania danych w terenie,
- zwiększeniu możliwości wieloaspektowego wykorzystania danych z sieci.

Innym obszarem prac jest monitoring rynku i możliwości uzyskania danych w terenie pod kątem oceny zasadności i możliwości wykorzystania danych transakcyjnych innych niż skanowane. Szczególnie trudnym obszarem pomiaru zmian cen jest sektor usług. Poszukuje się źródeł danych transakcyjnych innych niż dane skanowane, które mogą być wykorzystywane podczas profilowania usług objętych obserwacją cen.

Wykorzystanie danych pochodzących z nowych źródeł wymaga wielu prac przedwdrożeniowych, np. na ograniczonej próbie. Pomimo starań zmierzających do maksymalizacji ujednolicenia zestawów danych (zakres, format itd.) praktyka wskazuje, że struktury zbiorów otrzymywanych od ich gestorów znacznie się różnią. Wymaga to indywidualnego podejścia do włączania danych z poszczególnych sieci do obliczeń i opracowania koncepcji zarówno integracji danych od poszczególnych gestorów, jak i integracji danych ze źródeł alternatywnych z danymi pozyskiwanymi w tradycyjnym notowaniu. Prace w tym zakresie powinny identyfikować i wykorzystywać elementy systemów poszczególnych sieci, które są w jakimś stopniu do siebie zbliżone. Włączanie kolejnych sieci może się zatem okazać nieco łatwiejsze, ponieważ będzie można bazować na wypracowanych wcześniej ścieżkach współpracy z siecią i organizacji transferu danych.

Dane z nowych źródeł muszą być zintegrowane z tradycyjnie zbieranymi informacjami o cenach. Metoda agregacji zależy od zakresu pozyskanych danych. Dane mogą pokrywać całość wydatków grupy elementarnej COICOP lub stanowić uzupełniające źródło danych. Przykładowo w przypadku sieci handlowych przyjęto założenie, że do momentu włączenia wszystkich znaczących sieci handlowych, przy monitorowaniu zmian zachodzących na rynku dane skanowane będą stanowiły uzupełnienie notowań wykonywanych metodą tradycyjną. W związku z tym agregacja danych z sieci z danymi pochodzącymi z tradycyjnego notowania będzie się odbywała na poziomie elementarnym z wykorzystaniem informacji na temat udziałów w rynku poszczególnych sieci. Z uwagi na brak danych z tego zakresu z badania budżetów gospodarstw domowych udziały będą bazować na danych ze sprzedaży

detalicznej (nie będzie zatem możliwe określenie udziału w sprzedaży dla kategorii wyszczególnionych w klasyfikacji COICOP).

Metodologia wykorzystania danych z różnych źródeł w statystyce cen detalicznych jest ciągle rozwijana. W polskich warunkach wykorzystanie alternatywnych danych zwiększa się zarówno dzięki włączaniu danych z kolejnych sieci, jak i poszerzaniu zakresu grup produktów.

ROZDZIAŁ 4
Metodyka badania cen detalicznych z wykorzystaniem alternatywnych źródeł danych

4.1. Pozyskiwanie danych skanowanych i skrapowanych

Sposoby pozyskiwania danych skanowanych i skrapowanych są zasadniczo różne, ponieważ te pierwsze wymagają zawarcia porozumienia z sieciami handlowymi, te drugie zaś mogą swoim zasięgiem obejmować nie tylko sieci handlowe, ale również mniejsze markety, sklepy czy tylko elektronicznych wystawców produktów (sklepy internetowe). Potencjalnym dostawcą danych skrapowanych może być każda firma posiadająca stronę internetową z umieszczonym na niej cennikiem oferowanych towarów i usług. Kluczowe jednak jest wybranie tych podmiotów, które oferują jak najbardziej reprezentatywne dane dla potrzeb obliczania wskaźników cen towarów i usług konsumpcyjnych. Sam fakt posiadania szerokiej oferty produktowej nie oznacza rzeczywistych zakupów dokonywanych przez klientów. Ponieważ projekt INSTATCENY ogranicza się do kilku sieci handlowych co jest uzasadnione ich znacznymi udziałami sprzedaży na rynku, poniżej dokonano charakterystyki pozyskiwania danych skanowanych i skrapowanych od tego dostawców danych.

4.1.1. Współpraca z sieciami handlowymi

Pozyskiwanie danych skrapowanych z sieci handlowych wydaje się prostsze w konfrontacji z pozyskiwaniem danych skanowanych. Należy tu podkreślić, że gromadzenia danych ze stron internetowych sieci handlowych najlepiej jest dokonywać na podstawie źródła pierwotnego, a nie stron podmiotów trzecich, takich jak porównywarki cenowe czy agregatory cen. Co do zasady, dostęp do danych cenowych w sieci internetowej jest bezpłatny, jako że podmiotom (właścicielom stron internetowych) zależy na rozpowszechnianiu informacji zarówno o swojej ofercie, jak i cennikach. Należy jednak pamiętać, że podmiot może zablokować dostęp do swojej strony robotom przeszukującym Internet w poszukiwaniu informacji ze względu na często znaczne obciążenie witryn sklepowych procesem pobierania danych. Właściciel sieci

może zablokować dostęp do strony internetowej dla wybranych adresów IP, może zażądać zaniechania pobierania danych na drodze sądowej, jak również może wykorzystać protokół *robots.txt*, czyli oflagowanie strony internetowej jako tej, która nie może być przeglądana przez automatyczne skrypty, czyli scrapery. Aby zminimalizować ryzyko tego rodzaju blokady, w GUS przyjęto dwie fundamentalne zasady:

1. Przed rozpoczęciem skrapowania danych z witryn internetowych sieci handlowych wysyłano pismo do właścicieli sieci, które informowało o takim zamiarze i uzasadniało jego celowość.
2. Dane skrapowano tylko w nocy, tak aby ewentualne obciążenie witryn supermarketów nie wpływało na ich dzienną, normalną funkcjonalność.

Jeśli chodzi o dane skanowane, to proces pobierania danych musiało poprzedzić zawarcie porozumienia – w przypadku projektu INSTATCENY między GUS a siecią handlową. Praktyka GUS, ale też większości urzędów statystycznych w Europie, wskazuje, że od momentu przystąpienia do rozmów do finalnego porozumienia z siecią najczęściej mija od sześciu do ośmiu miesięcy, choć czasem okres ten może się znaczenie przedłużyć (np. jeśli nastąpiły zmiany personalne w zarządzie sieci). Porozumienie z siecią z jednej strony ma prawne podstawy[1], z drugiej strony, stosownie do art. 5 ust. 1a Ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej, „dostęp służb statystyki publicznej do danych oraz przekazywanie tym służbom danych następuje nieodpłatnie". Brak finansowej gratyfikacji i obawa przed potencjalną utratą konkurencyjności na wypadek ewentualnego wycieku danych o sprzedaży to dwa główne powody, dla których sieci bardzo ostrożnie przystępują do tego rodzaju rozmów. Porozumienie z siecią handlową musi być zatem tak skonstruowane, aby dawało obu jego stronom poczucie celowości i pożytku ze współpracy, a jednocześnie gwarantowało bezpieczeństwo przesyłanych danych.

Typowe porozumienie z siecią handlową określa zobowiązania dostawcy i odbiorcy danych. Z jednej strony dostawca (sieć) najczęściej deklaruje, że będzie dostarczał dane dotyczące określonego asortymentu produktów (np. produkty spożywcze czy kosmetyczne albo całość asortymentu) w określonym terminie (np. na koniec miesiąca) i za określony przedział danego miesiąca (np. za trzy pierwsze tygodnie). Precyzuje formę przekazywania danych (transfer elektroniczny lub API) i najczęściej wskazuje, że są to dane wrażliwe i poufne, niezawierające danych osobowych. Z kolei odbiorca (GUS) zaznacza, że otrzymywane dane będą wykorzystywane wyłącznie w celu obliczania wskaźników cen towarów i usług konsumpcyjnych i na potrzeby analiz statystycznych, ewentualnie na potrzeby prezentowania i publikowania wyni-

[1] Artykuł 5 Rozporządzenia Parlamentu Europejskiego i Rady UE (2016/792) z dnia 11 maja 2016 r. w sprawie zharmonizowanych wskaźników cen konsumpcyjnych oraz wskaźnika cen nieruchomości mieszkalnych i uchylające rozporządzenie Rady (WE) nr 2494/95 nakłada na jednostki statystyczne obowiązek przekazywania danych dotyczących transakcji do krajowych organów odpowiedzialnych za opracowywanie zharmonizowanych wskaźników.

ków badań w formie wyłącznie zagregowanej. Zakres przekazywanych danych najczęściej doprecyzowuje załącznik do porozumienia, który również określa format przekazywanych danych (np. CSV).

Czynnikiem ograniczającym pobieranie danych skrapowanych i skanowanych, zwłaszcza w pierwszej fazie projektu, jest konieczność zapewnienia niezbędnej wiedzy osób (wykształconych programistów oraz analityków), które będą dokonywać selekcji informacji. Dodatkowo wymagane jest zabezpieczenie infrastruktury technicznej – komputerów i/lub chmury obliczeniowej, zapewnienie bezpieczeństwa pobranych danych, a także przestrzeni dyskowej do przechowywania i archiwizowania dużej ilości danych.

4.1.2. Bezpieczny transfer danych

Jak już wspomniano, pobieranie danych skanowanych z sieci handlowych może odbywać się na dwa sposoby. Sposób wygodny i gwarantujący dostęp do całych zasobów sieci to udostępnienie przez sieć handlową swojego API. Jest to jednak forma, która z jednej strony generuje zespołowi IT po stronie sieci mniej pracy (organizacją przepływu danych zajmuje się zespół IT po stronie GUS), z drugiej zaś strony daje sieci mniejszą kontrolę nad tym, co jest pobierane. Z tego powodu część sieci decyduje się jednak dostarczać dane w wersji elektronicznej. Dzięki temu rozwiązaniu sieć może precyzyjniej ustalać, które konkretnie kolumny zbioru danych (zmienne) są przekazywanie (oczywiście w zgodzie z zawartym porozumieniem) oraz ewentualnie anonimizować wybrane pozycje tak, aby nie wpływać na uzyskane wartości wskaźników cen. Przykładowo jedna z sieci, z którymi GUS nawiązał współpracę, zdecydowała się przeskalowywać ceny i ilość sprzedawanych produktów tak, aby zapewnić sobie jeszcze większe bezpieczeństwo przekazywanych danych i nie wpływać na skalę obrotów i system ważenia wskaźników cen, które pozostały w ten sposób niezmienne.

Tylko jedna ze współpracujących z GUS sieci zdecydowała się na rozwiązanie w postaci API. Pozostałe sieci dostarczają dane w wersji elektronicznej, co oczywiście regulują zawarte porozumienia. Tego rodzaju dane przekazywane są wówczas za pośrednictwem bezpiecznego, szyfrowanego kanału komunikacyjnego TransGUS[2] raz w miesiącu. Szyfrowany kanał transferowy chroni interesy sieci handlowej, która najczęściej ma obawy dotyczące ewentualnego wycieku danych na zewnątrz i osłabienia w ten sposób swojej konkurencyjności, ale jest też swoistym wymogiem po stronie GUS, ponieważ pozyskiwane dane jednostkowe stanowią tajemnicę statystyczną zgodnie z art. 10 Ustawy z dnia 29 czerwca 1995 r. o statystyce publicznej

[2] Zbiory są przesyłane w kanale bezpiecznym z wykorzystaniem protokołu TLS 1.2 z zastosowaniem szyfrowania kluczem 256 bitów i z możliwością wskazania IP uprawnionych komputerów.

i podlegają ochronie. Dane pobierane za pomocą bezpiecznego kanału TransGUS[3] są następnie archiwizowane na serwerach GUS, do których dostęp również jest ograniczony i zabezpieczony. Ostatecznie uwierzytelniany dostęp do danych po stronie GUS uzyskują tylko pracownicy bezpośrednio zaangażowani w proces analizy danych skanowanych, przy czym ponoszą oni pełną odpowiedzialność (np. finansową) za ewentualne udostępnienie danych podmiotom czy osobom trzecim. Więcej technicznych informacji na temat zapewnienia bezpieczeństwa transferu danych znajduje się w podrozdz. 5.2.1.

4.1.3. Opracowanie i przetwarzanie zbiorów danych otrzymanych z sieci handlowych

Pliki CSV, które jakie GUS pobiera każdego miesiąca z sieci handlowych, różnią się między sobą zarówno pod względem zakresu przekazywanych informacji o produktach (liczba kolumn zbioru danych), jak i pod względem liczby uwzględnionych elementarnych grup produktów i ich różnorodności (liczba wierszy). Przykładowo jedna z sieci dostarcza każdego miesiąca dane transakcyjne dotyczące wszystkich oferowanych przez nią produktów, a z kolei inna zawęża zakres przesyłanych danych do 10 grup elementarnych z kategorii spożywczej. W przypadku tej pierwszej sieci GUS otrzymuje miesięcznie ok. 650 MB danych (5 661 586 rekordów, 148 290 różnych produktów), druga z sieci dostarcza już tylko nieco ponad 40 MB danych miesięcznie (364 950 rekordów, 2565 różnych produktów). Ponadto jedna z sieci w dostarczanej bazie danych uwzględnia osobno kolumny, w których zawarta jest informacja o jednostce sprzedaży produktu i jego gramaturze, z kolei w przypadku pozostałych sieci informacja ta znajduje się w opisie produktu i należy ją odpowiednio wydobyć. Podobną uwagę można poczynić dla danych skrapowanych, które najczęściej jednak pobierane są w formacie JSON, a następnie – jeśli jest taka potrzeba – pliki te konwertowane są do formatu CSV.

Zasadniczo pobrana z sieci baza danych poddawana jest na początku etapowi oczyszczenia, tzn. usuwane są z niej rekordy niekompletne, niewłaściwe (np. zawierające zerowe ceny lub ujemne ilości) lub duplikujące się. Następnie ujednolicony zostaje format kolumn (typy zmiennych i ich nazwy) oraz, jeśli zachodzi taka potrzeba, dokonywana jest ekstrakcja informacji o jednostce sprzedaży i gramaturze z etykiet produktów. Tak przygotowana ramka danych przechodzi do etapu klasyfikacji produktów. (ten etap zostanie dokładnie omówiony w podrozdz. 4.2). Warto nadmienić, że dzięki niemu główny plik z danymi zostaje podzielony na mniejsze pliki, które odpowiadają poszczególnym, elementarnym grupom produktów. W nowo utworzonych bazach danych pojawia się dodatkowa kolumna, która zawiera numer grupy COICOP 5 oraz jej podgrupy. Następnie aktywowany jest proces

[3] System umożliwia obsługę dowolnego formatu danych źródłowych, preferowane to XML, CSV, TXT i XLS.

dopasowywania w czasie produktów (matching), który ze względu na czasochłonność realizowany jest dla każdej grupy produktów z osobna (por. podrozdz. 4.3). Produkty z bieżącego miesiąca są parowane (matchowane) z produktami z poprzednich miesięcy analizowanego roku oraz z produktami z grudnia poprzedniego roku. W każdej z ramek danych, które odpowiadają grupom elementarnym, pojawia się ponownie nowa kolumna (prodID), która tym razem oznacza identyfikator dopasowanych do siebie produktów.

4.1.4. Zapis i archiwizacja ramki danych

Ramki danych, które poprawnie przeszły etap czyszczenia danych oraz klasyfikacji i dopasowywania produktów, są gotowe do obliczania na ich podstawie wskaźników cen. Mimo że tuż przed wyznaczeniem wskaźników cen dane poddane zostaną jeszcze odpowiedniemu filtrowaniu (por. podrozdz. 4.4), to na tym etapie w GUS wychodzi się z założenia, że większą wartość informacyjną ma cały zbiór i jego ewentualną redukcję można przeprowadzić dopiero wtedy, gdy zajdzie taka potrzeba. Uszczuplenie zbioru danych na tym etapie mogłoby wpływać np. na późniejsze dopasowywanie produktów lub przyszłą ocenę udziałów w sprzedaży poszczególnych produktów. Przetworzone dotychczas dane nadają się zatem do zapisu i archiwizacji.

W przypadku każdej grupy elementarnej przygotowane ramki danych są zapisywane w folderach odpowiadających danej sieci w ujednoliconym formacie (RDS) i pod nazwą odpowiadającą danej grupie elementarnej. Format plików RDS zapewnia oszczędność miejsca na dysku serwera (pliki zajmują od kilku do kilkunastu razy mniej pamięci dyskowej niż pliki CSV), a jednocześnie pozwala na bezpośrednią pracę z danymi w językach wysokiego poziomu (np. w środowisku R czy Python). Pliki RDS są powiększane wraz z napływem danych z kolejnych miesięcy i osiągają największy rozmiar na koniec roku. Dopiero wówczas staje się możliwe wyznaczenie wskaźników porównujących grudzień bieżącego roku z grudniem roku poprzedniego. Gdy pojawiają się dane za styczeń kolejnego roku, wówczas tworzone są nowe pliki RDS zawierające dodatkowo okres bazowy (grudzień poprzedniego roku), natomiast poprzednie pliki RDS (13 miesięcy obserwacji) zostają zarchiwizowane (w formacie ZIP). Następnie cały omówiony cykl zaczyna się nowa, tzn. zapisywane zbiory RDS są powiększane z każdym miesiącem, a okno czasowe analizy, mające swój początek w okresie bazowym, rozszerza się ponownie aż do grudnia bieżącego roku.

Tablica 4.1 przedstawia strukturę przykładowej ramki danych skanowanych zapisanej w pliku RDS i dotyczącej grupy ryż. Ramka została zapisana w postaci, która umożliwia wyznaczenie na jej podstawie wskaźników cen za pomocą pakietu *PriceIndices* w środowisku R.

Tabl. 4.1. Przykładowa struktura danych skanowanych w gotowej ramce danych

time	prices	quantities	retID	EAN	coicop	description	grammage	unit	prodID
2020-12-31	10,47	8,48	26-617	5906747171261	011111	ryż długoziarnisty	0.4	kg	1
2020-12-31	12,47	5,87	40-772	5906747171261	011111	ryż długoziarnisty	0.4	kg	1
2020-12-31	11,4	15,65	70-001	5906747171261	011111	ryż długoziarnisty	0.4	kg	1
2020-12-31	13,2	16,95	85-791	5906747171261	011111	ryż długoziarnisty	0.4	kg	1
2020-12-31	11,47	85,41	01-460	5906747171261	011111	ryż długoziarnisty	0.4	kg	1
2020-12-31	11,97	7,82	05-820	5906747171261	011111	ryż długoziarnisty	0.4	kg	1

Źródło: opracowanie własne.

4.2. Proces i metody klasyfikacji produktów do grup elementarnych

Klasyfikacja produktów do odpowiadających im grup elementarnych jest, obok dopasowania produktów w czasie, najważniejszym etapem przygotowywania danych skanowanych przed finalnym obliczeniem wskaźników cen. To trudny i złożony etap, wymagający manualnej pracy na początku, ale skutkujący niemal automatyczną klasyfikacją produktów. Dopiero po klasyfikacji produktów do grup elementarnych (COICOP 5) oraz niżej, do ich podgrup (lokalny COICOP 6), następuje etap dopasowania produktów (matching) i obliczania wskaźników cen dla tych grup. Z tego względu etap klasyfikacji jest nieodzowny, a jego poprawność ma realny wpływ na ostateczne oceny zmian cen (wartości indeksów cen). Warto podkreślić, że klasyfikacja produktów jedynie do poziomu COICOP 5 może się okazać niewystarczająca, ponieważ produkty, które znajdują się w grupie elementarnej (reprezentanty), nie zawsze tworzą homogeniczny zestaw. W toku realizacji projektu INSTATCENY przyjęto więc co do zasady, że reprezentant znajdujący się wewnątrz danej grupy elementarnej wyznacza, w przypadku danych skanowanych, bardziej jednorodną podgrupę produktów. To właśnie tego rodzaju podgrupy grup elementarnych stanowią docelowy kierunek klasyfikacji produktów w projekcie. Przykładowo w grupie elementarnej jogurty (COICOP: 011441) wyróżniono homogeniczne podgrupy: jogurt naturalny, jogurt owocowy oraz jogurt pitny. Podgrupy te, w zależności od sieci, liczą od kilkudziesięciu do nawet kilkuset produktów dostępnych w sprzedaży każdego miesiąca.

4.2.1. Klasyfikacja za pomocą słów kluczowych i fraz

Klasyfikacja produktów do homogenicznych zbiorów jest stosunkowo prostą i bardzo skuteczną techniką klasyfikacyjną, ale opierającą się na założeniu, że etykiety produktów (opisy) są wystarczająco szczegółowe i kompletne. Polega ona na tym, aby na podstawie zestawu słów kluczowych i/lub fraz, które występują lub nie występują w opisie produktu, przyporządkować dany produkt do odpowiadającej mu grupy produktów. Sama detekcja słów kluczowych i fraz w etykiecie produktu jest prostym zadaniem analitycznym – wystarczy użyć do tego celu odpowiedniej ko-

mendy języka wysokiego poziomu (np. *str_detect()* z pakietu w R o nazwie *stringr*). Dokładna klasyfikacja produktów wymaga jednak utworzenia całego zestawu fraz i słów kluczowych (wektorów łańcuchów tekstowych), które alternatywnie mogą pojawić się w etykiecie produktu, muszą się w niej pojawić lub też nie mogą być zawarte w etykiecie. Przykładowo w pakiecie *PriceIndices* (Białek, 2020) funkcja *data_selecting()*, która realizuje klasyfikację produktów poprzez ich selekcję na podstawie słów kluczowych i fraz, ma odpowiednio dobrane do tego celu parametry: include, must i exclude (por. tabl. 4.2).

Tabl. 4.2. Przykładowy zestaw parametrów klasyfikujących produkty poprzez selekcję opartą na frazach i słowach kluczowych

Homogeniczna grupa produktów	Parametr *include*	Parametr *must*	Parametr *exclude*
Mąka pszenna	„pszen", „razowa", „uniwersalna", „Uniwer", „Tradyc", „tradycyjna", „krupczatka", „tortowa", „pizz", „szczec", „Szczec", „wroc", „tort", „luks", „pozna", „Zamojska", „Tort", „Hetma", „chleb", „Wypiek", „gdańsk"	„mąka"	„żyt"
Ryż długoziarnisty	„ziarn", „dłogoziar", „długo", „risotto", „parboiled", „Basmati", „Jaśminowy", „paraboliczny"	„ry"	„płatki", „platki", „chrup", „britta", „natur"
Chusteczki dla dzieci	„chust"	„dzieci"	„platki", „patyczki", „podkłd", „płatki", „patyc", „mydło", „mydlo"
Cień do powiek	„cien", „cień", „EYESHADOW", „pow"	„cie"	„baza", „eyeliner", „tusz", „podkład", „baza", „pomada"
Grzebień	„grzeb"	„grz"	„pędzel", „pędzel", „grzywka", „grzyb"

Źródło: opracowanie własne.

Należy zaznaczyć, że niewątpliwą zaletą tego typu klasyfikacji jest to, że przygotowując zestawy słów kluczowych i fraz, na bieżąco śledzimy efekty klasyfikacji i dzięki temu upewniamy się, że ostateczny zestaw kombinacji wektorów *include*, *must* i *exclude* prowadzi do 100% poprawnych klasyfikacji (przynajmniej na zbiorze treningowym). Ten rodzaj klasyfikacji jest również stosunkowo szybki w realizacji, zwłaszcza jeśli liczba słów kluczowych i fraz nie jest zbyt duża. Pewną niedogodnością jest jednak konieczność opracowania słowników dla każdej sieci handlowej z osobna, ponieważ sieci z reguły różnią się sposobem etykietowania produktów. Poza tym należy pamiętać, że gdy pojawi się nowy produkt o bardzo szczątkowym opisie (np. zawężonym tylko do nazwy producenta), nie zostanie on sklasyfikowany.

4.2.2. Wykorzystanie do klasyfikacji metod uczenia maszynowego

Zdecydowana większość krajów, które korzystają z danych skanowanych i skrapowanych, dokonuje klasyfikacji produktów za pomocą metod uczenia maszynowego (ang. *machine learning methods*). Metody uczenia maszynowego mają ugruntowaną pozycję w statystyce publicznej (United Nations, 2021), co wynika z tego, że zapewniają automatyzację procesów klasyfikacyjnych przy jednoczesnym szerokim i darmowym dostępie do wyspecjalizowanych pakietów do uczenia maszynowego, stworzonych w językach wysokiego poziomu (np. pakiet *caret* w R czy *keras* lub *scikit-learn* w Pythonie). Metody uczenia maszynowego można podzielić na dwie podstawowe grupy[4]: metody z uczeniem nadzorowanym (ang. *supervised machine learning methods*) i metody z uczeniem nienadzorowanym (ang. *unsupervised machine learning methods*), przy czym wyłącznie te pierwsze, służące klasyfikowaniu produktów, będą rozważane w dalszej części pracy.

Do najpopularniejszych metod uczenia maszynowego wykorzystywanych przez urzędy statystyczne do klasyfikacji produktów do grup COICOP należą: naiwny klasyfikator bayesowski, regresja logistyczna, metoda wektorów nośnych (SVM), drzewa decyzyjne i lasy losowe, metoda k-najbliższych sąsiadów czy sieci neuronowe (United Nations, 2021). W ramach projektu INSTATCENY przeprowadzono na szeroką skalę eksperymenty klasyfikacyjne na zbiorach danych skanowanych pochodzących od dwóch sieci handlowych. Ostatecznie w aplikacji stworzonej w Instytucie Podstaw Informatyki PAN zaimplementowano metody o największej skuteczności (ang. *accuracy*): naiwny klasyfikator bayesowski, metodę SVM oraz lasy losowe. Dodatkowo do pakietu *PriceIndices* (GUS) dołączono algorytm XGBoost (Chen i in., 2001) realizujący klasyfikację produktów za pomocą drzew losowych, dokonując przy tym obliczeń równoległych z wykorzystaniem rdzeni procesora i tym samym kilkukrotnie przyspieszając proces klasyfikacyjny (funkcja *data_classifying()*). Dokładne omówienie tych metod znajduje się w podrozdz. 5.3, natomiast poniżej przedstawiona została krótka charakterystyka procesu uczenia maszynowego.

Dobór metody uczenia maszynowego, a także miar oceny jakości klasyfikacji podyktowany jest po pierwsze liczbą docelowych klas (grup), na jakie chcemy rozłącznie podzielić analizowane produkty, a po drugie skalą pomiarową, w jakiej wyrażone są zmienne (kolumny w ramce danych) użyte jako indykatory w modelu oraz zmienna celu (zmienna modelowana). W przypadku danych skanowanych i skrapowanych docelowa liczba klas jest praktycznie zawsze większa od 2 (nawet jeśli klasyfikujemy produkty do podgrup grup elementarnych), dlatego np. klasyczna regresja logistyczna nie znajduje tu zastosowania. Zauważmy także, że interesują nas w tym przypadku etykiety grup COICOP 5 lub COICOP 6, więc modelowana zmienna jest

[4] W praktyce, która jednak nie dotyczy problemów klasyfikacyjnych omawianych w tej monografii, wyróżnia się jeszcze uczenie częściowo nadzorowane oraz uczenie przez wzmacnianie.

kategoryczna – wyklucza to użycie średniego błędu kwadratowego (MSE) czy też średniego absolutnego błędu (MAE) jako miar jakości predykcji, natomiast rozważać można miary typu *accuracy*[5], *precision*, *recall* czy *F1*.

Punktem wyjścia do wdrożenia modelu uczenia maszynowego do klasyfikacji produktów jest utworzenie zbioru uczącego i zbioru testowego dla tego modelu. Zbiory te powinny być przygotowane niezależnie, niemiej jednak w praktyce będą pochodziły z tego samego zbioru danych skanowanych lub skrapowanych (od tej samej sieci) podzielonego według ustalonej proporcji. W toku realizacji projektu przyjęto, że wejściowy zbiór danych zostaje podzielony na zbiór uczący (90% najstarszych obserwacji) i testowy (10% najmłodszych obserwacji), przy czym w pracach eksperymentalnych zbiór uczący był dodatkowo dzielony na foldy w celu dokonania walidacji krzyżowej. Walidacja krzyżowa (k-stopniowa) polega na tym, że próba danych zostaje podzielona na k równolicznych podprób. Następnie model jest oddzielnie k razy uczony na próbach składających się z $k - 1/k\%$ obserwacji i walidacja odbywa się na pozostałych elementach zbioru (foldach). Za każdym razem inna podpróba jest wyłączona z uczenia i używana do walidowania. Błędem prognozy jest średnia z k prób walidacyjnych. Taka procedura minimalizuje ryzyko przetrenowania modelu oraz jego lokalnego dopasowania się do zbioru uczącego, dlatego w projekcie zastosowano 10-stopniową walidację krzyżową.

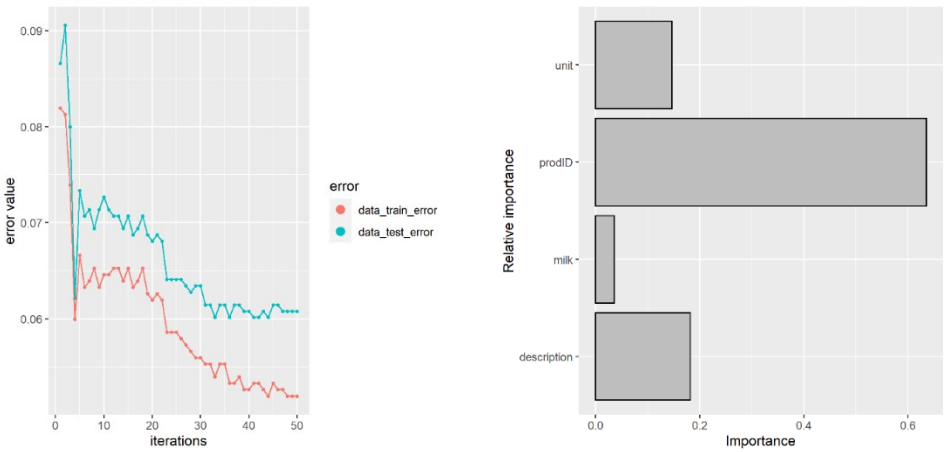
Należy podkreślić, że wybór modelu to pierwszy, a dobór jego indykatorów to drugi – równie ważny – etap procesu uczenia maszynowego. Atrybutami produktu mogą być w szczególności: *etykieta produktu*, *kod produktu*, *gramatura produktu*, *jednostka sprzedaży*, *cena* i *ilość sprzedaży* (*wolumen obrotu*) oraz opcjonalnie *stawka VAT*. Zaimplementowane w projekcie metody zostały tak oprogramowane, aby automatycznie dokonywały wyszukiwania skali pomiarowej indykatora (np. *cena* – cecha ilościowa wyrażona na skali mocnej, *jednostka sprzedaży* – cecha kategoryczna wyrażona na skali słabej) i dzięki temu poprawnie przeprowadzały proces uczenia się modelu, a tym samym budowę węzłów decyzyjnych. Ostatecznie wyboru modelu i zbioru przyjętych indykatorów dokonuje się na poziomie sieci handlowej (rekomendowane jest stworzenie odrębnych modeli dla każdej sieci). Przykładowa analiza błędów klasyfikacji (1-*accuracy*) produktów mlecznych za pomocą algorytmu drzew losowych (założono maksymalnie 50 iteracji) dokonana na zbiorze uczącym i testowym pochodzącym z jednej ze współpracujących z GUS sieci handlowych na podstawie indykatorów: jednostka sprzedaży (unit), kod produktu sieci (prodID), słowo kluczowe „mleko" (milk) oraz etykieta produktu (description) została zaprezentowana na wykr. 4.1A. Jak wynika z wykr. 4.1B, najlepszym indykatorem na podstawie

[5] W projekcie przyjęto, że naturalną miarą jakości klasyfikacji będzie wskaźnik *accuracy*, czyli iloraz liczby poprawnie zaklasyfikowanych produktów do liczby wszystkich rozważanych produktów.

indeksu Giniego (Menze i in., 2009), który służy wyznaczeniu wpływu („ważności") danego indykatora na ostateczny kształt modelu, jest w analizowanym przypadku zmienna prodID.

Wykr. 4.1. Proces uczenia modelu drzew losowych zastosowanego do klasyfikacji produktów mlecznych

a) Oceny błędów klasyfikacji modelu na zbiorze **b) Ocena ważności predyktorów uczącym i testowym w procesie budowy modelu**



Źródło: opracowanie własne w pakiecie *PriceIndices*.

4.2.3. Rekomendowana procedura klasyfikacji produktów

Obecnie wdrożona procedura klasyfikacyjna jest dwuetapowa i przebiega dla każdej z sieci z osobna (zarówno dla danych skanowanych, jak i skrapowanych). Na pierwszym etapie uruchomione zostaje klasyfikowanie produktów przez selekcję na podstawie słów kluczowych i fraz (por. pkt 4.2.1). Za pomocą funkcji *data_selecting()* z pakietu *PriceIndices* produktom zostaje przyporządkowana odpowiednia podgrupa grupy elementarnej (poziom COICOP 6) i nadana w ten sposób etykieta podgrupy. Ale – jak wspomniano w podrozdz. 4.2.1 – nie wszystkie produkty daje się w ten sposób zaklasyfikować ze względu na niekiedy niekompletny opis produktu. W połączeniu z manualnie przygotowanymi zbiorami uczącymi jest to jednak dobry punkt wyjścia do zaimplementowania metod uczenia maszynowego, ponieważ próba ucząca jest w ten sposób ciągle aktualizowana i rozbudowywana. Na drugim etapie modele wytrenowane na próbach uczących zostają zastosowane do klasyfikacji produktów z bieżącego miesiąca, w szczególności tych, których nie udało się zaklasyfikować na pierwszym etapie. Aby jednak wdrożyć dany model uczenia maszynowego, przyjęto, że jego jakość predykcyjna, mierzona odsetkiem poprawnych klasyfikacji (ang. *accuracy*), nie może być mniejsza niż 95% na zbiorze uczącym i 90% na zbiorze

testowym. W rzeczywistości uzyskiwano lepsze jakości modeli niż podane tutaj wartości progowe dla *accuracy* (por. 5.4.1). Zaakceptowany model, który był wystarczająco skuteczny dla danej sieci handlowej, jest zapisywany na dysku. W ten sposób, jeśli asortyment z nowego miesiąca okaże się zbliżony do asortymentu z poprzedniego miesiąca, trenowanie od nowa modelu można pominąć (dzięki temu zaoszczędzić czas) i użyć zarchiwizowanego modelu. Zaleca się jednak, aby przynajmniej raz na kwartał trenować modele z uwzględnieniem najświeższych danych.

4.3. Dopasowanie produktów w czasie (matching)

Kolejnym, po klasyfikacji, etapem przygotowania danych skanowanych jest ich dopasowanie w czasie (matching). Produkty zarówno stabilnie sprzedawane, jak i produkty, które pojawiają się i znikają z półek supermarketów (np. produkty sezonowe), a także wszystkie te, które w czasie trwania okna obserwacji zmieniły swój opis lub kod (a często jedno i drugie), powinny zostać dopasowane do siebie w porównywanych miesiącach analizy. Jeżeli produkt, którego dotyczą zmiany, o których mowa powyżej, nie uległ zmianie pod względem jakościowym, to obserwowany w różnych momentach powinien być identyfikowany przez ten sam charakteryzujący go numer (prodID). Na potrzeby projektu zostały utworzone dwie funkcje dopasowujące produkty: jedna została zaimplementowana w Pythonie przez zespół IPI PAN, druga została napisana w środowisku R i zaimplementowana w pakiecie *PriceIndices*. Generalnie sposób działania obu funkcji jest identyczny (por. pkt 4.4.1 i 4.4.2), różnicę stanowi jednak miara odległości tekstowej zastosowana do porównywania etykiet (w Pythonie jest to miara Jaccarda, w R jest to miara Jaro-Winklera). Idea tworzenia dwóch funkcji napisanych w różnych środowiskach i używających różnych miar odległości tekstowych jest uzasadniona potrzebą weryfikacji procesu dopasowywania produktów. To proces niezwykle ważny, ponieważ niepoprawnie dopasowane produkty w konsekwencji doprowadzą do niepoprawnych oszacowań wskaźników cen. Dlatego też matching odbywający się każdego miesiąca jest przeprowadzany z użyciem obu funkcji, a następnie wyniki są porównywane, także ze względu na czasochłonność całego procesu. Prace eksperymentalne, które wykonano w ramach projektu INSTATCENY, potwierdziły niezawodność obu funkcji, a także zbieżność wyników działania tych funkcji. Planowana do wdrożenia jest funkcja napisana w Pythonie przez zespół IPI PAN (funkcja *compare_products()*), jednak funkcja z pakietu *PriceIndices* będzie służyła jako weryfikator poprawności wyników oraz alternatywa na wypadek ewentualnej awarii aplikacji IPI PAN. Poniżej przedstawiono zaimplementowaną procedurę dopasowywania produktów, która zależy od atrybutów charakteryzujących produkty sprzedawane przez daną sieć.

4.3.1. Wykorzystanie etykiet produktów

Teoretycznie dopasowanie produktów może być przeprowadzone wyłącznie na podstawie etykiet produktów. Warunkiem koniecznym takiego dopasowania jest dokładny i jednoznaczny opis każdego produktu w bazie danych, natomiast ewentualną przesłanką – duży rozmiar bazy danych. Ten drugi argument jest o tyle istotny, że matching produktów już w kilkudziesięciotysięcznym zbiorze danych jest dość czasochłonnym procesem, zwłaszcza jeśli wykonujemy go dla dłuższego odcinka czasu (np. dla obserwacji z całego roku). W przypadku dopasowywania produktów wyłącznie na podstawie opisów próg odcięcia dla miary odległości tekstowej musi być bardzo niski, np. 0,005 czy 0,01. Oznacza to, że za dopasowane produkty uznamy takie produkty, których podobieństwo opisów wynosi odpowiednio 0,995 lub 0,99. W praktyce jednak matching jest wsparty o dodatkowe atrybuty produktu (por. pkt 4.4.2), chyba że sieć ich nie dostarcza lub dostarcza, ale tylko dla wąskiej grupy produktów.

4.3.2. Możliwość wykorzystania dodatkowych atrybutów produktów

W ramach projektu INSTATCENY wypracowana została procedura dopasowywania produktów bazująca zarówno na etykietach produktów, jak i ich kodach, tj. uwzględniono kod wewnętrzny sieci i kod zewnętrzny (GTIN, EAN lub SKU). Nie wszystkie sieci, z którymi zawarto porozumienia, dostarczają kody zewnętrzne produktu, dlatego procedura postępowania musi być elastyczna i uwzględniać każdy zakres przesyłanych danych. Najogólniej można wypracowaną procedurę przedstawić w następujących krokach:

- Krok 1. Za dopasowane produkty uznajemy te, które mają ten sam zewnętrzny kod kreskowy (codeOUT) oraz ten sam wewnętrzny kod produktu (kod produktu według sprzedawcy – codeIN);
- Krok 2. W przypadku produktów, które mają równy tylko jeden ze wspomnianych kodów, brane są dodatkowo pod uwagę etykiety tych produktów, np. w przypadku produktów, które mają ten sam kod produktu według sprzedawcy (codeIN), ale różny kod EAN (codeOUT), nadal jest duża szansa na to, że jakościowo jest to ten sam produkt. W takiej sytuacji więc porównujemy etykiety tekstowe obu produktów i za dopasowane uznajemy te produkty, które mają albo identyczne, albo bardzo podobne etykiety. Jak wspomniano, wdrożeniowo do oceny podobieństwa etykiet zastosowano miarę Jaccarda, przy czym ostatecznie zaimplementowano nieznacznie ulepszoną wersję tej metryki (Ratcliff i Obershelp, 1983). Podobieństwo graniczne (1-miara odległości), jakie prowadzi do stwierdzenia podobieństwa porównywanych produktów, jest regulowane przez odpowiedni parametr (np. text_sim_treshold = 0.95).

- Krok 3. W przypadku gdy produkty mają różny kod sprzedawcy i różny kod zewnętrzny, za dopasowane uznajemy te produkty, które mają identyczną etykietę (odległość tekstowa miary Jaccarda ich etykiet wynosi 0). Produkty, które nie spełniają tego wymogu, są uznawane za niepodobne (niedopasowane).

W zaimplementowanej funkcji *data_matching()* z pakietu *PriceIndices* przewidziano jeszcze jedną możliwość, polegającą na wskazaniu przez analityka tej zmiennej lub grupy zmiennych, które muszą być identyczne w przypadku dopasowanych produktów – parametr *variables*. Parametr ten może wskazywać np. na zmienną określającą przynależność do grupy elementarnej lub jednostkę sprzedaży. W tym ostatnim przypadku należy jednak być ostrożnym, ponieważ może się zdarzyć, że produkt sprzedawany dotychczas w kilogramach w pewnym momencie zaczyna być sprzedawany na sztuki lub w gramach.

4.4. Filtrowanie danych skanowanych i skrapowanych

Produkty, które zostały najpierw sklasyfikowane do grup COICOP i dopasowane do siebie w czasie, zostają następnie poddane procesowi filtrowania. Filtracja danych ma zapewnić usunięcie z próby tych produktów, których nienaturalna zmiana ceny mogłaby zaburzyć pomiar dynamiki cen. Podobnie, z punktu widzenia fundamentalnego założenia, że w koszyku produktów wykorzystywanych do pomiaru inflacji powinny znajdować się tylko najbardziej reprezentatywne, faktycznie konsumowane produkty, filtrowanie ma na celu eliminację produktów o relatywnie niskiej wartości sprzedaży. Pozostaje pytanie o ostateczny los tego typu odfiltrowanych produktów, które są flagowane (oznaczane). Niektóre kraje stoją na stanowisku, że filtrowanie danych w ogóle nie jest potrzebne w przypadku danych skanowanych, ponieważ po pierwsze niepotrzebnie redukuje się w ten sposób próbę badawczą, a po drugie ważone formuły indeksów (w tym konstruowane na podstawie metod multilateralnych) i tak dokonują odpowiedniego ważenia indeksów cząstkowych i promują najlepiej sprzedające się produkty. Innymi słowy, kraje te widzą ewentualną korzyść z filtracji tylko wtedy, gdy do pomiaru dynamiki cen jest używany łańcuchowy indeks Jevonsa (jako formuła nieważona, por. podrozdz. 4.5), co jest w zgodzie z klasyczną metodologią stosowaną na poziomie grup elementarnych. W literaturze podejście to jest zresztą powszechnie nazywane dynamicznym (ang. *dynamic approach*). Z drugiej strony, filtracja danych to oszczędność miejsca w przestrzeni dyskowej serwera operacyjnego i jednocześnie krótszy czas obliczeniowy dla formuł indeksowych (por. pkt 4.4.3). W związku z tym wiele krajów, w tym Polska, stosuje wstępną filtrację produktów, przy czym problem doboru rodzajów filtrów oraz ich wielkości progowych wydaje się nadal kwestią otwartą. Ostateczna decyzja o imputacji oflagowanych w wyniku filtracji danych o cenach i sprzedaży produktów jest również dyskusyjna. Należy pamiętać bowiem, że eliminacja tego rodzaju produktów prowadzi do reduk-

cji próby, ale imputacja wygenerowanych braków danych może prowadzić z kolei do obciążenia szacunków dynamiki cen.

Na potrzeby projektu zaimplementowano trzy rodzaje filtrów dla produktów, przy czym działają one niezależnie, tzn. kolejność ich uruchamiania nie ma znaczenia. A zatem każdy filtr z osobna usuwa dane z oryginalnego zbioru, a następnie z całego zbioru usuwana jest suma usuniętych wcześniej produktów. Należy dodać, że dla zadanego okna czasowego filtracja dotyczy wszystkich sąsiadujących ze sobą miesięcy, tzn. porównywany jest drugi miesiąc z pierwszym, trzeci z drugim, czwarty z trzecim itd.

4.4.1. Filtry ekstremalnych zmian cen

Filtr ekstremalnych cen (ang. *extreme price filter*) to filtr, który eliminuje z próby produkty, których miesięczne zmiany cen były zbyt duże. W praktyce przyjęto, że takimi zmianami będą miesięczny wzrost ceny o przynajmniej 200% (oznacza to trzykrotny wzrost ceny) oraz spadek ceny o ponad 75% (czterokrotny spadek ceny). Ekstremalne zmiany ceny mogą wynikać z kilku przyczyn. Mogą powstawać w wyniku nadzwyczajnych obniżek, przecen i rabatów, ale też mogą wynikać ze zwykłego błędu zapisu danych (np. gdy skraper źle pobrał cenę produktu ze strony internetowej). Filtr ekstremalnych zmian cen dotyczy zarówno cen skanowanych, jak i skrapowanych.

4.4.2. Filtry niskich poziomów sprzedaży i nieistotnych cen

Filtr niskich sprzedaży (ang. *low sales filter*) to filtr, który usuwa z próby produkty o relatywnie niskiej sprzedaży na tle innych produktów z tej samej homogenicznej grupy. Tym samym filtr niskich sprzedaży zapobiega uwzględnieniu w szacowaniu inflacji tych produktów, które na rynku konsumenckim znaczą relatywnie niewiele. Uściślając, w przypadku porównywania miesięcy $t-1$ i t produkt i obserwowany w grupie n dostępnych w tym czasie homogenicznych produktów (dopasowanych w czasie, czyli uzyskanych po matchingu) jest usuwany, jeśli zachodzi poniższa relacja:

$$\frac{s_i^{t-1}+s_i^t}{2} < \frac{1}{\lambda \cdot n}, \tag{4.1}$$

gdzie λ jest arbitralnie przyjętą stałą dodatnią (w projekcie przyjęto typową wartość $\lambda = 1{,}25$ (Van Loon i Roels, 2018), natomiast s_i^τ oznacza relatywny udział w sprzedaży danej grupy i-tego produktu w miesiącu τ. Ponieważ filtr ten wymaga znajomości poziomu konsumpcji analizowanych produktów, dotyczy zasadniczo danych skanowanych. Dane skrapowane nie dostarczają informacji o ilości sprzedawanych

produktów, choć podejmowane są w tym przypadku próby konstrukcji systemu wag, np. bazującego na liczbie kliknięć na podstronie opisującej dany produkt (Ayoubkhani i Heledd, 2022).

Czasami niskiej wartości sprzedaży produktu towarzyszy też wyraźny spadkowy trend jego ceny. Z założenia są to produkty, które za moment i tak byłyby wycofywane z oferty sprzedażowej. Zgodnie z rekomendacjami krajów, które stosują tego rodzaju filtr zrzucanych (nieistotnych) cen (ang. *dump price filter*), w implementacji tego filtru przyjęto 70% jako graniczną wartość dla spadku ceny i 75% dla spadku sprzedaży. Innymi słowy, spadki cen i wielkości sprzedaży większe od tych wartości progowych prowadzą do usunięcia tego rodzaju produktów z próby.

4.4.3. Wpływ filtrowania danych na redukcję rozmiaru bazy danych wartość indeksów cen

Każdy z omówionych filtrów danych (filtr ekstremalnych zmian cen, filtr niskich sprzedaży oraz filtr nieistotnych cen) prowadzi w mniejszym lub większym stopniu do redukcji wymiaru bazy danych, o ile tylko odfiltrowane dane są w konsekwencji usuwane. Jak wspomniano, redukcja ta w naturalny sposób prowadzi do oszczędności miejsca w przestrzeni dyskowej serwera operacyjnego i jednocześnie skraca czas obliczeniowy dla formuł indeksowych. Odrębną kwestią pozostaje oczywiście pytanie o wpływ filtracji danych na finalny poziom indeksów cen. W praktyce oczekujemy, że wpływ ten będzie niewielki, co dodatkowo uzasadniałoby stosowanie tego rodzaju filtrów (skrócenie czasu obliczeniowego wskaźników przy jednoczesnym braku istotnego obciążenia wyników). W tabl. 4.3 przedstawiono ocenę wpływu filtrowania przykładowych baz danych skanowanych na ich rozmiar, liczbę uwzględnionych produktów oraz wartość wybranych indeksów cenowych. Do celów ilustracyjnych wybrano homogeniczne grupy produktów pozyskane z jednej z sieci handlowej, przy czym analiza objęła okres od grudnia 2020 r. do grudnia 2021 r. Analizę przeprowadzono w pakiecie *PriceIndices* za pomocą funkcji *data_filtering()* oraz funkcji indeksowych: *jevons()*, *chjevons()*, *fisher()* oraz *geks()*.

Tabl. 4.3. Wpływ filtrowania przykładowych danych skanowanych na redukcję rozmiaru bazy danych i wartości wybranych indeksów cen (dane za okres: grudzień 2020–grudzień 2021)

Rodzaj filtru	Liczba wierszy w bazie	Liczba produktów w bazie	Indeks Jevonsa	Łańcuchowy indeks Jevonsa	Indeks Fishera	Indeks GEKS
Grupa produktów: ryż						
Brak filtru	48072	38	1,023846	1,023784	1,03148	1,028475
Filtr ekstremalnych zmian cen	48071	38	1,020172	1,023784	1,031476	1,028474
Filtr niskich sprzedaży	27190	20	1,024338	0,998925	1,039426	1,034023
Filtr zrzucanych cen	48071	38	1,020172	1,023784	1,031476	1,028474
Grupa produktów: środki papiernicze i higieniczne						
Brak filtru	388551	553	1,019182	0,9556525	1,080191	1,073590
Filtr ekstremalnych zmian cen	388084	506	1,017785	0,9480464	1,081081	1,074170
Filtr niskich sprzedaży	203457	190	1,050448	1,0436440	1,102698	1,097022
Filtr zrzucanych cen	388088	506	1,017968	0,955652	1,081104	1,074226
Grupa produktów: kosmetyki						
Brak filtru	48072	38	1,023846	1,023784	1,031480	1,028475
Filtr ekstremalnych zmian cen	48071	38	1,020172	1,023784	1,031476	1,028474
Filtr niskich sprzedaży	27190	20	1,024338	0,998925	1,039426	1,034023
Filtr zrzucanych cen	48071	38	1,020172	1,023784	1,031476	1,028474

Źródło: opracowanie własne w pakiecie *PriceIndices*.

Analiza wyników prezentowanych w tabl. 4.3 skłania do wniosku, że w przypadku ryżu oraz kosmetyków filtr ekstremalnych zmian cen oraz filtr zrzucanych cen nie miały znaczącego wpływu na redukcję wymiaru ramki danych skanowanych i tym samym liczbę produktów. W przypadku środków papierniczych i higienicznych filtry te prowadziły do identycznej, wynoszącej ponad 8%, redukcji liczby produktów w bazie danych, ale tylko nieznacznej zmiany wartości indeksów cenowych. Natomiast filtr niskich sprzedaży zdecydowanie redukował liczbę rekordów oraz liczbę produktów w przypadku wszystkich analizowanych grup produktów. Jednocześnie zastosowanie tego filtru prowadziło do widocznych zmian wartości indeksów cenowych.

4.5. Implementacja wybranych formuł obliczania wskaźników cen

Dane skanowane charakteryzuje przede wszystkim to, że zawierają informację o poziomie konsumpcji pojedynczego produktu, a więc na poziomie znacznie niższym niż elementarny COICOP 5, a co za tym idzie – nie ma najmniejszych metodologicznych przeszkód, aby stosować w ich przypadku ważone formuły indeksów cen. Co ważne, system wag nie musi być w tym przypadku oparty jedynie na okresie bazowym, jak w tradycyjnym pomiarze inflacji, lecz wagi mogą być ustalane dla dowolnego okresu badawczego włącznie z miesiącem bieżącym. W przypadku danych skrapowanych zasadniczo jesteśmy ograniczeni do formuł nieważonych (np. indeks

Jevonsa, łańcuchowy indeks Jevonsa czy indeks GEKS-J), niemniej jednak w literaturze przedmiotu funkcjonują metody ustalania wag dla danych skrapowanych (Ayoubkhani i Heledd, 2022). Teoretycznie istnieje zatem możliwość zastosowana formuł ważonych również w przypadku danych skrapowanych, ale nie jest to praktyka krajowych urzędów statystycznych. Poniżej omówione zostały formuły indeksowe, które zaimplementowano w projektowej aplikacji. Lista funkcjonujących w literaturze indeksów cen jest bardzo długa IMF i in., 2004, 2020; von der Lippe, 2007), naszym celem nie jest jednak omówienie wszystkich dostępnych formuł indeksowych. Aby przedstawić konstrukcję indeksów zaimplementowanych w ramach projektu, wprowadźmy niezbędne oznaczenia:

$0, t$ – odpowiednio okres bazowy (ang. *base period*) i okres badany (ang. *current period*),

$G_{0,t}$ – zbiór dopasowanych produktów z okresów $0, t$,

$N_{0,t} = card \;\; G_{(0,t)}$ – liczba dopasowanych produktów z okresów $0, t$,

p_i^τ – cena i-tego produktu w okresie τ (rozumiana jako *unit value,* czyli średnia miesięczna cena),

q_i^τ – ilość sprzedanego i-tego produktu w okresie τ,

$\{0,1,2,\ldots,T\}$ – okno czasowe analizy (z reguły 13 miesięcy, tj. $T = 12$).

4.5.1. Formuła Jevonsa

Indeks Jevonsa (1865) jest najbardziej rekomendowanym indeksem elementarnym w przypadku tradycyjnej kolekcji danych (IMF i in., 2004, 2020). Stanowi on średnią geometryczną z cząstkowych indeksów cen:

$$P_J^{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{N_{0,t}}}. \tag{4.2}$$

4.5.2. Formuła Fishera

Indeks Fishera (1922) jest najczęściej rekomendowanym bilateralnym indeksem ważonym w przypadku tradycyjnej kolekcji danych (IMF i in., 2020)., co wynika z jego aksjomatycznych własności oraz bardzo dobrego przybliżenia indeksu kosztów utrzymania (COLI). Indeks ten stanowi średnią geometryczną z indeksów Laspeyresa (1864) i Paaschego (1874), tj.:

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}}, \tag{4.3}$$

przy czym indeks Laspeyresa i Paaschego można wyrazić odpowiednio jako

$$I_{La}^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t \cdot q_i^0}{\sum_{i \in G_{0,t}} p_i^0 \cdot q_i^0}, \tag{4.4}$$

$$I_{Pa}^{0,t} = \frac{\sum_{i \in G_{0,t}} p_i^t \cdot q_i^t}{\sum_{i \in G_{0,t}} p_i^0 \cdot q_i^t}. \tag{4.5}$$

4.5.3. Indeksy łańcuchowe

Indeks bilateralny, np. indeks Jevonsa czy Fishera, porównuje ze sobą dwa okresy badawcze, czyli okres bieżący i okres bazowy. W przypadku danych skanowanych lub danych skrapowanych użycie indeksów bilateralnych generowałoby jednak zbyt duże obciążenie pomiaru, zwłaszcza dla dużych odległości czasowych porównywanych okresów. Wynika to z faktu, że na poziomie kodu kreskowego rotacja sprzedawanych przez sieci handlowe produktów jest bardzo duża, co wiąże się z sezonowością produktów oraz zmieniającymi się upodobaniami konsumentów. Pewną alternatywą, która stara się uchwycić dynamikę produktów w całym przedziale czasu, jest zastosowanie indeksów łańcuchowych. Ogólna idea takich indeksów może być wyrażona formułą:

$$P_{ch}^{0,t} = P^{0,1} \cdot P^{1,2} \cdot \ldots \cdot P^{t-1,t} = \prod_{\tau=0}^{t-1} P^{\tau,\tau+1}, \tag{4.6}$$

gdzie $P^{\tau,\tau+1}$ jest dowolną bilateralną formułą indeksu cenowego wyznaczoną dla sąsiadujących ze sobą okresów (miesięcy) τ i $\tau + 1$. Jeśli np. przyjmiemy $P^{\tau,\tau+1} = P_J^{\tau,\tau+1}$, wówczas wzór (4.6) opisywać będzie łańcuchowy indeks Jevonsa. Należy mieć na uwadze, że zastosowanie łańcuchowych indeksów ważonych może prowadzić do obciążenia pomiaru dynamiki cen wynikającego z efektu łańcuchowego dryfu. Efekt ten powstaje wówczas, gdy ceny i ilości produktów po pewnym czasie powracają do wyjściowych wartości (np. w przypadku dóbr sezonowych), ale wartość indeksu cenowego nie powraca mimo wszystko do 1.

4.5.4. Indeksy multilateralne

Indeksy multilateralne wydają się najlepszym rozwiązaniem w przypadku danych skanowanych, ponieważ nie tylko działają na całym oknie czasowym $\{0,1,2,\ldots,T\}$, ale również są wolne od efektu łańcuchowego dryfu (Chessa, 2015). Indeksy te mają swoją genezę w porównaniach międzynarodowych i międzyregionalnych, a obecnie przeżywają swoisty renesans w kontekście danych skanowanych i skrapowanych. W pakiecie *PriceIndices* została oprogramowana większość funkcjonujących indek-

sów multilateralnych (np. indeksy GEKS, CCDI, GEKS-W, GEKS-J, TPD czy SPQ; Białek, 2022b). Poniżej omówiono dwa przykładowe indeksy multilateralne, które uwzględniono w eksperymentalnych pracach projektu INSTATCENY:

- Indeks GEKS (por. Eltetö i Köves, 1964; Gini, 1931; Szulc 1964):

$$P_{GEKS}^{0,t} = \prod_{\tau=0}^{T} \left(\frac{P_F^{\tau,t}}{P_F^{\tau,0}}\right)^{\frac{1}{T+1}}, \tag{4.7}$$

- Indeks TPD (ang. *time-product dummy;* de Haan i Krsinich, 2018), dla którego punktem wyjścia jest budowa ekonometrycznego modelu opisującego zachowanie cen produktów w przedziale czasowym $\{0,1,2,\ldots,T\}$:

$$ln\, p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{j=1}^{N-1} \gamma_j D_i^j + \varepsilon_i^t\,, \tag{4.8}$$

gdzie D_i^t jest zmienną przyjmującą wartość 1, gdy i-ty produkt jest dostępy w okresie t, lub 0 w przeciwnym przypadku; D_i^j jest zmienną przyjmującą wartość 1, gdy i-ty produkt należy do j-ej homogenicznej grupy produktów, lub 0 w przeciwnym przypadku; $\alpha, \delta^t, \gamma_j$ to parametry modelu, ε_i^t to składnik losowy o średniej równej 0. Przy powyższych oznaczeniach, po przeprowadzeniu estymacji parametrów modelu (4.8) ważoną metodą najmniejszych kwadratów WLS (Diewert, 2004), indeks TPD można wyrazić następująco:

$$P_{TPD}^{0,t} = \frac{\prod_{i \in G_t} \left(\frac{p_i^t}{exp(\hat{\gamma}_i)}\right)^{s_i^t}}{\prod_{i \in G_0} \left(\frac{p_i^0}{exp(\hat{\gamma}_i)}\right)^{s_i^0}}, \tag{4.9}$$

gdzie G_τ oznacza zbiór produktów dostępnych w sprzedaży w okresie τ, a s_i^τ – udział i-tego produktu w sprzedaży dokonanej w okresie τ.

4.6. Agregacja wskaźników cen wyznaczonych na bazie różnych źródeł danych

W przypadku danych skanowanych i skrapowanych wskaźniki cen są wyznaczane dla każdej sieci z osobna. Dodatkowo w przypadku skaningu, jeśli sieć prowadzi regionalną politykę w zakresie cen produktów i zamówień, wskaźniki cen są wyznaczane niezależnie dla każdego punktu sprzedaży. Metodologiczną kwestią, jaką należy rozwiązać, jest ustalenie sposobu agregacji cenowych wskaźników cząstkowych wyznaczonych na bazie danych od wszystkich dostawców w obrębie danego źródła danych, co będziemy dalej nazywali *agregacją wewnętrzną.* Z kolei uwzględnienie wszystkich źródeł danych w pomiarze inflacji, tj. danych zbieranych metodą tradycyjną (przez ankieterów), danych skanowanych i skrapowanych, i w konsekwencji ustalenie wypadkowego wskaźnika cen na bazie otrzymanych w ten sposób trzech wskaźników cenowych stanowi wyzwanie metodologiczne określone przez nas mianem *agregacji zewnętrznej.*

4.6.1. Agregacja wewnętrzna wskaźników w obrębie źródła danych

W przypadku danych skanowanych tak naprawdę o ich agregacji możemy mówić już na etapie wyznaczania miesięcznych cen produktów. Cena produktu w tym przypadku jest rozumiana jako wartość jednostkowa (ang. *unit value*), stanowiąca łączną miesięczną wartość sprzedaży produktu podzieloną przez łączną sprzedaną liczbę jego jednostek. W ten sposób ceny są uśredniane w ramach miesiąca, przy czym z reguły sieć handlowa dostarcza danych za pierwsze trzy tygodnie kolejnych miesięcy. Jeśli sieć prowadzi regionalną politykę cenową to wartości jednostkowe powinny być wyznaczone dla każdego punktu sprzedaży z osobna. W dalszym opisie będziemy zakładać najbardziej ogólny przypadek, w którym każdy punkt sprzedaży (outlet) traktujemy niezależnie.

Na poziomie pojedynczej sieci handlowej, bez względu na ustaloną formułę indeksu cenowego (ważoną lub nieważoną), indeks ten należy wyznaczyć dla homogenicznej podgrupy produktów (lokalny poziom COICOP 6) dla każdego punktu sprzedaży z osobna. Następnie agregujemy wyznaczone w ten sposób indeksy cząstkowe w wypadkowy indeks dla podgrupy produktów, stosując jedną z dostępnych ważonych formuł agregatowych (np. formułę Laspeyresa, Paaschego czy Fishera). Wagi ustalane są w tym przypadku na podstawie relatywnych udziałów sprzedaży analizowanej podgrupy produktów w danym outlecie względem wszystkich outletów sieci. Etap ten kończy się uzyskaniem wskaźników cenowych dla podgrup grup elementarnych przy zastosowaniu agregacji względem punktów sprzedaży (ang. *aggregation over outlets*). Następnie, dokonując uśrednienia tych wskaźników na podstawie systemu wag opartego na udziałach sprzedaży podgrup produktów w sprzedaży

odpowiadających im grup elementarnych (w obrębie tej samej sieci), uzyskujemy wskaźniki cenowe dla grup elementarnych. Ten ostatni etap realizuje zatem agregację wskaźników dla podgrup produktów (ang. *aggregation over subgroups*). W efekcie dla każdej współpracującej z GUS sieci handlowej uzyskujemy finalne wskaźniki cen dla wszystkich grup elementarnych produktów, które znajdują się w ofercie sprzedażowej danej sieci. W pakiecie *PriceIndices* zaimplementowano do tego celu specjalne funkcje, tj. *final_index()* oraz *final_index2()*, przy czym użytkownik wybiera formułę agregacji względem punktów sprzedaży (parametr *aggrret*) oraz formułę agregacji względem podgrup produktów (parametr *aggrsets*). W projekcie INSTSTCENY przyjęto, że obie formuły agregujące będą formułami Fishera.

Następnie, posiadając wskaźniki cenowe wyznaczone dla poziomu COICOP 5 dla wszystkich sieci handlowych dostarczających dane skanowane, należy wyznaczyć ogólne wskaźniki dla grup elementarnych na poziomie źródła danych skanowanych. W tym celu stosujemy agregującą formułę Laspeyresa, przy czym wagami są w tym przypadku rynkowe udziały sprzedaży poszczególnych sieci w sprzedaży wszystkich sieci łącznie.

Podobną procedurę należy przeprowadzić na **poziomie źródła danych skrapowanych**, choć oczywiście – z uwagi na brak danych o poziomie konsumpcji – przebiega ona inaczej. Po pierwsze zamiast wartości jednostkowych za miesięczne ceny najczęściej przyjmuje się tu nieważone średnie geometryczne z dziennych zarejestrowanych poziomów cen. Po drugie nie mamy w tym przypadku do czynienia z agregacją względem punktów sprzedaży, natomiast agregacja względem podgrup przebiega według jednej z nieważonych formuł agregujących (w projekcie jako formułę agregującą przyjęto średnią geometryczną ze wskaźników cząstkowych). Ostatecznie, podobnie jak w przypadku danych skanowanych, dla każdej dostępnej grupy elementarnej, agregujemy wskaźniki wyznaczone dla poszczególnych sprzedawców w jeden wskaźnik wypadkowy. Jeżeli rynkowe udziały wartości sprzedaży poszczególnych sprzedawców w sprzedaży ogółem nie są znane, wówczas finalna agregacja wskaźników cząstkowych przebiega według formuły nieważonej. W przypadku projektu INSTATCENY, ponieważ skrapowanie cen dotyczyło znanych sieci handlowych, znajomość ich rynkowych udziałów pozwoliła jednak na zastosowanie agregującej formuły Laspeyresa.

4.6.2. Agregacja zewnętrzna wskaźników z różnych źródeł danych

Po przeprowadzeniu całego procesu agregacji wewnętrznej, na poziomie COICOP 5 otrzymujemy trzy wskaźniki cenowe: wskaźnik cen wyznaczony na podstawie tradycyjnej formy pozyskiwania danych (przez ankieterów), wskaźnik cen oparty na danych skanowanych i wskaźnik cen bazujący na danych skrapowanych. Dla wszystkich grup elementarnych produktów, których dotyczy omawiana sytuacja, należy

wyznaczyć ostateczny wskaźnik cen, który następnie – zgodnie z klasyczną procedurą – posłuży wyznaczeniu wskaźników cenowych na wyższych poziomach agregacji (aż do uzyskania CPI). Procedowanie alternatywnych źródeł danych w pomiarze inflacji kończy się zatem wraz z poziomem COICOP 5. Aby jednak wyznaczyć wypadkowe wskaźniki cen na tym poziomie, należy dokonać ważenia wskaźników uzyskanych ze wszystkich źródeł danych według formuły Laspeyresa. Aby ustalić poziom wag, w projekcie INSTATCENY udziały zakupu produktów przez Internet były szacowane na podstawie informacji pochodzących z badania budżetów gospodarstw domowych prowadzonego przez GUS, natomiast informacje o udziałach zakupów w sieciach handlowych uzyskano z baz danych Passport GMID, Euromonitor International oraz z badań rynku wewnętrznego prowadzonych przez GUS.

4.7. Ocena wpływu włączania nowych źródeł danych na pomiar dynamiki cen detalicznych

Decyzje o włączeniu lub niewłączeniu danego źródła danych do pomiaru inflacji, a także wybór formuły indeksu cenowego na poszczególnych poziomach agregacji mogą mieć wpływ na obciążenie wskaźników cen detalicznych (Nardo i in., 2005, 2011; Saisana i in., 2005; Sharpe i Salzman, 2004), a tym samym stanowią źródła niepewności co do uzyskiwanych przez nie wartości. Analiza odporności wskaźnika cen detalicznych pozwala na ocenę wpływu decyzji podejmowanych na poszczególnych etapach jego konstrukcji na wartość wskaźnika, przy czym obejmuje ona analizę niepewności oraz analizę wrażliwości.

Podstawą analizy niepewności jest budowa modelu łączącego zmienne wejściowe reprezentujące źródła niepewności (rodzaje założeń przyjmowanych na poszczególnych etapach konstrukcji wskaźnika cen detalicznych), wyznaczenie funkcji rozkładu zmiennych wejściowych, a następnie na tej podstawie ustalenie rozkładu wskaźnika cen detalicznych lub zmiennych wyjściowych stanowiących miary wpływu zmian założeń konstrukcji wskaźnika na jego wartości. Sama analiza niepewności polega na badaniu parametrów rozkładu wskaźnika cen detalicznych.

Celem analizy wrażliwości jest ocena udziału każdego ze zidentyfikowanych źródeł niepewności (każdego wejściowego założenia) w wariancji wskaźnika cen detalicznych. Odbywa się ona poprzez dekompozycję tej wariancji na wariancje wyjaśniane przez poszczególne zmienne wejściowe, reprezentujące typy założeń przyjmowanych na różnych etapach konstrukcji wskaźnika cen detalicznych. Analiza wrażliwości jest tym samym ściśle powiązana z analizą niepewności. Połączenie obu typów analiz umożliwia pomiar odporności wskaźników cen detalicznych na zmiany założeń przy ich konstrukcji, czyli analizę wpływu tych założeń na wartość wskaźnika cen detalicznych, co z kolei wpływa na uzyskiwane oceny zmian cen detalicznych. W praktyce struktura analizy niepewności i wrażliwości wskaźników cen

detalicznych zależy od tego, jakie źródła niepewności (etapy konstrukcji wskaźników) oraz jakie założenia odnoszące się do tych źródeł (warianty rozwiązań na tych etapach konstrukcji wskaźników) są przyjmowane w konkretnej analizie.

4.7.1. Próba badawcza

Ze względu na zakres dostępnych danych, aby móc wykorzystać w obliczeniach indeksów cen wszystkie trzy omawiane źródła, w badaniu uwzględniono ceny obserwowane w lutym i marcu 2021 r. Do badania wybrano siedem grup elementarnych obejmujących produkty spożywcze, tj. grupy: ryż, mleko pełne świeże, mleko świeże niskotłuszczowe, jogurt, napoje i inne produkty mleczne, cukier oraz kawa. W przypadku tradycyjnej metody zbierania danych (przez ankieterów) przeanalizowano ceny pochodzące z 207 rejonów notowań dla każdego reprezentanta z wymienionych grup, przy czym ceny – podobnie jak w przypadku cen skanowanych i skrapowanych – zostały przeskalowane do ustalonej jednostki miary (np. dla mleka był to litr, dla ryżu – kilogram itd.). W przypadku cen skanowanych i skrapowanych wykorzystano dane uzyskane z jednej z sieci handlowych współpracujących z GUS. Listę reprezentantów, na bazie których utworzono podgrupy grup elementarnych, rozszerzono o trzy nowe pozycje: jogurt czekoladowo-owocowy, cukier puder oraz kawa mielona. Podgrupy te były na tyle licznie reprezentowane i jednocześnie homogeniczne, że mimo rozbieżności w stosunku do obowiązującej listy reprezentantów postanowiono uwzględnić je przy ocenie zmian cen odpowiadających im grup ECOICOP 5.

4.7.2. Pozyskanie danych skanowanych i danych skrapowanych

Zakres pozyskiwanych danych zależy również od dostawcy, niemniej jednak w przypadku danych skanowanych użytych w niniejszym opracowaniu, poza kodami identyfikującymi produkt, transferowana w formacie CSV ramka danych zawierała: etykietę produktu, jednostkę sprzedaży, datę transakcji, cenę sprzedaży, wartość sprzedaży, flagę dotyczącą przecen i promocji oraz informację o stawce VAT. Siedem analizowanych grup elementarnych to w przypadku uwzględnionej w badaniu sieci handlowej ok. 32 MB danych skanowanych z każdego miesiąca.

Zakres pozyskiwanych danych skrapowanych jest bardzo zbliżony do tego, jaki bezpośrednio wysyła sieć. Okazało się, że wystawiane na stronie internetowej produkty uwzględnionej w badaniu sieci to między 40% a 90% produktów z tej samej kategorii, jakie można znaleźć w stacjonarnych punktach sprzedaży tej sieci. Przykładowo na początku 2021 r. w przypadku ryżu zarejestrowano 33 produkty po stronie danych skanowanych i tylko 27 wśród danych skrapowanych. Podobnie w przypadku kawy relacja ta wynosiła 275 do 152. Brak pełnego pokrycia produktów na półkach sklepowych przez produkty widniejące na stronach internetowych prawdo-

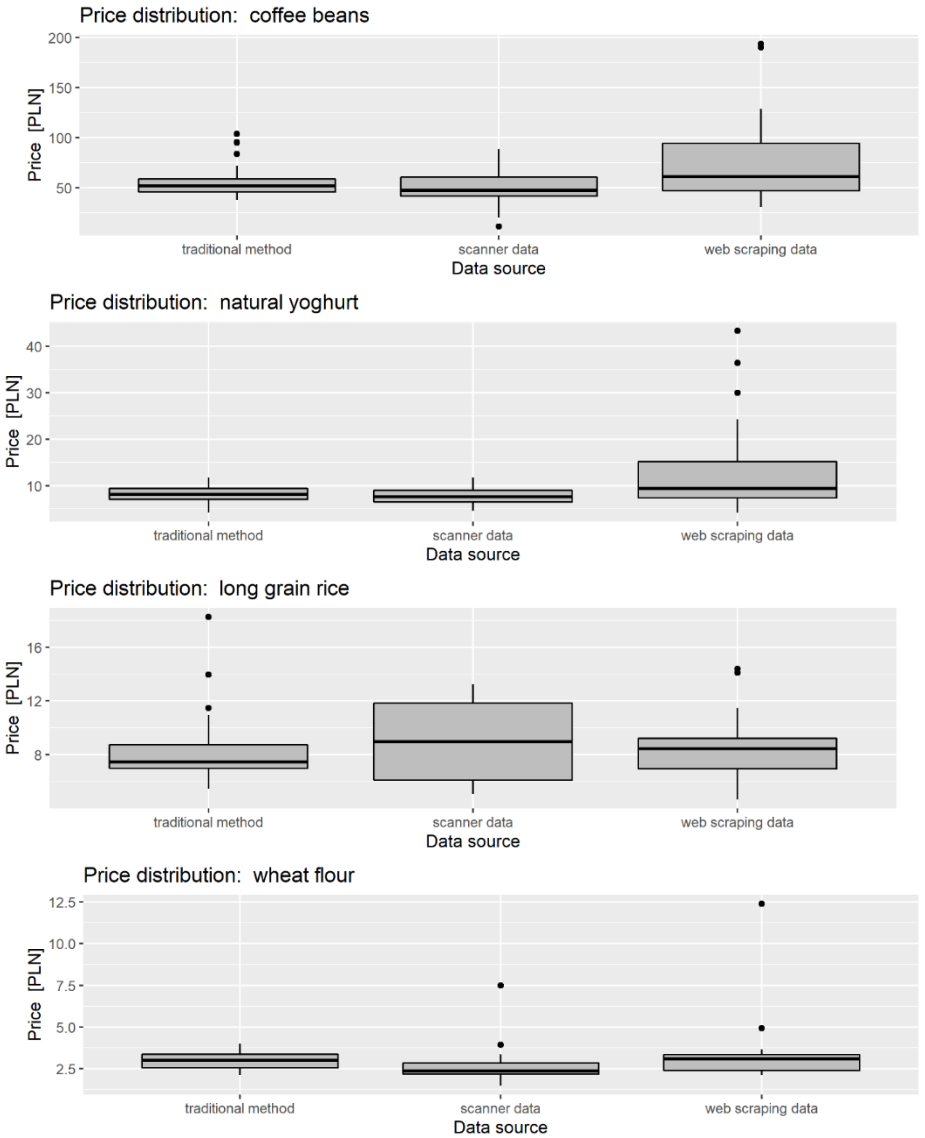
podobnie wynika z tego, że na stronach internetowych wystawiane są najbardziej popularne produkty. Siedem analizowanych grup elementarnych to w przypadku uwzględnionej w badaniu sieci handlowej ok. 4 MB danych skrapowanych z każdego miesiąca.

4.7.3. Przygotowanie danych ze źródeł alternatywnych

Po wstępnym wyczyszczeniu zbioru danych skanowanych i skrapowanych (tj. ujednoliceniu nazw, usunięciu błędnych danych i nietypowych cen) produkty zostały zaklasyfikowane do odpowiednich grup ECOICOP 5 i poziomu krajowego ECOICOP 6. W przypadku obydwu alternatywnych źródeł danych klasyfikacji produktów, jak również ich dopasowania dokonano na podstawie etykiet produktów i utworzonych wcześniej słowników słów kluczowych i fraz identyfikujących przynależność do tych grup. Zastosowano funkcje *data_selecting* oraz *data_matching* z pakietu *PriceIndices* w środowisku R. Etykiety tekstowe porównywano za pomocą miary odległości Jaro-Winklera (Jaro, 1989; Winkler, 1990), przy czym ustaloną graniczną odległością (powyżej której uznawano dwie etykiety za różne) było 0,02. Następnie próba produktów skanowanych została poddana filtrowaniu i w ten sposób usunięto z niej zarówno produkty charakteryzujące się ekstremalnymi miesięcznymi zmianami cen (3% przypadków), jak i produkty o relatywnie niskiej sprzedaży (nawet do 25% produktów w zależności od grupy). W przypadku danych skrapowanych zaimplementowano jedynie filtr ekstremalnych cen, z progami odcięcia 0,25 i 3 dla ilorazu cen z marca i lutego, co właściwie nie wpłynęło na wielkość próby (usunięto zaledwie dwa produkty z grupy jogurtów). Należy wyraźnie zaznaczyć, że pojęcie miesięcznej ceny jest inne w przypadku danych skanowanych i skrapowanych oraz odbiega od ceny reprezentanta, jaką ankieter notuje danego dnia w wybranym do badania punkcie sprzedaży. W przypadku danych skanowanych jako średnią cenę przyjmuje się wartości stanowiące iloraz łącznej wartości sprzedaży danego produktu i sumarycznej ilości jego sprzedaży z analizowanego miesiąca (ang. *unit value*). Z kolei w przypadku danych skrapowanych, które pobierane są każdego dnia miesiąca (bez względu na to, czy produkt się sprzedał czy nie), wyznacza się średnią arytmetyczną ze wszystkich obserwacji pozyskanych w danym miesiącu.

Przykładowe porównanie wyznaczonych w ten sposób cen dla marca 2021 r. pochodzących ze wszystkich omawianych źródeł danych przedstawiono wykr. 4.2, który uwzględnia cztery najliczniej reprezentowane grupy produktów (kawa ziarnista, jogurt naturalny, ryż długoziarnisty i mąka pszenna).

Wykr. 4.2. Wykresy pudełkowe dla cen wybranych czterech grup produktów spożywczych (porównanie trzech źródeł danych: marzec 2021 r.)



Źródło: opracowanie własne.

Generalnie ceny skrapowane zdają się charakteryzować największą zmiennością, a najmniejsza zmienność cen dotyczy danych pochodzących z tradycyjnej metody notowań cen przez ankieterów. Nie jest to jednak spostrzeżenie odnoszące się do wszystkich grup produktów, gdyż np. ceny ryżu długoziarnistego wykazują największe fluktuacje w przypadku danych skanowanych. Poza tym relatywnie najmniej

nietypowych wartości cen (ang. *outliers*) zaobserwowano w przypadku danych skanowanych, co nie jest zaskoczeniem, ponieważ dane skanowane przechodziły opisane wcześniej potrójne filtrowanie. Natomiast największe zaszumienie danych, mimo stosowanego filtru ekstremalnych cen na poziomie GTIN, dotyczyło cen skrapowanych. Wyjątek stanowił ryż długoziarnisty – w przypadku tego produktu wśród danych zebranych metodą tradycyjną zarejestrowano trzy wartości odstające cen w marcu. Nie ulega jednak wątpliwości, że różnice w średnich poziomach cen pozyskanych z analizowanych źródeł mogą być znaczne, przy czym ceny skrapowane (uśredniając) zdają się wyższe niż ceny z pozostałych źródeł danych (jedynym wyjątkiem jest ponownie ryż długoziarnisty).

4.7.4. Opis metodologii

W analizie odporności wskaźników cen detalicznych uwzględniono trzy źródła niepewności, czyli rodzaje założeń przyjmowanych przy ich konstrukcji, a mianowicie: wybór źródła danych o cenach detalicznych, wybór formuły indeksów do agregacji indeksów dla podgrup elementarnych do indeksów dla grup elementarnych w ramach poszczególnych źródeł danych, wybór formuły indeksu do agregacji wskaźników zmian cen w grupach elementarnych uzyskanych na podstawie różnych źródeł w indeksy łączne dla każdej z tych grup elementarnych. W analizie zbadano odporność wskaźników cen detalicznych dla siedmiu grup elementarnych zaliczanych do działu żywność.

Indeksy cen detalicznych dla wymienionych grup elementarnych zostały oszacowane na podstawie sześciu różnych źródeł informacji o cenach detalicznych produktów należących do tych grup w lutym i marcu 2021 r., a mianowicie:

- notowań cen detalicznych dokonywanych przez ankieterów;
- danych z systemu informatycznego sieci handlowej (dane skanowane);
- cenników sieci handlowej (web scraping);
- łącznie z notowań cen detalicznych dokonywanych przez ankieterów i cenników sieci handlowej;
- łącznie z notowań cen dokonywanych przez ankieterów i danych z systemu informatycznego sieci handlowej;
- wszystkich trzech źródeł danych o cenach łącznie.

Indeksy cen detalicznych dla wymienionych grup elementarnych liczone są poprzez agregację indeksów cen dla podgrup elementarnych tych grup. W przypadku operowania danymi z notowań dokonywanych przez ankieterów oraz z cenników sieci handlowych nie dysponujemy informacjami o ilościach zakupionych produktów w ramach podgrup elementarnych (ani też o udziałach wartości zakupów podgrup elementarnych w wartości zakupów ich grupy elementarnej). W związku z powyższym agregacja wskaźników cen detalicznych dla podgrup elementarnych do

wskaźnika ich grupy elementarnej (dla porównywanych okresów 0 i t) przebiegała przy wykorzystaniu średniej geometrycznej nieważonej, czyli z zastosowaniem indeksu Jevonsa (IMF i in., 2004):

$$ {}_{G}P_{J}^{0,t} = \sqrt[n]{\prod_{i=1}^{n} P_{g_i}^{0,t}}, \tag{4.10} $$

gdzie $P_{g_i}^{0,t}$ – wskaźniki cen dla i-tej podgrupy elementarnej G-tej grupy; ${}_{G}P_{J}^{0,t}$ – indeks Jevonsa dla G-tej grupy elementarnej, $G = \{g_1, g_2, \ldots, g_n\}$.

Gdy operujemy danymi skanowanymi, dysponujemy informacjami nie tylko o cenach, ale i o ilościach produktów zakupionych w ramach poszczególnych podgrup elementarnych danej grupy elementarnej. Wagi te zostały wykorzystane do dokonania agregacji podgrup elementarnych również dla pozostałych źródeł danych. Pozwoliło to na wykorzystanie w agregacji wskaźników cen podgrup elementarnych nie tylko nieważonej formuły Jevonsa, ale także ważonych formuł indeksowych, a mianowicie indeksów Laspeyresa, Paaschego, Fishera oraz Törnqvista (IMF i in., 2004):

$$ {}_{G}P_{La}^{0,t} = \sum_{i=1}^{n} W_{g_i}^{0} P_{g_i}^{0,t}, \tag{4.11} $$

$$ {}_{G}P_{Pa}^{0,t} = \frac{1}{\sum_{i=1}^{n} W_{g_i}^{1} \frac{1}{P_{g_i}^{0,t}}}, \tag{4.12} $$

$$ {}_{G}P_{F}^{0,t} = \sqrt{{}_{G}P_{La}^{0,t} \cdot {}_{G}P_{Pa}^{0,t}}, \tag{4.13} $$

$$ {}_{G}P_{T}^{0,t} = \prod_{i=1}^{n} (P_{g_i}^{0,t})^{\frac{w_{g_i}^{0} + w_{g_i}^{t}}{2}}, \tag{4.14} $$

gdzie ${}_{G}P_{La}^{0,t}$, ${}_{G}P_{Pa}^{0,t}$, ${}_{G}P_{F}^{0,t}$, ${}_{G}P_{T}^{0,t}$ – indeksy cen odpowiednio: Laspeyresa, Paaschego, Fishera i Törnqvista dla G-tej grupy elementarnej; $w_{g_i}^{0}, w_{g_i}^{t}$ – waga i-tej podgrupy G-tej grupy elementarnej odpowiednio w okresie bazowym (luty 2021) i w okresie badanym (marzec 2021).

Waga danej podgrupy w okresie bazowym otrzymywana jest przez podzielenie łącznej wartości sprzedaży wszystkich produktów z tej podgrupy w okresie bazowym przez łączną wartość sprzedaży produktów ze wszystkich podgrup należących do danej grupy elementarnej. Analogicznie liczone są wagi dla podgrup elementarnych w okresie badanym (wartości wag zostały przedstawione w tabl. 4.4).

Tabl. 4.4. Wagi podgrup elementarnych (w obrębie odpowiednich grup elementarnych) dla danych z sieci handlowych

Grupy i podgrupy elementarne	Wagi	
	luty 2021	marzec 2021
RYŻ		
ryż długoziarnisty	0,654	0,678
ryż biały	0,346	0,322
MLEKO PEŁNE ŚWIEŻE		
mleko pełne UHT	0,474	0,499
mleko pełne pasteryzowane	0,526	0,501
MLEKO ŚWIEŻE NISKOTŁUSZCZOWE		
mleko niskotłuszczowe UHT	0,420	0,421
mleko kozie	0,034	0,036
mleko niskotłuszczowe pasteryzowane	0,546	0,544
JOGURT		
Actimel	0,100	0,097
jogurt owocowy	0,317	0,307
jogurt czekoladowy i orzechowy	0,006	0,006
jogurt pitny	0,282	0,294
jogurt naturalny	0,294	0,297
NAPOJE I INNE PRODUKTY MLECZNE		
kefir	0,211	0,224
maślanka	0,096	0,101
Monte	0,229	0,201
serek homogenizowany	0,464	0,473
CUKIER		
cukier trzcinowy	0,136	0,126
cukier biały	0,790	0,778
cukier puder	0,075	0,096
KAWA		
kawa rozpuszczalna	0,083	0,073
kawa ziarnista	0,546	0,512
kawa mielona	0,371	0,415

Źródło: opracowanie własne.

W sytuacji gdy wykorzystujemy do obliczeń indeksów cen więcej niż jedno źródło danych, wskaźnik cen dla danej grupy elementarnej obliczany jest poprzez agregację wskaźników cen dla tej grupy elementarnej uzyskiwanych na podstawie każdego ze źródeł danych niezależnie. Powyższą agregację przeprowadzono na podstawie indeksu Younga lub geometrycznego indeksu Younga (Białek, 2017):

$$ {}_{G}P_{Y}^{0,t} = ({}_{G}P_{N}^{0,t})^{w_N^{\tau}} + ({}_{G}P_{W}^{0,t})^{w_W^{\tau}} + ({}_{G}P_{S}^{0,t})^{w_S^{\tau}}, \tag{4.15} $$

$$ {}_{G}P_{GY}^{0,t} = ({}_{G}P_{N}^{0,t})^{w_N^{\tau}} ({}_{G}P_{W}^{0,t})^{w_W^{\tau}} ({}_{G}P_{S}^{0,t})^{w_S^{\tau}}, \tag{4.16} $$

gdzie w_N^{τ} , w_W^{τ} , w_S^{τ} to wagi indeksu dla G-tej grupy elementarnej, ustalone na bazie okresu bardziej odległego niż okres bazowy (w naszym badaniu 2020 r.), obliczone odpowiednio na podstawie notowań cen ankieterów (N), danych skrapowanych (W) oraz danych skanowanych (S); ${}_{G}P_{Y}^{0,t}$, ${}_{G}P_{GY}^{0,t}$ – odpowiednio indeks cen Younga i geometryczny indeks cen Younga dla G-tej grupy elementarnej.

Wagami dla indeksów cen obliczanych na podstawie więcej niż dwóch źródeł były udziały zakupów poszczególnych grup elementarnych z danego źródła w ich łącznych zakupach ze wszystkich źródeł łącznie. Udziały zakupu produktów przez Internet były szacowane na podstawie informacji na ten temat pochodzących z badania budżetów gospodarstw domowych prowadzonego przez GUS. Natomiast informacje o udziałach zakupów w sieciach handlowych uzyskano z baz danych Passport GMID, Euromonitor International oraz z badań rynku wewnętrznego prowadzonych przez GUS (tabl. 4.5).

Tabl. 4.5. Wagi grup elementarnych według miejsca (sposobu) zakupu

Grupy elementarne	Wagi		
	notowania ankieterów (sklepy bez sieci handlowych)	dane skanowane (sieci handlowe)	dane scrapowane (Internet)
Ryż	34,65	64,36	0,99
Mleko pełne świeże	34,86	64,73	0,41
Mleko świeże niskotłuszczowe	34,85	64,72	0,43
Jogurt	34,85	64,72	0,43
Napoje i inne produkty mleczne	34,84	64,71	0,45
Cukier	39,82	59,72	0,46
Kawa	41,28	57,00	1,72

Źródło: opracowanie własne.

Indeksy cen dla analizowanych grup elementarnych według źródeł danych oraz formuł indeksowych zawiera tabl. 4.6, natomiast tabl. 4.7 zawiera uzupełniającą informację o dynamice cen podgrup grup elementarnych. W naszej analizie wskaźnik cen detalicznych dla danej grupy elementarnej jest wskaźnikiem złożonym, a jego wskaźnikami bazowymi są wskaźniki cen detalicznych dla jego podgrup elementarnych otrzymane dla poszczególnych źródeł danych. W sytuacji gdy dane pochodzą z więcej niż jednego źródła, jego szacunek odbywa się w dwóch krokach. W pierwszym kroku obliczane są wskaźniki cen dla danej grupy elementarnej na podstawie

danych z każdego źródła niezależnie. W kolejnym kroku wskaźniki cen dla danej grupy elementarnej obliczane w pierwszym kroku agregowane są w jeden wskaźnik cen dla tej grupy elementarnej dla wszystkich wykorzystywanych źródeł danych łącznie.

Tabl. 4.6. Indeksy cen towarów konsumpcyjnych według źródeł danych o cenach detalicznych oraz formuł indeksowych

Źródło danych i kombinacje formuł indeksowych	Indeksy cen marzec 2021						
	ryż	mleko pełne świeże	mleko świeże nisko-tłuszczo-we	jogurt	napoje i inne produkty mleczne	cukier	kawa
Wszystkie źródła danych							
Jevons x Young	101,023	100,425	100,692	99,966	102,809	99,323	99,473
Jevons x Young geom.	101,017	100,424	100,692	99,964	102,802	99,323	99,472
Laspeyres x Young	100,498	100,484	100,882	100,430	102,391	98,511	99,048
Laspeyres x Young geom.	100,495	100,483	100,882	100,430	102,383	98,511	99,046
Paasche x Young	100,372	100,418	100,880	100,381	102,121	98,510	99,039
Paasche x Young geom.	100,370	100,416	100,880	100,381	102,116	98,509	99,038
Fisher x Young	100,435	100,451	100,881	100,406	102,256	98,511	99,043
Fisher x Young geom.	100,433	100,449	100,881	100,405	102,250	98,510	99,042
Törnqvist x Young	100,434	100,451	100,881	100,405	102,252	98,510	99,043
Törnqvist x Young geom.	100,432	100,449	100,881	100,405	102,246	98,510	99,042
Dane ankieterów i dane skanowane							
Jevons x Young	101,034	100,421	100,695	99,966	102,821	99,333	99,501
Jevons x Young geom.	101,028	100,419	100,695	99,965	102,814	99,333	99,501
Laspeyres x Young	100,504	100,480	100,885	100,432	102,403	98,530	99,061
Laspeyres x Young geom.	100,502	100,478	100,885	100,432	102,396	98,530	99,060
Paasche x Young	100,378	100,413	100,883	100,383	102,132	98,528	99,052
Paasche x Young geom.	100,376	100,412	100,883	100,383	102,127	98,528	99,051
Fisher x Young	100,441	100,446	100,884	100,407	102,267	98,529	99,056
Fisher x Young geom.	100,439	100,445	100,884	100,407	102,261	98,529	99,055
Törnqvist x Young	100,441	100,446	100,884	100,407	102,264	98,529	99,056
Törnqvist x Young geom.	100,438	100,445	100,884	100,407	102,258	98,529	99,055
Dane ankieterów i dane skrapowane							
Jevons x Young	99,550	99,735	100,936	100,681	101,190	99,064	99,588
Jevons x Young geom.	99,550	99,735	100,936	100,681	101,190	99,064	99,588
Laspeyres x Young	99,530	99,718	100,948	100,200	100,754	98,715	99,528
Laspeyres x Young geom.	99,530	99,718	100,948	100,200	100,754	98,714	99,527
Paasche x Young	99,527	99,734	100,948	100,171	100,750	98,706	99,528
Paasche x Young geom.	99,527	99,734	100,948	100,171	100,750	98,705	99,528
Fisher x Young	99,529	99,726	100,948	100,185	100,752	98,711	99,528
Fisher x Young geom.	99,529	99,726	100,948	100,185	100,752	98,709	99,527
Törnqvist x Young	99,529	99,726	100,948	100,185	100,752	98,711	99,528
Törnqvist x Young geom.	99,529	99,726	100,948	100,185	100,752	98,709	99,527

Tabl. 4.6. Indeksy cen towarów konsumpcyjnych według źródeł danych o cenach detalicznych oraz formuł indeksowych (dok.)

Źródło danych i kombinacje formuł indeksowych	Indeksy cen marzec 2021						
	ryż	mleko pełne świeże	mleko świeże nisko-tłuszczo-we	jogurt	napoje i inne produkty mleczne	cukier	kawa
Dane skanowane i dane skrapowane							
Jevons x Young	101,809	100,806	100,556	99,577	103,667	99,482	99,341
Jevons x Young geom.	101,808	100,806	100,556	99,577	103,666	99,482	99,340
Laspeyres x Young	101,015	100,906	100,841	100,552	103,258	98,345	98,674
Laspeyres x Young geom.	101,015	100,906	100,841	100,552	103,258	98,344	98,674
Paasche x Young	100,825	100,795	100,838	100,492	102,847	98,347	98,661
Paasche x Young geom.	100,825	100,795	100,838	100,492	102,847	98,346	98,661
Fisher x Young	100,920	100,850	100,839	100,522	103,052	98,346	98,667
Fisher x Young geom.	100,920	100,850	100,839	100,522	103,052	98,345	98,667
Törnqvist x Young	100,919	100,850	100,839	100,522	103,048	98,346	98,667
Törnqvist x Young geom.	100,919	100,850	100,839	100,522	103,047	98,345	98,667
Dane ankieterów							
Jevons x Young	99,540	99,714	100,947	100,692	101,204	99,084	99,660
Jevons x Young geom.	99,540	99,714	100,947	100,692	101,204	99,084	99,660
Laspeyres x Young	99,522	99,697	100,959	100,202	100,768	98,763	99,579
Laspeyres x Young geom.	99,522	99,697	100,959	100,202	100,768	98,763	99,579
Paasche x Young	99,519	99,713	100,958	100,173	100,763	98,756	99,577
Paasche x Young geom.	99,519	99,713	100,958	100,173	100,763	98,756	99,577
Fisher x Young	99,520	99,705	100,958	100,188	100,765	98,760	99,578
Fisher x Young geom.	99,520	99,705	100,958	100,188	100,765	98,760	99,578
Törnqvist x Young	99,520	99,705	100,958	100,188	100,765	98,760	99,578
Törnqvist x Young geom.	99,520	99,705	100,958	100,188	100,765	98,760	99,578
Dane skanowane							
Jevons x Young	101,838	100,801	100,559	99,576	103,692	99,499	99,385
Jevons x Young geom.	101,838	100,801	100,559	99,576	103,692	99,499	99,385
Laspeyres x Young	101,033	100,901	100,846	100,556	103,283	98,374	98,685
Laspeyres x Young geom.	101,033	100,901	100,846	100,556	103,283	98,374	98,685
Paasche x Young	100,840	100,790	100,842	100,496	102,869	98,377	98,671
Paasche x Young geom.	100,840	100,790	100,842	100,496	102,869	98,377	98,671
Fisher x Young	100,937	100,846	100,844	100,526	103,076	98,376	98,678
Fisher x Young geom.	100,937	100,846	100,844	100,526	103,076	98,376	98,678
Törnqvist x Young	100,936	100,846	100,844	100,525	103,071	98,376	98,678
Törnqvist x Young geom.	100,936	100,846	100,844	100,525	103,071	98,376	98,678
Dane skrapowane							
Jevons x Young	99,885	101,522	100,060	99,757	100,104	97,319	97,865
Jevons x Young geom.	99,885	101,522	100,060	99,757	100,104	97,319	97,865
Laspeyres x Young	99,831	101,562	100,098	99,987	99,715	94,493	98,292
Laspeyres x Young geom.	99,831	101,562	100,098	99,987	99,715	94,493	98,292
Paasche x Young	99,823	101,521	100,098	99,978	99,709	94,434	98,339
Paasche x Young geom.	99,823	101,521	100,098	99,978	99,709	94,434	98,339
Fisher x Young	99,827	101,542	100,098	99,983	99,712	94,464	98,316
Fisher x Young geom.	99,827	101,542	100,098	99,983	99,712	94,464	98,316
Törnqvist x Young	99,827	101,542	100,098	99,983	99,712	94,462	98,316
Törnqvist x Young geom.	99,827	101,542	100,098	99,983	99,712	94,462	98,316

Źródło: opracowanie własne.

Tabl. 4.7. Indeksy cen towarów konsumpcyjnych dla podgrup grup elementarnych

Grupy i podgrupy elementarne	Indeksy cen marzec 2021		
	notowania ankieterów	dane skanowane	dane scrapowane
RYŻ			
ryż długoziarnisty	99,48	99,14	99,71
ryż biały	99,60	104,61	100,06
MLEKO PEŁNE ŚWIEŻE			
mleko pełne UHT	100,07	99,14	100,80
mleko pełne pasteryzowane	99,36	102,49	102,25
MLEKO ŚWIEŻE NISKOTŁUSZCZOWE			
mleko niskotłuszczowe UHT	100,84	100,57	100,00
mleko kozie	100,95	100,00	100,00
mleko niskotłuszczowe pasteryzowane	101,05	101,11	100,18
JOGURT			
Actimel	103,25	106,16	100,15
jogurt owocowy	99,61	100,49	100,59
jogurt czekoladowy i orzechowy	–	92,31	98,75
jogurt pitny	99,04	99,62	100,34
jogurt naturalny	100,92	99,79	98,97
NAPOJE I INNE PRODUKTY MLECZNE			
kefir	101,90	101,45	100,01
maślanka	102,01	102,14	101,31
Monte	101,08	110,91	100,00
serek homogenizowany	99,84	100,59	99,11
CUKIER	**99,08**	**99,50**	**97,32**
cukier trzcinowy	99,54	100,83	101,17
cukier biały	98,63	97,81	93,02
cukier puder	–	99,88	97,94
KAWA			
kawa rozpuszczalna	99,77	101,17	96,56
kawa ziarnista	99,55	98,27	98,13
kawa mielona	–	98,74	98,92

Źródło: opracowanie własne.

Od strony formalnej zależność wskaźnika cen detalicznych dla danej grupy elementarnej od zmiennych wejściowych reprezentujących założenia przyjmowane przy jego konstrukcji możemy przedstawić za pomocą następującego modelu:

$$ {}_{G}P^{0,t} = f_{rs}\left({}_{G}P_{Z}^{0,t},\ {}_{G}w_{Z}^{0,t} \right), \tag{4.17} $$

gdzie ${}_{G}P^{0,t}$ – wartość wskaźnika cen detalicznych dla G-tej grupy elementarnej; ${}_{G}P_{Z}^{0,t}$ – wektor indeksów cen dla podgrup produktów z G-tej grupy elementarnej, uzyskanych na podstawie różnych źródeł danych o cenach ($Z = N, W, S$); ${}_{G}w_{Z}^{0,t}$ – wektor wag indeksów cen (indeksy Laspeyresa, Paaschego, Fishera i Törnqvista) dla

podgrup produktów z G-tej grupy elementarnej, uzyskany na podstawie różnych źródeł danych o cenach ($Z = N, W, S$); f_{rs} – funkcja transformująca indeksy cen detalicznych dla podgrup elementarnych obliczanych na podstawie różnych źródeł danych o cenach detalicznych oraz wag przypisanych wskaźnikom cen dla podgrup elementarnych i grup elementarnych przy wykorzystaniu s-tego źródła informacji o cenach detalicznych (s-tej kombinacji źródeł danych o cenach detalicznych, gdy dane pochodzą z różnych źródeł) oraz r-tej formuły indeksowej (r-tej kombinacji formuł indeksowych, gdy dane pochodzą z więcej niż jednego źródła).

W ramach analizy niepewności i analizy wrażliwości przedmiotem naszego zainteresowania są zmiany wartości wskaźnika złożonego, którym jest wskaźnik cen detalicznych wyróżnionych w badaniu grup elementarnych na skutek zmian założeń jego szacunku.

4.7.4.1. Idea analizy niepewności

Możemy wyróżnić dwie podstawowe grupy metod analizy niepewności, a mianowicie metody probabilistyczne i metody deterministyczne. Metody probabilistyczne opierają się na symulacjach przeprowadzanych przy różnych założeniach dotyczących zasad konstrukcji wskaźnika złożonego (u nas – indeksu cen detalicznych dla danej grupy elementarnej), które stanowią wejściowe czynniki niepewności. Model symulacyjny powinien być tak skonstruowany, aby odtwarzać probabilistyczny charakter badanego zjawiska (OECD, 2008; Panek, 2016; Saisana i in., 2005).

Najczęściej stosowaną w praktyce metodą probabilistyczną analizy niepewności jest metoda Monte Carlo (OECD, 2008; Panek, 2016; Saisana i in., 2005). Podejście Monte Carlo polega na wielowymiarowej ocenie sformułowanego modelu z pseudolosowo wybranymi parametrami (zmiennymi wejściowymi określającymi założenia przyjmowane w konstrukcji wskaźnika złożonego). Procedura składa się z kilku etapów. W pierwszym kroku wyznaczamy funkcję rozkładu prawdopodobieństwa dla każdej zmiennej wejściowej (każdego założenia) X_k, $k = 1,2, \dots, z$. W naszym badaniu pierwsza zmienna wejściowa (X_1) reprezentuje wybór źródła danych o cenach detalicznych, druga zmienna wejściowa (X_2) określa formuły indeksowe do agregacji indeksów dla podgrup elementarnych w ramach poszczególnych źródeł danych w indeksy dla grup elementarnych, a trzecia (X_3) – formuły indeksowe użyte do agregacji indeksów dla poszczególnych grup elementarnych uzyskanych na podstawie różnych źródeł danych w indeksy łączne dla każdej z tych grup elementarnych. Wszystkie te założenia są zmiennymi losowymi o rozkładzie skokowym. Następnie generujemy w sposób losowy N kombinacji niezależnych zmiennych wejściowych X^l, $l = 1,2, \dots, N$. Zbiór $X^l = X_1^l, X_2^l, \dots, X_k^l$ kombinacji zmiennych wejściowych tworzy próbę. Generowanie takiej próby może odbywać się różnymi meto-

dami próbkowania, takimi jak próbkowanie losowe proste (*simple random sampling*), próbkowanie warstwowe (ang. *stratified sampling*) czy też próbkowanie quasi-losowe (ang. *quasi-random sampling*); (Saltelli i in., 2000). Dla każdej próby (kombinacji założeń) oceniamy nasz model, obliczając wartości zmiennych wyjściowych (w naszy badaniu – wartości indeksu cen detalicznych). Ciągi wektorów zmiennych wyjściowych $\boldsymbol{Y}^l$, $l = 1,2,\dots,N$ umożliwiają oszacowanie empirycznych rozkładów prawdopodobieństwa poszczególnych zmiennych wyjściowych oraz ich parametrów, takich jak wariancja i momenty wyższych rzędów.

W metodach deterministycznych szacujemy wartość wskaźnika złożonego (u nas indeksu cen detalicznych danej grupy elementarnej) dla wszystkich możliwych N kombinacji niezależnych założeń $\boldsymbol{X}^l$, $l = 1,2,\dots,N$. Dla każdej z tych kombinacji obliczamy wartość indeksu cen detalicznych dla danej grupy elementarnej (por. tabl. 4.6). Wartości tego indeksu tworzą jego rozkład wykorzystywany do oceny odporności wskaźnika cen detalicznych danej grupy elementarnej na zmiany założeń jego konstrukcji dotyczących źródła danych o cenach detalicznych oraz formuł indeksowych zastosowanych do jego konstrukcji.

4.7.4.2. Idea analizy wrażliwości

Celem analizy wrażliwości jest określenie, w jakich proporcjach przyjmowane przy konstrukcji wskaźnika cen detalicznych założenia (zmienne wejściowe) wpływają na jego wartości.

W sytuacji gdy kilka źródeł niepewności jest uwzględnianych jednocześnie do modelowania wskaźnika złożonego, może być stosowany model nieliniowy. Dobre rezultaty daje wtedy zastosowanie w analizie wrażliwości metod bazujących na analizie wariancji (Chan i in., 2000; Panek, 2016; Saltelli i in., 2008). Podobnie jak w przypadku analizy niepewności w analizie wrażliwości możemy stosować, w zależności od charakteru danych, podejście probabilistyczne albo podejście deterministyczne. W podejściu probabilistycznym wskaźniki wrażliwości najczęściej szacowane są metodą Sobola (1993), zmodyfikowaną przez Saltelliego (2002). Metoda Sobola, jak już wspomniano, stosuje quasi-losową metodę próbkowania dla wyznaczania rozkładów zmiennych wejściowych. Wrażliwość wskaźnika złożonego (indeksu cen detalicznych dla danej grupy elementarnej) na parametry (zmienne wejściowe, czyli założenia przyjmowane przy jego konstrukcji) oceniana jest na podstawie współczynników wrażliwości. Współczynniki wrażliwości są obliczane na podstawie dekompozycji wariancji zmiennej wyjściowej (w naszym badaniu – indeksu cen detalicznych):

$$D^2(Y) = \sum_{k=1}^{z} V_k + \sum_{k=1}^{z} \sum_{\substack{k'=1 \\ k<k'}}^{z} V_{k,k'} + \ldots + V_{1\ldots z}, \tag{4.18}$$

gdzie:

$$V_k = D^2_{X_k}\left[E_{X_{-k}}(Y|X_k)\right], \tag{4.19}$$

$$V_{kk'} = D^2_{X_k X_{k'}}\left[E_{X_{-kk'}}(Y|X_k, X_{k'})\right] - D^2_{X_k}\left[E_{X_{-k}}(Y|X_k)\right] - D^2_{X_{k'}}\left[E_{X_{-k'}}(Y|X_{k'})\right]. \tag{4.20}$$

Pierwszy z elementów równania (4.20) stanowi ocenę bezpośredniego wpływu na wariancję zmiennej wyjściowej zmiennej wejściowej X_k, czyli w naszym przykładzie zmiennej opisującej wybór konkretnej metody na danym etapie konstrukcji wskaźnika złożonego. Drugi element równania (4.20) określa wpływ interakcji pomiędzy k-tą i k'-tą zmienną wejściową na wariancję zmiennej Y (wpływ pośredni zmiennej X_k na wariancję zmiennej Y).

Bezpośredni wpływ zmiennych wejściowych na kształtowanie się wartości wskaźnika złożonego (przy założeniu braku interakcji pomiędzy zmiennymi wejściowymi) mierzony jest współczynnikami wrażliwości pierwszego rzędu:

$$S_k = \frac{V_k}{V} = \frac{D^2_{X_k}\left[E_{X_{-k}}(Y|X_k)\right]}{D^2(Y)}, \quad k = 1,2,\ldots,z \tag{4.16}$$

Model bez interakcji pomiędzy zmiennymi wejściowymi nazywany jest modelem addytywnym. W tym przypadku suma wszystkich współczynników wrażliwości pierwszego rzędu równa jest jedności ($\sum_{k=1}^{z} S_k = 1$).

Dla modelu nieaddytywnego należy oszacować wskaźniki wrażliwości wyższego rzędu, mierzące wpływ na wariancję zmiennej wyjściowej interakcji pomiędzy zmiennymi wejściowymi. Jednakże w praktyce są one rzadko obliczane, gdy w przypadku modelu z k zmiennymi wejściowymi liczba wskaźników wrażliwości koniecznych do oszacowania wynosiłaby aż $2^k - 1$. Z tej przyczyny wpływ na wariancję zmiennej wyjściowej interakcji pomiędzy zmiennymi wejściowymi obliczany jest w sposób pośredni. W pierwszym kroku obliczane są wówczas współczynniki wrażliwości całkowitej mierzące łączny wpływ na wartość zmiennej wyjściowej zmiennej wejściowej X_k, tzn. zarówno w sposób bezpośredni, jak i w interakcjach ze wszystkimi możliwymi kombinacjami pozostałych zmiennych wejściowych:

$$S_k^T = \frac{D^2(Y) - D_{X_{-k}}^2\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = \frac{E_{X_{-k}}\left(D_{X_k}^2(Y|X_{-k})\right)}{D^2(Y)}. \quad (4.22)$$

W prowadzonej analizie wrażliwości uwzględniane są trzy typy założeń, a tym samym możemy obliczyć trzy następujące wskaźniki wrażliwości całkowitej:

$$S_1^T = \frac{D^2(Y) - D_{X_{-k}}^2\left[E_{X_k}(Y|X_{-k})\right]\left[E_{X_k}(Y|X_{-k})\right]}{D^2(Y)} = S_1 + S_{12} + S_{13} + S_{123}, \quad (4.23)$$

$$S_2^T = S_2 + S_{23} + S_{23} + S_{123}, \quad (4.24)$$

$$S_3^T = S_3 + S_{13} + S_{23} + S_{123}. \quad (4.25)$$

Ostatecznie wpływ na wariancję zmiennej wyjściowej interakcji pomiędzy zmiennymi wejściowymi (wpływ pośredni zmiennych wejściowych) obliczamy jako różnicę pomiędzy ich wpływami łącznymi i wpływami pośrednimi ($S_k^T - S_k$). Znaczne różnice pomiędzy S_k^T oraz S_k wskazują na istotną rolę w kształtowaniu zmiennej Y interakcji k-tej zmiennej wejściowej z innymi zmiennymi wejściowymi, co świadczy o znacznej zależności między zmiennymi wejściowymi. Analiza interakcji pomiędzy zmiennymi wejściowymi jest pomocna w zrozumieniu struktury modelu. Metody analizy wrażliwości bazujące na analizie wariancji zarówno dla zmiennych wejściowych niezależnych, jak i zależnych zostały przedstawione m.in. w publikacji Saltelli i in. (2008).

W podejściu deterministycznym wskaźniki wrażliwości są obliczane na podstawie wartości indeksów cen dla poszczególnych grup elementarnych, oszacowane dla każdej kombinacji źródeł danych o cenach detalicznych oraz zastosowanych formuł indeksowych.

4.7.5. Wyniki

4.7.5.1. Analiza niepewności

W przeprowadzonym badaniu zastosowano podejście deterministyczne ze względu na to, że liczba założeń jest stosunkowo niewielka i można rozważyć wszystkie ich kombinacje.

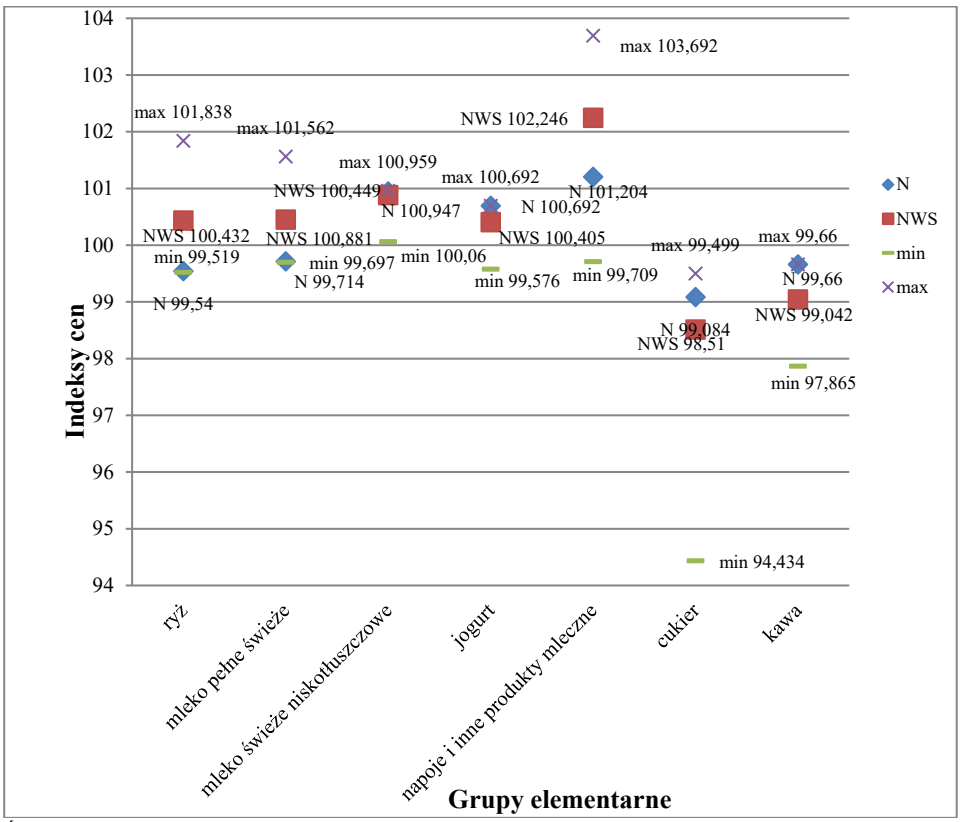
Wartości indeksów cen dla poszczególnych grup elementarnych, które zostały obliczone dla każdej kombinacji źródeł danych o cenach detalicznych oraz zastosowanych formuł indeksowych, zawiera tabl. 4.6. Na wykr. 4.3 dla każdej grupy elementarnej przedstawiono:

- wartość indeksu oszacowanego na podstawie obecnie stosowanej metodyki GUS (źródłem danych są notowania cen dokonywane przez ankieterów, agregacja wskaźników cen detalicznych dla podgrup elementarnych w ramach grupy elementarnej przebiega z zastosowaniem indeksu Jevonsa);
- wartość indeksu uzyskanego na podstawie wszystkich źródeł informacji łącznie oraz preferowanych formuł agregacji (agregacja wskaźników cen dla podgrup elementarnych do wskaźnika cen ich grupy elementarnej przebiegała za pomocą indeksu Törnqvista, agregacja wskaźników cen dla danej grupy elementarnej, uzyskanych na podstawie różnych źródeł danych, w jeden wskaźnik została przeprowadzona z wykorzystaniem geometrycznego indeksu Younga);
- wartość najwyższą i najniższą dla rozważanych indeksów.

Należy zaznaczyć, że w analizie odporności wskaźników cen detalicznych prezentujemy, ze względu na dostępność danych, zmiany cen detalicznych, jakie zaszły w ciągu miesiąca. W sytuacji wysokiej inflacji w Polsce odnotowywanej zarówno w 2000 r., jak i w 2001 r. różnice w ocenach inflacji w okresach rocznych za pomocą wskaźników cen bazujących na różnych źródłach informacji i różnych formułach wskaźników cen będą na pewno o wiele większe niż dla okresu miesięcznego (zakładając, że trend zmian zaobserwowany w okresie luty 2021–marzec 2021 utrzymałby się przez cały rok, zmiany cen w okresie rocznym byłyby 12-krotnie wyższe).

Największy rozstęp indeksów cen detalicznych (różnica pomiędzy największą i najmniejszą wartością z oszacowanych wartości indeksów cen), na podstawie kombinacji danych z różnych źródeł i różnych formuł, otrzymano dla grupy elementarnej cukier – 5,1 p.proc. Natomiast najmniejszy rozstęp indeksów cen obserwujemy dla grupy elementarnej mleko świeże niskotłuszczowe – 0,9 p.proc.

Wykr. 4.3. Analiza niepewności dla indeksów cen detalicznych



Źródło: opracowanie własne na podstawie danych z tabl. 4.6.

4.7.5.2. Analiza wrażliwości

W przeprowadzonym badaniu do analizy wrażliwości zastosowano podejście deterministyczne, w którym wskaźniki wrażliwości zostały obliczone na podstawie wartości indeksów cen dla poszczególnych grup elementarnych oszacowanych dla każdej kombinacji źródeł danych o cenach detalicznych oraz zastosowanych formuł indeksowych. W sumie rozpatrzono siedem wariantów źródeł danych, pięć formuł indeksowych dla agregacji indeksów dla podgrup w indeksy dla grup elementarnych oraz dwie formuły indeksowe dla agregacji indeksów uzyskanych z różnych źródeł danych. Dało to w sumie 70 kombinacji założeń.

W modelu deterministycznym całkowita wariancja V w modelu jest szacowana jako wariancja z wszystkich obliczonych wartości indeksu przy wszystkich kombinacjach założeń. W celu obliczenia wariancji związanej bezpośrednio z danym założeniem V_k (wzór 4.19) należy dla każdej możliwej wartości tego założenia obliczyć średnią wartość indeksu końcowego. Otrzymamy tyle średnich, ile wariantów

przyjmuje dane założenie. Wariancja z obliczonych średnich stanowi estymator wartości V_k.

W celu obliczenia wartości $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$ (wzór 4.22) dla k-tego założenia należy rozważyć wszystkie możliwe kombinacje pozostałych założeń (oznaczone jako k). Dla każdej takiej kombinacji należy obliczyć wariancję dla wartości indeksu końcowego. Indeksy te w ramach kombinacji będą się w swojej budowie między sobą różnić jedynie założeniem k. Średnia wartość ze wszystkich tak obliczonych wariancji stanowi estymator dla $E_{X_{-k}}\left(D^2_{X_k}(Y|X_{-k})\right)$.

Wyniki przeprowadzonych obliczeń zawiera tabl. 4.8. Wybór źródła danych odpowiadał w sposób bezpośredni aż za 85% i więcej całkowitej zmienności dla wszystkich źródeł danych poza jogurtem. Wpływ formuły indeksowej użytej do agregacji indeksów dla podgrup elementarnych w indeksy dla grup elementarnych odpowiadał za jedynie 0,09% zmienności dla mleka, 3% dla napojów i innych produktów mlecznych, 4% dla mleka świeżego, 5,4% dla kawy, 6,9% dla ryżu, 9,3% dla cukru oraz 16,5% dla jogurtu. Dla wszystkich grup elementarnych wpływ formuły indeksowej służącej do agregacji indeksów pochodzących z różnych źródeł danych w jeden indeks łączny był znikomy – mniejszy od 0,0001% całkowitej wariancji.

Tabl. 4.8. Wskaźniki wrażliwości indeksów cen konsumpcyjnych

Grupy elementarne	Wartości wskaźników wrażliwości w %								
	źródło danych o cenach			formuły indeksowe – podgrupy elementarne			formuły indeksowe – grupy elementarne		
	S_1	S_1^{T}	$S_1^{\mathrm{T}} - S_1$	S_2	S_2^T	$S_2^{\mathrm{T}} - S_2$	S_3	S_3^{T}	$S_3^{\mathrm{T}} - S_3$
Ryż	88,22	107,09	18,87	6,88	14,51	7,64	0,00004	0,00035	0,00031
Mleko pełne tłuste	99,82	114,89	15,07	0,09	0,22	0,13	0,00001	0,00008	0,00007
Mleko świeże	93,44	110,34	16,89	4,06	8,08	4,02	0,00000	0,00000	0,00000
Jogurt	21,89	96,01	74,12	16,52	96,24	79,73	0,00000	0,00007	0,00007
Napoje i inne produkty mleczne	96,69	111,51	14,82	3,04	4,08	1,04	0,00006	0,00036	0,00031
Cukier	85,45	104,28	18,83	9,32	17,93	8,61	0,00000	0,00001	0,00000
Kawa	84,42	108,75	24,33	5,44	19,20	13,77	0,00001	0,00006	0,00005

Źródło: opracowanie własne.

Dla wszystkich grup elementarnych i wszystkich rozpatrywanych założeń wartości wpływu łącznego (4,22) są wyższe niż wartości wpływu bezpośredniego (4,19). Im większe różnice między tymi dwoma wartościami, tym większy udział interakcji pomiędzy założeniami w kształtowaniu całkowitej zmienności wartości indeksów końcowych. Wśród rozpatrzonych grup elementarnych uwagę zwraca jogurt, gdzie wartości wpływu całkowitego dla dwóch pierwszych założeń wynoszą ponad 90%, a wartości ich wpływu bezpośredniego są równe odpowiednio 21,9% i 16,5%. Wskazuje to na wchodzenie w silne interakcje tych dwóch założeń w przypadku

rozważanej grupy elementarnej. Oznacza to, że o ile zmiana jednego z nich z osobna nie prowadzi do znacznych zmian wartości indeksu końcowego, o tyle istnieją takie kombinacje ich wartości, które mogą prowadzić do otrzymania relatywnie różnych wartości tego indeksu.

4.7.6. Wnioski

Oceny inflacji dokonane na podstawie aktualnie stosowanej przez GUS metodyki są wyższe niż oceny uzyskane na podstawie preferowanej metodyki dla czterech grup elementarnych (mleko świeże niskotłuszczowe, jogurt, cukier i kawa), a niższe dla trzech pozostałych analizowanych grup elementarnych (ryż, mleko pełne świeże oraz napoje i inne produkty mleczne). Największe niedoszacowanie wielkości inflacji za pomocą aktualnej metodyki w stosunku do preferowanej metodyki wystąpiło w przypadku grupy elementarnej napoje i inne produkty mleczne – wyniosło ono 1 p.proc. Natomiast największe przeszacowanie inflacji obserwujemy w grupie elementarnej kawa – wyniosło ono 0,6 p.proc.

Uzyskane wyniki świadczą o stosunkowo małej odporności indeksów cen detalicznych dla analizowanych grup elementarnych na zmiany założeń ich obliczeń dotyczących źródeł danych o cenach detalicznych oraz relatywnie wysokiej odporności na dobór formuł indeksowych do szacunku wskaźników cen detalicznych na poziomie podgrup i grup elementarnych.

Dla wszystkich grup elementarnych założenie pierwsze, tj. dobór źródła danych, w najwyższym stopniu determinuje końcową wartość indeksu cen. Dobór formuły indeksowej służącej do agregacji indeksów dla podgrup elementarnych w indeksy dla grup elementarnych już w zdecydowanie mniejszym stopniu wpływa na końcowe wyniki. Natomiast wpływ wyboru indeksu służącego do agregacji indeksów pochodzących z różnych źródeł danych jest znikomy.

Uzyskane wyniki są jedynie wstępną diagnozą wpływu nowych źródeł danych na pomiar dynamiki cen detalicznych. Wnioski nie mogą być w żaden sposób uogólniane, ponieważ po pierwsze dotyczą jedynie dwóch miesięcy obserwacji, a po drugie tylko wybranych grup elementarnych z kategorii spożywczej. Niniejsza analiza stanowi jednak pewien prolog i jednocześnie przyczynek do dalszych, szerszych badań w zakresie analizy odporności wskaźników cen w kontekście pomiaru inflacji.

Stosunkowo mała stabilność wyników ocen dynamiki cen detalicznych, przy zmianach założeń dotyczących pomiaru tej dynamiki, wskazuje na konieczność zachowania dużej ostrożności przy włączaniu do pomiaru nowych źródeł danych. Zmiany cen detalicznych mogą bowiem w dużym stopniu być wtedy wynikiem nie tyle rzeczywistych zmian tych cen, ile właśnie wprowadzenia do obliczeń tych zmian nowych źródeł danych.

ROZDZIAŁ 5
Zastosowanie nowoczesnych technologii informatycznych w korzystaniu z alternatywnych źródeł danych

5.1. Zbieranie i archiwizacja danych skrapowanych

Niniejszy podrozdział stanowi wprowadzenie do tematyki pająków internetowych i ich rodzajów, a także przedstawia specjalne oprogramowanie stworzone w ramach naszego projektu w celu zbierania cenowych danych internetowych.

5.1.1. Internet jako źródło danych o cenach detalicznych

Dostępne źródła danych można podzielić na dwa główne kanały:

- zewnętrzni gestorzy danych:
  – duże sieci handlowe (np. Lidl, Biedronka);
  – sieci franczyzowe (np. Żabka);
  – hipermarkety (np. Tesco, Auchan, Real);
- internet:
  – duże sklepy internetowe (np. www.morele.net, www.euro.com.pl);
  – porównywarki cen (np. www.ceneo.pl, www.skapiec.pl).

W przypadku kanału internetowego pierwszym wyzwaniem jest identyfikacja dostępnych sklepów internetowych. Nie ma bowiem spisu wszystkich sklepów funkcjonujących w sieci. Listę takich sklepów można stworzyć na podstawie następujących źródeł:

- porównywarki cen – serwisy tego typu posiadają znaczną liczbę dostępnych sklepów internetowych. Niestety adresy sklepów nie są łatwo dostępne i dla każdego niezbędne jest opracowanie lub dostosowanie osobnego systemu ekstrakcji stosownych informacji;
- korpus dokumentów internetowych – dzięki korpusowi internetowemu możliwe jest opracowanie prostego algorytmu regułowego, który z dużą skutecznością wy-

kryje serwisy internetowe będące sklepami internetowymi. Korpus dokumentów polskiego internetu można uzyskać z IPI PAN (wyszukiwarka NEKST) lub z ogólnodostępnego zrzutu danych internetowych Common Crawl[1];
- m.in. w IPI PAN opracowano oprogramowanie pozwalające na identyfikację witryn internetowych, które z dużym prawdopodobieństwem reprezentują sklepy internetowe.

Ważnym zagadnieniem związanym z zebieraniem danych o cenach w sieci jest ocena wiarygodności dostępnych źródeł. Z uwagi na dużą łatwość w tworzeniu treści internetowych należy się spodziewać sztucznie tworzonych sklepów z niewiarygodną informacją cenową. Umiejętność prawidłowej oceny wiarygodności witryn internetowych zajmuje poczesne miejsce w kreowaniu umiejętności informacyjnych (ang. *information literacy*; Czerwiński, 2019).

W celu ograniczenia powyższego problemu proponuje się m.in. następujące rozwiązania:
- ocena popularności sklepów względem ich globalnej popularności w internecie – wykorzystanie miary PageRank, intensywności ruchu w serwisie (dane z serwisu Gemius);
- ręczny wybór ograniczonej liczby zweryfikowanych sklepów internetowych.

Ponadto mamy do czynienia z wyzwaniami technologicznymi. W niniejszej monografii przedstawimy najważniejsze informacje dotyczące tej tematyki. Rekomendujemy jednakże pogłębienie wybranych zagadnień na podstawie dostępnej literatury. Polecamy m.in. pracę Bonato (2008), która jest krótkim wprowadzeniem do podstawowych własności sieci WWW, pracę Langville i Meyera (2006), która zawiera przegląd metod rangowania wartości stron WWW, oraz pracę Kłopotka i in. (2007), która zajmuje się metodami wizualizacji zawartości sieci WWW. Ponadto warto zapoznać się z systemami zbierania stron WWW. Na przykład Boldi i in. (2018) opisują nowego, publicznie dostępnego pająka (ang. *crawler*), czyli system masowego zbierania informacji ze stron WWW, w formie oprogramowania open-source napisanego w języku Java. Pojedyncza instancja tego systemu jest zdolna zbierać tysiące stron na sekundę przy przestrzeganiu reguł etykiety internetowej. System w przeciwieństwie do innych istniejących rozwiązań open-source nie jest oparty na technikach wsadowych (jak MapReduce), lecz na nowoczesnych protokołach rozpraszania zapewniających wysoką wydajność.

Ściągnięcie stron WWW sklepów to dopiero połowa sukcesu. Potrzebne są metody ekstrakcji informacji o produktach. Pomocne mogą być w tym celu systemy takie jak HTTrack[2], do ściągania stron, czy Scrapy[3] do skrapowania. Korzystanie z nich

[1] https://commoncrawl.org.
[2] https://www.httrack.com/.
[3] https://scrapy.org/.

nie zwalnia jednak z obowiązku dostarczenia własnego wkładu koncepcyjno-programistycznego z uwagi na konieczność dostosowania narzędzi do indywidualnych metod prezentacji informacji cenowych na stronach poszczególnych sklepów.

5.1.2. Pająki internetowe

Pająk to program (skrypt), który metodycznie i w sposób automatyczny porusza się po stronach internetowych (Kobayashi i Takeda, 2000; Manning i in., 2008). Przeszukiwanie zaczyna się od zdefiniowanej puli początkowych adresów URL, po czym pająk przechodzi do kolejnych lokalizacji przez zawarte na stronach hiperłącza (Yu i in., 2020). Celem pracy pająka jest ściągnięcie z internetu do własnego serwera jak najwięcej dokumentów, które zawierają interesującą treść.

Pająki określa się także nazwami: zbieracz, pełzacz, bot, robot internetowy, (web)robot, (web)*crawler*, (web)*spider*, *warm*, *ant*.

Pająki są zwykle powiązane z innymi modułami, np. w wyszukiwarce internetowej z indekserem i analizatorem, które pomagają przydzielać kolejne zadania pająkom.

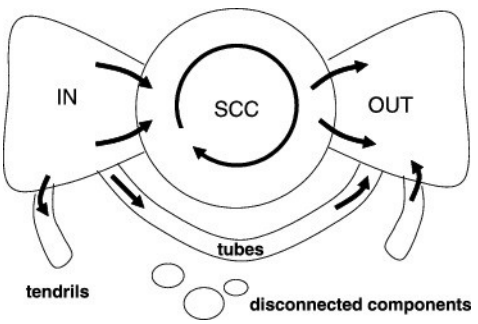
Pająki spełniają swe zadanie dzięki szczególnej strukturze sieci internetowej. Sieć jest spełnieniem XIX-wiecznej koncepcji książki, którą czytelnik miał czytać losowo. Książka wedle tego pomysłu miała się składać z leksji oraz łączników. Leksja to zamknięta w sobie historia obejmująca jedną lub dwie strony tekstu. Łączniki miały wskazywać czytelnikowi rekomendowane inne leksje do dalszego czytania. Sieć powstała dzięki integracji tej idei z technologią nawigacji za pomocą myszki, która umożliwiła czytelnikowi w łatwy sposób przechodzenie do rekomendowanych dalszych stron. Dokumenty HTML odgrywają tu rolę leksji, a linki – łączników.

Dość losowo przebiegający proces powstawania sieci stron HTML (czyli sieci WWW) doprowadził do powstania struktury grafu połączeń przypominającej sieci bezskalowe (popularne wprowadzenie w tę tematykę przedstawili Barabási i Bonabeau (2003).

Bardziej formalnie, sieć WWW składa się z szeregu rozłącznych grafów, z czego jedna składowa jest gigantyczna. Ponieważ linki (hiperłącza, odsyłacze, odnośniki) mają charakter jednokierunkowy, spójny komponent rozbija się na trzy części (por. schemat 5.1): IN, OUT i SCC. SCC to silnie połączony komponent: w jego ramach do dowolnej strony można przejść z dowolnej innej, podążając za ukierunkowanymi linkami. IN to ta część grafu, z której pająk może dojść do każdego węzła w SCC i OUT. Do każdego węzła z OUT można dojść z dowolnego węzła IN i SCC. Pozostała część sieci WWW składa się z rurek, wąsów i wysp. Rurki łączą komponent IN z komponentem OUT, z pominięciem SCC. Wąsy to te strony, do których można dotrzeć z komponentu IN lub prowadzą do komponentu OUT, ale nie znajdują się ani w rurkach ani w samym SCC. Wyspy to połączone komponenty wierzchołków, które nie są połączone żadnymi łączami z innymi komponentami. Bardziej szczegó-

łowe uwagi na ten temat można znaleźć w pracach Manning i in. (2008) i Broder i in. (2000).

Schemat 5.1. Główne komponenty sieci WWW



Źródło: Broder i in. (2000).

Taka struktura powoduje, że w wyszukiwarkach internetowych pająki zaczynają wędrówkę po sieci WWW z wielu różnych punktów startowych (Yu i in., 2020). Dlatego twórcy pająków muszą przedsięwziąć środki zabezpieczające przed zapętleniem się pająków przechodzących przez sieć.

Pająk odwiedza stronę i ekstrahuje z niej linki, które są podstawą do odwiedzania następnych stron. Strony bogate w linki z danej dziedziny, centra lub koncentratory (ang. *hubs*) są dobrym punktem startowym do przeszukiwania sieci. Pająk powinien się nie zapętlić, ale ponownie odwiedzać strony celem aktualizacji już ściągniętej strony.

Wyróżnia się następujące główne obszary zastosowań pająków (Kausar i in., 2013):

- gromadzenie stron w ogóle;
- utrzymywanie zbioru stron w stanie aktualnym;
- tematyczne gromadzenie stron;
- losowe spacery i próbkowanie;
- przeszukiwanie ukrytego/głębokiego webu.

Z punktu widzenia przedmiotowego zastosowania istotne wydają się punkty drugi, trzeci i piąty, jednak dla kompletności omówimy je wszystkie po kolei.

Gromadzenie stron w ogóle to zastosowanie służące budowie dużych wyszukiwarek (np. Google, NEKST) lub archiwów internetowych (np. Internet Archive). Punktem startowym dla pająków jest mały zbiór stron początkowych. Sieć WWW jest przeszukiwana wszerz. Stosuje się liczne modyfikacje tej strategii, np. jeden pająk przeszukuje tylko jedną witrynę (Kumar i in., 2016; Yu i in., 2020). Bardziej formalne podejście do tego zagadnienia przedstawiono w pracy Wolf i in. (2002).

Zgromadzone w ten sposób zasoby mogą być przydatne dla celów zbierania cen, ponieważ na ich podstawie można wytypować listę stron internetowych reprezentujących potencjalnie sklepy, w tym np. strony z cenami.

W serwisach użytkowanych na bieżąco (jak np. wyszukiwarki, archiwa, wyszukiwarki specjalizowane i inne zastosowania, w tym zbieranie cen) zgromadzone strony muszą być odwiedzane wielokrotnie i aktualizowane. Podejście prymitywne polegałoby na tym, by ponownie zbierać strony od początku. Jednak takie rozwiązanie jest zbyt kosztowne, dlatego stosuje się heurystyki ponownego odwiedzania ważniejszych stron, witryn lub domen, np. oparte o techniki określania i optymalizacji świeżości kolekcji na bazie obserwacji historii aktualizacji strony lub witryny. Jest to niezwykle ważne w sytuacji, gdy mamy wąskie gardło w postaci ograniczonej przepustowości. W praktyce stosuje się kilka kolejek stron do odwiedzania (co najmniej dwie), które znacznie różnią się częstością modyfikacji, np. strony z wiadomościami (gazety internetowe itp.) winny być odwiedzane co najmniej raz na dzień, a inne strony zmieniają się zwykle dość rzadko. W kontekście akwizycji cen konieczna byłaby inna strategia odwiedzania, bazująca na częstości raportowania cen, a przy tym gwarantująca, że fluktuacje cen w międzyczasie będą adekwatnie reprezentowane.

Tematyczne gromadzenie stron (ang. *focused crawling*) to skupienie pracy pająka na wybranym temacie (np. ceny detaliczne) lub języku (np. polskim), określonych typach plików (np. pliki graficzne, dźwiękowe), klasie dokumentów (np. artykuły naukowe, dokumentacje produktów, zarządy/pracownicy firm, witryny sklepów internetowych) itp. Do wytypowania stron do odwiedzenia stosuje się m.in. heurystyki oparte o analizę linków i techniki uczenia maszynowego. Cel stawiany przed takim pająkiem to znalezienie wielu interesujących stron bez obciążania szerokości pasma internetowego. Dość specyficzny temat cen każe podejrzewać, iż raczej wykaz cen jednej firmy nie będzie wskazywał na inne strony z cenami. Natomiast być może będzie zawierać linki do opisów produktów, które ułatwią ich klasyfikację. Ponadto można oczekiwać, że istnieją strony zbiorcze, z których linki prowadzą do różnych stron z cenami, np. ceneo.pl ma opcję „porównaj ceny".

Losowe spacery i próbkowanie (ang. *random walking and sampling*) polega na błądzeniu po (wcześniej uzyskanym innymi technikami) grafie WWW lub jego modyfikacji w celu badań naukowych nad tym grafem np.:

- próbkowania stron;
- oceny rozmiarów WWW;
- oceny jakości wyszukiwarek WWW.

Nie jest to istotne zastosowanie z punktu widzenia projektu InstatCeny.

Zagadnienie eksploracji ukrytej sieci, czyli *crawling the hidden Web* lub *deep Web*, jest konsekwencją tego, że wiele danych na stronach WWW znajduje się w bazach danych. Dostęp do nich jest możliwy jedynie poprzez kwerendy lub wypełnianie

formularzy (ew. wymaga znajomości hasła dostępu). Często jest to połączone z kwestią analizy innych struktur strony WWW, która nie jest zakodowana w HTML, lecz z użyciem np. JavaScript, opartego na nim AJAX, czy innych technik skryptowania (w starych systemach – apletów czy też Flash, technik, z których się obecnie wycofuje ze względu na istniejące luki bezpieczeństwa). Zwykle wyszukiwanie danych wymaga dedykowanych pająków, choć prowadzi się prace nad metodami uogólnień. W kontekście analizy cen możemy się spodziewać, że będą istnieć katalogi produktów, zwłaszcza dla dużych sklepów, gdzie do cen można będzie dotrzeć jedynie z wykorzystaniem zapytań do odpowiedniej bazy danych. Może to wymagać konieczności tworzenia pająków przeznaczonych do takiego zadania dla konkretnej strony (lub stron wykorzystujących podobne technologie prezentacji sklepu).

5.1.2.1. Wymagania względem pająków

Pająkom stawia się wiele wymagań technicznych:

- elastyczność (ang. *flexibility*) – możliwość użycia w różnych scenariuszach;
- niski koszt i wysoka wydajność (ang. *low cost and high performance*) – skalowalność do co najmniej kilkuset stron na sekundę i setki milionów stron w jednym przebiegu;
- odporność na zakłócenia (ang. *robustness*):
  - tolerancja wobec złych HTML, dziwnych zachowań serwerów i konfiguracji;
  - tolerancja załamań sieci i przerwań jej działania bez straty danych;
- „etykieta" w internecie i sterowanie szybkością pobierania stron (ang. etiquette andsSpeed control):
  - przestrzeganie standardowych konwencji, robot exclusion (specyfikacja czynności dopuszczalnych dla danej strony zawarta jest w pliku robots.txt, a sposób indeksowania konkretnej strony opisuje się w robots metatags);
  - nieprzeciążanie serwerów, np. kontaktowanie nie częściej niż co 30 sekund;
- zarządzalność i rekonfigurowalność (ang. manageability and reconfigurability):
  - interfejs nadzorujący działanie, m.in. prędkość, statystyki hostów i stron, rozmiary danych;
  - administracyjne regulowanie prędkości, dodawanie, usuwanie komponentów pająka, zamykanie systemu;
  - wymuszanie punktu kontrolnego, dodawanie hostów i domen na „czarną listę";
  - restart po załamaniu i/lub modyfikacji oprogramowania.

Technicznie dla projektu InstatCeny dodatkowo należy podkreślić znaczenie odpowiednich technik aktualizacji zebranych informacji. Z merytorycznego punktu widzenia w projekcie InstatCeny znaczenie mają przede wszystkim:

- tematyczne gromadzenie stron;
- ekstrakcja informacji z zebranych stron.

5.1.2.2. Inteligentne pająki

Tematyczne gromadzenie stron wymaga w praktyce stosowania inteligentnego pająka. Jest to ważny i dynamicznie rozwijający się kierunek badań, któremu poświęca się wiele uwagi. Powstające na ten temat prace to głównie referaty prezentowane na konferencjach i artykuły w różnych czasopismach (np. Aggarwal, 2005; Agre i Dongre, 2015; Pande i Singh, 2015).

Główne zadanie dla inteligentnego pająka to wstępne przewidywanie przed ściąganiem dokumentu, czy może on być interesujący, na podstawie:

- krótkich informacji o danym dokumencie – jeśli pająk ma dostęp do dokumentu przez bazę danych (np. Google), to pająk również posiada kilka zdań na temat tego dokumentu;
- opisu linku do danego dokumentu w innym dokumencie w bazie pająka – linki są stowarzyszone z krótkimi opisami, dzięki czemu czasem można dużo się dowiedzieć o stronie na podstawie tych opisów;
- informacji wyciąganej z adresu URL (nazwa serwera, nazwa pliku itp.).

W swej pracy inteligentny pająk stosuje zwykle następujące strategie:

- focused crawling – hipoteza o ograniczonym obszarze skupiającym informacje na dany temat;
- model Hubs → Authorities (Manning i in., 2008; por. pkt 21.3);
- hipoteza linkage locality – strony na dany temat częściej wskazują na strony na ten sam temat;
- hipoteza sibling locality – jeśli strona wskazuje na strony na dany temat, to prawdopodobnie wskazuje na inne strony na dany temat;
- start z reprezentatywnej strony i poruszanie się po stronach klasyfikowanych do interesujących kategorii przez uprzednio wyuczony klasyfikator hipertekstowy.

Celem inteligentnego pająka jest odwiedzanie w pierwszej kolejności tych stron kandydujących, które najprawdopodobniej będą spełnić predykat ciekawości. Aby ten cel zrealizować, potrzebne jest zdefiniowanie predykatu stopnia ciekawości strony. Inteligentny pająk w trakcie pracy uczy się rozpoznawać interesujące strony na bazie już zebranych informacji poprzez statystyczną ocenę dotychczas odwiedzanych stron. Ocena bazuje na następujących źródłach informacji:

- zawartość stron wskazujących na kandydata;
- zawartość opisów linków;
- struktura linku (nazwy komputera, ścieżki, pliku);
- stopień satysfakcji predykatu przez strony wskazujące;
- stopień satysfakcji predykatu przez strony „siostrzane” już odwiedzone.

Na podstawie tych danych pająk zmienia priorytety odwiedzanych stron.

5.1.2.3. Ekstrakcja informacji

Elementem inteligentnego pająka a zarazem także końcowym wynikiem jego pracy jest ekstrakcja informacji ze stron WWW. Ekstrakcja informacji wykorzystuje analizę języka, specjalnie przygotowane wyrażenia regularne czy też reguły oraz strukturę strony WWW (np. tabele).

Rozważmy przykładową regułę stosowaną w systemie ekstrakcji informacji WHISK (Soderland, 1999)[4]:
Wzorzec: * (Digit) 'BR' * '$' (Number).
Wynik: Rental Bedrooms $1, Price $2,
gdzie elementy wzorca oznaczają:

- * – pomiń znaki aż do zauważenia wzorca, np. (Digit), czyli cyfra;
- pojedyncze cudzysłowy zawierają dokładny tekst, który ma nastąpić;
- Digit – pojedyncza cyfra;
- Number – liczba wielocyfrowa.

Zastosujmy tę regułę do następującej strony HTML
Capitol Hill - 1 br twnhme. Fplc D/W W/D. Undrgrnd Pkg incl $675. 3 BR, upper flr of turn of ctry HOME. incl gar, grt N. Hill loc $995. (206) 999-9999.

Wynik jej zastosowana wyglądać będzie następująco:
Rental:
Neighborhood: Capitol Hill
Bedrooms: 1
Price: 675
Rental:
Bedrooms: 3
Price: 995.

Należy zauważyć, że istnieje wiele gotowych systemów do ekstrakcji informacji z sieci WWW, w tym takich, które m.in. wyszukują informacje o cenach. Warto tu wymienić system Lixto (Baumgartner i in., 2005, 2007) czy WebSpy (Fong i in., 2002).

W projekcie InstatCeny zastosowano nieco odmienne podejście (por. ppkt 5.1.4.3), mianowicie kreowanie oddzielnego programu przeznaczonego do każdej pojedynczej sieci sklepów.

5.1.2.4. Podejścia do masowego ściągania danych

Choć w pilotażowym projekcie zagadnienie masowego ściągania danych nie występuje, to warto nadmienić, że w literaturze wyróżnia się następujące typy metod zrównoleglania pracy pająka:

[4] Przegląd metod ekstrakcji informacji w internecie można znaleźć m.in. w pracach: Chang i in. (2006), Khder (2021) i Sarawagi (2007).

- podział przestrzeni według URL – zaletą tego podejścia jest równomierne obciążenie pająków, wadą – wysoki poziom wymiany rozproszonych tablic haszujących;
- podział według nazwy witryny (hosta) – oznacza stosowanie na nazwę hosta jednego crawlera wyposażonego w punkt kontrolny przeciążania hosta. Pająki mogą działać w wysokim stopniu niezależnie. Niestety zachodzi tu ryzyko nierównomiernego obciążenia pająków, gdy mamy pojedynczy punkt awarii, a pechowy crawler obniża liczbę ściągniętych stron na witrynę. Stosuje się tu czasem wariant polegający na stosowaniu N pająków na witrynę;
- przekierowania – jest to modyfikacja powyższych strategii polegająca na tym, że pająk przekazuje część swoich zadań innemu, przez co osiąga się równoważenie obciążeń.

5.1.2.5. Alternatywne sieciowe systemy zbierania informacji o cenach detalicznych

Warto zwrócić uwagę, że istnieją systemy zbierające ceny detaliczne na podstawie technologii telefonów mobilnych (i współpracy społecznej). System MobiShop (Dong i in., 2008; Sehgal i in., 2008) to rozproszony system pozwalający użytkownikom na stosowanie swoich telefonów mobilnych do zbierania i przetwarzania oraz przekazywania potencjalnym klientom informacji o produktach i cenach w lokalnych sklepach – wykorzystuje się tu technologie OCR do zbierania danych. System PetrolWatch (Dong i in., 2011) działa na podobnych zasadach, ale specjalizuje się w cenach paliw. Z dokładnością wynoszącą 70% identyfikuje miejsca, gdzie znajdują się tablice z cenami paliw, a z dokładnością równą 80% odczytuje ceny. System LiveCompare (Deng i Cox, 2009) stworzono dla sklepów spożywczych. Użytkownik skanuje informację o cenie, a system identyfikuje produkt na podstawie kodu kreskowego UPS, traktowanego jako globalny unikalny identyfikator produktu. Z telefonu uzyskuje się też GPS i informację GSM w celu identyfikacji i lokalizacji sklepu, z którego pochodzi cena.

5.1.3. Implementacja scrapera

Moduł IC.Scraper, czyli pająk (scraper) stworzony w ramach projektu InstatCeny, powstał jako alternatywa dla ankieterów fizycznie odwiedzających sklepy celem notowania informacji o cenach produktów znajdujących się w koszyku inflacyjnym GUS.

Najprościej ujmując, zadaniem IC.Scrapera jest odwiedzanie stron internetowych sieci handlowych, a następnie parsowanie pobranego kodu źródłowego strony celem ekstrakcji istotnych cech produktu.

W tej sekcji opisano logikę IC.Scrapera z koncepcyjnego punktu widzenia. Oprogramowanie pracuje jako zespół wzajemnie współpracujących mikroserwisów

i w taki sposób pozwala na tworzenie ciągów przetwarzania danych (komunikacja przez zapytania w protokole HTTP, czyli możliwość pracy rozproszonej w sieci WWW). Techniczne szczegóły implementacyjne, sposób konfiguracji, modele struktur danych itd. są szerzej opisane w dokumentacji technicznej (Czerski i in., 2022).

5.1.3.1. Logika IC.Scrapera

Celem IC.Scrapera jest zebranie informacji o poszczególnych produktach dostępnych online. Konieczne w takiej sytuacji jest skupienie się wyłącznie na witrynach internetowych sieci handlowych, które umożliwiają zakupy online, a tym samym udostępniają w pełni istotne, z punktu badania inflacji, informacje o produktach, w szczególności cenę produktu.

Założono, że produkty dostępne w ofercie sieci handlowych są pogrupowane w odpowiednie kategorie (np. nabiał, kawy, owoce itp.). Parsowanie kodów źródłowych stron internetowych kategorii produktów pozwala na wyłuskanie adresów URL pojedynczych produktów. Następnie parsowanie kodów źródłowych stron pojedynczych produktów pozwala wyłuskać informacje dotyczące konkretnego produktu.

Uruchomienie aplikacji automatycznie inicjuje główną logikę IC.Scrapera, tj. pobieranie danych o produktach. Z punktu widzenia metodyki działania, IC.Scraper po uruchomieniu pobiera z odpowiedniej tabeli w bazie danych konfigurację sieci handlowych do przetworzenia. Każda z uprzednio zdefiniowanych sieci handlowych, a dokładniej konfiguracja tych sieci przechowywana w bazie danych, zawiera szereg adresów URL definiujących kategorie produktów. Zdecydowano się na takie rozwiązanie, ponieważ wielobranżowe sieci handlowe posiadają w swojej ofercie kilkadziesiąt tysięcy produktów. W celu odwiedzenia wszystkich produktów dostępnych w ofercie należałoby więc wykonać kilkadziesiąt tysięcy zapytań HTTP, aby dotrzeć do stron internetowych każdego z produktów. Tak częsta i duża liczba zapytań z jednego adresu IP mogłaby skutkować zablokowaniem tego adresu IP przez serwer sieci handlowej, a w konsekwencji sprawić, że działanie IC.Scrapera stałoby się niemożliwe.

Czas trwania procesu pobierania danych o produktach dostępnych w sprzedaży sieci zależy wprost od liczby produktów (liczby zapytań HTTP do serwerów sieci) oraz okresu zwłoki pomiędzy kolejnymi zapytaniami. Logika IC.Scrapera zostaje uruchamiana co 24 godziny od momentu rozpoczęcia procesu (ten czas jest tożsamy z czasem startu całej aplikacji). Jednym z parametrów aplikacji jest flaga definiująca możliwość wyłączenia lub włączenia procesu pobierania. Flaga umożliwia również przerwanie głównego procesu IC.Scrapera pobierającego dane o produktach.

Chociaż metodologia badania inflacji wymaga porównania ceny produktu raz w miesiącu, zdecydowano się na uruchamianie procesu pobierania raz na dobę. Wy-

nika to z potencjalnych problemów, jakie może napotkać IC.Scraper, tj. zmiany szaty graficznej (kodu źródłowego) odwiedzanych stron internetowych, zbanowania (zablokowania) IP, z którego IC.Scraper wysyła zapytania HTTP, błędów po stronie sieci handlowej czy po prostu braku internetu. Częste uruchomienia IC.Scrapera pozwalają wcześnie zasygnalizować potencjalne problemy i dają czas administratorom procesu na odpowiednie rozwiązanie problemu.

Wspomniany okres zwłoki jest sposobem zabezpieczenia IC.Scrapera przed zablokowaniem jego IP przez serwer sieci; stosuje się niejako symulację zachowania zwykłego klienta przeglądającego strony sieci handlowej. Okres zwłoki jest definiowany indywidualnie dla każdej z sieci handlowej, ponadto jest on modyfikowalny w trakcie działania aplikacji.

Co więcej, IC.Scraper dla każdej z sieci handlowych przechowuje i wysyła wraz z zapytaniami HTTP, uprzednio zwrócone przez serwer sieci handlowej, *cookies* (z ang. ciasteczka). Jednym z zadań *cookies* jest optymalizacja procesu korzystania ze stron WWW. Wraz z zapytaniami wysyłane są również standardowe nagłówki (headery) HTTP mające sprawić, że z punktu widzenia serwera sieci handlowej zapytania wysyłane przez IC.Scraper będą tożsame z zapytaniami wysyłanymi z poziomu przeglądarki internetowej. Jeden z kluczowych headerów dotyczy obsługi kompresji zapytań. Odpowiednia logika kompresująca zapytania wysyłane oraz dekompresująca odpowiedzi serwerów sieci handlowej została zaimplementowana w IC.Scraperze.

Kolejną z optymalizacji procesu pobierania danych o produktach jest kilkukrotna próba wysyłania zapytań HTTP w przypadku wystąpienia błędu. Błędy mogą być spowodowane np. długim czasem odpowiedzi przez serwer sieci handlowej, przesłaniem niepełnej odpowiedzi, bądź też zgubieniem ramki danych. Jakikolwiek błąd uniemożliwia pobranie danych o produkcie. Jeśli pomimo kilkukrotnego ponawiania zapytań za każdym razem kończy się ono brakiem sukcesu, IC.Scraper pomija ten URL i płynnie przechodzi do dalszych stron WWW. Następna próba pobrania strony WWW, w przypadku której uprzednio kończyło się to błędem, odbywa się przy kolejnym uruchomieniu procesu pobierania, tj. następnego dnia.

Z założenia IC.Scraper obsługuje wiele sieci handlowych. Aby skrócić czas działania całego procesu oraz wydłużyć okresy zwłoki pomiędzy zapytaniami dla konkretnej sieci pobieranie danych o produktach dla każdej z sieci odbywa się równolegle w osobnych wątkach.

Wraz z opisaną powyżej specyfikacją produktu jest zapisywany również pobrany kod źródłowy strony internetowej dotyczącej danego produktu. Takie działanie umożliwia w sytuacji błędnego parsowania kodu źródłowego do cech produktu przetestowanie logiki parsowania, a następnie jej ewentualną poprawę. Kod źródłowy zapisywany w bazie danych jest uprzednio kompresowany celem zmniejszenia używanych zasobów systemowych.

5.1.3.2. Potencjalne problemy i administrowanie aplikacją

Działanie IC.Scrapera może natrafić na wiele problemów. Część z nich dotyczy generalnie pająków internetowych. Najbardziej prawdopodobnymi problemami są:

- zmiana kodu źródłowego stron internetowych sieci handlowych;
- zbanowanie IP IC.Scrapera przez serwer sieci handlowej;
- błędy komunikacji HTTP – po stronie klienta 4xx oraz po stronie serwera 5xx;
- brak dostępu do internetu.

IC.Scraper odwiedza ogólnie dostępne zasoby sieci internetowej. Stąd jest możliwa sytuacja, w której dana sieć handlowa zmieni kod źródłowy strony internetowej i wpłynie tym samym na parsowanie kodu źródłowego. Co ważne, sama zmiana szaty graficznej niekoniecznie niesie ze sobą błędne parsowanie strony WWW przez IC.Scraper. Szata graficzna może zostać zmieniona za pomocą wczytywanych plików CSS odpowiadających za warstwę prezentacji i sposób renderowania po stronie przeglądarki internetowej bez wpływu na samą strukturę kodu źródłowego. Błędne parsowanie stron WWW może zostać poprawione przez administratorów IC.Scrapera dzięki zapisanemu wraz z produktem w bazie danych kodowi źródłowemu.

Omawiając zaimplementowane rozszerzenia IC.Scrapera, nie sposób nie wspomnieć o mechanizmie proxy. Umożliwia on uruchomienie IC.Scrapera na maszynie pozbawionej bezpośredniego połączenia z siecią internetową i wysyłanie ich właśnie poprzez maszynę stanowiącą proxy. Dynamiczne ustawianie adresu IP przez proxy umożliwia ukrycie i zanonimizowanie działanie IC.Scrapera.

Do zadań administratora należy głównie monitorowanie działania aplikacji. Wraz ze startem aplikacji są generowane na bieżąco logi opisujące bieżące działanie IC.Scrapera, między innymi odwiedzane adresy URL i ekstrahowane informacje o produktach.

Logi aplikacji są zapisywane do plików w środowisku wykonawczym (kontenerze) oraz na standardowym wyjściu aplikacji. Istnieją trzy podsystemy dopisywania informacji do logów (appendery logowania), każdy z innym poziomem logów (zapisujący wyszczególniony poziom oraz wszystkie wyżej), mianowicie:

- ERROR+WARN – File appender zapisujący błędy i wszystkie potencjalnie sytuacje mogące powodować błędy podczas działania aplikacji;
- INFO – Console/Stdout appender informujący na bieżąco o działaniu aplikacji;
- DEBUG – File appender zapisujący wszystkie informacje podczas działania aplikacji.

Najbardziej istotnymi plikami z logami są oczywiście pliki oznaczone jako „error", ponieważ to one zawierają krytyczne informacje na temat działania aplikacji. Dodatkowo domyślny strumień wyjściowy aplikacji zawierający logowanie może być przekierowany do dowolnej lokalizacji za pomocą komend systemu Linux.

5.1.3.3. Metodologia dodawania nowej sieci handlowej

Sieć handlowa jest opisana w ramach niniejszego systemu przez dwa komponenty: (1) uzupełnienie zawartości bazy danych IC.Scrapera oraz (2) uzupełnienie kodu źródłowego IC.Scrapera o metody obsługi nowej sieci handlowej.

Po wybraniu odpowiedniej sieci handlowej jako źródło danych należy w pierwszej kolejności dodać tę sieć jako jedno ze źródeł IC.Scrapera. Służy do tego odpowiednie zapytanie opisane w dokumentacji technicznej. Następnie należy określić kategorie pobieranych produktów podając adresy internetowe (URL) stron WWW sieci handlowej, z tymi kategoriami produktów. Adresy URL muszą zostać podane wprost z racji mnogości dostępnych produktów oraz skorelowanego z tym czasem działania IC.Scrapera. Adresy można określić bezpośrednio z paska przeglądarki internetowej (schemat 5.2) bądź za pomocą analizy kodu źródłowego (schemat 5.3).

Adresy te należy dodać do konfiguracji sieci handlowej poprzez API IC.Scrapera.

Schemat 5.2. Adres URL kategorii z poziomu paska przeglądarki internetowej

Źródło: opracowanie własne.

Schemat 5.3. Adres URL kategorii z poziomu paska przeglądarki internetowej

```
▼<div class="ui-slider_inner">
  ▼<div id="slider-rum-id-89" class="ui-slider_inner_content" style="transform: translateX(-1230px);"> event flex overflow
    <a href="/c,467/cat,spozywcze-sypkie/stn,searchResults">Sypkie</a> event
    <a href="/c,10998/cat,spozywcze-cukier-i-slodziki/stn,searchResults">Cukier i słodziki</a> event
    <a href="/c,465/cat,spozywcze-sniadaniowe/stn,searchResults">Śniadaniowe</a> event
    <a href="/c,468/cat,spozywcze-dzemy-miody-i-kremy/stn,searchResults">Dżemy, miody i kremy</a> event
    <a href="/c,9/cat,spozywcze-do-wypiekow/stn,searchResults">Do wypieków</a> event
    <a href="/c,2761/cat,spozywcze-przetwory-owocowe-i-warzywne/stn,searchResults">Przetwory owocowe i warzywne</a> event
    <a href="/c,3650/cat,spozywcze-przetwory-rybne/stn,searchResults">Przetwory rybne</a> event
    <a href="/c,3646/cat,spozywcze-przetwory-miesne/stn,searchResults">Przetwory mięsne</a> event
    <a href="/c,83/cat,spozywcze-dania-gotowe/stn,searchResults">Dania gotowe</a> event
    <a href="/c,4168/cat,spozywcze-pasty-kanapkowe-i-salatki/stn,searchResults">Pasty kanapkowe i sałatki</a> event
    <a href="/c,3670/cat,spozywcze-sosy-i-dressingi/stn,searchResults">Sosy i dressingi</a> event
    <a href="/c,94/cat,spozywcze-przyprawy/stn,searchResults">Przyprawy</a> event
    <a href="/c,24/cat,spozywcze-oleje-oliwy-octy/stn,searchResults">Oleje, oliwy, octy</a> event
  </div>
</div>
```

Źródło: opracowanie własne.

W dalszej kolejności należy utworzyć klasę w kodzie źródłowym IC.Scrapera[5], której nazwa musi być tożsama z nazwą sieci handlowej wpisaną do bazy danych. Klasę tę definiuje się jako rozszerzenie klasy abstrakcyjnej „Store", przy czym wymagane jest zaimplementowanie poniższych metod:

- parseProductsLinks;
- parseProduct.

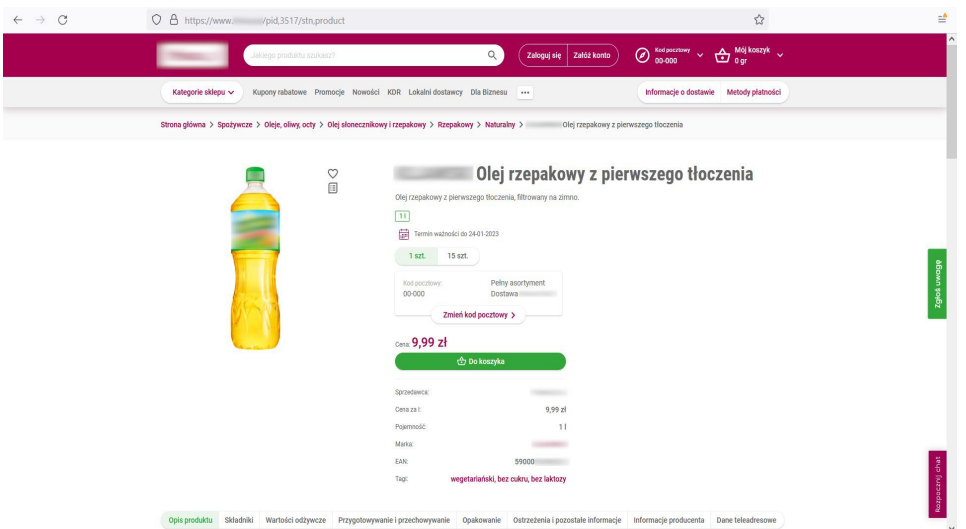
[5] IC.Scraper jest napisany w języku programowania Java.

Nazwę stworzonej klasy należy dodać do bazy danych IC.Scrapera. Kluczem głównym rekordu w bazie danych przechowującym konfigurację sieci handlowej jest wprost nazwa zaimplementowanej klasy w kodzie źródłowym scrapera.

Zilustrujmy kroki potrzebne do zaimplementowania tych klas na przykładzie wybranej sieci handlowej, której produkt pokazuje schemat 5.4.

Pierwszą z metod do zaimplementowania jest parseProductsLinks – w języku programowania Java – czyli parsowanie kodu źródłowego strony WWW opisującego kategorię celem ekstrakcji adresów URL do poszczególnych produktów tej kategorii (por. schemat 5.5). Zwykle wystarczy zidentyfikować selektor CSS dla odpowiedniego znacznika (tagu) HTML zawierającego żądaną informację (por. schemat 5.6). Aby uzyskać znaczniki, warto wspierać się rozbudowanymi funkcjonalnościami przeglądarek internetowych określanymi jako narzędzia dla deweloperów i pozwalającymi na efektywną analizę kodu źródłowego.

Schemat 5.4. Przykładowy produkt oferowany przez sieć handlową



Źródło: strona sklepu internetowego.

Schemat 5.5. Kod źródłowy strony wskazujący adres URL do modelowego produktu

```
Inspector  Console  Debugger  Network  Style Editor  Performance  Memory  Storage  Accessibility  Application
Search HTML
▸ <div class="product-box_holder">⋯</div>
▸ <div class="product-box_holder">⋯</div>
▸ <div class="product-box_holder">⋯</div>
▸ <div class="product-box_holder">⋯</div>
▸ <a class="catalog-products_banner-wrapper tracked-side-clip-1" href="/stn,akcja-seans_filmowy/n,seans_filmowy?side">⋯</a> event
▾ <div class="product-box_holder">
  ▾ <div class="product-box with-switcher"> event
    ▸ <div class="product-box_switcher">⋯</div>
      <div style="width: 0px; height: 0px;"></div>
    ▾ <div class="product-box_content"> flex
        ::before
    ▸ <div class="product-box_labels">⋯</div> flex
    ▸ <div class="product-box_actions">⋯</div> flex
    ▸ <a class="product-box_image " href="/pid,3517/n,[redacted]-olej-rzepakowy-z-pierwszego-tloczenia/stn,product" title="[redacted] Olej rzepakowy
      z pierwszego tłoczenia">⋯</a> event
    ▸ <div class="product-box_desc">⋯</div> flex
        ::after
      </div>
    ▸ <div class="product-box_additional-info">⋯</div>
```

Źródło: strona sklepu internetowego, widok narzędzia przeglądarki.

Schemat 5.6. Przykładowy selektor CSS tagu HTML wskazujący adres URL produktu

```
iv.catalog-page › div.catalog-products.list-view_browser-r... › div.list-view_content › div.product-box_holder › div.product-box.with-switcher › div.product-box_content › a.product-box_image
```

Źródło: opracowanie własne.

Druga metoda, parseProduct, wymagająca implementacji w Javie, określa sposób ekstrakcji danych o pojedynczym produkcie (por. schemat 5.7). Należy uwzględnić wszystkie jawnie dostępne, a tym samym możliwe do skrapowania dane, tj. nazwę, markę, standardową oraz promocyjną cenę, opis produktu, jednostkę sprzedaży oraz wewnętrzny i zewnętrzny kod produktu (GTIN/EAN). Podobnie jak dla kategorii produktów i tutaj zwykle wystarczy wskazać selektory CSS odpowiednich tagów HTML (por. schemat 5.8). Ramka danych produktu, którą ta metoda musi zwracać, została szczegółowo wyjaśniona w dokumentacji technicznej (Czerski i in., 2022).

Schemat 5.7. Kod źródłowy strony modelowego produktu

```
Inspector  Console  Debugger  Network  Style Editor  Performance  Memory  Storage  Accessibility  Application
Search HTML
▾ <div class="product-page_content full-width-container"> overflow
  ▾ <div class="container">
    ▾ <div class="product-page_product" itemscope="" itemtype="http://data-vocabulary.org/Product">
      ▸ <div class="product-page_image">⋯</div>
      ▾ <div class="product-page_info">
        ▾ <div class="product-page_name-box">
            <meta itemprop="brand" content="[redacted]">
            <h1 itemprop="name">[redacted] Olej rzepakowy z pierwszego tłoczenia</h1>
          ▸ <h3 class="subname">⋯</h3>
          </div>
        ▸ <div class="product-page_inf">⋯</div>
        </div>
      ▸ <div class="product-page_details">⋯</div>
      </div>
    ▸ <div class="trustmate-comments">⋯</div>
    </div>
  </div>
▸ <div style="">⋯</div>
</div>
```

Źródło: strona sklepu internetowego, widok narzędzia przeglądarki.

Schemat 5.8. Przykładowy selektor CSS tagu HTML wskazujący na nazwę produktu

ne › div › div.product-page › div.product-page_content.full-width-cont... › div.container › div.product-page_product › div.product-page_info › div.product-page_name-box

Źródło: opracowanie własne.

Powyższe działania należy zakończyć ponownym uruchomieniem aplikacji. Czynność ta jest konieczna, ponieważ jednym z kroków jest rozszerzenie kodu źródłowego scrapera, który musi zostać ponownie poddany kompilacji. Zatem należy w pierwszej kolejności zatrzymać aplikację, następnie skompilować kod źródłowy i wygenerować nowy obraz kontenera przechowującego aplikację, aby ostatecznie ponownie ją uruchomić. Poszczególne kroki tej procedury są zautomatyzowane poprzez odpowiednie skrypty, które zawierają również szczegóły implementacyjne, opis i sposób użycia.

5.1.4. Funkcjonalność IC.Scrapera

Moduł IC.Scrapera obejmuje następujące funkcjonalności:

- przerwanie i ponowne uruchomienie procesu skrapowania;
- informacja o stanie procesu skrapowania;
- zbiorcze informacje o pobranych produktach:
  - zbiorcze zestawienie wszystkich produktów w bazie danych;
  - zbiorcze zestawienie produktów w bazie danych spełniające kryterium konkretnej nazwy sieci handlowej;
- rozszerzona informacja o produktach jednostkowych:
  - jednostkowy produkt w bazie danych spełniający kryterium konkretnej sieci handlowej oraz numeru identyfikacyjnego wewnątrz tej sieci;
  - jednostkowy produkt w bazie danych spełniający kryterium numeru identyfikacyjnego;
- konfiguracja skrapowanych sieci handlowych obejmująca funkcje:
  - dodanie nowej lub usunięcie istniejącej sieci handlowej;
  - zbiorcza prezentacja konfiguracji sieci handlowych;
  - zmiana konfiguracji indywidualnej sieci handlowej (specyfikacja, jakie rodzaje produktów mają być pobierane).

5.1.5. Dostęp do API IC.Scrapera

API, czyli interfejs aplikacji (ang. *application programming interface*), to specjalny sposób wymiany informacji między aplikacjami (API musi mieć zawsze dokumentację i specyfikację). Współcześnie dominują (zwłaszcza w serwisach internetowych) dwa rodzaje API: SOAP i REST. Serwisy internetowe działają zasadniczo w trybie klient-serwer. SOAP to rodzina interfejsów „stanowych", tzn. zakłada się z góry serię

wymiany komunikatów, które muszą następować w określonej kolejności, natomiast REST to rodzina interfejsów bezstanowych, tzn. jeden komunikat podaje komplet informacji potrzebnych do wykonania usługi, a usługa (serwer) zwraca cały wynik działania jednoetapowo.

Choć funkcjonalność oprogramowania stworzonego w ramach projektu InstatCeny nie jest wystawiana do użytku publicznego, to została zrealizowana w konwencji usług (serwisów) intranetowych, w wewnętrznej sieci lub w pseudosieci na pojedynczym komputerze.

Dostęp do usług w tym kontekście uzyskuje się przez endpointy, czyli kanały wymiany informacji, opisane (rozbudowanymi) adresami URL usług.

Aplikacja IC.Scrapera komunikuje się z otoczeniem przez API z rodziny REST endpoint.

5.1.6. Format i zakres ramki danych

5.1.6.1. Założenie o strukturze witryny internetowej sieci handlowej

Dostępne w ofercie produkty sieci handlowych są pogrupowane w odpowiednie kategorie (np. nabiał, kawy, owoce), a każda kategoria jest reprezentowana przez oddzielną stronę internetową. Każdy produkt na stronie kategorii charakteryzowany jest przez link, prowadzący do strony pojedynczego produktu. Na stronie konkretnego produktu są dotyczące go informacje, takie jak opis, cena itd.

5.1.6.2. Konfiguracja sieci handlowej

W bazie danych jest opisanych szereg sieci handlowych. Opis każdej z nich obejmuje nazwę sieci, oraz wykaz adresów URL definiujących kategorie produktów będące przedmiotem zainteresowania użytkownika. Ramka danych (wejściowych i wyjściowych) zawiera elementy:

- store – nazwa sieci handlowej;
- productsCategoriesUrls – tablica z linkami do kategorii, z których są pobierane produkty;
- minimumFetchDelay – minimalna zwłoka pomiędzy żądaniami (requestami) do serwera sieci [ms];
- maximumFetchDelay – maksymalna zwłoka pomiędzy żądaniami (requestami) do serwera sieci [ms].

5.1.6.3. Opis produktu

Ramka danych (wyjściowych) obejmuje elementy:

- id – id produktu według tabeli „product” w bazie danych;
- storeId – id produktu według sieci handlowej;

- store – nazwa sieci handlowej;
- name – nazwa produktu;
- regularPrice – standardowa cena;
- salePrice – cena promocyjna (pole opcjonalne);
- description – opis produktu;
- amount – obiekt opisujący ilość produktu;
- amount.id – id ilości produktu według tabeli „amount" w bazie danych;
- amount.pack – liczba paczek produktu;
- amount.quantityPerPack – ilość w jednej paczce lub opakowaniu produktu;
- amount.totalQuantity – łączna ilość produktu;
- amount.unit – jednostka ilości;
- amount.inputString – ciąg znaków będący wejściem do parsowania i wyznaczenia ilości produktu;
- amount.asFormattedString – ilość jako sformatowany napis „<totalQuantity ><unit> [<pack>x<quantityPerPack><unit>]" (np. „500g [4x125g]");
- foreignProductCode – obiekt charakteryzujący zewnętrzny kod produktu (pole opcjo- nalne);
- foreignProductCode.id – id kodu zewnętrznego według tabeli „foreign_product _code" w bazie danych;
- foreignProductCode.code – zewnętrzny kod produktu;
- foreignProductCode.identifier – rodzaj kodu zewnętrznego („EAN" lub „GTIN");
- fetchingDate – data pobrania informacji o produkcie z sieci w formacie „yyyy-MM-dd HH:mm";
- url – link do produktu w serwisie sieci handlowej;
- html – pobrany podczas skrapowania HTML będący podstawą do parsowania informacji o produkcie;
- available – dostępność produktu.

5.1.6.4. Przykładowe wielkości asortymentu produktów zbieranych przez IC.Scraper

Dla uzmysłowienia bogactwa produktów oferowanych w interesujących GUS kategoriach podajemy statystyki wybranych kategorii produktów dla dwóch przykładowych sieci handlowych.

Sieć handlowa SiecSigma:

- produkty spożywcze – 10 251 unikalnych produktów dostępnych w sieci;
- chemia/higiena/produkty dla dzieci i niemowląt/produkty dla zwierząt itp. – 4 963 unikalnych produktów.
- Sieć handlowa SiecXi:
- produkty spożywcze – 11 645 unikalnych produktów;

- chemia/higiena/produkty dla dzieci i niemowląt/produkty dla zwierząt itp. – 5194 unikalnych produktów.

5.2. Uzyskiwanie i archiwizacja danych skanowanych

Głównym wyzwaniem w uzyskiwaniu danych skanowanych od zewnętrznych gestorów jest odpowiednie zdefiniowanie zakresu otrzymywanych danych. Zadanie to jest istotne ze względu na niechęć zewnętrznych podmiotów do częstej zmiany uzgodnionego formatu przekazywanych danych. W razie błędnie zdefiniowanego modelu danych próba jego zmiany może być bardzo utrudniona, co będzie paraliżowało procesy analizy.

Dla celów analizy gestorzy powinni udostępniać co najmniej następujące informacje na temat produktów (por. ppkt 2.4.2.3):

- kod EAN produktu;
- kod producenta (lub kod nadany przez gestora);
- nazwa produktu;
- opis produktu (zwykle dostępny jako tekst w języku naturalnym polskim);
- wolumen sprzedaży;
- cena sprzedaży;
- okres (od dnia do dnia), którego dotyczy wolumen i cena.

Dostępność innych informacji, jak granulacja ze względu na miejsce sprzedaży, charakterystyka jednostki sprzedaży itp., może być pomocna.

Informacje dotyczące wolumenu i ceny sprzedaży są krytyczne dla prowadzenia działalności biznesowej gestorów i mogą wpłynąć na konkurencyjność podmiotu udostępniającego dane. Dlatego współpraca z gestorami wymaga zadbania przez GUS o maksymalne bezpieczeństwo przekazanych danych, tzn. zapewnienie, że dane jednostkowe, szczególnie te bardzo czułe, nie dostaną się w ręce osób czy firm trzecich.

Z jednaj strony istotne jest, aby dostępu do danych nie zyskały osoby postronne. To zagadnienie jest omawiane w niniejszym podrozdziale.

Z drugiej – GUS gromadzi te dane po to, aby je przetwarzać i wykorzystywać, a także publikować wyniki przetwarzania. Udostępniane mogą być tylko dane przetworzone, w których ślad do gestora został całkowicie zatarty. Kolejne podrozdziały (5.3 i 5.4) przedstawiają formę, do jakiej dane będą przetwarzane w GUS, aby uzmysłowić, że tu również mamy do czynienia z maksymalną ochroną danych.

Warto dodać, że ochrona danych leży nie tylko w interesie gestorów, lecz także GUS. Wgląd w szczegóły procesu przetwarzaniach danych w GUS, np. jakie towary są użyte w koszykach przy wyliczaniu inflacji, mógłby być podstawą do stosowania przez zewnętrzne czynniki mechanizmów fałszowania wskaźników inflacji dla wro-

gich celów. Stąd ochrona musi dotyczyć nie tylko danych od gestorów, lecz także tych skrapowanych z internetu.

5.2.1. Bezpieczny transfer danych

Problematyka bezpieczeństwa danych to istotny element działalności wszelkich przedsiębiorstw i instytucji dysponujących wrażliwymi danymi (Castano i in., 1995). Standardowo podstawą zabezpieczania danych jest autoryzacja dostępu oraz utworzenie zapór sieciowych (Szmit i in., 2005). Pomimo że są to zabezpieczenia istotne, a nawet koniecznie, niezbędne są zapewnienie niezawodności dostępu do danych oraz dodatkowa ochrona w postaci metod kryptograficznych.

5.2.1.1. Zapewnienie bezpieczeństwa danych

Obecnie stosowaną architekturą systemów bazodanowych jest wielowarstwowość, w której poszczególne warstwy systemu odpowiadają za pojedyncze funkcje systemu – stanowią kolejne zapory przeciw potencjalnym atakom. Bezpieczeństwo wymaga stosowania następujących procedur (Busłowska, 2013; Kasprowski i in., 2016):

- autentyfikacja i autoryzacja użytkowników;
- transfer danych zabezpieczonymi kanałami systemu;
- wypracowanie procedur bezpieczeństwa i bezwzględne przestrzeganie ich przez wszystkich użytkowników systemu.

Autentyfikacja to proces weryfikacji tożsamości użytkownika – mówiąc wprost, czy użytkownik jest tym, za kogo się podaje. Natomiast autoryzacja to proces sprawdzenia, czy dany użytkownik posiada odpowiednie uprawnienia do wykonania żądanej przez niego akcji.

Najbardziej istotna jest kontrola dostępu do bazy danych i jej obiektów poprzez nadawanie użytkownikom odpowiednich uprawnień do wykonywania określonych operacji. Uprawnienia ograniczają dostęp do danych, możliwości zmiany danych lub struktur bazy danych lub możliwości wykonywania funkcji systemowych.

Transfer danych odbywa się za pośrednictwem sieci z wykorzystaniem protokołu IP. Komunikacja może przebiegać w sieciach zaufanych bądź też publicznie dostępnych. Charakter komunikacji w sieciach opartych na protokole IP wymusza założenie, że komunikacja w sieci internet może być w całości podsłuchana, zarejestrowana, przechwycona lub sfałszowana.

5.2.1.2. Bezpieczeństwo wielowarstwowych systemów bazodanowych

Bazy danych, jako systemy informatyczne narażone są na szereg zagrożeń. Do najważniejszych z nich można zaliczyć:

- odczyt danych przez nieuprawnionych użytkowników;
- modyfikacje istniejących danych;

- zniszczenie danych w skutek poważnych awarii sprzętu komputerowego.

Wypracowano środki ochrony baz danych realizowane w poszczególnych warstwach systemu bazodanowego.

5.2.1.2.1. Autentyfikacja i autoryzacja

Autentyfikacja i autoryzacja użytkowników, określana jest też jako kontrola dostępu. Kontrola dostępu jest realizowana według określonych reguł, nazywanych polityką bezpieczeństwa (Stokłosa i in., 2001). Przywileje nadaje się pojedynczym lub wszystkim użytkownikom oraz pełnionym rolom. Uprawnienia użytkownikom nadaje się w zależności od zajmowanego przez nich stanowiska i wykonywanych czynności. Zawsze jednak zaczyna się od nadania minimalnych uprawnień, które są stopniowo, według potrzeb, rozszerzane. Ten system jest określany terminem poufności – polega na braku dostępu do danych dla użytkowników oraz aplikacji nieuprawnionych do ich odczytywania. Klauzulę bezpieczeństwa w postaci: ściśle tajne, tajne, poufne, nadaje się danym newralgicznym z punktu widzenia instytucji.

5.2.1.2.2. Limity zasobów

Oprócz ograniczeń dostępu do danych (wynikających z przyznanych uprawnień) na użytkowników mogą być nakładane ograniczenia dotyczące wykorzystania zasobów systemowych, kontrolowanych przez system zarządzania bazą danych. Należą do nich np.:

- ilość czasu procesora przeznaczonego na obsługę zadań określonego użytkownika;
- liczba równoczesnych sesji otwartych przez użytkownika;
- liczba odczytów (logicznych) z dysku przez użytkownika;
- dopuszczalny czas bez wykonywania operacji na bazie danych.

Ograniczenia nałożone na użytkownika tworzą profil użytkownika.

5.2.1.2.3. Integralność danych

Wśród mechanizmów zapewniających integralność baz danych można wydzielić mechanizmy integralności semantycznej oraz mechanizmy integralności transakcyjnej. Semantyczne więzy integralności określają poprawność stanu bazy danych pomiędzy kolejnymi operacjami na bazie, stąd umożliwiają ochronę przed celową lub przypadkową (niepoprawną) modyfikacją danych. Integralność danych oznacza pewność, że dane nie zostały podmienione, zniekształcone, zmodyfikowane lub usunięte bez wiedzy ich właściciela. Stan bazy danych pozostaje zgodny ze stanem wymaganym i oczekiwanym przez adresata, do którego są przesyłane. Naruszenie integralności następuje przy nieupoważnionym dostępie i ataku, potknięciach i zaniedbaniach podczas normalnej pracy użytkowników uprawnionych, nieposiadających odpowiedniego przygotowania, wiedzy lub przeszkolenia. Stąd zakres uprawnień poszczególnych użytkowników powinien być stopniowo zwiększamy wraz z wzro-

stem doświadczenia. Mechanizmy integralności transakcyjnej chronią spójność bazy danych w warunkach współbieżnie realizowanych operacji na bazie przez wielu użytkowników, a także w przypadku błędów oprogramowania. Mogą również być użyteczne w sytuacjach awarii sprzętu, zakłóceń transmisji, błędów w oprogramowaniu lub wirusów. W systemach informatycznych integralność danych powinna być zapewniona podczas przechowywania danych ich przetwarzania i przesyłania.

5.2.1.2.4. Monitoring operacji na bazie danych

Działania użytkowników, a w szczególności wszystkie działania istotne z punktu widzenia bezpieczeństwa systemu winny być monitorowane i rejestrowane w celu możliwości powrotu do nich i ich analizy. Analiza wyników takich logów może służyć do oceny poprawności zaimplementowanego systemu zabezpieczeń. Poszczególne operacje jako kolejne wpisy w rejestrach logów stanowią ślad dyskretnego procesu zmiany. Wprowadzane, modyfikowane i usuwane informacje muszą spełniać warunki narzucone na dane podczas definicji bazy danych tak, aby baza była zgodna z modelowaną rzeczywistością. W każdym momencie baza danych znajduje się w określonym stanie. Spójny stan to taki, w którym wszystkie wartości, które są zawarte w bazie danych w tym stanie, mogą zaistnieć również w świecie rzeczywistym. Warunki spójności mogą być dynamiczne lub statyczne. Warunki dynamiczne różnią się od statycznych tym, że pamiętają poprzedni stan. Spełnienie warunków spójności zapewnia poprawność zgromadzonych danych. Naruszenie spójności danych następuje w wyniku semantycznie niepoprawnych operacji, niewłaściwej synchronizacji działania transakcji współbieżnych lub w wyniku awarii systemu.

Z uznaniową kontrolą dostępu powiązana jest dostępna w niektórych serwerach baz danych możliwość audytu, czyli obserwacji i rejestrowania informacji o działaniach użytkowników. Taka funkcja, którą określamy jako monitorowanie, może obejmować monitoring:

- operacji – śledzenie wskazanych operacji SQL (wykonanych bądź odrzuconych);
- uprawnień – śledzenie wykorzystania uprawnień systemowych przez konkretnych użytkowników;
- obiektów – śledzenie wykonania wskazanych operacji SQL na konkretnych obiektach.

Monitoring może być wykorzystany do podniesienia bezpieczeństwa systemu, przede wszystkim poprzez kontrolę działań podjętych przez użytkowników, którzy próbują przekraczać przyznane im uprawnienia i badają w ten sposób system nałożonych uprawnień na konkretne grupy użytkowników, role, czy też określone zasoby. Poza tym monitorowanie jest wykorzystywane do strojenia serwera bazy danych.

5.2.1.2.5. Szyfrowanie zawartości bazy danych

Szyfrowanie danych w aplikacjach, w szczególności tych wrażliwych to konieczność wynikająca nie tylko z dobrych praktyk inżynierskich, lecz także z przepisów prawa (ochrona danych osobowych) lub wymagań branżowych (PCI Security Standard Council, b.r.). Pod kątem technicznym mechanizmy szyfrowania baz danych można koncepcyjnie przyrównać do technik szyfrowania dysków. Istnieje więc możliwość szyfrowania zarówno całych baz danych, jak i operowania nieco bardziej precyzyjnie na poszczególnych obiektach. Algorytmy szyfrujące nie pozostają jednak bez wpływu na wydajność pracy baz danych, jak również często na zajmowaną przez nie przestrzeń dyskową. Decyzja o zastosowaniu szyfrowania powinna być poparta wcześniejszą analizą potencjalnych negatywnych skutków jej zastosowania w określonym wariancie. Wybierając model szyfrowania, trzeba rozważyć, co dokładnie chcemy szyfrować. Może się okazać, że nie ma potrzeby szyfrowania całych baz. Istotną kwestią jest też decyzja o szyfrowaniu pola indeksowanego; wówczas można spodziewać się znacznego wzrostu wymagań na zasoby systemowe, gdzie każdorazowe odwołanie do takiego pola będzie wymagało przeprowadzenia deszyfracji.

5.2.2. Bezpieczeństwo transmisji danych

W architekturach wielowarstwowych komunikacja pomiędzy warstwami odbywa się za pośrednictwem sieci z wykorzystaniem protokołu IP (Sojak i in., 2009). W zależności od rozwiązania niektóre z tych sieci mogą być sieciami zaufanymi (m.in. wyodrębniony intranet) inne natomiast publicznie dostępne – m.in. sieć internet. Ze względu na charakter komunikacji w sieciach opartych na protokole IP można założyć, że komunikacja w sieci internet może być w całości podsłuchana, zarejestrowana, przechwycona lub sfałszowana. Dlatego należy przedsięwziąć środki mające na celu odpowiednie zabezpieczenie komunikacji. W internecie stosowane są dwa standardowe protokoły zabezpieczeń: SSL i IPSec.

5.2.2.1. SSL – warstwa gniazd bezpiecznych

Protokół SSL (ang. *Secure socket layer*, warstwa gniazd bezpiecznych), opracowany przez firmę Netscape, oraz nowszy protokół TLS (ang. *Transport layer security*, zabezpieczenia warstwy transportu) opracowany przez IETF spełniają analogiczne funkcje. SSL działa wraz z protokołami warstwy aplikacji takimi jak Telnet, M.IN., FTP, a także protokołem TCP/IP realizując cztery główne zadania:

- identyfikacja serwera – potwierdzenie tożsamości serwera. Oprogramowanie klienta, przy wykorzystaniu algorytmów kryptograficznych, może jednoznacznie stwierdzić, że serwer posiada certyfikat i klucz publiczny wydany przez zaufanego wystawcę certyfikatów (CA – Cerficate Authority);

- identyfikacja klienta (opcjonalna) – weryfikacja certyfikatu klienta przez serwer. Identyfikację klienta przeprowadza się na podobnej zasadzie, na jakiej następuje identyfikacja serwera. Metoda ta może być stosowana zamiast pary użytkownik – hasło. Ze względu na małą popularność certyfikatów wśród osób fizycznych ta metoda nie jest powszechnie stosowana;
- ochrona transmisji przed podsłuchem – komunikacja między klientem a serwerem jest szyfrowana w całości od pierwszego zapytania klienta, co uniemożliwia przechwycenie przekazywanych danych;
- ochrona transmisji przed modyfikacją – dodatkowo przesyłane zestawy są chronione przez funkcję skrótu umożliwiającą weryfikację ich poprawności i zabezpieczającą przed nieuprawnioną modyfikacją.

  Po nawiązaniu połączenia, w protokole SSL następuje negocjacja wersji protokołu oraz wybór algorytmów kryptograficznych (algorytmy szyfrowania oraz funkcja skrótu). Następnie protokół uzgadniania uwierzytelnia oba lub jeden punkt końcowy sesji i ustanawia unikalny symetryczny klucz używany do generowania kluczy służących do szyfrowania i deszyfrowania danych w sesji. Dalsza komunikacja jest szyfrowana za pomocą szyfru symetrycznego, co w małym stopniu wpływa na szybkość transmisji. W czasie transmisji możliwa jest renegocjacja parametrów oraz powtórne wygenerowanie kluczy sesyjnych. Uzgadnianie SSL, ze względu na operacje szyfrujące z użyciem kluczy publicznych i prywatnych, jest działaniem wymagającym dużej wydajności (negocjacje algorytmów, uwierzytelnianie i wymiana kluczy), natomiast sama transmisja danych dzięki przechowywaniu informacji o sesji SSL w pamięci chronionej jest już szybsza.

Podkreślić należy, że protokół SSL umożliwia zabezpieczenie wyłącznie transmisji realizowanych za pomocą protokołu TCP/IP i jego implementacja wymaga modyfikacji aplikacji (ewentualnie możliwe jest tunelowanie). Do poprawnego działania protokołu SSL wymagany jest certyfikat X.509 serwera (oraz opcjonalnie klienta). Protokół ten jest najczęściej stosowany do bezpiecznego dostępu do stron WWW, rzadziej do szyfrowania komunikacji z bazami danych.

5.2.2.2. IPSec – protokół bezpieczeństwa w internecie

Protokół IPSec (ang. *Internet protocol security*) jest protokołem opracowanym przez IETF. Działa w warstwie sieciowej, pomiędzy protokołami TCP i IP, i umożliwia bezpieczne tunelowanie pakietów w sieci internet. Protokoły tej grupy mogą być wykorzystywane do tworzenia sieci VPN (wirtualnej sieci prywatnej). Zadania realizowane przez protokół są następujące:

- uwierzytelnienie obu stron komunikacji wobec siebie;
- nawiązanie bezpiecznego kanału;
- bezpieczne uzgodnienie kluczy kryptograficznych oraz parametrów tuneli;

- renegocjacja kluczy kryptograficznych.
  W skład IPSec wchodzą trzy klasy protokołów: ESP (ang. *Encapsulating security payload*, protokół kapsułkowania), AH (ang. *Authentication header*, protokoł nagłówków uwierzytelniania), będące protokołami niskiego poziomu, oraz IKE (ang. *Internet key exchange*, protokół zarządzania kluczami). Podobnie jak w protokole SSL, w protokole IPSec również jest możliwa negocjacja funkcji skrótu i algorytmów kryptograficznych stosowanych do szyfrowania transmisji. Samo szyfrowanie jest dokonywane za pomocą szyfrów symetrycznych. Bezpieczeństwo zapewniane przez Ipsec może być dwojakie, w zależności od stosowanego protokołu. Bezpieczeństwo może opierać się na certyfikatach X.509 lub dodatkowo współdzielonych kluczach. Protokół IPSec działa na poziomie systemu operacyjnego i umożliwia stworzenie bezpiecznego kanału komunikacyjnego w dwóch trybach:
- trybie transportowym – pomiędzy nagłówek IP a TCP dokładamy jeszcze dodatkowo nagłówek Ipsec; w tym trybie szyfrowane są tylko dane z pakietu IP, ponieważ komunikacja odbywa się pomiędzy dwoma komputerami;
- trybie tunelowym – zestaw użytkownika jest w całości kapsułkowany w ESP, a na początek jest dokładany zupełnie nowy nagłówek IP. W tym trybie szyfrowane są wszystkie dane oraz nagłówki oryginalnych zestawów IP, co oznacza, że nadawca i odbiorca nie są znani. Tryb ten stosuje się do utworzenia wirtualnych sieci prywatnych pomiędzy dwoma sieciami.
  Tunelowanie IPSec jest przezroczyste dla działających aplikacji i nie wymaga ich modyfikacji. W wielowarstwowych systemach bazodanowych protokół stosuje się do komunikacji serwera aplikacji i serwera bazy danych, jeżeli nie znajdują się one w zaufanej i wydzielonej sieci.

5.2.2.3. Systemy zarządzania bazą danych

Systemy zarządzania bazą danych (ang. *Database management systems* – DBMS) są zorganizowanym zbiorem narzędzi umożliwiającym zdefiniowanie struktury bazy, jej stworzenie i później wykonywanie wszystkich operacji modyfikujących strukturę bazy oraz same dane. Stanowią warstwę pośredniczącą pomiędzy bazą danych, a użytkownikami. Na zarządzanie bazami danych składa się m.in.:

- organizacja struktury danych;
- wprowadzanie danych;
- wyszukiwanie danych według zadanych kryteriów;
- modyfikacja danych;
- zachowanie integralności (ochrona przed błędami);
- administrowanie bezpieczeństwem;
- organizowanie pracy wielodostępnej;

- łączenie i wymiana danych z innymi systemami baz danych;
- zarządzanie transakcjami.

Popularne modele baz danych i ich systemów zarządzania obejmują:

- system zarządzania relacyjną bazą danych (ang. Relational database management system – RDBMS) – przystosowany do większości przypadków użycia;
- system zarządzania nierelacyjną bazą danych SQL (NoSQL DBMS) – dobrze nadaje się do luźno zdefiniowanych struktur danych, które mogą z czasem ewoluować;
- system zarządzania bazą danych w pamięci (ang. In-memory database management system – IMDBMS) – zapewnia krótszy czas reakcji i lepszą wydajność;
- system zarządzania kolumnową (korelacyjną) bazą danych (ang. Column (correlation) database management system – CDBMS) – odpowiedni dla jednostek, które mają dużą liczbę podobnych pozycji danych;
- system zarządzania bazą danych w chmurze (ang. Cloud database management system) – dostawca usług w chmurze jest odpowiedzialny za zapewnienie i utrzymanie DBMS.

Współcześnie najczęściej wykorzystywane są systemy oparte na architekturze klient-serwer, jednakże istnieją również systemy deskoptowe. Najpopularniejszymi systemami są:

- Oracle[6] – jeden z dwóch najpopularniejszych w Polsce systemów komercyjnych ze względu na swoją niezawodność i funkcjonalność;
- Microsoft SQL Server[7] – system stworzony przez największą na świecie firmę informatyczną, jeden z dwóch najpopularniejszych w Polsce systemów komercyjnych;
- PostgreSQL[8] – darmowy system opracowany na Uniwersytecie Kalifornijskim. System jest znacznie prostszy w porównaniu z dwoma poprzednimi, posiada za to pewne rozszerzenia obiektowe;
- MySQL[9] – system jest prostszy od PostgreSQL, pewne ważne mechanizmy stosowane w bazach relacyjnych nie są zaimplementowane, jednak jest to system bardzo szybki, czasami dorównujący nawet Oracle czy Microsoft SQL Server; bardzo popularny w społeczności open-source;
- DB2 firmy IBM[10] – w wielu testach uznawany za najszybszy;

[6] http://www.oracle.com/.
[7] https://www.microsoft.com/en-us/sql-server.
[8] http://www.postgresql.com/.
[9] http://www.mysql.com/.
[10] https://www.ibm.com/analytics/db2.

- Microsoft Access – deskoptowy system wchodzący w skład pakietu biurowego MS Office, ze względu na ograniczoną wydajność nadający się do niewielkich projektów.

5.2.2.4. Istniejące rozwiązania i narzędzia dotyczące bezpieczeństwa danych

W teorii bazy danych są chronione przed złośliwą aktywnością atakujących przez firewall oraz inne systemy wykrywania włamań. Jednakże bazy danych wymagają własnej ochrony z racji połączenia ich z siecią internet.

Stare strategie ochrony danych oparte na ręcznych metodach nie pozwalają egzekwować separacji obowiązków, rozpoznawać podejrzanej aktywności w czasie rzeczywistym, ani ujednolicać wyników. Nie pomagają też w podejmowaniu działań. Wymagania formalno-prawne i potencjalne audyty jedynie zwiększają złożoność. Dlatego opracowano sposoby zapewnienia bezpieczeństwa, do których należą m.in.:

- kontrola dostępu w celu zapobiegania nieautoryzowanemu dostępowi poprzez wdrożenie kilkuetapowej weryfikacji;
- system monitorowania luk w zabezpieczeniach i opracowywanie planu ich eliminacji;
- fizyczne bezpieczeństwo bazy danych i serwerów przed kradzieżą i klęskami żywiołowymi.

Narzędzia te monitorują choćby ruch sieciowy czy zapytania do bazy danych, wykrywają potencjalne zagrożenia, wskazują na niebezpieczeństwo z nich wynikające, a często podsuwają polecane rozwiązania. Generowane przez te narzędzia audyty umożliwiają zespołowi administrującemu bazę danych podjęcie decyzji o usunięciu wykrytych luk albo kontakt z obcymi instytucjami w celu rozwiązania problemu, jeśli dotyczy on np. przesyłu danych z zewnętrznych źródeł.

Wszelkie operacje na danych oraz próby autentyfikacji i autoryzacji użytkowników odbywają się w czasie rzeczywistym, co umożliwia administratorom podjęcie natychmiastowych działań. Wiele narzędzi pozwala zarządzanemu systemowi na ustalanie własnych zasad klasyfikacji oraz definiowania działań krytycznych użytkowników, a dodatkowo ustawianie sposobu powiadamiania w przypadku złamania kolejnych poziomów bezpieczeństwa.

5.2.2.5. Oracle Audit Vault and Database Firewall

Oracle AVDF (Oracle audit, 2021)[11] łączy możliwości zapory sieciowej baz danych z alarmowaniem i raportowaniem w formie aplikacji. Oracle AVDF zapewnia kompletne rozwiązanie monitorowania aktywności bazy danych, które łączy wraz z kontrolą przechwytywanego ruchu sieciowego. Zapora bazy danych wykorzystuje zaawansowany silnik analizy gramatycznej do sprawdzania instrukcji SQL wysyłanych

[11] Zob. specyfikację: https://docs.oracle.com/cd/E69292_01/doc.122/e49588/toc.htm.

do bazy danych i z dużą dokładnością określa, czy zezwalać, logować, ostrzegać, zastępować, a wręcz blokować przychodzące zapytania SQL.

AVDF współpracuje m.in. z bazami danych: Oracle Database, Oracle MySQL, Microsoft SQL Server, IBM DB2, SAP Sybase, Linux, AIX czy Solaris. Narzędzie to jest używane na różnych platformach internetowych, w informacyjnych bazach danych oraz w celu pobierania lub przesyłania nowych informacji. Dodatkowo oprogramowanie dostarcza szczegółową listę zagrożeń, ich opis oraz poziom niebezpieczeństwa.

Oracle AVDF najlepiej sprawdza się w branżach obarczonych wieloma regulacjami prawnymi, jak choćby branża finansowa, gdzie wymagana jest dodatkowa ochrona wrażliwych danych klientów, wynikająca z przepisów PCI-DSS czy RODO (Ustawa z dnia 10 maja 2018 r. o ochronie danych osobowych). Minusem tego rozwiązania jest fakt, że oprogramowanie działa najlepiej na bazie danych Oracle.

5.2.2.6. IBM Security Guardium

IBM Security Guardium (IBM security, b.r.) to przydatne narzędzie do klasyfikowania wrażliwych danych w relacyjnych i nierelacyjnych bazach danych oraz analizowania zachowań użytkowników końcowych, którzy mogą uzyskać dostęp do baz danych, samych administratorów bazy danych czy też deweloperów aplikacji[12]. Korzystanie z oprogramowania pozwala przygotować się organizacjom na zewnętrzne kontrole i audyty, np. odnośnie do bezpieczeństwa przechowywania wrażliwych danych osobowych.

Zaletą IBM Security Guardium są elastyczne rozwiązania szyfrujące, chroniące dane znajdujące się w środowiskach lokalnych pojedynczej chmury, wielu chmur lub hybrydowych. Oprogramowanie oferuje możliwości szyfrowania pojedynczych plików i ich woluminów, tokenizację, szyfrowanie aplikacji, szyfrowanie Teradata i funkcje zarządzania kluczami bezpieczeństwa, co pozwala nadzorować wrażliwe dane, egzekwować strategie dostępu i spełniać wymogi w zakresie zgodności.

Klienci korzystający z IBM Security Guardium wskazują jednak na wady produktu, takie jak trudne do zrozumienia dzienniki logów, z których trzeba wydobyć istotną dla użytkownika informację, np. dotyczącą komunikatów o błędach. Ponadto skomplikowany interfejs użytkownika utrudnia administratorom odpowiednią konfigurację oprogramowania.

5.2.2.7. Sophos Intercept X Server

Sophos Intercept X Server (Sophos Intercept, b.r.) to zestaw ochrony przed oprogramowaniem *ransomware*[13] oraz atakami i wirusami. Sophos Intercept X stosuje

[12] Por. także: https://www.ibm.com/pl-pl/products/ibm-guardium-data-protection.
[13] Złośliwe oprogramowanie blokujące dostęp do komputera do czasu zapłacenia okupu.

podejście do ochrony punktów końcowych oparte na kompleksowej ochronie, zamiast polegać jedynie na podstawowej technice bezpieczeństwa.

Sztuczna inteligencja wbudowana w Intercept X to głęboko ucząca się sieć neuronowa z zaawansowaną formą uczenia maszynowego, która wykrywa zarówno znane, jak i nieznane złośliwe oprogramowanie bez polegania na sygnaturach. Konfiguracja Sophos Intercept X jest łatwa dla początkujących użytkowników, a mimo to narzędzie posiada bogaty zakres funkcji, które zadowolą doświadczonych specjalistów ds. bezpieczeństwa. Szeroka gama wyrafinowanych funkcji chroniących przed złośliwym oprogramowaniem sprawiła, że Sophos Intercept X zdobył uznanie kilku niezależnych laboratoriów.

Samo narzędzie posiada jednak także wady, z których najbardziej istotną są duże wymagania na zasoby systemowe, jak również częste poprawki i zmiany w narzędziu, które niosą za sobą konieczność poświęcenia czasu i środków na utrzymanie i administrację aplikacji.

5.2.3. Zasady przetwarzania danych

Dane nie tylko muszą być bezpieczne w bazie danych GUS, lecz także podczas przetwarzania. Stąd procesy przetwarzania powinny odbywać się na komputerach całkowicie odciętych od dostępu z sieci, w trakcie gdy baza danych jest również odcięta od świata zewnętrznego. Wyniki powinny być składowane na powrót w bazie danych z odpowiednio przypisanymi uprawnieniami dostępu, a dopiero po zakończeniu pracy oprogramowania baza danych może ponownie zostać połączona z siecią.

5.2.4. Format i zakres ramki danych

Format i zakres danych jest zwykle wynikiem negocjacji GUS oraz poszczególnych gestorów, jednak dla procesu wyliczeń wskaźników cen powinno się mieć dostęp do co najmniej następujących danych:

- sieć – jednoznaczny w ramach GUS identyfikator sieci (stały w czasie);
- identyfikator sklepu – jednoznaczny w ramach sieci identyfikator sklepu, którego dotyczą dane sprzedażowe (identyfikator powinien przekładać się w dodatkowych danych np. na region sprzedaży);
- rok, miesiąc, dzień początkowy, dzień końcowy – okres czasu, którego dotyczą dane sprzedażowe;
- hierarchia sieci – informacja nt. poziomu zagnieżdżenia raportowanych informacji (np. dane na poziomie całej sieci, regionu bądź indywidualnego sklepu), która służy do interpretowania identyfikatora sklepu, jak i do stosowanych procesów podsumowań (aby nie liczyć dwukrotnie tych samych informacji na różnych poziomach agregacji);

- identyfikator produktu sieci – jednoznaczny kod produktu wewnątrz sieci handlowej;
- opis produktu – tekst zawierający istotne z punktu widzenia sprzedawcy informacje o produkcie;
- gramatura – masa jednostki sprzedaży produktu wyrażona w gramach;
- jednostka miary – jednostka sprzedażowa produktu (np. sztuki, litry itd.);
- cena jednostki sprzedaży produktu;
- ilość sprzedana w jednostkach sprzedaży produktu;
- obrót – wartość sprzedana w złotych;
- VAT – zastosowany podatek VAT;
- udział VAT w obrocie – wartość VAT w ramach obrotu;
- procent promocji w obrocie – procent sprzedanej wartości objętej promocjami.

Opis wykorzystania tych informacji podano na zakończenie omawiania poszczególnych elementów oprogramowania w podrozdz. 5.4.

5.3. Statystyczne metody zaimplementowane do klasyfikacji produktów

Jak łatwo można zauważyć na podstawie porównania metodyki badania cen detalicznych, opisanej w rozdz. 4, oraz informacji uzyskiwanych z danych skrapowanych (pkt 5.1.7) i danych skanowanych (pkt 5.2.4), istnieje pewien rozdźwięk między zebranymi danymi a zapotrzebowaniem na informację w celu analizy cen detalicznych. Dotyczy on w szczególności:

- przypisania kategorii COICOP do produktów (z sieci WWW i od gestorów);
- stwierdzenia ciągłości lub braku ciągłości sprzedaży produktów;
- doboru reprezentantów kategorii COICOP;
- stwierdzenia pojawienia się nowych relewantnych produktów;
- radzenia sobie z różnymi zabiegami biznesowymi sprzedawców i/lub producentów, polegającymi na zmianach jednostek sprzedaży, nazwy produktu itp.

Z uwagi na masowość przetwarzanych danych, jak i sporą częstość wymienionych zjawisk, potrzebne jest co najmniej częściowe automatyczne wsparcie analityków.

Wsparcie to w znakomitej części może być realizowane znanymi z dziedziny uczenia maszynowego technikami klasyfikacji.

5.3.1. Czym jest klasyfikacja

Klasyfikacja w rozumieniu systemów uczenia maszynowego to wykorzystanie opracowanych modeli zjawisk do przypisania tych zjawisk do jednej ze zdefiniowanych

z góry klas. W naszym przypadku chodzi o przypisanie produktom klas COICOP, lub też przypisanie innych produktów, które są z nimi tożsame.

Modelem (klasyfikatorem) będzie więc funkcja $f(\cdot)$, która opisowi $\mathbf{x}$ pewnego obiektu przypisywać będzie klasę $y(y = f(\boldsymbol{x}))$. W naszym przypadku opis obiektu (produktu) $\mathbf{x}$ będzie wektorem cech dostarczonym przez gestorów lub ekstrahowanym z internetu. W większości przypadków będzie to opis tekstowy. Opis tekstowy będzie traktowany jako „worek słów", więc wektor cech $\boldsymbol{x} = (x_1, \ldots, x_m)$ będzie wskazywać, czy i-te słowo ze słownika słów występuje w opisie $(x_i = 1)$, czy nie $(x_i = 0)$. x_i może być także liczbą wystąpień słowa lub inną miarą, np. uwzględniającą, czy słowo i jest słowem popularnym, czy rzadkim (miara tfidf). Zamiast słów wektor cech może zawierać n-gramy słów (np. bigram słów to dwa kolejno po sobie występujące słowa w tekście), n-gramy liter (przykładowo trigram liter to trzy kolejne litery w słowie, np. słowo „kwadrat" może być reprezentowane przez następujące trigramy liter: „kwa", „wad", „adr", „dra", „rat"). Bardziej szczegółowe uwagi na ten temat można znaleźć monografii Manning i in. (2008, podrozdz. 3.2).

Modele mogą powstawać jako wynik pewnego manualnego opracowania przez człowieka lub w wyniku procesu uczenia maszynowego.

W każdym z tych przypadków podstawą do opracowania modelu jest istnienie praw natury, które kojarzą ze sobą opis $\mathbf{x}$ z odpowiadającą mu klasą y. Gdyby taki związek między nimi nie istniał, zadanie konstrukcji modelu byłoby bezprzedmiotowe.

Co więcej, związek ten musi być w miarę prosty, tak abyśmy byli w stanie manualnie czy automatycznie wykryć go na podstawie obserwacji znanych związków między pewnymi $\mathbf{x}$ i y.

W praktyce takie związki są dość skomplikowane, ale dają się często dobrze przybliżyć za pomocą stosunkowo prostych modeli.

Zatem celem jest znalezienie funkcji f z pewnej rodziny (typu) funkcji F takiej, aby z wysokim prawdopodobieństwem przewidywać związek między cechami zjawiska czy obiektu a jego klasą.

W ramach dziedziny uczenia maszynowego wyróżnia się dwa główne typy modeli: czarnej skrzynki (ang. *blackbox*) i białej skrzynki (ang. *whitebox*). Modele białej skrzynki to takie, których istota modelu jest zrozumiała dla człowieka. Np. modele naiwnego Bayesa przyznają poszczególnym składowym obserwacji prawdopodobieństwa przypisania do klasy. Człowiek może zatem ocenić wizualnie, które z komponentów obserwacji (np. słów w opisie produktu) mają większy wpływ na wynik procesu klasyfikacji niż inne. Modele czarnej skrzynki, choć zwykle bardziej precyzyjne w działaniu, są trudne do zrozumienia, ponieważ składają się z szeregu współczynników powiązanych w uwikłany sposób tak, że związek między składową ob-

serwacji a klasą jest trudny do określenia. Ten typ modeli to np. sieci neuronowe czy maszyny wektorów nośnych.

Choć modele typu białej skrzynki mają zwykle niższą skuteczność, to ich wartość wykracza poza zadanie samej klasyfikacji – mogą wzbogacić rozumienie zjawisk lub przyczynić się do uproszczenia procesu modelowania (np. przez ignorowanie mało ważnych komponentów obserwacji).

W niniejszym rozdziale omawiamy możliwości automatycznej konstrukcji stosownego klasyfikatora. Proces tej konstrukcji nazywany jest uczeniem klasyfikatora.

Punktem wyjścia do automatycznego modelowania interesujących nas zależności jest posiadanie zbioru produktów $p_1, \ldots, p_n$, z których każdy jest opisany pewnym przetwarzalnym cyfrowo wektorem odpowiednio $\mathbf{x}_1, \ldots, \mathbf{x}_n$ oraz każdy z nich jest opisany dodatkowo klasą COICOP $y_1, \ldots, y_n$, do której należy. Nazywa się go zbiorem etykietowanym.

Pierwszym etapem pracy jest zastosowanie dla tych danych algorytmu uczącego, który wygeneruje interesujący nas klasyfikator. Tak powstały klasyfikator f możemy stosować do innych produktów $p'_1, \ldots, p'_m$, z których każdy jest opisany przetwarzalnym cyfrowo wektorem $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$, złożonym z takich samych cech, które użyliśmy do uczenia (trenowania) naszego modelu (klasyfikatora), aby przewidywać ich przynależność klasową:
COICOP $y'_1, \ldots, y'_m$.

Jednak aby czynić to w sposób odpowiedzialny, trzeba sprawdzić, na ile system uczący jest w stanie wykryć prawdziwą zależność w danych. Niezdolność do wykrycia może wynikać z kilku przyczyn, z których wymieńmy trzy najważniejsze: (1) liczne błędy w zbiorze etykietowanym (co do wartości klasy y i/lub wartości cech $\mathbf{x}$), (2) tendencyjny dobór zbioru etykietowanego, (3) nieadekwatność rodziny funkcji F do opisu interesującego nas zjawiska. Inne przyczyny to np. zbyt mały zbiór etykietowany dla celów algorytmu uczącego lub opisu zjawiska albo zmienność w czasie poszukiwanej zależności.

Pierwszą i drugą z wymienionych przyczyn chcemy w tym rozdziale wykluczyć, choć generalnie przy gromadzeniu danych trzeba mieć je na uwadze.

Występowanie trzeciego problemu w dziedzinie systemów uczących wykrywa się w sposób zautomatyzowany poprzez walidację. Przez walidację rozumie się ocenę poprawności działania systemu uczącego się klasyfikatora dla konkretnego zbioru danych.

Aby móc dokonać walidacji, należy losowo lub w inny adekwatny do zadania sposób (np. w sekwencji czasowej) podzielić zbiór etykietowany na rozłączne zbiory: zbiór uczący (zwany też trenującym) $\mathbf{x}_{u1}, \ldots, \mathbf{x}_{un}$, $y_{u1}, \ldots, y_{un}$ i walidujący (zwany też testowym) $\mathbf{x}_{w1}, \ldots, \mathbf{x}_{wn}$, $y_{w1}, \ldots, y_{wn}$, $u_n + w_n = n$, $u_n : w_n = 2 : 1$.

Trenowanie klasyfikatora odbywa się na zbiorze trenującym $\mathbf{x}_{u1},\ldots,\mathbf{x}_{un}$, $y_{u1},\ldots,y_{un}$. Otrzymany model f_u testuje się następnie na zbiorze testującym, obliczając: $y'_{w_1} = f_u(\mathbf{x}_{w_1}),\ldots,y'_{w_n} = f_u(\mathbf{x}_{w_n})$. Następnie konfrontujemy $y'_{w_1},\ldots,y'_{w_n}$, z $y_{w_1},\ldots,y_{w_n}$ na zgodność przewidywań klasyfikatora f_u ze stanem faktycznym. Dokładność modelu (ang. *precision*) to liczba przypadków, gdy $y'_{w_i} = y'_{w_i}$, podzielona przez w_n. Często porównuje się dokładność modelu z modelem losowym („zgadującym" przynależność do klas), aby stwierdzić, czy rzeczywiście model zawiera jakąś wiedzę o zjawisku.

Jeżeli zbiór etykietowany jest zbyt mały, to zamiast opisanej jednokrotnej walidacji stosuje się technikę k-krotnej walidacji krzyżowej lub kroswalidacji (ang. *cross validation*). Polega ona na losowym podziale zbioru etykietowanego na k (prawie) równych rozłącznych części $\mathbf{x}_{11},\ldots,\mathbf{x}_{1n}$ $y_{11},\ldots,y_{1n}$, $\mathbf{x}_{k1},\ldots,\mathbf{x}_{kn}$, $y_{k1},\ldots,y_{kn}$. Następnie uczy się k modeli $f_{u1},\ldots,f_{uk}$, przy czym dla funkcji f_{uj} zbiorem walidacyjnym jest $\mathbf{x}_{j1},\ldots,\mathbf{x}_{jn}$ $y_{j1},\ldots,y_{jn}$, a zbiorem uczącym są pozostałe dane etykietowane. Dokładność modelu (ang. *precision*) to liczba przypadków, gdy $y'_{j_i} = y_{j_i}$, zliczona po wszystkich $j = 1,\ldots,k$, podzielona przez n.

W projekcie wykorzystano metody (algorytmy) uczenia czterech typów klasyfikatorów: naiwny klasyfikator bayesowski, regresja logistyczna, lasy losowe, oraz maszyny wektorów nośnych. Opisujemy je krótko poniżej. Znakomite wprowadzenie do tych metod można znaleźć w monografii Koronackiego i Ćwika (2005), natomiast bardziej szczegółowy i pogłębiony matematycznie opis znajduje się w pracach Gareth i in. (rozdziały: 4 – regresja logistyczna, 8 – lasy losowe i 9 – klasyfikatory oparte o maszyny wektorów nośnych) oraz Bishop (2006, rozdział 1 – naiwny klasyfikator bayesowski).

W eksperymentach oraz implementacji wykorzystano ogólnodostępny moduł Scikit-Learn[14], który zawiera napisane w języku Python implementacje większości algorytmów stosowanych w uczeniu maszynowym. Moduł ten może być z powodzeniem używany do analizy danych, wizualizacji i konstrukcji praktycznych modeli. Z tego potężnego zestawu wybrano cztery wspomniane wcześniej algorytmy.

Regresja logistyczna oraz maszyny wektorów nośnych w wersji bazowej działają tylko dla dwóch klas lub kategorii (dychotomizatory). Aby zapewnić ich działanie dla większej liczby kategorii, buduje się docelowy klasyfikator z wielu dychotomizatorów, np. za pomocą metod:

- OVO (ang. *one versus one*): tworzy się dychotomizator dla każdej pary różnych kategorii i wektor x jest zaklasyfikowany do klasy, która ma najwięcej „wygranych" z innymi klasami;
- OVA (ang. *one versus all*): tworzy się tyle dychotomizatorów, ile jest kategorii, a decyzja to klasa „wygrywająca" (np. największa wartość logitowa w regresji logi-

[14] https://scikit-learn.org/stable/.

stycznej lub największa pozytywna wartość funkcji celu w klasyfikatorze na bazie maszyn wektorów nośnych itd.).

Poniżej przedstawiamy charakterystykę wymienionych klasyfikatorów.

5.3.2. Regresja logistyczna

Dla dwóch klas C_1, C_2 symbol $p(\boldsymbol{x})$ oznacza prawdopodobieństwo, że wektor atrybutów $\mathbf{x}$ reprezentuje obserwację należącą do klasy C_1.

Metoda regresji logistycznej (ang. *logistic regression*) modeluje prawdopodobieństwo $p(\boldsymbol{x})$ jako $\frac{e^{\mathbf{a}^T\mathbf{x}+b}}{1+e^{\mathbf{a}^T\mathbf{x}+b}}$, gdzie wektor $\mathbf{a}$ i skalar b są parametrami do wyuczenia przez model.

Stąd wynika funkcja logitowa, $\log\frac{p(x)}{1-p(x)}$, czyli logarytm z ilorazu prawdopodobieństw, że $\mathbf{x}$ reprezentuje klasę C_1 (licznik) i klasę C_2 (mianownik). Jest to funkcja liniowa: $\mathbf{a}^T\boldsymbol{x} + b$ parametrów $\mathbf{a}$ i b. Do ich wyznaczenia stosuje się np. metodę największej wiarogodności (Bishop, 2006, pkt 4.3.2).

W przypadku wielu klas stosuje się opisane wcześniej podejścia OVO lub OVA.

Dokumentacja wykorzystywanej implementacji klasyfikatora regresji logistycznej (wraz z jego algorytmem uczącym) znajduje się pod adresem: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

5.3.3. Naiwny klasyfikator bayesowski

W tym przypadku zakłada się na ogół, że wszystkie atrybuty są nominalne, tzn. przyjmują wartości (etykiety), dla których nie istnieje wynikające z natury danego zjawiska uporządkowanie. Przykładami takich cech są: kolor oczu, płeć, wyznanie, miasto zamieszkania, itp. W naszym przypadku może to być występowanie bądź brak konkretnych słów w opisie produktu, ponadto miejsce sprzedaży, typ opakowania, występowanie promocji itp.

Dany jest zbiór treningowy T składający się z N n-wymiarowych wektorów atrybutów (cech), $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$, $i = 1, \ldots, N$. Tutaj x_{ij} oznacza wartość j-ego atrybutu (cechy) X_j zaobserwowaną w i-tym obiekcie.

Zarówno atrybuty $X_1, \ldots, X_n$, jak i atrybut decyzyjny $Y = [y_1, \ldots, y_k]$ (tzn. klasyfikujemy do jednej z k klas) traktuje się jako zmienne losowe, przy czym naiwność klasyfikatora polega na założeniu, że cechy są niezależnymi zmiennymi losowymi.

Do zaklasyfikowania wektora $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ do jednej z klas decyzyjnych korzysta się z twierdzenia Bayesa:

$$P(Y = y_i | X = x) = \frac{P(X = X | Y = y_i) P(Y = y_i)}{P(X = x)}.$$

Wektor $\mathbf{x}$ przydziela się do klasy (wartość atrybutu decyzyjnego) y, dla której to prawdopodobieństwo jest największe. Innymi słowy, wektor $\mathbf{x}$ należy do klasy y_{i^*}, takiej że

$$P(Y = y_{i^*}|X = x) = \max_{i=1,\dots,k} P(Y = y_i|X = x).$$

Przy uwzględnieniu faktu, że zmienne X_j są statystycznie niezależne, oraz że w powyższym wzorze wartość $P(\boldsymbol{x})$ jest identyczna dla wszystkich $i = 1, \dots, k$, wyznacza się wartość formuły:

$$f(y_i) = \left(\hat{P}(x_1|Y = y_i) \cdots \hat{P}(x_n|Y = y_i)\right) \cdot \hat{P}(y_i),$$

gdzie $\hat{P}(x_1|Y = y_i)$ oraz $\hat{P}(Y = y_i)$ to oszacowania odpowiednich prawdopodobieństw wyliczone na podstawie zbioru trenującego.

W wypadku, gdy zakładamy, że dla każdej pary (i, j) rozkład prawdopodobieństwa $\hat{P}(x_1|Y = y_i)$ zmiennej x_j względem klasy y_i jest rozkładem wielomianowym (tzn. zmienna x_j przyjmuje dwie lub więcej, ale skończoną liczbę różnych wartości), będziemy mówić o *wielomianowym naiwnym klasyfikatorze bayesowskim* (ang. *Multinomial Naive Bayes Classifier*, w skrócie *MultinomialNB*).

Dokumentacja wykorzystywanego naiwnego klasyfikatora bayesowskiego znajduje się pod adresem: https://scikit-learn.org/stable/modules/ generated/sklearn.naive _bayes.MultinomialNB.html#sklearn.naive_ bayes.MultinomialNB.

5.3.4. Metoda wektorów nośnych

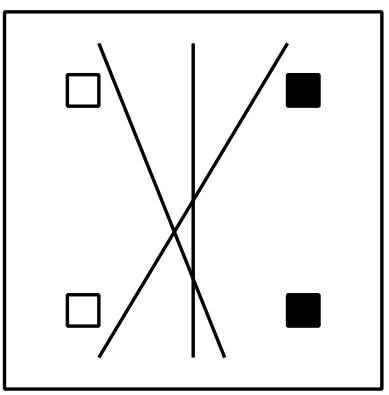
Jest to uogólnienie idei klasyfikacji z pomocą hiperpłaszczyzn dyskryminacyjnych podane przez Vladimira Vapnika (por. podrozdz. 6.2 w pracy Koronackiego i Ćwika, 2005). Punktem wyjściowym tej idei jest nowe spojrzenie na zadanie wyboru hiperpłaszczyzny rozdzielającej (separującej) grupy obiektów. W przypadku zbiorów *liniowo* separowalnych istnieje na ogół wiele hiperpłaszczyzn oddzielających poszczególne klasy, co ilustruje lewa strona schematu 5.9. Puste kwadraty oznaczają tu elementy zbioru C_1, a wypełnione kwadraty – elementy zbioru C_2. Liniami przerywanymi zaznaczono potencjalne hiperpłaszczyzny (w tym wypadku są to odcinki prostych) rozdzielające obie klasy. Pomysł Vapnika polega na wyznaczeniu[15] lokalizacji zapewniającej, że rozdzielająca obie klasy hiperpłaszczyzna jest umieszczona środku marginesu, czyli najszerszego możliwego do uzyskania pasa[16] rozdzielającego

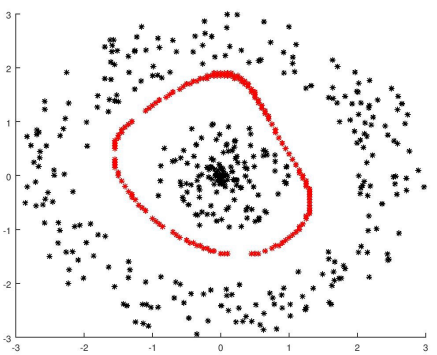
[15] Poprzez rozwiązanie odpowiednio sformułowanego zadania optymalizacji z kwadratową funkcją celu i liniowymi ograniczeniami.
[16] Jest to swego rodzaju ziemia niczyja: w obszarze wewnątrz pasa nie ma żadnych obserwacji.

obie klasy. Ta wersja metody nosi nazwę liniowego klasyfikatora na bazie wektorów nośnych (*LinearSVC*).

W przypadku zbiorów, które nie są liniowo separowalne, wyznaczanie takiej hiperpłaszczyzny wymaga wstępnego umieszczenia oryginalnych danych w przestrzeni cech. Operację tę wykonuje się z użyciem funkcji jądrowych (maszyna wektorów nośnych z jądrem nieliniowym – *SVM/SVC*). Najpopularniejszą funkcja jądrową jest jądro gaussowskie, nazywane też radialną funkcją jądrową, w skrócie RBF (ang. *radial based function*). Taka operacja powoduje „uliniowienie" zbioru danych w nowej przestrzeni, co w konsekwencji wymaga użycia prostego algorytmu wyznaczania liniowej hiperpłaszczyzny separującej. Udowadnia się, że zastosowanie jądra RBF pozwala klasyfikować nierozdzielne liniowo klasy. Przykładowy efekt działania takiej procedury przedstawiono na prawej stronie schematu 5.9, gdzie kolorem czerwonym zaznaczono hiperpłaszczyznę rozdzielającą.

Schemat 5.9. Przykłady zbiorów liniowo i nieliniowo separowalnych





Źródło: opracowanie własne.

W przypadku wielu klas stosuje się opisane wcześniej podejścia OVO lub OVA.

Dokumentacja wykorzystywanych klasyfikatorów maszyny wektorów nośnych znajduje się dla wersji liniowej pod adresem: https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html, a dla wersji z jądrem radialnym: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC.

5.3.5. Lasy losowe

Podejście do klasyfikacji na podstawie lasów losowych (ang. *random forest classifier*) polega na zbudowaniu klasyfikatora zespołowego, czyli klasyfikatora złożonego z wielu prostych klasyfikatorów nazywanych też *słabymi uczniami*. Zbiór uczniów obsługujących tworzony klasyfikator stanowi *rodzinę*. Od członków rodziny wymaga

się tylko, aby podejmowali prawidłowe decyzje z prawdopodobieństwem nieco większym od decyzji losowej, np. z prawdopodobieństwem równym 0,55. Zapewnia to, że jeżeli rodzina jest dostatecznie duża, np. zawiera 1000 członków, możemy być prawie pewni, że większość słabych, ale niezależnych uczniów dokona poprawnej klasyfikacji. Specyfikę takich rodzin klasyfikatorów przystępnie Koronacki i Ćwik (2005, podrozdz. 4.6). W przypadku lasów losowych[17] prostymi klasyfikatorami są drzewa decyzyjne. Takie podejście ma na celu m.in. redukcję wariancji docelowego klasyfikatora, czyli podniesienie jego odporności na szum i zmiany w danych uczących.

W przypadku lasów losowych każde składowe drzewo decyzyjne jest budowane w ten sposób, że przy wyborze atrybutów na każdym poziomie budowanego drzewa losuje się tylko $m < n$ spośród wszystkich n atrybutów. Dzięki temu wynikowe drzewa są nieskorelowane co daje niższą wariancję wynikową zespołu drzew. Typowym wyborem wobec wartości m jest $m = \sqrt{n}$.

Dokumentacja wykorzystywanego klasyfikatora lasów losowych znajduje się pod adresem: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

5.4. Implementacje metod informatycznych do procedowania danych ze źródeł alternatywnych

W ramach projektu InstatCeny zaimplementowano trzy moduły z interfejsami w konwencji REST endpoints, wspomagające przetwarzanie danych pochodzących ze źródeł alternatywnych:

- klasyfikator produktów do kategorii COICOP IC.Classifier (pkt 5.4.1);
- moduł dopasowania produktów albo śledzenia między okresami czasu IC.Matcher (pkt 5.4.2);
- moduł obliczania wskaźników zmienności cen IC.PriceIndexer (pkt 5.4.3).

5.4.1. Klasyfikator produktów do grup COICOP

Moduł IC.Classifier oferuje następujące usługi:

- klasyfikacja produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu;
- zaawansowana klasyfikacja produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu oraz dodatkowych cech;
- trenowanie klasyfikatora produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu;

[17] Zaproponował je w 2001 r. Leo Breiman. Prowadzi on stronę http://www.stat. berkeley.edu/users/breiman/RandomForests, na której można znaleźć uwagi o zastosowaniach w klasyfikacji, grupowaniu, regresji, oraz analizie przeżycia.

- trenowanie zaawansowanego klasyfikatora produktów do drzewa kategorii COICOP na podstawie tekstowego opisu produktu oraz dodatkowych cech;
- zarządzanie modelami wytrenowanych klasyfikatorów.
- Klasyfikator może być jednego z następujących typów:
- wielomianowy naiwny klasyfikator Bayesa (MultinomialNB), opisany w pkt 5.3.3;
- liniowa maszyna wektorów nośnych (LinearSVC), opisana w pkt 5.3.4;
- regresja logistyczna (Logistic Regression), opisana w pkt 5.3.2;
- lasy losowe (Random Forest Classifier), opisane w pkt 5.3.5.

Domyślnie wybierany jest wielomianowy naiwny klasyfikator Bayesa.

Dane wejściowe są przechowywane w pliku CSV. Klasyfikacja może odbywać się także dla danych wprowadzanych z terminala.

Dane wyjściowe są zapisywane do pliku CSV lub wypisywane na ekranie.

Modele są archiwizowane w bazie danych modeli.

Szczegółowy opis można znaleźć w dokumentacji technicznej (Czerski i in., 2022).

IC.Classifier wystawia swoje usługi jako endpointy REST-owe (podobnie jak scraper w pkt 5.1.6).

5.4.1.1. Format i zakres ramki danych

5.4.1.1.1. Klasyfikacja produktów

Wejście stanowi nazwa pliku wraz z jego pełną ścieżką w formacie CSV (oddzielany przecinkami). Plik musi zawierać pole *desc* zawierające opis produktu. Ponadto musi zawierać inne pola, jeśli były użyte w procesie trenowania.

Wyjście stanowi plik w postaci CSV o nazwie prediction.csv, w którym znajduje się kopia pliku wejściowego z dodatkową kolumną *coicpo.predicted*, w której znajdują się predykowane kategorie. Plik ten jest zapisywany w katalogu /katalog_na_wyniki.

5.4.1.1.2. Trenowanie klasyfikatora produktów

Wejście stanowi nazwa pliku wraz z jego pełną ścieżką w formacie CSV (oddzielany przecinkami). Plik musi zawierać pole *desc* zawierające opis produktu oraz pole *coicop* zawierające docelową kategorię dla każdego produktu. Ponadto musi zawierać inne pola, jeśli mają być użyte w procesie trenowania. Poza tym przekazywane są następujące dane:

- typ_klasyfikatora – wybierany typ klasyfikatora – 0: MultinomialNB, 1: LinearSVC, 2: Logistic Regression, 3: Random Forest Classifier; domyślnie wybierany jest MultinomialNB;
- opis_modelu – dodatkowy opis tekstowy dla danego klasyfikatora, który pojawi się w nazwie pliku z modelem i ułatwi jego identyfikację;

- encoding – kodowanie w jakim jest zapisany plik CSV, domyślnie: 'utf8'. (uwaga: jeśli 'utf-8' nie zadziała, warto spróbować z wartością 'cp1250');
- dodatkowe_zm_uczace – lista pól, które będą użyte w trenowaniu prócz pola descr.
- Wyjście stanowi wytrenowany klasyfikator, który będzie zapisany w katalogu /katalog_na_wynik. Format jest specyficzny dla użytego typu klasyfikatora i parametrów wejściowych.

5.4.2. Dopasowanie produktów w czasie (matcher)

Moduł IC.Matcher oferuje następujące usługi:

- kojarzenie (ang. matching) produktów (por. ppkt 5.4.2.1);
- grupowanie produktów w produkty tożsame (por. ppkt 5.4.2.2).

IC.Matcher wystawia swoje usługi jako endpointy REST-owe (podobnie jak scraper w podrozdz. 5.1.6).

5.4.2.1. Kojarzenie (matching) produktów

Algorytm sprawdza, czy dwie zadane reprezentacje opisują ten sam produkt. Zakłada się, że reprezentacje te posiadają następujące pola:

- ean – kod EAN;
- seller_code – kod produktu dostarczany przez sprzedawcę;
- desc – opis produktu (etykieta).

Procedura porównywania produktów wykonywana jest w poniższych krokach:

- Za dopasowane produkty uznajemy te, które mają ten sam kod kreskowy EAN (cecha „ean") i kod produktu według sprzedawcy (cecha „seler_code").
- W przypadku produktów, które mają ten sam kod produktu według sprzedawcy, ale różny kod EAN, nadal jest duża szansa na to, że jest to jakościowo ten sam produkt. Wówczas za dopasowane uznajemy te produkty, które mają tę samą etykietę (cecha „desc") produktu lub etykietę bardzo podobną. W tym celu oblicza się podobieństwa pomiędzy opisami produktów za pomocą metody bazującej na algorytmie Ratcliffa i Obershelpa (Black, 2004)[18] (implementacja pochodzi z pakietu Python *difflib.SequenceMatcher*). Przy tak wyliczonym podobieństwie za produkty podobne uważane są te o podobieństwie powyżej parametru *precision*.

[18] Podobieństwo dwóch ciągów znaków S_1 i S_2 obliczana jest jako iloraz podwojonej liczby K_m znaków zgodnych w obu tekstach oraz sumy długości obu tekstów: $D_{RaOb} = \frac{2K_m}{|S_1|+|S_2|}$, przy czym $0 \leq D_{RaOb} \leq 1$. Liczbę wspólnych znaków K_m określa się w sposób następujący: poszukujemy najdłuższego wspólnego podciągu znaków w obu tekstach W, dzielimy S_1 i S_2 na $S_1 = S_{1l} + W + S_{1p}$, $S_2 = S_{2l} + W + S_{2p}$ i powtarzamy to poszukiwanie w parach (S_{1l}, S_{2l}) i (S_{1p}, S_{2p}) aż do wyczerpania. K_m jest sumą długości wszystkich wspólnych podciągów. W wypadku istnienia kilku wspólnych podciągów W równej długości wybieramy ten, który jest najbardziej z lewej strony w ciągu S_1. Stąd miara nie jest symetryczna.

- W przypadku gdy produkty mają różny kod sprzedawcy i różne kody EAN, za dopasowane uznamy te produkty, które mają identyczną etykietę.

5.4.2.2. Grupowanie produktów w produkty tożsame

Procedura grupowania produktów w produkty tożsame bazuje na algorytmie porównywania produktów parami (por. ppkt 5.4.2.1), tak aby dokonać grupowania tych produktów dla całej ramki danych. Podobnie jak wcześniej zakłada się, że każdy produkt jest opisany cechami *ean*, *seller_code*, *desc*.

W omawianej procedurze zaimplementowano dwa różne algorytmy, które mogą zostać użyte do wykonania tego zadania:

- procedura dokładna, która bazuje na obliczaniu podobieństw dla wszystkich par zadanych produktów; dla przyspieszenia tej operacji obliczanie podobieństw wykonywane jest w sposób równoległy na wszystkich dostępnych rdzeniach procesora:
  - produkty przetwarzamy po kolei;
  - dany produkt porównujemy po kolei z już stworzonymi grupami (porównanie z grupą polega na porównaniu z tylko jednym jej elementem, który nazywamy jej reprezentantem);
  - jeżeli produkt nie jest podobny do żadnej z już istniejących grup, wtedy tworzona jest dla niego nowa grupa, a produkt ten staje się jej reprezentantem.

Schemat 5.10. Ramka danych używana w procedurze matchingu z dodaną kolumną wynikową

1	seller_code	ean	desc	coicop	group
2	[illegible]	[illegible]	[illegible]	[illegible]	0
3	[illegible]	[illegible]	[illegible]	[illegible]	0
4	[illegible]	[illegible]	[illegible]	[illegible]	0
5	[illegible]	[illegible]	[illegible]	[illegible]	1
6	[illegible]	[illegible]	[illegible]	[illegible]	1
7	[illegible]	[illegible]	[illegible]	[illegible]	1
8	[illegible]	[illegible]	[illegible]	[illegible]	0
9	[illegible]	[illegible]	[illegible]	[illegible]	0
10	[illegible]	[illegible]	[illegible]	[illegible]	0
11	[illegible]	[illegible]	[illegible]	[illegible]	0
12	[illegible]	[illegible]	[illegible]	[illegible]	1
13	[illegible]	[illegible]	[illegible]	[illegible]	0
14	[illegible]	[illegible]	[illegible]	[illegible]	0
15	[illegible]	[illegible]	[illegible]	[illegible]	1
16	[illegible]	[illegible]	[illegible]	[illegible]	0

Źródło: opracowanie własne.

Implementacja jest opisana bardziej szczegółowo w dokumentacji technicznej (Czerski i in., 2022).

5.4.2.3. Format i zakres ramki danych

5.4.2.3.1. Grupowanie produktów w produkty tożsame

Wejściem jest plik CSV, opisujący produkty. Dla każdego produktu należy podać:

- ean – kod EAN;
- seller_code – kod produktu dostarczany przez sprzedawcę;
- desc – opis produktu.

Na wyjściu pojawi się plik CSV będący kopią wejściowego z dodatkowym polem *group* oznaczającym grupę, do której należy produkt. Przykład grup przypisanych produktom pokazuje schemat 5.10 (zob. kolumna *group*).

5.4.3. Analiza wskaźników cen detalicznych (PriceIndices)

Aplikacja IC.PriceIndexer udostępnia funkcje z projektu *PriceIndices*[19] utworzonego przez prof. Jacka Białka. Aplikacja wystawia wybrane funkcje z projektu jako endpointy REST-owe (podobnie jak w pkt 5.1.6).

Wystawione funkcje umożliwiają przygotowanie danych skanowanych do obróbki pod kątem liczenia wskaźników inflacji. Obejmują funkcje wyboru danych, klasyfikacji, doboru i filtracji, a także obliczania samych wskaźników inflacji.

Zestaw pozwala m.in. na uczenie klasyfikatora do klas COICOP (funkcja *model_classification()*) na bazie algorytmu XGBoost do uczenia modelu dla danych numerycznych. Nauczony model może być następnie wykorzystany do klasyfikacji produktów do klas COICOP (funkcja *data_classifying()*).

Ponadto aplikacja wystawia funkcje o podobnej funkcjonalności do IC.Classifiera – są to funkcje:

- data_matching() – kojarzenie produktów w grupy produktów tożsamych;
- data_classifying() – przypisanie produktom klasy COICOP.

Więcej informacji o samym projekcie bazowym (m.in. pełen wykaz funkcji i przykłady użycia) jest dostępnych pod adresem: https://cran.r-project.org/web/packages/PriceIndices/index.html. Podstawowa funkcjonalność zestawu z punktu widzenia obliczania wskaźników cen detalicznych została omówiona w podrozdz. 7.3. Natomiast implementacja jest opisana bardziej szczegółowo w dokumentacji technicznej (Czerski i in., 2022).

Dane wejściowe mają zwykle postać szeregów czasowych. Podobną postać mają dane wyjściowe. Ich szczegółowa forma opisana jest na stronie https://cran.r-project.org/web/packages/PriceIndices/index.html. Szczegółowe dane dotyczące produktów będące danymi wejściowymi funkcji są w pełni kompatybilne z modelem produktu zwracanym przez moduły IC.Classifier oraz IC.Matcher.

Format danych wejściowych jest tożsamy dla wszelkich metodyk liczenia inflacji (zgodnie ze wspomnianą biblioteką PriceIndices), np. fisher, jevons czy young:

[19] Projekt dostępny pod adresem https://github.com/JacekBialek/PriceIndices.

- data – tablica produktów będąca bazą do liczenia inflacji; format ramki produktu:
  – time – data uzyskania informacji;
  – prices – cena produktu;
  – quantities – ilość produktu;
  – prodID – zewnętrzny kod produktu, np. GTIN, EAN;
  – retID – wewnętrzny kod produktu w obrębie sieci handlowej;
  – description – opis produktu;
- start – data będąca podstawą obliczeń;
- end – data końcowa analizowanego okresu;
- interval – flaga definiująca, czy w ramce zwrotnej zawrzeć pośrednie wskaźniki inflacji, np. w stosunku msc/msc.

Rezultatem powyższych obliczeń jest wartość, a w przypadku flagi „interval" tablica wartości opisująca wielkość inflacji w zadanym okresie.

ROZDZIAŁ 6
Przetwarzanie i analiza danych uzyskiwanych ze źródeł alternatywnych

W niniejszym rozdziale zaprezentowano eksperymenty, których celem było sprawdzenie, na ile wybrane metody uczenia maszynowego sprawdzą się przy dopasowywaniu (klasyfikacji) produktów z danych skanowanych do właściwych pozycji 5-cyfrowej klasyfikacji COICOP w różnych scenariuszach użycia. Dotyczy on w szczególności:

- przypisania kategorii COICOP do produktów (zarówno w przypadku danych pobranych z internetu, jak i uzyskanych od gestorów);
- stwierdzenia ciągłości / braku ciągłości sprzedaży produktów;
- doboru reprezentantów kategorii COICOP;
- stwierdzenia pojawienia się nowych, relewantnych produktów;
- radzenia sobie z różnymi zabiegami biznesowymi sprzedawców i/lub producentów, polegającymi na zmianach jednostek sprzedaży, nazwy produktu itp.

Pytamy, czy klasa COICOP może być w ogóle przewidziana z opisu produktu. Ponadto chcemy wiedzieć, czy można wyróżnić słowa charakterystyczne dla poszczególnych klas COICOP i na ich podstawie dokonać klasyfikacji. Wreszcie chcemy wiedzieć, czy wytrenowany klasyfikator dobrze zachowuje się w czasie, tzn. czy nowe produkty pojawiające się w kolejnych miesiącach mogą mieć automatycznie przypisaną klasę COICOP.

Ponadto przedmiotem zainteresowania autorów jest kwestia ciągłości sprzedaży produktów, w szczególności to, na ile nowo pojawiające się produkty są tożsame z już istniejącymi.

6.1. Opis próby badawczej

W ramach prac eksperymentalnych wykorzystano pięć zbiorów danych, z tego trzy krótkookresowe oraz dwa zbiory długookresowe.

6.1.1. Dane – zestaw krótkookresowy

Do eksperymentów użyto danych, które zawierały informacje o poszczególnych produktach, takie jak „Nr Produktu", „Nazwa Produktu" i kod COICOP. Dane te pochodziły z czterech sieci supermarketów: GestorAlfa ($N = 3298$), GestorBeta ($N = 954$), GestorGamma ($N = 1293$) i GestorKappa ($N = 4\ 888\ 221$). Dodatkowo stworzono też zbiór, który powstał z tych trzech zestawów danych. Zbiór ten nazwany został „3 sieci" ($N = 5545$). Uzyskane zbiory danych różnią się liczbą klas k. W pierwszym zbiorze wyróżniono 25 klas, w drugim – 15, w trzecim – 13, a liczba klas w agregowanym zbiorze jest równa 26. Przy wyznaczaniu liczebności klas pomijane były te z nich, które zawierały mniej niż pięć różnych produktów. Jest to motywowane użyciem omawianej dalej kroswalidacji (walidacji krzyżowej) z parametrem $K = 5$.

Zestawienia liczebności poszczególnych klas dla ww. zbiorów danych znajdują się w tabl. 6.1 i 6.2. Dla wygody w tabl. 6.3 zamieszczono łączne zestawienie częstości w poszczególnych zbiorach danych.

6.1.2. Dane – zestaw długookresowy

Druga porcja danych to duże dane sprzedażowe pochodzące z sieci GestorAlfa i GestorDelta. Badane dane z sieci GestorAlfa dotyczą okresu 12.2017–08.2019, a dane z sieci GestorDelta obejmują okres 10.2017–10.2018. Zbiory te w dalszej części raportu oznaczane będą jako GestorAlfa 2018/9 i GestorDelta 2018.

Tabl. 6.1. Liczebności poszczególnych klas COICOP z poziomu 5 dla zbiorów danych opisanych w podrozdz. 6.1.1 – GestorGamma i GestorBeta

GestorGamma ($k = 13$)	
Klasa COICOP	Liczebność
CP012111	565
CP011441	350
CP011111	68
CP011421	52
CP011462	49
CP011811	47
CP011121	38
CP011411	35
CP011461	23
CP011431	17
CP011991	10
CP011861	9
CP011122	9

GestorBeta ($k = 15$)	
Klasa COICOP	Liczebność
CP012111	296
CP011441	261
CP011462	102
CP011111	50
CP011811	35
CP011121	31
CP011421	30
CP011431	21
CP011411	20
CP011861	19
CP011461	19
CP012131	18
CP011452	13
CP011122	11
CP011181	6

Źródło: opracowanie własne.

Dane te zawierały informacje o poszczególnych produktach, takich jak: siec, id_sklepu, rok, miesiac, dzien_poczatek, dzien_koniec, hierarchia_sieci, id_towaru_sieci, opis_towaru, gramatura, jednostka_miary, cena, obrot, ilosc, vat, inne, symbol_coicop, nazwa_coicop, id, obrot_procent_promocji i obrot_vat.

W eksperymentach zajmowano się tylko rekordami, w których występowały niepuste wartości przypisanych im klas COICOP (kolumna symbol_coicop), co przełożyło się na następującą liczbę badanych przypadków: GestorAlfa 2018/9 $N = 637\,339$, GestorDelta 2018 $N = 2\,682\,354$.

Zestawienia liczebności poszczególnych klas wraz z ich kodami i etykietami dla ustalonych zbiorów danych znajdują się w tabl. 6.4 dla sieci GestorAlfa i tabl. 6.5 dla sieci GestorDelta.

6.1.3. Dane – zestaw do dopasowania produktów

Trzecia porcja danych zawiera duże dane sprzedażowe pochodzące z sieci GestorKappa. Badane dane z tej sieci dotyczą okresu 12.2020–10.2021. Zbiór ten w dalszej części będzie oznaczany przez GestorKappa 2020/21.

Tabl. 6.2. Liczebności poszczególnych klas COICOP z poziomu 5 dla zbiorów danych opisanych w podrozdz. 6.1.1 – GestorAlfa i 3 sieci

GestorAlfa ($k = 25$)		3 sieci ($k = 26$)	
Klasy COICOP	Liczebność	Klasy COICOP	Liczebność
CP011441	624	CP012111	1245
CP011462	598	CP011441	1235
CP012111	384	CP011462	749
CP011461	268	CP011461	310
CP011991	217	CP011991	231
CP011631	209	CP011631	212
CP011181	127	CP011111	198
CP011121	107	CP011121	176
CP011123	107	CP011421	171
CP011421	89	CP011181	138
CP011452	87	CP011811	126
CP011111	80	CP011123	107
CP011841	64	CP011411	107
CP011411	52	CP011452	100
CP012131	47	CP012131	66
CP011811	44	CP011431	65
CP011732	39	CP011841	65
CP011431	27	CP011122	41
CP011921	26	CP011861	40
CP011122	21	CP011732	39
CP011141	15	CP011921	26
CP011861	12	CP012221	20
CP012221	12	CP011141	16
CP011821	11	CP011821	13
CP056121	6	CP011911	9
		CP056121	6

Źródło: opracowanie własne.

Dane te zawierały następujące informacje o produktach: time, prices, quantities, seller_code, ean, desc i coicop[1]. Zostały użyte do eksperymentów związanych z dopasowywaniem produktów (por. ppkt 6.2.8.2), dlatego wykorzystywane były w tych eksperymentach tylko informacje zawarte w kolumnach: time, seller_code, ean, desc i coicop.

Tabl. 6.3. Liczebności poszczególnych klas COICOP z poziomu 5 dla czterech zbiorów danych z podrozdz. 6.1.1

Klasy COICOP	Liczebność grupy			
	GestorGamma	GestorBeta	GestorAlfa	3 sieci
CP012111	565	296	384	1245
CP011441	350	261	624	1235
CP011462	49	102	598	749
CP011461	23	19	268	310
CP011991	10	–	217	231
CP011631	–	–	209	212
CP011111	68	50	80	198
CP011121	38	31	107	176
CP011421	52	30	89	171
CP011181	–	6	127	138
CP011811	47	35	44	126
CP011123	–	–	107	107
CP011411	35	20	52	107
CP011452	–	13	87	100
CP012131	–	18	45	66
CP011431	17	21	27	65
CP011841	–	–	64	65
CP011122	9	11	21	41
CP011861	9	19	12	40
CP011732	–	–	39	39
CP011921	–	–	26	26
CP012221	–	–	12	20
CP011141	–	–	15	16
CP011821	–	–	11	13
CP011911	–	–	–	9
CP056121	–	–	6	6

Uwaga. Wskazano liczebności większe niż 5.
Źródło: opracowanie własne.

W zbiorze tym łącznie było $N = 4\,888\,221$ obserwacji, co oznacza, że każdego miesiąca dysponujemy średnio 45 000 obserwacji.

Dane reprezentowały dzienną sprzedaż każdego z produktów przez różnych sprzedawców.

1 Dane dostarczane przez różnych gestorów różnią się formatami, w tym nazwami kolumn. Stąd w praktyce istnieje potrzeba dokładnego ustalenia definicji zawartości.

Tabl. 6.4. Liczebności poszczególnych klas COICOP dla zbioru GestorAlfa 2018/19 ($N = 637339$, $k = 26$)

Klasy COICOP	Etykieta	Liczebność
CP011123	Kasze i ziarna zbóż	92 181
CP01111	Ryż	81 018
CP011462	Napoje i inne produkty mleczne	79 991
CP01211	Kawa	69 449
CP01144	Jogurt	64 245
CP011121	Mąka pszenna	59 411
CP01199	Inne art. żywnościowe, gdzie indziej niesklas.	37 781
CP01163	Owoce suszone i orzechy	37 411
CP01118	Pozostałe produkty zbożowe	28 583
CP011122	Pozostałe mąki	17 924
CP011461	Śmietana	17 493
CP01186	Sztuczne substytuty cukru	11 667
CP011452	Twarogi	9963
CP01142	Mleko świeże niskotłuszczowe	5238
CP01184	Wyroby cukiernicze	5183
CP011921	Sól	4537
CP011732	Pozostałe przetwory warzywne i grzybowe	3209
CP01213	Kakao i czekolada w proszku	2473
CP01116	Makarony i produkty makaronowe	2310
CP01114	Pozostałe wyroby piekarskie	2277
CP01181	Cukier	2187
CP01141	Mleko pełne świeże	1650
CP01182	Dżemy, marmolady i miód	721
CP01117	Płatki śniadaniowe	267
CP01143	Mleko zagęszczone i w proszku	167
CP05612	Pozostałe nietrwałe art. gospodarstwa domowego	3

Źródło: opracowanie własne.

6.2. Metodyka eksperymentów

6.2.1. Unikatowość kodów i etykiet produktów

Z punktu widzenia omawianych zagadnień ważnym elementem jest zbadanie cechy oznaczonej jako „id". Jest ona kodem producenta nadawanym danemu produktowi.

Badanie cechy „id" przeprowadzono dla danych GestorAlfa2018/9 (por. pkt 6.1.2).

Tabl. 6.5. Liczebności poszczególnych klas COICOP dla zbioru GestorDelta 2018 (N = 2 682 354, k = 45)

Klasa COICOP	Etykieta	Liczebność
CP01151	Masło	297458
CP011613	Jabłka	248283
CP01122	Mięso wieprzowe	239736
CP011123	Kasze i ziarna zbóż	233825
CP01111	Ryż	227333
CP01193	Żywność dla dzieci	131381
CP011741	Ziemniaki	111462
CP011271	Wędliny z wyjątkiem drobiowych	109784
CP011121	Mąka pszenna	101209
CP01147	Jaja	99971
CP011718	Cebula	97254
CP011242	Pozostały drób	89034
CP01126	Podroby i przetwory podrobowe	85801
CP011716	Marchew	61320
CP011241	Kury, koguty, kurczęta	60911
CP011211	Mięso wołowe	46530
CP011281	Mięso mielone mieszane	46196
CP01116	Makarony i produkty makaronowe	38524
CP01222	Napoje bezalk., gdzie indziej niesklasyf.	36946
CP01114	Pozostałe wyroby piekarskie	32273
CP121321	Środki kosmetyczne i higieniczne	26167
CP011122	Pozostałe mąki	24411
CP011717	Buraki	24379
CP01176	Pozostałe warzywa bulwiaste i przetwory z w.bul.	24373
CP01118	Pozostałe produkty zbożowe	24298
CP012232	Soki warzywne i owocowo-warzywne	22647
CP011462	Napoje i inne produkty mleczne	22017
CP011922	Przyprawy korzenne i zioła kulinarne	19996
CP01184	Wyroby cukiernicze	18218
CP01175	Chipsy	16485
CP012231	Soki owocowe	14488
CP01183	Czekolada	13730
CP09342	Artykuły dla zwierząt domowych	9019
CP01113	Pieczywo	8004
CP011212	Mięso cielęce	7126
CP011272	Wędliny drobiowe	2936
CP01117	Płatki śniadaniowe	2858
CP01194	Gotowe dania	2608
CP05612	Pozostałe nietrwałe art. gosp. domowego	1285
CP01191	Sosy, przyprawy	1107
CP011719	Pozostałe warzywa i grzyby	521
CP09521	Gazety	238
CP01155	Pozostałe tłuszcze zwierzęce	201
CP01185	Lody	9
CP011732	Pozostałe przetwory warzywne i grzybowe	2

Źródło: opracowanie własne.

Zmienna „id" pojawia się w dwóch miejscach: w danych sprzedażowych (por. pkt 6.1.2) oraz w dodatkowym zbiorze danych (zawartym w pliku grant_hicp_gtin.csv), w którym znajdują się zależności pomiędzy zmienną id a zmienną „gtin". Zbadanie

tych zależności jest interesujące z punktu widzenia tego, w jaki sposób kod producenta dla danego produktu można powiązać z kodem „gtin'.

Metoda badania polegała na obliczeniu statystyk współwystępowania takich samych kodów producenta dla różnych GTIN (ang. Global Trade Item Number) oraz różnych kodów producenta dla takich samych GTIN.

6.2.2. Wykorzystane metody klasyfikacji

W projekcie wykorzystano cztery typy klasyfikatorów: naiwny klasyfikator bayesowski, regresję logistyczną, lasy losowe oraz maszyny wektorów nośnych (SVM/SVC), będące częścią modułów IC.Classifier i IC.Matcher. W eksperymentach wykorzystano ogólnodostępny moduł Scikit-Learn[2], który zawiera napisane w języku Python implementacje interesujących nas rodzajów algorytmów: (i) regresja logistyczna (por. pkt 5.3.2), (ii) naiwny klasyfikator bayesowski (por. pkt 5.3.3), (iii) lasy losowe (por. pkt 5.3.5), (iv) maszyny wektorów nośnych: wersje liniowa i z jądrem radialnym (por. pkt 5.3.4).

6.2.3. Eksperyment klasyfikacji produktów na krótkookresowym zestawie danych

W toku prac eksperymentalnych zbadano opisane wyżej metody klasyfikacji z modułu IC.Classifier przeznaczone do przypisywania produktów do klas wyznaczanych przez kody COICOP, wykorzystując dane opisane w pkt 6.1.1.

Atrybuty (cechy, zmienne opisujące) dla poszczególnych produktów konstruowane były na podstawie zmiennej „nazwa produktu", która stanowi jego tekstowy opis (np. „Cukier z prawdziwą wanilią 10 g").

Uwzględniono różne opcje: pojedyncze słowa, *n*-gramy słów, *n*-gramy liter. Dla reprezentacji za pomocą pojedynczych słów stosowano ich ważenie częstościami występowania lub wyliczanymi dla nich wagami tf-idf (Manning i in., 2008). Dla pozostałych reprezentacji stosowano ważenie ich częstością. Należy odnotować, że sama zmienna „Nr Produktu" nie była używana w eksperymentach. Klasyfikacja przeprowadzana była dla każdej z sieci handlowych osobno: GestorAlfa ($N = 3298$), GestorBeta ($N = 954$), GestorGamma ($N = 1293$) oraz w zbiorze powstałym z ich połączenia: 3 sieci ($N = 5545$). Do klasyfikacji wykorzystano algorytmy wymienione w pkt 6.2.2.

Do podziału danych na zbiór uczący i testowy zastosowano metodę kroswalidacji warstwowej z parametrem K = 5 (Hastie i in., 2001, pkt 7.10; Koronacki i Ćwik, 2005, pkt 2.3). Losowy podział wykonano w taki sposób, aby w tworzonych podzbiorach zachować proporcje z klas z oryginalnego zbioru.

2 https://scikit-learn.org/stable/ – na tej stronie dostępne są stosowne opisy wraz z kodami źródłowymi.

Do oceny wyników klasyfikacji użyto miary dokładności (Koronacki i Ćwik, 2005, s. 102), czyli odsetka obserwacji prawidłowo zaklasyfikowanych. Dla uzyskanych w ten sposób wartości miar dokładności po wszystkich przebiegach kroswalidacji obliczono ich średnią oraz odchylenie standardowe.

6.2.4. Eksperyment klasyfikacji produktów na długookresowym zestawie danych

Metodologia przeprowadzonych dalej eksperymentów była bardzo podobna do omówionej w pkt 6.2.3. W szczególności użyto tych samych metod, które zostały tam omówione. Wszystkie eksperymenty klasyfikacji wykonane były na danych GestorAlfa 2018/9 (por. pkt 6.1.2).

Tak jak poprzednio, zbadano różne metody klasyfikacji pod nadzorem, które służyły do przypisywania produktów do klas wyznaczanych przez kody COICOP. Do klasyfikacji użyto tylko tekstowego opisu produktu – zmienna ta nazywała się „opis_towaru" w badanym zbiorze danych. Eksperyment został przeprowadzony w dwóch wariantach: pierwszy to klasyfikacja z użyciem wszystkich 26 klas wyznaczanych przez kod COICOP. W drugim wariancie wyszczególniono dziewięć klas, które odpowiadały grupom: ryż, mąka pszenna, pozostałe mąki, mleko pełne świeże, mleko świeże niskotłuszczowe, mleko zagęszczone i w proszku, jogurt, cukier, kawa, a z pozostałych klas stworzono jedną wspólną klasę. W efekcie zadanie sprowadza się do klasyfikacji do 10 klas.

Badano różne opcje: pojedyncze słowa, *n*-gramy słów, *n*-gramy liter, stosując reprezentację jak w podrozdz.6.2.3.

Ewaluację przeprowadzono jak w podrozdz. 6.2.3.

6.2.5. Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP

Ogólnym celem podjętych badań było stworzenie metody, która – bazując na danych uczących – w automatyczny sposób tworzyłaby zbiory słów pozytywnych i negatywnych. Przez *słowa pozytywne* rozumiemy słowa, których obecność w opisie produktu zwiększa przynależność do danej klasy COICOP; *słowa negatywne* zdefiniowane są analogicznie. Słowa te służyłyby do klasyfikacji produktów za pomocą reguł decyzyjnych, np. mleko_kozie<-select_labels(d,include=c(„kozie",”Kozie”),exclude=c(„proszku”)).

Badania prowadzono na zbiorze opisanym w pkt 6.1.2.

6.2.5.1. Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP z wykorzystaniem naiwnego klasyfikatora bayesowskiego

Pierwsze z zaproponowanych podejść bazuje na koncepcji wyliczania prawdopodobieństw słów powiązanych z danymi klasami, na której opiera się naiwny klasyfikator bayesowski (Kibriya i in., 2004, pkt 5.3.3). Metodę tę wybrano z uwagi na naturalną jej własność odwoływania się do częstości słów w tekstach i prostą probabilistyczną interpretację, a przy tym lepszą skuteczność generowanego klasyfikatora w porównaniu z regresją logistyczną, lasami losowymi czy maszynami wektorów nośnych.

W zaproponowanej metodzie identyfikacji słów pozytywnych i negatywnych wykonywane są następujące kroki:

- na danych trenujących następuje wytrenowanie naiwnego klasyfikatora bayesowskiego. Przy trenowaniu jako atrybuty dla klasyfikatora użyte były tylko słowa pochodzące z etykiety opisującej dany produkt (w szczególności nie używano żadnej dodatkowej wiedzy, która pomogłaby zidentyfikować, które ze słów są pozytywne czy negatywne);
- dla każdej klasy COICOP (na podstawie prawdopodobieństw słów uzyskanych z modelu Bayesa) wybierany jest zestaw słów pozytywnych i negatywnych z odpowiednim progowaniem wyników uzyskanych w trakcie trenowania.

Wyzwaniem w tym podejściu jest znalezienie punktu odcięcia dla prawdopodobieństw przynależnych słowom w wytrenowanym modelu. W tym celu przeanalizowano rozkłady prawdopodobieństwa słów z poszczególnych klas COICOP uzyskanych z modelu, co szczegółowo opisano są w ust. 6.3.3.3.1[3].

6.2.6. Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP na bazie podejścia teoriomnogościowego

Drugie z badanych podejść opiera się na podejściu teoriomnogościowym, kreowanym w sposób autorski. W opisanym poniżej podejściu tworzone są różne zbiory słów, które służą do konstrukcji klasyfikatorów. Właściwy klasyfikator bazuje na sprawdzeniu, w jaki sposób tekst opisujący dany produkt przynależy do tych zbiorów.

Tworzenie zbioru słów pozytywnych polega na:

- (P1) obliczeniu częstości występowania słów w opisach produktów danej klasy i wybraniu tych, które występują w (prawie) każdym opisie danej klasy;
- (P2) użyciu słów niewystępujących w żadnym produkcie innej klasy.

3 Dla pozostałych metod identyfikacja słów pozytywnych i negatywnych nie jest tak oczywista. W wypadku lasów losowych konieczna jest identyfikacja częstości występowania w sensie pozytywnym i negatywnym poszczególnych słów w budowanych drzewach lasu. W wypadku regresji logistycznej o przynależności do grupy słów pozytywnych decydują występowanie w równaniu regresji, znak współczynnika oraz technika radzenia sobie z klasami niebinarnymi. Podobnie jest w wypadku maszyn wektorów nośnych.

Tworzenie zbioru słów negatywnych polega na wybraniu słów niewystępujących w żadnym opisie produktu danej klasy.

6.2.7. Klasyfikator K1 na bazie zidentyfikowanych słów pozytywnych i negatywnych związanych z klasami COICOP

Zbadano jeden z zaproponowanych klasyfikatorów w grupie opartej o teoriomnogościową identyfikację słów kluczowych (nazwany K1) . Korzysta on tylko ze zbioru P1 (por. podrozdz. 6.2.6)), a działa w następujący sposób:

- bazując na danych uczących osobno dla każdej klasy, tworzony jest zbiór słów pozytywnych w oparciu o regułę P1. Przy czym zapamiętywany jest cały wektor częstości występowania słów w danej klasie ($p1$) obliczany na podstawie słów w opisach produktów tej klasy; klasyfikacja produktu polega na przedstawieniu opisu tego produktu jako wektora słów $w1$. Następnie obliczany jest iloczyn skalarny tego wektora z wektorem $p1$;
- klasa, która uzyska największą wartość produktu ($p1 \cdot w1$), jest wskazywana jako wynik klasyfikacji.

6.2.8. Śledzenie produktów

Powodem badań nad śledzeniem produktu między GTIN jest znany problem występujący przy tworzeniu indeksu cen, a mianowicie kwestia „wystarczającej" jednorodności produktów dla celów porównania ich cen. GTIN w danym momencie i w danym sklepie reprezentuje jednorodne produkty. Ale wiele firm wielokrotnie wprowadza ten sam produkt z innym GTIN-em z powodu zmian drugorzędnych (np. koloru opakowania). Przypisanie nowego GTIN może też wynikać z komponowania nowej linii produktów i wpisania produktu do niej, nowego opakowania czy częściowej zmiany składu produktu bez istotnej zmiany własności z punktu widzenia konsumenta. Zwykle związane jest to z przypisaniem nowej ceny. Rozmiar problemu zależy od branży i może być na tyle poważny, że indeksy cen oparte na samym GTIN-ie mogą spaść do zera w ciągu roku. Dlatego kluczowe jest przypisywanie produktów do szerszych, bardziej stabilnych kategorii (Feenstra i Shapiro, 2000).

Przeprowadzono dwa rodzaje eksperymentów:

- klasyfikację produktów miesiąc po miesiącu, którego celem było stwierdzenie, czy na podstawie klasyfikatora COICOP wytrenowanego na danych z jednego miesiąca można poprawnie klasyfikować produkty do klas COICP w kolejnym miesiącu;
- dopasowanie produktów miesiąc po miesiącu, którego celem było stwierdzenie, ile rzeczywistych nowych produktów można się spodziewać w kolejnych miesiącach.

6.2.8.1. Klasyfikacja produktów miesiąc po miesiącu

Przeprowadzono eksperyment klasyfikacji produktów miesiąc po miesiącu za pomocą modułu IC.Classifier (pkt 5.4.1). Wykorzystano dane opisane w pkt 6.2.4.

Jako danych do klasyfikacji użyto tylko tych rekordów, dla których zostały przypisane etykiety COICOP. Niepuste wartości COICOP występowały dla miesięcy o etykietach: m_0 = „12/2017" do m_{10} = „10/2018" (n = 11).

Za dane uczące brano wszystkie rekordy z danego miesiąca (począwszy od miesiąca o etykiecie m_0), a za dane testowe – rekordy z kolejnego miesiąca. Procedurę taką wykonano dla kolejnych par miesięcy: (m_0,m_1), (m_1,m_2), ... (m_9,m_{10}).

Eksperyment został przeprowadzony w dwóch konfiguracjach:

- klasyfikacja z użyciem wszystkich 26 klas wyznaczanych przez kod COICOP;
- klasyfikacja polegająca na wyszczególnieniu dziewięciu klas (klasy te odpowiadały grupom: ryż, mąka pszenna, pozostałe mąki, mleko pełne świeże, mleko świeże niskotłuszczowe, mleko zagęszczone i w proszku, jogurt, „cukier, kawa), a z pozostałych klas stworzono jedną wspólną klasę. W efekcie zadanie to polegało na klasyfikacji do 10 klas.

Do klasyfikacji użyto następujących metod: (i) regresja logistyczna (por. pkt 5.3.2), (ii) naiwny klasyfikator bayesowski (por. pkt 5.3.3), (iii) lasy losowe (por. pkt 5.3.5), (iv) maszyny wektorów nośnych: wersje liniowa (model LinearSVC pakietu scikit-learn) i z jądrem radialnym (modele SVM/SVC/RBF, por. pkt 5.3.4).

6.2.8.2. Dopasowanie produktów miesiąc po miesiącu

Przeprowadzono również eksperyment dopasowania produktów miesiąc po miesiącu za pomocą modułu IC.Matcher (por. pkt 5.4.2). Wykorzystano dane opisane w pkt 6.1.3. Danymi wejściowymi do modułu były dane GestorKappa 2020/21.

6.2.9. Wsparcie informatyczne przy wyborze reprezentantów dla kategorii COICOP

Ważnym problemem przy ocenie wskaźników inflacji jest dobór reprezentantów produktów i usług (koszyka) odpowiadających określonym kategoriom COICOP. W ramach projektu InstatCeny zaproponowano metodę polegającą na pomiarze podobieństwa opisu produktu przynależnego do danej kategorii COICOP do nazwy tejże kategorii (z poziomu 6). Na podstawie miar podobieństwa rangowano produkty.

Podobieństwo liczono, używając podobieństwa cosinusowego i traktując opis każdego produktu jako wektor reprezentujących go słów[4]:

4 Reprezentacja ta zakłada, że mamy pewien słownik uporządkowany wszystkich słów z wszystkich opisów produktów, np. S = („biały", „cukier", „kryształ"). Wtedy opis produktu „cukier kryształ" będzie postaci [0,1,1], a „cukier biały" – postaci [1,1,0].

$$\text{cosine similarity}\,(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}.$$

6.3. Wyniki przeprowadzonych eksperymentów

6.3.1. Unikatowość kodów i etykiet produktów

Metodykę eksperymentu opisano w pkt 6.2.1.

Tablice 6.6 i 6.7 przedstawiają proste statystyki opisowe relacji między zmiennymi „id" (kod produktu) oraz „gtin" (plik grant_hicp_gtin.csv), który odnosi się do danych GestorAlfa2018/9. W zbiorze tym jest $N_{id} = 9733$ unikalnych identyfikatorów „id" i $N_{gtin} = 10333$ unikalnych kodów „gtin".

Tablica 6.6 pokazuje, ile kodów „gtin" odpowiada pojedynczemu kodowi „id".

Tabl. 6.6. Zestawienie częstości dla relacji: liczba kodów kodów „gtin" odpowiadająca pojedynczemu kodowi „id"

Liczba kodów „gtin" odpowiadająca pojedynczemu „id"	Liczba przypadków
1	1353
2	5800
3	1026
4	916
5	273
6	181
7	59
8	39
9	27
10	30
11	6
12	11
13	2
14	4
15	3
16	1
19	2

Źródło: opracowanie własne.

Jak widać, w zaledwie 14% przypadków produkt ma przypisany jeden jedyny kod GTIN. W prawie 60% przypadków dwa różne kody GTIN oznaczają ten sam produkt.

Tablica 6.7 dotyczy relacji odwrotnej: liczby kodów „id" odpowiadających pojedynczemu „gtin".

Tabl. 6.7. Zestawienie częstości dla relacji: liczba kodów kodów „id" odpowiadająca pojedynczemu kodowi „gtin"

Liczba kodów „id" odpowiadająca pojedynczemu kodowi „gtin"	Liczba przypadków
1	3620
2	3258
3	1779
4	818
5	412
6	266
7	94
8	57
9	17
10	7
11	5

Źródło: opracowanie własne.

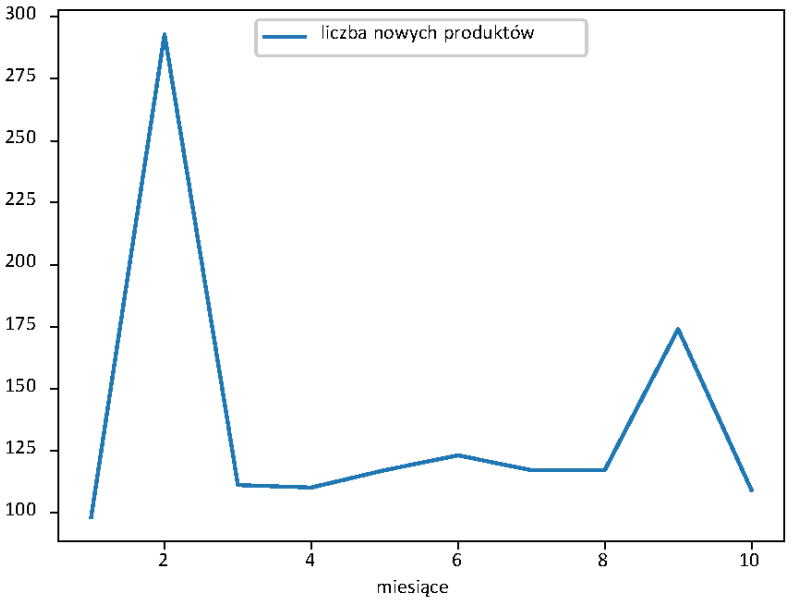
Okazuje się, że w mniej niż 35% przypadków kod GTIN pozwala na identyfikację kodu produktu. Oznacza to, że w badaniu statystyk cen GTIN nie spełnia swego zadania jako identyfikator produktu, którego śledzenie może być wykorzystane do śledzenia produktów w koszyku inflacyjnym.

Dodatkowo zbadano też relacje pomiędzy cechą „id" oraz cechą „opis towaru". Okazało się, że te dwie cechy są w relacji 1–1. Oznacza to, że jeden opis towaru może być powiązany z wieloma kodami GTIN, a jeden kod GTIN – z wieloma opisami towarów. Z drugiej strony oznacza to, że gestorzy starannie opisują oferowane produkty. Jest to pozytywna przesłanka do uznania opisu produktu jako dobrej podstawy ich klasyfikacji.

6.3.2. Skala rotacji produktów

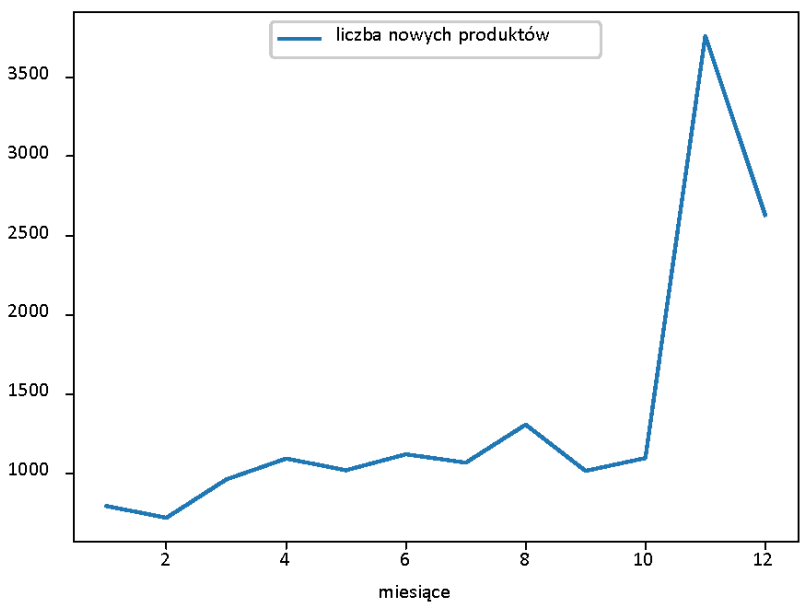
Zmienność portfela produktów badano dla danych GestorAlfa 2018/9 (por. pkt 6.1.2) oraz danych GestorDelta. Dla kolejnych miesięcy wyznaczono zestaw unikalnych zbiorów produktów, które występowały w danym miesiącu. Bazowano przy tym na liczbowych identyfikatorach produktów. Dysponując tak obliczonymi zbiorami, sprawdzano, ile nowych produktów pojawia się przy przechodzeniu z danego miesiąca do następnego. Rezultaty zaprezentowano na poniższych wykresach: na wykr. 6.1 – liczbę nowych produktów pojawiających się w każdym miesiącu dla danych GestorAlfa, a na wykr. 6.2 – liczbę nowych produktów dla GestorDelta pojawiających się w kolejnych miesiącach. Wykresy pokazują, że zwykle pojawia się co najmniej 5% produktów z nowymi identyfikatorami w każdym miesiącu, a liczba nowych produktów w tym sensie może czasem sięgnąć 40% sprzedaży.

Wykr. 6.1. Liczba nowych towarów określanych na podstawie cechy „id" w kolejnych miesiącach (m_1 = „01/2018" do m_{10} = „10/2018")



Uwaga. Liczba produktów w miesiącu m_0 wynosi 2016.
Źródło: opracowanie własne.

Wykr. 6.2. Liczba nowych towarów określanych na podstawie cechy „id" w kolejnych miesiącach (m_1 = „10/2017" do m_{12} = „10/2018")



Uwaga. Liczba produktów w miesiącu m_0 wynosi 8635.
Źródło: opracowanie własne.

6.3.3. Jakość zaimplementowanych metod klasyfikacyjnych

Badania jakości klasyfikatorów prowadzono na zbiorach danych opisach w pkt 6.1.1 i 6.1.2.

6.3.3.1. Eksperyment klasyfikacji do COICOP na krótkookresowym zestawie danych

Wyniki eksperymentów przeprowadzonych w sposób omówiony w pkt 6.2.3 na danych opisanych w pkt 6.1.1 znajdują się w tabl. 6.8 (dla danych z sieci GestorBeta i GestorGamma) i tabl. 6.9 (dla sieci GestorAlfa oraz wszystkich dostępnych danych). W tabelach zamieszczono wartości średnie i odchylenia standardowe miar dokładności uzyskanych za pomocą kroswalidacji.

Okazuje się, że dane najłatwiejsze do analizy dostarczył GestorGamma, a najtrudniejsze – GestorAlfa.

Metoda klasyfikacji lasów losowych najwyraźniej radzi sobie najgorzej we wszystkich przypadkach. Klasyfikator liniowy maszyn wektorów nośnych radzi sobie najlepiej.

Tabl. 6.8. Średnie i odchylenia standardowe miar dokładności różnych metod klasyfikacji uzyskanych dla 5-krotnej kroswalidacji na zbiorach GestorBeta i GestorGamma

GestorBeta			GestorGamma		
coicop lvl: 3	coicop lvl: 4	coicop lvl: 5	coicop lvl: 3	coicop lvl: 4	coicop lvl: 5
Klasyfikator LinearSVC					
0,933(0,049)	0,909(0,046)	0,917(0,026)	0,988(0,001)	0,930(0,007)	0,932(0,006)
Klasyfikator MultinomialNB					
0,936(0,063)	0,887(0,042)	0,892(0,029)	0,987(0,003)	0,928(0,012)	0,921(0,015)
Klasyfikator Regresja logistyczna					
0,921(0,063)	0,880(0,058)	0,887(0,027)	0,980(0,005)	0,907(0,009)	0,902(0,007)
Klasyfikator Random Forest					
0,911(0,065)	0,869(0,045)	0,883(0,039)	0,977(0,005)	0,928(0,011)	0,917(0,004)

Źródło: opracowanie własne.

Tabl. 6.9. Średnie i odchylenia standardowe miar dokładności różnych metod klasyfikacji uzyskanych dla 5-krotnej kroswalidacji na zbiorach GestorAlfa i 3 sieci

GestorAlfa			3 sieci		
coicop lvl: 3	coicop lvl: 4	coicop lvl: 5	coicop lvl: 3	coicop lvl: 4	coicop lvl: 5
Klasyfikator LinearSVC					
0,898(0,059)	0,892(0,024)	0,896(0,023)	0,925(0,055)	0,912(0,017)	0,909(0,014)
Klasyfikator MultinomialNB					
0,883(0,054)	0,840(0,021)	0,844(0,026)	0,908(0,052)	0,870(0,017)	0,865(0,013)
Klasyfikator Regresja logistyczna					
0,850(0,068)	0,842(0,022)	0,845(0,028)	0,886(0,062)	0,876(0,013)	0,866(0,021)
Klasyfikator Random Forest					
0,832(0,071)	0,814(0,028)	0,819(0,036)	0,854(0,074)	0,852(0,030)	0,844(0,031)

Źródło: opracowanie własne.

Dodatkowo w tabl. 6.10 i 6.11 umieszczono prawdopodobieństwa słów dla ustalonych klas COICOP (z poziomu 5) obliczone dla naiwnego klasyfikatora bayesowskiego. Do obliczenia tych prawdopodobieństw użyto danych pochodzących ze zbioru 3 sieci, przy czym zaprezentowane tylko te klasy, które posiadały co najmniej 50 obserwacji.

Tabl. 6.10. Prawdopodobieństwa słów dla ustalonych klas COICOP (z poziomu 5) obliczone dla naiwnego klasyfikatora bayesowskiego dla danych pochodzących ze zbioru 3 sieci poszczególnych klas COICOP uzyskanych z modelu

CP011111	Prawd.	CP011121	Prawd.	CP011123	Prawd.	CP011181	Prawd.
ryż	0,0321	mąka	0,0321	kasza	0,0231	ciasto	0,0081
4x100g	0,0150	1kg	0,0296	gryczana	0,0072	oetker	0,0067
ryz	0,0117	pszenna	0,0088	jęczmienna	0,0070	babeczki	0,0057
500g	0,0072	tortowa	0,0070	4x100g	0,0057	dr	0,0057
biały	0,0061	maka	0,0070	500g	0,0045	delecta	0,0057
kupiec	0,0061	450	0,0070	400g	0,0035	ziemniaczana	0,0033
1kg	0,0054	500	0,0067	1kg	0,0035	dolce	0,0031
basmati	0,0052	typ	0,0049	manna	0,0030	mąka	0,0031
parboiled	0,0047	pzz	0,0047	cenos	0,0030	na	0,0031
sonko	0,0040	luksusowa	0,0047	melvit	0,0027	stilla	0,0029

Tabl. 6.10. Prawdopodobieństwa słów dla ustalonych klas COICOP (z poziomu 5) obliczone dla naiwnego klasyfikatora bayesowskiego dla danych pochodzących ze zbioru 3 sieci poszczególnych klas COICOP uzyskanych z modelu (dok.)

CP011411	Prawd.	CP011421	Prawd.	CP011431	Prawd.	CP011441	Prawd.
mleko	0,0256	mleko	0,0388	mleko	0,0145	jogurt	0,0903
1l	0,0183	1l	0,0269	zag	0,0052	150g	0,0244
uht	0,0073	uht	0,0115	gostyń	0,0047	naturalny	0,0132
5l	0,0040	folia	0,0059	light	0,0036	activia	0,0131
pet	0,0040	pet	0,0056	słodz	0,0026	400g	0,0127
świeże	0,0033	butelka	0,0044	150g	0,0026	jogobella	0,0116
folia	0,0023	5l	0,0042	350g	0,0026	del	0,0115
karton	0,0020	świeże	0,0040	gos	0,0026	dan	0,0108
łowicz	0,0020	9l	0,0028	tuba	0,0023	pitny	0,0089
butelka	0,0020	łaciate	0,0028	tł	0,0023	bakoma	0,0089

Uwaga. Niech $P(x_i|y$ oznacza dla zadanego słowa x_i przy ustalonej klasie y prawdopodobieństwo występowania słowa x_i w opisach produktów należących do klasy y. Dla każdej klasy y wyznaczono wielkość przesunięcia $shift_y = \frac{1}{2}(\max_i \left(\ln\left(P(x_i|y)\right)\right) - \min_i \left(\ln\left(P(x_i|y)\right)\right)$, gdzie ln to logarytm naturalny. Każdej parze (słowo, klasa) (x_i, y) przypisano wskaźnik pozytywności $WP(x_i, y) = \ln\left(P(x_i|y)\right) - shift_y$.

Tabl. 6.11. Prawdopodobieństwa słów dla ustalonych klas COICOP (z poziomu 5) obliczone dla naiwnego klasyfikatora bayesowskiego dla danych pochodzących ze zbioru 3 sieci

CP011452	Prawd.	CP011461	Prawd.	CP011462	Prawd.	CP011631	Prawd.
twaróg	0,0143	śmietana	0,0371	serek	0,0224	bakalland	0,0130
1kg	0,0058	18	0,0234	kefir	0,0140	200g	0,0103
250g	0,0058	12	0,0199	del	0,0125	100g	0,0085
200g	0,0048	śmietanka	0,0173	mleko	0,0106	suszone	0,0066
półtłusty	0,0043	200g	0,0147	maślanka	0,0104	500g	0,0064
ser	0,0040	kubek	0,0133	danio	0,0103	orzechy	0,0064
serek	0,0038	400g	0,0111	dan	0,0097	farmers	0,0059
naturalny	0,0033	30	0,0107	400g	0,0094	snack	0,0055
sernikowy	0,0030	uht	0,0070	dele	0,0086	rodzynki	0,0050
delikatny	0,0025	180g	0,0042	wanilia	0,0086	słonecznik	0,0039
CP011841	Prawd.	CP011991	Prawd.	CP012111	Prawd.	CP012131	Prawd.
dr	0,0081	oetker	0,0106	kawa	0,1094	kakao	0,0070
oetker	0,0078	do	0,0102	250g	0,0281	25g	0,0036
100g	0,0047	dr	0,0102	100g	0,0146	czekolada	0,0036
baton	0,0031	cukier	0,0071	mielona	0,0141	napój	0,0029
polewa	0,0031	delecta	0,0050	500g	0,0139	nesquik	0,0023
30g	0,0023	aromat	0,0041	gold	0,0108	kakaowy	0,0023
lukier	0,0021	masa	0,0037	200g	0,0101	picia	0,0023
mleczny	0,0018	proszek	0,0035	tchibo	0,0084	puchatek	0,0021
kinder	0,0018	krem	0,0035	espresso	0,0080	kruger	0,0021
gellwe	0,0016	gellwe	0,0035	jacobs	0,0075	do	0,0021

Źródło: opracowanie własne.

Tabele te ilustrują użyteczność naiwnego klasyfikatora bayesowskiego dla kojarzenia klas COICOP ze słowami z tekstu opisu produktu. Mimo że metoda maszyn

wektorów nośnych daje lepsze wyniki klasyfikacyjne, to z punktu widzenia opisowego, zrozumiałego dla człowieka, metoda naiwnego klasyfikatora bayesowskiego sprawdza się najlepiej. Jak widzimy, wiodące słowa przypisane do poszczególnych klas COICOP (czyli te, którym klasyfikator bayesowski przypisuje najwyższe prawdopodobieństwa) odzwierciedlają istotę klasy, np. w przypadku klasy 01.1.4 (mleko, ser, jaja) najczęstszym słowem jest „mleko", a w przypadku klasy 1 (chleb i zboża) – „mąka", „kasza" i „ciasto".

6.3.3.2. Eksperyment klasyfikacji do COICOP na długookresowym zestawie danych

Wyniki eksperymentów klasyfikacji przeprowadzonych w sposób przedstawiony w pkt 6.2.4 na danych opisanych w pkt 6.1.2 znajdują się w tabl. 6.12. Tym razem widać, że wszystkie metody klasyfikacji radzą sobie bardzo dobrze, zarówno przy podziale na 26 klas, jak i na 9+1, i to dla gestora, który wypadł najgorzej dla mniejszych danych. Tę poprawę wyjaśnia zwiększenie ilości rekordów w zbiorze danych.

Tabl. 6.12. Średnie i odchylenia standardowe miar dokładności różnych metod klasyfikacji dla zbioru GestorAlfa w zależności od liczby docelowych kategorii COICOP, uzyskanych w 5-krotnej kroswalidacji

GestorAlfa	
26 klas	9 klas vs reszta
Klasyfikator LinearSVC	
0,996(0,003)	0,996(0,003)
Klasyfikator MultinomialNB	
0,994(0,004)	0,992(0,007)
Klasyfikator Regresja logistyczna	
0,996(0,003)	0,996(0,003)
Klasyfikator Random Forest	
0,996(0,003)	0,996(0,003)

Źródło: opracowanie własne.

W tabl. 6.13, 6.14 i 6.15 przedstawione zostały prawdopodobieństwa najczęstszych słów dla wszystkich klas COICOP obliczonych dla naiwnego klasyfikatora bayesowskiego z wykorzystaniem danych pochodzących ze zbioru GestorAlfa2018/19.

Tabl. 6.13. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorAlfa2018/19

CP01111	Prawd.	CP011121	Prawd.	CP011122	Prawd.	CP011123	Prawd.
ryż	0,2127	mąka	0,1790	mąka	0,1970	kasza	0,1881
4x100g	0,0730	1kg	0,1321	melvit	0,1316	4x100g	0,1055
risana	0,0620	basia	0,0655	1kg	0,1164	gryczana	0,0612
1kg	0,0511	pszenna	0,0506	żytnia	0,0499	cenos	0,0507
cenos	0,0450	450	0,0430	owsiana	0,0384	jęczmienna	0,0480
parboiled	0,0428	typ	0,0370	500g	0,0360	risana	0,0466
biały	0,0418	tortowa	0,0345	kukurydziana	0,0280	melvit	0,0360
klc	0,0408	lubella	0,0272	kbio	0,0267	klc	0,0345
długoziarnisty	0,0357	kg	0,0255	ryżowa	0,0256	jaglana	0,0293
sawi	0,0327	500	0,0221	wyp	0,0253	prażona	0,0257
CP01114	Prawd.	CP01116	Prawd.	CP01117	Prawd.	CP01118	Prawd.
wafelki	0,2542	kuskus	0,2224	orkiszowe	0,1210	ciasto	0,1144
schaer	0,2542	300g	0,2220	płatki	0,1210	delecta	0,1077
kakaowe125g	0,1444	klc	0,1999	kbio	0,1210	mąka	0,0515
cytrynowe	0,1099	kasza	0,1999	500g	0,1210	380g	0,0318
125g	0,1099	janex	0,0222	dzikim	0,0005	1kg	0,0306
żółty	0,0001	karton	0,0222	drożdże	0,0005	oetker	0,0299
dziki	0,0001	alfabet	0,0005	duet	0,0005	schaer	0,0289
duet	0,0001	razowy	0,0005	dunkel	0,0005	chleb	0,0235
dunkel	0,0001	eko	0,0005	duża	0,0005	ziemniaczana	0,0227
duża	0,0001	400g	0,0005	dyni40g	0,0005	melvit	0,0224
CP01141	Prawd.	CP01142	Prawd.	CP01143	Prawd.	CP01144	Prawd.
mleko	0,2047	mleko	0,2162	mleko	0,0926	jogurt	0,1460
1l	0,1157	uht	0,1127	proszku	0,0920	bakoma	0,0456
pet	0,0673	5l	0,1000	4kg	0,0920	naturalny	0,0389
uht	0,0521	łowickie	0,1000	krasnystaw	0,0920	bakomamen	0,0339
piątnica	0,0482	1l	0,0890	niesł7	0,0011	140g	0,0285
ekologiczne	0,0482	pet	0,0609	puszka411g	0,0011	bio	0,0284
kbio	0,0439	krasnystaw	0,0248	zagęszcz	0,0011	200g	0,0252
tł	0,0397	tł	0,0222	dele	0,0011	pitny	0,0228
głubczyce	0,0322	butelka	0,0195	duet	0,0006	230g	0,0203
mlekpol	0,0234	10	0,0136	dunkel	0,0006	wysokobiał	0,0203

Źródło: opracowanie własne.

Tabl. 6.14. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorAlfa2018/19

CP011452	Prawd.	CP011461	Prawd.	CP011462	Prawd.	CP01163	Prawd.
serek	0,0891	śmietana	0,1817	müller	0,0581	100g	0,1048
danio	0,0780	18	0,0983	ryż	0,0568	poex	0,0989
135g	0,0780	400g	0,0952	na	0,0559	bio	0,0707
twaróg	0,0587	piątnica	0,0868	mleku	0,0551	bakalland	0,0517
200g	0,0583	12	0,0778	200g	0,0456	selection	0,0475
naturalny	0,0503	200g	0,0463	wanilia	0,0265	fresco	0,0335
delikatny	0,0462	30	0,0416	truskawka	0,0241	orzechy	0,0334
brzoskwinia	0,0462	śmietanka	0,0394	kefir	0,0213	suszone	0,0267
gruszka	0,0462	krasnystaw	0,0326	mix	0,0201	mieszanka	0,0255
szarlotka	0,0318	kubek	0,0295	maślanka	0,0189	żurawina	0,0254

Tabl. 6.14. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorAlfa2018/19 (dok.)

CP011732	Prawd.	CP01181	Prawd.	CP01182	Prawd.	CP01184	Prawd.
fasola	0,1506	1000g	0,1811	zott	0,1792	baton	0,1371
jaś	0,1251	trzcinowy	0,1811	galaretka	0,1792	mleczny	0,1369
piękny	0,1251	cukier	0,1811	pomarańczowa	0,1792	milino	0,1369
cenos	0,1126	drobny	0,1811	175g	0,1792	iwt	0,1369
tyczny	0,1126	goldpack	0,1811	delgellwe	0,0002	4x28g	0,0692
370g	0,1126	dżungli	0,0001	droetker	0,0002	miód	0,0692
400g	0,0503	dziki	0,0001	dżungli	0,0002	kakao	0,0678
biała	0,0255	drożdże	0,0001	dżemofix	0,0002	4x30g	0,0678
drobna	0,0255	duet	0,0001	długoziarnisty	0,0002	350g	0,0235
ludwiczyn	0,0254	dunkel	0,0001	długoziarnist	0,0002	torciki	0,0235
CP01186	Prawd.	CP011921	Prawd.	CP01199	Prawd.	CP01211	Prawd.
listek	0,1664	sól	0,1903	dr	0,1057	kawa	0,1339
zielony	0,1664	kotanyi	0,1903	oetker	0,1057	100g	0,0384
stevia	0,1087	wieliczka	0,0952	mix	0,0442	inka	0,0348
słodzik	0,0762	1kg	0,0952	delecta	0,0317	espresso	0,031
250g	0,0749	kopalni	0,0952	40g	0,0295	aroma	0,0307
ksylitol	0,0749	morska	0,0951	dekoracji	0,0244	cafissimo	0,0303
250	0,0384	gruboziarnista	0,0951	decorada	0,0242	black	0,0297
tab	0,0384	500g	0,0951	czekoladowe	0,0239	tchibo	0,0292
efferta	0,0384	delel	0,0002	30ml	0,0225	zbożowa	0,029
150	0,0377	dietetyczna	0,0002	cytrynowy	0,0215	ziarnista	0,0287

Źródło: opracowanie własne.

Tabl. 6.15. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorAlfa2018/19

CP01213	Prawd.	CP05612	Prawd.
kakao	0,1022	kawy	0,0034
200g	0,1013	jn	0,0034
wawel	0,1013	filtry	0,0034
do	0,0996	rozm	0,0034
25g	0,0996	pudełko	0,0026
picia	0,0996	80szt	0,0026
czekolada	0,0996	folia	0,0017
krüger	0,0996	40szt	0,0017
bananowa	0,0996	delr	0,0009
ciemne	0,0008	dziki	0,0009

Źródło: opracowanie własne.

Prawdopodobieństwa obliczone w analogiczny sposób dla zbioru GestorDelta2018 znajdują się w tabl. 6.16 i 6.17.

Tabl. 6.16. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorDelta2018

CP01111	Prawd.	CP011121	Prawd.	CP011122	Prawd.	CP011123	Prawd.	CP01113	Prawd.
ryż	0,17	mąka	0,15	mąka	0,17	kasza	0,16	chleb	0,29
supreme	0,12	1kg	0,11	500g	0,08	plony	0,14	500g	0,24
4x100g	0,08	bie	0,1	jaglana	0.08	natury	0,14	pasterski	0,21
mix	0,06	pola	0,1	ryżowa	0,08	4x100g	0,07	żytni	0,05
biały	0,05	złote	0,1	bezglute-nowa	0,08	bie	0,07	krojony	0,04
CP01114	Prawd.	CP01116	Prawd.	CP01117	Prawd.	CP01118	Prawd.	CP011211	Prawd.
wafle	0,17	plony	0,12	mix	0,14	popcorn	0,17	mięso	0,25
ryż	0,14	300g	0,12	ryż	0,14	100g	0,17	gulasz	0,13
100g	0,12	kuskus	0,12	kuk	0,14	micropop	0,17	500g	0,13
musli	0,12	kasza	0,12	bezglut	0,14	sól	0,17	wołowy	0,13
arroza	0,12	natury	0,12	vitanella	0,14	bie	0,17	wołowe	0,12
CP011212	Prawd.	CP01122	Prawd.	CP011241	Prawd.	CP011242	Prawd.	CP01126	Prawd.
mięso	0,20	mięso	0,19	mięso	0,2	mięso	0,18	kaszanka	0,16
cielęce	0,20	kg	0,11	kurczaka	0,2	indyka	0,18	600g	0,09
kg	0,14	wp	0,06	uda	0,08	na	0,13	wędlin	0,08
kością	0,11	500g	0,04	500g	0,08	500g	0,08	kraina	0,08
km	0,07	wieprzowe	0,04	pulpety	0,06	kotlety	0,05	mix	0,07

CP011271	Prawd.	CP011272	Prawd.	CP011281	Prawd.	CP011462	Prawd.	CP01147	Prawd.
szynka	0,21	konserwowa	0,24	mięso	0,18	ryżowy	0,24	jaja	0,26
konserwowa	0,17	szynka	0,24	500g	0,18	1l	0,24	szt	0,13
kraina	0,05	drobiowa	0,24	miel	0,1	vitanella	0,24	kl	0,13
wędlin	0,05	600g	0,24	wp	0,1	napój	0,24	ale	0,12
200g	0,05	indyka	0	łopatki	0,1	kokosowo	0,05	wybieg	0,09
CP01151	Prawd.	CP01155	Prawd.	CP011613	Prawd.	CP011716	Prawd.	CP011717	Prawd.
masło	0,20	mięso	0,13	jabłko	0,25	marchew	0,34	luz	0,33
200g	0,13	wieprzowa	0,13	luz	0,20	luz	0,17	czerwone	0,33
mleczna	0,10	kg	0,13	polskie	0,15	opak	0,13	buraki	0,33
dolina	0,09	słonina	0,13	opak	0,05	kg	0,11	frostino	0
ekstra	0,09	kawałek	0,06	gala	0,04	młoda	0,06	gazeta	0

Źródło: opracowanie własne.

Tabl. 6.17. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorDelta2018

CP011718	Prawd.	CP011719	Prawd.	CP011732	Prawd.	CP011741	Prawd.	CP01175	Prawd.
cebula	0,36	luz	0,24	zupa	0,01	ziemniaki	0,30	chipsy	0,16
kg	0,20	masłowa	0,24	express	0,01	opak	0,14	wiejskie	0,11
opak	0,14	fasolka	0,24	620g	0,01	luz	0,08	masło	0,11
mix	0,09	gaz	0	marchew	0,01	2kg	0,06	twaroż	0,11
siatka	0,09	garmażeryj-ne	0	danie	0,01	3kg	0,06	ziemn	0,11

Tabl. 6.17. Prawdopodobieństwa słów dla wszystkich klas COICOP obliczone dla naiwnego klasyfikatora bayesowskiego dla danych ze zbioru GestorDelta2018 (dok.))

CP01176	Prawd.	CP01183	Prawd.	CP01184	Prawd.	CP01185	Prawd.	CP01191	Prawd.
ziemniaki	0,25	luz	0,18	cukierki	0,2	zbóż	0,01	ocet	0,15
bataty	0,25	jaja	0,13	luz	0,2	ml	0,01	ryżowy	0,15
luz	0,25	czekolado-we	0,1	mix	0,14	czek	0,01	of	0,15
słodkie	0,25	duello	0,1	toffi	0,08	lody	0,01	house	0,15
frutti250	0	mix	0,1	dobromiej-skie	0,08	ryż	0,01	asia	0,15
CP011922	**Prawd.**	**CP01193**	**Prawd.**	**CP01194**	**Prawd.**	**CP01222**	**Prawd.**	**CP012231**	**Prawd.**
grill	0,12	kaszka	0,17	kasza	0,19	5l	0,18	sok	0,20
20g	0,12	ryżowa	0,17	warz	0,19	jabłko	0,15	jabłko	0,12
mix	0,12	mix	0,17	mięsem	0,19	polaris	0,14	dnia	0,08
kamis	0,12	mleczno	0,13	820	0,19	gaz	0,11	5l	0,08
kiełb	0,06	nestle	0,11	gołąbki	0,19	napój	0,11	cymes	0,08
CP012232	**Prawd.**	**CP05612**	**Prawd.**	**CP09342**	**Prawd.**	**CP09521**	**Prawd.**	**CP121321**	**Prawd.**
sok	0,19	dancoal	0,19	puffi	0,20	regionalny	0,13	ciała	0,16
marchew	0,14	brykiet	0,19	marchew	0,20	pt	0,13	masło	0,16
wycisk	0,14	5kg	0,19	karma	0,20	gazeta	0,13	mix	0,11
świeżo	0,14	węgla	0,14	1240g	0,20	dodatek	0,13	do	0,08
25l	0,10	drzewnego	0,14	dziczyzna	0,20	wyborcza	0,13	200ml	0,08

Źródło: opracowanie własne.

Manualna analiza wskazuje, że słowa kluczowe uznane przez te metody klasyfikacyjne za miarodajne dla produktów należących do poszczególnych kategorii COICOP wydają się sensowne. Przykładowo dla kategorii CP011123 „Kasze i ziarna zbóż" słowa przypisane na podstawie opisów produktów to: „kasza", „gryczana", „cenos", „jęczmienna", „risana", „melvit" itp.

6.3.3.3. Klasyfikacja na bazie słów pozytywnych i negatywnych

Przeprowadzone eksperymenty bazowały przede wszystkim na danych pochodzących z sieci GestorAlfa, opisanych w pkt 6.1.2.

W eksperymentach ograniczono się do następujących dziewięciu grup elementarnych: ryż (011111), mąka pszenna (011121), pozostałe mąki (011122), mleko pełne świeże (011411), mleko świeże niskotłuszczowe (011421), mleko zagęszczone i w proszku (011431), jogurt (011441), cukier (011811), kawa (012111).

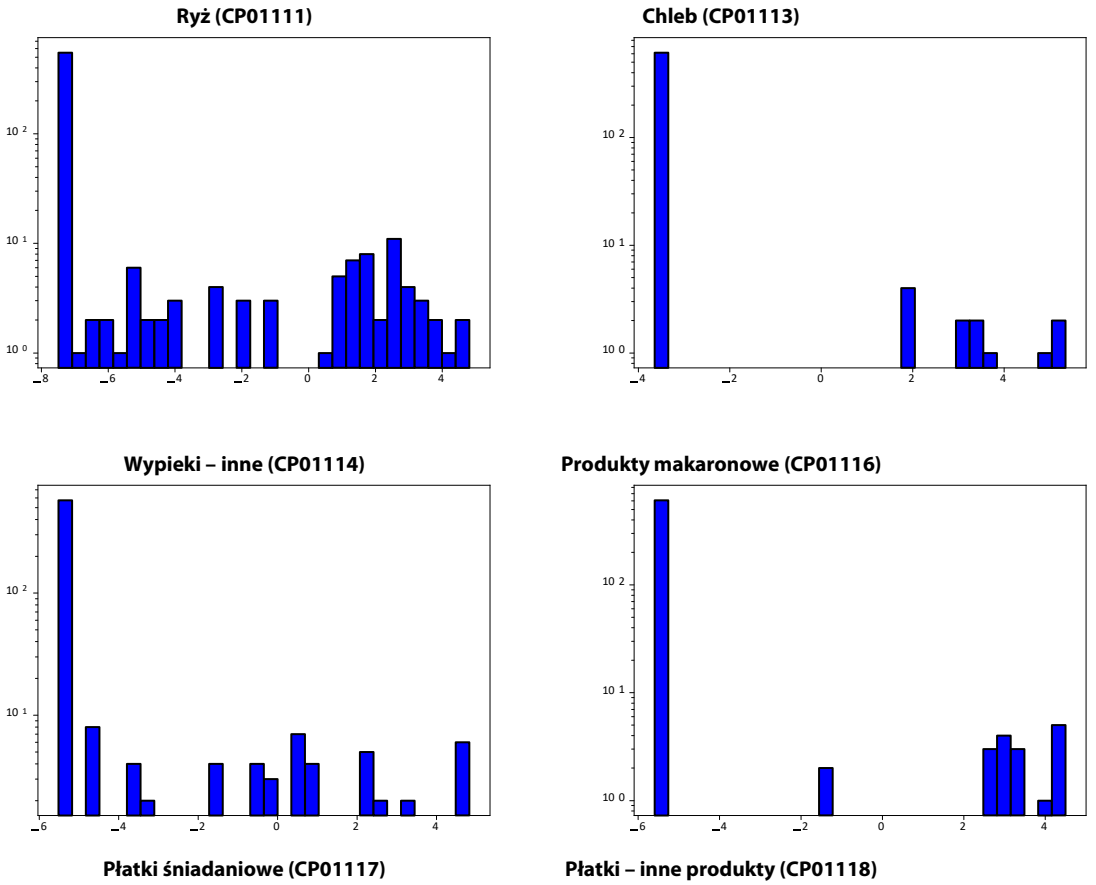
6.3.3.3.1. Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP na bazie wielomianowego naiwnego klasyfikatora bayesowskiego

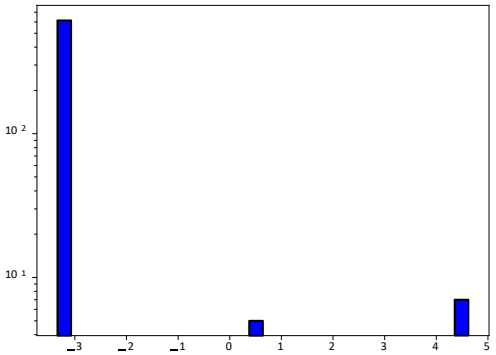
Wyzwaniem w podejściu opisanym w ppkt 6.2.5.1 jest znalezienie punktu odcięcia dla prawdopodobieństw przynależnych słowom w wytrenowanym modelu. W tym celu przeanalizowano warunkowe rozkłady prawdopodobieństwa słów w każdej klasie.

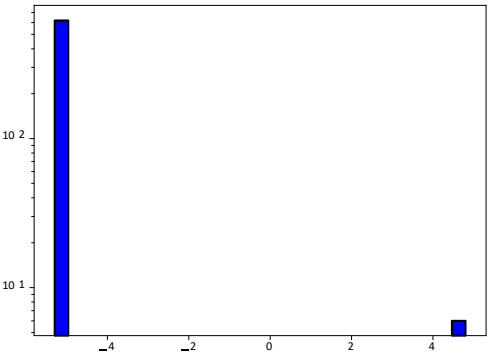
Na wykr. 6.3 i 6.4 przedstawiono histogramy wskaźników pozytywności słów w obrębie pojedynczej klasy. Histogramy stworzono dla danych odpowiednio z sieci GestorDelta oraz GestorAlfa.

Za słowa *pozytywne* dla danej klasy uznaje się te słowa, dla których wskaźniki pozytywności są dodatnie. Słowa, dla których wskaźniki pozytywności mają wartości ujemne, ustala się jako *negatywne*.

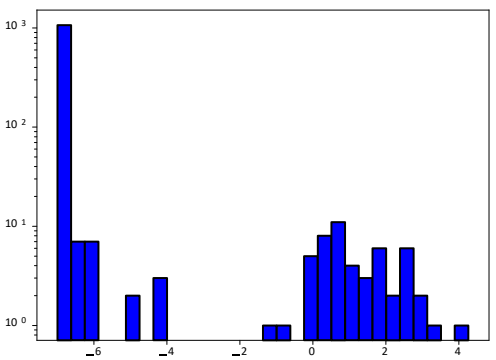
Wykr. 6.3. Histogramy wskaźników pozytywności – GestorDelta



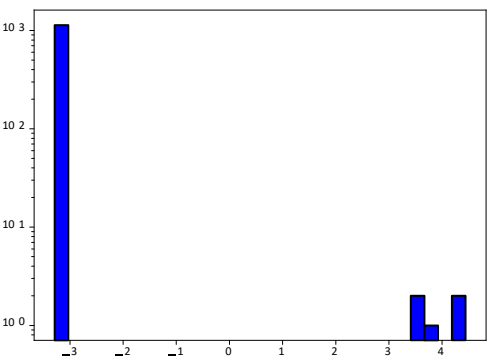




Ryż (CP01111) Wypieki – inne (CP01114)



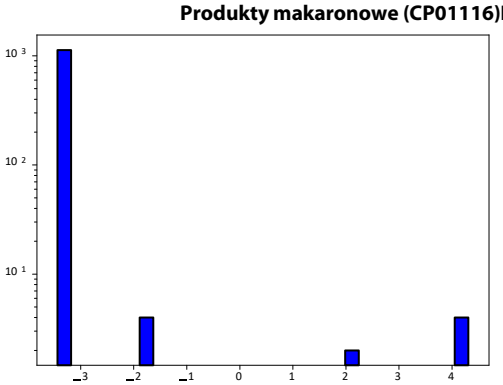
Ryż (CP01111) Wypieki – inne (CP01114)

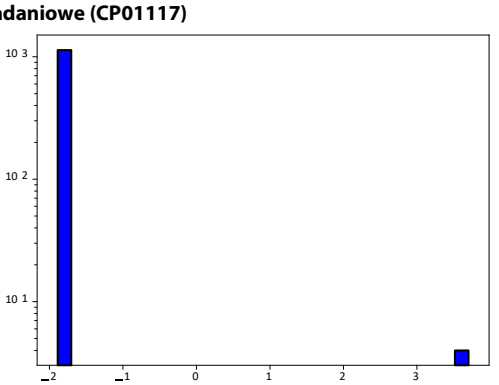


Źródło: opracowanie własne.

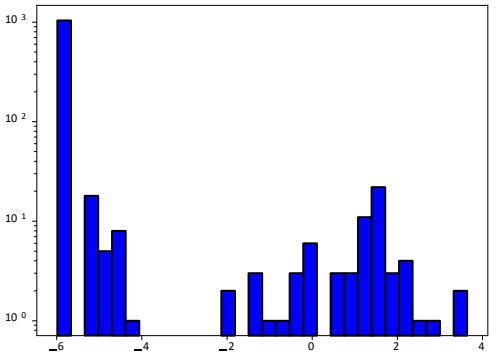
Wykr. 6.4. Histogramy wskaźników pozytywności – GestorAlfa

Produkty makaronowe (CP01116)Płatki śniadaniowe (CP01117)





Płatki – inne produkty (CP01118)



Źródło: opracowanie własne.

6.3.3.3.2. Identyfikacja słów pozytywnych i negatywnych związanych z klasami COICOP na bazie podejścia teoriomnogościowego

W tabl. 6.18 znajdują się przykłady słów pozytywnych (P1) dla wybranych klas, uzyskanych zgodnie z metodologią opisaną w pkt 6.2.6.

Tabl. 6.18. Częstość występowania słów w opisach danej klasy (GestorAlfa) – używane do wyboru słów pozytywnych metoda P1

CP01111	freq	CP01114	freq	CP01116	freq	CP01117	freq
Ryż		Wypieki – inne		Makaron, kuskus		Płatki śniadaniowe	
ryż	0,959	wafelki	1,000	kuskus	1,000	orkiszowe	1,000
4x100g	0,329	schaer	1,000	300g	0,998	płatki	1,000
risana	0,280	kakaowe125g	0,568	klc	0,899	500g	1,000
1kg	0,231	125g	0,432	kasza	0,899	kbio	1,000
.cenos	0,203	cytrynowe	0,432	janex	0,100	.	.
parboiled	0,193	.	.	karton	0,100	.	.
biały	0,189	.	.	alfabet	0,002	.	.
klc	0,184	.	.	400g	0,002	.	.
długoziarnisty	0,161	.	.	eko	0,002	.	.
sawi	0,148	.	.	razowy	0,002	.	.

Źródło: opracowanie własne.

Generowanie zbiorów słów za pomocą metody P2 pozwala stworzyć zbiory pozytywne tylko dla niektórych klas. Poniżej wymienione są wszystkie klasy, dla których wygenerowano niepuste zbiory słów pozytywnych:

- CP1181: drobny, cukier, goldpack, trzcinowy;
- CP1142: hula, międzybórz, śmietana, mlektar, bychawa, wieluń, jarocin, homog, myszków, 10, łask, krasulamlekopaster, 406237, kar, jędrzejów, łowickie, grudziądz, osm, spoż, spożywcze, głubczyckie, osowa, litr;
- CP1143: zagęszcz, niesł7, puszka411g, 4kg, proszku;
- CP11122: 997, owsiana, żytnia, żytniat, schaer, dom, proso, gryczana, it, vitacorn, 720, 2000, provena, ryżowa, uniwersaln, owies, kukurydzia, jaglana;

- CP11142: hula, międzybórz, śmietana, mlektar, bychawa, wieluń, jarocin, homog, myszków, 10, łask, krasulamlekopaster, 406237, kar, jędrzejów, łowickie, grudziądz, osm, spoż, spożywcze, głubczyckie, osowa, litr;
- CP1111: hom, kuchnia, sys, 4001561, pełnoziarn, 5x100g, biała, mal, parabolicz, ze, 2x100, 2x150g, amerykańsk, długoziarn, jaśminowy, dzikim, torebka, gosposi, czerwony, długi, risotto, dziki, delebiały, arroz, carnaroli, tajlandii, parboiled, kupiec, ryż, cenos, bomba, arborio, basmati, szpinakiem, kolory, 4x100, sawi, brązowy, 8x100g, 4x100g, okrągły, riso, łam, gallo, długoz, długoziarn, sushi1kg, risana, biały, gołąbków, okrągłozia, quinoa, czarny, okrągłozia, janex, sushi;
- CP11121: pełnoziarnista, 3600061, 450, wrocław, xxlt405, włoską, razowa, t1850, pizza, przemiału, 1k, pszenica1000g, orkiszowa, 9kg, t480, krupczatka, 480 psz, szczepanki, białystok, werbkowice, na, tortowa1kg, wypiek, t405, 1850 delella, poznańska, t450, wrocławska, pizzę, gdańskiemłynymąkawrocławskat500, młynpol, pszen, zamojska, delea, t550, poznań, peł, młyny, mąk, lubelska, tortowych, orkisz, 405 0 miłosna, puszysta, szyman, mąkapszenna, orkiszu, luksusowa, szczecińska, włoska, koszalińska, reszelska, bianpol, 5kg, stanisz, tortowt, 750 650 delemąka, typ500, młyn, hetmańska, zamość, płońsk, szadek, pszenna, t500, królowa, 400 40203 extra, graham, kruszwica, tortowa, polskie, lubella, 550 luksusowa1kg, pzz, delr, 4x1kg, xxl, basia;
- CP1211: class, cafe, sasz, 110g, cafes, arabica, rozpu, costa, 25g, smak, segafredokawmiel, instant, white500g, 10x7g, wapń, pusz, vacum, 36tx4g, ulubiona, 100 dobra, kawa, caffe, palona, mielon, white2x250g, mielona, cafissimo, fine, błonnikiem, rozpuszcz, fix, karmelowa, 10x8, gratis, 10sztx7g, zbożowa, saszetka, classic, zbóż, rica, rozpuszczalna, frappe, 275g, ziarn, davidoff, mokasidamo, green, family, vac, maślanka, 57 guilis, strac, costarica, elegant, exquisit, 250 intenso, zboż, całe, azera, 100g, ziarno, 35tx4, sidamo, 10x8g, gala, zbozowa, szklanka, puszka, magne, rich, ziarnista, moka, americano, espresso, mocca, aroma, rozp, creme, bag, espres, inka, 10x16g, nescafeintensokawa, sido, mezcla, mk, coffee, sensazione, 3w1, mrągowska, 18g, ziarna, 2g, mokate, caramel, 5xlatte, witaminy, woseba, brasil, delikatny, anatol, india, cafea, astra, nescafe, 225g, ziar, ziarni500g, crema, marila, tchibo, especial, orkiszem, kaffee;
- P1144: gęsty, activia, leśne, bałkański, brzos, bażanowice, fantasia, grec, prudnik, froop, 98g, 1kg9, kokos, nap, pinacolada, środziemnomorski, ananas, świecie, pitny, jogutr, brzosk, 450g, zott, 122g, kubek, owcz, picia, 180g, brzoskwi, bakoma, mango, cytr, lekki, płatki, truskawkami, 220g, wiśn, 380g, wanil, koktajl, special, kesem, kub, lim, ow, ziarn280g, piatnica, but, laktozy, naturalna, lekki330g, brzo, ozorków, truskawka1l, pitny250g, jagodowy, pozi, lakt, bakomazbóż, grecki, standard, truskawka, bieluch, cuiavia, figand, ml, augustowski, bieluchjogurt, 8zbóż, gruszka, bio, 400gbut, czekoladowe2x105g, mueller, 140g, natur, anan, nadbużań-

ski, środziemnomor, ale, somlek, ca, truskawska, 120g, bez, dan, truskawk, 2x120g, napój, cyn, jovi, piątuś, 300g, naturalny350g, jogurt, krasnyst, delja, jeżyna, 2x122g, jog, 70g, brzoskwinia, grusz, saszetce, drażami, nat, joguś, jag, duet, aleowoc, jogobella, truskaw, deleg, nadbużański200g9, pit, typu, leśn, morela200g, trus, 290g, dele2, 135 mar, 7zbóżmen, malina, owoce, leś, maliny, czekoladowe, 7zbóż, arbuz, grecki150g, mśś, kri, morelami, 340g, 99g, maluta, wanilia70g, śliwka, wiśnia, wysokobiał, sącz, banan, 175g, iwt, pomarań, 310g, 350g, garwolin, greckiego, łowicz, nowy, 125 maguś, wiaderko, jab, 230g, brzoskw, marak, bakomamenjogurt, 330g, opole, bakomamen, jogobellabreakfasttropmusli125g, calpro, owoc, sokólski, prob, owocami, light, danonki, tł170g, naturalny150g, kulki.

6.3.3.3.3. Klasyfikator K1 na bazie zidentyfikowanych słów pozytywnych i negatywnych związanych z klasami COICOP

Wyniki uzyskane z zastosowaniem klasyfikatora K1 (opisanego w pkt 6.2.7) znajdują się w tabl. 6.19. Klasyfikator wykorzystujący wyłączne wybrane słowa pozytywne odznacza się bardzo dobrą skutecznością, porównywalną z metodami tradycyjnych klasyfikatorów.

Tabl. 6.19. Średnie i odchylenia standardowe miar dokładności uzyskanych dla klasyfikacji na bazie wyłącznie słów pozytywnych, ocenionych z zastosowaniem 10-krotnej kroswalidacji (*cv* = 10)

GestorAlfa	
Cechy do klasyfikacji: „opis towaru"	
Klasyfikator: K1	0,959(0,005)

Źródło: opracowanie własne.

6.3.4. Skuteczność metod dopasowania produktów

6.3.4.1. Skuteczność klasyfikacji miesiąc po miesiącu

Przeprowadzono badania śledzenia produktów – eksperyment klasyfikacji produktów miesiąc po miesiącu na danych opisanych w pkt 6.1.2 metodą przedstawioną w ppkt 6.2.8.1.

W tabl. 6.20 znajdują się wyniki klasyfikacji produktów bazowego eksperymentu, w którym dane jednego miesiąca służyły jako dane trenujące do przewidywania kategorii COICOP, a dane kolejnego – jako dane testujące. Obliczono średnią dokładność oraz odchylenie standardowe.

Wszystkie klasyfikatory radzą sobie bardzo dobrze z postawionym zadaniem.

Tabl. 6.20. Średnie i odchylenia standardowe miar dokładności uzyskanych dla klasyfikacji „miesiąc po miesiącu"

GestorAlfa	
26 klas	9 klas vs reszta
Cechy do klasyfikacji: „opis towaru"	
Klasyfikator LinearSVC	
0,9983(0,0013)	0,9983(0,0013)
Klasyfikator MultinomialNB	
0,9962(0,0015)	0,9952(0,0021)
Klasyfikator Regresja logistyczna	
0,9982(0,0013)	0,9985(0,0011)
Klasyfikator Random Forest	
0,9976(0,0027)	0,9977(0,0021)

Źródło: opracowanie własne.

W tabl. 6.21 znajdują się wyniki analogicznych eksperymentów, z tą różnicą że do obliczania miary dokładności używano tylko nowych przypadków pojawiających się w kolejnych miesiącach. Wyniki tego eksperymentu w rozbiciu na poszczególne miesiące znajdują się w tabl. 6.22 (klasyfikacja z użyciem zmiennej „opis towaru") oraz w tabl. 6.23 (klasyfikacja z użyciem zmiennych „opis towaru" i „hierarchia sieci").

Tabl. 6.21. Średnie i odchylenia standardowe miar dokładności uzyskanych dla klasyfikacji „miesiąc po miesiącu"

GestorAlfa		
Średnia	26 klas	9 klas vs reszta
Cechy do klasyfikacji: „opis towaru"		
Klasyfikator LinearSVC		
Po miesiącach	0,9130(0,1115)	0,9527(0,0537)
Ważona	0,9638(0,1869)	0,9660(0,1811)
Klasyfikator MultinomialNB		
Po miesiącach	0,8666(0,1251)	0,8903(0,1775)
Ważona	0,9464(0,2251)	0,9322(0,2515)
Klasyfikator regresja logistyczna		
Po miesiącach	0,9067(0,1146)	0,9686(0,0484)
Ważona	0,9633(0,1879)	0,9838(0,1262)
Klasyfikator Random Forest		
Po miesiącach	0,8026(0,2033)	0,8832(0,1926)
Ważona	0,9276(0,2591)	0,9318(0,2521)

Tabl. 6.21. Średnie i odchylenia standardowe miar dokładności uzyskanych dla klasyfikacji „miesiąc po miesiącu" (dok.)

GestorAlfa		
Średnia	26 klas	9 klas vs reszta
Cechy do klasyfikacji: „opis towaru"		
Klasyfikator LinearSVC		
Po miesiącach	0,9444 (0,0921)	0,9645 (0,0477)
Ważona	0,9437 (0,2304)	0,9626 (0,1896)
Klasyfikator MultinomialNB		
Po miesiącach	0,9129 (0,1389)	0,9365 (0,0990)
Ważona	0,9113 (0,2844)	0,9118 (0,2836)
Klasyfikator regresja logistyczna		
Po miesiącach	0,9286 (0,0932)	0,9643 (0,0476)
Ważona	0,9433 (0,2313)	0,9625 (0,1901)
Klasyfikator Random Forest		
Po miesiącach	0,8816 (0,1020)	0,9766 (0,0377)
Ważona	0,9257 (0,2622)	0,9801 (0,1395)

Uwaga. Klasyfikowano tylko produkty uznane na podstawie cechy „opis towaru" w danym miesiącu za nowe

Źródło: opracowanie własne.

Tabl. 6.22. Wartości średnie miar dokładności dla kolejnych miesięcy uzyskanych dla klasyfikacji „miesiąc po miesiącu"

	1	2	3	4	5	6	7	8	9	10
LinearSVC	0,997	1,000	0,842	0,998	0,875	1,000	0,939	0,992	0,946	0,938
MultinomialNB	0,995	1,000	0,842	0,380	0,938	0,999	0,879	0,987	0,946	0,938
Log. Regression	0,997	0,999	0,999	0,998	0,875	1,000	0,879	0,992	0,946	1,000
Random Forest	0,995	0,999	0,842	0,680	0,875	0,999	0,970	0,992	0,946	1,000

Uwaga. Klasyfikowano tylko produkty uznane na podstawie cechy „opis towaru" w danym miesiącu za nowe.

Źródło: opracowanie własne.

Tabl. 6.23. Wartości średnie miar dokładności dla kolejnych miesięcy uzyskanych dla klasyfikacji „miesiąc po miesiącu"

	1	2	3	4	5	6	7	8	9	10
LinearSVC	0,998	0,989	0,843	1,000	1,000	0,999	0,939	0,992	0,946	0,938
MultinomialNB	0,998	0,994	0,686	1,000	1,000	0,999	0,939	0,990	0,821	0,938
Regr. logistyczna	0,997	0,989	0,843	1,000	1,000	0,999	0,939	0,992	0,946	0,938
Random Forest	0,997	0,988	1,000	0,998	1,000	0,999	0,970	0,992	0,946	0,875

Uwaga. Klasyfikowano tylko nowe produkty w danym miesiącu dla schematu „9 klas vs reszta" z użyciem cech: „opis towaru" i „hierarchia sieci" miesiąc liczba błędów opis towaru oryginalna nazwa coicop predykowana nazwa coicop

Źródło: opracowanie własne.

Zadanie klasyfikacji miesiąc po miesiącu jest trudniejsze i tu widać zróżnicowanie skuteczności klasyfikatorów. Najlepiej radzi sobie z zadaniem klasyfikator oparty na liniowych maszynach wektorów nośnych, najgorzej – lasy losowe. Jak można się było spodziewać, zadanie klasyfikacji do 26 kategorii jest trudniejsze od 9+1 kategorii. Dzięki wykorzystaniu informacji „hierarchia sieci" zadanie klasyfikacji było łatwiejsze.

W tabl. 6.24 wymieniono produkty, które zostały źle sklasyfikowane. Dla kolejnych miesięcy zaprezentowane są poszczególne przypadki (opis towaru oraz przynależny im kod COICOP) wraz z predykcją kodu COICOP uzyskanego w wyniku klasyfikacji naiwnym klasyfikatorem bayesowskim.

Tabl. 6.24. Przypadki nieprawidłowo zaklasyfikowanych produktów (naiwny klasyfikator bayesowski)

Miesiąc	Liczba błędów	Oryginalna nazwa COICOP	Predykowana nazwa COICOP
0	1	Pozostałe nietrwałe artykuły	Kasze i ziarna zbóż
0	1	Napoje i inne produkty mleczne	Twarogi
0	1	Pozostałe przetwory warzywne	Śmietana
0	1	Napoje i inne produkty mleczne	Twarogi
0	1	Napoje i inne produkty mleczne	Twarogi
1	1	Napoje i inne produkty mleczne	Jogurt
2	206	Napoje i inne produkty mleczne	Jogurt
2	205	Napoje i inne produkty mleczne	Jogurt
2	3	Wyroby cukiernicze	Inne artykuły żywnościowe
2	206	Napoje i inne produkty mleczne	Kasze i ziarna zbóż
3	1	Inne artykuły żywnościowe	Pozostałe produkty zbożowe
3	1	Kasze i ziarna zbóż	Makarony i produkty makaronow
5	1	Napoje i inne produkty mleczne	Twarogi
5	1	Napoje i inne produkty mleczne	Jogurt
6	1	Pozostałe nietrwałe artykuły	Kasze i ziarna zbóż
6	1	Kakao i czekolada w proszku	Wyroby cukiernicze
6	1	Ryż	Napoje i inne produkty mleczne
6	1	Napoje i inne produkty mleczne	Jogurt
7	1	Napoje i inne produkty mleczne	Jogurt
7	1	Pozostałe nietrwałe artykuły	Pozostałe przetwory warzywne
7	1	Wyroby cukiernicze	Kasze i ziarna zbóż
7	2	Kawa	Inne artykuły żywnościowe
7	2	Kasze i ziarna zbóż	Makarony i produkty makaronow
7	1	Kasze i ziarna zbóż	Makarony i produkty makaronow
8	204	Pozostałe produkty zbożowe	Pozostałe mąki
8	1	Inne artykuły żywnościowe	Pozostałe mąki
8	169	Pozostałe produkty zbożowe	Pozostałe mąki
8	1	Jogurt	Napoje i inne produkty mleczne
9	1	Napoje i inne produkty mleczne	Twarogi
9	1	Napoje i inne produkty mleczne	Jogurt

Uwaga. Opisy produktów zostały ukryte, aby zachować anonimowość gestorów.
Źródło: opracowanie własne.

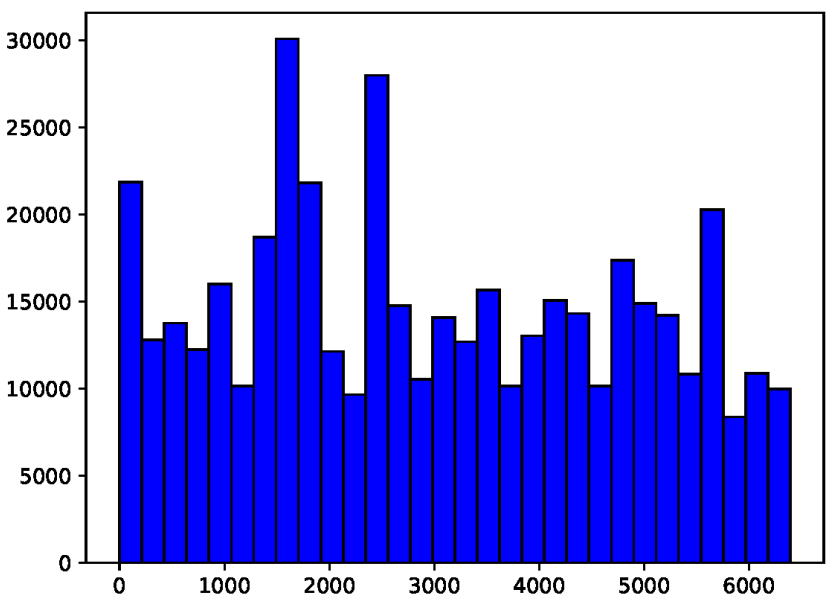
6.3.5. Eksperyment dopasowania

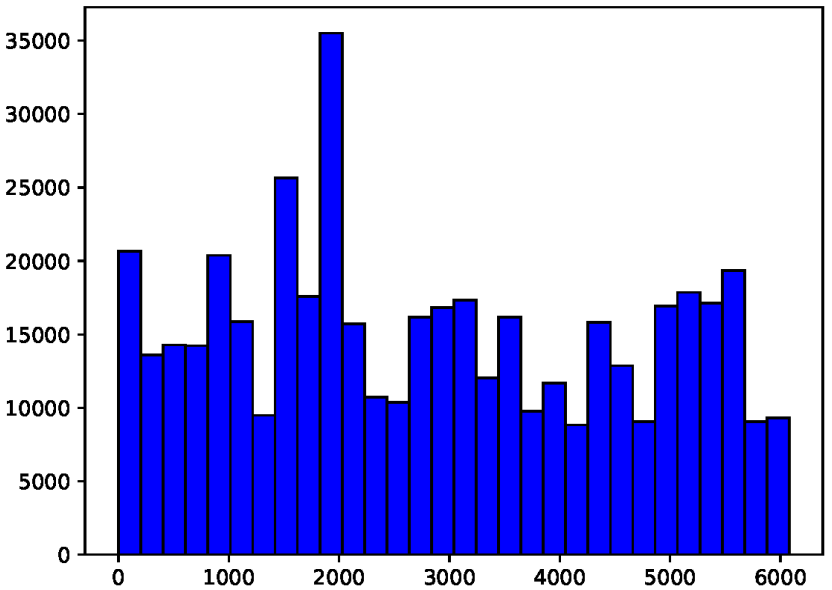
Przeprowadzono badania śledzenia produktów – eksperyment dopasowania produktów miesiąc po miesiącu na danych opisanych w pkt 6.1.3 metodą przedstawioną w ppkt 6.2.8.2.

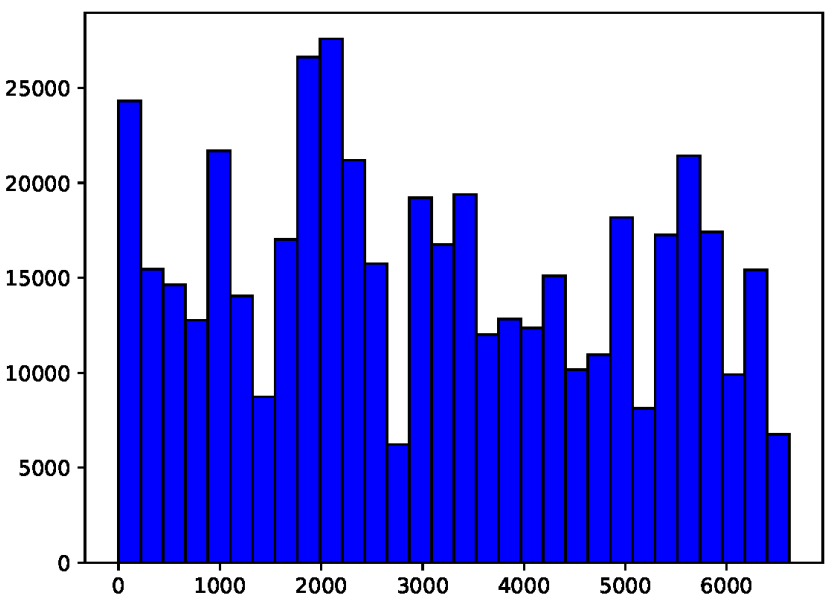
Na wykr. 6.5 i 6.6 znajdują się histogramy przedstawiające liczbę grup powstałych ze matchowania produktów występujących w kolejnych miesiącach dla danych GestorKappa. Użyto szybkiej metody matchowania z parametrem podobieństwa dla tekstów równym: *precision* = 0,95. Na każdym histogramie oś pozioma oznacza liczebność grupy tożsamych produktów (kategorie liczebności co 200 elementów grupy), a oś pionowa – liczbę grup w poszczególnych kategoriach liczebności.

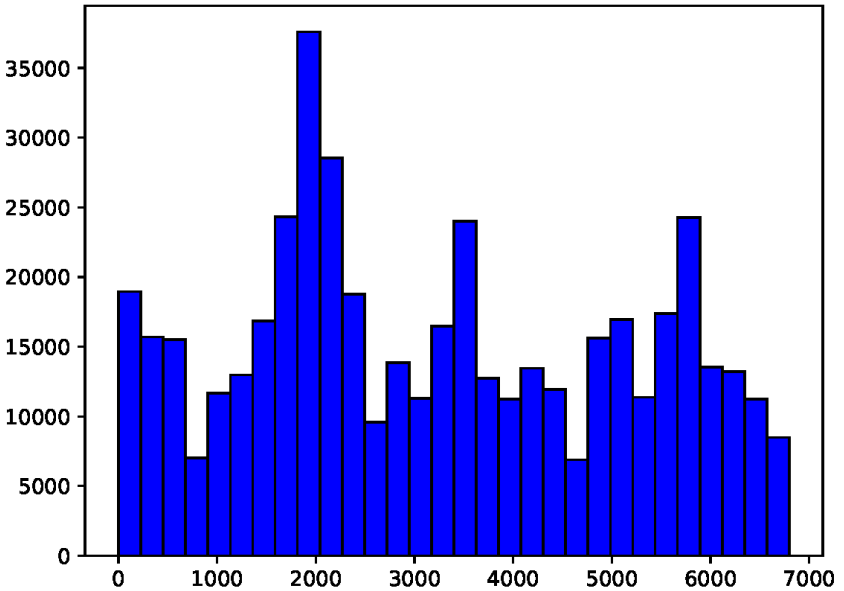
Liczebność grup utożsamianych z tym samym wyrobem waha się od 1 do ponad 6000. Liczba grup w każdej kategorii liczebności jest mniej więcej taka sama, z wyjątkiem pików o liczebności ok. 2000, a histogramy nie zmieniają się szczególnie z miesiąca na miesiąc.

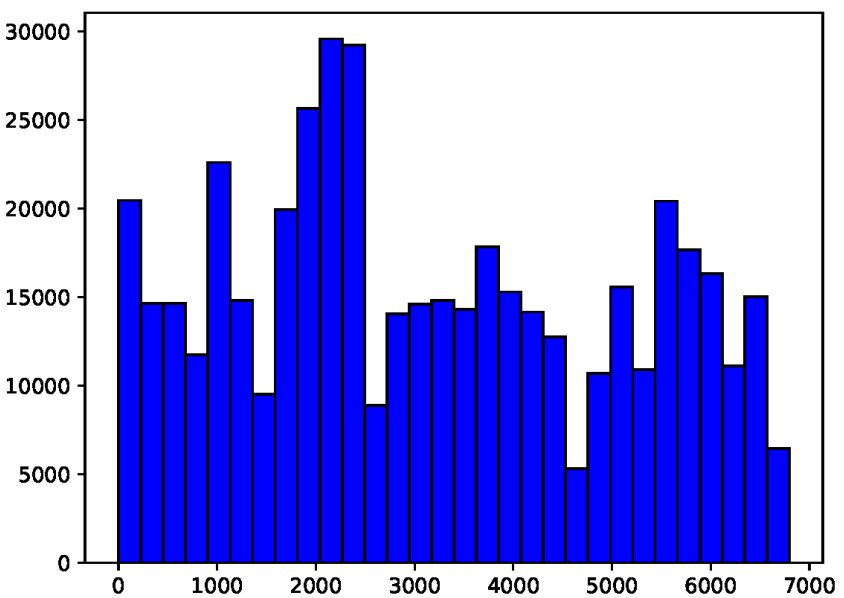
Wykr. 6.5. Miesięczne histogramy dopasowania produktów (GestorKappa)




35000
30000
25000
20000
15000
10000
5000
0
0
1000
2000
3000
4000
5000
6000

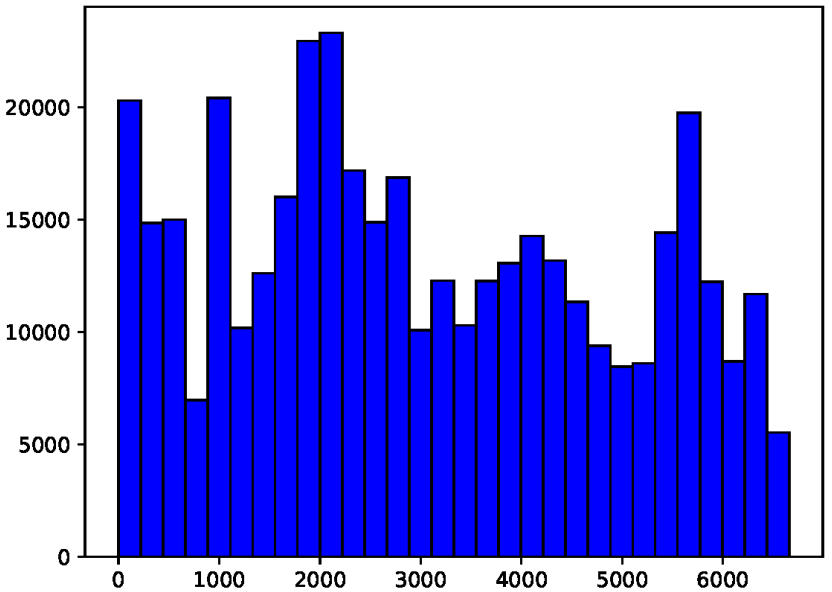

25000
20000
15000
10000
5000
0
0
1000
2000
3000
4000
5000
6000

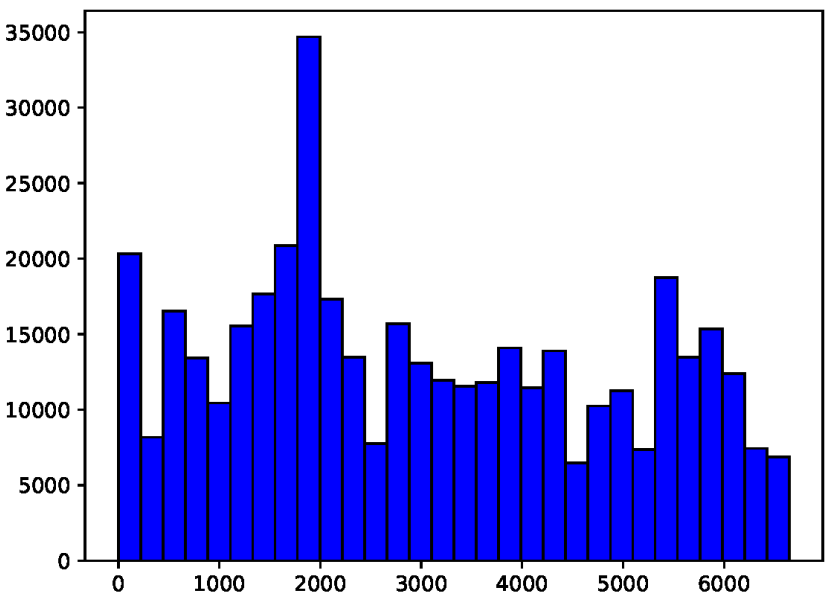


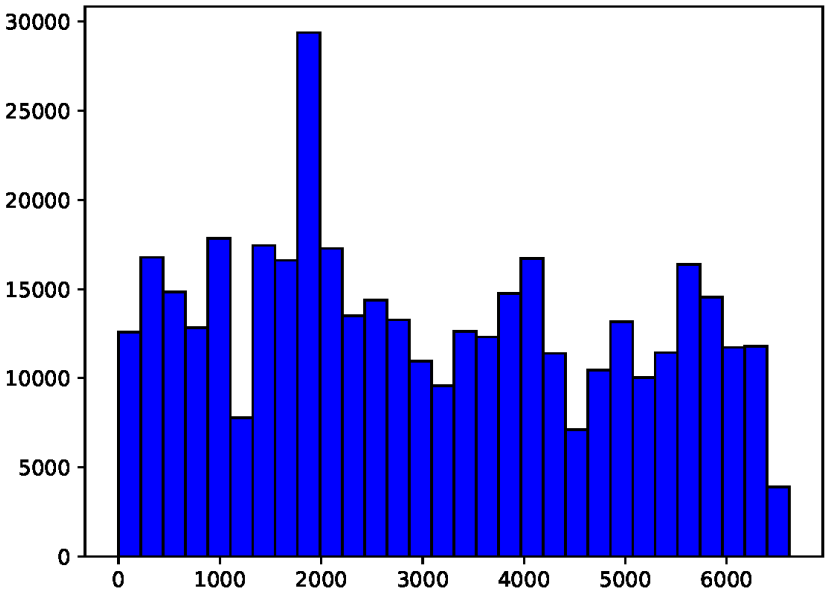


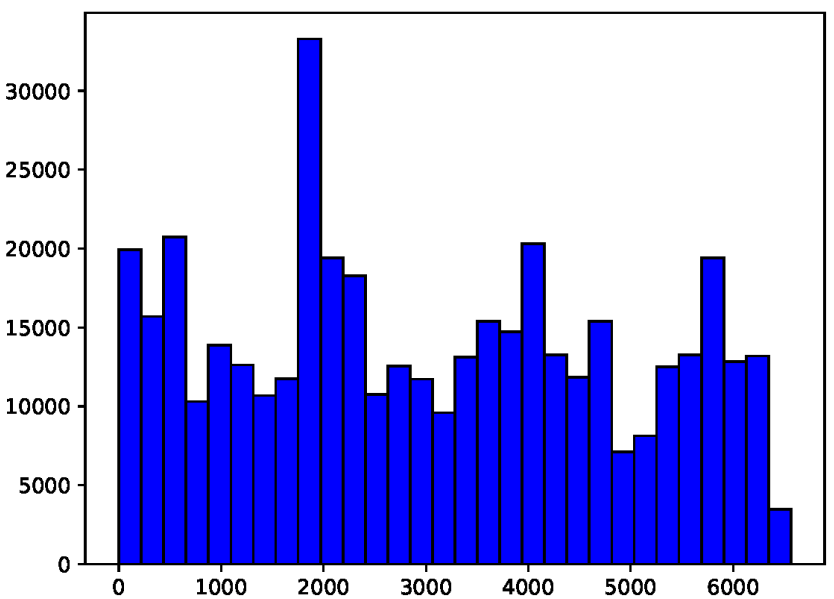
Źródło: opracowanie własne.

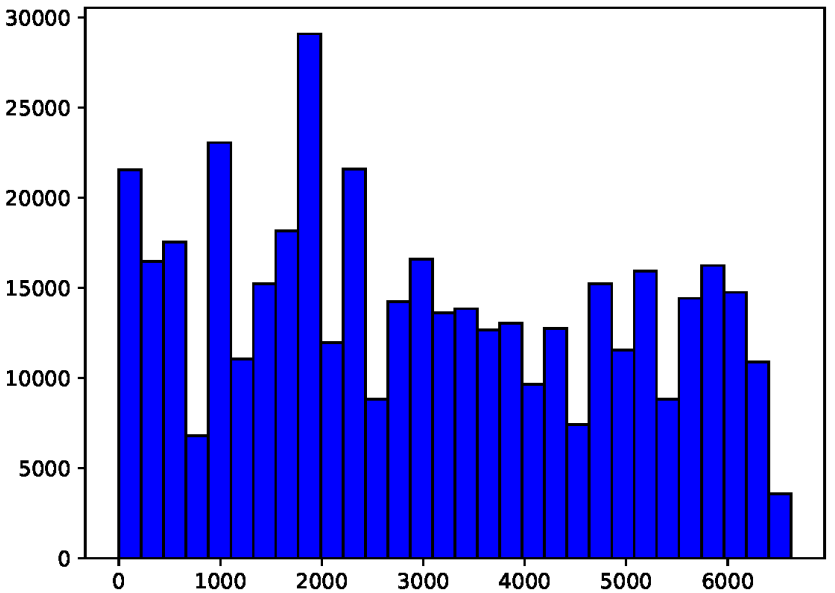
Wykr. 6.6. Miesięczne histogramy dopasowania produktów (GestorKappa)






30000
25000
20000
15000
10000
5000
0
0
1000
2000
3000
4000
5000
6000


30000
25000
20000
15000
10000
5000
0
0
1000
2000
3000
4000
5000
6000



Źródło: opracowanie własne.

6.3.6. Eksperyment dotyczący wsparcia informatycznego przy wyborze reprezentantów dla kategorii COICOP

Eksperyment przeprowadzono metodą przedstawioną w pkt 6.2.9 na danych GestorAlfa 2018/9 opisanych w pkt 6.1.2. Polegał on na wyliczeniu rankingu produktów, które byłyby proponowane jako reprezentanci dla kategorii poziomu 6 – im wyżej na liście, tym mocniej rekomendowane. Ranking polegał na wyliczeniu podobieństw tekstowych opisu produktu do nazwy kategorii z poziomu 6 COICOP. Przykładowe rankingi znajdują się w tabl. 6.25, 6.26, 6.27, 6.28 i 6.29.

Tabl. 6.25. Sugestie reprezentantów w ramach kategorii „Cukier" (CP01181) dla podkategorii „Cukier biały, kryształ"

Podobieństwo	Opis produktu
0,77	03200110 Cukier kryształ biały 1kg
0,58	3/Cukier biały drobny 750g
0,33	Diamant Cukier trzcinowyMuskovado500g
0,33	Cukier żelujący 3:1 350g
0,33	Cukier żelujący 2:1 500g
0,33	Cukier trzcinowy 500g
0,29	Tereos Cukier puder 1kg
0,29	GOLDPACK cukier trzcinowy 1kg
0,29	Diamant cukier rafinada 1000g
0,29	Diamant Cukier żelujący 1:1 1kg
0,29	Diamant Cukier w saszetkach 100x5g
...	...

Źródło: opracowanie własne.

Tabl. 6.26. Sugestie reprezentantów w ramach kategorii „Kawa" (CP01211) dla podkategorii „Kawa naturalna ziarnista"

Podobieństwo	Opis produktu
0,58	Ex.Kawa ziarnista TurriAlba500g
0,58	00064198Tchibo CaffeCrema kawa ziarnista
0,52	Woseba kawa ziarnista Espresso 1 KG
0,52	Woseba Unica kawa ziarnista 500g
0,52	Woseba Kawa ziarnista Arabica 500g
0,52	Woseba Gold kawa ziarnista 500 g
0,52	Woseba Espresso kawa ziarnista 500 g
0,52	Woseba Arabica kawa ziarnista 1 Kg
0,52	Tchibo kawa ziarnista Black&White 500G
0,52	Tchibo Family kawa ziarnista 500g.
0,52	Tchibo Exclusive kawa ziarnista 500g
...	...

Źródło: opracowanie własne.

Tabl. 6.27. Sugestie reprezentantów w ramach kategorii „Jogurt" (CP01144) dla podkategorii „Jogurt owocowy"

Podobieństwo	Opis produktu
0,63	KBio.Jogurt owocowy 2,7 mix 150g
0,63	KBio Jogurt owocowy 2,7 mix 150g
0,58	dele/ B/ BIO jogurt owocowy mix 150g
0,41	R/Ozorków Jogurt 330g
0,41	R/Cuiavia Jogurt naturalny150g
0,41	G/Activia jogurt malina4x120g
0,41	G/ Jogurt naturalny b/c 165g
0,41	B/G/Activia jogurt jagoda2x120g
0,41	Activia jogurt jagoda120 g
0,35	R/Świecie Jogurt naturalny 150g
0,35	R/Łobżenica jogurt naturalny 150g

...	...

Źródło: opracowanie własne.

Tabl. 6.28. Sugestie reprezentantów w ramach kategorii „Mąka pszenna" (CP011121) dla podkategorii „Mąka pszenna Poznańska"

Podobieństwo	Opis produktu
0,71	Basia Mąka pszenna poznańska T 500 1kg
0,58	Mąka pszenna T-500 1kg
0,52	R/Werbkowice Mąka Poznańska T-500 4 kg
0,52	R/Werbkowice Mąka Poznańska 1kg T-500
0,52	R/Werbkowice Mąka Poznańska 1kg T-450
0,52	R/Stanisz.Mąka pszenna t550 luksusowa1kg
0,52	R/Stanisz.Mąka pszenna t.450 tortowa1kg
0,52	R/Młyny Szczepanki Mąka poznańska 1kg
0,52	R/Młyn Miłosna Mąka poznańska 1kg
0,52	R/Kruszwica Mąka pszenna t500 wrocławska
0,52	R/Kruszwica Mąka pszenna t-550 luksusowa
...	...

Źródło: opracowanie własne.

Tabl. 6.29. Sugestie reprezentantów w ramach kategorii „Ryż" (CP01111) dla podkategorii „Ryż biały"

Podobieństwo	Opis produktu
0,71	Janex Ryż biały 4x100g
0,71	Cenos Ryż biały 4x100g
0,63	Sawi Ryż biały do sushi1kg
0,63	Risana ryż biały kartonik 4x100g
0,63	Risana Ryż jaśminowy biały 4x100g
0,63	Risana Ryż amerykański biały 4x100g
0,63	Kuchnia Gosposi Ryż biały 1kg
0,63	Cenos Ryż biały długi 8x100g
0,63	Cenos Ryż biały długi 1kg
0,58	Janex Ryż biały długoz. 400g folia
0,35	Ryż Kuchnia Gosposi 4x100g
...	...

Źródło: opracowanie własne.

6.4. Wnioski wynikające z przeprowadzonych eksperymentów

6.4.1. Unikatowość kodów i etykiet produktów

Kody produktów nie są unikatowe w relacji do GTIN ani odwrotnie. Oznacza to poważny problem merytoryczny w ramach pobierania danych od gestorów. Z przeprowadzonych eksperymentów wynika, że żaden kod nie może być wykorzystywany do stwierdzenia pojawiania się nowych produktów ani zanikania sprzedaży już istniejących, w szczególności należących do koszyka. Konieczne jest zatem wykorzystanie modułu IC.Matcher w celu wskazania sytuacji wątpliwych.

To, że opis produktu jest unikatowy dla wszystkich produktów w ramach jednego gestora, wskazywałoby, że gestorzy przywiązują większą wagę do nadania własnego kodu produktu niż do norm międzynarodowych. Jednak również kod produktu od gestora nie daje pewności ciągłości/nieciągłości sprzedaży tego samego produktu z punktu widzenia analizy cen, ponieważ zmiana tego kodu może być związana z ruchem cen, a nie wprowadzeniem nowego produktu do obrotu.

6.4.2. Zmienność produktów

Powyższy problem staje się tym bardziej istotny, gdy zwrócimy uwagę na dość wysoką zmienność oferty produktów, która z miesiąca na miesiąc wynosi zwykle powyżej 5%, ale może sięgać nawet 40% oferowanych produktów. W tym kontekście zdolność do automatycznej klasyfikacji do COICOP nowych produktów zyskuje na znaczeniu.

6.4.3. Klasyfikacja

Przeprowadzone eksperymenty prowadzą do następujących wniosków:

- dla dostępnych danych najlepiej sprawdzały się metody: linowa maszyna wektorów nośnych (LinearSVC) (dokładność 93% dla poziomu 3 COICOP), naiwny klasyfikator bayesowski (93%), regresja logistyczna (92%);
- klasyfikacja do wyższych poziomów hierarchii wypada nieco lepiej (93% dla poziomu 3 COICOP, 91% dla poziomu 5 COICOP dla maszyn wektorów nośnych);
- stosowane metody pozwalają na ekstrakcję słów z nazw produktów charakteryzujących klasy (por. tabl. 6.10);
- klasy z małą liczbą produktów są czynnikiem ryzyka;
- wyzwaniami będą: dane z większą liczbą rekordów / z większą liczbą klas;

Wspomnianym problemom mogą zaradzić:

- zwiększenie liczby klasyfikatorów (komitety);
- klasyfikatory hierarchiczne (klasyfikacja na kolejne poziomy hierarchii);
- klasyfikacja nie kategorii, lecz słów należących do nazwy kategorii.

6.4.4. Ekstrakcja słów kluczowych i klasyfikacja na ich podstawie

Przeprowadzone badania wskazują, że najbardziej skuteczną, a zarazem intuicyjną metodą ekstrakcji słów kluczowych jest wykorzystanie modelu stworzonego na bazie naiwnego klasyfikatora bayesowskiego. Pozwala on na zidentyfikowanie zarówno pozytywnych, jak i negatywnych słów kluczowych, charakteryzujących klasę COICOP. Wykorzystanie tych słów kluczowych daje bardzo dobre wyniki przy identyfikacji klasy na podstawie opisu produktu.

Podejście teoriomnogościowe, polegające na wykrywaniu słów różnicujących klasy, jest mniej efektywne, choć daje także dobre rezultaty, ponieważ dla niektórych

klas COICOP nie da się takich słów ustalić ze względu na współdzielenie z innymi klasami. Niemniej klasyfikator klasyfikujący wyłącznie na podstawie słów pozytywnych uzyskuje dobre wyniki klasyfikacji do COICOP (dokładność ponad 95% dla dziewięciu wybranych klas).

6.4.5. Śledzenie produktów

Trenując klasyfikatory na danych z bieżącego miesiąca, bardzo dobrze (dokładność ponad 99%) przewiduje się klasyfikację COICOP na miesiąc następny. Dla nowo pojawiających się produktów najlepiej z zadaniem klasyfikacji radzi sobie klasyfikator oparty na liniowych maszynach wektorów nośnych (dokładność ponad 90%), najgorzej – lasy losowe (ok. 80%). Stopień trudności zadania zależy od liczby kategorii, do których klasyfikuje się produkty (maszyny wektorów nośnych osiągają dokładność ponad 90% dla 26 klas, ponad 95% dla 10 klas).

Zagadnienie automatycznego śledzenia produktów wydaje się dość istotne z praktycznego punktu widzenia, ponieważ grupy produktów tożsamych, różniących się jedynie ceną, kodem produktu lub opisem, może sięgać kilku tysięcy i manualny dobór produktów do analizy cen jest zadaniem zbyt trudnym dla wykonania.

6.4.6. Wsparcie wyboru reprezentantów klas

Uzyskane wyniki są obiecujące, a proponowane produkty wydają się dobrze reprezentować odpowiednie klasy COICOP. Jest to jednak ocena subiektywna i korzystanie z takiego narzędzia powinno odbywać się na zasadzie uzyskiwania pierwszej iteracji – propozycji, którą powinien zweryfikować ekspert. Ocena mierzalnych efektów wymagałaby współpracy większego zespołu ekspertów i wykracza poza zakres projektu.

ROZDZIAŁ 7
Zastosowanie wyników projektu w badaniu cen detalicznych – możliwości wdrożeniowe i wyzwania

7.1. Nowa platforma informatyczna InstatCeny umożliwiająca przetwarzanie danych pochodzących z alternatywnych źródeł

Platforma InstatCeny umożliwia przetwarzanie danych skrapowanych i skanowanych. Została utworzona w języku Java i składa się z następujących modułów:

- price – moduł główny, umożliwiający dostęp do pozostałych modułów;
- price-user – moduł do zarządzania użytkownikami systemu;
- price-log – moduł przechowujący listę zdarzeń;
- price-scraper – moduł do obsługi danych skrapowanych;
- price-scanner – moduł do obsługi danych skanowanych.

Platforma komunikuje się z modułami opracowanymi przez IPI PAN za pomocą API (ang. *application programming interface*). Dostępny jest również interfejs graficzny dla użytkowników.

7.1.1. Pozyskiwanie danych skrapowanych za pomocą narzędzi informatycznych

Do pozyskiwania danych ze sklepów internetowych służy moduł opracowany przez IPI PAN. Moduł pobiera dane ze sklepów i zapisuje we własnej bazie danych. Skrapowanie odbywa się codziennie. Moduł price-scraper obserwuje proces skrapowania, a po jego zakończeniu w danym dniu importuje dane do własnej bazy danych, gdzie zachodzi dalsze przetwarzanie. Importowane są następujące informacje o produkcie:

- URL (adres);
- identyfikator produktu;
- nazwa;
- opis;

- marka;
- data skrapowania;
- dostępność produktu;
- cena;
- cena promocyjna;
- GTIN;
- liczba opakowań;
- ilość w opakowaniu;
- ilość całkowita;
- jednostka miary.

Po zakończeniu importu dane z danego dnia są klasyfikowane. Po zakończeniu miesiąca wykonywane jest kojarzenie produktów. Do obydwu operacji wykorzystywane są moduły opracowane przez IPI PAN.

7.1.2. Pozyskiwanie danych skanowanych

Dane skanowane są to dane pochodzące z systemów informatycznych sieci handlowych. Transfer tego rodzaju danych odbywa się raz w miesiącu. Każda z sieci może stosować nieco inny format danych, więc dla każdej z nich wymagane jest utworzenie osobnego procesu ETL (ang. *extract transform load*). Dane skanowane mogą zawierać następujące informacje:

- pierwszy dzień, w którym obowiązywała dana cena;
- ostatni dzień, w którym obowiązywała dana cena;
- identyfikator sklepu;
- kod pocztowy;
- identyfikator kategorii produktu;
- nazwa kategorii produktu;
- identyfikator produktu;
- nazwa;
- ilość;
- jednostka miary;
- stawka VAT;
- cena;
- obrót;
- liczba sprzedanych sztuk;
- procent obrotu, jaki stanowiła promocja;
- GTIN.

Po zakończeniu importu produkty są klasyfikowane, a następnie kojarzone. Do obydwu operacji wykorzystywane są moduły opracowane przez IPI PAN.

7.1.2.1. Dane ściągane bezpośrednio z systemów informatycznych sieci handlowych

Sieć handlowa może udostępnić dane skanowane w ramach swojej infrastruktury informatycznej. W praktyce stosuje się w tym celu serwer SFTP, na którym sieć handlowa umieszcza dane w postaci pliku ZIP. Platforma InstatCeny wczytuje ten plik za pomocą procesu ETL przeznaczonego dla danej sieci handlowej. Uwierzytelnianie do serwera SFTP odbywa się za pomocą klucza publicznego, a do transmisji danych używany jest bezpieczny protokół SSH.

7.1.2.2. Zbiory dostarczane przez sieci za pośrednictwem systemu informatycznego TransGUS

Jeżeli sieć handlowa nie chce lub nie może udostępniać danych skanowanych w ramach swojej infrastruktury informatycznej, może skorzystać z systemu TransGUS. Jest to system, który działa na serwerze GUS i umożliwia transmisję danych w postaci plików. Sieć handlowa przygotowuje raz w miesiącu plik z danymi i wysyła go do systemu TransGUS. Może to zrobić ręcznie za pomocą przeglądarki internetowej lub automatycznie za pomocą API. Sieć handlowa uwierzytelnia się w systemie TransGUS za pomocą loginu i hasła, a do transmisji danych jest używany bezpieczny protokół TLS. TransGUS umieszcza otrzymany plik na serwerze FTP znajdującym się w sieci GUS. Platforma InstatCeny wczytuje ten plik za pomocą procesu ETL przeznaczonego dla danej sieci handlowej.

7.2. Zastosowanie nowych formuł obliczania wskaźników cen detalicznych z wykorzystaniem big data

Alternatywne źródła danych w pomiarze inflacji stwarzają dodatkowe możliwości w zakresie wyboru formuły indeksu do analizy dynamiki cen. W przypadku danych skanowanych nawet na najniższym poziomie agregacji danych (czyli na poziomie kodu kreskowego) dysponujemy informacją o cenie sprzedawanego produktu oraz wielkości jego sprzedaży. Dzięki zastosowaniu ważonego indeksu cen do porównania okresu bieżącego z okresem badanym możemy oprzeć system wag na dowolnym okresie z analizowanego przedziału czasowego. To umożliwia zastosowanie nie tylko bilateralnych ważonych formuł indeksów cen, lecz także formuł indeksowych działających na całym oknie czasowym – indeksów łańcuchowych lub indeksów multilateralnych. Jest to szczególnie ważne z uwagi na dużą rotację produktów oferowanych przez sieci handlowe (por. pkt 7.2.1). W przypadku danych skrapowanych rotacja produktów w trakcie miesięcy badanego przedziału czasowego jest porównywalnie duża, a co za tym idzie potencjalnie dobrym wyborem formuły indeksu są również indeksy multilateralne (w tym przypadku nieważone). Formuły indeksów cen zaimplementowane w ramach projektu zostały szczegółowo omówione w podrozdz. 4.5. W kontekście organizacji pracy obecnego systemu informatycznego GUS oraz ocze-

kiwanej kompatybilności produkcji wskaźników cen opartych na alternatywnych źródłach danych z produkcją wskaźników cen obliczonych na podstawie danych zbieranych przez ankieterów planowane jest obecnie wdrożenie podejścia dynamicznego (ang. *dynamic approach* – por. podrozdz. 4.4), uwzględniającego z łańcuchową formułę Jevonsa. Niemniej prace eksperymentalne wykonane w ramach projektu miały na celu również rozpoznanie zasadności stosowania indeksów multilateralnych w produkcji wskaźników cen opartych na alternatywnych źródłach danych. W pakiecie PriceIndices zaimplementowano niemal wszystkie znane metody multilateralne (indeksy GEKS, GEKS-W, GEKS-J, CCDI, GK, czyli Geary-Khamis, TPD; zob. Chessa, 2015, 2016), a także pewne propozycje multilateralnych metod (GEKS-L, GEKS-GL, GEKS-AQI, GEKS-AQU; por. Białek, 2022). Ostatecznie aplikacja stworzona przez IPI PAN uwzględniła formuły: GEKS, GEKS-L, GK i SPQ. Analiza literatury przedmiotu, a także wyników przeprowadzonych eksperymentalnych prac projektowych pozwoliła autorom projektu na ustalenie metodologicznych i technologicznych uwarunkowań potencjalnego wdrożenia do praktyki GUS indeksów multilateralnych w przyszłości. Tematem niniejszego podrozdziału są zasadność stosowania tego rodzaju indeksów, uwarunkowania wdrożenia, jak również kryteria wyboru indeksu multilateralnego.

7.2.1. Zasadność stosowania multilateralnych wskaźników cen

Indeksy multilateralne, w odróżnieniu od bilateralnych, biorą pod uwagę nie tylko porównywane ze sobą okresy bieżący i bazowy, lecz także wszystkie okresy pośrednie. Ma to szczególne znaczenie w przypadku produktów sprzedawanych przez sieci handlowe, ponieważ takie sieci charakteryzuje duża rotacja asortymentu, mnogość dóbr nowych i znikających, a także znaczny odsetek produktów sezonowych. Z konstrukcji indeksów multilateralnych wynika[3], że ich użytkownik ma możliwość sterowania oknem czasowym $[0, T]$, jakie obejmuje porównywane miesiące. Co za tym idzie wybór 13-miesięcznego okna (rzadziej jest to 25 miesięcy) pozwala uwzględnić zjawisko sezonowości produktów.

Z aksjomatycznego punktu widzenia (por. pkt 7.2.3), cenną własnością, jaka charakteryzuje indeksy multilateralne, jest tranzytywność. Za pomocą oznaczenia przez $P_M^{0,t}$ indeksu multilateralnego, jaki mierzy dynamikę cen, porównując okres bieżący t z okresem bazowym 0, możemy wyrazić aksjomat tranzytywności następującą równością:

$$P_M^{0,t} = P_M^{0,s} \cdot P_M^{s,t}, \tag{7.1}$$

[3] Wyjątkiem jest indeks SPQ, który nie działa na oknie czasowym.

jaka powinna zachodzić dla dowolnego $0 < s < t \leq T$. Tranzytywność jest kluczową własnością, która powoduje, że indeksy multilateralne są wolne od zjawiska łańcuchowego dryfu (ang. *chain drift bias*). Innymi słowy, gdy ceny i ilości sprzedawanych produktów powracają do wyjściowego poziomu (np. po upływie sezonu), wartości indeksów multilateralnych konsekwentnie powracają do jedności. Jest to główna zaleta indeksów multilateralnych w kontekście wykorzystania danych skanowanych lub skrapowanych, ponieważ nie generują dzięki temu obciążenia pomiaru dynamiki cen wynikającego z łańcuchowego dryfu.

7.2.2. Uwarunkowania związane z wdrożeniem indeksów multilateralnych

Przed podjęciem decyzji o wdrożeniu indeksów multilateralnych do produkcji wskaźników cen urząd statystyczny musi uwzględnić uwarunkowania tego wdrożenia, jakie wynikają z konstrukcji indeksów multilateralnych. Po pierwsze jak wielokrotnie wcześniej podkreślano, indeksy multilateralne (z wyjątkiem indeksu SPQ) operują na zadanym z góry oknie czasowym. Najczęściej jest to okno przynajmniej 13-miesięczne, dlatego urząd statystyczny musi dysponować dostatecznie długimi szeregami czasowymi dla danych dostarczanych przez współpracującą sieć handlową. Po drugie w przypadku pojawienia się nowych danych, tj. za kolejny miesiąc, rozszerzenia okna czasowego zmieniłyby poprzednio wyznaczony indeks cen. Publikowane wskaźniki CPI nie podlegają rewizji, dlatego należy zastosować odpowiednią metodę łączenia starego okna czasowego z nowym (ang. *linking method*). W literaturze przedmiotu funkcjonuje wiele metod, które eliminują konieczność rewizji poprzednio wyznaczonych indeksów cen poprzez stosowanie ich rozszerzeń (ang. *multilateral extensions*). Pierwszą grupę metod stanowią metody przesuwanego okna (ang. *rolling-window methods* lub *splicing methods*), w ramach których wraz z każdym miesiącem okno o ustalonej długości (z reguły 13 lub 25 miesięcy) jest przesuwane o jeden miesiąc do przodu. W zależności od wyboru miesiąca, który służy do połączenia indeksu wyznaczonego na starym oknie z indeksem bazującym na nowym oknie (ang. *linking month*), możliwe są następujące warianty: (1) metoda *movement splice* – miesiącem łącznikiem jest tu ostatni miesiąc z poprzedniego okna (de Haan i van der Grient, 2011); (2) metoda *window splice* miesiącem łącznikiem jest drugi miesiąc z poprzedniego okna (Krsinich, 2014); (3) metoda *half splice* – miesiącem łącznikiem jest miesiąc znajdujący się w połowie poprzedniego okna (de Haan, 2015); (4) metoda *mean splice* – miesiącem łącznikiem jest każdy potencjalny miesiąc z poprzedniego okna, a wyniki łączenia okien są uśredniane za pomocą średniej geometrycznej (Diewert i Fox, 2018). Dodatkowo niektóre kraje (np. Holandia) stosują modyfikacje wymienionych metod bazujące na publikowanych wskaźnikach, np. metodę WISP (ang. *window splice on published indices*) lub HASP (ang. *half splice on published indices*), gdzie bazowe okno składa się z 25 miesięcy

(Chessa, 2019). Należy mieć na uwadze, że metody łączenia okien (ang. *window splice methods*) mogą prowadzić do obciążenia pomiaru wynikającego z efektu łańcuchowego dryfu. Dlatego w literaturze przedmiotu można spotkać jeszcze inne podejścia. Jednym z nich jest metoda rozszerzania okna czasowego przy ustalonym miesiącu bazowym – *fixed base expanding window* (FBEW). W tej metodzie w każdym roku analizy okno czasowe rozszerza się z każdym nowym miesiącem przy zachowaniu grudnia poprzedniego roku jako miesiąca bazowego i dzieje się tak aż do osiągnięcia grudnia bieżącego roku (Chessa, 2016). Możliwe są również kombinacje omawianych metod, np. metoda *fixed base moving window* (FBMW), stanowiąca wypadkową metody FBEW i metody *movement splice* (Lamboray, 2017). Wszystkie omówione w tej części pracy metody rozszerzeń indeksów multilateralnych zostały oprogramowane w pakiecie PriceIndices (Białek, 2021, 2022a).

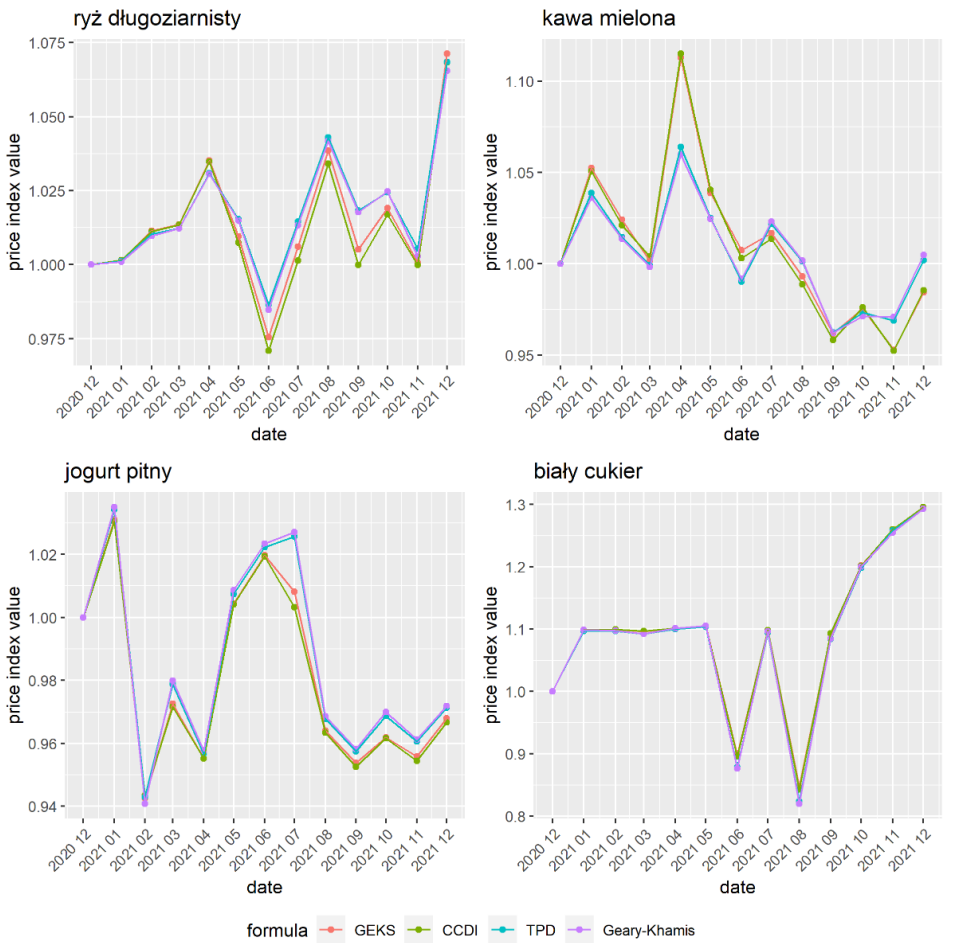
Poza omówionymi konsekwencjami stosowania indeksów multilateralnych warto wspomnieć jeszcze o uwarunkowaniach po stronie systemu IT. Indeksy te wykorzystują całe okno czasowe analizy, dlatego pełne ramki danych (z identyfikatorami dopasowanych produktów oraz identyfikatorami outletów) muszą być dostępne przez cały czas pracy na danym oknie czasowym. Dopiero gdy informacje z danego okna czasowego są już niepotrzebne do wyliczania bieżącego wskaźnika cen (a niezbędne są jedynie informacje o poprzednio wyznaczonych wskaźnikach), dane z tego okna mogą zostać skompresowane, a następnie przeniesione do innej lokalizacji lub nawet usunięte. Stosowanie indeksów multilateralnych wymaga zatem większej przestrzeni dyskowej na serwerze operacyjnym niż np. stosowanie bilateralnych indeksów cen, porównujących bieżący miesiąc do miesiąca poprzedniego. Dodać należy, że kalkulacja indeksów multilateralnych jest kilkukrotnie lub nawet kilkunastokrotnie bardziej czasochłonna, niż ma to miejsce w przypadku klasycznych indeksów bilateralnych (Białek i Beręsewicz, 2021). Ma to szczególne znaczenie w przypadku dużych rozmiarów baz danych skanowanych, gdzie obliczenie indeksu multilateralnego dla pojedynczej grupy elementarnej produktów dla danych pochodzących od jednej tylko sieci może trwać nawet kilkadziesiąt minut.

7.2.3. Kryteria wyboru indeksu multilateralnego

Problem wyboru optymalnej metody multilateralnej wydaje się nadal otwarty. W literaturze przedmiotu funkcjonuje co najmniej kilka kryteriów (podejść), które mogą być brane pod uwagę przez urząd statystyczny przed ostatecznym wyborem multilateralnego indeksu cen. Choć kryteria te nie prowadzą do zbieżnych wskazań co do najlepszego wyboru indeksu, to budują przynajmniej świadomość użytkowników w zakresie zalet i wad rozważanych indeksów. Niestety indeksy multilateralne mogą prowadzić do wyników, które w ciągu roku różni nawet kilka punktów procentowych. Dzieje się tak zwłaszcza w przypadku dużych fluktuacji w poziomie kon-

sumpcji produktów, choć zmienność cen produktów również odgrywa tu ważną rolę. Gdy zmiany cen i ilości sprzedawanych produktów są dość stabilne, wówczas z reguły nie rejestrujemy znacznych różnic pomiędzy wskazaniami indeksów multilateralnych. Przykładowe porównanie wybranych formuł multilateralnych dla grup: ryż długo-ziarnisty, kawa mielona, jogurt pitny oraz cukier biały, uzyskanych od jednej ze współpracujących z GUS sieci handlowych, przedstawiono na wykr. 7.1.

Wykr. 7.1. Porównanie wybranych indeksów multilateralnych dla czterech grup produktów spożywczych od grudnia 2020 do grudnia 2021 r.



Źródło: obliczenia własne w pakiecie PriceIndices.

Najbardziej powszechne przy wyborze indeksu multilateralnego jest podejście aksjomatyczne (ang. *axiomatic approach*). Listę aksjomatów (testów), jakich wymaga się od poprawnie skonstruowanego indeksu multilateralnego, można znaleźć w pra-

cach Australijskiego Urzędu Statystycznego (ABS, 2016) lub np. w pracy Zhang i in. (2019). Choć zbiór testów spełnianych jednocześnie przez wszystkie opisane w literaturze indeksy multilateralne jest dość szeroki, niektóre aksjomaty okazują się bardziej restrykcyjne od innych. Na przykład test jednoznaczności (ang. *identity test*) jest spełniony przez indeksy SPQ, GEKS-L czy GEKS-GL, ale już nie przez formuły GEKS, CCDI, GK czy TPD (por. Białek, 2022b). Test ten jest silniejszy od wymogu braku łańcuchowego dryfu, ponieważ orzeka, że w sytuacji powrotu jedynie cen do poziomu wyjściowego (ilości sprzedaży mogą być dowolne) indeks cen przyjmuje wartość równą 1.

Innymi podejściami ugruntowanymi w literaturze są podejście ekonomiczne oraz podejście stochastyczne. W ramach tego pierwszego zakłada się, że ilości sprzedawanych produktów są funkcjami cen, a każde gospodarstwo domowe dąży do minimalizacji wydatków zakupowych przy jednoczesnym utrzymaniu swojej indywidualnej użyteczności z koszyka inflacyjnego na stałym poziomie. W podejściu stochastycznym buduje się ekonometryczny model wyjaśniający zachowanie cen produktów w przedziale czasowym, a następnie konstruuje indeks na podstawie estymowanych parametrów tego modelu. Oba podejścia mają swoich faworytów wśród indeksów multilateralnych – są nimi odpowiednio indeksy GEKS oraz TPD.

Ostatnio w literaturze przedmiotu dyskusji poddano dwa nowe kryteria wyboru indeksu multilateralnego. Pierwsze to nowe podejście stochastyczne (ang. *new stochastic approach*), które zakłada, że w eksperymencie symulacyjnym generowane są ceny i ilości produktów według zadanego z góry, znanego rozkładu prawdopodobieństw (np. log-normalnego). Rozkład jest tak dobierany, aby można było znaleźć teoretyczną (oczekiwaną) wartość indeksu cenowego na zadanym odcinku czasu. Następnie dla wygenerowanych procesów cen i ilości oblicza się indeksy multilateralne i sprawdza, na ile są one rozbieżne w stosunku do wyznaczonej wartości teoretycznej (por. Białek i Bobel, 2019). Zgodnie z tym kryterium preferowanym indeksem jest taki indeks, który generuje tu najmniejsze obciążenie pomiaru. Drugim kryterium wyboru może być wreszcie kryterium czasochłonności (ang. *time-consuming criterion*). W ramach tego kryterium wybieramy formułę indeksu najszybciej prowadzącą do wyniku oszacowania (Białek i Beręsewicz, 2021). Pod tym względem najlepszy okazuje się indeks SPQ, dalej indeksy GEKS-L, GEKS-GL czy GEKS, natomiast najbardziej czasochłonne obliczenia dotyczą indeksów GK czy TPD (por. Białek, 2022b).

7.3. Wykorzystanie pakietu PriceIndices do bieżących obliczeń wskaźników cen detalicznych

Program PriceIndices został napisany przez Białka (2021) w środowisku R, a następnie zaimplementowany w projektowej aplikacji stworzonej przez IPI PAN (por. pkt

5.4.3). Program, stanowiący pakiet środowiska R, funkcjonuje niezależnie od tej aplikacji i można go pobrać bezpośrednio z serwera CRAN za pomocą komendy: install.packages("PriceIndices"), lub z serwera GitHub: remotes::install_github ("JacekBialek/PriceIndices"). Na serwerze CRAN są dostępne obszerne dokumentacje dotyczące pakietu[4] i demonstracyjny plik README[5]. Z kolei szczegółowe omówienie funkcjonalności pakietu znaleźć można w pracy Białka (2022a).

Aktualna wersja pakietu PriceIndices (wersja 0.0.7) składa się z ponad 170 funkcji. Zasadniczo pakiet ten realizuje następujące zadania związane z procedowaniem danych skanowanych oraz skrapowanych:

- funkcje do przygotowania danych do analizy:
  - data_preparing() – przygotowanie ramki danych do analizy;
  - data_agregating() – agregacja cen i ilości produktów względem miesięcy i punktów sprzedaży;
  - data_unit() – wydobywanie informacji o gramaturze i jednostce sprzedaży na podstawie opisu produktu;
  - data_norm() – przeskalowanie cen i ilości produktów, tak aby odpowiadały ustalonej jednostce sprzedaży;
- funkcje do opisu zbioru danych:
  - available() – ustalenie dostępnych produktów, kodów produktów lub kodów punktów sprzedaży w przedziale czasowym;
  - matched() – ustalenie listy dopasowanych kodów produktów, kodów punktów sprzedaży lub opisów produktów w przedziale czasowym;
  - matched_index(), matched_fig() – liczbowe lub graficzne prezentowanie udziału dopasowanych elementów, jakie zwraca funkcja matched() w ogólnej liczbie dostępnych produktów;
  - prices(), quantities(), sales() – wyznaczenie średnich miesięcznych cen, ilości i wartości sprzedaży określonej grupy produktów;
  - sales_groups() – tabelaryczne lub graficzne prezentowanie udziałów w sprzedaży podgrup produktów z określonej grupy produktów;
  - pqcor(), pqcor_fig() – wyznaczenie korelacji pomiędzy cenami i ilościami produktów odpowiednio dla zadanego miesiąca lub dla zadanego przedziału czasowego;
  - dissimilarity(), dissimilarity_fig() – wyznaczenie miary niepodobieństwa pomiędzy cenami i ilościami produktów odpowiednio dla zadanego miesiąca lub dla zadanego przedziału czasowego;
- funkcje do klasyfikacji produktów:
  - model_classification(), save_model(), load_model() – odpowiednio: budowa modelu uczenia maszynowego do klasyfikacji produktów do grup COICOP, zapis modelu na dysku, odczyt modelu z dysku;

[4] https://cran.r-project.org/web/packages/PriceIndices/PriceIndices.pdf.
[5] https://cran.r-project.org/web/packages/PriceIndices/readme/README.html.

  - data_classifying() – klasyfikacja produktów do grup COICOP na podstawie wytrenowanego wcześniej modelu uczenia maszynowego (zaimplementowano algorytm drzew losowych);
  - data_selecting() – wybór produktów i klasyfikacja do grup COICOP na podstawie zestawu fraz i słów kluczowych w opisie produktu;
- funkcje do dopasowania i filtrowania danych:
  - data_matching() – dopasowanie produktów w czasie na podstawie etykiet i kodów;
  - data_filtering() – filtrowanie danych, m.in. implementacja filtru niskich sprzedaży i ekstremalnych cen;
- funkcje do wyznaczania nieważonych indeksów bilateralnych:
  - w pakiecie PriceIndices zaimplementowano sześć nieważonych indeksów bilateralnych, m.in. formułę Jevonsa (funkcja jevons()), Carliego (funkcja carli()) oraz Dutota (funkcja dutot());
- funkcje do wyznaczania ważonych indeksów bilateralnych:
  - w pakiecie PriceIndices zaimplementowano 26 ważonych indeksów bilateralnych, m.in. formuły Laspeyresa (funkcja laspeyres()), Paaschego (funkcja paasche()) oraz Fishera (funkcja fisher());
- funkcje do wyznaczania indeksów łańcuchowych:
  - w pakiecie PriceIndices zaimplementowano 32 nieważonych i ważonych indeksów łańcuchowych, m.in. łańcuchową formułę Laspeyresa (funkcja chlaspeyres()), Paaschego (funkcja chpaasche()) oraz Fishera (funkcja chfisher());
- funkcje do wyznaczania indeksów multilateralnych:
  - w pakiecie PriceIndices zaimplementowano 18 indeksów multilateralnych, m.in. indeksy GEKS (funkcja geks()), CCDI (funkcja ccdi()), TPD (funkcja tpd()), SPQ (funkcja SPQ()) czy Geary-Khamisa (funkcja gk()); pakiet zawiera także funkcje do rozszerzeń indeksów multilateralnych (ang. *extensions*), w szczególności metody łączenia okienkowego, metodę rozszerzanego okna oraz przesuwanego okna, a dodatkowo ogólne funkcje do wyznaczania wielu różnych indeksów na tym samym zbiorze danych (m.in. funkcja price_indices());
- funkcje do agregacji indeksów cząstkowych:
  - final_index(), final_index2() – agregacja indeksów cząstkowych wyznaczonych na podstawie wskazanych podgrup produktów oraz identyfikatora punktów sprzedaży; agregacja względem podgrup produktów oraz względem outletów odbywa się przy zastosowaniu wskazanej przez użytkownika funkcji agregującej (np. formuły Laspeyresa lub Fishera);
- funkcje do porównywania indeksów cen:
  - compare_indices() – graficzne porównanie indeksów cen wyznaczonych na tym samym zbiorze danych;

- compare_final_indices() – graficzne porównanie indeksów cen wyznaczonych na różnych zbiorach danych;
- compare_distances() – porównanie indeksów cen za pomocą miary MAD (ang. *mean absolute distance*) lub RMSD (ang. *root mean square distance*), czyli odpowiednio średniej absolutnej różnicy lub średniokwadratowej różnicy z ich wartości;
- compare_to_target() – porównanie wskazanych indeksów cen z indeksem stanowiącym punkt odniesienia za pomocą miary MAD lub RMSD.

W zrealizowanych pracach projektowych oraz bieżących pracach GUS związanych z zastosowaniem danych skanowanych i skrapowanych do pomiaru inflacji pakiet PriceIndices wspiera wszystkie omówione etapy procedowania danych z tego typu źródeł. W szczególności pakiet realizuje zadania związane z klasyfikacją, dopasowaniem i filtracją produktów. Za pomocą pakietu przygotowywane i formatowane są ramki danych, które następnie przechodzą do etapu wyznaczania wskaźników cen. W pakiecie PriceIndices są wyznaczane miesięczne wskaźniki cen dla danych skanowanych i skrapowanych, zarówno według formuł obecnie wdrażanych (np. łańcuchowy indeks Jevonsa czy indeks Fishera), jak i według formuł, nad którymi na razie prowadzone są prace eksperymentalne (np. indeks GEKS czy TPD).

Wykr. 7.2 przedstawia porównanie wybranych indeksów cen dla zbioru *milk,* który został zaimplementowany w pakiecie PriceIndices. Porównano ze sobą indeks łańcuchowy Jevonsa, bilateralne indeksy Laspeyresa i Fishera oraz multilateralne indeksy GEKS i TPD dla 13-miesięcznego okna czasowego (od grudnia 2018 r. do grudnia 2019 r.), przy czym jako okres bazowy przyjęto pierwszy miesiąc analizy. Porównanie to zostało wywołane następującą komendą z pakietu:

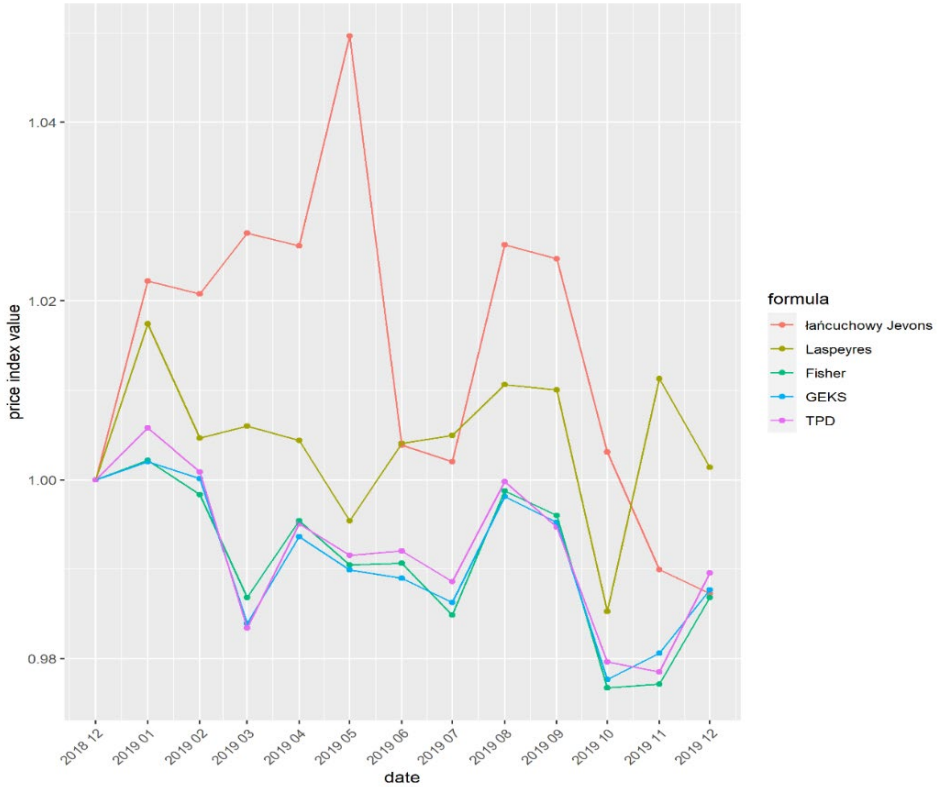
```
compare_indices(milk,
start = "2018-12", end = "2019-12",
bilateral = c("chjevons", "laspeyres", "fisher"),
namebilateral = c("łańcuchowy Jevons", "Laspeyres", "Fisher"),
fbmulti = c("geks", "tpd"),
fbwindow = c(13, 13),
namefbmulti = c("GEKS", "TPD"))
```

Wyznaczenie średnich absolutnych różnic pomiędzy porównywanymi indeksami realizowane jest za pomocą komendy:

```
indeksy <- price_indices(milk,
start = "2018-12", end = "2019-12",
bilateral = c("chjevons", "laspeyres", "fisher"),
fbmulti = c("geks", "tpd"),
fbwindow = c(13, 13),
interval = TRUE)
compare_distances(indeksy),
```

w efekcie czego otrzymujemy wyniki wyrażone w punktach procentowych (tabl. 7.1).

Wykr. 7.2. Porównanie wybranych indeksów cen dla zbioru danych *milk* z pakietu PriceIndices



Źródło: opracowanie własne w pakiecie PriceIndices.

Tabl. 7.1. Średnie absolutne różnice pomiędzy indeksami prezentowanymi na wykr. 7.2. [p.proc.]

##		chjevons	laspeyres	fisher	geks	tpd
##	chjevons	0,000	1,712	2,497	2,504	2,407
##	laspeyres	1,712	0,000	1,429	1,429	1,300
##	fisher	2,497	1,429	0,000	0,141	0,212
##	geks	2,504	1,429	0,141	0,000	0,181
##	tpd	2,407	1,300	0,212	0,181	0,000

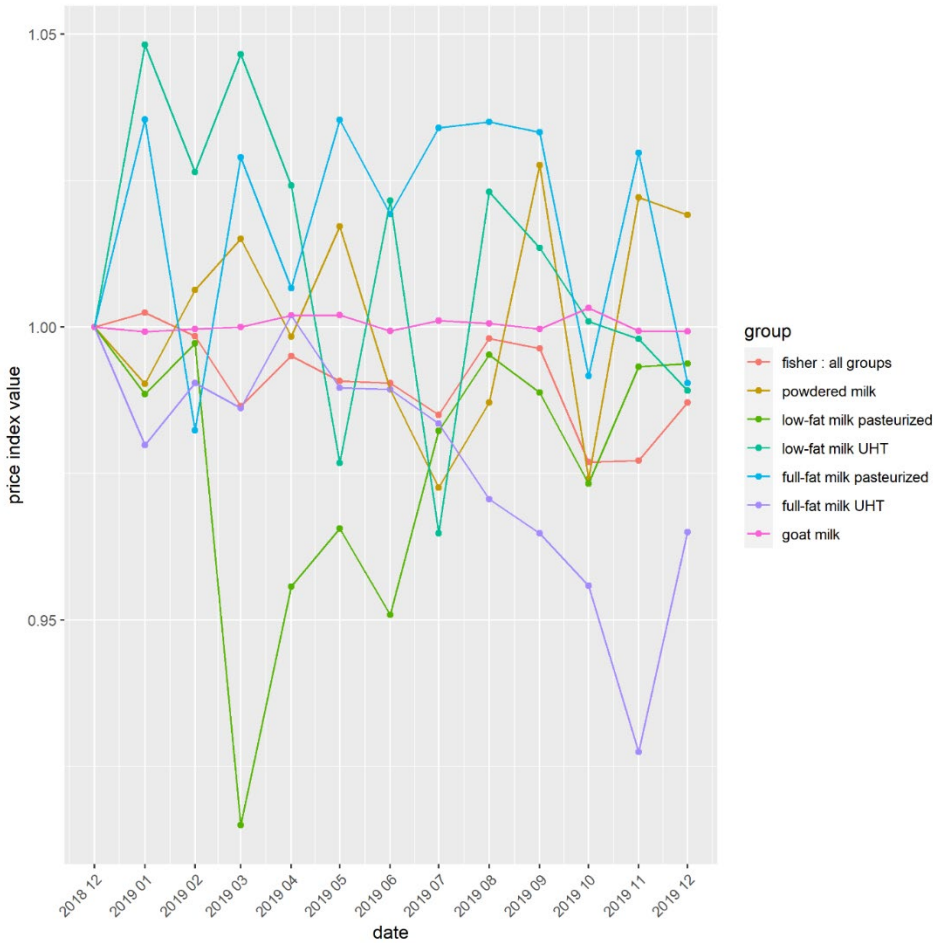
Źródło: opracowanie własne w pakiecie PriceIndices.

Natomiast agregację cząstkowych indeksów Fishera dla podgrup grupy mleko, przy zastosowaniu agregującej formuły Laspeyresa, można uzyskać za pomocą komendy (por. wykr. 7.3):

final_index2(**data** = milk,
by = "description",
all = TRUE,

start = "2018-12", **end** = "2019-12",
formula = "fisher",
interval = TRUE,
aggrsets = "none",
aggrret = "laspeyres",
figure = TRUE)

Wykr. 7.3. Porównanie indeksu Fishera dla różnych rodzajów mleka (podgrupy grupy *milk* z pakietu PriceIndices) z indeksem Fishera dla całego zbioru *milk* (agregacja według Laspeyresa)



Źródło: opracowanie własne w pakiecie PriceIndices.

7.4. Doświadczenia zebrane w trakcie realizacji projektu – szanse i zagrożenia

Prace realizowane w ramach projektu miały na celu przede wszystkim wdrożenie do badań cen nowych źródeł danych pozyskiwanych bezpośrednio z sieci handlowych oraz stron internetowych. Proces zbierania danych ze stron internetowych jest automatyczny i zasadniczo samo zebranie danych przy użyciu odpowiednich narzędzi informatycznych powinno być stosunkowo proste. Również wygenerowanie zbiorów z systemów informatycznych przedsiębiorstw posiadających duże sklepy (z sieci handlowych) nie jest zbyt skomplikowane przy zastosowaniu odpowiedniej aplikacji. Takie zebranie danych umożliwia stopniowe ograniczanie pracy ankieterów statystycznych, którzy dotychczas przeprowadzali notowania w sklepach o różnej wielkości (zarówno dużych, jak i mniejszych punktach detalicznych, a także na targowiskach). Należy zaznaczyć, że notowania cen przez ankieterów bezpośrednio w punktach sprzedaży były tradycyjnie podstawowym źródłem danych do obliczeń wskaźników obrazujących ruchy cen.

Pomimo że uzyskanie danych z wymienionych alternatywnych źródeł jest stosunkowo proste i nie jest obarczone dużymi kosztami, to w porównaniu z tradycyjnymi metodami gromadzenia cen ich przetworzenie i podporządkowanie standardom statystyki publicznej (Białek i Beręsewicz, 2021) jest już bardzo pracochłonne i metodologicznie trudne.

Warto zaznaczyć, że sieci handlowe dokonują zmian w swoich systemach informatycznych, a sklepy internetowe w swoich witrynach, co wynika m.in. z wewnętrznej polityki marketingowej. Z tego względu pozyskiwane zbiory mogą mieć w poszczególnych okresach różne struktury, co wymaga zmian w programach komputerowych przetwarzających tego typu dane. W efekcie zmienia się profil kompetencyjny pracowników realizujących badania w obszarze statystyki cen, jak również pojawia się konieczność rozwoju umiejętności w obszarze analitycznym (data science, big data). Duże znaczenie przy zbieraniu danych oraz ich przetwarzaniu ma również stworzenie stałego zespołu informatyków, którzy rozwijaliby oprogramowanie w miarę rozszerzania zakresu pozyskiwania danych z nowych źródeł.

Nowe źródła danych, zwłaszcza skanowanych, są dostępne jedynie dzięki współpracy z różnorodnymi sieciami handlowymi oferującymi produkty z możliwe jak największej liczby segmentów rynku. Aby zebrane dane jak najlepiej odwzorowywały schemat zakupów gospodarstw domowych, niezbędne jest śledzenie zarówno trendów w sprzedaży detalicznej, jak i zmian w infrastrukturze handlowej oraz monitorowanie rozwoju sieci handlowych, w tym sieci franczyzowych. Ponadto ze względu na zróżnicowane rozmieszczenie w skali kraju punktów sprzedaży sieci handlowej oraz specyficzną dla danej sieci politykę cenową korzystanie z dostarczonych przez

sieci zbiorów danych wymaga zmian w metodyce obliczania wskaźników cen w przekroju terytorialnym.

Równocześnie z dotychczasowych doświadczeń wynika, że ze względu na strukturę realizacji zakupów przez konsumentów niezbędne jest uwzględnienie również danych z małych sklepów, w których klienci robią szacunkowo 30% zakupów.

Z tego względu nie jest możliwe całkowite zrezygnowanie z notowań ankieterskich na rzecz informacji uzyskiwanych z nowych źródeł. Utrzymanie notowań ankieterskich pozwoli także na porównanie danych pochodzących z różnych źródeł oraz zapewnienie informacji w przypadku braku danych z jakiegoś kanału dystrybucji lub wybranej sieci.

Istotnym czynnikiem, który może stanowić wyzwanie w badaniach cen za pomocą nowych źródeł danych, jest inne podejście do określenia poziomu ceny. Charakterystyczna dla badań w tym obszarze jest koncentracja na dynamice, a nie na poziomach cen, co nie zawsze spełnia oczekiwania odbiorców informacji. W przypadku badań opartych na danych skanowanych i skrapowanych najczęściej zamiast poziomów cen wykorzystuje się kategorię ceny jednostkowej (ang. *unit value*) uzyskanej jako informacja pochodna wartości i ilości sprzedanych towarów.

Należy podkreślić, że wdrożenie nowych źródeł danych jest możliwe jedynie przy założeniu utrzymania współpracy z sieciami handlowymi oraz rozszerzania ich listy. Działania w tym zakresie, które były realizowane w ramach prac projektowych, wskazywały, że nie jest łatwo nakłonić sieci handlowe do współpracy, nawet przy uwzględnieniu odpowiednich zapisów legislacyjnych. Praktycznie nie ma narzędzi, które mogłyby zmotywować firmy prywatne do współpracy na rzecz statystyki publicznej. Zagrożeniem dla kontynuacji badań z wykorzystaniem nowych źródeł danych może być również ograniczona możliwość kontroli danych przekazywanych przez sieci handlowe. Z uwagi na krótki okres, w jakim jest realizowany proces obliczeń, jedynym rozwiązaniem jest automatyczna kontrola danych. Taka kontrola jest utrudniona m.in. ze względu na zróżnicowany zakres danych transferowanych do GUS przez poszczególne sieci (zakres danych jest negocjowany indywidualnie).

Jednak mimo tego, że proces włączania nowych źródeł danych do badania cen detalicznych musi być stopniowy i jest uwarunkowany postępem w pracach nad metodyką badania i jego organizacją, niewątpliwie wpływ tych źródeł na podniesienie jakości badania jest ogromny. Wynika to z uwzględnienia w obliczeniach danych o cenach różnych kategorii produktów dostępnych na rynku. Dużą zaletą alternatywnych źródeł jest również fakt, że pozwalają one na bieżące śledzenie trendów zakupowych oraz na włączanie do badania nowych produktów już w momencie ich pojawienia się w sprzedaży. Wskazane zalety stanowią przesłankę do kontynuacji prac w tym obszarze. W tym celu będzie jednak konieczne rozwijanie automatycznych procesów przetwarzania danych i tworzenie zespołów specjalistów, którzy stale

będą podnosić swoje kompetencje w zakresie obsługi informatycznej oraz analizy danych. A przede wszystkim niezbędne jest stworzenie procedur pozwalających na nawiązywanie i utrzymywanie ciągłej współpracy z sieciami handlowymi i zabezpieczenie statystyki przed odmową przekazania danych.

Podsumowując doświadczenia zebrane podczas realizacji projektu, należy podkreślić, że wyniki przeprowadzonych badań eksperymentalnych świadczą o ogromnym potencjale tkwiącym w nowych źródłach danych, który może w przyszłości zrewolucjonizować statystyczne badania cen. Trzeba mieć przy tym świadomość zarówno ograniczeń każdego z tych źródeł, jak i możliwości nowoczesnych technologii zbierania i przetwarzania danych. Jednym z nich jest reprezentatywność danych skrapowanych i skanowanych ograniczona tylko do dużych sklepów i supermarketów. Inną barierą w przypadku cen skrapowanych jest brak informacji, czy produkty faktycznie się sprzedały po tych cenach. Jeśli je sprzedano, to nie wiadomo, jaka była wielkość tej sprzedaży. Wykorzystanie tego źródła danych o cenach będzie więc musiało być wspomagane informacjami z innych statystyk, takich jak raporty i sprawozdania dotyczące sprzedaży detalicznej.

Ponadto ceny zebrane z nowych źródeł danych mogą różnić się od cen w mniejszych, osiedlowych sklepach, nieprowadzących sprzedaży w internecie. Jest więc konieczne zachowanie proporcji co do ilości zebranych poziomów cen metodą tradycyjną (przez ankieterów w sklepach) i za pośrednictwem skrapowania oraz transmisji danych z sieci handlowych, tak aby dane uzyskane z alternatywnych źródeł nie wpłynęły negatywnie na średnie ceny w badaniu: nie zaniżyły ich lub zawyżyły w sposób nieuzasadniony wielkością zakupów dokonywanych przez konsumentów w poszczególnych kanałach dystrybucji. Dotychczasowe doświadczenia na podstawie eksperymentalnych obliczeń przeprowadzonych podczas udziału w projekcie pokazały, że takie potencjalne zagrożenie może zaistnieć, choć przyczyną zawyżania lub zaniżania cen w relacji do danych zebranych metodą tradycyjną mógł być zbyt krótki szereg czasowy w badaniu eksperymentalnym (por. Białek i in., 2021).

Dalsze badanie problemu reprezentatywności danych ze źródeł alternatywnych jest istotne m.in. z punktu widzenia postrzegania danych o inflacji przez użytkowników danych statystycznych. Na konieczność odpowiedniego, zrozumiałego dla powszechnego odbiorcy danych, upowszechniania informacji o zmianach metodologicznych i ich wpływie na wielkość wskaźnika inflacji zwracają uwagę instytucje międzynarodowe. Podczas realizacji projektu opracowano m.in. publikację *Co warto wiedzieć o inflacji,* opisującą metodykę badania cen konsumpcyjnych, która dostępna jest na stronie GUS. Pełne wdrożenie alternatywnych źródeł danych do obliczania CPI i ich przetwarzanie przy wykorzystaniu nowoczesnych technologii będzie wymagało przygotowania tego typu publikacji lub serii artykułów wyjaśniających rolę, znaczenie i sposób wykorzystania nowych źródeł danych do obliczania wskaźników

cen konsumpcyjnych. Zarówno przeciętny odbiorca danych, jak i profesjonalni użytkownicy informacji o inflacji muszą mieć niepodważalne zaufanie do oficjalnych wskaźników cen konsumpcyjnych ogłaszanych przez statystykę publiczną. Stąd niezmiernie ważnym wyzwaniem jest precyzyjne opracowanie metadanych opisujących nowe zasady metodyczne obliczania inflacji.

Całkowite przejście na automatyczne zbieranie danych (skanowanych i skrapowanych) na potrzeby obliczania wskaźników cen w zakresie niektórych kategorii wydatków konsumenckich może spowodować zagrożenie ich pozyskania w razie awarii lub innych przerw w dostępie do internetu albo poszczególnych stron internetowych. Trzeba mieć świadomość, że dysponowanie pełnym zbiorem danych o cenach według harmonogramu opracowania wskaźników cen towarów i usług konsumpcyjnych jest warunkiem absolutnie koniecznym do ich terminowej publikacji. Należy również zdawać sobie sprawę z potencjalnego zagrożenia pobrania zawirusowanego oprogramowania podczas transmisji danych z sieci handlowych lub skrapowania. Niezmiernie ważna jest tu konieczność zapewnienia odpowiedniego zabezpieczenia przez pracowników IT. Dlatego implementacja nowoczesnych technologii pozyskiwania i przetwarzania danych wymaga m.in. organizacji i utrzymania zespołu specjalistów stale doskonalących swoje kompetencje w zakresie obsługi informatycznej oraz analizy danych statystycznych.

W trakcie prac projektowych pojawiła się konieczność zapewnienia odpowiednich rozwiązań prawnych w kwestii gromadzenia danych skrapowanych (w tym: czy i jak zawiadamiać sklepy internetowe, że ich dane będą ściągane i wykorzystywane w badaniu cen) oraz wypracowania prawnej formuły współpracy z sieciami handlowymi odnośnie do transferu danych skanowanych. Doświadczenia innych krajów w tym zakresie różnią się od siebie. Niektóre urzędy statystyczne zawierają umowy z sieciami, a inne płacą wyspecjalizowanym firmom (jak Nielsen, GfK) za przekazywane im dane. Wynika to głównie z różnego usytuowania prawnego statystyki w poszczególnych krajach. Wśród wskazówek i rekomendacji instytucji międzynarodowych można znaleźć przykładowe wzory takich umów. W GUS, mając na uwadze obowiązek statystyczny w zakresie przekazywania danych na potrzeby statystyki publicznej nakładany na podmioty gospodarcze ustawą o statystyce oraz programem badań statystycznych corocznie wprowadzanym rozporządzeniem Rady Ministrów, po konsultacjach z prawnikami zdecydowano o uzgadnianiu (odrębnie z każdą siecią) i podpisywaniu dokumentu w formie porozumienia, które zawiera szczegółowo opisane zasady transferu oraz upewnia sieć o zachowaniu całkowitej poufności przekazywanych do GUS danych. W miarę rozwoju znaczenia źródeł alternatywnych w pomiarze inflacji i włączania do badania wielu kolejnych sieci handlowych będzie niezbędne wypracowanie takich rozwiązań prawnych, które spełniałyby oczekiwania sieci co do bezpieczeństwa i poufności transferowanych danych oraz upraszczałyby

procedurę transferu i określałyby jego standardowe ramy, a w konsekwencji – zachęcały sieci do tego rodzaju współpracy ze statystyką.

Ważnym metodologicznym wyzwaniem pozostaje doskonalenie metody integracji danych z różnych źródeł. Doświadczenia zebrane podczas realizacji projektu zaowocowały zapewnieniem możliwości wykorzystywania danych skanowanych przekazywanych przez sieci handlowe (te, które udało się nakłonić do współpracy) i skrapowanych, jednak dotyczy to tylko niektórych grup wydatków i jedynie danych o cenach. Pozostaje rozwijanie metodologii w kierunku rozszerzenia tej metody gromadzenia danych na kolejne kategorie ECOICOP oraz szerszego włączenia do bieżących obliczeń danych o ilości sprzedaży produktów. Projekt wykazał, że trzeba wypracować formuły pełnego wykorzystania informacji o bieżącej wielkości sprzedaży produktów do obliczania średnich cen oraz opracowania systemów wag, tak aby spełniały zarówno standardy przyjęte w polskiej statystyce, jak i wymogi statystyki międzynarodowej.

Niezbędne jest stałe monitorowanie, czy wprowadzane modyfikacje dotyczące metodyki i organizacji gromadzenia oraz przetwarzania danych o cenach detalicznych są zgodne z prawodawstwem UE w zakresie zharmonizowanych wskaźników cen konsumpcyjnych, do których stosowania jest zobowiązana polska statystyka. Podstawą do obliczania obydwu wskaźników są te same zbiory danych o poziomach cen. W trakcie realizacji projektu stało się też jasne, że dalsze prace powinny być prowadzone w celu określenia, na ile może lub powinna różnić się metodyka obliczania krajowego CPI od zasad opracowywania HICP.

Konieczne jest kontynuowanie śledzenia rozwoju metodologii międzynarodowej w zakresie wykorzystania big data w obliczaniu wskaźników cen konsumpcyjnych. Istotnym elementem tej metodologii jest zastosowanie wskaźników multilateralnych. Decyzja o wyborze algorytmu obliczania wskaźników na poziomie grup elementarnych spośród formuł wskaźników multilateralnych może być kluczowa dla zmiany metodyki opracowywania wskaźników cen konsumpcyjnych opartej przez wiele dekad głównie na formule Jevonsa i – na wyższych poziomach agregacji – Laspeyresa. Udział w tym projekcie dał możliwość uzyskania wyników badań eksperymentalnych w tym zakresie, które – prezentowane na międzynarodowych konferencjach – stanowią istotny wkład polskiej statystyki w dalszy rozwój nowoczesnych technologii w pomiarze inflacji.

Bibliografia

Aggarwal, C. C. (2005). *On Learning Strategies for Topic Specific Web Crawling*. http://charuaggarwal.net/collab.pdf.

Agre, G., Dongre, S. (2015). A Keyword Focused Web Crawler Using Domain Engineering and Ontology. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(3). https://doi.org/10.17148/IJARCCE.2015.43111.

Australian Bureau of Statistics. (2016). *Making Greater Use of Transactions Data to Compile the Consumer Price Index* (ABS Information Paper No. 6401.0.60.003). https://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/FE1AE4B7443728E5CA258079000EAF99/$File/6401060003_2016.pdf.

Ayoubkhani, D., Heledd, T. (2022). Estimating Weights for Web-Scraped Data in Consumer Price Indices. *Journal of Official Statistics*, *38*(1), 5–21. https://doi.org/10.2478/jos-2022-0002.

Balk, B. M. (1995). Axiomatic Price Index Theory: A Survey. *International Statistical Review*, *63*(1), 69–95. https://doi.org/10.2307/1403778.

Baumgartner, R., Frölich, O., Gottlob, G. (2007). The Lixto systems applications in business intelligence and semantic web. In E. Franconi, M. Kifer, W. May (Eds.), *The Semantic Web: Research and Applications*, LNCS (vol. 4519), (s. 16–26). Springer Berlin Heidelberg.

Baumgartner, R., Frölich, O., Gottlob, O., Harz, P., Herzog, M., Lehmann, P. (2005). Web Data Extraction for Business Intelligence: the Lixto Approach. W: Vossen, G., Leymann, F., Lockemann, P. Stucky, W. (Eds.), *Datenbanksysteme in Business, Technologie und Web (BTW), 11. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme"* (s. 48–65). Gesellschaft für Informatik. https://dl.gi.de/bitstream/handle/20.500.12116/28294/GI-Proceedings.65-4.pdf?sequence=1&isAllowed=y.

Białek, J. (w druku) .Scanner data processing in a newest version of the PriceIndices package. *Statistical Journal of the IAOS*.

Białek, J. (2017). Approximation of the Fisher price index by using Lowe, Young and AG Mean indices. *Communications in Statistics – Simulation and Computation*, *46*(8), 6454–6467. https://doi.org/10.1080/03610918.2016.1205608.

Białek, J. (2020). Comparison of elementary price indices. *Communications in Statistics – Theory and Methods*, *49*(19), 4787–4803. https://doi.org/10.1080/03610926.2019.1609035.

Białek, J. (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations. *Statistika. Statistics and Economy Journal*, (2), 122–141. https://www.czso.cz/documents/10180/143550797/32019721q2_bialek.pdf/3cd5bf11-22f4-4ee5-b294-1d7d5909e4b4?version=1.1.

Białek, J. (2022a). Elementary price indices under the GBM price model. *Communications in Statistics – Theory and Methods*, *51*(5), 1232–1251. https://doi.org/10.1080/03610926.2021.1938127.

Białek, J. (2022b). *The general class of multilateral indices and its two special cases*. 17th Meeting of the Ottawa Group on Price Indices, Rome.

Białek, J., Beręsewicz, M. (2021). Scanner data in inflation measurement: From raw data to price indices. *Statistical Journal of the IAOS*, *37*(4), 1315–1336. https://doi.org/10.3233/SJI-210816.

Białek, J., Bobel, A. (2019). *Comparison of Price Index Methods for CPI Measurement Using Scanner Data*. 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro.

Białek, J., Panek, T., Zwierzchowski, J. (w druku). Uncertainty and sensitivity analysis of the CPI based on scanner and scraped data. *Statistics in Transition new series*.

Bishop, Ch. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

Black, P. E. (2004). *Ratcliff/obershelp pattern recognition*.

Boldi, P., Andrea, M., Santini, M., Vigna, S. (2018). BUbiNG: Massive crawling for the masses. *ACM Transations on the Web*, *12*(2). https://doi.org/10.1145/3160017.

Bonato. A. (2008). *A Course on Web Graph. Graduate Studies in Mathematics* (vol. 89). American Mathematical Society.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, *33*(1-6), 309–320. https://doi.org/10.1016/S1389-1286(00)00083-9.

Busłowska, .E. (2013). Metody zabezpieczania logistycznych baz danych. *Logistyka*, *6*, 67–71.**Błąd! Nieprawidłowy odsyłacz typu hiperłącze.**

Carli, G. R. (1804). Del valore e della proporzione de'metalli monetati. W: P. *Scrittori Classici Italiani di Economia Politica. Parte Moderna* (vol. 13) (s. 297–336). Destefanis.

Castano, S., Fugini, M. G., Martella, G., Samarati,. P. (1995). *Database security*. ACM Press.

Cavallo. A. (2017). Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers. *American Economic Review*, *107*(1), 283–303. https://doi.org/10.1257/aer.20160542.

Cavallo. A. (2018). Scraped data and sticky prices. *The Review of Economics and Statistics*, *100*(1), 105–119. https://doi.org/10.1162/REST_a_00652.

Cavallo, A., Erwin Diewert, W., Feenstra, R. C., Inklaar, R., Timmer, M. P. (2018). Using Online Prices for Measuring Real Consumption across Countries. *AEA Papers and Proceedings*, *108*, 483–487. https://doi.org/10.1257/pandp.20181037.

Cavallo, A., Rigobon, R.. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, 30(2—Spring), 151–178. https://doi.org/10.1257/jep.30.2.151.

Chan, K., Tarantola, S., Saltelli, A., Sobol, I. M. (2000). Variance based methods. W: A. Saltelli, K. Chan, E. M. Scott (Eds.), *Sensitivity analysis* (s. 167–197). John Wiley & Sons.

Chang, C.-H., Kayed, M., Girgis, M. R., Shaalan, K. F. (2006). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1411–1428. https://doi.org/10.1109/TKDE.2006.152.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li. Y. (b.r.). *xgboost: Extreme Gradient Boosting. R package version 1.3.2.1*. https://CRAN.R-project.org/package=xgboost.

Chessa, A. G. (2015). *Towards a generic price index method for scanner data in the Dutch CPI*. Ottawa Group Meeting, Urayasu City.

Chessa, A. (2016). *A new methodology for processing scanner data in the Dutch CPI*. Meeting of the Group of Experts on Consumer Price Indices, Geneva.

Chessa, A. G. (2019). *A comparison of index extension methods for multilateral methods*. 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro.

Chessa, A. G. (2021). A Product Match Adjusted R Squared Method for Defining Products with Transaction Data. *Journal of Official Statistics*, *37*(2), 411–432. https://doi.org/10.2478/jos-2021-0018.

Czerski, D., Rąkoski, R., Borkowski, P. (2022). *Dokumentacja techniczna oprogramowania, klasyfikacja oraz kojarzenie produktów, zbieranie cen z Internetu oraz obliczenia wskaźników cen. projekt gospostrateg/instat-ceny. v. 3.1.1.*

Czerwiński, A. (2019). Ważność kryteriów oceny wiarygodności witryn internetowych na podstawie badań. *e-mentor*, (4), 39–46. https:https://doi.org/10.15219/em81.1433.

Deng, L., Cox, L. P. (2009). *LiveCompare: grocery bargain hunting through participatory sensing*. HotMobile '09: 10th workshop on Mobile Computing Systems and Applications, Santa Cruz.

Dilmegani, C. (2020, 26 grudnia). *Top 18 Web Scraping Applications & Use Cases in 2022*. https://research.aimultiple.com/web-scraping-applications/.

Dong, Y. F., Kanhere, S., Chou, C. T., Bulusu, N. (2008). Automatic Collection of Fuel Prices from a Network of Mobile Cameras. *Distributed Computing in Sensor Systems, 4th IEEE International Conerence, Santorini Island*.

Dong, Y., Kanhere, S., Chou, C.-T., Liu, R. (2011). *Automatic image capturing and processing for PetrolWatch*. 17th IEEE International Conference on Networks, Washington.

Diewert, W. E., Fox, K. J. (2018). *Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data* (UNSW Economics Working Paper No. 2018-13). http://dx.doi.org/10.2139/ssrn.3276457.

Diewert, W. E., Silver, M. (2008). Elementary indices. W: M. Silver (Ed.), *Export and Import Price Index Mannual. Theory and Practice*. International Monetary Fund. https://www.imf.org/external/np/sta/tegeipi/ch21.pdf.

Drobisch, M. W. (1871). Ueber einige Einwürfe gegen die in diesen Jahrbüchern veröffentlichte neue Methode, die Veränderungen der Waarenpreise und des Geldwerths zu berechten. *Jahrbücher für Nationalökonomie und Statistik*, *16*(1), 420–431. https://doi.org/10.1515/jbnst-1871-0134.

Dutot, C. (1738). *Reflexions Politiques sur les Finances et le Commerce*. La Haye, Chez les Freres Vaillant & Nicolas Prevost.

Eltetö, Ö., Köves, P. (1964). Egy nemzetközi összehasonlításoknál fellépő indexszámítási. *Statisztikai Szemle*, *42*(10), 507–518. https://www.ksh.hu/statszemle_archive/all/1964/1964_05/1964_05_0507_0518.pdf.

European Union. (2021). *Eurostat review of National Accounts and Macroeconomic Indicators*, 1, 49–69. https://ec.europa.eu/eurostat/documents/3888793/13018876/KS-GP-21-001-EN-N.pdf/32280440-89b3-49a4-415b-00dba0ed818f?t=1630504343578.

Feenstra, R. C., Shapiro, M. D. (Eds.) (2000). *Scanner Data and Price Indexes*. University of Chicago Press. https://www.nber.org/books-and-chapters/scanner-data-and-price-indexes.

Fisher, I. (1922). *The Making of Index Numbers. A Study of Their Varieties, Tests, and Reliability*. Houghton Mifflin.

Fong, S., Siu, S., Sun, A. (2002). *WebSpy: Retrieving Web Contents for e-Business Intelligence*. The First International Conference on Information Technology and Applications (ICITA 2002), Bathurst.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7.

Geary, R. G. (1958). A Note on the Comparisons of Exchange Rates and Purchasing Power Between Countries. *Journal of the Royal Statistical Society. Series A*, *121*(1), 97–99. https://doi.org/10.2307/2342991.

Gini, C. (1931). On the circular test of index numbers. *Metron*, *9*(9), 3–24.

de Haan, J. (2015). A framework for large scale use of scanner Data in the Dutch CPI. *Report from 14th Ottawa Group meeting on Price Indices, Tokyo*.

de Haan, J., van der Grient, H. A. (2011). Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics*, *161*(1), 36–46. https://doi.org/10.1016/j.jeconom.2010.09.004.

de Haan, J., Krsinich, F. (2018). Time Dummy Hedonic and Quality-Adjusted Unit Value Indexes: Do They Really Differ?. *Review of Income and Wealth*, *64*(4), 757–776. https://doi.org/10.1111/roiw.12304.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer. https://doi.org/10.1007/978-0-387-21606-5.

Hiremath. O. S. (2022, 9 kwietnia). *A Beginner's Guide to learn web scraping with python!*. https://www.edureka.co/blog/web-scraping-with-python/.

IBM. (b.r.). *IBM Security Guardium solutions*. https://www.ibm.com/security/data-security/guardium.

International Labour Organization, International Monetary Fund, Organisation for Economic Co-operation and Development, Statistical Office of the European Communities, United Nations, The International Bank for Reconstruction and Development, The World Bank. (2004). *Consumer Price Index Manual. Theory and practice*. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf.

Jevons, W. S. (1865). On the Variation of Prices and the Value of the Currency since 1782. *Journal of the Statistical Society of London*, *28*(2), 294–320. https://doi.org/10.2307/23 38419.

Kasprowski, P., Kozielski, S. Kuźniacki, P., Pietraszek, T. (2003). Bezpieczeństwo systemów bazodanowych dostępnych przez Internet. W: A. Grzywak. (red.), *Internet w społeczeństwie informacyjnym* (s. 33–65). Wyższa Szkoła Biznesu w Dąbrowie Górniczej.

Kausar, M. A., Dhaka, V. S., Singh, S. K. (2013). Web Crawler: A Review. *International Journal of Computer Applications*, *63*(2), 31–36. https://doi.org/10.5120/10440-5125.

Khamis, S. H. (1970). Properties and Conditions for the Existence of a New Type of Index Number. *Sankhyā: The Indian Journal of Statistics. Series B*, *32*(1/2), 81–98.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing & Its Applications*, *13*(3), 144–169. https://doi.org/10.15849/IJASCA.211128.11.

Kibriya, A. M., Frank, E. Pfahringer, B., Holmes, G. (2004). Multinomial Naive Bayes for Text Categorization Revisited. W: G. I. Webb, X. Yu (Eds.), *AI 2004: Advances in Artificial Intelligence* (vol. 3339) (s. 488–499). Springer. https://doi.org/10.1007/978-3-540-30549-1_43.

Kireyev, P., Kumar, V., Ofek, E. (2017). Match Your Own Price? Self-Matching as a Retailer's Multichannel Pricing Strategy. *Marketing Science*, *36*(6), 908–930. https://doi.org/10.1287/mksc.2017.1035.

Kłopotek, M. A., Wierzchoń, S. T., Ciesielski, K., Dramiński, M., Czerski, D. (2007). *Conceptual Maps of Document Collections in Internet and Intranet. Coping with the Technological Challenge*. Instytut Podstaw Informatyki PAN.

Kobayashi, M., Takeda, K. (2000). Information Retrieval on the Web. *ACM Computing Surveys*, *32*(2), 144–173. https://doi.org/10.1145/358923.358934.

Koronacki J., Ćwik, J. (2005). *Statystyczne systemy uczące się*. Wydawnictwa Naukowo-Techniczne.

Krsinich, F. (2014). *The FEWS index: fixed effects with a window splice – non-revisable quality-adjusted price indixes with no characteristic information*. Meeting of the group of experts on consumer price indices, Genewa.

Kumar, R., Jain, A., Agrawal, C. (2016). Survey of web crawling algorithms. *Advances in Vision Computing: An International Journal (AVC)*, *3*(3), 1–7. https://doi.org/10.5121/avc.2016.3301.

Lamboray, C. (2017). *The Geary Khamis index and the Lehr index: how much do they differ?*. 15th meeting of the Ottawa Group, Eltville.

Langville, A. N., Meyer, C. D. (2006). *Google's PageRank and Beyond. The Science of Search Engine Rankings*. Princeton University Press.

Laspeyres E. (1871). Die Berechnung einer mittleren Waaren-preissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, *16*(1), 296–314. https://doi.org/10.1515/jbnst-1871-0124.

Levell, P. (2015). Is the Carli index flawed?: assessing the case for the new retail price index RPIJ. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *178*(2), 303–336.

von der Lippe, P. (2007). *Index Theory and Price Statistics*. Peter Lang.

Van Loon, K., Roels, D. (2018). *Integrating big data in the Belgian CPI*. Meeting of the Group of Experts on Consumer Price Indices, Geneva.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071.

Marshall, A. (1887). Remedies for Fluctuations of General Prices. *Contemporary Review*, *51*, 355–375.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*, 1–16. https://doi.org/10.1186/1471-2105-10-213.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S. (2011). *Tools for Composite Indicators Building*. Dictus Publishing.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2005). Handbook on Constructing Composite Indicators: Methodology and User Guide (OECD Statistics Working Paper no. 2005/03). https://doi.org/10.1787/533411815016.

Oracle. (b.r.). *Oracle Audit Vault and Database Firewall*. https://www.oracle.com/database/technologies/security/audit-vault-firewall.html.

Organisation for Economic Co-operation and Development. (2008). *Handbook on Constructing Composite Indicators. Methodology and user guide*. https://www.oecd.org/sdd/42495745.pdf.

Paasche, H. (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, *23*(2/4), 168–178.

Pande, V., Singh, P. (2015). Focused and Adaptive Crawling for Topic Specific and Hidden Web Entries. *International Journal of Science and Research (IJSR)*, *4*(12), 2212–2215. https://www.ijsr.net/get_abstract.php?paper_id=NOV152532.

Panek, T. (2016). *Jakość życia od koncepcji do pomiaru*. Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1.

Saltelli, A., Chan K., Scott, E. M. (Eds.) (2000). *Sensitivity Analysis*. John Wiley & Sons.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. (2008). *Global sensitivity analysis. The primer*. John Wiley & Sons.

Sarawagi, S. (2007). Information Extraction. *Foundations and Trends® in Databases*, *1*(3), 261–377. https://doi.org/10.1561/1900000003.

Sharpe, A., Salzman, J. (2004). *Methodological choices encountered in the construction of composite indices of economic and social well-being*. Center for the Study of Living Standards.

Silver, M., Heravi, S. (2007). Why elementary price index number formulas differ: Evidence on price dispersion. *Journal of Econometrics*, *140*(2), 874–883. https://doi.org/10.1016/j.jeconom.2006.07.017.

Sobol, I. M. (1993). Sensitivity Estimates for Non-linear mathematical Models. *Mathematical Modelling and Computational Experiment*, *1*(4), 407–414.

Szulc, B. (1964). Indices for multiregional comparisons. *Przegląd Statystyczny*, (3), 239–254.

United Nations Economic Commission for Europe. (2021). *Machine Learning for Official Statistics*. United Nations. https://unece.org/sites/default/files/2022-01/ECECESSTAT20216.pdf.

Walsh, C. M. (1901). *The Measurement of General Exchange-Value*. The Macmillan Company.

Zhang, L.-C., Johansen, I., Nyagaard, R. (2019). Tests for Price Indices in a Dynamic Item Universe. *Journal of Official Statistics*, *35*(3), 683–697. https://doi.org/10.2478/jos-2019-0028.