STATBEL
België in cijfers

Het gebruik van scannerdata van supermarkten in de consumptieprijsindex

- Ken Van Loon en Dorien Roels -

nr. 01

ANALYSE

12.2017

economie
FOD Economie, K.M.O., Middenstand en Energie

.be

Het gebruik van scannerdata van supermarkten in de consumptieprijsindex

Dorien Roels, Ken Van Loon[1]

[1] Statistici bij Statbel (Algemene Directie Statistiek – Statistics Belgium)

ABSTRACT

Deze analyse licht het gebruik van scannerdata van supermarkten in de consumptieprijsindex toe. Wat zijn scannerdata? Hoe zien deze data er uit? En hoe verwerkt de AD Statistiek – Statistics Belgium (Statbel) deze gegevens?

Sinds 2015 gebruikt Statbel scannerdata in de berekening van de consumptieprijsindex (CPI). Scannerdata zijn kassascangegevens van handelszaken (geaggregeerd tot op productniveau) en zijn naast webscraping een big data bron die gebruikt wordt in de prijsopvolging van de indexkorf. Scannerdata worden gebruikt voor producten die courant aangekocht worden in supermarktketens en vervangen de klassieke prijsopnames via enquêteurs.

Wekelijks ontvangt Statbel gegevens van supermarktketens via een beveiligde transferoverdracht. Deze data bestaan uit zowel productinformatie (productcodes, -omschrijvingen, omzet,…) als de interne classificatie van de ketens. In eerste instantie moeten de producten geklasseerd worden in vooraf bepaalde consumptiesegmenten/productgroepen, welke gelinkt zijn aan de Europese (E)COICOP-nomenclatuur. Via de interne classificatie-gegevens en machine learning krijgt elk 'nieuw' product (lees 'nieuwe productcode') een COICOP-classificatie toegewezen, die eventueel manueel overschreven kan worden.

Met scannerdata wordt niet de geafficheerde prijs (cfr. traditionele prijsopnames), maar de reële gemiddelde transactieprijs geobserveerd. Dit geeft een conceptueel verschil, maar het effect in prijsevolutie blijkt verwaarloosbaar. Daarnaast leert ervaring met scannerdata dat het handiger is om interne productcodes (SKU's) te gebruiken i.p.v. barcodes.

Voor alle supermarktketens worden eerst elementaire indexcijfers (zonder gewichten) berekend op productniveau (COICOP 6). Per niveau worden de representatieve producten (=steekproef) bepaald a.d.h.v. een omzetdrempel, dit zorgt voor een dynamische korf. Indien een product niet wordt opgenomen in de steekproef, worden prijsimputaties berekend. Omwille van statistische en economische redenen worden de elementaire indexcijfers berekend met een Jevons-index (= meetkundig gemiddelde). Hoewel met het gebruik van scannerdata ook omzetinformatie tot op productniveau beschikbaar is, wordt op het elementaire niveau niet gewerkt met gewogen indices omwille van chain drift. De omzetcijfers (per consumptiesegment en per keten) worden nadien wel gebruikt om de indices te aggregeren naar hogere niveaus.

INHOUD

Het gebruik van scannerdata van supermarkten in de consumptieprijsindex 1

Abstract 2

Inhoud 3

1. Inleiding 4

2. Definitie en toepassing 5

2.1. Wat zijn scannerdata of kassascangegevens ? 5

2.2. Voor welke productgroepen worden scannerdata gebruikt? 5

2.3. Hoe ziet scannerdata er uit? 6

3. Methodologie 8

3.1. Classificatie 8

3.1.1. Scannerdata en COICOP-classificatie: opstartfase 9

3.1.2. Scannerdata en COICOP-classificatie: recurrent 10

3.1.3. Machine learning 10

3.2. Prijsconcept 11

3.2.1. Eenheidsprijzen (of unit values) 11

3.2.2. Aggregatie van prijzen over verschillende tijdsperioden 13

3.2.3. Aggregatie van prijzen over verschillende winkels 15

3.2.4. Interne code in plaats van barcode 16

3.3. Sampling en aggregatie 17

3.3.1. Berekening van elementaire indexcijfers 17

3.3.2. Keuze voor een Jevons index 22

3.3.3. Kettingindex en matched modelling 23

3.3.4. Waarom ongewogen indexcijfers? 24

3.3.5. Dumping- en outlierfiltering 27

3.4. Rebranding en replacements 27

3.5. Seizoensproducten 30

4. Stratificatiemodel 32

5. Conclusie 35

Bijlage 36

1. INLEIDING

De consumptieprijsindex (CPI) is een maandelijkse statistiek die opgemaakt wordt door de AD Statistiek – Statistics Belgium van de FOD Economie (Statbel). Het is een economische indicator die de prijsevolutie meet van consumptieve bestedingen van Belgische consumenten. Het is de belangrijkste maatstaf voor inflatie. De CPI vormt in België via de gezondheidsindex en de afgevlakte index rechtstreeks de basis voor de indexering van pensioenen, sociale uitkeringen, fiscale barema's, huur en sommige lonen en wedden.

De CPI wordt berekend op basis van een korf van goederen en diensten die door de gezinnen worden aangekocht en als representatief voor hun verbruiksgewoonten kan worden beschouwd. Aangezien het aanbod van goederen en diensten continu verandert, wijzigt ook de steekproef van de opgenomen prijzen regelmatig. Momenteel worden prijzen van goederen en diensten opgevolgd voor 229 productcategorieën.

De prijsopvolging gebeurt op basis van verschillende databronnen. Zo worden prijzen ingezameld door enquêteurs die winkels bezoeken verspreid over het land. De gegevensinzameling voor de huurenquête gebeurt via papier of web. De prijzen met het grootste gewichtsaandeel worden echter centraal ingezameld via websites, catalogussen, telefoon of via bestanden verkregen van regulatoren of privébedrijven. Recentelijk worden ook meer bigdata-bronnen gebruikt in de berekening van de consumptieprijsindex, namelijk scannerdata van supermarktketens en webscraping.

Naast de nationale consumptieprijsindex (CPI) berekent Statbel ook een Europees geharmoniseerde consumentenprijsindex (Harmonized Index of Consumer Prices, HICP). De HICP maakt een vergelijking tussen de inflatiegraden van de lidstaten van de Europese Unie mogelijk. De toegepaste bestedingsoptiek en methoden zijn daartoe gecoördineerd en in Europese regelgeving vastgelegd. De uitkomsten van de CPI en de HICP zijn echter niet gelijk. Dat komt vooral door een andere weging en samenstelling van het pakket goederen en diensten waarop deze indices zijn gebaseerd.

Dit artikel geeft meer inzicht in de aanwending van scannerdata, een databron die sinds 2015 gebruikt wordt in de berekening van de CPI en vanaf 2016 in de HICP. Het hele productieproces wordt stap per stap beschreven:

- welke scangegevens zijn er en wat is de datastructuur?
- classificatie van scannerdata: hoe kunnen producten worden ondergebracht in de correcte productcategoriehiërarchie (de COICOIP-nomenclatuur[2]) van de CPI?
- hoe worden de indexcijfers berekend op basis van die data?
- misverstanden over het gebruik van deze gegevens in de berekening van de CPI;
- detail van de berekeningswijze vanaf productniveau tot op het geaggregeerde niveau;
- het stratificatiemodel dat toelaat om de resulterende indexcijfers te integreren met andere prijzen van andere databronnen;
- voordelen van het gebruik van scannerdata ten opzichte van de vroegere methode.

[2] Classification of Individual Consumption by Purpose, een classificatie van de consumptieve bestedingen ontwikkeld door de Verenigde Naties (UNSD – United Nations Statistics Division).

2. DEFINITIE EN TOEPASSING

2.1. Wat zijn scannerdata of kassascangegevens ?

Eurostat definieert scannerdata als:

Transaction data obtained from retail chains containing data on turnover, quantities per item code based on transactions for a given period and from which unit value prices can be derived at item code level[3].

Met scannerdata worden dus de kassascangegevens van handelszaken bedoeld, geaggregeerd tot op productniveau. Het zijn niet de individuele kassatickets per klant. Momenteel ontvangt Statbel wekelijks van de drie grootste supermarktketens deze geaggregeerde verkoopgegevens (productspecificaties, omzet en prijsinformatie) op individueel productniveau, dus per streepjescode.

De eerste scannerdata werden ontvangen in oktober 2013, samen met historische gegevens vanaf januari 2012. Na één jaar testen werden scannerdata gefaseerd geïntegreerd in de consumptieprijsindex (CPI), met een ruime integratie voor 70 productgroepen in januari 2016 en een verdere uitbreiding met drie productgroepen in 2017. In de Belgische geharmoniseerde consumptieprijsindex (HICP) worden scannerdata sinds 2016 gebruikt voor dezelfde productgroepen als in de CPI.

Statbel is het 5de nationale statistiekbureau dat het gebruik van scannerdata geïmplementeerd heeft in de berekening van de consumptieprijsindices, daarmee is het één van de Europese koplopers in de toepassing van scannerdata. Bovendien zijn de Eurostat-methodes voor het gebruik van scannerdata (deels) gebaseerd op de methodologie die Statbel toepast.

Het is alvast de bedoeling om het aantal ketens dat scannerdata aanlevert uit te breiden, alsook de scannerdata verder uit te bouwen met andere sectoren dan supermarkten zoals bijvoorbeeld kleding of consumentenelektronica.

2.2. Voor welke productgroepen worden scannerdata gebruikt?

Het betreft producten die courant in supermarktketens aangekocht worden. De in totaal 73 productgroepen op basis van scannerdata zijn samen goed voor 23% van het gewicht van de indexkorf in de consumptieprijsindex (tabel 1).

Tabel 1: Productgroepen waarvoor vanaf januari 2017 scannerdata gebruikt worden

COICOP	Omschrijving	Gewicht 2017
01	Voeding en niet-alcoholische dranken	16.4%
02	Alcoholische dranken en tabak	2.5%
05.5.2.2	Kleine toebehoren voor het huis en de tuin	0.3%
05.6.1	Niet-duurzame huishoudartikelen	1.1%
09.3.4.2	Artikelen voor huisdieren	0.7%
09.5.4.1	Papierartikelen	0.1%
09.5.4.9	Materiaal om te schrijven en te tekenen	0.2%
12.1.3	Producten voor lichaamsverzorging	1.7%
	Totaal	23.0%

De COICOP-nomenclatuur is hiërarchisch opgebouwd waardoor de 73 productgroepen worden bekomen. De nomenclatuur is een verdeling van de totale consumptieve bestedingen (niveau 1) in 12 hoofdgroepen (niveau 2) met daaronder nog verschillende subgroepen op 2 onderliggende niveaus (niveau 3 en 4). Op Europees niveau werd een verdere harmonisering tot op niveau 5 geïntroduceerd. Dit laagste ECOICOP-niveau is tevens het publicatieniveau van zowel de CPI als de HICP. Een volledig overzicht van alle productgroepen waar scannerdata gebruikt worden, is terug te vinden in de bijlage. Verder kunnen de lidstaten zelf nog beslissen om lagere niveaus toe te voegen. Statbel maakt gebruik van een 6de COICOP-niveau, de consumptiesegmenten. De Europese COICOP-nomenclatuur (ECOICOP) wordt – naast de CPI – ook gebruikt bij de HICP, de Nationale Rekeningen en het Huishoudbudgetonderzoek.

[3] Wat vertaald kan worden als: transactiedata verkregen van handelszaken die gegevens bevatten over omzet, hoeveelheden per product gebaseerd op transacties voor een bepaalde periode en waaruit eenheidsprijzen kunnen worden afgeleid op code niveau van het product.

Elk consumptiesegment (niveau 6) krijgt een wegingsfactor, waarbij de wegingsfactor van de bovenliggende niveaus gelijk is aan die van de onderliggende niveaus. De 73 productgroepen waarvoor scannerdata gebruikt worden, refereren dan ook naar het aantal productgroepen op het laagste ECOICOP-niveau.

Een voorbeeld van een opsplitsing van consumptieve bestedingen staat in tabel 2. Zo is de groep 'Voeding en niet-alcoholische dranken' één van de 12 hoofdgroepen. Een gedetailleerder niveau is vervolgens 'Voeding', waarbinnen één van de categorieën 'Brood en granen' is. Die categorie wordt vervolgens verder opgesplitst in acht productgroepen.

Tabel 2: Voorbeeld van classificatie van consumptieve bestedingen

COICOP	Benaming	Niveau
0	Totale bestedingen	1
01	Voeding en niet-alcoholische dranken	2
01.1	Voeding	3
01.1.1	Brood en granen	4
01.1.1.1	Rijst	5
01.1.1.2	Meel en andere granen	5
01.1.1.3	Brood	5
01.1.1.4	Andere bakkerijproducten	5
01.1.1.5	Pizza en quiche	5
01.1.1.6	Deegwaren en couscous	5
01.1.1.7	Ontbijtgranen	5
01.1.1.8	Andere producten op basis van granen	5

Bovenop de scannerdata worden voor de voedingsproducten nog steeds aanvullend klassieke prijsopnames uitgevoerd in een aantal speciaalzaken (vb. bakkers en slagers) en discounters. In de HICP wordt ook webscraping gebruikt, waarbij gegevens automatisch van webpagina's worden gehaald ("*gescrapet*"). Bij de uiteindelijke indexberekening worden die verschillende datastromen gecombineerd op basis van omzet- of bestedingsinformatie. Zo wordt gegarandeerd dat de prijsevolutie op een representatieve manier gemeten wordt. Door die combinatie van verschillende datastromen bedraagt het feitelijke gewicht van de scannerdata ongeveer 18 à 19%.

2.3. Hoe ziet scannerdata er uit?

Statbel ontvangt wekelijks data van de voorgaande week van drie supermarktketens en dit via SFTP (SSH File Transfer Protocol). De data is opgesplitst in twee delen, het ene deel bevat alle productinformatie, het tweede deel bestaat uit de interne classificatie van de keten. Een fictief voorbeeld van beide datasets wordt gegeven in onderstaande tabellen.

De productinformatie bevat:

- een tijdsaanduiding voor de periode waarnaar de gegevens verwijzen;
- uiteenlopende productcodes, omzet en hoeveelheidsgegevens;
- gedetailleerde productomschrijvingen;
- een link naar de dataset met de interne classificatie waarin de supermarkt de producten klasseert.

Tabel 3: Voorbeeld productinformatie

Variabele	Beschrijving	Voorbeeld
DT_STRT	Startdatum week	2/feb/15
DT_STOP	Einddatum week	8/feb/15
CD_PROD_CLASS1	Interne classificatie level 1	D
CD_PROD_CLASS2	Interne classificatie level 2	E
CD_PROD_CLASS3	Interne classificatie level 3	I
CD_PROD_CLASS4	Interne classificatie level 4	K
NR_ITRL	Interne product code – 1	8523
NR_ART	Interne product code – 2	1568
NR_EAN	GTIN code van het product	5449000000286
TX_BRAND_NL	Merk beschrijving – Nederlands	Coca-Cola
TX_BRAND_FR	Merk beschrijving – Frans	Coca-Cola
TX_TYPE_NL	Beschrijving product type - Nederlands	2L
TX_TYPE_FR	Beschrijving product type - Frans	2L
TX_INFO_NL	Beschrijving product info - Nederlands	Regular (PET)
TX_INFO_FR	Beschrijving product info – Frans	Regular (PET)
MS_VAT_RT	BTW-tarief	6
MS_TRNOVR	Omzet	10000
MS_SALES_UNIT	Verkochte hoeveelheid	4000
CD_TYPE	Verkocht per eenheid of per gewicht	Units
MS_PKGG	Waarde verpakking	2
MS_PKGG_DESCR	Beschrijving verpakking (Liter, Kilo, Stuks, …)	L
MS_ALC_RT	Alcoholpercentage (%)	0
MS_AV_PRC	Gemiddelde prijs (MS_TRNOVR/MS_SALES_UNIT)	2,5

Tabel 4: Voorbeeld interne classificatie

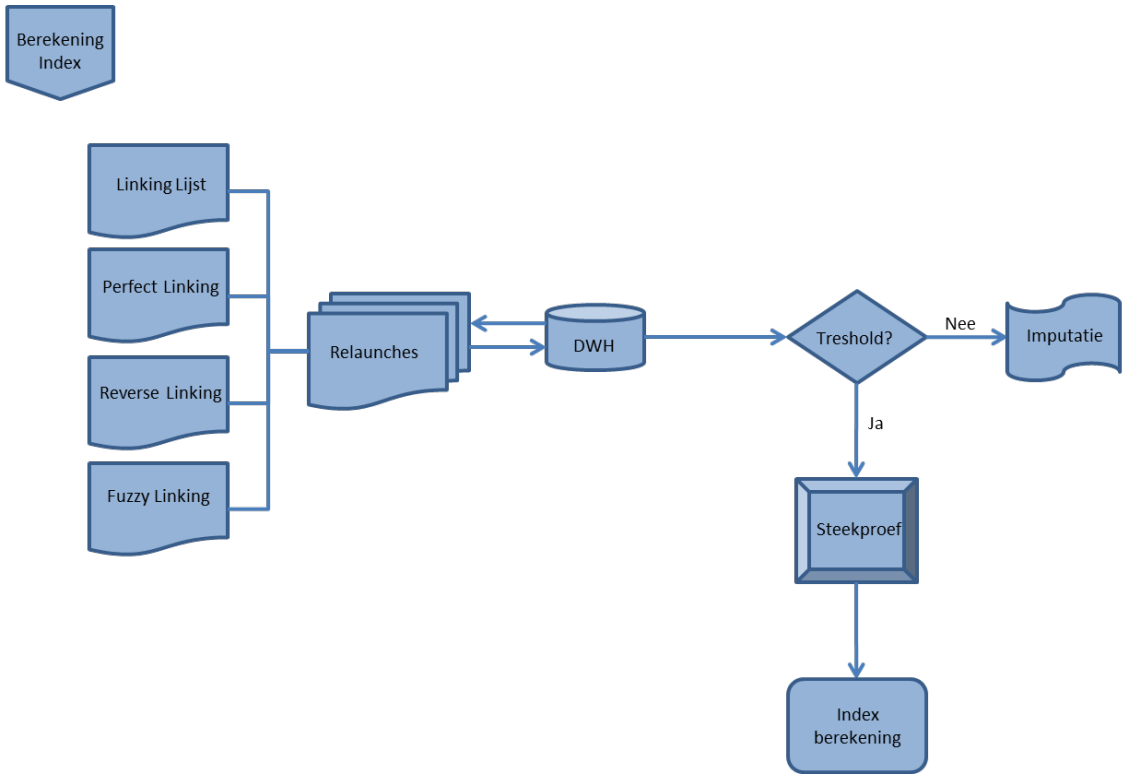
Variabele	Beschrijving	Voorbeeld
CD_PROD_CLASS1	Interne classificatie level 1	D
TX_PROD_CLASS1	Beschrijving classificatie level 1	Food
CD_PROD_CLASS2	Interne classificate level 2	E
TX_PROD_CLASS2	Beschrijving classificatie level 2	Drinks
CD_PROD_CLASS3	Interne classificatie level 3	I
TX_PROD_CLASS3	Beschrijving classificatie level 3	Lemonades
CD_PROD_CLASS4	Interne classificatie level 4	K
TX_PROD_CLASS4	Beschrijving classificatie level 4	Regular Cola

3. METHODOLOGIE

Kort samengevat maakt Statbel voor de verwerking van scannerdata tot indexcijfers gebruik van een dynamische korf (behalve voor seizoensproducten, cfr. 3.5 Seizoensproducten) met een Jevons kettingindex. Die dynamische steekproef wordt bepaald op basis van omzetcijfers van individuele producten in twee opeenvolgende maanden.

Aan de hand van een treshold wordt er vervolgens bepaald of een product al dan niet in de steekproef wordt opgenomen. Voor producten die buiten de steekproef vallen worden prijsimputaties berekend. Een individueel product wordt bepaald aan de hand van de interne code in plaats van barcodes.

Relaunches van producten worden gelinkt om "verborgen" prijswijzigingen mee in rekening te brengen. Indien nodig wordt er een hoeveelheidsaanpassing gedaan om een vergelijk te kunnen maken tussen het oude en het nieuwe product. Hieronder staat een schematische voorstelling dit proces:



Figuur 1: Schematische voorstelling indexberekening

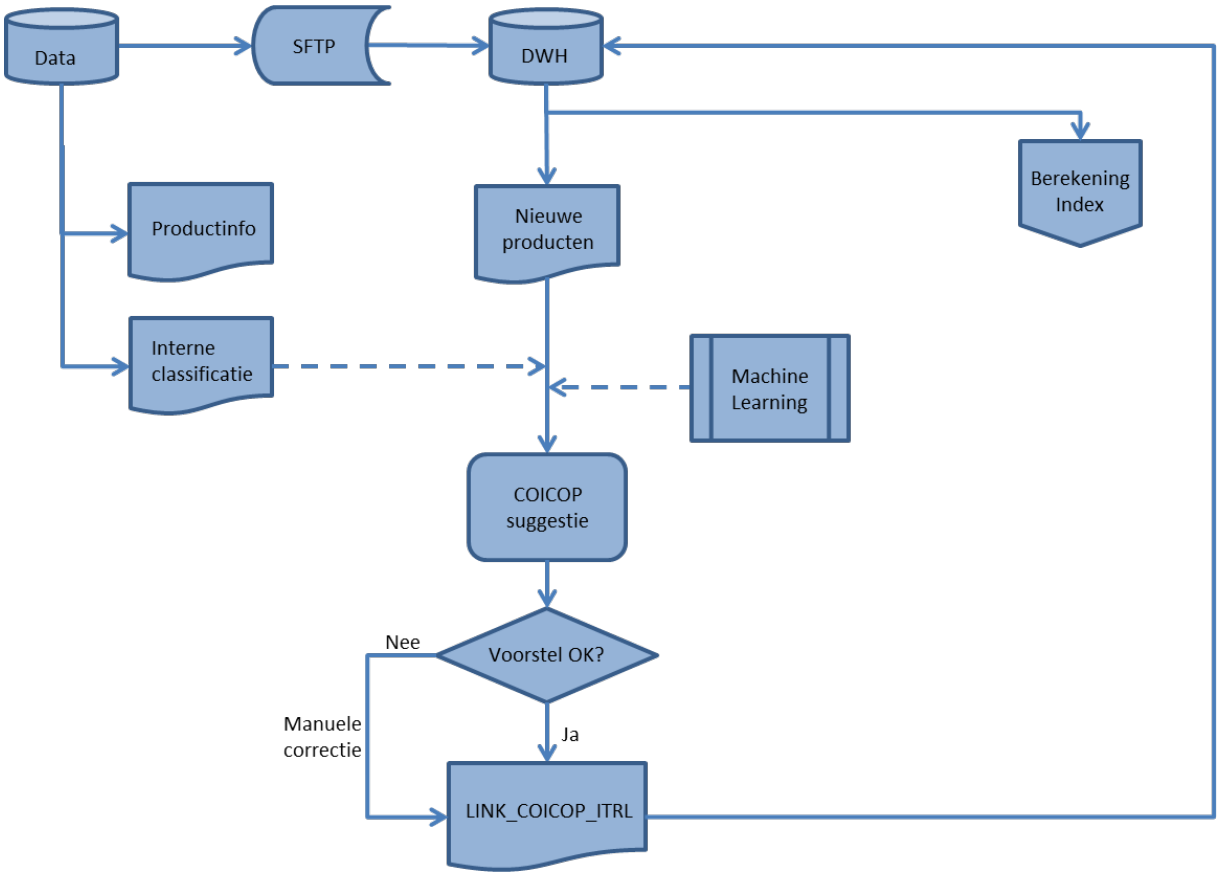
Voor elke supermarktketen wordt op deze wijze een index berekend, deze indices worden dan gecombineerd met andere data (o.a. klassieke prijsopnames) aan de hand van een stratificatiemodel.

In de volgende secties volgt een meer gedetailleerde uitleg voor elke van de hierboven vermelde punten. Alvorens indexcijfers berekend kunnen worden dienen producten gecombineerd te worden in productgroepen/consumptiesegmenten. Deze productgroepen dienen gelinkt te worden aan de juiste ECOICOP categorie.

3.1. Classificatie

De classificatie van scannerdata in de COICOP-nomenclatuur kan opgesplitst worden in een opstartfase en een recurrente verwerking. In de opstartfase wordt de interne supermarktclassificatie zo goed mogelijk gelinkt aan de ECOICOP-classificatie en worden er verschillende consumptiesegmenten gecreëerd op COICOP 6-niveau. Tijdens de wekelijkse verwerking wordt er een controle uitgevoerd voor de nieuwe producten.

Het classificatieproces wordt hieronder schematisch weergegeven:



Figuur 2: Schematische voorstelling classificatie

3.1.1. Scannerdata en COICOP-classificatie: opstartfase

De eerste stap is het linken van de interne classificatie van de supermarktketens aan het ECOICOP 5-niveau (voor elke keten apart). Vervolgens worden per keten onderverdelingen gemaakt op het zogenaamde COICOP 6-niveau. Dat gebeurt per consumptiesegment. Bijvoorbeeld voor de ECOICOP-categorie frisdranken kan dit bestaan uit cola, limonade, ice-tea, etc. Het is niet de bedoeling om voor elke keten dezelfde segmenten te creëren, maar om prijsevoluties te kunnen vergelijken wordt er wel gestreefd naar een zo groot mogelijke conformiteit tussen de ketens. Hoewel de scangegevens later gecombineerd moeten worden met de klassieke prijsopnames, is het ook niet de bedoeling om een kopie te maken van de consumptiesegmenten van die klassieke methode. Aangezien de klassieke methode ook een sample betreft, zou het limiteren van scannerdata tot dezelfde segmenten leiden tot het niet gebruiken van veel informatie en producten. Zo wordt er bij klassieke prijsopname gewerkt met ongeveer 173 consumptiegroepen per keten in COICOP 01 (Voeding en niet-alcoholische dranken), terwijl er hiervoor met scannerdata ongeveer 450 groepen gemaakt worden.

Elke keten heeft ongeveer 3.500 interne klassen. De consumptiesegmenten worden soms gebaseerd op één interne ketenclassificatie, waarbij deze interne klasse dan in zijn geheel kan ingedeeld worden onder het consumptiesegment. Maar meestal is het een combinatie van verschillende interne klassen. Als voorbeeld is er het huismerk van een keten, dat wordt soms als aparte interne klasse beschouwd, soms worden zelfs de verschillende merken van een productgroep apart geklasseerd. In dit geval worden de verschillende interne classificaties samengenomen in één COICOP 6-consumptiesegment. Ook zijn de interne klasseringen van de keten soms té specifiek, waardoor er slechts één product in de klasse zit. In dit geval zullen er dan ook verschillende interne klassen gecombineerd worden. Bij de latere aggregatie worden er op COICOP 6-niveau gewichten gebruikt, waardoor consumptiesegmenten met slechts één product vermeden moeten worden. In het geval het product niet meer beschikbaar zou zijn en er geen vervanging voorhanden is, zou de prijsevolutie van zo'n groep volledig bestaan uit een imputatie, wat volgens de HICP regelgeving slechts toegelaten is voor één maand, maar niet langer.

Ook is het mogelijk dat een interne classificatie te heterogeen is om goede consumptiesegmenten te maken. Bijvoorbeeld de interne klasse 'Rode wijn', wat kan opgesplitst worden in 'Franse rode wijn', 'Italiaanse rode wijn',… Of de interne klasse koffie die bijvoorbeeld opgesplitst kan worden in 'pads', 'gemalen' en 'oploskoffie'. Aan de hand van de data of naar analogie

met de andere ketens, worden er dan extra subklassen op COICOP 6-niveau gemaakt. Afhankelijk van de keten wordt ongeveer 10 à 15 procent van de producten individueel gelinkt. In principe wordt linken op productniveau vermeden, maar in dergelijke gevallen is het noodzakelijk.

3.1.2. Scannerdata en COICOP-classificatie: recurrent

De mapping van scannerdata naar de COICOP-classificatie vraagt aanvankelijk veel tijd. Maar na de opstartfase hoeft er enkel wekelijks gecontroleerd te worden of de nieuwe producten uit de scannerdata gelinkt zijn aan de correcte COICOP 6-groep. In het geval van een foute linking, wordt het product individueel gelinkt aan de juiste groep. Afhankelijk van het aantal nieuwe producten in de datasets, neemt dit proces ongeveer 1 à 2 dagen per week in beslag voor de drie supermarktketens. Er wordt gebruik gemaakt van een aparte database om de interne productcodes te linken met een COICOP 6-groep/consumptiesegment. Als data in deze database zit wordt de andere linking (o.b.v. interne classificatie) verworpen. Met andere woorden: linking op productniveau heeft voorrang op linking met interne klassen.

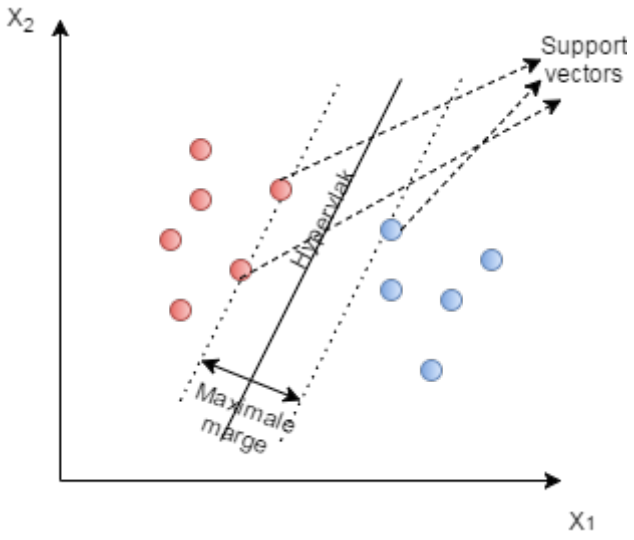
3.1.3. Machine learning

Om producten in het juiste COICOP-consumptiesegment te klasseren wordt “supervised machine learning” gebruikt.

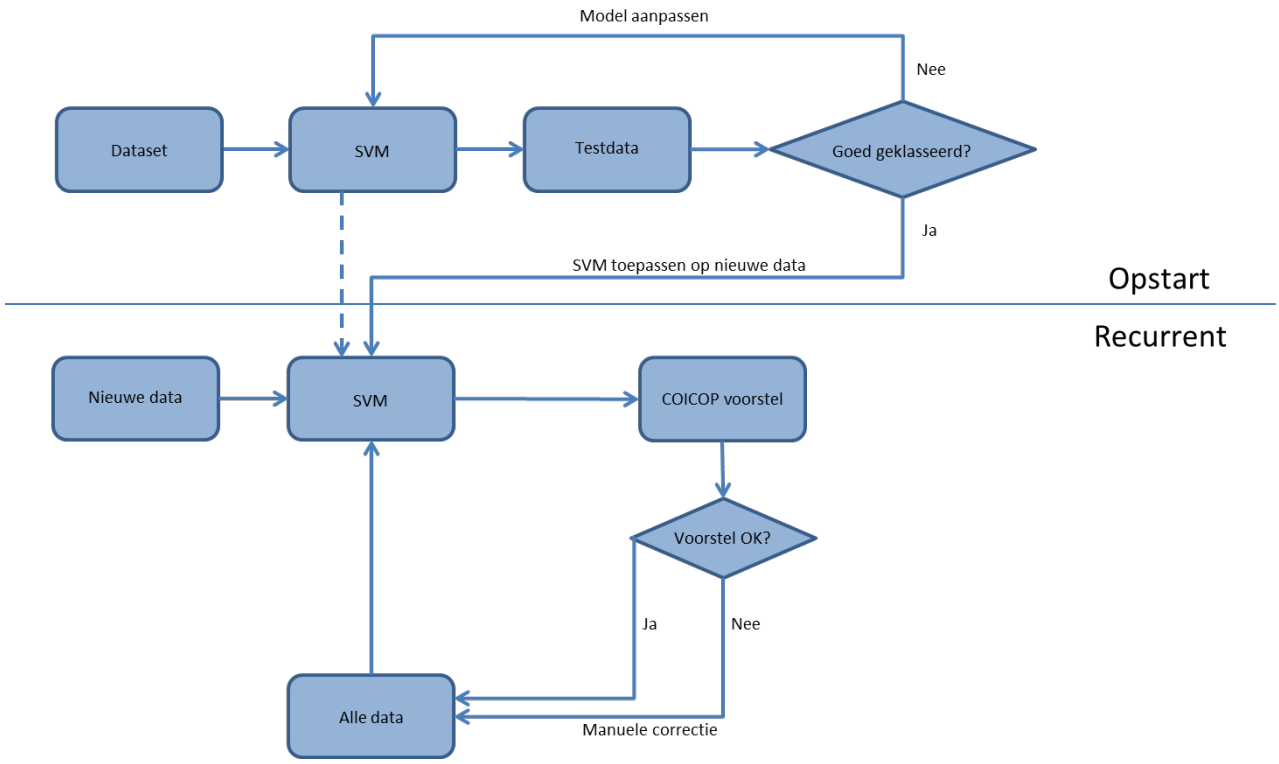
De supervised machine learning die Statbel hanteert, maakt gebruik van een “support vector machine”-algoritme (SVM). Aan de hand van een pregedefinieerde trainingsdataset kan het algoritme een model bouwen om nieuwe data te classificeren op basis van de gelijkenissen tussen de trainingsdataset en de nieuwe data. Theoretisch gezien maakt een SVM-model een ‘zo goed mogelijke’ scheiding (hypervlak) tussen de verschillende categorieën. Door te kijken aan welke kant van het hypervlak het nieuwe data-object ligt, kan het SVM-model het object in de juiste categorie onderbrengen. De ‘zo goed mogelijke’ scheiding houdt in dat de afstand tussen het scheidingsvlak en de dichtsbijgelegen objecten (support vectors) van elke klasse (de marge) zo groot mogelijk is.

Onderstaande grafiek geeft schematisch weer hoe de classificatie van observaties in twee klassen kan gebeurd via lineaire SVM.

Figuur 3: Voorstelling van classificatie via Support Vector Machine-algoritme



In eerste instantie wordt een trainingsdataset aangelegd door producten manueel aan een consumptiesegment toe te wijzen (“supervised” gedeelte). Het support vector machine-algoritme zal dan, op basis van de productomschrijvingen en de toegewezen classificatie, een model creëren dat nieuwe producten toewijst aan één van de categorieën. Vooraleer het algoritme in gebruik wordt genomen, wordt het getest op testdata die ook manueel geklasseerd werd. Wanneer de foutenmarge beperkt blijft, kan het algoritme elke week opnieuw toegepast worden op nieuwe data. Waarna de voorgestelde productcategorie voor elk product manueel bevestigd of gecorrigeerd wordt. Deze nieuwe geklasseerde data wordt de week nadien toegevoegd aan de trainingsdata waardoor het model steeds intelligenter en betrouwbaarder wordt. Het onderstaande schema geeft een voorstelling van deze procedure:



Figuur 4: Machine learning voor COICOP-classificatie

Stel bijvoorbeeld dat een supermarkt slechts één interne klasse heeft voor koffie (COICOP 5-groep 01.2.1.1), maar dat er op basis van de producten vier subklassen kunnen gemaakt worden: pads, geroosterd, gemalen en oploskoffie. De bedoeling is aan de hand van de beschrijving (tekst) de producten in de vier categorieën te verdelen. Een deel van de koffiedataset wordt eerst manueel geklasseerd over de vier categorieën (trainingsdataset). Daarna maakt het algoritme een model aan, wat op zijn beurt getest wordt op de testdata. Als het algoritme goed geëvalueerd wordt, kan het 'in productie' worden toegepast op nieuwe data. Nadien wordt deze correct geklasseerde nieuwe data gebruikt om het model opnieuw te schatten voor de classificatie van toekomstige data.

Er bestaat ook "unsupervised machine learning" waarbij er geen pregedefinieerde trainigsdataset nodig is. Het algoritme bepaalt dan zelf de categorieën op basis van de data. Het voordeel van deze methode is dat geen trainingsdataset manueel moet worden opgesteld. Het nadeel is dat het resultaat waarschijnlijk niet de categorieën aanmaakt zoals verwacht zou worden. In het voorbeeld van de koffie kan er mogelijks een consumptiesegment gemaakt worden dat zowel geroosterde koffie als pads bevat (ipv twee aparte categorieën). Een ander minpunt kan zijn dat, aangezien deze methode niet zoekt naar gelijkenissen tussen de omschrijvingen, er tussen twee perioden verschillende consumptiesegmenten worden aangemaakt, doordat er nieuwe producten werden toegevoegd.

Die minpunten maken dat Statbel gebruik maakt van supervised machine learning.

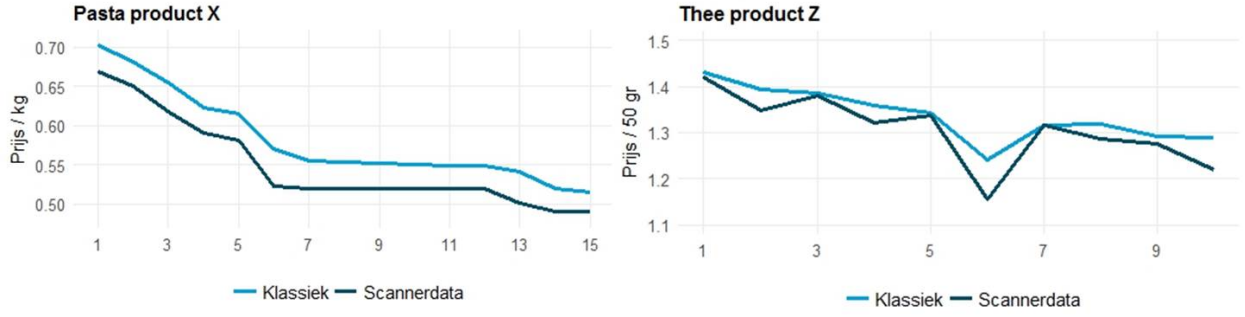
3.2. Prijsconcept

3.2.1. Eenheidsprijzen (of unit values)

In tegenstelling tot de geafficheerde prijzen bij traditionele prijsopnames, wordt via scannerdata de reële gemiddelde transactieprijs geobserveerd. Deze prijs wordt berekend als quotiënt van de totale omzet en verkochte hoeveelheid van een product voor een bepaalde periode. In België is dit meestal één of twee weken voor de CPI en drie weken voor de HICP.

De omzetten en hoeveelheden worden per product geaggregeerd op het niveau van de supermarktketens. Daardoor is het resultaat een gemiddelde transactieprijs (of eenheidsprijs) per product en per keten. Ook bij de traditionele prijsopname wordt een gemiddelde prijs per product en keten bekomen, maar slechts op één bepaalde dag van de maand. Scannerdata

leveren prijsinformatie over een langere tijdsperiode. Hoewel er dus een conceptueel verschil is in de waargenomen prijzen, is het effect wat betreft de prijsevolutie echter verwaarloosbaar. Dit blijkt in de volgende grafieken[4]:



Figuur 5: Vergelijking prijsevolutie klassieke methode en scannerdata

De scannerdataprijzen liggen over het algemeen lager dan de klassiek opgenomen prijzen. Dit is een gevolg van de grotere hoeveelheid aan prijzen die met scannerdata gecapteerd wordt. Zo zijn ook promo's en kortingen inbegrepen. Dat zal bij de klassieke methode minder waargenomen worden. Volgens de HICP-regelgeving moeten discriminatoire kortingen uitgesloten worden, indien geen informatie beschikbaar is over het aantal mensen dat een product tegen een korting heeft aangekocht.

Discriminatoire kortingen zijn kortingen die niet voor iedereen beschikbaar zijn (enkel van toepassing met een klantenkaart) of slechts een beperkte dag geldig zijn. Bij klassieke prijsopnames door enquêteurs is deze informatie niet beschikbaar en worden deze kortingen dan ook niet in rekening gebracht in lijn met de HICP-regelgeving. Met scannerdata zijn deze gegevens wel beschikbaar omdat de gemiddelde prijs het aantal transacties met korting omvat. Deze kortingen mogen, in lijn met de HICP-regelgeving, dan ook in rekening gebracht worden. Ongeacht deze lagere prijzen is de prijsevolutie nagenoeg identiek.

Het verschil in het aantal weken dat gebruikt wordt voor de berekening in de HICP en CPI is te wijten aan het verschil in publicatiedatum tussen beide indexcijfers. De CPI wordt gepubliceerd de voorlaatste werkdag van de maand. De HICP wordt gepubliceerd door Statbel en Eurostat, ten laatste twee weken na het einde van de maand. De HICP-regelgeving stelt ook dat prijzen van de week die de 15de van de maand bevat, opgenomen moeten worden in de indexberekening. Door de vroege publicatiedatum van de CPI is het normalerwijze onmogelijk om hier aan te voldoen voor de HICP. Bovendien is de aanbeveling van Eurostat ook om prijzen van drie weken te gebruiken in de berekening van indexcijfers met scannerdata. Het voorstel om de publicatiedatum van de CPI te verlaten en die van de HICP te vervroegen – in lijn met andere Europese landen - zodat beide indexcijfers op hetzelfde moment gepubliceerd zouden worden, en dezelfde prijsinformatie zouden gebruiken, werd afgewezen door de Indexcommissie[5] omwille van implicaties op indexeringsmechanismen die van toepassing zijn in België.

Het verschil in aggregatie van prijzen over kortere tijdsperioden in de CPI heeft op lange termijn geen impact op de indexevolutie. Op korte termijn zijn er wel verschillen zoals in volgende paragraaf wordt toegelicht.

[4] De onderzoeken zijn gebaseerd op meerdere verschillende perioden, voor de algemeenheid worden deze steeds genummerd startend vanaf 1. Elke periode komt overeen met een maand.

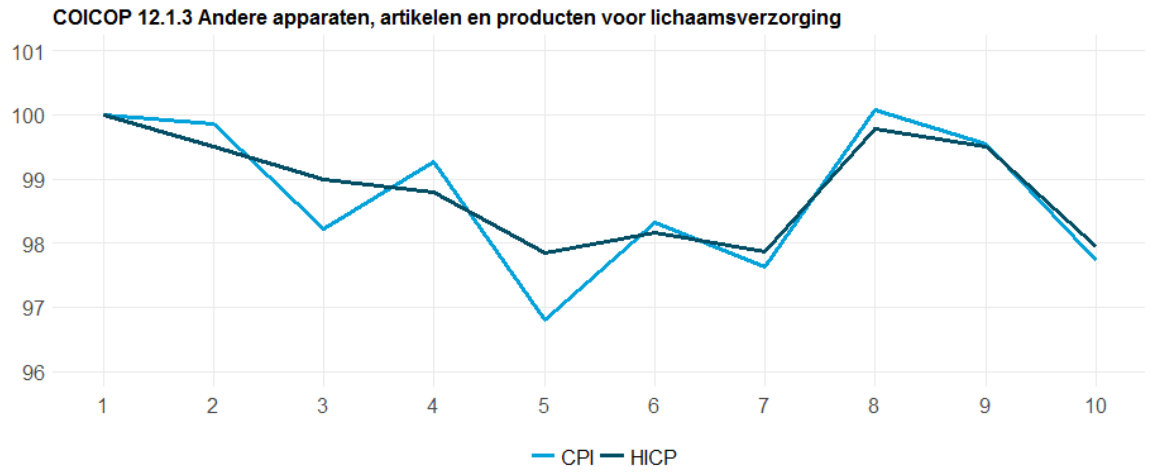
[5] De Indexcommissie is een paritair samengestelde commissie, bestaande uit academici en werknemers- en werkgeversorganisaties. De commissie wordt ondersteund door statistici van Statbel. Deze commissie geeft aan de minister van Economie advies over alle vraagstukken rond de consumptieprijsindex en geeft ook maandelijks advies over het door Statbel berekende indexcijfer. De Indexcommissie geeft ook haar advies over de jaarlijkse actualisering aan de minister van Economie.

3.2.2. Aggregatie van prijzen over verschillende tijdsperioden

Prijzen op individueel productniveau worden dus berekend door de omzet en de verkochte hoeveelheden te aggregeren over de tijdsperiode waarop de indexberekening is gebaseerd en vervolgens het quotiënt te nemen.

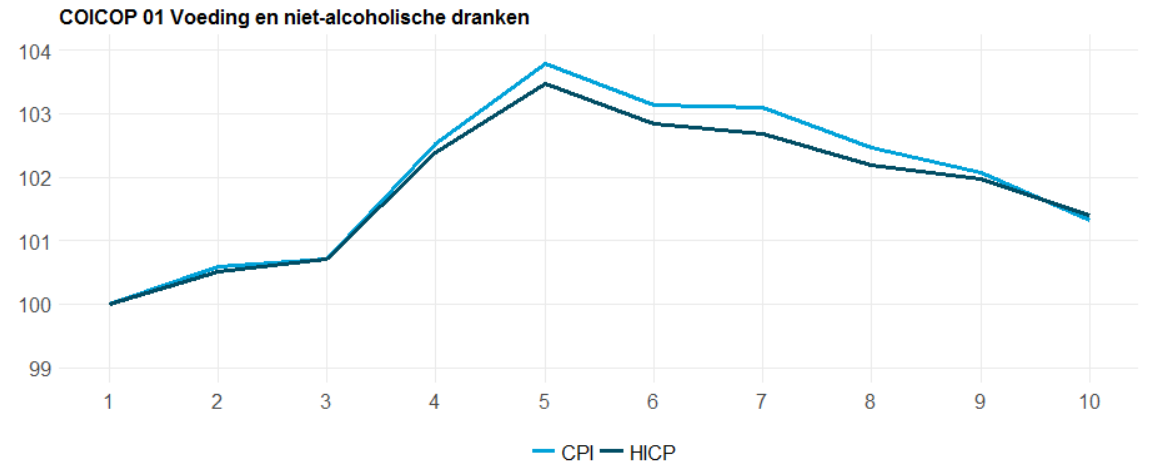
Voor de HICP en CPI worden er, zoals eerder werd beschreven, verschillende weken gebruikt in de berekening. Hoe de indexberekening met scannerdata concreet gebeurt, wordt toegelicht in hoofdstuk 3.3. Toch wordt al even stilgestaan bij de manier waarop de eenheidsprijs berekend wordt met scannerdata, en het eventuele effect hiervan op de gemeten prijsevolutie**.**

Het gevolg van het opnemen van meerdere weken in de berekening van de HICP zorgt er namelijk voor dat kortingen en promoties – die meestal één week duren - minder zwaar doorwegen in de eenheidsprijs omdat het effect hiervan uitgevlakt wordt over meerdere weken. Dat resulteert in stabielere maand-op-maand indexcijfers zoals blijkt uit volgende grafiek voor COICOP 12.1.3. De prijsevolutie op lange termijn is echter identiek.



Figuur 6: Indexcijfers (CPI en HICP) voor COICOP 12.1.3

Op geaggregeerde niveaus is het verschil tussen CPI en HICP echter kleiner. Zo is het verschil voor de globale COICOP 01 (Figuur 7) verwaarloosbaar omdat promoties op lagere COICOP niveaus elkaar opheffen van maand tot maand. Het einde van één promotie zorgt voor een opwaarts effect terwijl een nieuwe promotie in een andere productgroep voor een tegenstelde beweging zorgt, waardoor beide effecten in aggregatie worden opgeheven.

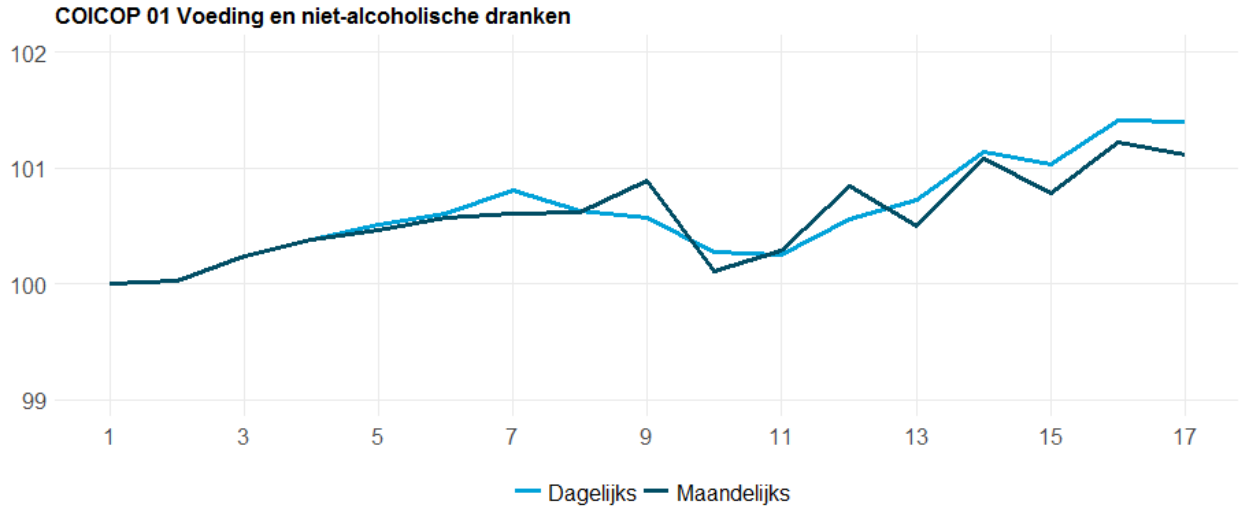


Figuur 7: Geaggregeerde indexcijfers (CPI en HICP) voor COICOP 01

De berekening met verschillende tijdsperioden resulteert in eenzelfde langetermijn-prijsevolutie. De methodologie die gebruikt wordt, garandeert dus consistentie, ongeacht het aantal weken dat wordt gebruikt.

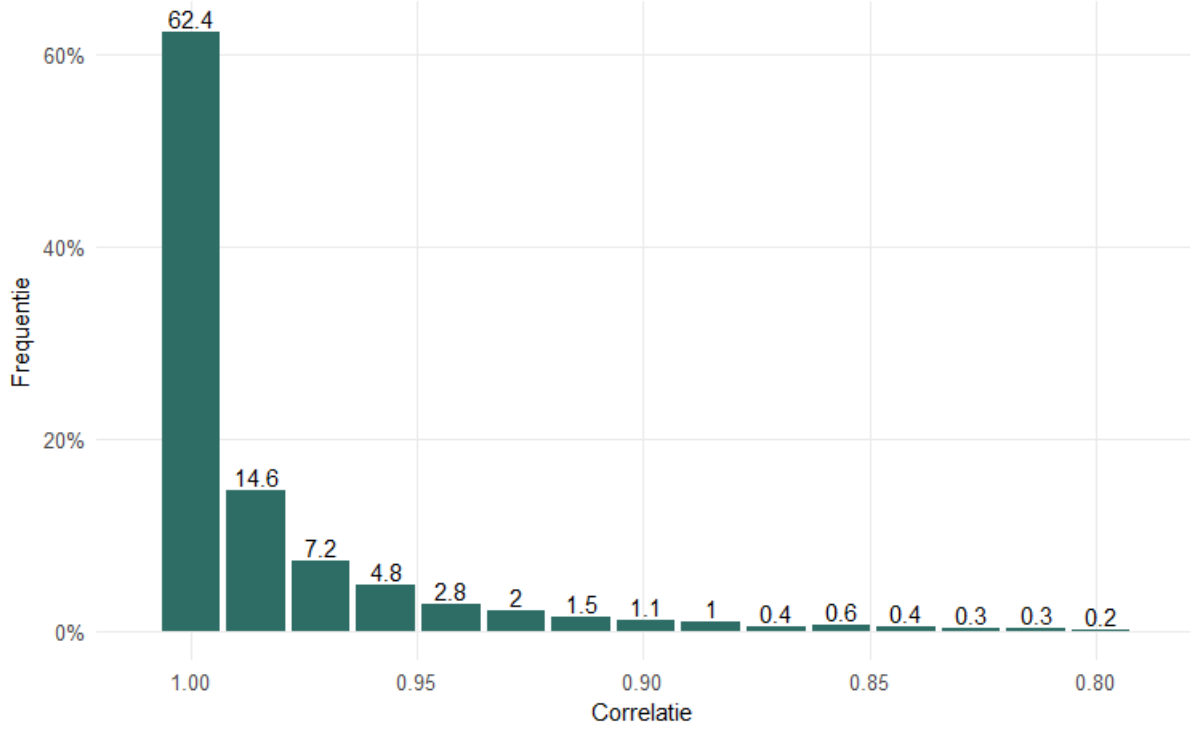
Het is ook aangewezen om even stil te staan bij de keuze om een eenheidsprijs per product te berekenen over een tijdsperiode van een aantal weken. Die berekening geeft de werkelijke transactieprijs weer waartegen een product is gekocht, maar wijkt wel sterk af van de manier waarop prijzen manueel worden ingezameld door enquêteurs. Met manuele prijsinzameling ontbreekt informatie over verkochte hoeveelheden (of omzet). Enkel de ongewogen prijzen van een individueel product worden gebruikt.

Indien geprobeerd wordt deze methode na te bootsen met scannerdata via een ongewogen gemiddelde van dagelijkse prijzen gedurende een periode van 3 weken ten opzichte van een eenheidsprijs over dezelfde tijdsperiode, wordt globaal gezien weer eenzelfde prijsevolutie bekomen (Figuur 8). De index op basis van prijzen per dag is uiteraard stabieler dan de index die prijzen over een langere tijdsperiode aggregeert, omdat het effect van kortingen ook hier weer afgevlakt wordt.



Figuur 8: Prijsevolutie van dagelijkse en maandelijkse eenheidsprijzen

Er is trouwens ook een hoge correlatie tussen de prijzen per dag en die per week. In 83% van de gevallen is die groter dan 99% (zie figuur 9).

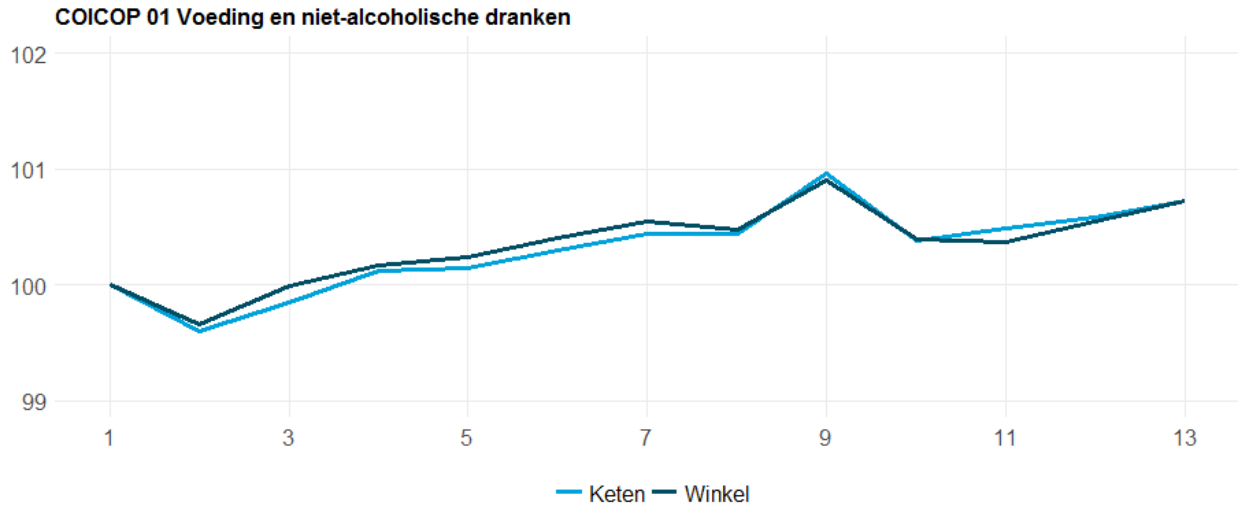


Figuur 9: Correlatie tussen prijzen per dag en prijzen per week

3.2.3. Aggregatie van prijzen over verschillende winkels

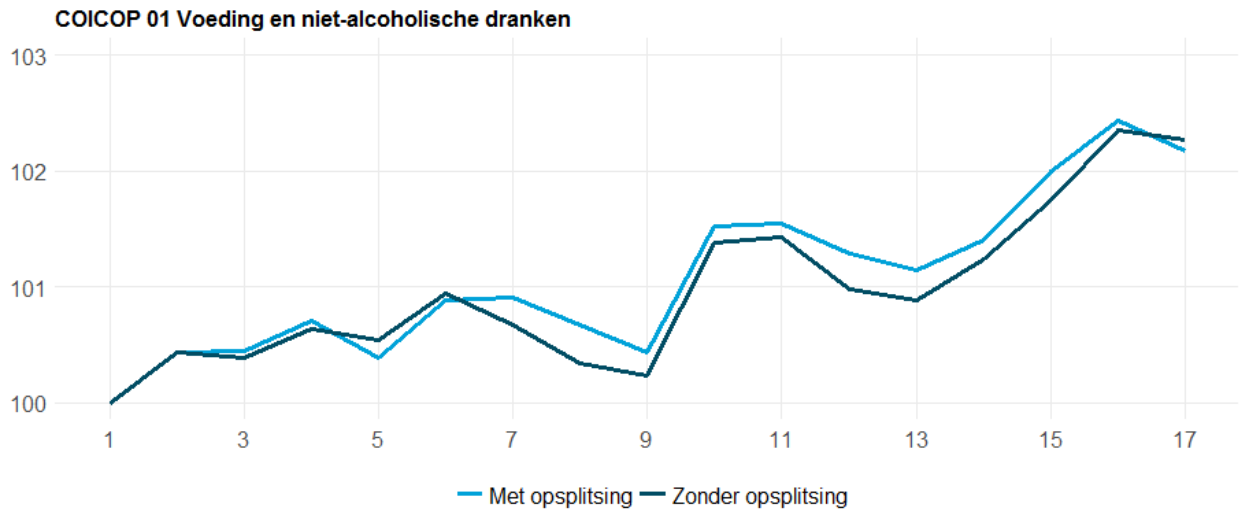
Met scannerdata worden de prijzen geaggregeerd op niveau van een supermarktketen. Dat vereenvoudigt de berekeningswijze, maar kan bij een verschil in "service level" (dienstverleningsniveau) leiden tot een potentiële vertekening van de gemeten prijsevolutie. Het kan hierbij gaan om een verschil tussen winkels van hetzelfde winkelmerk van een keten (bv. eenzelfde type van winkel in verschillende gemeentes), of een verschil tussen de winkelmerken van een keten (bv. kleine buurtwinkels en grote supermarkt). Dit omdat de prijsevolutie per winkel of per segment kan verschillen en de overstap van klanten naar andere winkels binnen dezelfde supermarktketen in theorie dan geneutraliseerd zou moeten worden door middel van opsplitsing en het berekenen van een index per winkel of per winkelsegment.

In praktijk levert een berekening die gebeurt op individueel winkelniveau, waarna vervolgens geaggregeerd wordt naar het niveau van de keten, gelijkaardige resultaten op als een berekening die rechtstreeks gebeurt op het winkelketenniveau. De aggregatie over winkels stelt dus geen probleem.



Figuur 10: Prijsevolutie op ketenniveau en geaggregeerd op winkelniveau

Ook de aggregatie over verschillende winkelsegmenten van eenzelfde keten levert geen vertekening op. De methode waarbij eerst een opsplitsing zou gedaan worden per segment en deze vervolgens aggregeert per segment levert geen vertekening op van de gemeten prijsevolutie t.o.v. de methode waarbij de eenheidsprijs meteen per product wordt berekend over de winkelsegmenten heen.



Figuur 11: Prijsevolutie met en zonder opsplitsing tussen winkelsegmenten

3.2.4. Interne code in plaats van barcode

Het meten van inflatie heeft als doel het meten van de prijsevolutie van eenzelfde product doorheen de tijd. Indien verschillende producten met elkaar vermengd worden, dan zou eerder de gemiddelde bestedingsevolutie gemeten worden en niet de prijsevolutie. Om dezelfde producten maand na maand met elkaar te linken wordt er internationaal gebruik gemaakt van GTIN (Global Trade Item Number), beter bekend als de barcode. Dat is een unieke code die aan een bepaald product wordt toegewezen. Ervaring met scannerdata door Statbel heeft echter uitgewezen dat het handiger is om de interne codes van de supermarktketens (stock keeping units of SKUs) te gebruiken. Die zijn stabieler en ook uniek voor de prijsindexberekening en worden ook in andere landen, zoals Zwitserland en Australië, gebruikt. De GTIN-codes worden wel nog gebruikt om de consistentie tussen de verschillende supermarkten te controleren voor wat betreft de linking met de COICOP-classificatie. De voorkeur voor interne code boven barcodes wordt duidelijk met de volgende voorbeelden.
De onderstaande tabel toont het voorbeeld van een product dat een aantal weken met verschillende barcodes werd verkocht, maar wel gelinkt was aan eenzelfde interne code.

Tabel 5: Voorbeeld van product met verschillende barcodes

Week	Interne code	Productomschrijving	Eenheid	Verkochte eenheden	Omzet	Prijs	EAN
3713	12345	Merk x - 40 stuks	0,375	380	2755	7,25	#8000565755675
3813	12345	Merk x - 40 stuks	0,375	561	3540	6,31	#8000565755675
3913	12345	Merk x - 40 stuks	0,375	1289	7657	5,94	#8000565755675
4013	12345	Merk x - 40 stuks	0,375	763	4288	5,62	#8000565755675
4113	12345	Merk x - 40 stuks	0,375	1128	6757	5,99	#8000565755675#8000508890089
4213	12345	Merk x - 40 stuks	0,375	912	5591	6,13	#8000565755675#8000508890089
4313	12345	Merk x - 40 stuks	0,375	621	4229	6,81	#8000565755675#8000508890089
4413	12345	Merk x - 40 stuks	0,375	848	5080	5,99	#8000565755675#8000508890089
4513	12345	Merk x - 40 stuks	0,375	2120	12699	5,99	#8000565755675#8000508890089
4613	12345	Merk x - 40 stuks	0,375	6728	44270	6,58	#8000565755675#8000508890089

De interne code toont aan dat het om hetzelfde product gaat. Hoewel er verschillende barcodes worden gebruikt, gaat het hier toch steeds over eenzelfde product, dat aan eenzelfde prijs wordt verkocht. Dit volgt ook uit de volgende tabel.

Tabel 6: Voorbeeld van product met verschillende barcodes in eenzelfde week

EAN	Interne code	Week	Productomschrijving	Eenheid	Verkochte eenheden	Omzet	Prijs
8000565755675	12345	4113	Merk x - 40 stuks	0,375	410	2455,9	5,99
8000508890089	12345	4113	Merk x - 40 stuks	0,375	718	4300,82	5,99

Daarnaast wordt een barcode soms opnieuw gebruikt voor een totaal verschillend product. Ook multipacks krijgen vaak een andere barcode om de kassaverwerking te versnellen, maar die barcode is dan wel soms gelinkt aan eenzelfde interne productcode. Die interne code kan ook gebruikt worden om de producten via de website van de supermarktketen te raadplegen. Dat is handig bij het linken van nieuwe en oude producten in het geval van vervangingen. Door gebruik te maken van de interne codes is het ook mogelijk om een index te berekenen voor verse producten (bv. vlees, fruit en groenten). Die producten worden verkocht in verschillende hoeveelheden (met verschillende gewichten), waarbij elk gewicht een aparte barcode heeft. De gebruikte barcodes voor deze producten zijn trouwens ook geen officiële GTIN-codes, maar intern gegenereerde barcodes of PLU (price look-up) codes die ingevoerd worden aan de kassa. Door de verschillende hoeveelheden te aggregeren op basis van de interne code kan de gemiddelde prijs per kilo wel bepaald worden.

3.3. Sampling en aggregatie

In dit onderdeel zal de berekening van indexcijfer op het laagste niveau binnen een supermarktketen worden uiteengezet.

3.3.1. Berekening van elementaire indexcijfers

In eerste instantie worden voor elke keten apart elementaire indexcijfers bepaald. Elementaire indexcijfers zijn die waarbinnen geen weging wordt gehanteerd. Aan het resulterende indexcijfer zelf wordt wel een weging gekoppeld. De elementaire indexcijfers voor scannerdata bevinden zich op COICOP 6-niveau en worden berekend op basis van de individuele prijsgegevens, op productniveau (SKU-niveau).

Aangezien alle producten binnen een bepaalde groep hetzelfde gewicht hebben, is het noodzakelijk om een steekproef samen te stellen die enkel representatieve producten omvat. Dat gebeurt met behulp van een drempel (treshold). Zoniet zou de index beïnvloed worden door producten die weinig gekocht worden. Er wordt immers vastgesteld (zie verder) dat er binnen de groepen een groot marktaandeel is van een beperkt aantal producten. Een product wordt opgenomen in het staal als door het gemiddelde marktaandeel over twee maanden een drempel wordt overschreden die afhankelijk is van het aantal producten per groep.

Een product wordt opgenomen wanneer:

$$\frac{s_m + s_{m-1}}{2} > \frac{1}{n * \lambda} \qquad \textit{Vgl. 1}$$

met λ = 1,25

n = aantal producten per groep

s_m = marktaandeel van elk product in maand m

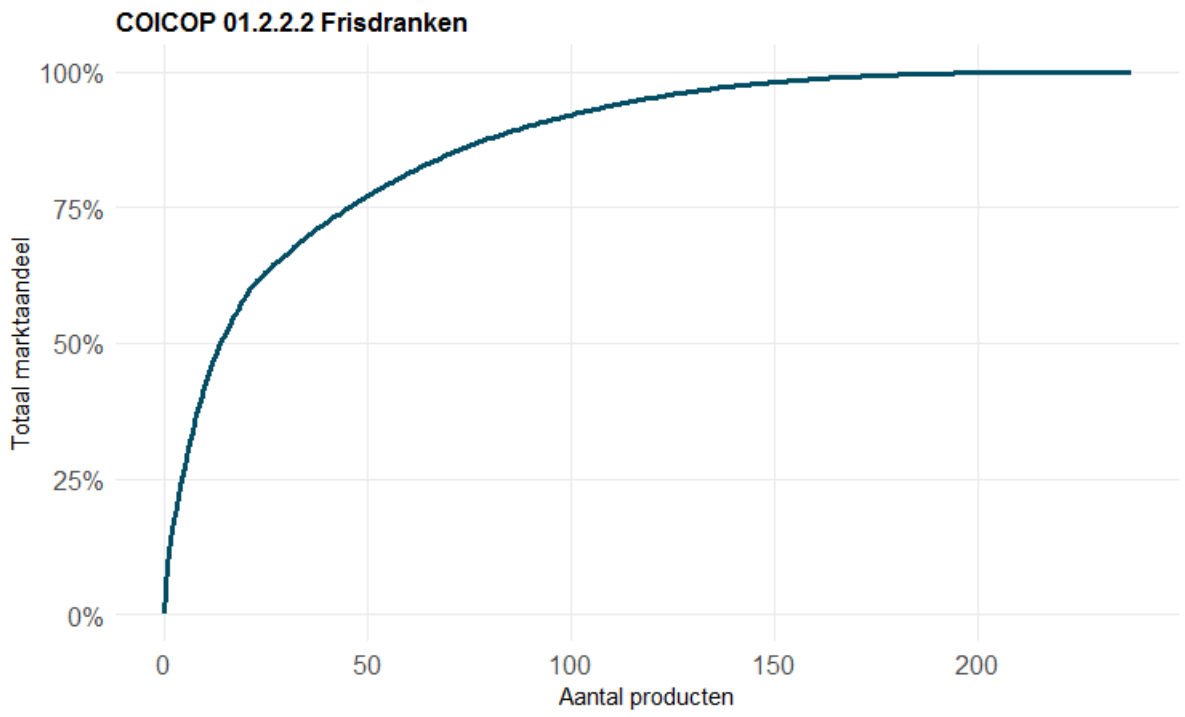
s_{m-1} = marktaandeel van elk product in maand m-1

Uit de praktijk blijkt dat een relatief klein aantal producten verantwoordelijk is voor een groot deel van de omzet van een COICOP 6 consumptiesegment. Gemiddeld 40 à 45% van de producten stellen ongeveer 80% van het marktaandeel voor. In de volgende tabel staan een paar voorbeelden voor de ECOICOP 5 groep Yoghurt (1.1.4.4):

Tabel 7: Marktaandeel COICOP 1.1.4.4

COICOP	Beschrijving	% producten	Totaal marktaandeel
01.1.4.4.01	Magere yoghurt (natuur)	46.67%	81.72%
01.1.4.4.02	Volle yoghurt (natuur)	45.00%	79.05%
01.1.4.4.03	Magere yoghurt met fruit	37.50%	71.56%
01.1.4.4.04	Volle yoghurt met fruit	37.74%	81.96%
01.1.4.4.05	Functionele yoghurt	72.22%	92.70%
01.1.4.4.06	Yoghurt voor kinderen	80.00%	85.32%
01.1.4.4.07	Drinkyoghurt	44.00%	70.77%

Voor de ECOICOP 5-groep frisdrank wordt 80% van het marktaandeel zelfs bepaald door minder dan 30% van de producten:



Figuur 12: Marktaandeel t.o.v. het aantal producten in COICOP 1.2.2.2

Uit de formule volgt dat de indexberekening gebeurt met relatieve prijzen (of de prijsevolutie) van producten die twee opeenvolgende maanden in het staal zitten op basis van een dynamische drempel.

De relatieve prijzen worden bepaald op productniveau (interne code). Er gebeurt telkens een vergelijking van de prijs voor hetzelfde product tussen twee maanden. Om tot een elementair indexcijfer te komen, worden die relatieve prijzen binnen hun consumptiesegment geaggregeerd via een meetkundig gemiddelde (Jevons-index), zoals dat sinds 2014 ook gebeurt voor de klassieke prijsopnames. Dat groepsniveau, dat niet gepubliceerd wordt wegens eigen aan de betreffende keten, ligt onder het gepubliceerde COICOP 5-niveau.

Uit dat alles blijkt dat het prijsindexcijfer op basis van scannerdata, net zoals voorheen, steeds de prijsevolutie van een identiek product (bv. een blikje limonade van 33 cl van merk A in supermarkt X) in een bepaalde maand ten opzichte van de prijs in de voorafgaande maand weergeeft. Steeds worden zowel eerste-prijs-producten, producten van het huismerk en merkproducten in de index gevolgd, voor zover ze een representatieve verkoop halen in de beschouwde periode en dus in de steekproef zijn opgenomen.

Het is echter een vaak voorkomende misvatting dat door het gebruik van scannerdata bij een toenemend verbruik van bijvoorbeeld goedkopere producten ten koste van duurdere producten, de index zou dalen of in het omgekeerde geval zou stijgen. De consumptieprijsindex weerspiegelt het zuivere prijsverloop. Wijzigingen in consumptiegedrag mogen en zullen dus de gemeten prijsevolutie niet rechtstreeks beïnvloeden. Op productniveau (niveau van de streepjescodes) wordt geen rekening gehouden met verkochte volumes of omzetcijfers bij de aggregatie van de verschillende producten tot een indexcijfer voor een productgroep. Het aggregaat is een ongewogen meetkundig gemiddelde. Indien prijzen worden geaggregeerd met omzetcijfers wordt een zogenaamd “unit value index” bekomen die niet het zuivere prijsverloop meet, maar wel een combinatie van wijzigende prijzen én verkochte hoeveelheden.

De bekomen elementaire indexcijfers voor elke keten worden geaggregeerd aan de hand van de omzetcijfers van de betreffende consumptiesegmenten voor het voorgaande jaar. De aggregatie gebeurt gewogen op basis van de gewichten van elke elementaire groep. Die gewichten worden jaarlijks aangepast en zijn het resultaat van de omzet van elke elementaire groep in het voorgaande jaar. De gewichten van de elementaire indices blijven onveranderd doorheen het jaar. De veranderingen wat betreft representatieve producten speelt zich af op productniveau, een subniveau van de elementaire indices. Door het gebruik van scannerdata wordt de keuze van representatieve producten objectief.

Voor de volgende aggregatiestap worden de indices per keten samengenomen aan de hand van de omzet van de respectievelijke ketens. Voor de gewichten, die een maat voor de omzet zijn, worden voor de verschillende ketens de jaarrekeningen of de BTW-aangiften gebruikt.

Voor producten die tijdelijk ontbreken, worden prijzen geïmputeerd. Dat is nodig om rekening te kunnen houden met de prijsevolutie tussen de laatste maand wanneer een product beschikbaar was en de maand wanneer het product terug beschikbaar wordt. De imputatie van de ontbrekende prijzen gebeurt met behulp van de prijsevolutie van de overeenkomstige producten. Er is geen impact op de index wanneer het product afwezig is, maar er wordt wel rekening gehouden met de "gemiste" prijsevolutie wanneer het product terug in het staal komt.

Hierna volgen twee voorbeelden over de hierboven beschreven procedure. Een eerste voorbeeld behandelt het bepalen van de steekproef. In het tweede voorbeeld wordt de imputatie bekeken.

Voorbeeld 1: steekproefbepaling

Volgende gegevens worden verondersteld: vier verschillende producten, waarbij product 2 slechts beschikbaar is vanaf februari en product 4 van de markt verdwijnt in maart.

Prijs	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1	6,00	6,30	6,20	6,40	6,00
product 2			4,50	4,60	4,40
product 3	5,20	5,40	5,20	5,30	5,10
product 4	5,70	5,40	5,80		

De omzetcijfers van de vier producten worden gegeven in de volgende tabel:

Omzet	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1	170	200	140	120	110
product 2			90	180	170
product 3	80	100	60	70	140
product 4	90	120	70		
Totaal	340	420	360	370	420

Product 2 kan in februari nog niet in de steekproef opgenomen worden aangezien er geen prijsevolutie is met betrekking tot januari. Op elk moment is n (aantal producten) dus gelijk aan drie, ook in februari waar vier producten beschikbaar zijn. Via de vermelde formule bedraagt de drempel voor het gemiddelde marktaandeel ongeveer 26,7%.

Met bovenstaande gegevens kan het marktaandeel voor de perioden sm en sm-1 berekend worden. Voor februari moet de berekening twee keer uitgevoerd worden. De eerste keer (a) voor de indexberekening tot en met februari, waarbij product 2 nog niet in de steekproef zit. De tweede keer (b) wordt het marktaandeel van februari herberekend in maart, omdat product 2 in maart wel gematched kan worden met februari. Product 4 valt dan echter uit de steekproef omdat het niet meer beschikbaar is in maart.

Marktaandeel	Dec-16	Jan-17	Feb-17 (a)	Feb-17 (b)	Mrt-17	Apr-17
product 1	50,0%	47,6%	51,9%	48,3%	32,4%	26,2%
product 2				31,0%	48,6%	40,5%
product 3	23,5%	23,8%	22,2%	20,7%	18,9%	33,3%
product 4	26,5%	28,6%	25,9%			

Nu kan bepaald worden welke producten twee opeenvolgende maanden (m en m-1) een gemiddeld marktaandeel hebben dat boven de drempel van 26,7% ligt. Die producten zullen opgenomen worden in de indexberekening van maand m.

Gemiddelde marktaandeel	Dec-16	Jan-17	Feb-17		Mrt-17	Apr-17
product 1		48,8%	49,7%		40,4%	29,3%
product 2					39,8%	44,6%
product 3		23,7%	23,0%		19,8%	26,1%
product 4		27,5%	27,2%			

Dat betekent dat product 1 de hele periode in de steekproef zit, terwijl vanaf maart product 4 in de steekproef zal vervangen worden door product 2.

De elementaire prijsindex op COICOP 6-niveau wordt berekend met de Jevons-kettingindex:

Index	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1		1,05	0,98	1,03	0,94
product 2				1,02	0,96
product 4		0,95	1,07		
Maandelijkse index		1,00	1,03	1,03	0,95
Kettingindex	100,00	99,74	102,54	105,33	99,75

<u>Voorbeeld 2</u>: imputatie

Zoals eerder aangehaald, worden de prijzen van producten die uitgesloten zijn van de steekproef geïmputeerd. Die imputatie gebeurt door vermenigvuldiging van de laatste prijs (wanneer het product wel nog tot de steekproef behoorde) met de Jevons-index van de COICOP 6-groep waartoe het product behoort. Ook wanneer een product een reële prijs heeft, maar uitgesloten is omdat het gemiddelde marktaandeel niet groot genoeg is (treshold), wordt de productprijs geïmputeerd. Dat is noodzakelijk omdat anders de prijswijziging tussen de laatste maand waarin het product in de steekproef zit en de maand waarin het opnieuw in de steekproef komt, genegeerd wordt.

Theoretisch gezien zorgt imputatie er ook voor dat de index zowel aan de identiteitstest als de transitiviteitstest voldoet. De identiteitstest controleert dat: indien de prijzen voor eenzelfde product in de steekproef in de huidige periode t gelijk zijn aan de prijzen in de referentie-/basisperiode 0, dan moet de index gelijk zijn aan 100. De index is transitief wanneer de koppeling van de maandelijkse kettingindices gelijk is aan de directe prijsvergelijking tussen periode 0 en t. Wanneer er geen imputaties zouden worden uitgevoerd, zou aan geen van beide tests voldaan zijn en is de gemeten prijsevolutie niet correct.

Opnieuw worden 4 producten verondersteld, waarbij product 2 en product 4 respectievelijk in februari en maart uitgesloten zijn van de steekproef hoewel ze toch verkocht worden. Voor de eenvoud wordt in dit voorbeeld geen rekening gehouden met de omzet. De prijzen in december en april zijn dezelfde, dus volgens de identiteitstest moet de kettingindex van december tot april gelijk zijn aan 100.

Prijs	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1	6,00	6,30	6,20	6,40	6,00
product 2	4,60	4,70		4,30	4,60
product 3	5,20	5,40	5,20	5,30	5,20
product 4	5,70	5,40	5,80		5,70

In de volgende tabel staan de maandelijkse indices zonder imputatie. De kettingindex in april is 104,02, dit is dus een afwijking van 4,02% van de index 100 die volgens de identiteitstest moet bekomen worden.

Indices	Dec-16	Jan-17	Feb-17	Mrt-7	Apr-17
product 1		1,05	0,98	1,03	0,94
product 2		1,02			1,07
product 3		1,04	0,96	1,02	0,98
product 4		0,95	1,07		
Maandelijkse index		1,01	1,01	1,03	0,99
Kettingindex	100,00	101,36	101,96	104,58	104,02

Eerst wordt de prijs van product 2 in februari geïmputeerd. Dat gebeurt door de laatste prijs, wanneer product 2 in de steekproef zat (januari), te vermenigvuldigen met de maandelijkse Jevons-index van februari.

Prijs	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1	6,00	6,30	6,20	6,40	6,00
product 2	4,60	4,70	4,73	4,30	4,60
product 3	5,20	5,40	5,20	5,30	5,20
product 4	5,70	5,40	5,80		5,70

Met die prijs voor product 2 in februari kan er nu in maart een index berekend worden voor dit product.

Indices	Dec-16	Jan-17	Feb-17	Mrt-7	Apr-17
product 1		1,05	0,98	1,03	0,94
product 2		1,02	1,01	0,91	1,07
product 3		1,04	0,96	1,02	0,98
product 4		0,95	1,07		
Maandelijkse index		1,01	1,01	0,99	0,99
Kettingindex	100,00	101,36	101,96	100,47	99,93

Nu wordt de ontbrekende prijs in maart van product 4 geïmputeerd:

Prijs	Dec-16	Jan-17	Feb-17	Mrt-17	Apr-17
product 1	6,00	6,30	6,20	6,40	6,00
product 2	4,60	4,70	4,73	4,30	4,60
product 3	5,20	5,40	5,20	5,30	5,20
product 4	5,70	5,40	5,80	5,72	5,70

Een verdere berekening van de index toont dat de index van april gelijk is aan die van december, waardoor de index voldoet aan de identiteitstest.

Indices	Dec-16	Jan-17	Feb-17	Mrt-7	Apr-17
product 1		1,05	0,98	1,03	0,94
product 2		1,02	1,01	0,91	1,07
product 3		1,04	0,96	1,02	0,98
product 4		0,95	1,07	0,99	1,00
Maandelijkse index		1,01	1,01	0,99	1,00
Kettingindex	100,00	101,36	101,96	100,47	100,00

Wanneer een directe prijsvergelijking wordt uitgevoerd tussen de prijzen van maart en december, resulteert dat in een index van 100,47. Dat is gelijk aan de hierboven berekende kettingindex, dus er is ook voldaan aan de transitiviteitstest.

3.3.2. Keuze voor een Jevons index

De bepaling van elementaire indexcijfers op het laagste geaggregeerde niveau gebeurt via een meetkundig gemiddelde of Jevons-index op niveau van elke keten:

$$P_J = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{1/n} = \frac{\prod_{i=1}^{n} (p_i^t)^{1/n}}{\prod_{i=1}^{n} (p_i^0)^{1/n}} \qquad \textit{Vgl. 2}$$

De reden waarom werd geopteerd voor een meetkundig gemiddelde (Jevons-index) in plaats van een rekenkundig gemiddelde (Dutot-index) is als volgt. Een Jevons-index heeft het voordeel dat het prijsniveau van de observaties geen invloed heeft op de index. Als een duurder product even sterk stijgt als een goedkoper product dan wegen beide prijsstijgingen evenveel door in de berekening van het indexcijfer. Dit is niet het geval bij de Dutot-index. Die index is namelijk gevoelig aan het prijsniveau van de observaties. Het gewicht van een prijs in de berekening van een rekenkundig gemiddelde is proportioneel aan het relatieve belang in het rekenkundige gemiddelde. Dat betekent dat de Dutot-index een groter gewicht geeft aan hogere prijzen dan aan lagere prijzen, denk bijvoorbeeld aan duurdere merkproducten ten opzichte van goedkopere producten. Ook zouden grotere verpakkingen belangrijker doorwegen in de inflatiemeting dan kleinere verpakkingen in de berekening van een Dutot-index.

Hieronder volgt een voorbeeld ter verduidelijking:

Veronderstel dat product B een duidelijk ander prijsniveau heeft dan producten A en C. En stel dat product A in prijs halveert terwijl product B verdubbelt in prijs en de prijs van product C constant blijft.

	December	Januari
Product A	4	2
Product B	20	40
Product C	5	5
Rekenkundig gemiddelde	9,67	15,67
Dutot-index	100	162,07

Door de tegenovergestelde prijsevolutie van product A en B zou de inflatie stabiel moeten blijven. Met het rekenkundig gemiddelde weegt product B echter zwaarder door dan producten A en C.

Bij gebruik van een Jevons-index wordt het resultaat:

	December	Januari
Product A	4	2
Product B	20	40
Product C	5	5
Meetkundig gemiddelde	7,37	7,37
Jevons-index	100	100

De Jevons-index heeft dus het voordeel dat de prijsniveaus van de prijsopnames niet doorwegen op de gemeten prijsevolutie, waardoor de geregistreerde inflatie niet vertekend wordt door de keuze voor goedkopere of duurdere producten.

Het is ook eenvoudig aan te tonen dat het meetkundig gemiddelde van de Jevons-index altijd kleiner is dan het rekenkundig gemiddelde van de Dutot-index (behalve in het geval wanneer alle prijzen gelijk zijn, dan zijn de twee gemiddelde gelijk). Neem het rekenkundig gemiddelde $a = \frac{x_1+x_2}{2}$, en het meetkundig gemiddelde $b = \sqrt{x_1 x_2}$ van twee prijzen x_1 en x_2. Dan is $b \leq a$ als $a - b \geq 0$. Invullen van de gemiddelden geeft:

$$a - b = \frac{x_1 + x_2}{2} - \sqrt{x_1 x_2} = \frac{1}{2}\left[\left(\sqrt{x_1}\right)^2 + \left(\sqrt{x_2}\right)^2 - 2\sqrt{x_1}\sqrt{x_2}\right] = \frac{1}{2}\left(\sqrt{x_1} - \sqrt{x_2}\right)^2$$

Dit is altijd ≥ 0, waardoor het gestelde bewezen is.

Ook vanuit economisch oogpunt gaat de voorkeur uit naar de Jevons-index. Die veronderstelt namelijk een constante substitutie-elasticiteit van 1. Substitutie-elasticiteit drukt uit in welke mate hoeveelheden relatief veranderen ten opzichte van relatieve prijswijzigingen. Bij de Jevons-index wordt deze substitutie-elasticiteit dus als constant beschouwd. Bij de Dutot-index wordt verondersteld dat er geen substitutie plaatsvindt (substitutie-elasticiteit gelijk aan 0), wat economisch onlogisch is.

Omwille van deze statistische en economische eigenschappen werd verkozen om te werken met een Jevons-index.

3.3.3. Kettingindex en matched modelling

De resulterende Jevons-index op elementair niveau is er dus een die berekend wordt met producten die in twee aanliggende perioden beschikbaar zijn en waarbij het bestedingsaandeel voor het product in kwestie boven de dynamisch bepaalde drempel ligt. De formule van de Jevons-index wordt dan de volgende:

$$P_J = \prod_{i \epsilon G_{0,1}} \left(\frac{p_i^1}{p_i^0}\right)^{1/N_{0,1}} \qquad \textit{Vgl. 3}$$

waarbij $G_{0,1}$ de steekproef is. Dat is de set van alle producten die matchen in twee opeenvolgende perioden (periode 0 en periode 1) en zich boven de drempel bevinden en waarbij $N_{0,1}$ het totale aantal producten in de steekproef is. In een kettingindex van meerdere perioden worden die maandelijkse indexcijfers aan elkaar gekoppeld door de resulterende indexcijfers met elkaar te vermenigvuldigen. Voor een Jevons-kettingindex met drie perioden wordt de formule dan de volgende:

$$P_J^{02} = \prod_{i \epsilon G_{0,1}} \left(\frac{p_i^1}{p_i^0}\right)^{1/N_{0,1}} \prod_{i \epsilon G_{1,2}} \left(\frac{p_i^2}{p_i^1}\right)^{1/N_{1,2}} \qquad \textit{Vgl. 4}$$

Deze indexcijfers voor een productgroep k, in het lopende jaar uitgedrukt met basisperiode december van het voorgaande jaar (=100), kunnen geschreven worden als $P_{J,k,dec=100}^{0t}$. Dat komt omdat de wegingscoëfficiënten telkens in december worden aangepast. De resulterende indexcijfers worden vervolgens geaggregeerd met de bijhorende gewichten die berekend worden op basis van de omzet die gehaald wordt uit de scannerdata van de desbetreffende productgroep in het voorgaande jaar:

$$P_A^{0t} = \frac{\sum_{k=1}^{n} w_k P_{J,k,dec=100}^{0t}}{\sum_{k=1}^{n} w_k} \qquad \textit{Vgl. 5}$$

De jaarlijkse aanpassing van de gewichten in december heeft geen rechtstreekse impact op het resulterende indexcijfer. Er is wel een onrechtstreekse impact omdat die aanpassing uiteraard wel een impact heeft op het aandeel van die productgroep in de gemeten prijsevolutie van het komende jaar.

3.3.4. Waarom ongewogen indexcijfers?

Op het elementaire niveau wordt er gebruik gemaakt van een ongewogen index. Dat kan in eerste instantie vreemd lijken omdat scannerdata omzetinformatie hebben tot op individueel productniveau. Met die omzetinfo zou overgestapt kunnen worden naar een superlatieve indexformule. In een superlatieve indexformule worden prijzen en hoeveelheden van zowel de referentieperiode als de verslagperiode symmetrisch behandeld.

Waarom wordt een superlatieve indexformule niet gebruikt en waarom wordt omzetinformatie überhaupt niet rechtstreeks gebruikt in de indexberekening, maar enkel indirect om op basis van de dynamische drempel de maandelijkse steekproef samen te stellen? De reden hiervoor is dat gebruik maken van die omzetinformatie in de indexberekening leidt tot chain drift. Chain drift betekent dat de kettingindex niet terugkeert naar eenheid (of een indexniveau van 100) wanneer de prijzen in de verslagperiode terugkeren naar het niveau van de referentieperiode. Het concept van chain drift zal verduidelijkt worden aan de hand van een Törnqvist-index, P_T , waarbij de formule gelijk is aan:

$$P_T^{0t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\left(\frac{p_i^0 q_i^0}{\sum_{j=1}^{n} p_j^0 q_j^0} + \frac{p_i^t q_i^t}{\sum_{j=1}^{n} p_j^t q_j^t}\right)}$$

$$= \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\left(s_i^0 + s_i^t\right)}$$

Vergelijking 6 geeft de Törnqvist-index weer voor de periode 0 tot t en is gelijk aan het gewogen meetkundig gemiddelde van de prijsratio's (huidige periode ten opzichte van basis), waarbij gewogen wordt op basis van het rekenkundig gemiddelde van de uitgavenaandelen in de basisperiode en de huidige periode. Voor een twee-perioden Törnqvist-kettingindex wordt de formule dan als volgt:

$$P_{T,ketting}^{02} = \prod_{i=1}^{n} \left(\frac{p_i^1}{p_i^0}\right)^{0.5\left(s_i^0 + s_i^1\right)} \prod_{i=1}^{n} \left(\frac{p_i^2}{p_i^1}\right)^{0.5\left(s_i^1 + s_i^2\right)}$$

Stel nu dat er slechts 1 product *a* is (*i* = *a*) gevolgd wordt. Dan kan de vergelijking herleid worden naar volgende vorm:

$$P_{T,ketting}^{02} = \left(\frac{p_a^1}{p_a^0}\right)^{0.5\left(s_a^0 + s_a^1\right)} \left(\frac{p_a^2}{p_a^1}\right)^{0.5\left(s_a^1 + s_a^2\right)}$$

In het geval dat dit product *a* tijdelijk in promotie is in periode 1 ($p_a^0 > p_a^1$), maar terugkeert naar het prijsniveau van periode 0 in periode 2 ($p_a^0 = p_a^2$) dan kan vergelijking 8 herschreven worden als:

$$
\begin{aligned}
\mathrm{P}_{\mathrm{T,ketting}}^{02} &= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 + s_a^1)} \left(\frac{p_a^0}{p_a^1}\right)^{0.5\,(s_a^1 + s_a^2)} \\
&= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 + s_a^1) - 0.5\,(s_a^1 + s_a^2)} \\
&= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 - s_a^2)}
\end{aligned}
$$

Aangezien de index in theorie (cfr. identiteitstest) terug naar eenheid dient te komen omdat het prijsniveau in periode 2 gelijk is aan dat van periode 0 kan dit, zoals uit vergelijking 9 blijkt, enkel gebeuren in het geval dat $s_a^0 = s_a^2$. Dat betekent dus dat de uitgaven in periode 2 en periode 0 gelijk moeten zijn omdat de prijsniveaus in die perioden identiek zijn. In standaard micro-economische theorie wordt er vanuit gegaan dat de hoeveelheden voor een gegeven set van prijzen uniek bepaald zijn. Indien dat het geval zou zijn, dan zouden in periode 2 en periode 0 de gewichten identiek zijn.

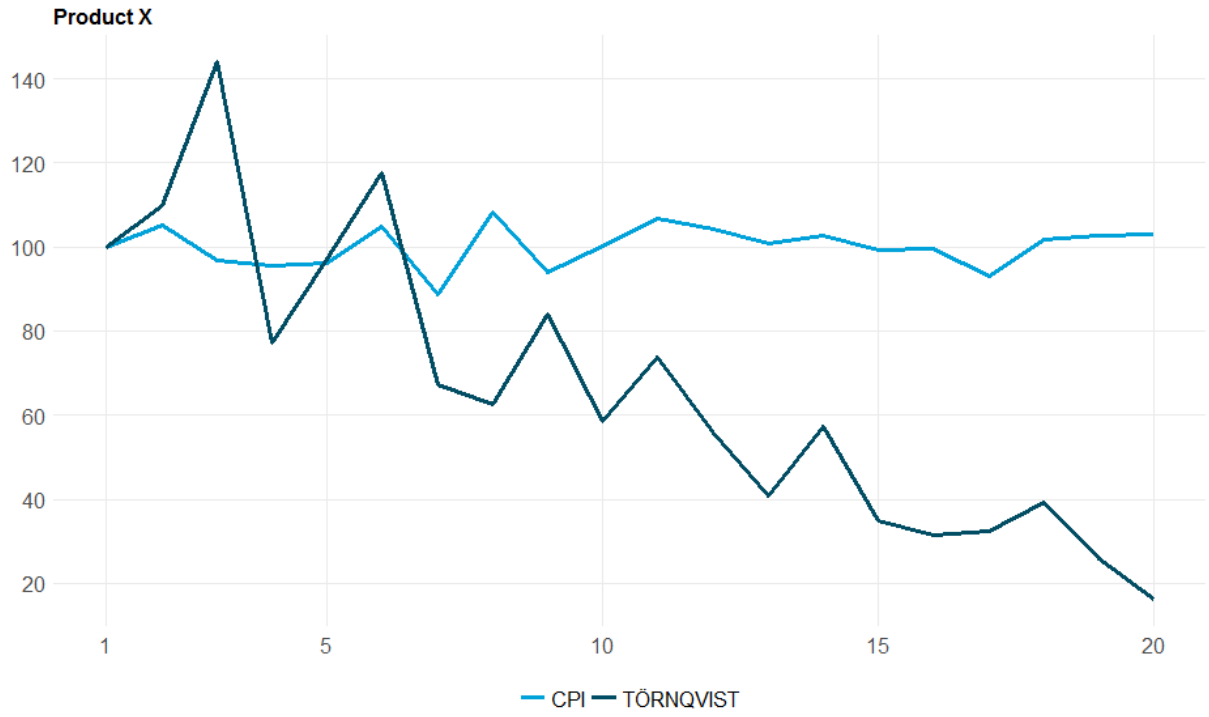
In het geval deze theoretische voorspelling niet gerealiseerd wordt dan is $P_{T,ketting}^{02} < 1$ **wanneer** $s_a^0 > s_a^2$ en $P_{T,ketting}^{02} > 1$ in het geval dat $s_a^0 < s_a^2$. Een eenvoudig fictief voorbeeld waar de Törnqvist-kettingindex niet terugkeert naar eenheid, wordt in onderstaande tabel gegeven voor producten A en B voor een periode van vier maanden. De prijs en omzet van product B wordt stabiel gehouden gedurende de hele periode, de prijs en omzet van product A fluctueert daarentegen.

Tabel 8: Fictief voorbeeld van chain drift bij Törnqvist kettingindex

	Januari		Februari		Maart		April	
	Prijs	Omzet	Prijs	Omzet	Prijs	Omzet	Prijs	Omzet
Product A	2,50	10.000	2,00	500.000	2,50	2.000	2,50	10.000
Product B	3,00	10.000	3,00	10.000	3,00	10.000	3,00	10.000
Törnqvist kettingindex	100,00		84,78		96,35		96,35	

In het voorbeeld uit de tabel is product A in februari in promotie waardoor de omzet sterk stijgt in die maand, met een dalend effect op de index tot gevolg (een daling van 100 tot 84,78). In maart zijn de prijsniveaus teruggekeerd naar het niveau van januari. De omzet voor product A is echter niet gelijk aan die van januari (bijvoorbeeld ten gevolge van voorraadvorming door consumenten tijdens de promotie in februari), waardoor de index niet terugkeert naar het verwachte niveau van 100. Dat laat voorts een permanent effect na zoals uit de index van april is af te leiden. Om dat permanente effect aan te tonen zijn in april zowel de omzetgegevens als de prijzen identiek aan die van januari, waardoor er geen twijfel kan ontstaan dat dat een index van 100 zou moeten opleveren. Dat is echter niet het geval. De index blijft op de waarde 96,35 omdat de prijsverhouding niet wijzigt. Dat fenomeen wordt *chain drift* genoemd.

In de onderstaande figuur wordt de chain drift van de Törnqvist-kettingindex visueel voorgesteld, de vermelde CPI-cijfers zijn de officiële indexcijfers van het product X:



Figuur 13: Chain drift bij Törnqvist-kettingindex

Hierin is duidelijk dat bij de Törnqvist-index een neerwaartse drift optreedt.

De vraag is nu of die *chain drift* eerder in opwaartse dan in neerwaartse zin zal gebeuren. Gaat het over een index die naar beneden tendeert (richting nul) over lange perioden of één die opwaarts gaat (richting oneindig)?

In eerste instantie wordt verwacht dat de index naar nul zou naderen, omdat de component waar de prijsdaling optreedt een hoger gewicht krijgt door de sterk gestegen omzet. Dat in tegenstelling tot de prijsstijging in de volgende periode, die minder gewicht krijgt omdat consumenten minder aankopen dan normalerwijze verwacht zou worden (door bijvoorbeeld voorraadvorming of substitutie). Het is ook mogelijk dat er opwaartse drift is, alhoewel die in praktijk minder vaak voorkomt. Opwaartse drift kan gebeuren wanneer een promotie plaatsvindt aan het einde van de referentieperiode en wanneer die deels doorloopt in de verslagperiode. De gestegen prijs in de verslagperiode krijgt op die manier een hoger gewicht met een opwaartse drift tot gevolg.

Omwille van chain drift werd er dus geopteerd om te werken met een ongewogen Jevons-index op het laagste niveau. De Jevons-index is in feite een speciale variant van de hierboven vermelde Törnqvist-index, waarbij de gewichten van alle goederen in een elementair aggregaat overeenkomen. Hierdoor is het gewicht van goed j in periode 0 $\left(s_j^0\right)$ gelijk aan het gewicht in periode t $\left(s_j^t\right)$, alsook de gewichten van alle goederen 1 tot n in het elementair aggregaat zijn gelijk aan elkaar. Dat betekent dan ook dat $s_j^0 = s_j^t = 1/N$ waardoor de Törnqvist-formule in dit geval gelijk is aan:

$$P_T^{0t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\left(s_j^0 + s_j^t\right)} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{\left(s_j^0\right)} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{n}}, \qquad \textbf{Vgl. 10}$$

wat resulteert in de Jevons-index.

3.3.5. Dumping- en outlierfiltering

Er wordt ook nog een dumpingfilter en outlierfilter toegepast, alvorens de index wordt berekend. De dumpingfilter zorgt dat producten die een sterke daling in zowel prijs als verkochte hoeveelheid kunnen uitgesloten worden van de steekproef. Daarmee wordt voorkomen dat die producten de index negatief beïnvloeden (downward bias). Omdat het een matchedmodel-kettingindex is (zie hierboven), verlaten die producten de index tegen een lage prijs, maar komen later nooit terug. Het opnemen van die producten in de berekening zou een neerwaartse drift in de index veroorzaken.

Op een analoge manier worden producten die maandelijks extreme prijsveranderingen via een outlierfilter uitgesloten van de steekproef. In praktijk vangt de outlierfilter voornamelijk producten op die verkregen kunnen worden op basis van een spaarkaart.

Indien de outlier- en dumpingfilter onterecht producten zouden uitsluiten heeft het geen langetermijnimpact op de index omdat de prijzen van uitgesloten producten geïmputeerd worden. Het is dus beter om te voorzichtig te zijn en producten uit te sluiten en te imputeren dan onvoorzichtig te werk te gaan en producten op te nemen in de berekening die een drift kunnen veroorzaken.

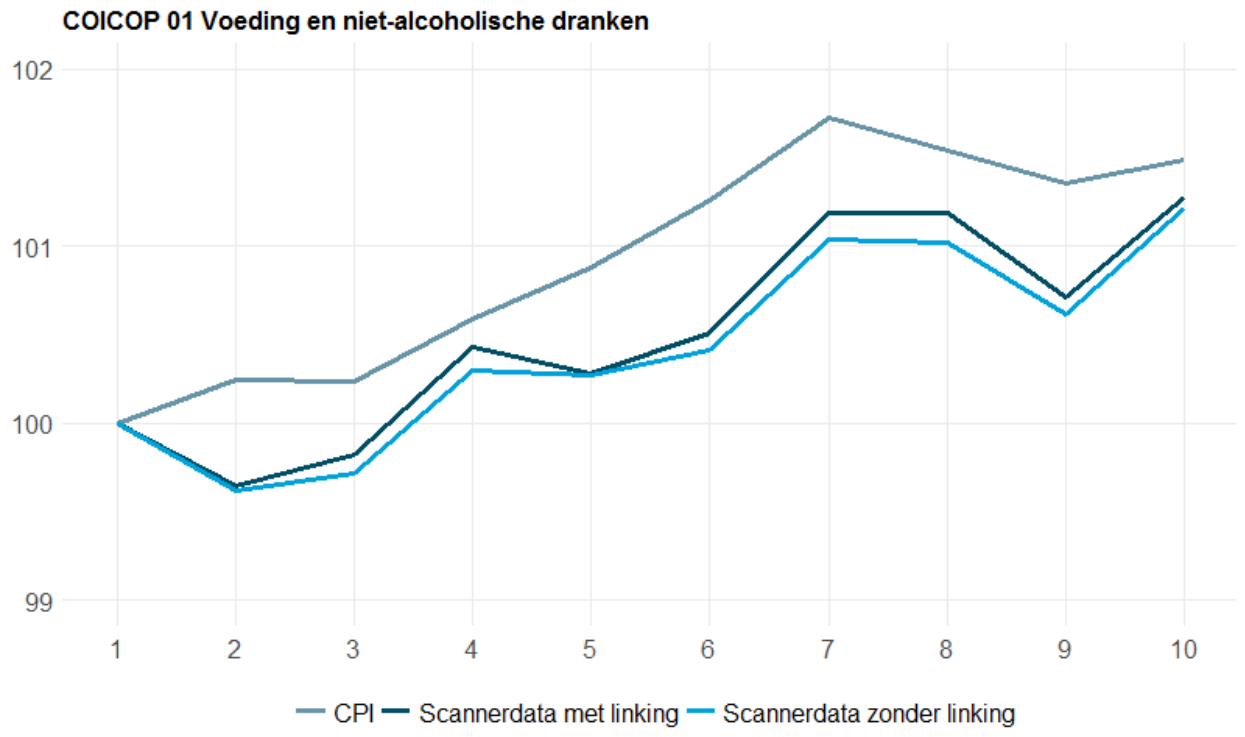
3.4. Rebranding en replacements

Maandelijks wordt ook de link gelegd tussen verdwenen en nieuwe producten. Een product kan immers een andere barcode krijgen, terwijl het nog steeds identiek hetzelfde product is. Ook kan een product een andere barcode krijgen door gewijzigde inhoud (bv. x gram extra) of promotie (bv. x% gratis). Soms krijgen producten een andere verpakking waardoor het product een nieuwe interne code krijgt. Die moet dan gelinkt worden met de oude code aangezien het wel nog steeds over hetzelfde product gaat.

Hieronder staan twee voorbeelden van producten die vervangen zijn:

Maand	COICOP	Groep	Oud product	Nieuw product	Coëff.
1	01.1.1.4.02	Koekjes	Merk X - 6 + 2 gratis - Chocolade	Merk X - 225 g - Chocolade	0.75
2	01.1.4.5.10	Verse Kaas	Merk Y - 200 g - Verse kaas light	Merk Y - 300 g - Verse kaas light	1.50

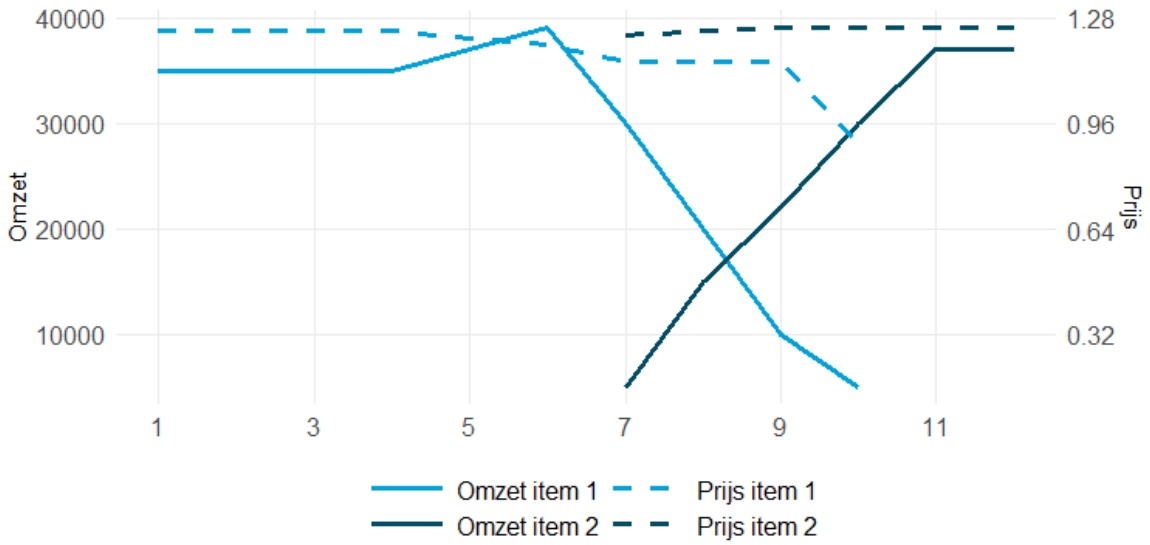
In het eerste voorbeeld had het oude product een tijdelijke promotie van twee gratis koekjes, de coëfficiënt is dan 0,75 (nieuwe hoeveelheid / oude hoeveelheid = 6 / 8). Bij het tweede voorbeeld is de inhoud gewijzigd van 200gr naar 300gr. Wanneer de linking van het eerste voorbeeld niet zou gemaakt worden, zou dat niet zo’n grote invloed hebben op de langetermijnindex omdat het slechts gaat over een tijdelijke promotie. Het tweede voorbeeld kan echter wel een impact hebben op de langetermijnindex wanneer die linking niet zou gemaakt worden.



Figuur 14: Invloed van linking op index (COICOP 01)

Bij de klassieke prijsopnames weet de enquêteur welk product vorige maand gekozen werd en kan dus een opvolger bepaald worden, met eventueel een kwaliteits- of kwantiteitsaanpassing. Voor scannerdata wordt die manier van handelen gereproduceerd.

Vervangingen en herlanceringen van producten verlopen vaak volgens een bepaald patroon, zoals blijkt uit onderstaande figuur.



Figuur 15: Voorbeeld van vervanging en herlancering van producten

Als een product (Item 1) vervangen wordt door een nieuw product (Item 2), met een iets hogere prijs, daalt de verkochte hoeveelheid en de omzet van het oude product en stijgt dit voor het nieuwe product.

De vraag is: in welke maand moet de vervanging doorgevoerd worden? Producten blijven slechts representatief als de omzet in de nieuwe maand groot genoeg is. Bij traditionele prijsopnames moet de enquêteur zelf beslissen welke producten representatief zijn, waardoor het mogelijk is dat prijzen verzameld worden zolang het product beschikbaar is hoewel het eigenlijk niet meer representatief is. In tegenstelling tot de traditionele methode, is via scannerdata meer info beschikbaar, waarmee duidelijk is welke producten representatief zijn. Door de beschikbaarheid van omzetinformatie wordt het duidelijk dat de representativiteit van item 2 stijgt in het nadeel van item 1, en kan er beslist worden om een vervanging door te

voeren en een linking te maken tussen item 1 (oud product) en item 2 (nieuw product). Scannerdata maken zo vervangingen zichtbaar.

Elke maand wordt er voor elke supermarktketen voor elke COICOP 6-groep een lijst aangemaakt, die zowel de 'nieuwe' producten in de steekproef als de verdwenen producten bevat. Voor de verdwenen producten wordt er gekeken naar de huidige maand en de drie voorgaande maanden omdat het verdwijnen van een product en het op de markt brengen van een nieuw (de opvolger) niet noodzakelijk op hetzelfde moment gebeurt (de stock van het oude product moet eerst verkocht worden). Tijdens die overgangsperiode verkleint het marktaandeel van het oude product en verdwijnt het uit de sample, terwijl het nieuwe product een groeiend marktaandeel heeft maar nog niet in de sample voorkomt.

Om de codes te linken is er een applicatie gecreëerd. Ook wordt er bij hoeveelheidsverschillen automatisch een coëfficiënt berekend op basis van de gegevens van de scannerdata. De coëfficiënt is gelijk aan het quotiënt van de nieuwe hoeveelheid en de oude. De oude prijs wordt dan vermenigvuldigd met deze coëfficiënt om het hoeveelheidsverschil te corrigeren. Indien nodig kan die coëfficiënt aangepast worden. Als er geen link wordt gemaakt tussen het oude en het nieuwe product wordt er verondersteld dat er een kwaliteitsaanpassing is. De berekening gebeurt dan aan de hand van standaard bridged overlap (analoog aan imputatie). In dat geval wordt de prijsevolutie van vergelijkbare producten als basis genomen voor de schatting van het prijsverloop van het nieuwe en het vervangen product. Er wordt dan voor het nieuwe product een prijs geschat (geïmputeerd) voor de vorige maand op basis van de prijsevolutie tussen de huidige en de vorige maand voor vergelijkbare producten.

Naast de algemene lijst met nieuwe en verdwenen producten, worden er nog 3 andere lijsten aangemaakt. Een eerste lijst is de 'perfect linking'-lijst. Daarin staan de nieuwe en verdwenen producten waarvan de productbeschrijvingen identiek zijn. Producten uit die lijst kunnen altijd worden gelinkt en de linking wordt automatisch doorgevoerd. Een tweede lijst is de 'reverse linking'-lijst. Die lijst toont de linkings die in het verleden gebeurd zijn en nu in de omgekeerde zin gebeuren. Bijvoorbeeld een product met een tijdelijke promotie (6+2) in maand t en het normale product (6) keert terug in maand t+1. In maand t is het product (in tijdelijke promotie) gelinkt aan het (normale) product in maand t-1. De omgekeerde (reverse) linking verschijnt dan automatisch in maand t+1 op deze lijst en kan in de applicatie ingevoerd worden. De derde lijst is een 'fuzzy linking'-lijst. Deze lijst probeert gelijkaardige producten te linken op basis van de beschrijvingen (die niet identiek zijn). Hiervoor worden eerst onbelangrijke woorden uit de beschrijving verwijderd waarna een score berekend wordt op basis van het aantal gemeenschappelijke woorden tussen twee producten (X en Y):

$$\text{score} = \frac{\text{\# gemeenschappelijke woorden}}{\sqrt{\text{\# woorden in X} * \text{\# woorden in Y}}}$$

De score ligt tussen 0 en 1 en enkel de scores hoger dan 30% worden opgenomen in de 'fuzzy linking'-lijst.

3.5. Seizoensproducten

Seizoensproducten zijn de goederen en diensten die niet beschikbaar zijn voor aankoop (of slechts in kleine hoeveelheden) voor bepaalde maanden in een jaar. Er moet sprake zijn van een typisch patroon, dus producten die toevallig niet beschikbaar zijn tijdens een maand behoren niet tot de seizoensproducten. Voor scannerdata van supermarktketens gaat het over verse producten (verse groenten, vers fruit en verse zeevruchten). Andere seizoensproducten die slechts een heel korte periode beschikbaar zijn zoals kerstartikelen of paaseieren worden uitgesloten van de indexberekening. De reden hiervoor is dat deze producten slechts heel kort beschikbaar zijn, verder zou het opnemen van deze producten in de indexberekening een neerwaartse drift veroorzaken.

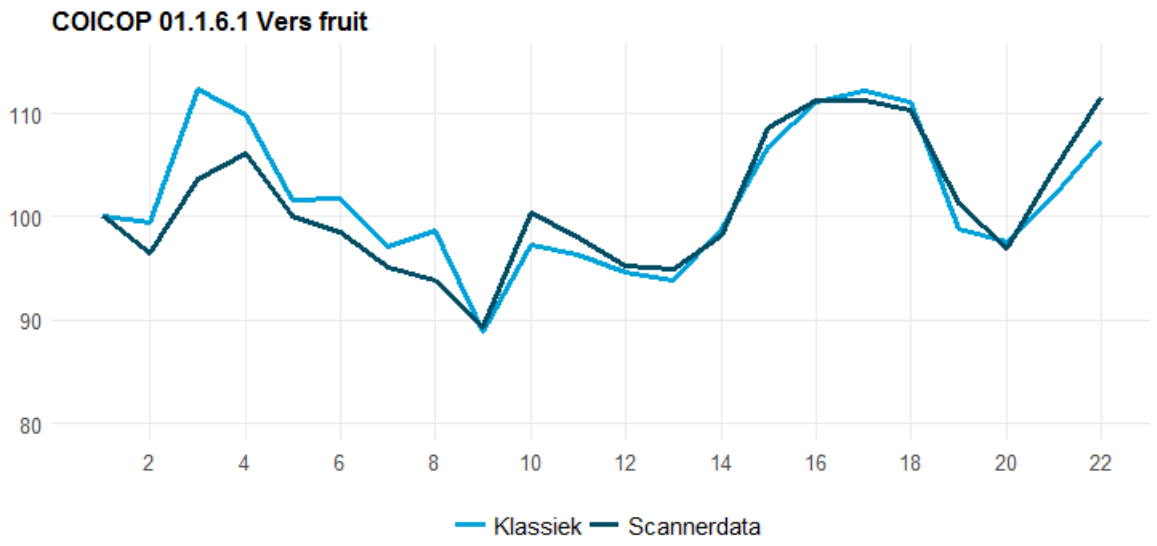
De neerwaartse drift wordt veroorzaakt omdat relatieve prijzen van twee opeenvolgende maanden gebruikt worden in de indexberekening (matched model, zie hoofdstuk 3.3.3). Voor deze artikelen zou de prijs die in de indexberekening komt eigenlijk die zijn van de periode nadat het “seizoen” reeds voorbij is en de prijzen dus dumpingprijzen zijn (en niet de representatieve prijzen tijdens “het seizoen”). De kans dat dezelfde producten het jaar nadien terugkomen is vrij klein, dus de index zou geen “bounce” terug naar boven krijgen.

Omdat niet alle verse producten maandelijks beschikbaar zijn, is het gebruik van een dynamische korf hier minder interessant. De COICOP 6-digits groepen zouden dan heel heterogeen (bv. een gemengde groep van aardbeien en bananen) moeten samengesteld worden, want anders zijn er groepen die gedurende een bepaalde periode (seizoen) geen producten bevatten. De COICOP 6-digits groepen hebben echter vaste gewichten waardoor er dan maandenlange imputatie nodig is, wat in strijd is met de HICP-regelgeving. Voor verse producten wordt daarom gebruik gemaakt van de “class confined seasonal weights”-methode in plaats van een dynamische korf. Die methode heeft als doel de variatie in de wegingscoëfficiënten per maand te beperken en tegelijkertijd er voor te zorgen dat een product als wegingscoëfficiënt 0 krijgt in de maanden dat het niet beschikbaar is. Dat wordt vertaald in de volgende eigenschappen, terwijl in bijlage een fictief voorbeeld wordt geschetst:

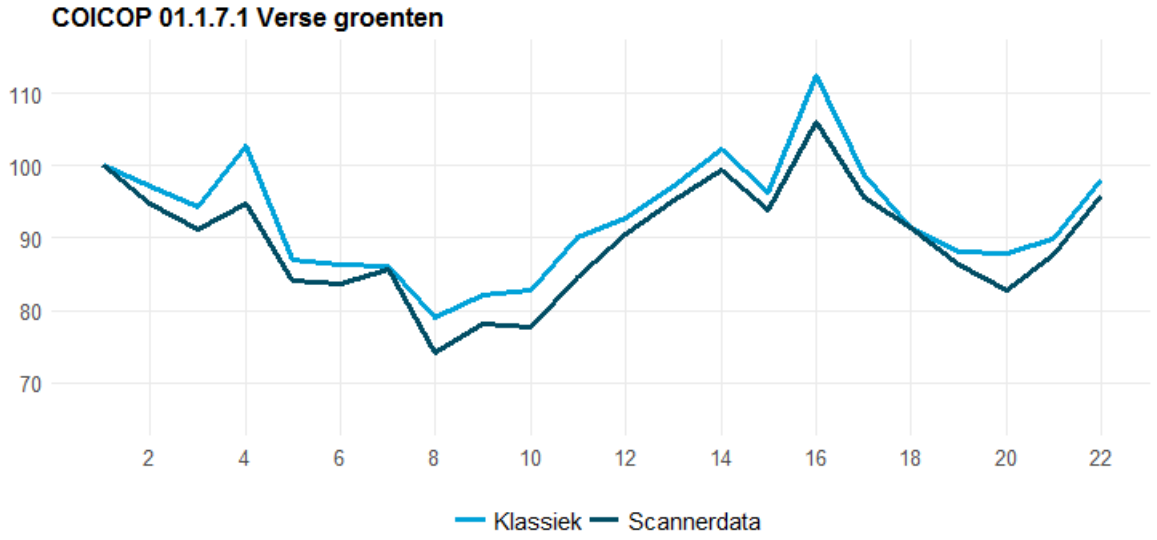
- een COICOP 6-digits groep heeft gewicht 0 wanneer het niet beschikbaar is (out-of-season);
- het gemiddelde gewicht van de COICOP 6-digits groep gedurende het jaar is overeenkomstig met het jaarlijkse aandeel van deze productgroep in het totaal van de bestedingen van de bovenliggende productgroep. Dat wil dus zeggen dat het gemiddelde gewicht op jaarbasis van aardbeien in de korf gelijk is aan het procentuele aandeel van aardbeien in de totale bestedingen van de opgenomen productgroepen in de COICOP 5-digits groep voor vers fruit voor het ganse jaar;
- de ratio van de gewichten van de COICOP 6-digits groepen die ‘in-season’ zijn blijft gelijk zolang beide groepen beschikbaar zijn, ook wanneer de individuele gewichten wijzigen doordat andere groepen al dan niet ‘in-season’ zijn. Zo zal de verhouding van het gewicht van bananen en aardbeien altijd hetzelfde zijn wanneer beide producten ‘in-season’ zijn, ook al komen andere producten ‘in-season’ of verdwijnen producten ‘out-season’. De stabiliteit van de ratio van gewichten zorgt dat enkel zuivere prijsevolutie gemeten wordt in de index, wijzigingen in de gewichten van maand tot maand hebben geen rechtstreekse impact op de gemeten prijsevolutie.

De gewichten worden bepaald op basis van de omzetcijfers van het voorgaande jaar. Om te bepalen of een COICOP 6-digits groep al dan niet ‘in-season’ is, wordt er gekeken naar de data van de twee voorgaande jaren. Deze “class confined seasonal weights”-methode wordt ook gebruikt voor verse producten in de klassieke prijsopnames.

De onderstaande grafieken geven de indexevolutie weer van de COICOP-groepen vers fruit en verse groenten, berekend via de klassieke methode en via de gegevens verzameld met scannerdata. Voor beide COICOP-groepen valt een gelijkaardige evolutie op te merken tussen de twee methodes.



Figuur 16: Vergelijking prijsevolutie tussen klassieke methode en scannerdata voor COICOP 01.1.6.1 Vers fruit

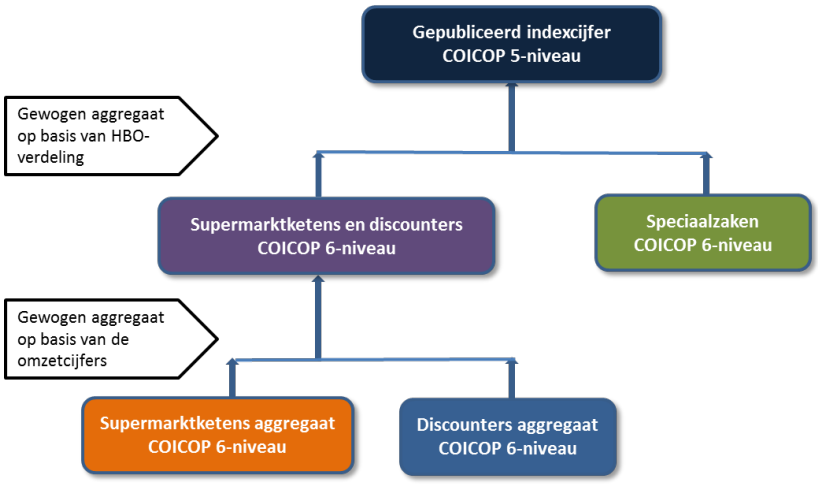


Figuur 17: Vergelijking prijsevolutie tussen klassieke methode en scannerdata voor COICOP 01.1.7.1 Verse groenten

4. STRATIFICATIEMODEL

Om de effectieve combinatie tussen scannerdata en klassieke prijsopnames uit te voeren wordt een stratificatiemodel gebruikt. De gegevens worden gecombineerd op COICOP 5-niveau.

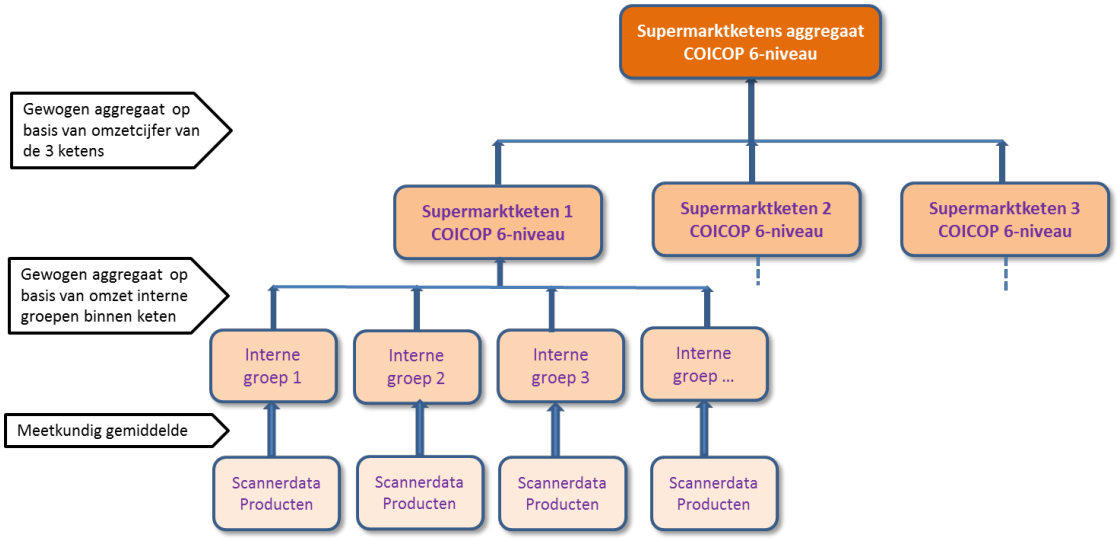
Er wordt een onderscheid gemaakt tussen “supermarkten en discounters” en “speciaalzaken” (slagers, bakkers,…). Elk van beide strata heeft een gewicht op basis van het aankoopgedrag van de consument. Aan de hand van het tweejaarlijkse huishoudbudgetonderzoek (HBO) worden deze gewichten elke twee jaar geüpdatet.



Figuur 18: Stratificatiemodel op COICOP 5-niveau

De tak “supermarkten en discounters” wordt opgesplitst in twee groepen, 1 voor supermarkten (waar scannerdata worden gebruikt) en 1 voor discounters ((nog) geen scannerdata beschikbaar). Elke supermarkt en discounter krijgen een gewicht toegekend op basis van omzetcijfers. Die gewichten worden jaarlijks vernieuwd.

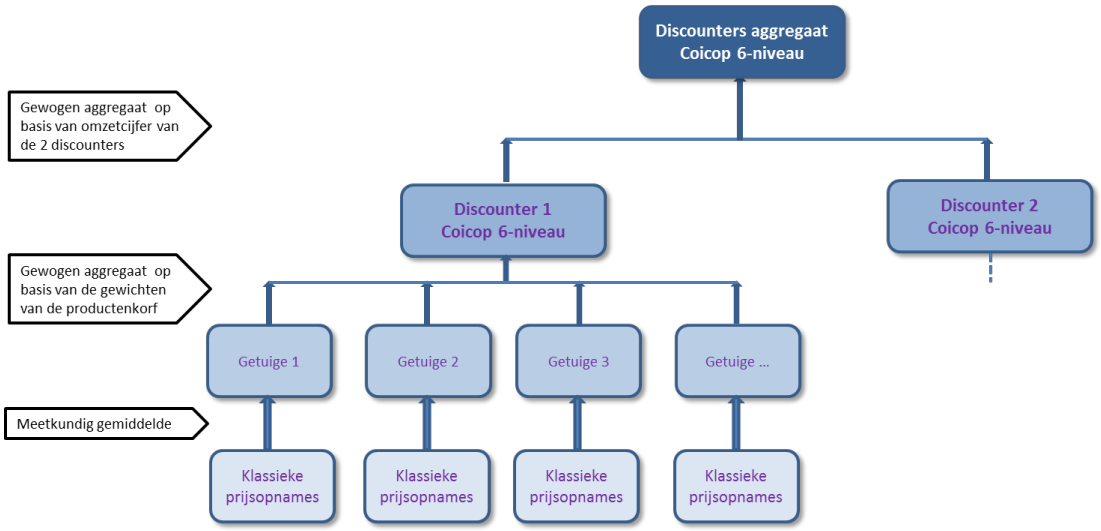
De groep supermarkten wordt onderverdeeld in de drie ketens waarvoor er scannerdata worden ontvangen. Elke keten krijgt een gewicht op basis van hun jaarrekeningen (jaarlijkse update). Zoals hierboven reeds vermeld worden er consumptiesegmenten opgesteld voor elke keten. Elk van die COICOP-groepen krijgt een gewicht op basis van de scannerdata. Dat gewicht wordt jaarlijks vernieuwd gebruik makend van de omzetcijfers van het voorgaande jaar. De index op consumptiesegment-niveau wordt berekend met een Jevons-index (meetkundig gemiddelde).



Figuur 19: Stratificatiemodel van supermarktketens op COICOP 6-niveau

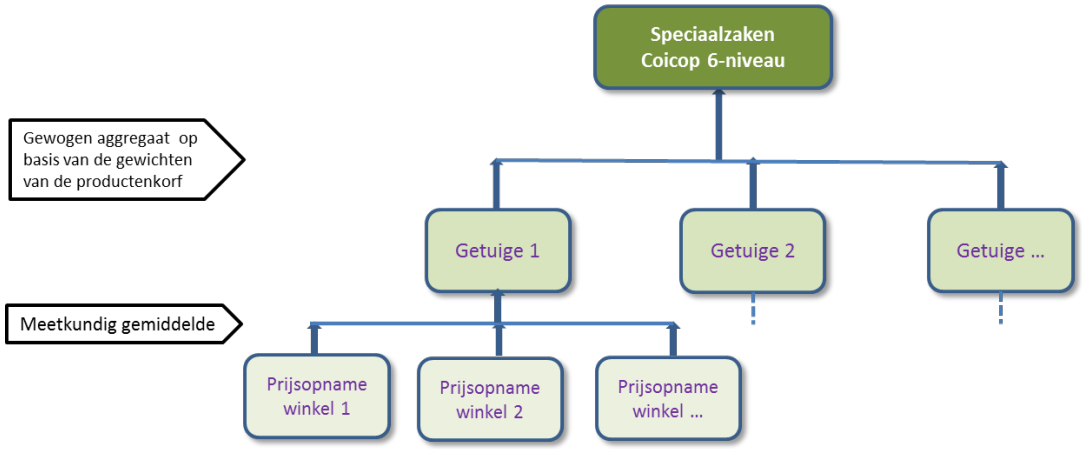
In België zijn er twee discounters (Aldi en Lidl). Ook hun gewicht wordt jaarlijks geüpdatet. Aan de hand van de klassieke methode worden de prijzen verzameld voor de traditionele COICOP 6-aggregaten. Die krijgen gewichten op basis van het

huishoudbudgetonderzoek. Jaarlijks worden deze gewichten herzien als gevolg van een prijsupdate of toevoeging van een nieuw COICOP 6-aggregaat. Voor de HICP wordt er webscraping toegepast voor één discounter.



Figuur 20: Stratificatiemodel van discounters op COICOP 6-niveau

Naast de supermarkten en discounters zijn er ook nog speciaalzaken, waar de enquêteurs prijzen opnemen volgens de traditionele methode. Een Jevons-index wordt vervolgens gebruikt voor aggregatie. Daarna worden de gewichten, gebaseerd op het HBO, toegepast op het COICOP 6-niveau.

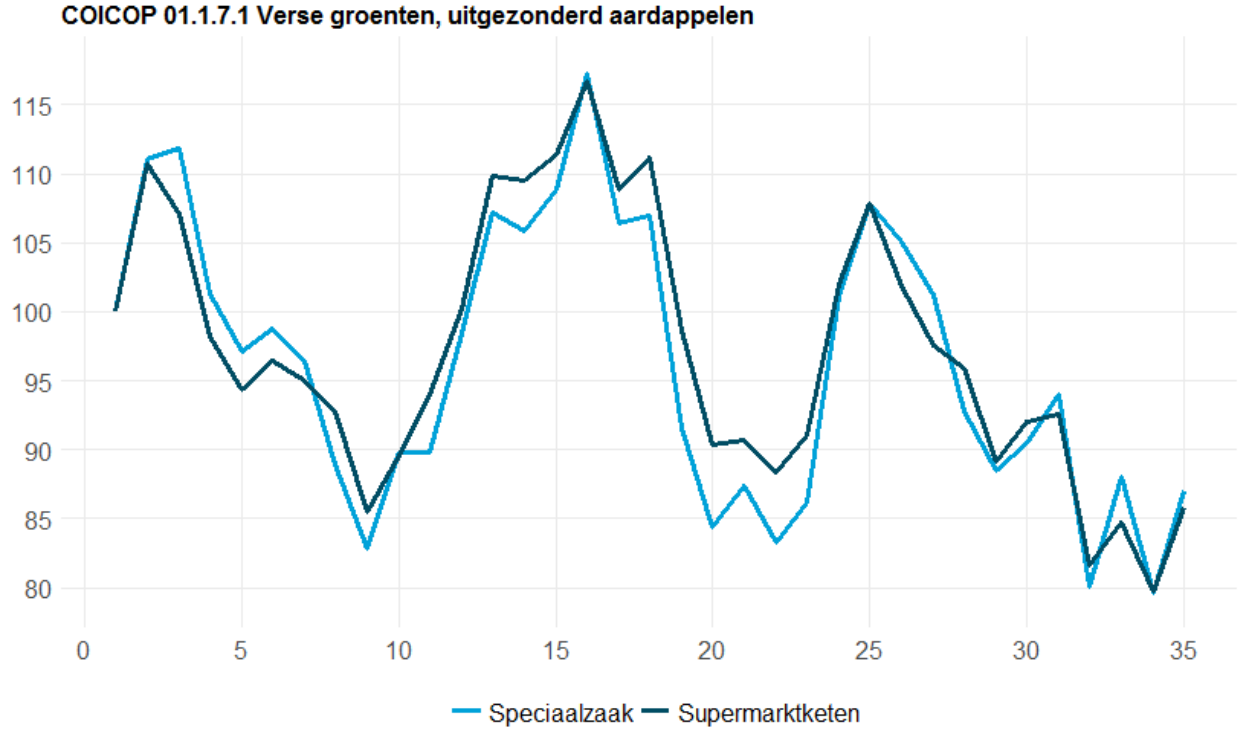


Figuur 21: Stratificatie van speciaalzaken op COICOP 6-niveau

Bij de speciaalzaken worden echter slechts prijzen opgenomen in het geval het marktaandeel (volgens HBO) groter is dan 10%. De onderstaande tabel geeft een overzicht van de COICOP-groepen waarvoor nog traditionele prijsopnames gebeuren:

COICOP	Groep	Supermarkten & discounters	Speciaalzaken
01.1.1.3	Brood	53,80%	46,20%
01.1.2.8	Andere vleesproducten	74,37%	25,63%
01.1.2.1	Runds- en kalfsvlees	76,23%	23,77%
01.1.2.5	Ander vlees	76,45%	23,55%
01.1.2.2	Varkensvlees	77,67%	22,33%
01.1.2.7	Bereide vleeswaren	78,55%	21,45%
01.1.1.4	Andere bakkerijproducten	80,19%	19,81%
01.1.2.3	Schaaps- en lamsvlees	80,87%	19,13%
01.1.2.4	Gevogelte	84,35%	15,65%
01.1.3.1	Verse vis	87,73%	12,27%

Voor de groepen waar het marktaandeel van de speciaalzaken onder 10% ligt (bv groenten en fruit), wordt er wel een sterke correlatie geobserveerd tussen de index gebaseerd op de prijzen uit supermarktketens en de index gebaseerd op prijzen uit de speciaalzaken. Hieronder staat een voorbeeld voor COICOP 01.1.7.1 Verse groenten, uitgezonderd aardappelen. Hieruit kan geconcludeerd worden dat de gemeten prijsevolutie tussen supermarktketens en speciaalzaken nagenoeg identiek is.



Figuur 22: Vergelijking prijsevolutie tussen supermarktketens en speciaalzaken

5. CONCLUSIE

Statbel ontvangt van drie supermarktketens data via SFTP. Automatisch wordt er een lijst gemaakt met nieuwe producten, die gelinkt moeten worden aan de juiste COICOP. Door gebruik te maken van de interne classificatie wordt er een COICOP-consumptiesegment voorgesteld. Aan de hand van machine learning wordt er een tweede lijst met COICOP-suggesties opgemaakt. Na controle wordt er een bestand aangemaakt met de productcodes en hun overeenkomstige COICOP-classificatie. Daarna wordt er een voorafgaande indexberekening (op de scannerdata) uitgevoerd om de dynamische korf te bepalen. Hierbij worden ook de lijsten aangemaakt om de herlanceringen te linken. De correcte herlanceringen van producten worden daarna opgeladen in de data warehouse, waarna de finale indexberekening kan worden uitgevoerd die gebaseerd is op een Jevons-kettingindex. De resulterende scannerdata-indexcijfers worden geïntegreerd in de volledige indexberekening via een stratificatiemodel.

Het gebruik van scannerdata heeft er voor gezorgd dat de inflatiemeting verbeterd is ten opzichte van vroeger. Door gebruik te maken van de omzetinformatie, die beschikbaar is in scannerdata, worden vrijwel alle representatieve producten opgenomen in de indexberekening via de dynamische korf. Bovendien kan de gewichtsbepaling nauwkeuriger gebeuren op het laagste niveau in de indexberekening. Een eenvoudig voorbeeld illustreert dit . In de groep ‘suiker’ werd tot eind 2015 de prijsevolutie voor één product gevolgd (weliswaar voor verschillende merken en verschillende verkooppunten), namelijk de kristalsuiker in een verpakking van 1 kg. Vanaf januari 2016 wordt de prijsevolutie van alle suikersoorten en verpakkingen, die een representatief verkoopcijfer halen, gevolgd. Dat betekent dat, naast de prijsevolutie voor kristalsuiker, ook de prijsevolutie van klontjessuiker, rietsuiker, kandijsuiker, bruine suiker, poedersuiker en vloeibare suiker opgemeten en verrekend wordt onder de deelindex voor ‘suiker’. Hierbij wordt ook rekening gehouden met de verschillende verpakkingen (250 g, 500 g, 750 g, 1 kg, …) die bestaan. Het is duidelijk dat deze werkwijze de representativiteit van het prijsindexcijfer voor de betreffende groepen sterk ten goede komt.

Voor de overige supermarktketens en kleinhandelaars, waarvan momenteel nog geen scannerdata worden ontvangen, blijven gedeeltelijk klassieke prijswaarnemingen in de winkels (of webscraping) noodzakelijk om een betrouwbaar beeld te krijgen van de totale prijsevolutie in België. De scannerdata worden momenteel ook gebruikt om de steekproef van de producten in die klassieke prijsopnames representatief te houden en hebben op die manier dus ook daar een impact. Voorts lopen er gesprekken met de overige supermarktketens om ook van hen scannerdata te kunnen verwerken. Ten slotte bekijkt Statbel ook een uitbreiding met andere sectoren.

De huidige methode gebruikt op het laagste niveau nog geen expliciete wegingsinformatie, echter wordt hier onderzoek over uitgevoerd door Statbel en andere statistiekbureaus om eventueel deze toch te kunnen gebruiken op het laagste niveau zonder dat “chain drift” optreedt. Deze methoden werden hier buiten beschouwing gelaten, aangezien deze zich nog in de onderzoeksfase bevinden.

BIJLAGE

In de onderstaande tabel staan een overzicht van alle COICOP-groepen waarvoor scannerdata gebruikt worden en het gewicht (2017) dat bij de groep hoort:

COICOP	Benaming	Gepubliceerde gewichten 2017 (‰)
01.1.1.1	Rijst	0,43
01.1.1.2	Meel en andere granen	0,77
01.1.1.3	Brood	9,94
01.1.1.4	Andere bakkerijproducten	10,64
01.1.1.5	Pizza en quiche	1,53
01.1.1.6	Deegwaren en couscous	2,55
01.1.1.7	Ontbijtgranen	1,29
01.1.1.8	Andere producten op basis van granen	0,61
01.1.2.1	Runds- en kalfsvlees	6,02
01.1.2.2	Varkensvlees	2,62
01.1.2.3	Schaaps- en lamsvlees	1,03
01.1.2.4	Gevogelte	4,81
01.1.2.5	Ander vlees	0,95
01.1.2.7	Bereide vleeswaren (gezouten, gedroogd of gerookt vlees)	9,57
01.1.2.8	Andere vleesproducten	14,08
01.1.3.1	Verse vis	3,54
01.1.3.2	Diepgevroren vis	0,82
01.1.3.3	Verse zeevruchten	2,01
01.1.3.4	Diepgevroren zeevruchten	0,66
01.1.3.5	Gerookte vis en zeevruchten	1,17
01.1.3.6	Vis in conserven en bereidingen op basis van vis of zeevruchten	2,37
01.1.4.1	Volle melk	0,88
01.1.4.2	Magere en halfvolle melk	1,62
01.1.4.3	Geconcentreerde melk, melk in poedervorm	0,17
01.1.4.4	Yoghurt	2,48
01.1.4.5	Kaas	9,76
01.1.4.6	Andere melkproducten	2,52
01.1.4.7	Eieren	1,11
01.1.5.1	Boter	2,09
01.1.5.2	Margarine en plantaardige vetten	1,14
01.1.5.3	Olijfolie	0,75
01.1.5.4	Andere eetbare oliën	0,59
01.1.6.1	Vers fruit	10,69
01.1.6.3	Gedroogd fruit en noten	1,13
01.1.6.4	Fruit in conserven	0,38
01.1.7.1	Verse groenten, uitgezonderd aardappelen	9,57
01.1.7.2	Diepgevroren groenten, uitgezonderd aardappelen	0,61
01.1.7.3	Gedroogde groenten, bereidingen en conserven van groenten	2,26
01.1.7.4	Aardappelen	3,23
01.1.7.5	Chips	1,17
01.1.8.1	Suiker	0,46
01.1.8.2	Confituur, marmelade en honing	1,18
01.1.8.3	Chocolade	4,72

01.1.8.4	Snoepgoed	2,10
01.1.8.5	Roomijs en sorbets	1,52
01.1.9.1	Sauzen en kruiden	2,20
01.1.9.2	Zout, specerijen en keukenkruiden	1,03
01.1.9.3	Voeding voor baby's	1,06
01.1.9.4	Bereide maaltijden	1,40
01.1.9.9	Andere voedingswaren (dieetproducten inbegrepen)	1,57
01.2.1.1	Koffie	3,26
01.2.1.2	Thee	0,63
01.2.1.3	Cacao en chocolade in poedervorm	0,14
01.2.2.1	Mineraal- of bronwater	3,47
01.2.2.2	Frisdranken	7,35
01.2.2.3	Fruitsap en groentesap	2,40
02.1.1.1	Likeuren en sterke dranken	2,08
02.1.1.2	Alcoholhoudende frisdranken	0,07
02.1.2.1	Wijn op basis van druiven	8,61
02.1.2.2	Wijn op basis van ander fruit	0,11
02.1.2.3	Likeurwijnen	0,91
02.1.3.1	Lichte bieren	2,40
02.1.3.2	Zware bieren	2,01
02.2.0.1	Sigaretten	6,92
02.2.0.3	Andere tabakswaren	2,04
05.5.2.2	Kleine toebehoren voor het huis en de tuin	3,03
05.6.1.1	Reinigings- en onderhoudsproducten	6,45
05.6.1.2	Andere kleine niet-duurzame huishoudartikelen	4,32
09.3.4.2	Artikelen voor huisdieren	6,70
09.5.4.1	Papierartikelen	1,05
09.5.4.9	Materiaal om te schrijven en te tekenen	1,58
12.1.3.1	Niet-elektrische apparaten voor lichaamsverzorging	0,93
12.1.3.2	Producten voor lichaamsverzorging & schoonheidsproducten	16,11

De onderstaande tabel bevat een fictief voorbeeld van de gewichten voor vers fruit:

	Jan	Feb	Mrt	Apr	Mei	Jun	Jul	Aug	Sep	Okt	Nov	Dec	Gem
Vers fruit	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Appelen: Jonagold	3%	3%	3%	2%	2%	2%	1%	2%	1%	2%	2%	2%	2%
Appelen: Golden	4%	4%	4%	3%	3%	2%	2%	2%	2%	3%	3%	3%	3%
Appelen: Granny	16%	16%	16%	12%	12%	8%	7%	9%	7%	12%	12%	12%	12%
Ronde peren	4%	4%	4%	3%	3%	2%	2%	2%	2%	3%	3%	3%	3%
Peren: Conference	3%	3%	3%	3%	3%	0%	0%	0%	1%	2%	2%	2%	2%
Druiven	0%	0%	0%	0%	0%	0%	0%	0%	16%	26%	26%	26%	8%
Perziken	0%	0%	0%	0%	0%	6%	5%	7%	6%	0%	0%	0%	2%
Abrikozen	0%	0%	0%	0%	0%	6%	5%	0%	0%	0%	0%	0%	1%
Pruimen	0%	0%	0%	0%	0%	0%	0%	13%	11%	0%	0%	0%	2%
Kersen	0%	0%	0%	0%	0%	13%	12%	0%	0%	0%	0%	0%	2%
Nectarines	0%	0%	0%	0%	0%	0%	14%	20%	16%	0%	0%	0%	4%
Sinaasappelen	19%	19%	19%	15%	15%	9%	8%	11%	9%	14%	14%	14%	14%
Citroenen	3%	3%	3%	2%	2%	1%	1%	2%	1%	2%	2%	2%	2%
Pompelmoezen	3%	3%	3%	2%	2%	1%	1%	1%	1%	2%	2%	2%	2%
Mandarijnen	17%	17%	17%	0%	0%	0%	0%	0%	0%	13%	13%	13%	8%
Bananen	18%	18%	18%	14%	14%	9%	7%	10%	8%	14%	14%	14%	13%
Aardbeien	0%	0%	0%	37%	37%	23%	19%	0%	0%	0%	0%	0%	10%
Meloenen	0%	0%	0%	0%	0%	14%	12%	16%	13%	0%	0%	0%	5%
Kiwi's	9%	9%	9%	7%	7%	4%	4%	5%	4%	6%	6%	6%	6%

OVER STATBEL

Statbel, het Belgische statistiekbureau, verzamelt, produceert en verspreidt betrouwbare en pertinente cijfers over de Belgische economie, de samenleving en het territorium.

Het verzamelen gebeurt op basis van administratieve databronnen en enquêtes, het produceren gebeurt op een wetenschappelijke, kwaliteitsvolle wijze en de statistieken worden tijdig en op een klantvriendelijke wijze verspreid.

De privacy, de bescherming van vertrouwelijke gegevens en het gebruik voor louter statistische doeleinden worden hierbij gegarandeerd.

Bezoek onze website
www.statbel.fgov.be

Of contacteer ons
e-mail: statbel@economie.fgov.be

Statbel (Algemene Directie Statistiek - Statistics Belgium)
North Gate - Koning Albert II-laan, 16 - 1000 Brussel
E-mail: statbel@economie.fgov.be

Ondernemingsnummer
0314.595.348

Verantwoordelijke uitgever
Nico Waeyaert

North Gate
Koning Albert II-laan 16
1000 Brussel