Centraal Bureau voor de Statistiek

Divisie Macro-economische statistieken en publicaties
Sector Ontwikkeling en ondersteuning

Postbus 24500
2490 HA Den Haag

Het gebruik van supermarkt scannerdata in de Nederlandse CPI

Heymerik van der Grient en Jan de Haan

Projectnummer: *KOO-204759*
BPA-nummer: *2010-138-KOO*
Datum: *14 juli 2010*

HET GEBRUIK VAN SUPERMARKT SCANNERDATA IN DE NEDERLANDSE CPI

Samenvatting: Sinds januari 2010 past het CBS in de CPI een nieuwe methode toe bij de berekening van prijsindexcijfers op basis van scannerdata. Tegelijkertijd is het aantal bedrijven dat scannerdata aan het CBS levert uitgebreid van één supermarkt tot zes supermarkten. Dit rapport beschrijft de nieuwe methode inclusief de daarbij gebruikte formules en de beschikbare scannerdata. Enige resultaten worden gepresenteerd en vergeleken met indexcijfers berekend via een recent voorgestelde 'benchmark' methode.

Trefwoorden: consumenten prijsindex, indexcijfer theorie, scannerdata.

1. Inleiding

De CPI-Manual (ILO, 2004; 54, 92, 478) geeft aan dat "scanner data constitute a rapidly expanding source of data with considerable potential for CPI purposes. …. Scanner data obtained from electronic points of sale include quantities sold and the corresponding value aggregates on a very detailed level. …. Scanner data are up to date and comprehensive."

Enkele landen in Europa gebruiken al scannerdata bij het samenstellen van hun CPI, zij het op verschillende manieren. Het statistische bureau van Noorwegen maakt sinds augustus 2005 gebruik van scannerdata om de sub-index voor voeding en niet-alcoholische dranken te berekenen (Rodriguez en Haraldsen, 2006). Het CBS heeft scannerdata van supermarkten in juni 2002 geïntroduceerd ten behoeve van de CPI (Schut, 2002), in eerste instantie van twee supermarktketens. Eén daarvan is na verloop van tijd weggevallen. In Noorwegen en Nederland worden zowel prijzen als gewichten (voor een grote steekproef van artikelen per productgroep) afgeleid uit de scannerdata. Het Zwitserse Federale Statistische Bureau hanteert een meer pragmatische aanpak. Scannerdata van enkele grote ketens wordt gebruikt als bron voor de prijswaarneming als vervanging van de veldwaarneming. De aan de berekening van prijsindexcijfers ten grondslag liggende principes blijven bij deze aanpak ongewijzigd (Becker-Vermeulen, 2006).

In januari 2010 heeft het CBS het gebruik van scannerdata bij de berekening van de CPI uitgebreid. Zeven extra supermarktketens zijn bereid gevonden samen te werken en op reguliere basis scannerdata te verstrekken; data van vijf daarvan is vanaf januari 2010 bij de indexberekening betrokken. De zes

ketens waarvan scannerdata momenteel wordt gebruikt, hebben een gezamenlijk marktaandeel van ongeveer 50% en hebben een belang van iets meer dan 5% in de totale CPI. Het is de bedoeling dat scannerdata van de twee andere supermarktketens gedurende 2010 of uiterlijk in 2011 wordt geïmplementeerd.

Voor het CBS zijn er twee potentiële voordelen aan het gebruik van scannerdata verbonden. Ten eerste kan de kwaliteit van de CPI en de HICP worden verbeterd omdat op een zeer gedetailleerd niveau voor alle transacties zowel prijzen als hoeveelheden beschikbaar zijn. Ten tweede kan het leiden tot een efficiëntere werkwijze. Het bezoeken van supermarkten voor de prijswaarneming is een belangrijke kostenpost bij het op de traditionele manier produceren van een CPI. Het gebruik van scannerdata heeft geleid tot een reductie van circa 15 000 prijswaarnemingen per maand voor de zes ketens samen. Ook voor de supermarkten zelf brengt het verstrekken van scannerdata voordelen met zich mee. Zij worden niet langer lastig gevallen door interviewers die in de winkel rondlopen en het personeel om hulp vragen. Het verlagen van de (administratieve) lastendruk voor bedrijven is een belangrijk issue voor het CBS.

Het breder toepassen van scannerdata is onderdeel van een algemener re-design van de Nederlandse CPI, een project dat in het begin van deze eeuw begon (De Haan, 2006). Sleutelbegrippen in dit project zijn: kwaliteit, efficiency en flexibiliteit. De implementatie van nieuwe werkwijzen begon in 2007 toen de gebruikelijke vijfjaarlijkse basisverlegging vervangen werd door het jaarlijks actualiseren van de wegingsschema's en het berekenen van een kettingindex. De gewichten, die gebruikt worden om indexcijfers van productgroepen te aggregeren, worden afgeleid uit de nationale rekeningen en niet meer uit het budgetonderzoek. In 2009 is een tweedimensionaal wegingsschema in gebruik genomen. Naast de bekende COICOP-classificatie voor goederen en diensten (de eerste dimensie) is een classificatie van verkoopkanalen toegevoegd (de tweede dimensie). In de traditionele betekenis omvat een verkoopkanaal alle bedrijven of winkels die hetzelfde assortiment voeren, zoals supermarkten, warenhuizen, slagers, kappers enz. Een verkoopkanaal kan daarnaast ook worden gedefinieerd als een enkel bedrijf of keten waarvoor indexcijfers worden berekend. Dit laatste is het geval voor scannerdata.

Dit rapport is als volgt opgezet. De tot en met 2009 bij één supermarkt toegepaste handelwijze wordt beschreven in hoofdstuk 2. Die handelwijze bleek te arbeidsintensief en te kostbaar om bij meer dan één of twee ketens toe te passen. Het CBS heeft daarom besloten een nieuwe berekenings-methodiek te implementeren die veel minder arbeidsintensief is. De nieuwe

methode wordt uitvoerig toegelicht in hoofdstuk 3. Hoofdstuk 4 beschrijft de beschikbare scannerdata en enkele activiteiten, zoals het opschonen van de ontvangen data, die worden uitgevoerd alvorens indexcijfers worden berekend. In hoofdstuk 5 worden de formules behorend bij de nieuwe methode gepresenteerd, gevolgd door enkele resultaten in hoofdstuk 6. In hoofdstuk 7 worden de resultaten vergeleken met die op basis van een veelbelovende methode die recent door een groep van academische onderzoekers is ontwikkeld.

2. Waarom een nieuwe methode?

2.1 De oude methode

In de oude methode werd aan het begin van ieder jaar een grote steekproef van artikelen samengesteld die representatief was voor het voorafgaande jaar. Ieder artikel, geïdentificeerd door middel van het *European Article Number* (EAN) kreeg een gewicht dat het relatieve belang ervan weergaf, i.e. het bestedingsaandeel binnen de supermarktketen.[1] De maandelijkse prijsindex voor het artikel werd berekend als de verhouding van de gemiddelde prijs (unit value) in de verslagmaand[2] en de unit value in het referentiejaar (het voorafgaande jaar). Vervolgens werden voor iedere 4-digit COICOP-groep elementaire prijsindexcijfers berekend als gewogen gemiddelde van de indexcijfers van de onderliggende artikelen. Gedurende het kalenderjaar werden de indexcijfers voor productgroepen dus berekend volgens de formule van Laspeyres. Dit was ook het geval voor de indexcijfers op hogere niveaus van aggregatie. Jaarlijks werden op ieder aggregatieniveau deze kortlopende indexcijfers aan elkaar gekoppeld waardoor langlopende indexreeksen (kettingindex) ontstonden.

Gedurende het jaar verdwenen artikelen uit het assortiment waardoor de steekproef kromp. De mate waarin dit plaatsvond, verschilde per productgroep en hing af van de ontwikkelingen op de markt. Om de representativiteit te handhaven en de omvang van de steekproef op peil te houden werden nieuwe artikelen als vervangers geselecteerd. Hierbij kwam de vraag aan de orde of expliciet voor mogelijke kwaliteitsverschillen moest

[1] Een deel van de bestedingen zullen een zakelijk karakter hebben of toegeschreven kunnen worden aan buitenlanders. We nemen aan dat dit slechts een fractie is van de bestedingen door binnenlandse huishoudens en daarom verwaarloosbaar.

[2] Gebaseerd op de bestedingen en verkochte hoeveelheden in de eerste twee volle weken van de maand.

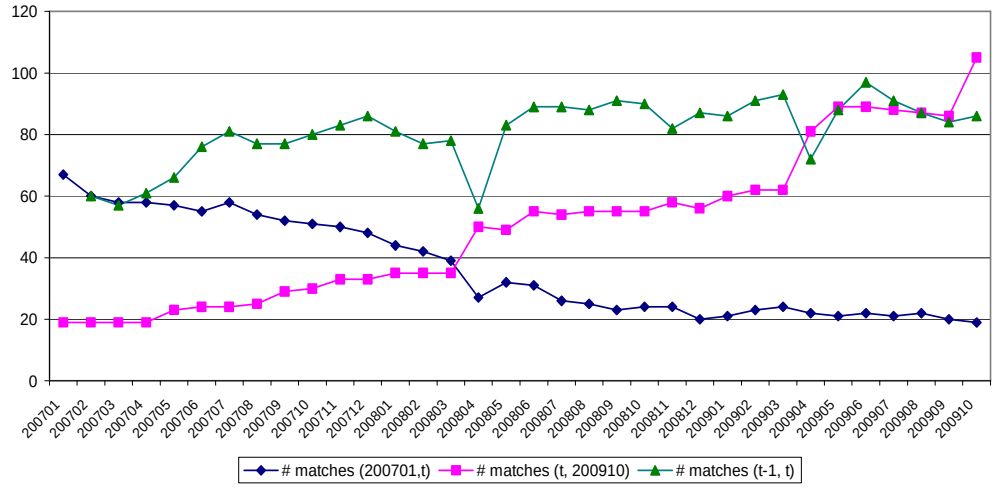
worden gecorrigeerd. In de praktijk werden meestal impliciete methodes, vooral ‘bridged overlap’ toegepast; in een beperkt aantal gevallen zijn aanpassingen in verband met hoeveelheidsveranderingen doorgevoerd.

2.2 Overwegingen bij de nieuwe berekeningsmethodiek

Een belangrijk probleem bij de oude methode was het feit dat verdwijnende artikelen werden vervangen door vergelijkbare artikelen. Dit verminderde weliswaar de noodzaak tot het uitvoeren van kwaliteitscorrecties, maar bracht met zich mee dat echt nieuwe artikelen niet in de steekproef werden opgenomen, tenminste niet eerder dan bij een jaarlijkse revisie van de steekproef. Dit is een ongewenste situatie. Kwaliteitsveranderingen lijken echter niet het belangrijkste probleem te zijn geweest. Veel belangrijker was waarschijnlijk het gebrek aan representativiteit. Uit de scannerdata blijkt dat de marktdynamiek groot is. Iedere maand verdwijnen er vele artikelen uit het assortiment van supermarkten en vele nieuwe worden erin opgenomen.

Een vast pakket artikelen, dat in feite de grondslag vormt van de oude methode, verliest dan snel zijn representativiteit. Een voorbeeld van het grote verloop wordt gegeven in figuur 1.

Figuur 1. Aantal te matchen artikelen; afwasmiddelen



Deze figuur laat op maandbasis en op drie manieren het aantal overeenkomstige artikelen bij afwasmiddelen zien. De dalende lijn geeft aan hoe snel artikelen die aan het begin van de periode (januari 2007) in het assortiment zitten daaruit verdwijnen. Aan het eind van de periode (oktober 2009) worden nog maar 19 van de 67 oorsponkelijke artikelen verkocht. De stijgende lijn geeft de omgekeerde situatie weer: deze toont het aantal overeenkomstige artikelen tussen de laatste maand en iedere voorafgaande maand. Een vergelijking met de stijgende lijn laat zien dat het totaal aantal

soorten afwasmiddelen over de jaren is toegenomen. Kennelijk zijn er meer artikelen in het assortiment opgenomen dan er uit verdwenen zijn.

De derde lijn geeft het aantal maandelijkse matches weer, dat wil zeggen het aantal artikelen waarvoor in twee opeenvolgende maanden verkopen plaatsgevonden hebben. Er is hierbij sprake van enkele opvallende veranderingen. In april 2008 bijvoorbeeld lijkt de supermarkt een deel van het assortiment afwasmiddelen vervangen te hebben.

Een ander, meer praktisch probleem was dat het jaarlijks samenstellen van de artikelsteekproef en het maandelijks bijhouden daarvan een erg arbeids-intensieve activiteit bleek te zijn vanwege het grote aantal artikelen dat hierbij betrokken was. Dit deed een groot beroep op de beschikbare capaciteit bij de sector Consumentenprijzen van het CBS. De conclusie was daarom dat het toepassen van deze methode bij meer dan één of twee supermarkten een onmogelijke zaak zou zijn gezien de beschikbare tijd en middelen. Om scannerdata van meer supermarkten op een efficiënte manier te verwerken zou een berekeningsmethodiek noodzakelijk zijn met minder handmatige interventies. Het gebruik van een maandelijkse matched-item index op het elementaire niveau is dan een voor de hand liggende keus. Bij een dergelijke aanpak is het niet meer nodig jaarlijks een verzameling artikelen vast te stellen en tegelijkertijd wordt de dynamiek in het assortiment actief gevolgd.

Omdat maandelijks de bestedingen op artikelniveau beschikbaar zijn, ligt het gebruik van een superlatieve indexformule zoals die van Fisher of Törnqvist erg voor de hand bij het berekenen van indexcijfers op basis van scannerdata. Het is echter bekend dat superlatieve kettingindexcijfers zeker op maandbasis het risico op vertekening met zich meebrengen. Dit verschijnsel staat bekend als *chain link bias* of *chain drift* (ILO, 2004; 283)

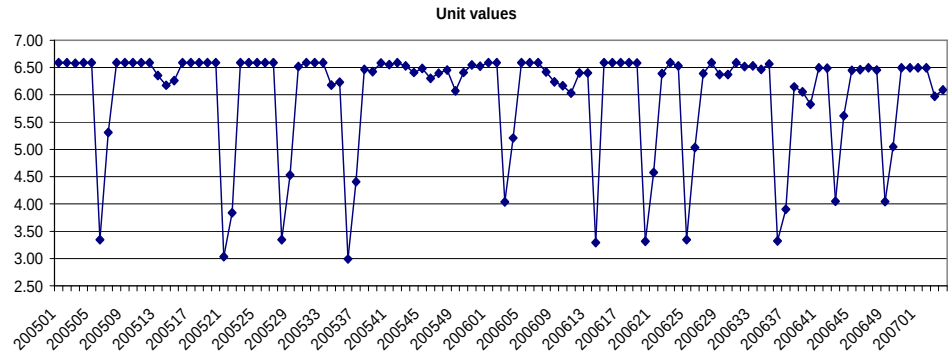
Deze vertekening wordt veroorzaakt door zeer sterke schommelingen in prijzen en verkochte hoeveelheden als gevolg van aanbiedingen; huishoudens slaan grootschalig in bij aanbiedingen en verbruiken de aangelegde voorraad gedurende de tijd dat het betrokken goed tegen de reguliere prijs in de schappen ligt.[3]

De hoeveelheden die tegen een aanbiedingsprijs worden gekocht zijn soms het honderdvoudige van die in tijden van reguliere prijzen. Een kenmerkend voorbeeld van dit verschijnsel wordt gegeven in figuren 2 en 3 waar de prijs

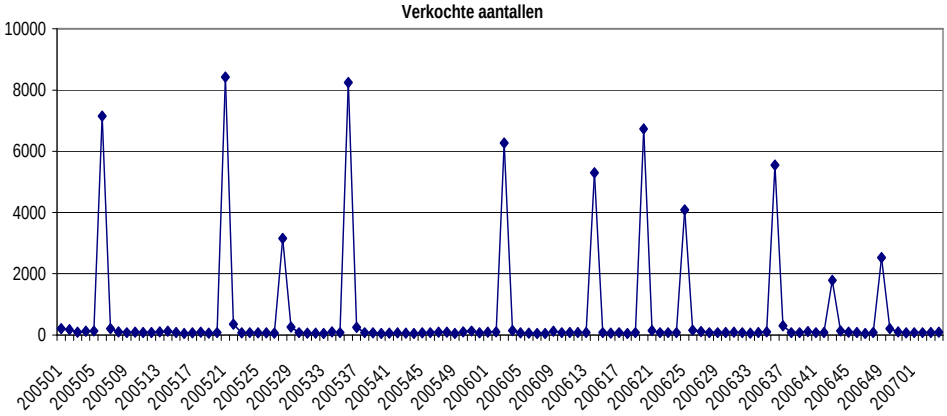
[3] Het onderliggende probleem is de asymmetrie van de gewichten van de in aanbieding zijnde artikelen. Stel dat enkele artikelen in maand t in de aanbieding zijn. ls het gewicht van die artikelen in maand t+1 gelijk aan het gewicht in maand t-1, dan zou bij een maandelijkse ketting superlatieve index geen sprake van vertekening zijn. Zie De Haan en Van der Grient (2009) voor een gedetailleerde beschrijving.

en de verkochte hoeveelheid van een specifiek vaatwasmiddel bij één supermarktketen wordt weergegeven. Het meest valt op dat verkopen tegen de reguliere prijs van circa 6.50 euro verwaarloosbaar zijn.

Figuur 2. Gemiddelde weekprijzen; vaatwastabletten XYZ



Figuur 3. Wekelijks verkochte aantallen; vaatwastabletten XYZ



Het risico op vertekening bij gewogen maandkettingindexcijfers heeft geleid tot de keus voor het ongewogen middelen van prijsmutaties op het elementaire niveau ofwel de formule van Jevons. In hoofdstukken 3 en 4 wordt uitgelegd dat enkele verfijningen hierop noodzakelijk zijn om tot acceptabele resultaten te komen.

3. Een overzicht van de nieuwe methode

De bestedingen aan artikelen binnen een productgroep is meestal erg scheef verdeeld; vaak genereert de helft van de artikelen minder dan 15% van de totale bestedingen. Anders gezegd, een relatief klein aantal artikelen is verantwoordelijk voor het grootste deel van de bestedingen.

Door het gebruik van de (ongewogen) Jevons index zouden we daar geen rekening mee houden. We hebben daarom besloten een soort ruwe impliciete

weging toe te passen in de vorm van een cut-off steekproef: belangrijke artikelen binnen een elementair aggregaat worden met zekerheid in de steekproef opgenomen en onbelangrijke artikelen worden buiten beschouwing gelaten. Meer specifiek, een artikel wordt betrokken bij de indexberekening voor twee opeenvolgende maanden indien het gemiddelde bestedingsaandeel (met betrekking tot de artikelen die in beide maanden een prijs hebben) in die maanden boven een bepaalde drempelwaarde uitkomt. De drempel is zodanig gekozen dat ruwweg 50% van de artikelen in een elementair aggregaat bij de berekening wordt betrokken. Deze artikelen vertegenwoordigen gemiddeld 80-85% van de totale bestedingen.

Een nadeel van een strikte matched-item benadering is dat artikelen waarvoor tijdelijk geen prijzen beschikbaar zijn buiten de berekening blijven. Dit heeft tot gevolg dat een prijsmutatie tussen de laatste maand waarin het artikel in de steekproef was opgenomen en de maand waarin het artikel weer terugkeert zou worden genegeerd. Om dit te voorkomen worden ontbrekende prijzen op de gebruikelijke manier geïmputeerd door de laatst waargenomen prijs te vermenigvuldigen met de (Jevons) prijsindex van het aggregaat waartoe het betrokken artikel behoort. In zekere zin wordt daarmee een panelelement opgelegd aan de dynamische matched-item aanpak. De prijsmutatie die optreedt na een periode van ontbrekende prijzen wordt nu in de index meegenomen.

Zoals iedere andere matched-item methode houdt de nieuwe methode geen rekening met kwaliteitsveranderingen. Daar in de Nederlandse CPI impliciete methodes om voor kwaliteitsveranderingen te corrigeren de overhand hebben, is de nieuwe methode op dit aspect vergelijkbaar met de oude. Het nieuwe computersysteem biedt echter de mogelijkheid om expliciete correcties uit te voeren voor het geval dat noodzakelijk mocht zijn. Hierbij valt te denken aan hoeveelheidsveranderingen. De verwachting is echter dat van deze mogelijkheid slechts spaarzaam gebruik zal (hoeven te) worden gemaakt.

De procedure om indexcijfers voor hogere aggregaten te berekenen wordt niet veranderd. Deze blijven berekend worden als jaarlijkse Laspeyres kettingindexcijfers, waarbij voor de kortlopende indices het voorafgaande jaar als index- en wegingsreferentiejaar fungeert. Om zo optimaal mogelijk gebruik te kunnen maken van de beschikbare informatie over bestedingen, worden de elementaire aggregaten zo gedetailleerd mogelijk gedefinieerd.

Om het handmatig indelen van EAN’s te voorkomen, wordt een ondergrens aan deze detaillering opgelegd door de productgroepindeling die door de supermarkt wordt meegeleverd. Vaak zijn deze laatste productgroepen samengevoegd om ervoor te zorgen dat elementaire aggregaten structureel voldoende EAN’s bevatten waardoor deze als robuust kunnen worden

beoordeeld. De aldus ontstane elementaire aggregaten zijn meestal nog steeds supermarkt-specifiek en het niveau ervan is vergelijkbaar met 6-digit COICOP.

Voor iedere supermarktketen is bestedingsinformatie beschikbaar op ieder COICOP-niveau. Deze informatie die vóór het gebruik van scannerdata niet beschikbaar was, wordt gebruikt om over de verschillende ketens heen te aggregeren waardoor de kwaliteit van de productgroepindexcijfers verhoogd wordt: het relatieve belang van productgroepen blijkt namelijk significant te verschillen tussen de diverse supermarktketens[4].

4. Scannerdata en het opschonen daarvan

4.1 Beschikbare scannerdata

Wekelijks stuurt iedere supermarktketen een bestand naar het CBS dat de gescande gegevens bevat over de bestedingen en verkochte hoeveelheden van alle individuele artikelen die worden geïdentificeerd met het *European Article Number* (EAN).[5]

Deze gegevens worden verstrekt voor alle individuele verkooppunten (filialen) van de keten, voor een representatieve selectie daarvan of geaggregeerd over alle verkooppunten. Het CBS neemt aan dat beide laatste varianten geen groot probleem vormen. Het is redelijk om te veronderstellen dat verkooppunten die tot dezelfde keten behoren dezelfde dienstverlening kennen, waardoor het per artikel aggregeren over verkooppunten acceptabel is. De meeste supermarktketens in Nederland kennen een nationale prijspolitiek: voor veruit de meeste artikelen zijn de prijzen in ieder verkooppunt gelijk. Zelfs bij een steekproef van verkooppunten, is het onwaarschijnlijk dat de unit values veel zullen afwijken van de 'echte' waardes[6].

---

[4] Hoewel het een voordeel is dat gegevens over bestedingen kunnen worden gebruikt als gewichten, kan het tegelijkertijd leiden tot inconsistenties met gegevens uit andere bronnen (Nationale rekeningen, budgetonderzoeken bij huishoudens, detailhandelsstatistieken, enz.). Het zal nodig zijn aanpassingen door te voeren om tot een consistent overall wegingsschema te komen.

[5] Wekelijks wordt ongeveer 570 Mb aan data ontvangen van de zes supermarktketens die momenteel bij de indexberekening zijn betrokken. Het aantal ontvangen records loopt op tot 300 mln per jaar. Met nadruk wordt vermeld dat het CBS niet voor deze scannerdata betaalt.

[6] Op de korte termijn zijn wel enkele malen kleine verschillen in maandelijkse indexmutaties gesignaleerd. Op de lange termijn echter leidt het gebruik van een steekproef van verkooppunten niet tot een andere indexontwikkeling.

Ieder record in het databestand heeft betrekking op een individuele EAN en bevat de wekelijkse bestedingen, het aantal verkochte eenheden en een (meestal korte) productomschrijving, die vaak het gewicht, de inhoud of de verpakkingsvorm van het artikel aanduidt. Speciaal voor de afgeleide CPI en HICP (‘constant tax HICP’) is het noodzakelijk de hoeveelheid of het percentage alcohol in flessen bier, wijn enz. te kennen. Deze gegevens staan niet altijd in de scannerdata en moeten daarom soms worden geschat.

Iedere supermarktketen voegt eigen classificatiecodes toe die aangeven tot welke productgroep een EAN behoort. Vanwege het grote aantal nieuwe EAN’s dat maandelijks op de markt verschijnt, is het voor een efficiënt productieproces noodzakelijk dergelijke classificatiecodes beschikbaar te hebben. Wanneer de relatie tussen de supermarkt-specifieke classificatie en de COICOP-indeling is vastgelegd, kunnen EAN’s automatisch worden toegekend aan de 4-digit COICOP-groep waartoe zij behoren. Om te voorkomen dat deze relatie steeds moet worden aangepast is het een vereiste dat de supermarkt-specifieke classificatie stabiel in de tijd is. En vanzelfsprekend dient deze gedetailleerder te zijn dan het laagste COICOP-niveau dat op het CBS wordt gebruikt (in de praktijk 4-digit).

Het komt voor dat supermarkten allerlei artikelen groeperen die verbonden zijn met speciale gelegenheden, zoals Kerstmis, Pasen, verjaardagen e.d. Ook worden artikelen gegroepeerd die een bepaald doel dienen. Een voorbeeld is ‘artikelen voor een barbecue’ waarin vlees, houtskool en sauzen zijn opgenomen. In dit soort gevallen kunnen EAN’s niet meer automatisch aan COICOP-groepen worden toegekend. Besloten is deze EAN’s buiten de berekening te laten.

Records die met dezelfde EAN’s worden geïdentificeerd, worden geacht betrekking te hebben op (fysiek) identieke artikelen. Dit is inderdaad het geval voor de overgrote meerderheid van producten omdat die door de producent van een EAN zijn voorzien. Aan sommige producten, zoals vers fruit, mogen supermarkten zelf een EAN toekennen. Ook komt het voor dat voor lastig te scannen producten (kratten bier) verkorte codes worden gebruikt om tikfouten bij de kassa te voorkomen. Een speciale verzameling EAN’s is voor dat doel beschikbaar. Dit is allemaal geen probleem zolang de supermarkt dezelfde code voor hetzelfde artikel blijft gebruiken. Helaas is dit niet altijd het geval. Dit heeft tot gevolg dat prijzen van verschillende artikelen (met dezelfde EAN) worden vergeleken. Het aantal gevallen is echter zodanig beperkt gebleken dat het effect verwaarloosd kan worden. Bovendien vinden maandelijks routinecontroles plaats waarbij situaties

waarin een fout gemeten prijsmutatie wel veel invloed zou hebben, zullen worden gesignaleerd.[7]

Artikelen met verschillende EAN's worden behandeld als verschillende, dus niet vergelijkbare, producten. Voor CPI-doeleinden kan het EAN-niveau soms te gedetailleerd zijn. Artikelen die vanuit het oogpunt van de consument identiek zijn, zouden in principe als hetzelfde artikel moeten worden gezien, ook al is de EAN verschillend. Als een artikel verdwijnt en een volledig vergelijkbaar artikel, maar met een andere EAN, op de markt verschijnt, zouden de prijzen van beide artikelen direct moeten worden vergeleken. Voorbeeld is een pak koffie dat gewoonlijk in rood papier is verpakt maar op een bepaald moment vanwege promotie activiteiten in een blauwe verpakking gepresenteerd wordt. Bij de nieuwe methode worden de prijzen van deze twee varianten echter niet direct met elkaar vergeleken. Zoals eerder aangegeven, biedt het nieuwe computersysteem echter wel de mogelijkheid een expliciete aanpassing door te voeren. Als de bedoelde situatie dus voorkomt heeft de gebruiker de mogelijkheid alsnog een directe vergelijking te maken[8].

In de afgelopen decennia hebben supermarkten in Nederland hun assortiment aanmerkelijk uitgebreid. Vaak bevat het assortiment tegenwoordig ook zaken als kleding, glas, aardewerk en huishoudelijke gebruiksvoorwerpen. Het CBS heeft besloten de indexberekening op basis van scannerdata vooralsnog te beperken tot de meer traditionele productgroepen. In tabel 1 staan de categorieën waarvoor scannerdata indexcijfers worden berekend.

Tabel 1. COICOP groepen waarvoor scannerdata indexcijfers worden berekend

COICOP-code	Omschrijving
010000	Voeding en niet-alcoholische dranken
021200	Wijn
021300	Bier
055000	Gereedschappen en werktuigen voor huis en tuin
056000	Goederen en diensten voor het dagelijks onderhoud van de woning
061000	Niet verzekerbare medische en farmaceutische producten
093400	Huisdieren en producten voor huisdieren
121300	Toiletartikelen, schoonheidsartikelen en andere artikelen voor de lichaamsverzorging

[7] Als het bekend is welke EAN's gebruikt worden door individuele supermarkten, kan ook worden besloten al deze EAN's uit te sluiten van de indexberekening.

[8] Het nieuwe computersysteem voorziet de gebruiker van indicatoren die wijzen op grootschalige aanpassingen in het assortiment van de supermarkt.

4.2 Het voorbereiden en opschonen van de data

De prijsindexcijfers voor maand *t* worden door het CBS gepubliceerd in de eerste week van maand *t*+1. Deze snelle publicatie brengt met zich mee dat gegevens over de laatste week van maand *t* niet kunnen worden gebruikt. De prijswaarneming is daarom beperkt tot de eerste drie volle weken van iedere maand. Deze aanpak geldt ook voor de scannerdata. Dat houdt in dat de gemiddelde prijs ('unit value') van een artikel gebaseerd is op de scannerdata van de eerste drie volle weken per kalendermaand en van alle vestigingen van de supermarktketen waarvan scannerdata wordt ontvangen.

Deze gemiddelde prijzen worden aan twee automatische controles onderworpen. Ten eerste wordt gekeken naar de verandering ten opzichte van de gemiddelde prijs in de voorafgaande maand. Is deze mutatie groter dan een factor 4 dan wordt de mutatie onwaarschijnlijk geacht en buiten de verdere berekening gelaten. Het gaat dus om artikelen waar de huidige prijs 300% hoger of 75% lager is dan de prijs in de vorige maand.

Ten tweede is een algoritme ontwikkeld, dumpfilter genaamd, dat artikelen buiten de berekening laat in situaties dat een sterke prijsdaling optreedt in combinatie met een grote omzetdaling. 'Dumpen' komt soms voor als artikelen uit het assortiment worden gehaald en de laatste restanten tegen bijzonder lage prijzen worden verkocht. Tegenover een dergelijke prijsdaling staat geen terugkeer naar een reguliere prijs zoals bij aanbiedingen. Dergelijke dumpprijzen kunnen dan een onaanvaardbare neerwaartse vertekening in de prijsindex van de betrokken productgroep veroorzaken, zoals uit analyses is gebleken. In de praktijk is het mogelijk dat door dit dumpfilter meer prijzen buiten beschouwing worden gelaten dan alleen die van de echte dumpgevallen. Dit heeft geen ernstige consequenties omdat ontbrekende prijzen worden geïmputeerd.

5. Formules voor de indexberekening

Nadat de data zijn opgeschoond, worden op diverse aggregatieniveaus prijsindexcijfers berekend. In dit hoofdstuk worden de daarbij gebruikte formules besproken. We beginnen op het elementaire niveau waarop ongewogen meetkundig gemiddelde prijsindexcijfers worden berekend.

We gebruiken de volgende notatie. De prijs en het bestedingsaandeel van artikel *i* in maand *m* van jaar *y* worden aangegeven met respectievelijk $p_i^{y,m}$ en $s_i^{y,m}$. Laat *a* een bepaald elementair aggregaat zijn en $N_a^{(y,m-1),(y,m)}$ het daarbij horende aantal te matchen artikelen tussen maanden *m* en *m*-1 van jaar *y*. Om een ruwe vorm van wegen te introduceren krijgt ieder artikel *i* een

kans $w_i^{y,m}$ om in de steekproef te worden opgenomen voor de berekening van de prijsverandering tussen maand *m*-1 en maand *m*. Deze insluitkansen of impliciete gewichten zijn:

$$w_i^{y,m} = 1 \qquad \text{als } \frac{s_i^{y,m-1} + s_i^{y,m}}{2} > \frac{1}{N_a^{(y,m-1),(y,m)}\chi};$$

$$w_i^{y,m} = 0 \qquad \text{anders.}$$

Dit houdt in dat een artikel in de steekproef wordt opgenomen indien het gemiddelde bestedingsaandeel in de maanden *m*-1 en *m* de drempel $1/N_a^{(y,m-1),(y,m)}\chi$ overschrijdt. De uiteindelijke steekproefomvang is dus gelijk aan de som van alle impliciete gewichten: $\sum_{i=1}^{N(y,m-1),(y,m)} w_i^{y,m} = n_a^{(y,m-1),(y,m)}$. Aan parameter χ kan iedere willekeurige positieve waarde toegekend worden, hoewel er in de praktijk wel een ondergrens is; als de waarde te laag wordt gekozen, zal de steekproef leeg zijn. Op basis van diverse simulaties is gekozen voor $\chi = 1{,}25$ bij alle productgroepen; het was niet noodzakelijk om tussen productgroepen te differentiëren. Om een voorbeeld te geven, als $N_a^{(y,m-1),(y,m)} = 80$, worden alle artikelen geselecteerd met een gemiddeld bestedingsaandeel groter dan 1%.

De prijsverandering tussen *y,m*-1[9] en *y,m* voor elementair aggregaat *a* wordt dan berekend als

$$\pi_a^{y,m/y,m-1} = \prod_{i=1}^{n_a^{(y,m-1),(y,m)}} \left(\frac{p_i^{y,m}}{p_i^{y,m-1}} \right)^{1/n_a^{(y,m-1),(y,m)}} . \qquad (1)$$

Vergelijking (1) is een maand-op-maand Jevons prijsindex op steekproef-basis. Deze maand-op-maand mutaties worden vervolgens met elkaar vermenigvuldigd (‘gekettingd’) wat leidt tot langlopende tijdreeksen met een bepaalde referentie- of startmaand y_0, m_0:

$$P_a^{y,m/y_0,m_0} = P_a^{y,m-1/y_0,m_0} * \pi_a^{y,m/y,m-1}, \qquad (2)$$

waarin $P_a^{y,m-1/y_0 m_0}$ de ketting matched-items prijsindex aangeeft tussen de startmaand en maand *m*-1 van jaar *y*.

Voor artikelen die niet in maand *y,m* verkocht zijn, maar wel in voorafgaande periodes, wordt een prijs geïmputeerd[10]:

$$\hat{p}_i^{y,m} = p_i^{y,m-1} * \pi_a^{y,m/y,m-1}. \qquad (3)$$

---

[9] *y,m*-1 is gelijk aan *y*-1,12 in geval *m*=1.

[10] De voorafgaande prijs (in maand *y,m*-1) kan ook een geïmputeerde prijs zijn.

Voor hogere aggregaten *A* worden kortlopende prijsindexcijfers berekend volgens de formule van Laspeyres met index referentie periode *y*-1:

$$P_A^{y,m/y-1} = \frac{\sum_{a \in A} w_a^{y-1} * P_a^{y,m/y-1}}{\sum_{a \in A} w_a^{y-1}}. \qquad (4)$$

De gewichten w_a^{y-1} in (4) zijn gebaseerd op de jaarlijkse bestedingen van alle artikelen die tot elementair aggregaat *a* behoren, ongeacht of artikelen in de steekproef zijn opgenomen of niet.[11] De kortlopende tijdreeksen worden vervolgens in december (de koppelmaand) gekettingd tot langlopende tijdreeksen met index referentie periode 0.[12]

$$P_{ch,A}^{y,m/0} = \left(\frac{P_A^{y,m/y-1}}{P_A^{y-1,12/y-1}} \right) * \left[\prod_{\tau=1}^{y-1} \frac{P_A^{\tau,12/\tau-1}}{P_A^{\tau-1,12/\tau-1}} \right] * P_A^{0,12/0}. \qquad (5)$$

Kortlopende indices $P_A^{y,m/y-1}$ en kettingindexcijfers $P_{ch,A}^{y,m/0}$ worden berekend voor alle COICOP-categorieën. Dit wordt gedaan per supermarktketen die scannerdata levert, steeds gebruik makend van formules (4) en (5).

Via dezelfde procedures worden uiteindelijk de supermarktindexcijfers op ieder COICOP-niveau samengenomen met de indexcijfers die zijn berekend voor andere relevante branches zoals bakkers en slagers. Deze laatste indexcijfers worden nog berekend op basis van in het veld waargenomen prijzen.

6. Enige resultaten

Er bestaat een groot verschil tussen de traditionele manier waarop indexcijfers op basis van veldwaarneming worden berekend en de nieuwe manier die op scannerdata is gebaseerd. De op traditionele wijze in de winkel waargenomen prijzen zijn gewoonlijk schapprijzen, terwijl uit de scannerdata gemiddelde transactieprijzen (unit values) worden afgeleid. De omvang van de steekproef bij de veldwaarneming is erg klein vergeleken met scannerdata. De traditionele steekproeven zijn meestal panels met vaste omvang, terwijl bij scannerdata een dynamische matched-item aanpak wordt gevolgd. En tot slot, er wordt een andere indexformule gebruikt. Traditioneel wordt op het elementaire niveau de verhouding van ongewogen gemiddelde prijzen

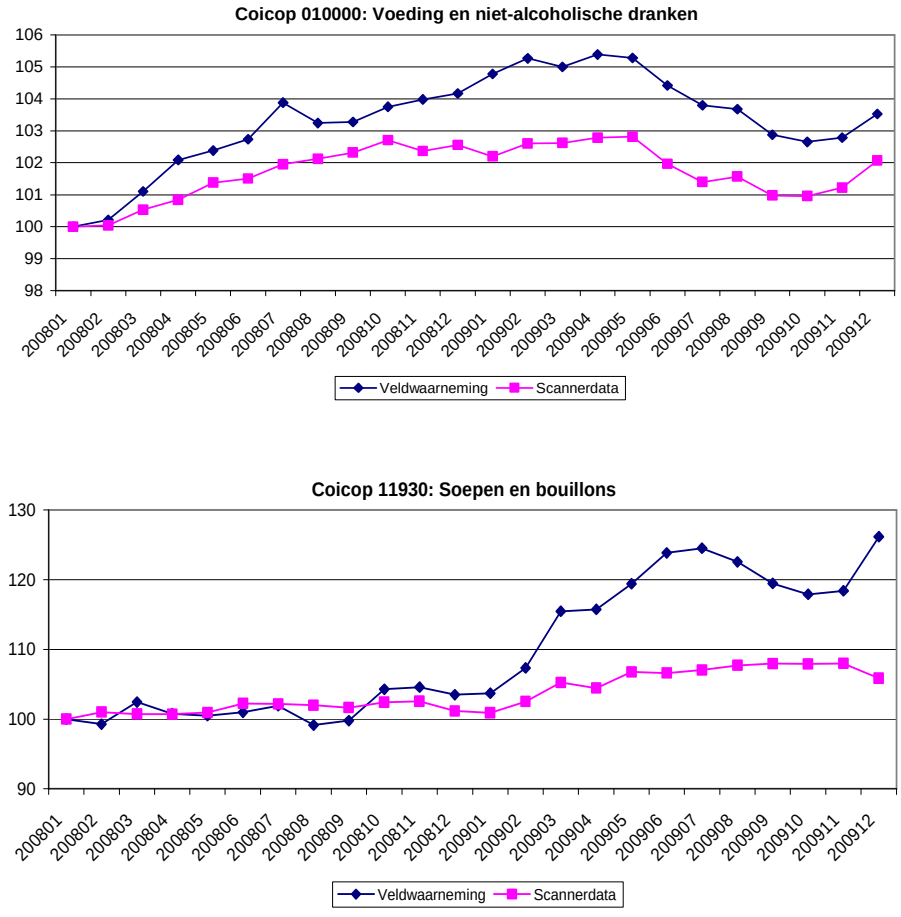
[11] In de berekening van indexcijfers voor seizoenproducten, zoals vers fruit en verse groente, worden maandelijks wisselende gewichten toegepast.

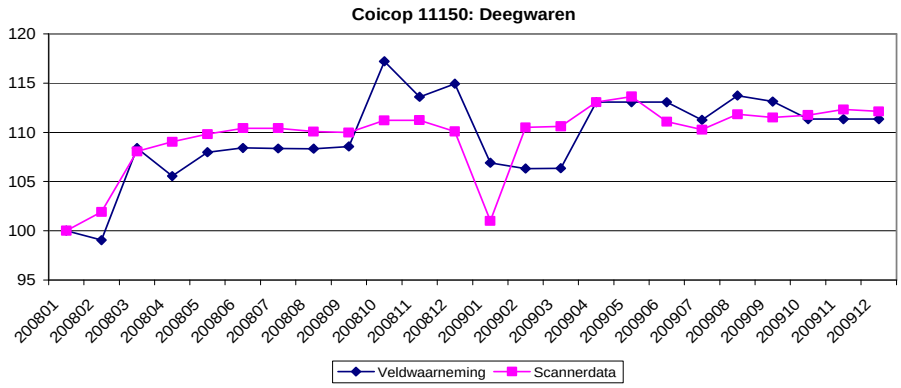
[12] Momenteel heeft de CPI 2006 als index referentie periode.

gebruikt (de formule van Dutot), terwijl bij scannerdata gebruik gemaakt wordt van de ongewogen meetkundige Jevons index.

Vanwege al deze verschillen is het niet verrassend dat ook de resulterende indexcijfers van elkaar afwijken. In figuur 4 worden de indexcijfers op basis van ieder van beide methodes vergeleken voor voeding en niet-alcoholische dranken (COICOP-groep 010000) en voor twee specifieke productgroepen. Deze indexcijfers hebben betrekking op het totaal van de vijf nieuwe supermarkten waarvan scannerdata vanaf januari 2010 voor de CPI-berekening wordt gebruikt. De scannerdata index voor voeding en dranken is duidelijk lager dan de index die gebaseerd is op de in de winkels waargenomen prijzen, hoewel het verschil in de tweede helft van 2009 wel kleiner wordt. Voor soepen en bouillons is sprake van een systematisch verschil in 2009. Bij deegwaren daarentegen vertonen de indexcijfers dezelfde trend, al is op korte termijn wel sprake van enkele opvallende verschillen.

Figuur 4. Priisindexcijfers; veldwaarneming versus scannerdata





7. Vergelijking met een benchmark index

Tegen de nieuwe aanpak bij het verwerken van supermarkt scannerdata door het CBS kan worden ingebracht dat op het elementaire niveau geen rekening wordt gehouden met relatieve belangen, niettegenstaande het feit dat bestedingsinformatie beschikbaar is op het niveau van individuele artikelen. Zoals eerder aangegeven, is de belangrijkste reden het vermijden van structurele vertekeningen die het gevolg kunnen zijn van het gebruik van maandelijkse ketting Fisher of Törnqvist prijsindexcijfers.[13]

Onlangs hebben enkele academische onderzoekers een nieuwe, zeer veelbelovende aanpak voorgesteld om superlatieve ketting indexcijfers te construeren op basis van scannerdata waarbij optimaal gebruik wordt gemaakt van alle in de data beschikbare matches en waarbij toch geen vertekening optreedt (Ivancic, Diewert en Fox, 2009). Zij stellen voor de GEKS-procedure, bekend van de ruimtelijke prijsvergelijkingen (koopkracht-pariteiten) toe te passen bij de prijsvergelijking in de tijd. Vanwege de constructie leidt de GEKS-methode tot transitieve indexcijfers wat inhoudt dat de kettingindex gelijk is aan de directe index. Daardoor wordt een prijsverandering gemeten die vrij is van vertekening. De auteurs stellen een voortschrijdende ketting variant voor (*RYGEKS*: *rolling-year GEKS*) om het probleem te vermijden dat eerder gepubliceerde indexcijfers herzien zouden moeten worden.

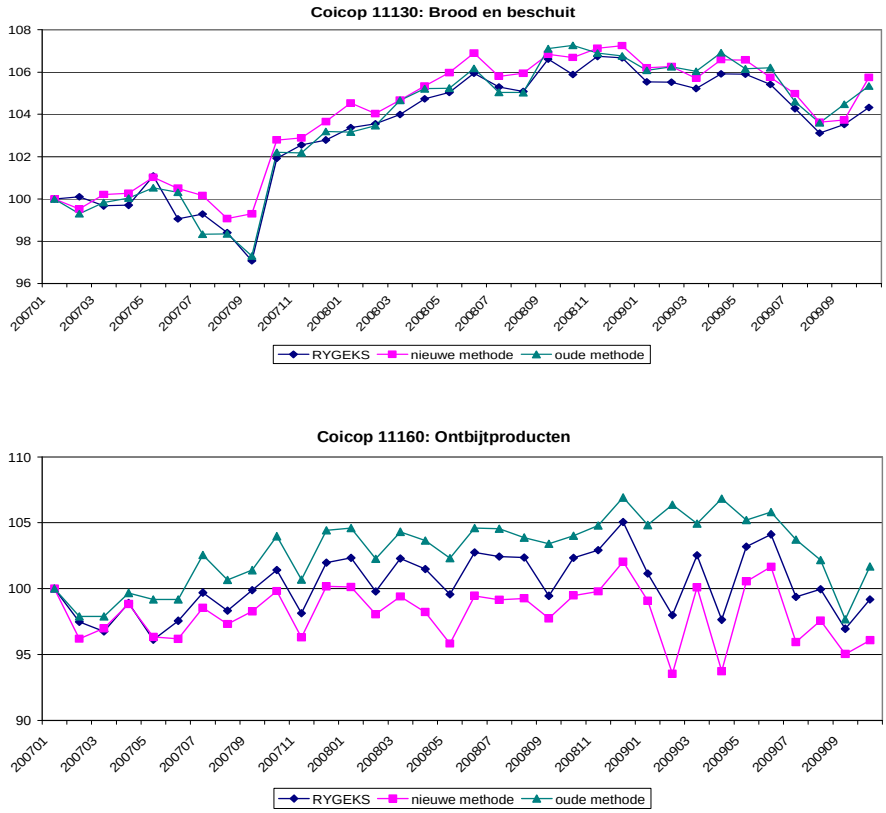
Wij hebben deze aanpak toegepast op een grote Nederlandse dataset (De Haan en Van der Grient, 2009), waarbij één van de oogmerken was de

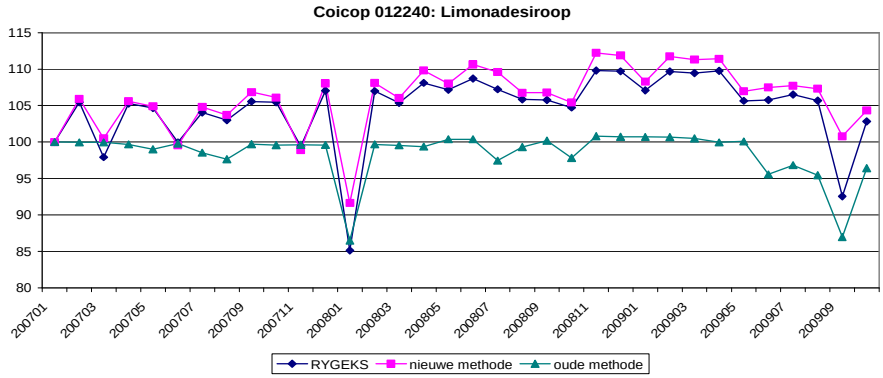
[13] Schut (2001) geeft aan dat het CBS oorspronkelijk van plan was een maandelijkse ketting Fisher index voor scannerdata te introduceren op het elementaire niveau. Zij liet zien dat bij enkele indexcijfers sprake was van enige vertekening wat de reden was dat het CBS indertijd koos voor een jaarlijkse ketting Laspeyres index.

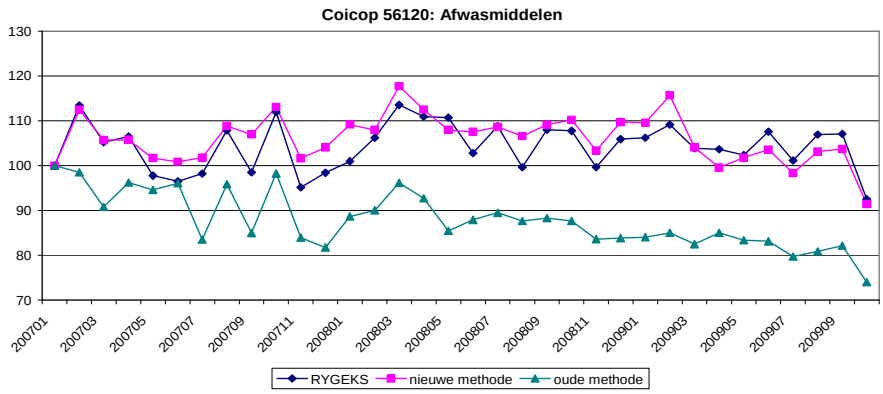
nieuwe CBS-methodiek te valideren. Onze conclusie was dat de indexcijfers berekend via de nieuwe CBS-methodiek over het algemeen goed overeen komen met de RYGEKS-indexcijfers. Af en toe was sprake van verschillen in ontwikkeling maar deze waren steeds tijdelijk. Essentieel hierbij waren de aanpassingen die het CBS heeft toegepast op een strikte matched-items Jevons aanpak: de juiste waarde voor de parameter (χ) die de steekproef-omvang bepaalt, het imputeren van tijdelijk ontbrekende prijzen en het toepassen van een dumpfilter.

Figuur 5 vergelijkt voor vier productgroepen de prijsindexcijfers berekend volgens de oude en de nieuwe CBS-methodiek met RYGEKS-indexcijfers. Al deze indexcijfers hebben betrekking op de supermarkt waarvan ook vóór januari 2010 scannerdata gebruikt werd bij de CPI-berekening. Wanneer we de RYGEKS-index als referentieindex ('benchmark') zien dan presteren de oude en de nieuwe CBS-methodiek even goed voor productgroep brood en beschuit. Voor ontbijtproducten wijken de indices volgens oude en nieuwe CBS-methodiek beide, en in verschillende richting, af van de referentie-index. Voor limonadesiroop en afwasmiddelen leidt de nieuwe CBS-aanpak tot duidelijk betere indexcijfers dan de oude.

Figuur 5. Prijsindexcijfers; oude en nieuwe CBS-methode en RYGEKS







Er zijn tal van argumenten waarom de RYGEKS-aanpak aanbevelenswaardig is, al is de methode niet zo eenvoudig uit te leggen aan gebruikers en makers van prijsindexcijfers. Toch heeft het CBS deze methode (nog) niet toegepast. Een van de redenen is het beleid van het CBS dat bepaalt dat alleen methodieken worden toegepast die breed in de internationale wereld van prijsstatistici worden gedragen. Wanneer dit het geval is – bij diverse statistische bureaus vindt momenteel research plaats – kan het CBS op termijn op de RYGEKS-methodiek overstappen.

Referenties

Becker-Vermeulen, C. (2006), Recent developments in the Swiss CPI: scanner data, telecommunications and health price collection, paper presented at the ninth meeting of the Ottawa Group, 14-16 May 2006, London.

Haan, J. de (2006), The re-design of the Dutch CPI, *Statistical Journal of the the United Nations Economic Commission for Europe* 28, 101-118.

Haan, J. de and H.A. van der Grient (2009), Eliminating chain drift in price indexes based on scanner data, paper presented at the eleventh meeting of the Ottawa Group, 27-29 May 2009, Neuchâtel.

ILO, IMF, OECD, Eurostat, United Nations, World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, Geneva: ILO Publications.

Ivancic, L., E.W. Diewert and K.J. Fox (2009), Scanner data, time aggregation and the construction of price indexes, paper presented at the eleventh meeting of the Ottawa Group, 27-29 May 2009, Neuchâtel.

Rodriguez, J. and F. Haraldsen (2006), The use of scanner data in the Norwegian CPI: The “new” index for food and non-alcoholic beverages, *Economic Survey* 4, 21-28.

Schut, C. (2001), Using Scanner Data to Compile Price Indices: Experiences and Practical Problems, Paper presented at the Joint ECE/ILO Meeting on Consumer Price Indices, 1-2 November 2001, Geneva.

Schut, C. (ed.) (2002), Gebruik van scannerdata van supermarkten in de consumentenprijsindex, Statistics Netherlands, Voorburg. Beschikbaar op www.cbs.nl.