Centraal Bureau voor de Statistiek
Postbus 4000
2270 JM Voorburg

Gebruik van scannerdata van supermarkten in de consumentenprijsindex

Redactie:
Cecile Schut

Bijdragen:
Bert Balk
Jan Eefting
Tjalling Gelsema
Jan de Haan
Wim Kiewiet
Gwen Krul
Corien Ooms
Kees van Popele
Peter van Poppel
Cecile Schut
Bert van Zanten
Lianne de Zeeuw

Datum: *4 juli 2002*

Gebruik van scannerdata van supermarkten in de consumentenprijsindex

1. Introductie en leeswijzer

Bij de samenstelling van de consumentenprijsindex (CPI) voor juni 2002 maakt het CBS voor het eerst grootschalig gebruik van kassascangegevens van supermarkten, verder scannerdata genoemd. De consumentenprijsindex, waarvan het inflatiecijfer direct wordt afgeleid, behoort sinds jaar en dag tot de belangrijkste economische indicatoren die het CBS publiceert. Het proces van waarneming, verwerking en controle van gegevens voor de samenstelling van de CPI heeft in de loop van de tijd talrijke verbeteringen ondergaan, waarbij steeds gebruik is gemaakt van alle moderne hulpmiddelen. Door het op grote schaal inzetten van scannerdata van supermarkten vindt nu een substantiële innovatieslag plaats, die tevens tot een inhoudelijke verbetering leidt.

Met de introductie van scannerdata wordt een aantal beleidsdoelstellingen bereikt. Een belangrijke centrale doelstelling is de beperking van de administratieve lasten- of enquêtedruk. Tot medio 2002 werden maandelijks in supermarkten door heel Nederland prijzen van een groot aantal producten verzameld door interviewers. Die activiteit wordt door het inzetten van scannerdata in belangrijke mate gereduceerd. Het gebruik van scannerdata levert niet alleen efficiencywinst op, het leidt ook tot een kwaliteitsverbetering van de CPI, omdat het aantal waargenomen prijzen een veelvoud is van voorheen en er meer dan tot dusverre rekening wordt gehouden met allerlei klantenkortingen en speciale acties. Het gebruik van scannerdata in de CPI kan leiden tot een grotere volatiliteit van prijsindexcijfers op het laagste niveau van aggregatie. Het is gebleken dat de doorwerking hiervan op de totale CPI binnen de nauwkeurigheidsmarge valt.

Aan de ingebruikneming van scannerdata ging een langdurig, zorgvuldig en intensief onderzoekstraject vooraf. Het onderzoek is door het CBS geplaatst in het kader van een serie onderzoeken die alle zijn gericht op de verdere verbetering van de CPI in de lijn van de zogenoemde Boskin-commissie die zich met de kwaliteit van de Amerikaanse CPI heeft beziggehouden. Over de opgedane ervaringen en de problemen die in een eerdere fase van het onderzoek werden ontmoet, is in een tussentijdse nota verslag gedaan[1]. Dij de berekening van de CPI wordt gebruik gemaakt van de Laspeyres-indexformule, die de prijsontwikkeling van een gedurende een aantal jaren constant gehouden pakket goederen en diensten weergeeft. Die benadering is ook voor de inzet van scannerdata gekozen, zij het dat het betrokken “mandje” jaarlijks wordt aangepast. Daarmee wordt alvast rekening gehouden met de jaarlijkse basisverlegging voor de gehele CPI, waarmee het CBS na de komende basisverlegging van de CPI naar 2000 wil starten. In dit rapport wordt beschreven op welke wijze de scannerdata nu daadwerkelijk worden gebruikt.

---

[1] “Using scanner data to compile price indices: experiences and practical problems”, Cecile Schut, 2001. Dit paper is gepresenteerd tijdens de Joint ECE/ILO Meeting on Consumer Price Indices op 1 en 2 november 2001 te Geneve.

De enorme massaliteit van de data dwingt tot een vooraf vastgestelde, zeer gestructureerde werkwijze om het productieproces qua doorlooptijd en kwaliteit beheersbaar te houden. Beheersbaarheid geldt als een absolute voorwaarde voor het productieproces van de CPI dat gekenmerkt wordt door korte productiecycli, harde deadlines en een output waaraan hoge kwaliteitseisen worden gesteld vanwege de grote maatschappelijke betekenis. Het onderzoek heeft geresulteerd in statistisch verantwoorde en in de praktijk goed hanteerbare oplossingen met gebruikmaking van relatief eenvoudige algoritmes voor de afbakening van het omvangrijke pakket artikelen, zodat de beoogde mate van beheersbaarheid zonder meer is bereikt.

Leeswijzer

In hoofdstuk 2 wordt een beschrijving gegeven van de structuur van de CPI zoals die wordt berekend. In hoofdstuk 3 wordt vervolgens een beschrijving gegeven van de gebruikte data. In hoofdstuk 4 wordt een beschrijving gegeven van de wijze waarop gebruik wordt gemaakt van de supermarktscannerdata. Deze beschrijving geeft inzicht in de gehanteerde werkwijze, zonder gebruik van technische formules. In de appendix wordt een meer formele beschrijving van de werkwijze gepresenteerd aan de hand van formules. De appendix kan worden beschouwd als verdieping van hoofdstuk 4, maar bevat geen nieuwe informatie. Tenslotte wordt in hoofdstuk 5 aandacht besteed aan verder onderzoek door het CBS naar het gebruik van scannerdata ten behoeve van de CPI.

2. Huidige werkwijze CPI[2]

2.1 Structuur[3]

De CPI beschrijft het gemiddelde prijsverloop van goederen en diensten (kortweg producten) die door een gemiddeld huishouden worden geconsumeerd. Uitgangspunt bij de berekening van de Nederlandse CPI is de Laspeyres-formule, die de prijsontwikkeling weergeeft van een consumptiepakket dat in de tijd constant wordt gehouden en is gebaseerd op gegevens uit een bepaalde basisperiode[4]. De Laspeyres-prijsindex is te zien als een gewogen gemiddelde van de prijsindexcijfers van de producten in het consumptiepakket. Het consumptiepakket op basis waarvan de CPI wordt bepaald, wordt ook wel het “mandje” genoemd.

---

[2] Deze paragraaf is gebaseerd op de nota “CBS-onderzoeksprogramma ter verbetering van de Consumentenprijsindex”, J. de Haan en L. Hoven, 2001, CBS, Voorburg.

[3] In de appendix wordt de huidige werkwijze nogmaals beschreven, echter inclusief de gebruikte formules.

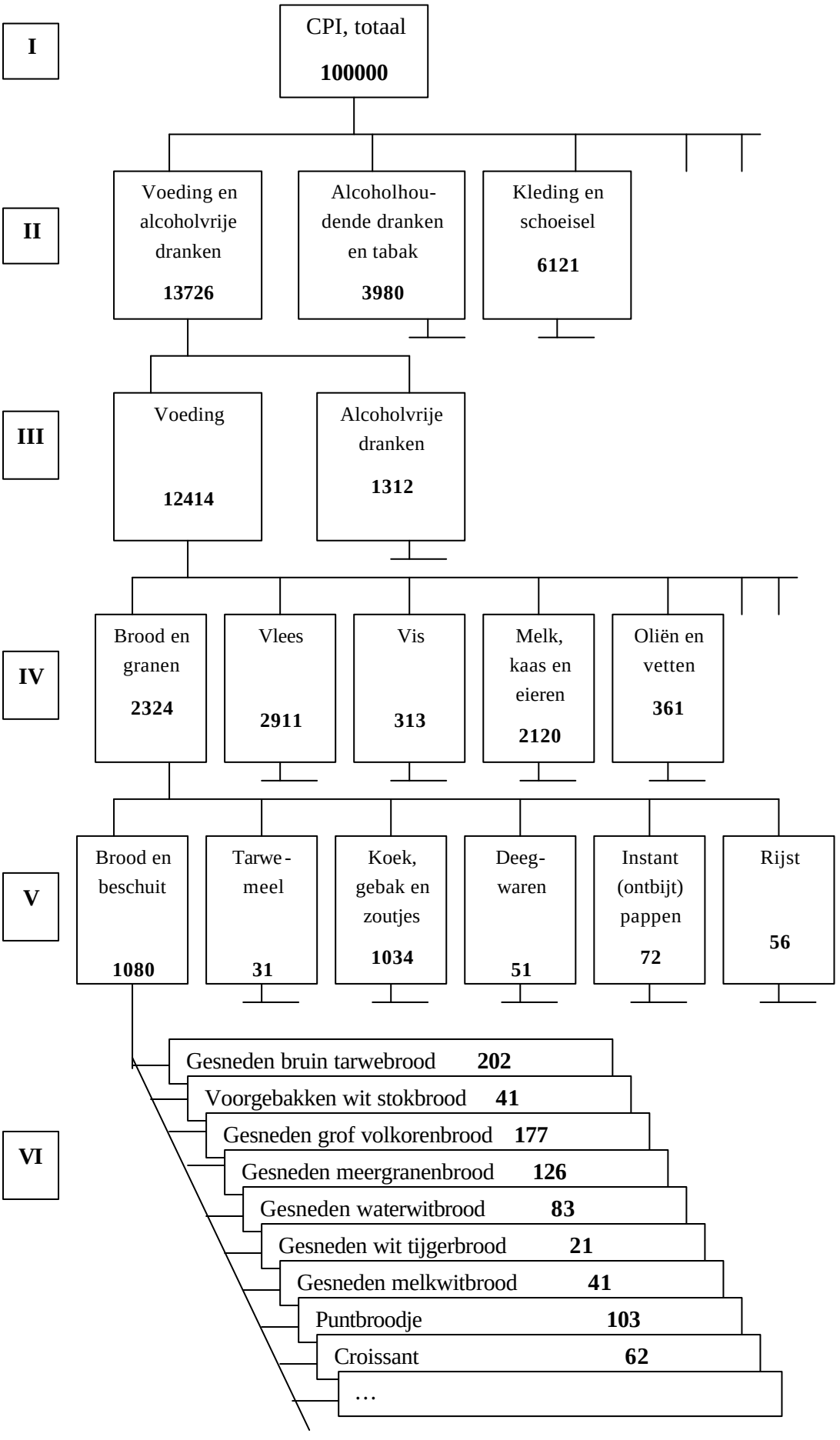
[4] De CPI die momenteel wordt gepubliceerd is gebaseerd op het consumptiepakket uit 1995. Vanaf januari 2003 zal worden overgegaan op het consumptiepakket uit 2000.

De prijsindexcijfers per product worden maandelijks vastgesteld door het vergelijken van prijzen van het product met prijzen uit de vorige maand. Hiervoor worden prijzen van dit product bij verschillende verkoopkanalen (winkels, markt etc.) gemeten. In de praktijk is het natuurlijk ondoenlijk om voor alle afzonderlijke producten die aan consumenten worden verkocht prijsindexcijfers te berekenen. Dat heeft te maken met de beschikbaarheid van gegevens, de belasting van berichtgevers en met kosten. Er wordt daarom volstaan met een steekproef van producten op basis waarvan de CPI wordt geschat. Om een en ander te illustreren verwijzen we naar het schema in figuur 1. Hierin zijn zes aggregatieniveaus van de consumptieve bestedingen weergegeven, genummerd I tot en met VI. Bovendien zijn de relatieve aandelen van de bestedingsgroepen in de CPI weergegeven (in honderdduizendsten). We geven een voorbeeld. De totale bestedingen (niveau I) worden op niveau II naar 13 hoofdcategorieën onderscheiden, waaronder 'Voeding en alcoholvrije dranken'. Deze hoofdgroep valt op niveau III uiteen in twee subgroepen, namelijk 'Voeding' en 'Alcoholvrije dranken'. Een nadere uitsplitsing daarvan gebeurt op niveau IV, met onder de voedingsmiddelen 'Brood en granen' als een van de onderdelen. Dat onderdeel wordt op niveau V weer verder uitgesplitst, met 'Brood en beschuit' als een van de productgroepen. Uit niveau V wordt vervolgens een steekproef van producten getrokken. In het voorbeeld betreft het op niveau VI meer dan tien producten, waaronder 'Gesneden bruin tarwebrood'.

Van de geselecteerde producten worden maandelijks in een steekproef van verkooppunten verspreid over het hele land prijzen waargenomen. Aan de hand van deze waargenomen prijzen worden productprijsindexcijfers berekend. De prijsindex van een gesneden bruin tarwebrood wordt bijvoorbeeld bepaald door elke maand de prijs van dit brood te meten bij circa 150 supermarkten en bakkers in het hele land. Deze prijzen worden verzameld door interviewers, die maandelijks in winkels prijzen noteren.

De prijsindexcijfers per product worden door het CBS niet gepubliceerd. Publicatie van prijsindexcijfers gebeurt wel op de niveaus I tot en met V. De berekening geschiedt getrapt: de productindexcijfers (niveau VI) worden met vaste gewichten samengewogen tot indexcijfers per productgroep (niveau V), die op hun beurt met vaste gewichten worden geaggregeerd tot indexcijfers op niveau IV, enzovoort, totdat de totaal-CPI wordt verkregen.

Figuur 1 - Structuur CPI



2.2 Vaststellen van het wegingschema van de CPI

Het wegingschema van de CPI wordt thans vastgesteld op basis van het Budgetonderzoek (BO). Voor dit onderzoek worden gedurende een kalenderjaar onder een steekproef uit de in Nederland wonende particuliere huishoudens gegevens verzameld over hun bestedingen. Bij de samenstelling van het wegingschema voor de CPI vormt het budgetonderzoek in het basisjaar het uitgangspunt. Regelmatig, tot dusverre iedere vijf jaar, wordt de basis van de CPI herzien. Op dit moment is 1995 het basisjaar. Vanaf januari 2003 wordt overgegaan op een actueler wegingsschema, gebaseerd op gegevens uit het jaar 2000.

Bij de samenstelling van het wegingschema worden in ieder geval alle bestedingscategorieën uit het budgetonderzoek van het basisjaar opgenomen met een aandeel van 0,5 promille of meer in de totale consumptieve bestedingen. Bestedingscategorieën die beneden die norm vallen, worden in het algemeen toegerekend aan verwante categorieën. Bestedingscategorieën uit het budgetonderzoek die erg omvangrijk zijn, worden – indien mogelijk – juist verder uitgesplitst; de onderdelen worden vervolgens als productgroepen (niveau V in figuur 1) opgenomen. Hiervoor maakt het CBS gebruik van het BO, maar ook van externe gegevens, zoals marktonderzoekgegevens en informatie van producenten en importeurs. Dergelijke gegevens worden ook gebruikt om de wegingsaandelen per product te bepalen.

2.3 Productkeuze

Het aantal geselecteerde productsoorten per productgroep is voornamelijk afhankelijk van het gewicht van de groep en deels ook van de heterogeniteit. Vuistregel is dat voor elke 50 wegingspunten (op een totaal van 100.000) één productsoort gekozen wordt. Waar sprake is van een grote heterogeniteit, wordt het aantal producten iets uitgebreid. In principe wordt met een zogenaamde *cut-off* steekproef gewerkt: per productgroep worden de qua omzet belangrijkste producten gekozen. In totaal neemt het CBS van ongeveer 1.700 producten maandelijks de prijzen waar. Hiervan zijn circa 400 producten verkrijgbaar in supermarkten.

3. Beschrijving van de scannerdata van supermarkten

Sinds een aantal jaren beschikken vrijwel alle Nederlandse supermarkten over kassasystemen waarbij verkochte producten worden gescand. Deze systemen zijn ingevoerd om het voorraadbeheer en de boekhouding van supermarkten te automatiseren. Veelal worden systemen gehanteerd waarbij automatisch producten worden besteld bij leveranciers precies op het moment dat een bepaald product (bijna) is uitverkocht. Het is duidelijk dat dergelijke systemen in belangrijke mate hebben bijgedragen aan de verbetering van de efficiency in de bedrijfsvoering van supermarkten. Alle supermarkten maken bij het scannen gebruik van dezelfde

standaard, namelijk de zogenaamde “European Article Numbers” (EAN), de welbekende streepjescodes.

Traditioneel worden voor de samenstelling van de consumentenprijsindex prijzen verzameld van een breed scala van goederen en diensten (zie hoofdstuk 2 voor een uitgebreide beschrijving van de werkwijze voor de berekening van de CPI). In Nederland worden, verspreid over het hele land, maandelijks circa 90.000 prijzen verzameld van ongeveer 1.700 verschillende producten. De opkomst van scannerdata leidde tot het idee om niet meer handmatig prijzen te verzamelen, maar om gebruik te maken van elektronische bestanden van supermarkten. Halverwege de jaren ’90 van de vorige eeuw, is het CBS begonnen met een oriëntatie op het verkrijgen van dergelijke gegevens. Hiertoe zijn een aantal supermarktketens in Nederland benaderd. Dit heeft geresulteerd in de reguliere levering van scannerdata van een aantal winkelketens die filialen hebben door heel het land.

Het CBS ontvangt wekelijks elektronisch bestanden waarin de omzet en de verkochte aantallen per EAN-code van tientallen filialen van supermarktketens zijn opgeslagen. Per EAN-code wordt tevens een korte omschrijving van het product gegeven. Het CBS ontvangt deze bestanden inmiddels geruime tijd. In totaal bevatten de bestanden over deze periode ruim 20.000 verschillende EAN-codes. Een flink deel van deze codes betreft overigens producten die inmiddels niet meer verkrijgbaar zijn, of oude codes die in de loop van de tijd door nieuwe codes zijn vervangen.

Voor het CBS levert de elektronische beschikbaarheid van scannerdata minder enquêtewerk in het veld op. Bovendien leidt het tot een aanzienlijke kwaliteitsverbetering van de CPI, enerzijds omdat er veel meer producten en prijzen kunnen worden meegenomen in de berekening van het indexcijfer waardoor het populatieconcept veel dichter kan worden benaderd, anderzijds omdat de scannerdata werkelijke transactieprijzen opleveren.

4. Beschrijving van het gebruik van scannerdata in de CPI

Met de beschikbaarheid van de grote hoeveelheden scannerdata van supermarkten, is het niet langer nodig om het prijsindexcijfer voor een productgroep te baseren op slechts een beperkt aantal producten. Echter, het werken met grote hoeveelheden data is minder eenvoudig dan het op het eerste gezicht lijkt. In de volgende paragrafen wordt verder uiteengezet op welke wijze de scannerdata worden gebruikt in de CPI.

4.1 Structuur

Natuurlijk is het voor veel producten zo dat ze niet alleen in supermarkten worden verkocht, maar ook via andere verkoopkanalen (bijvoorbeeld brood bij de warme bakker en vlees bij de slager op de hoek). Om een betrouwbaar beeld te krijgen van de prijsontwikkeling in Nederland, blijven de prijswaarnemingen bij deze groep van

"overige" verkoopkanalen bestaan[5]. Op basis van deze prijsgegevens worden deelindexcijfers voor productgroepen op niveau V berekend. Apart hiervan worden deelindexcijfers berekend aan de hand van de scannerdata. Deze deelindexcijfers worden per supermarktketen afzonderlijk berekend. Het is duidelijk dat niet voor elke productgroep deelindexcijfers beschikbaar kunnen komen uit de scannerdata, simpelweg omdat supermarkten slechts een deel van het totale consumptiepakket in hun assortiment hebben. De deelindexcijfers voor de "overige" verkoopkanalen en die gebaseerd op scannerdata worden daarna samengewogen met vaste gewichten tot indexcijfers op productgroepniveau V. De gewichten die hierbij worden gebruikt variëren per productgroep en zijn vastgesteld aan de hand van de gewichten die uit het grondmateriaal van de CPI worden afgeleid. Indirect zijn deze gewichten afkomstig van informatie over marktaandelen van supermarkten voor de desbetreffende productgroep. In de volgende paragrafen wordt uiteengezet op welke wijze de deelindexcijfers die zijn gebaseerd op de scannerdata worden samengesteld.

4.2 Bepalen van indexcijfers op productgroepniveau op basis van supermarktscannerdata

Bij het samenstellen van indexcijfers op productgroepniveau op basis van de scannerdata wordt een aantal stappen doorlopen. Er worden indexcijfers bepaald voor elke supermarktketen afzonderlijk. Uitgangspunt hierbij is dat van een vaste groep EAN-codes het prijsverloop wordt gevolgd. Hiertoe worden eerst de gegevens van de verschillende filialen van een supermarktketen bij elkaar geteld. Daarna worden de afzonderlijke EAN-codes ingedeeld in productgroepen. Omdat er veel producten, of beter gezegd EAN-codes, verdwijnen en verschijnen is het praktisch ondoenlijk om alle data uit de scannerdatabestanden mee te nemen bij de bepaling van prijsindexcijfers. Daarom wordt op basis van de gegevens uit een vaste periode een "basismandje" met een beperkt aantal EAN-codes samengesteld. Op basis van dit mandje worden dan tenslotte Laspeyres-prijsindexcijfers berekend. Daarbij worden de totaalomzetten van de EAN-codes in het basismandje gebruikt als gewichten.

In de onderstaande paragrafen worden deze stappen verder beschreven.

4.2.1 Filiaalgegevens optellen

Per supermarktketen worden weekgegevens van verschillende filialen bij elkaar geteld. Dit gebeurt door de gemeten omzetten en de verkochte hoeveelheden per week voor elke afzonderlijke EAN-code over de filialen heen op te tellen. Hierdoor ontstaan bestanden waarin per EAN-code de omzetten en verkochte hoeveelheden in alle filialen waarvan data zijn ontvangen van een bepaalde supermarktketen op weekbasis terug zijn te vinden.

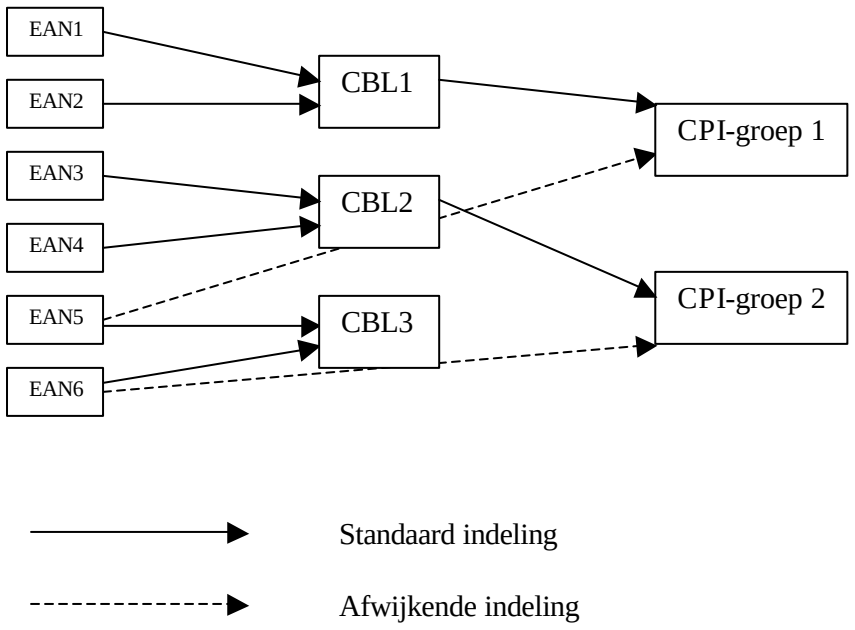
[5] Dit geldt overigens ook voor waarnemingen bij supermarktketens waarvan we geen scannerdata ontvangen.

4.2.2 Indelen van de gegevens

Zoals in hoofdstuk 3 is aangegeven, bevatten de bestanden circa 20.000 verschillende EAN-codes. De eerste stap bij het maken van prijsindexcijfers voor productgroepen is dan ook het indelen van de EAN-codes in groepen, die overeenkomen met de productgroepen op niveau V (zie figuur 1).

Het handmatig indelen van de EAN-codes is niet alleen veel werk, het is bovendien lastig, omdat slechts een korte productomschrijving beschikbaar is. Gelukkig bleek dat het Centraal Bureau voor Levensmiddelenhandel (CBL) de EAN-codes indeelt in een productclassificatie die in samenwerking met haar leden is opgesteld. Deze productclassificatie is wat gedetailleerder dan de indeling van de CPI op het laagste niveau. Door het aanbrengen van een koppeling tussen de circa 290 CBL-productgroepen en de 53 productgroepen op niveau V van de CPI, kunnen de EAN-codes uiteindelijk worden ingedeeld in CPI-productgroepen. Een beperkt aantal CBL-groepen was overigens niet eenduidig in te delen in een CPI-groep. De EAN-codes in deze CBL-groepen zijn handmatig ingedeeld in de CPI-groepen. In figuur 2 wordt het indelen van EAN-codes geïllustreerd.

Figuur 2 – Indeling van de EAN-codes in CBL-groepen en CPI-productgroepen.



4.2.3 Samenstelling van het basismandje

Zoals eerder opgemerkt, is het ondoenlijk om alle gegevens uit de scannerdatabestanden mee te nemen bij de berekening van deelprijsindexcijfers. Het is daarom nodig een selectiemechanisme te hebben dat bepaalt welke producten wel en welke producten niet mogen worden gebruikt.

Voor het samenstellen van het basismandje en het vaststellen van gewichten van de producten in het basismandje wordt een vol kalenderjaar aan gegevens gebruikt. Het

jaar waarop het mandje gebaseerd is noemen we het basisjaar. Door de snelle beschikbaarheid van de scannerdata, is het mogelijk voor de berekening van de indexcijfers over 2002, het jaar 2001 als basisjaar te gebruiken.

Zoals beschreven in paragraaf 2.2, is het wegingsschema van de huidige CPI gebaseerd op bestedingsaandelen van huishoudens in 1995. Dit betreft de gewichten op de niveaus I tot en met V zoals weergegeven in figuur 1, die na toevoegen van de scannerdataindexcijfers niet worden aangepast. Het basisjaar van de CPI blijft dus 1995, alleen de gewichten van de individuele producten waarvoor prijzen worden waargenomen (niveau VI in figuur 1) zijn op recentere gegevens gebaseerd. Dit is een aanmerkelijke verbetering; er wordt op deze wijze immers beter rekening gehouden met het echte uitgavenpatroon van consumenten.

De gewichten op niveau I tot en met V zullen vanaf januari 2003 worden gebaseerd op basisjaar 2000. Het scannerdatagedeelte zal jaarlijks worden aangepast, en vanaf 2003 dus zijn gebaseerd op basisjaar 2002.

In de rest van deze paragraaf wordt geschetst hoe het selectiemechanisme werkt om te komen tot een basismandje met producten.

Eerst wordt bepaald welke producten als “echte” supermarktproducten kunnen worden beschouwd. Als we de producten uit de scannerdata classificeren volgens de CBL-indeling, dan valt een grote diversiteit aan producten die bij supermarkten worden verkocht op. In eerste instantie hebben we ons echter geconcentreerd op die productgroepen waarvoor consumenten vooral naar de supermarkt gaan. Dit zijn met name voedingsmiddelen en daarnaast een aantal huishoudelijke artikelen en verzorgingsproducten. De selectie van deze productgroepen is gebeurd door analyse van de huidige waarneming voor de CPI. Buiten de boot vielen bijvoorbeeld sokken, theedoeken, cd’s, boeken, serviesgoed en speelgoed. Ook specifieke seizoensproducten zoals aardappelen, verse groenten, vers fruit en bloemen, maar ook rookwaren zijn in dit stadium buitengesloten. Voor seizoensproducten geldt dat ze niet het gehele jaar door verkrijgbaar zijn. Om dergelijke producten op een juiste wijze opgenomen te krijgen in de CPI, is het gebruik van een basismandje dat het hele jaar hetzelfde is, niet direct geschikt. Voor deze producten is ervoor gekozen om de huidige wijze van prijswaarneming voor een beperkte groep producten in winkels te vervangen door het waarnemen van gemiddelde transactieprijzen uit de scannerdatabestanden.

De volgende stap is gebaseerd op de gedachte dat het voor het bepalen van een prijsindexcijfer vooral van belang is die producten mee te nemen die consumenten “veel” kopen. Dit idee is uitgewerkt door het toepassen van een aantal regels om te komen tot een basismandje. Deze regels houden kortgezegd in, dat alleen producten met EAN-codes die 48 of meer weken in het basisjaar zijn verkocht, in aanmerking komen voor het mandje. Om praktische redenen is bovendien de aanvullende eis gesteld dat de codes in elk geval in de laatste 4 weken van het basisjaar aanwezig moeten zijn.

De geselecteerde groep van EAN-codes wordt vervolgens ingedeeld in CBL-groepen. Per CBL-groep worden vervolgens de in het basisjaar behaalde omzetten

bepaald. Deze CBL-groepen worden vervolgens geordend aflopend op omzet, en alleen die CBL-groepen die cumulatief 80% van de totale omzet in het basisjaar vormen, worden geselecteerd. Dit betekent dat betrekkelijk kleine CBL-groepen met slechts een of enkele producten die bovendien slechts een zeer beperkt deel van de totaalomzet vormen, niet worden meegenomen in de verdere berekeningen.

Daarna wordt gekeken in hoeverre de geselecteerde CBL-groepen voldoende vulling opleveren voor de CPI-productgroepen op niveau V. Wanneer de CPI-productgroepen onvoldoende blijken te zijn gevuld, dan worden alsnog CBL-productgroepen toegevoegd aan het basismandje. Het criterium hiervoor is dat elke CPI-productgroep tenminste voor 50% gevuld is.

Het toepassen van de bovengenoemde regels leiden tot basismandjes per supermarkt die tussen de 6.000 en 7.000 EAN-codes bevatten. Hoewel bovengenoemde regels om te komen tot een basismandje in eerste instantie wellicht wat rigoureus lijken, blijkt in de praktijk dat de geselecteerde EAN-codes in het basisjaar meer dan 80% van de in het basisjaar behaalde omzet vormen.

4.2.4 Berekening van indexcijfers

Na selectie van de EAN-codes in het basismandje, worden er prijsindexcijfers berekend per geselecteerde CPI-groep. Deze indexcijfers zijn een gewogen gemiddelde van de prijsveranderingen per EAN-code, waarbij de (relatieve) gewichten per EAN-code worden bepaald door de totale omzet van het desbetreffende product in het basisjaar. Voor de berekening van de prijsverandering per EAN-code worden gemiddelde transactieprijzen berekend, door de totale omzet van een product over alle beschikbare filialen te delen door de totale hoeveelheid.

De CPI wordt maandelijks gepubliceerd, meestal binnen 10 dagen na afloop van een maand. Dit betekent dat er slechts een beperkte tijd beschikbaar is voor het productieproces, controle en analyse van de indexcijfers. Dit maakt het niet mogelijk om alle scannerdata over een bepaalde maand te gebruiken. Er is dan ook voor gekozen om alleen de eerste twee weekbestanden die volledig in een maand[6] vallen te gebruiken voor het berekenen van de gemiddelde transactieprijzen.

4.3 Omgaan met producten die verdwijnen uit het assortiment

De werkwijze zoals hierboven beschreven werkt natuurlijk alleen als alle EAN-codes in het basismandje in elke verslagperiode in het assortiment van de supermarkten blijven. In de praktijk blijkt dat er vrijwel elke maand een aantal

[6] De weekbestanden hebben betrekking op gegevens van maandag tot en met zondag. Dit betekent dat sommige bestanden gegevens bevatten uit twee verschillende kalendermaanden. Deze weekbestanden komen niet in aanmerking voor de berekening van het Europees geharmoniseerde indexcijfer (HICP), omdat in Europees verband is afgesproken dat de prijsgegevens die worden gebruikt om indexcijfers over een bepaalde maand te meten, afkomstig moeten zijn uit diezelfde maand. Hoewel de dekking van de HICP anders is dan die van de CPI, worden voor de samenstelling van de HICP grotendeels dezelfde indexcijfers op productniveau gebruikt als in de CPI

EAN-codes niet meer in de bestanden zijn terug te vinden. Dit kan verschillende oorzaken hebben. In sommige gevallen krijgt een product een andere EAN-code, bijvoorbeeld na restyling. In andere gevallen verandert een product wezenlijk, er komt bijvoorbeeld meer of juist minder in een verpakking, of de kwaliteit van het product verandert. Ook komt het voor dat supermarkten besluiten producten niet langer in hun assortiment aan te bieden. Een test op gegevens uit 2000 en 2001 heeft uitgewezen dat in 2001 ruim 20% van de EAN-codes in het basismandje dat was gebaseerd op het jaar 2000, verdween.

Analyse van producten die verdwijnen, heeft uitgewezen dat het in ruim de helft van de gevallen gaat om producten die een relatief klein omzetaandeel vertegenwoordigen. Daarom is gekozen voor een werkwijze waarbij alleen EAN-codes die een wegingsaandeel van meer dan 2,5% hebben binnen de CBL-groep waarin ze horen, verder onderzocht worden. Voor deze groep van EAN-codes worden opvolgers of vervangende EAN-codes aangewezen. Wanneer het producten betreft die zijn gewijzigd in kwaliteit of een andere verpakkingshoeveelheid hebben gekregen, dan wordt de prijsverandering hiervoor gecorrigeerd. In gevallen waarin het product uit het assortiment van de supermarkt is verdwenen, wordt er een nieuwe EAN-code opgevoerd van een vergelijkbaar product. Het is duidelijk dat in dat geval de prijs van het verdwenen product niet direct vergelijkbaar is met de prijs van het gekozen vervangende product. In die gevallen wordt dan ook op het moment van overgang op het nieuwe product gebruik gemaakt van de gemiddelde prijsverandering van andere producten in dezelfde CBL-groep.

Voor verdwijnende EAN-codes die slechts een klein omzetaandeel binnen de CBL-groep hebben, wordt tot aan de overgang op een nieuw basismandje de gemiddelde prijsverandering van de andere producten in dezelfde CBL-groep gebruikt. Deze methode wordt ook toegepast wanneer een EAN-code slechts tijdelijk afwezig is, bijvoorbeeld omdat het betreffende product tijdelijk niet leverbaar is geweest. Tenslotte wordt elke maand nog gecontroleerd of tenminste 80% van de EAN-codes (gemeten in omzetten in het basisjaar) per CBL-groep nog aanwezig is (al dan niet inmiddels vervangen door een andere EAN-code). Wanneer dit niet het geval is, kan het zijn dat er veel EAN-codes zijn weggevallen die een relatief klein wegingsaandeel (minder dan 2,5%) hadden, maar die bij elkaar wel een grote groep vormen. In dat geval wordt er nog aanvullend naar vervangers gezocht.

5. Toekomstige ontwikkelingen

5.1 Verbeteringen aan de nieuwe werkwijze

In de tweede helft van dit jaar zal het onderzoek worden voortgezet naar het gebruik van de reeds beschikbare scannerdata voor seizoensproducten. Daarnaast zal onderzoek worden gedaan naar het beste moment waarop een verdwijnend artikel moet worden vervangen. Een artikel verdwijnt immers meestal niet van het ene moment op het andere uit het assortiment, maar vertoont veelal zowel qua prijs als

qua verkochte hoeveelheid en dus ook qua omzet gedurende kortere of langere tijd een aflopende tendens. De “afstervingscurves” verdienen nader onderzoek om het optimale tijdstip van vervanging vast te stellen.

EAN-coderingen spelen een sleutelrol bij de verwerking van scannerdata. Door EAN Nederland, de beheerder van de EAN-codering, is een systeem ontwikkeld onder de naam EAN DAS (DAS staat voor Digital Alignment System), waarbinnen fabrikanten alle logistieke gegevens behorende bij hun producten kwijt kunnen. Dit systeem bevat niet alleen informatie met betrekking tot de afmetingen van het product, maar ook gegevens over de inhoud, de EAN code, welke code een nieuwe product vervangt, vanaf wanneer het leverbaar is etc. Met behulp van deze informatie kunnen verdwijnende producten een stuk eenvoudiger worden vervangen door passende alternatieven. EAN Nederland heeft toegezegd dat het CBS op korte termijn de beschikking krijgt over een applicatie waarmee het EAN DAS-systeem kan worden benaderd. Daardoor zal de hanteerbaarheid van scannerdata aanzienlijk toenemen.

5.2 Scannerdata voor andere artikelgroepen

In de komende jaren zet het CBS hoog in op het langs elektronische weg verzamelen van prijsgegevens. Het gebruik van kassascangegevens en van informatie op en via het internet (websites en e-mailwaarneming) zal de traditionele prijswaarneming met behulp van interviewers en papieren enquêteformulieren geleidelijk vervangen. Het ligt voor de hand dat waar mogelijk het gebruik van scannerdata van supermarkten verder wordt uitgebreid, omdat de penetratiegraad van kassascanners in deze branche nu al zeer hoog is. Kassascanners worden echter ook in andere branches op grote schaal toegepast, zodat de acquisitie van scannerdata voor andere artikelgroepen dan supermarktartikelen hoge prioriteit krijgt. Het gebruik van scannerdata voor het zogenaamde wit- en bruingoed is reeds in voorbereiding. Daarna zal de aandacht worden verlegd naar drogisterij- en doe-het-zelf-producten. Het voornaamste doel van deze continue procesvernieuwing is de verbetering van de kwaliteit van de CPI. Daarnaast wordt de enquêtedruk op het bedrijfsleven verder gereduceerd en biedt het innovatieproces diverse mogelijkheden efficiencywinsten te behalen.

APPENDIX – Technische beschrijving van het gebruik van scannerdata in de CPI

In de appendix wordt een technische beschrijving van het gebruik van scannerdata gegeven. Allereerst wordt de huidige structuur van de CPI, zoals die is beschreven in hoofdstuk 2, opnieuw weergegeven, echter dit keer uitgebreid met formules. Daarna wordt de structuur van hoofdstuk 4 gevolgd, waarbij dieper wordt ingegaan op rekentechnische details en de specifieke wijze van de berekening van de CPI.

A Huidige werkwijze CPI[7]

De CPI beschrijft het gemiddelde prijsverloop van goederen en diensten (kortweg producten) die door een gemiddeld huishouden worden geconsumeerd. Uitgangspunt bij de Nederlandse CPI is de Laspeyres-formule die de prijsontwikkeling weergeeft van een consumptiepakket dat in de tijd constant is[8]. De (partiële) prijsindex van product i (i= 1,...,N) voor periode t vergeleken met de basisperiode 0 geven we weer met $p_i^{t,0}$. De Laspeyres-formule weegt de indices van de producten met hun bestedingsaandelen uit de basisperiode w_i^0. De Laspeyres-prijsindex voor periode t vergeleken met de basisperiode 0 is:

$$P^{t,0} = \sum_{i=1}^{N} w_i^0 p_i^{t,0} \,. \qquad (1)$$

Formule (1) is het populatiemodel van de CPI, dat geldt voor alle producten die in principe zijn te onderscheiden. In de praktijk is het ondoenlijk om prijsindexcijfers te berekenen voor alle producten die consumenten kopen. Dat heeft te maken met de beschikbaarheid van gegevens, de belasting van berichtgevers en de kosten. Er wordt daarom volstaan met een steekproef van producten op basis waarvan de CPI wordt geschat. Ter illustratie verwijzen we naar het schema in figuur 1. Hierin zijn zes aggregatieniveaus van de consumptieve bestedingen weergegeven, genummerd I tot en met VI. Ook zijn de relatieve aandelen van de bestedingsgroepen in de CPI vermeld (in honderdduizendsten). We geven een voorbeeld. De totale bestedingen (niveau I) worden op niveau II naar 13 hoofdcategorieën onderscheiden, waaronder 'Voeding en alcoholvrije dranken'. Deze hoofdgroep valt op niveau III uiteen in twee subgroepen, namelijk 'Voeding' en 'Alcoholvrije dranken'. Een nadere uitsplitsing daarvan gebeurt op niveau IV, met 'Brood en granen' als een van de onderdelen van voedingsmiddelen. Dat onderdeel wordt op niveau V weer verder uitgesplitst, met 'Brood en beschuit' als een van de productgroepen. Hieruit wordt

---

[7] Deze paragraaf is gebaseerd op de nota "CBS-onderzoeksprogramma ter verbetering van de Consumentenprijsindex", J. de Haan en L. Hoven, 2001, CBS, Voorburg.

[8] De CPI die momenteel wordt gepubliceerd is gebaseerd op het consumptiepakket uit 1995. Vanaf januari 2003 zal worden overgegaan op het consumptiepakket uit 2000.

vervolgens een steekproef van producten getrokken. In het voorbeeld betreft het op niveau VI meer dan tien producten, waaronder ‘Gesneden bruin tarwebrood’.

Het CBS neemt van de geselecteerde producten maandelijks de prijzen waar in een steekproef van verkooppunten verspreid over het land. De prijs van een gesneden bruin tarwebrood wordt bijvoorbeeld waargenomen bij circa 150 supermarkten en bakkers. Door de waargenomen prijzen in de verslagperiode te middelen en te delen door de gemiddelde prijs in de basisperiode ontstaan prijsindexcijfers per product. Het middelen van de prijzen gebeurt ongewogen rekenkundig. In symbolen geldt voor de schatter van $p_i^{t,0}$:

$$\hat{p}_i^{t,0} = \frac{\sum_{b \in \hat{B}_i} p_{bi}^t / n_{\hat{B}_i}}{\sum_{b \in \hat{B}_i} p_{bi}^0 / n_{\hat{B}_i}}, \tag{2}$$

waarin $\hat{B}_i$ de steekproef (ter grootte $n_{\hat{B}_i}$) van verkooppunten *b* is waar de prijs $p_{b,i}^s$ ($s = 0, t$) van product *i* wordt waargenomen. Het CBS publiceert de prijsindices per product overigens niet.

Stel, er zijn in totaal *n* producten geselecteerd. Voor elk product *i* in de steekproef is een schatting gemaakt van het toegerekende aandeel $\hat{w}_i^0$ in de bestedingen van de basisperiode. De schatter van de CPI kan dan worden geschreven als:

$$\hat{P}^{t,0} = \sum_{i=1}^{n} \hat{w}_i^0 \hat{p}_i^{t,0}. \tag{3}$$

De berekening vindt in de praktijk getrapt plaats. Eerst worden de prijsindexcijfers per product (niveau VI) geaggregeerd met vaste gewichten tot deelindexcijfers per productgroep op niveau V. Deze worden op hun beurt geaggregeerd met vaste gewichten tot indexcijfers op niveau IV, enzovoort, totdat de CPI (niveau I) wordt verkregen. Voor een willekeurige CPI-productgroep *c* is het prijsindexcijfer te berekenen als:

$$\hat{P}_c^{t,0} = \sum_{i \in c} \left(\frac{\hat{w}_i^0}{\sum_{i \in c} \hat{w}_i^0} \right) \hat{p}_i^{t,0}. \tag{4}$$

B Technische beschrijving van het gebruik van scannerdata in de CPI

Met de beschikbaarheid van de grote hoeveelheden scannerdata van supermarkten is het niet langer nodig om het prijsindexcijfer voor een productgroep te baseren op slechts een beperkt aantal producten. Het werken met grote hoeveelheden data is minder eenvoudig dan het op het eerste gezicht lijkt. Hierna wordt uiteengezet op welke wijze scannerdata in de CPI worden verwerkt.

B.1 Structuur

Veel producten die in supermarkten te koop zijn, zijn ook verkrijgbaar bij andere verkoopkanalen, bijvoorbeeld brood bij de warme bakker en vlees bij de slager op de hoek. Om een betrouwbaar beeld te krijgen van de prijsontwikkeling blijft het CBS ook bij deze groep van "overige" verkoopkanalen prijzen waarnemen. Analoog aan formule (2) worden productprijsindices voor de overige verkoopkanalen geschat volgens:

$$\hat{p}_{i,O}^{t,0} = \frac{\sum_{b \in \hat{B}_{i,O}} p_{b,i}^{t} / n_{\hat{B}_{i,O}}}{\sum_{b \in \hat{B}_{i,O}} p_{b,i}^{0} / n_{\hat{B}_{i,O}}} \tag{5}$$

waarin $\hat{B}_{i,O}$ de steekproef van overige verkoopkanalen ter grootte $n_{\hat{B}_{i,O}}$ voorstelt. In het geval van producten die zowel in supermarkten als in andere winkels verkocht worden geldt $\hat{B}_{i,O} \subset \hat{B}_i$; voor de resterende producten geldt uiteraard $\hat{B}_{i,O} = \hat{B}_i$. Analoog aan (4) worden prijsindexcijfers per productgroep op niveau V voor de overige verkoopkanalen per CPI-productgroep *c* berekend als:

$$\hat{P}_{c,O}^{t,0} = \sum_{i \in c} \left(\frac{\hat{w}_i^0}{\sum_{i \in c} \hat{w}_i^0} \right) \hat{p}_{i,O}^{t,0} \ . \tag{6}$$

Los hiervan worden prijsindexcijfers op productgroepniveau berekend op basis van scannerdata. Dat gebeurt voor iedere supermarktketen afzonderlijk. De indexcijfers gebaseerd op scannerdata en die voor de overige verkoopkanalen worden daarna samengewogen met vaste gewichten tot indexcijfers op productgroepniveau V (zie hiervoor paragraaf B.2.6). De gewichten variëren per productgroep en berusten op informatie over marktaandelen van supermarkten voor de desbetreffende productgroep. In de volgende paragrafen wordt uiteengezet hoe de deelindexcijfers die zijn gebaseerd op de scannerdata worden samengesteld.

B.2 Bepalen van indexcijfers op productgroepniveau op basis van supermarktscannerdata

Bij het samenstellen van indexcijfers op productgroepniveau op basis van de scannerdata worden een aantal stappen doorlopen. Er worden indexcijfers bepaald voor elke supermarktketen afzonderlijk. Hiertoe worden eerst de gegevens van de verschillende filialen van een supermarktketen bij elkaar geteld. Daarna worden de afzonderlijke EAN-codes ingedeeld in productgroepen. Omdat er veel producten, of beter gezegd EAN-codes, verdwijnen en verschijnen is het praktisch ondoenlijk om alle data uit de scannerdatabestanden mee te nemen bij de bepaling van prijsindexcijfers. Daarom wordt op basis van de gegevens uit een vaste periode een "basismandje" met slechts een beperkt aantal EAN-codes samengesteld. Op basis van dit mandje worden dan tenslotte Laspeyres-prijsindexcijfers berekend.

In de onderstaande paragrafen worden deze stappen verder beschreven. Eerst introduceren we de volgende notaties:

- B^* als de verzameling van *berichtgevers* die supermarktscannerdata leveren ten behoeve van de CPI; laten $b, b', \text{K} \in B^*$ variëren over deze verzameling.
- F_b als de verzameling van *filialen* van berichtgever b waarvan scannerdata worden geleverd; laten $f, f', ... \in F_b$ variëren over deze verzameling.
- I_b als de verzameling van alle *producten* (EAN-codes) die betrokken zijn geweest in een transactie van enig filiaal van berichtgever b; laten $i, i', \text{K}\ j, j', ... \in I_b$ variëren over deze verzamelingen. Merk op dat niet noodzakelijkerwijs $I_b \cap I_{b'} = \varnothing$, voor $b \neq b'$, met andere woorden dat bepaalde producten i verkocht kunnen worden bij verschillende berichtgevers.
- $v_{i,f,b}^t$ de *omzet* van product $i \in I_b$ in filiaal $f \in F_b$ in week t.
- $q_{i,f,b}^t$ de *verkochte hoeveelheid* (aantal stuks) van product $i \in I_b$ in filiaal $f \in F_b$ in week t.

B.2.1 Filiaalgegevens optellen

Per supermarktketen worden weekgegevens van verschillende filialen bij elkaar geteld volgens:

$$v_{i,b}^t = \sum_{f \in F_b} v_{i,f,b}^t \tag{7}$$

en

$$q_{i,b}^t = \sum_{f \in F_b} q_{i,f,b}^t \tag{8}$$

en dus is $v_{i,b}^t$ de totale omzet voor product i bij alle filialen in het scannerdatabestand van berichtgever b in week t en $q_{i,b}^t$ de bijbehorende hoeveelheid.

B.2.2 Indelen van gegevens

Zie voor de werkwijze van het indelen van de EAN-codes in productgroepen paragraaf 4.2.2. Hier introduceren we verder de volgende notatie:

- C de totale verzameling van *CPI-productgroepen* en $C_b \subseteq C$ de verzameling van (vooraf) *geselecteerde CPI-productgroepen* die van betekenis zijn voor berichtgever b; laten $c, c', \text{K} \in C$ variëren over deze verzamelingen.

- L de totale verzameling van *CBL-productgroepen*; laten $l, l', \text{K} \in L$ variëren over deze verzameling.
- c_i de CPI-*productgroep waartoe product* i *behoort*, en l_i de *CBL-groep waartoe* i *behoort*. Merk op dat, voor producten $i, i' \in I_b$, $i \neq i'$, niet noodzakelijkerwijs $c_i = c_{i'}$ indien $l_i = l_{i'}$. Dus, hoewel de indeling van producten in CBL-productgroepen in het algemeen een gedetailleerder is dan een indeling in CPI-productgroepen, is het niet zo dat een CBL-productgroep altijd in zijn geheel binnen een CPI-productgroep valt (zie ook figuur 2 op pagina 8).

B.2.3 Samenstelling van het basismandje

In deze paragraaf wordt in formele termen herhaald wat is vastgelegd in paragraaf 4.2.3. Om te kunnen vaststellen in welke mate het beoogde basismandje een representatieve doorsnede is van alle in een jaar verkochte producten, is het noodzakelijk eerst vast te stellen welke producten we *a priori* beschouwen en welke we uitsluiten. Dit geeft een verzameling producten die we vanaf nu het *EAN-domein* zullen noemen. Dit domein – en preciezer: de omzetten die zijn gemoeid met de producten eruit – zal bepalend zijn voor de uiteindelijke samenstelling van het basismandje: we gebruiken dit domein als ijkpunt voor de *vullingsgraad* van het basismandje.

Allereerst beschouwen we natuurlijk alleen producten die ondergeschikt zijn aan een CPI-productgroep welke we vertegenwoordigd willen zien in de indexcijfers gebaseerd op de scannerdata, zoals beschreven in paragraaf 4.2.3; het is onzinnig om te spreken over een totaalomzet van alle producten als maat voor de vullingsgraad van het basismandje, wanneer in die totaalomzet producten zijn vertegenwoordigd die nooit in het basismandje zullen terechtkomen. Een tweede criterium is minder voor de hand liggend: ook producten die niet ingedeeld zijn in productgroepen, zullen niet worden meegenomen in het EAN-domein (dat wil zeggen: producten waarvan de EAN-codering nog niet is gekoppeld met een CBL- of CPI-productgroep). Het komt namelijk voor dat de bestanden producten bevatten die (nog) niet door het CBL zijn ingedeeld in een productgroep. Wanneer dit om producten gaat die een aanzienlijk omzetaandeel hebben, dan wordt geprobeerd zelf een codering aan te brengen. Tenslotte sluiten we uit van ons EAN-domein alle producten met een omschrijving ‘artikel onbekend’ (of die een leeg veld hebben op de plek van hun productomschrijving).

De eerste rekenstap die nodig is voor het samenstellen van het basismandje is het bepalen van de totale omzet over alle transacties van elk afzonderlijk product i in alle filialen van berichtgever b in het basisjaar T^0. Deze totaalomzet noteren we als $v_{i,b}^{T^0}$ en is gedefinieerd als:

$$v_{i,b}^{T^0} = \sum_{t \in T^0} v_{i,b}^t \qquad (9)$$

waarbij t varieert over alle weken in basisjaar T^0.

Definieer verder $d_{i,b}^{T^0}$ het *aantal weken* (of: de *duur*) dat product i bij berichtgever b werd verkocht in het basisjaar T^0. Definieer $I_b^{48} = \left\{ i \in I_b \mid d_{i,b}^{T^0} \geq 48 \right\}$ als de verzameling van *48-weken-producten* (van berichtgever b in het basisjaar). Laat bovendien I_b^{12} de verzameling zijn van *laatste-12-weken-producten*; dit zijn *product*en die de laatste 12 weken van het basisjaar continu werden verkocht.

Laat, voor willekeurige $C' \subseteq C$, $I_b(C') = \{ i \in I_b \mid c_i \in C' \}$ en $I_b^{48}(C') = I_b(C') \cap I_b^{48}$, en laat voor willekeurige $L' \subseteq L$, $I_b(L') = \{ i \in I_b \mid l_i \in L' \}$ en $I_b^{48}(L') = I_b(L') \cap I_b^{48}$ (laat ook, voor $c \in C$, $I_b(c) = I_b(\{c\})$ en definieer $I_b^{48}(c)$, $I_b(l)$ en $I_b^{48}(l)$ overeenkomstig).

Het *EAN-domein* E_b *voor berichtgever* b definiëren we nu als $I_b(C_b)$ (en het *EAN-domein* E als $\bigcup_{b \in B} E_b$, maar omdat in het algemeen $E_b \cap E_{b'} \neq \varnothing$, maken we van E verder geen gebruik).

Het *basismandje* $M_b^{T^0}$ uit basisjaar T^0 (voor berichtgever b) wordt nu gedefinieerd als de vereniging $M_{b,1}^{T^0} \cup M_{b,2}^{T^0} \cup M_{b,3}^{T^0}$ van *product*verzamelingen $M_{b,k}^{T^0}$. De drie afzonderlijke *product*verzamelingen worden als volgt samengesteld:

1. Kies de kleinste verzameling van CBL-productgroepen zodat de totale omzet van alle 48-weken-producten die eraan ondergeschikt zijn minimaal 80% is van de totale omzet in het EAN-domein (wanneer deze verzameling niet bestaat, kiezen we *alle* CBL-productgroepen). Verzamel al deze 48-weken-producten uit die verzameling van CBL-productgroepen in het voorlopige basismandje. (Dit komt neer op het aflopend ordenen van CBL-productgroepen naar de omzet van alle 48-weken-producten die eraan ondergeschikt zijn en de grootste te kiezen die samen aan de 80%-eis voldoen.). In formele termen:

   $M_{b,1}^{T^0} = I_b^{48}\left(L_b^{80}\right) \cap E_b$, waarbij $L_b^{80} \subseteq L$ de kleinste verzameling X is, zodanig dat $\dfrac{\sum_{i \in I_b^{48}(X) \cap E_b} v_{i,b}^{T^0}}{\sum_{i \in E_b} v_{i,b}^{T^0}} \geq 0.8$ mits deze bestaat, en $L_b^{80} = L$ anders.

2. Voor CPI-productgroepen die niet voor 50% zijn gevuld (voor wat betreft de totale omzet van producten uit het voorlopige basismandje, beperkt tot deze productgroep, in verhouding tot de totale omzet van alle producten uit het EAN-domein, beperkt tot deze productgroep) verzamelen we wederom alle

CBL-productgroepen (waarvan we in stap 1 de producten nog niet aan het voorlopige basismandje hebben toegevoegd) waarvan de 48-weken-producten bijdragen aan de totaalomzet van de CPI-productgroep. We ordenen deze CBL-quanta weer naar hun aandeel in deze totaalomzet en we kiezen weer de bovenste in deze lijst, zodanig dat aan de 50%-eis wordt voldaan (of ze zijn uitgeput, in welk geval we ze alle kiezen). We verzamelen alle 48-weken-producten die hiermee zijn gemoeid in het voorlopige basismandje. In formele termen:

$M_{b,2}^{T^0} = I_b^{48}\left(L_b^{50}\right) \cap E_b$, waarbij $L_b^{50} = \bigcup_{c \in C_b} L_{b,c}^{50}$ en $L_{b,c}^{50} \subseteq L$ de kleinste verzameling Y_c is zodat $\frac{\sum_{i \in \left(I_b^{48}(Y_c) \cup M_b^1\right) \cap I_b(c)} v_{i,b}^{T^0}}{\sum_{i \in I_b(c)} v_{i,b}^{T^0}} \geq 0.5$ mits deze bestaat, en $L_{b,c}^{50} = \left\{l \in L \mid I_b^{48}(l) \cap I_b^{48}(c) \neq \varnothing\right\}$ anders.

3. Voor productgroepen die nog steeds niet voldoen aan de 50%-regel, verzamelen we in het voorlopige basismandje de grootste (qua omzet) producten uit het EAN-domein die de laatste 12 weken uit het basisjaar beschikbaar waren. We putten net zolang uit deze lijst totdat de CPI-productgroep aan de 50%-eis voldoet, of totdat al deze producten zijn uitgeput. In formele termen:

$M_{b,3}^{T^0} = I_b^{50} \cap E_b$, met $I_b^{50} = \bigcup_{c \in C_b} I_{b,c}^{50}$ en $I_{b,c}^{50} \subseteq I_b^{12}$ de kleinste verzameling Z_c zodat $\frac{\sum_{i \in \left(Z_c \cup M_b^1 \cup M_b^2\right) \cap I_b(c)} v_{i,b}^{T^0}}{\sum_{i \in I_b(c)} v_{i,b}^{T^0}} \geq 0.5$ mits deze bestaat, en $I_{b,c}^{50} = I_b^{12} \cap I_b(c)$, anders.

B.2.4 Berekening van indexcijfers

Na selectie van de EAN-codes in het basismandje, worden per geselecteerde CPI-productgroep voor elke scannerdataberichtgever afzonderlijk, indexcijfers op basis van de scannerdata berekend. Hiertoe worden eerst gemiddelde transactieprijzen $\overline{p}_{i,b}^{T^0}$ per product *i* voor berichtgever *b* berekend in het basisjaar T^0 als:

$$\overline{p}_{i,b}^{T^0} = \frac{v_{i,b}^{T^0}}{q_{i,b}^{T^0}} \tag{10}$$

waarin $q_{i,b}^{T^0} = \sum_{t \in T^0} q_{i,b}^t$.

De volgende stap is het bepalen van productprijsindexcijfers voor maand t voor elk afzonderlijk product i uit het basismandje. Hiervoor wordt een transactieprijs $\overline{p}_{i,b}^t$ in maand t berekend volgens:

$$\overline{p}_{i,b}^{t} = \frac{v_{i,b}^{t}}{q_{i,b}^{t}} \tag{11}$$

waarbij $v_{i,b}^{t}$ de totale omzet is van product i bij berichtgever b in de eerste twee volledige weken van maand t (zie paragraaf 4.2.4). De verkochte hoeveelheid in die weken van maand t wordt weergegeven met $q_{i,b}^{t}$.

Een prijs die berekend wordt door omzetten te delen door verkochte hoeveelheden, wordt ook wel een *unit value* genoemd. Als productprijsindex wordt vervolgens de *unit value* index berekend:

$$U_{i,b}^{t,T^0} = \frac{\overline{p}_{i,b}^{t}}{\overline{p}_{i,b}^{T^0}} \tag{12}$$

Voor elke CPI-productgroep worden deze productprijsindexcijfers samengewogen met vaste gewichten uit het basisjaar tot een CPI-productgroepprijsindex[9]. De gewichten per product worden gedefinieerd als:

$$w_{i,b}^{T^0} = \frac{v_{i,b}^{T^0}}{\sum_{i \in M_b^{T^0}} v_{i,b}^{T^0}} \tag{13}$$

Voor berichtgever b en CPI-productgroep c ($c \in C_b$) wordt dan de productgroepprijsindex $\hat{P}_{b,c,S}^{t,T^0}$ in maand t ten opzichte van basisjaar T^0 op basis van de scannerdata berekend volgens:

$$\hat{P}_{b,c,S}^{t,T^0} = \sum_{i \in c} \left(\frac{w_{i,b}^{T^0}}{\sum_{i \in c} w_{i,b}^{T^0}} \right) U_{i,b}^{t,T^0} \tag{14}$$

B.2.5 Omgaan met producten die verdwijnen uit het assortiment

De werkwijze zoals hierboven beschreven werkt natuurlijk alleen als alle EAN-codes in het basismandje in elke verslagperiode in het assortiment van de supermarkten blijven. In de praktijk blijkt dat er vrijwel elke maand een aantal EAN-codes niet meer in de bestanden zijn terug te vinden. Dit kan verschillende oorzaken hebben. In sommige gevallen krijgt een product een andere EAN-code, bijvoorbeeld na restyling. In andere gevallen verandert een product wezenlijk, er komt bijvoorbeeld meer of juist minder in een verpakking, of de kwaliteit van het product verandert. Ook komt het voor dat supermarkten besluiten producten niet langer in hun assortiment aan te bieden. Een test op gegevens uit 2000 en 2001 heeft uitgewezen dat in 2001 ruim 20% van de EAN-codes in het basismandje dat was gebaseerd op het jaar 2000, verdween.

Analyse van producten die verdwijnen, heeft uitgewezen dat het in ruim de helft van de gevallen gaat om producten die een relatief klein omzetaandeel vertegenwoordigen. Daarom is gekozen voor een werkwijze waarbij alleen EAN-codes die een wegingsaandeel van meer dan 2,5% hebben binnen de CBL-groep waarin ze horen, verder onderzocht worden. Voor deze groep van EAN-codes worden opvolgers of vervangende EAN-codes aangewezen. Wanneer het producten betreft die zijn gewijzigd in kwaliteit of een andere verpakkingshoeveelheid hebben gekregen, dan wordt de prijsverandering hiervoor gecorrigeerd. In gevallen waarin het product uit het assortiment van de supermarkt is verdwenen, wordt er een nieuwe EAN-code opgevoerd van een vergelijkbaar product. Het is duidelijk dat in dat geval de prijs van het verdwenen product niet direct vergelijkbaar is met de prijs van het gekozen vervangende product. In die gevallen wordt dan ook op het moment van overgang op het nieuwe product gebruik gemaakt van de gemiddelde prijsverandering van andere producten in dezelfde CBL-groep.

Voor verdwijnende EAN-codes die slechts een klein omzetaandeel binnen de CBL-groep hebben, wordt tot aan de overgang op een nieuw mandje de gemiddelde prijsverandering van de andere producten in dezelfde CBL-groep gebruikt. Deze methode wordt ook toegepast wanneer een EAN-code slechts tijdelijk afwezig is, bijvoorbeeld omdat het product tijdelijk niet leverbaar is geweest. Dit wordt ook wel imputeren genoemd. Stel dat er in maand t g^t ontbrekende prijzen van producten $j_1,...,j_{g^t}$ zijn welke worden geïmputeerd. Dan wordt voor een ontbrekend product j_x een prijs $\hat{p}^t_{j_x,b}$ berekend door de prijs van dit product uit de vorige maand (merk op dat deze prijs zelf ook al geïmputeerd kan zijn) te vermenigvuldigen met het prijsverloop van alle producten i waarvoor wel een prijs beschikbaar is en die in dezelfde CBL-productgroep vallen. In formulevorm:

$$\hat{p}^t_{j_x,b} = \overline{p}^{t-1}_{j_x,b} \times \sum_{i \in l_{j_x},\, i \notin \left\{ j_1,\ldots j_{g^t} \right\}} \left(\frac{w^{T^0}_{i,b}}{\sum_{i \in l_{j_x},\, i \notin \left\{ j_1,\ldots j_{g^t} \right\}} w^{T^0}_{i,b}} \right) U^{t,t-1}_{i,b} \qquad (15)$$

waarin $U^{t,t-1}_{i,b} = \dfrac{\overline{p}^t_{i,b}}{\overline{p}^{t-1}_{i,b}}$

De geïmputeerde prijs voor product j_x, wordt vervolgens gebruikt om de unit value voor het betreffende product in maand t te berekenen (zie formule 12).

In die gevallen waarin een CBL-groep over meer dan een CPI-productgroep is verdeeld, wordt geïmputeerd met het prijsverloop van producten uit dat gedeelte van de CBL-groep dat in dezelfde CPI-productgroep valt als het product waarvoor een prijs moet worden geïmputeerd.

---

[9] Op dezelfde wijze worden overigens ook per CBL-productgroep indexcijfers samengesteld.

B.2.6 Samenvoegen van scannerdataindexcijfers en indexcijfers van overige verkoopkanalen

In de voorafgaande paragrafen is uiteengezet op welke wijze de productgroepprijsindexcijfers op niveau V voor de supermarkten waarvan scannerdata worden ontvangen en voor de overige verkoopkanalen worden berekend. De volgende stap in het rekenproces is het samenvoegen van deze gegevens tot één totaalindexcijfer per productgroep op niveau V.

Zoals beschreven in paragraaf 4.1 worden de indexcijfers per productgroep samengewogen met gewichten die variëren per productgroep. Deze gewichten representeren de marktaandelen van de verschillende supermarkten waarvan scannerdata beschikbaar zijn en de overige verkoopkanalen. Zij:

- $m_{c,b}$ het marktaandeel van supermarkt b ($b \in B^*$) voor CPI-productgroep c waarvoor scannerdata van berichtgever b zijn ingezet. Er geldt dat $m_{c,b} = 0$ voor productgroepen c waarvoor bij berichtgever b geen scannerdata worden gebruikt (dus wanneer $c \notin C_b$). Voor alle andere productgroepen (dus voor alle $c \in C_b$) geldt dat $m_{c,b} > 0$.

- $m_{c,O} = 1 - \sum_{b \in B^*} m_{c,b}$ het marktaandeel van de overige verkoopkanalen.

- $\hat{P}_{c,O+S}^{t,0}$ het totaal prijsindexcijfer in maand t ten opzichte van basisjaar 0 op niveau V voor productgroep c. Dit indexcijfer is gebaseerd op alle prijswaarnemingen die beschikbaar zijn, dus zowel de scannerdatagegevens als de overige waarnemingen.

Stel nu dat vanaf maand t^* indexcijfers worden bepaald inclusief het gebruik van scannerdata. Tot en met maand $t^* - 1$ worden de productgroepprijsindexcijfers $\hat{P}_c^{t,0}$ bepaald op de "traditionele" wijze, zoals beschreven in deel A van de Appendix (formule 4).

Voor alle CPI-productgroepen c wordt het totaalprijsindexcijfer dan vanaf maand t^* berekend volgens:

$$\hat{P}_{c,O+S}^{t,0} = m_{c,O} \hat{P}_{c,O}^{t,0} + \sum_{b \in B^*} m_{c,b} \hat{P}_{b,c,S}^{t,0} \qquad (16)$$

Merk op dat in formule (16) de CPI-productgroepprijsindexcijfers gebaseerd op de scannerdata, indexcijfers zijn ten opzichte van het basisjaar 0 van de CPI (dit basisjaar is op dit moment 1995). De deelprijsindexcijfers zoals berekend volgens formule (14), zijn echter indexcijfers ten opzichte van basisjaar T^0 (op dit moment is het basisjaar 2001). Het is dus nodig om de deelprijsindexcijfers gebaseerd op de scannerdata hetzelfde basisjaar te geven als dat van de totale CPI. Hiertoe wordt voor de maand voorafgaand aan de introductie van de scannerdata in de CPI, het

niveau van de tot dan toe gepubliceerde CPI-cijfers op productgroepniveau gebruikt om de scannerdatareeks met basisjaar 0 een juiste start te geven. In formule:

$$\hat{P}_{b,c,S}^{t^*-1,0} \equiv \hat{P}_{c}^{t^*-1,0} \tag{17}$$

Vanaf maand t^* worden de indexcijfers gebaseerd op scannerdata ten opzichte van basisjaar 0 berekend door telkens het indexcijfer van de vorige maand te vermenigvuldigen met het maand-op-maand-verloop dat kan worden afgeleid uit de scannerdataprijsindexcijfers ten opzichte van basisjaar T^0. In formule (vanaf maand t^*):

$$\hat{P}_{b,c,S}^{t,0} = \hat{P}_{b,c,S}^{t-1,0} \frac{\hat{P}_{b,c,S}^{t,T^0}}{\hat{P}_{b,c,S}^{t-1,T^0}} \tag{18}$$

Wanneer de totaalprijsindexcijfers op CPI-productgroepniveau V volgens formule (16) zijn berekend, volgt de berekening van de indexcijfers op de hogere niveaus, door aggregatie met vaste gewichten, zoals weergegeven in figuur 1.

B.2.7 Jaarlijks basisverleggen

Voor de scannerdata vindt jaarlijks een basisverlegging plaats. Dit betekent dat elk jaar de Laspeyresprijsindex op basis van de scannerdata voor elk van de winkelketens moet worden gekoppeld aan de indexcijferreeks op het nieuwe basisjaar. Deze basisverlegging zal jaarlijks in januari ingaan. Er dient dan een koppeling gelegd te worden tussen de indexcijferreeks van december en januari op CPI-productgroepniveau V. Hiervoor is het nodig het prijsverloop te kennen tussen december van jaar T en januari van jaar T+1. Dit prijsverloop zal worden gebaseerd op het basismandje uit jaar T (op basis waarvan indexcijfers over jaar T+1 worden samengesteld). Dit maand-op-maand prijsverloop van december op januari zal worden gebruikt om de indexcijfers op basis van de scannerdata te berekenen (zie formule 18).