

Åsa Bilius, Olivia Ståhl, Can Tongur

Alternativa beräkningsmetoder för kassaregisterdata: kaffe-index

För diskussion

I takt med ökad användning av kassaregisterdata i svenska KPI fortsätter Prisenheten utredningen av alternativa indexmetoder framtagna för kassaregisterdata. I denna PM exemplifieras tre olika typer av metoder genom en tillämpning på kaffepriser. Inlägget utgör en sammanfattning av ett av Prisenhetens bidrag till UNECE-konferensen i Genève 2018.

Innehåll

1	BAKGRUND	2
2	SYFTE	3
3	DATAMATERIALET	3
4	JÄMFÖRANDE STUDIE MELLAN ALTERNATIVA METODER	5
4.1	STATISK ANSATS: DIREKT INDEX BASERAT PÅ EN FAST KORG	5
4.2	DYNAMISK ANSATS: MÅNADSKEDJNING AV PARVIS MATCHANDE ARTIKLAR	7
4.3	MULTILATERAL ANSATS: GEARY-KHAMIS INDEX.....	98
5	DISKUSSION	11
	KÄLLOR	12



1 Bakgrund

Kassaregisterdata används i KPI sedan 2012 då det introducerades för delar av dagligvaruurlvalet. Idag används det i undersökningar som sammantaget utgör cirka 30 procent av KPI. Kassaregisterdata har många fördelar jämfört med exempelvis intervjuarinsamlad eller webbskrapad data eller data insamlad via enkäter till företagen. Till skillnad från de övriga indatatyperna så innehåller kassaregistermaterialet *faktiska transaktionspriser* och försäljningskvantiteter. En övergång till kassaregisterdata kan dessutom vara ekonomiskt fördelaktigt för SCB samt i vissa fall minska uppgiftslämnarbördan hos företagen. Att söka efter kassaregisterdata inom fler produktområden är därför en prioriterad uppgift för SCB framöver.

I och med införandet av kassaregisterdata för dagligvaror gjordes inte några principiella metodändringar vad gäller indexberäkningen. Index per produktgrupp beräknas, precis som inom övriga delar av KPI, baserat på en fast korg av produkter med bas i december föregående år. Urvalet av produkter är ett sannolikhetsurval där produkternas urvalssannolikheter är proportionella mot deras totala försäljning under närmast föregående helår. Utvalda produkter följs därefter hela året, eller så länge de bedöms som ”tillräckligt representativa”; Vid påtaglig minskning i försäljning analyseras materialet för att identifiera bästa tänkbara ersättare.

En begränsning med dagens metod är att dynamiken på marknaden inte adresseras fullt ut. Produkter adderas till korgen enbart då de utgör ersättare till utgående produkter. Det innebär att helt nya produkter introduceras först i och med den årliga uppdateringen av urvalet. Risken är att urvalets representativitet sett i relation till målpopulationen minskar gradvis under året.

Alternativa indexmetoder för kassaregisterdata frångår idén om en fast korg och baserar istället indexberäkningarna på de produkter som finns tillgängliga under de aktuella månaderna. På så sätt säkerställs urvalets representativitet vid varje enskild tidpunkt, och utmaningen ligger istället i att uppnå prisutvecklingar som är jämförbara över tiden. Risken finns att faktiska prisförändringar missas när dessa sker samtidigt med ändringar i produkternas artikelnummer och därför inte länkas samman på ett korrekt sätt.

Enheten för prisstatistik har sedan en tid tillbaka arbetat med olika analyser kring indexformuleringen för kassaregisterdata (se exempelvis Sammar, 2011; Norberg m.fl. 2012; Jonsson och Jansen, 2015; Tongur och Sandén 2016 och Bilius och Tongur, 2017). Analyserna har främst utgått från dagligvaror även om kassaregisterdata, som redan nämnts, används inom fler områden av KPI. Även i denna PM används data från dagligvaruområdet för att exemplifiera. Tanken är emellertid att bygga erfarenheter och kunskaper som kan vara till nytta också inom andra delar av KPI, då mängden kassaregisterdata väntas bli större framöver. Liknande analysarbete sker också internationellt i numera ganska många andra länder.

2 Syfte

I denna PM redovisas en mindre studie där dagens indexmetod inom dagligvaror jämförs mot ett antal (mer eller mindre) tänkbara alternativ. Det underlag som används för jämförelsen är begränsat till priser på kaffeprodukter under 2017. Syftet med analysen har varit att öka förståelsen för hur kassaregisterdata kan hanteras och redan hanteras i vissa andra länder idag.

Denna PM utgör en rapportering från ett pågående arbete. Prisenheten ämnar återkomma längre fram med mer fullständig analys, om det bedöms relevant. Nämnden inbjuds redan nu att komma med synpunkter på utformningen av Prisenhetens fortsatta arbete med analyser av kassaregisterdata. Enhetens avsikt är på längre sikt att ha en fullständig orientering kring vilka metoder som bör användas i KPI i de fall kassaregisterdata finns tillgängligt.

3 Datamaterialet

Det kassaregisterunderlag som SCB använder för dagligvaror innehåller veckovis total omsättning och antal sålda enheter per artikelnummer (EAN-kod¹) över samtliga transaktioner per butik. I materialet finns även kedjornas egna gruppindelningar av produkter, vars detaljeringsnivå varierar mellan kedjor.

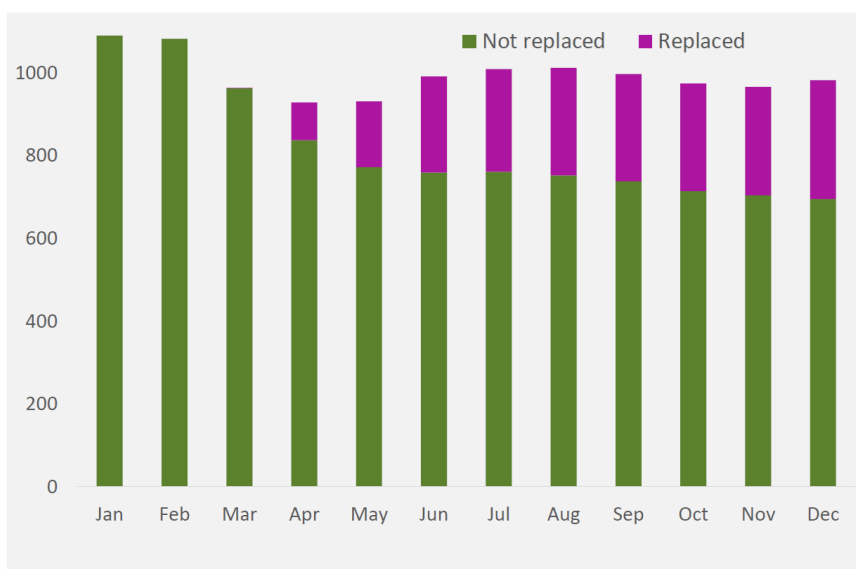
För denna analys används data från ett begränsat urval av butiker från det fullständiga kassaregistermaterialet. Resultaten är därför inte jämförbara med publicerade indextal. För de utvalda butikerna används all data som kunnat identifieras som kaffe. När vi fortsättningsvis pratar om ”hela materialet” syftar vi alltså på samtliga produkter från dessa utvalda butiker. I de fall där den data som används utgör ett *urval* så har samma butiker använts som för det större materialet, medan valet av produkter inom dessa gjorts med hjälp av sannolikhetsurval.

Anledningen till att produktgruppen Kaffe valdes för denna preliminära studie är att många manuella byten gjordes inom den under 2017 (jmf figur 1). Skälet till bytena var i många fall simultana ändringar i EAN-kod och förpackningsstorlek. Produkten ansågs därför kunna vara en lämplig kandidat för att visa på likheter och skillnader mellan olika beräkningsmetoder för den aktuella perioden.

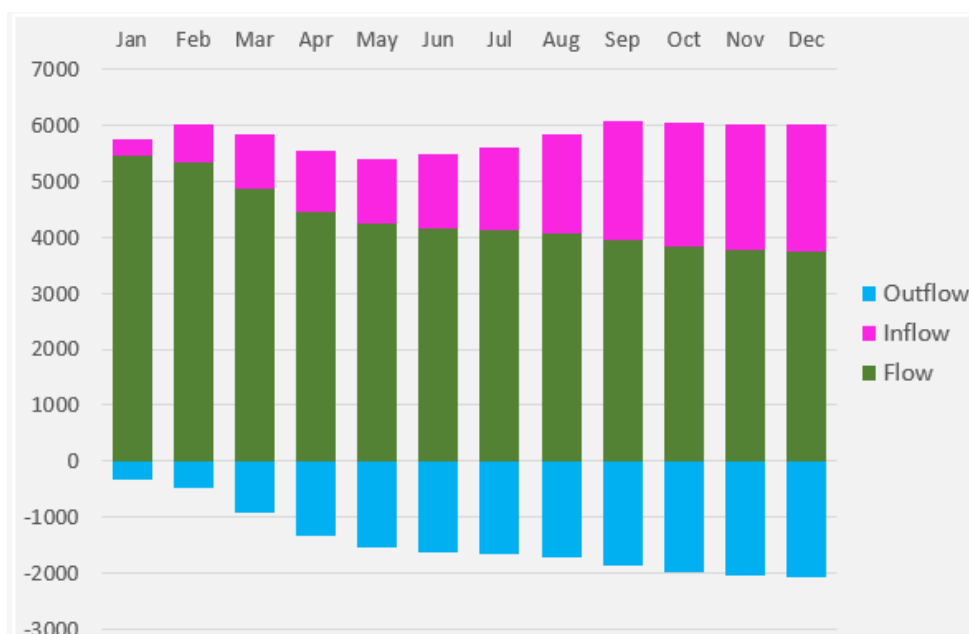
Graden av ”genomströmning” för kaffeprodukter i hela materialet presenteras i figur 2. Genomströmning (*flow*) definieras här som det antal artikelnummer i en viss butik som säljs både under den aktuella perioden och under basperioden. Inflödet (*inflow*) utgörs i sin tur av antalet sålda produkter under den specifika månaden som inte fanns i basperioden, medan utflöde (*outflow*) är antal försvunna produkter relativt basperioden. Av de ca 5500

¹ Numera används oftast beteckningen GTIN internationellt, men i de data SCB får från kedjorna används fortfarande den tidigare benämningen EAN.

”produkterbjudanden” (ett visst artikelnummer i en viss butik) som säljs i basperioden är det knappt 4000 som fortfarande finns till försäljning i december.



Figur 1: Antal manuella ersättningar av produkterbjudanden (viss produkt i viss butik) inom produktgruppen Kaffe under 2017. *Replaced* innebär att produkten ersatts med ett nytt artikelnummer och *Not replaced* att artikelnumret är oförändrat jämfört med basmånaden.



Figur 2: Genomströmningen av produkter under 2017 för hela materialet. *Flow* betecknar det antal artikelnummer som säljs både under den aktuella perioden och under basperioden, *Inflow* betecknar antalet sålda produkter under den specifika månaden vilka inte fanns i basperioden, *Outflow* betecknar antalet försvunna produkter relativt basperioden.

4 Jämförande studie mellan alternativa metoder

Med utgångspunkt i kassaregistermaterialet för kaffe har resultatet av ett antal olika indexmetoder jämförts. Metoderna representerar tre klasser av index; Index baserat på en fast korg (även kallat *statisk ansats*), månadskedjning (*dynamisk ansats*) och den *multilaterala ansatsen*. Den statiska ansatsen används för dagligvaror i Sveriges KPI idag medan den dynamiska ansatsen är relativt vanligt förekommande internationellt och bland annat används i Norge och Belgien. Den multilaterala ansatsen har rönt stort internationellt intresse under de senaste åren och har redan implementerats för dagligvaror i bland annat Nederländerna. Det bör dock poängteras att de metoder som prövas i denna PM inte direkt överensstämmer med de som används i andra länder idag, utan utgör mer eller mindre förenklade varianter av dessa.

I det följande redogörs för de olika metoder som prövats samt för resultatet. I avsnitt 5 illustreras de tre typerna av metoder mot varandra. Vissa preliminära slutsatser inför det fortsatta arbetet dras även.

4.1 Statisk ansats: Direkt index baserat på en fast korg

Direkt index baserat på en fast korg är en etablerad metod och används i KPI idag såväl för kassaregisterdata som för merparten av övriga elementäraggregat. Index beräknas som ett viktat geometriskt medelvärde:

$$I(FK)_0^{t,m} = \prod_{i=1}^n \left(\frac{p_i^{t,m}}{p_i^0} \right)^{w_i} \quad (1)$$

där n är antalet produkterbjudanden (en viss produkt i en viss butik) i korgen, och w_i är en vikt kopplad till det aktuella produkterbjudandet. När $w_i = \frac{1}{n}$ så benämns (1) *Jevons index*. Basperioden (december föregående år) betecknas 0 och innevarande period betecknas t,m .

I studien har fastkorgsindex enligt (1) tillämpats i fyra varianter:

(a) Nuvarande metod (baserat på ett urval, inklusive byten)

Korgen omfattar ca 1000 produkterbjudanden, valda med πps . Vikterna, w_i , är proportionella mot omsättningen år $t-2$ multiplicerat med de inverterade inklusions-sannolikheterna, vilket i många fall innebär att produkterna är självvägande. Byten genomförs kontinuerligt av personal på prisenheten.

(b) Metod baserad på urval, men utan byten

Samma produkter och vikter som i (a) används, men här utan att de manuella byten som görs under året räknats med. Produkter finns alltså med i korgen så länge de finns till försäljning och försvinner sedan helt.

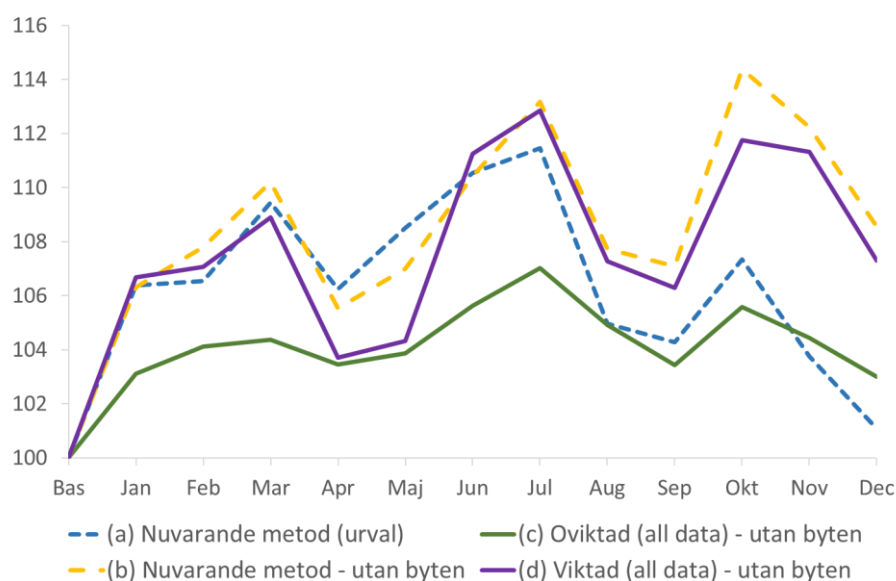
(c) Index baserat på hela materialet, oviktat

Korgen omfattar samtliga ca 5500 "produkterbjudanden" som påträffas i basperioden. Inga vikter används. Precis som i (b) så görs inga byten.

(d) Index baserat på hela materialet, med årsvikter

Korgen omfattar samma produktbjudande som metod (c). Vikterna är proportionella mot årlig omsättning innevarande år, d.v.s. 2017. Precis som i (b) och (c) görs inga byten.

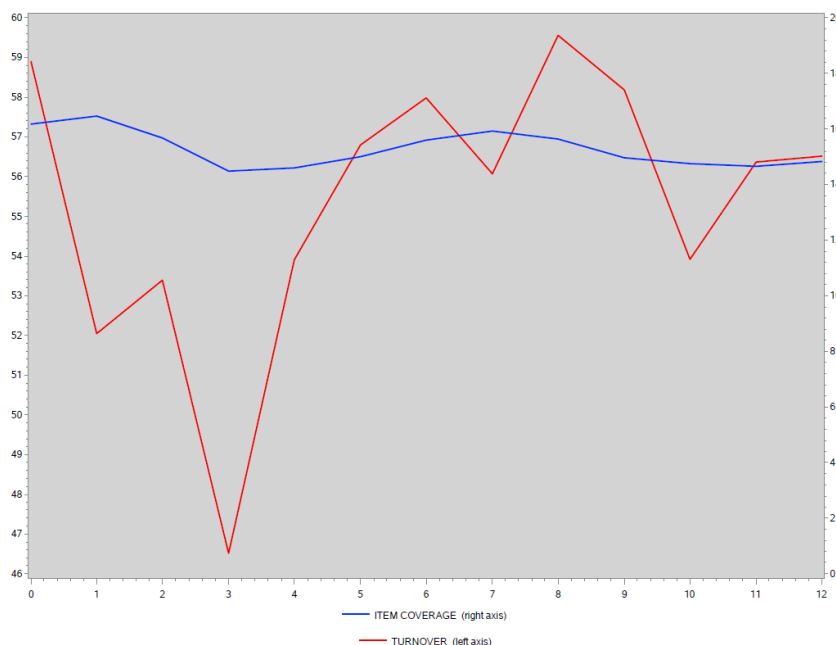
Resultatet för den statistiska ansatsen redovisas i figur 3. Under början av året betar sig nuvarande metod (a) på ungefär samma sätt som det viktade indexet som använder sig av hela datamaterialet (d), men alltefter året går så ger den ett allt lägre index. De manuella ersättningar som görs verkar ha en stor effekt på index och leder till att index sjunker i slutet av året, vilket kan ses genom jämförelse mellan (a) och (b).



Figur 3: Fastkorgsindex för kaffepriser under 2017.

Vi har även undersökt den nuvarande korgens (den som används i metod a ovan) täckning i relation till hela datamaterialet för de enskilda månaderna. Resultatet presenteras i figur 4. Den röda linjen avser här en jämförelse i termer av omsättning (oviktad) och den blå i termer av antal produktbjudanden. Sett över 2017 ingår ungefär 15% av antalet produktbjudanden i urvalet medan andelen ingående omsättning varierar mellan 47% och 59%.

Preliminära analyser av mikrodata tyder på att flera av de produkter som manuellt bytts in under året har haft en nedåtgående prisutveckling under andra halvåret. Detta kan förklara varför metod (a) ger lägre årsförändring i december än övriga tre alternativ vilka samtliga bygger på en ”krympande korg”.



Figur 4: Korgens täckning i procent, i termer av omsättning (röda linjen, vänstra axeln) och antal produkterbjudanden (blå linjen, högra axeln).

4.2 Dynamisk ansats: Månadskedjning av parvis matchande artiklar

En vanlig metod för kassaregisterdata internationellt är kedjning av index baserat på parvis matchande artiklar månad för månad. Index mot basperioden beräknas indirekt där varje länk baseras enbart på produkter som finns tillgängliga i de två intilliggande månaderna. Vanligtvis används någon form av tröskelvärde avseende omsättningen i de båda perioderna för att undvika alltför stor volatilitet i korgen, samt imputeringar för ytterligare stabilitet för bland annat säsongprodukter. Index i den dynamiska ansatsen beräknas enligt

$$I(MK)_0^{t,m} = \prod_{k=1}^m I_{k-1}^k \quad (2)$$

där I_{k-1}^k betecknar de månatliga indexlänkarna. För dessa används vanligen geometriska index:

$$I_{k-1}^k = \prod_{i=1}^{n_{(k,k-1)}} \left(\frac{p_i^{t,m}}{p_i^{t,m-1}} \right)^{w_{k,i}} \quad (3)$$

där $n_{(k,k-1)}$ betecknar det antal produkter som finns både månad k och månad $k-1$ år t , och $w_{k,i}$ är en vikt kopplad till produkterbjudande i och perioden $(k-1,k)$. I bland annat Norge används oviktade månadsindex, vilket ger $w_{k,i} = \left(\frac{1}{n_{(k,k-1)}} \right)$ i (3). Användningen av oviktade månadsindex är även konsistent med beskrivningen i Eurostat (2017), vilken innehåller praktisk vägledning för användandet av "supermarket scanner data".

I den preliminära studien har fyra olika typer av dynamiska index beräknats i enlighet med (2) och (3). I två av fallen beräknas index oviktat, medan övriga två scenarier innehåller vikter som antingen är konstanta över året eller varierar för varje månad. De fyra olika indexformuleringarna kan sammanfattas enligt följande:

(a) Oviktat – utan cutoff

Index beräknas oviktat. Inget tröskelvärde (cutoff) tillämpas.

(b) Oviktat - med cutoff

Samma beräkning som i (a) förutom att mängden artiklar som ligger till grund för månadsindex begränsats till de som har störst omsättning: En artikel måste stå för minst 1% av omsättningen i såväl period $k-1$ som k för att tas med i månadsindexet.

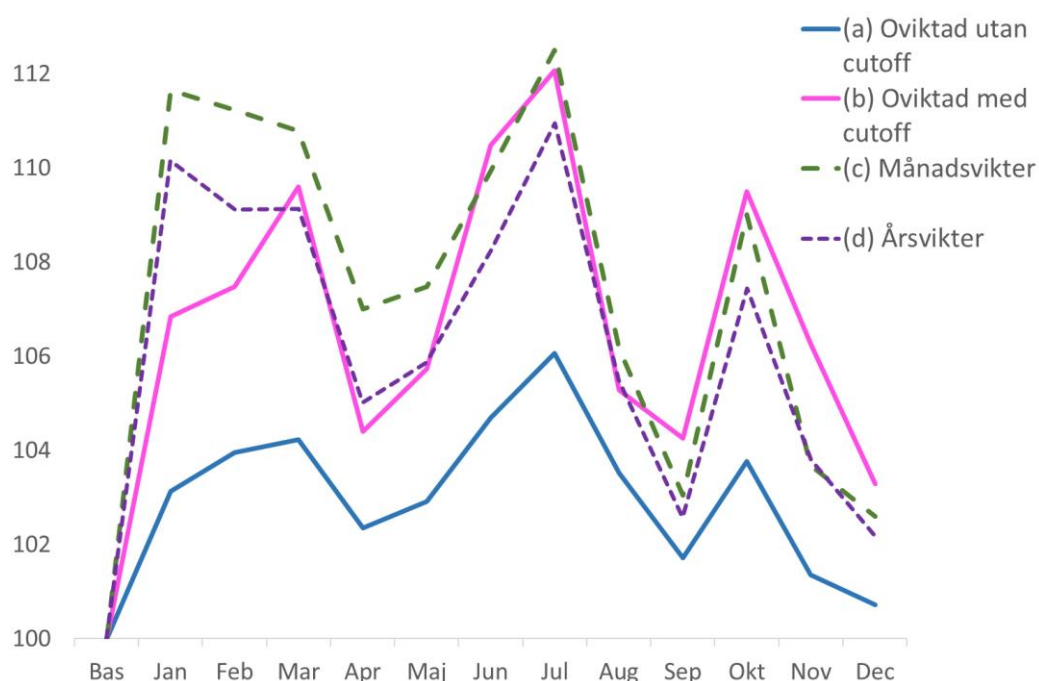
(c) Månatliga vikter

Vikterna varierar över året utifrån omsättningen under de två aktuella månaderna ($k-1$ och k). Inget tröskelvärde (cutoff) tillämpas.

(d) Årsvikter

Vikterna är konstanta över året och baseras på total omsättning under 2017. Inget tröskelvärde (cutoff) tillämpas.

Resultaten illustreras i figur 5. Införandet av ett tröskelvärde har en liknande effekt på index som att införa vikter.



Figur 5: Dynamisk ansats (månadskedjning av matchande artiklar).

4.3 Multilateral ansats: Geary-Khamis Index

En nackdel med att använda sig av vikter som är konstanta över året är att produkter som säljs i väldigt liten utsträckning under en viss del av året ändå kan ha stor effekt på index. Det är dock väl känt att vikter som varierar över året kan ge upphov till kedjedrift när den dynamiska metoden används. En ansats som undviker kedjedriften men samtidigt tillåter varierande vikter är den *multilateral* ansatsen, vilken nu även verkar rekommenderas i den nya versionen av den internationella KPI-manualen² (ILO, 2018). Vi har därför även inkluderat en multilateral indexmetod i studien och har här valt en förenklad variant av den som används av den nederländska statistikbyrån³ nämligen *Geary-Khamis* (GK) index översatt till jämförelser över tiden. Index kan i detta fall skrivas

$$I(GK)_0^{t,m} = \frac{\sum_{i \in (t,m)} P_i^{t,m} Q_i^{t,m} / \sum_{i \in (t,m)} (Q_i^{t,m} \cdot v_i)}{\sum_{i \in 0} P_i^0 Q_i^0 / \sum_{i \in 0} (Q_i^0 \cdot v_i)} \quad (4)$$

där v_i betecknar så kallade ”kvalitetsjusteringsfaktorer” vilka kan tolkas som nytta/kvaliteten hos produkt i i relation till övriga produkter. Notera att täljaren i (4) innehåller en typ av ”kvalitetsjusterat enhetsvärde” över alla produkter som finns i korgen vid tidpunkt (t,m) , medan nämnaren utgörs av samma sak för de produkter som fanns i korgen vid bas tidpunkten 0.

Kvalitetsjusteringsfaktorerna skattas i sin tur utifrån ett antagande om att de relativa snittpriserna sett över en längre tidsperiod utgör ett mått på den relativa nyttan som konsumenterna får ut av de enskilda produkterna:

$$v_i = \frac{\sum_{z \in T} \frac{P_i^z Q_i^z}{I_0^z}}{\sum_{z \in T} Q_i^z} \quad (5)$$

Även (5) är alltså en typ av enhetsvärde, men denna gång sett över alla tidsperioder, z , som ingår i T . I denna preliminära studie har vi använt oss av hela 2017 som vårt T i formel (5).

Vi har testberäknat flera olika varianter av GK enligt (4) och (5), varav två redovisas här. I den första varianten används artikelnumret (EAN-koden) som identifikator för produkterna, d.v.s. P_i och Q_i i formlerna ovan avser pris och kvantitet kopplat till en viss EAN-kod. I den andra varianten sker först en aggregering upp till en lägre nivå, kallad ”EAN-grupp”, med

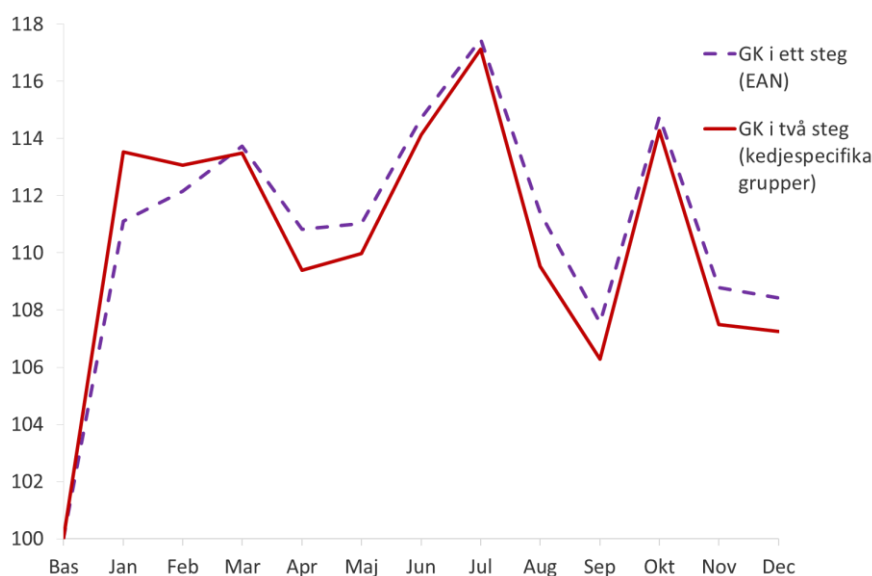
² Följande citat är hämtat från ILO (2018), kapitel 10. På sid.11 står att läsa: “When using a census of products, hence without sampling, a weighted index number formula should be used. [...] Product turnover poses a significant problem. To maximize the number of matches in the data, chaining at high frequency will be needed. This, however, can lead to significant drift in the index. Multilateral price index number methods, which are drift-free by construction, are most suitable to handle a census of products from scanner data.” På sid. 20 står vidare: “Only a handful of NSOs have actually implemented multilateral price indexes in the CPI, displaying caution in implementation and a divergence in methods and practices”.

³ Förenklingen består främst i att samtliga perioder använts för att skatta kvalitetsjusteringsfaktorerna (jmf formel 6), något som inte vore möjligt i praktiken annat än för reviderade indextal.

hjälp av ett vanligt enhetsvärde. Index beräknas sedan med dessa enhetsvärden som priser. I formlerna (4) och (5) avser med andra ord P_i och Q_i istället pris och kvantitet för en viss *grupp* av EAN-koder. Tanken med denna två-steps-metod är att index på ett bättre sätt ska kunna ta hänsyn till prisändringar som sker samtidigt med ändringar i artikelnumren (se exempelvis Chessa, 2016).

EAN-grupperna bör formas så detaljerat att de kan betraktas som homogena, men samtidigt tillräckligt breda för att prisändringar som sker över tiden ska kunna plockas upp. Vi har i denna studie skapat grupperna utifrån tre dimensioner: Storlek, märke och ”typ av kaffe”. Storleksdimensionen innehåller fem kategorier: ”0-99 g”, ”100-399 g”, ”400-749 g”, ”>750 g” respektive ”Styck-förpackningar”. För dimensionen ”typ av kaffe” användes de gruppindelningar som redan fanns väldefinierade i datamaterialet.⁴ I samtliga fall har vi även räknat om alla priser och kvantiteter till kilo- eller styckpriser/kvantiteter.

Resultaten för de olika GK-indexen återfinns i figur 6. Man kan se en skillnad mellan index på EAN-nivå och två-steps-metoden i bland annat januari och april. Närmare analys av mikrodata tyder emellertid på att denna skillnad främst beror på att de använda EAN-grupperna inte är tillräckligt homogena. Effekterna av simultana ändringar i artikelnummer och kvantitet verkar med andra ord inte plockas upp av två-steps-metoden enligt denna preliminära utformning. En detaljstudie av ett par sådana ändringar som skedde i april till juni indikerar att för dessa var relationen mellan en- och två-steps-metoderna snarare den motsatta, vilket också är mer i linje med vad man hade kunnat vänta sig.

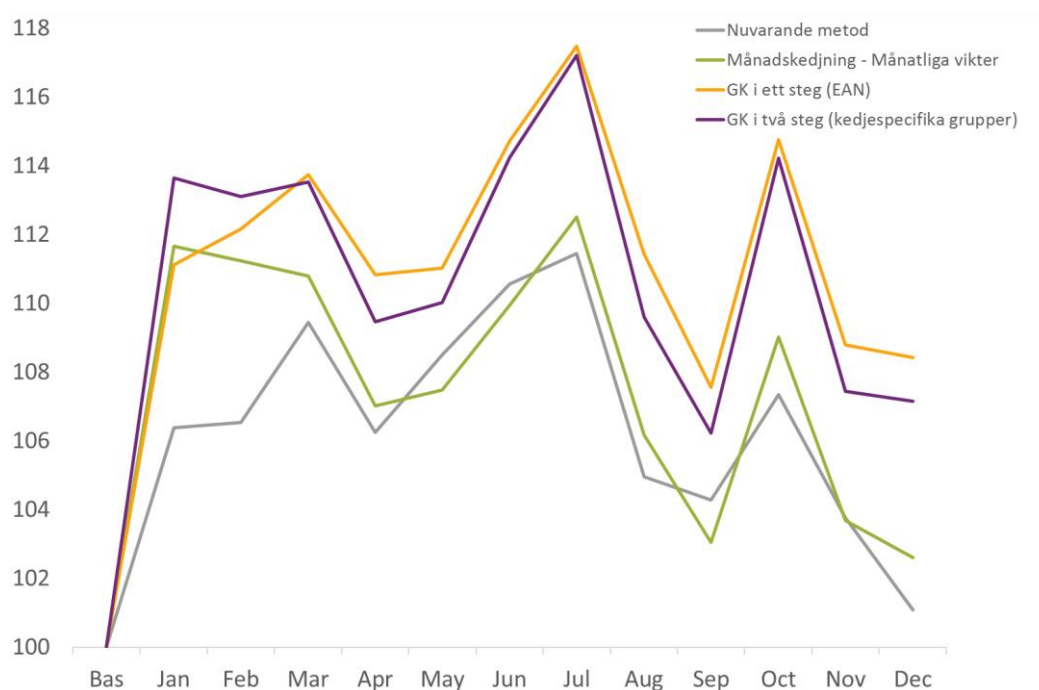


Figur 6: Geary-Khamis index beräknat i ett respektive två steg.

⁴ Ett antal andra grupperingar prövades även i analysen men visade sig ge väldigt likartade resultat. Ingen av grupperingarna visade sig dock vara detaljerad nog för att kunna plocka upp de simultana ändringar i artikelnummer och kvantitet som skedde under 2017 på ett bra sätt, då ”mixproblemen” var alltför stora (grupperna inte tillräckligt homogena).

5 Diskussion

I figur 7 jämförs resultatet av den statistiska, den dynamiska och den multilaterala metoden, representerade av dagens metod, det kedjade indexet som använder sig av månadsvikter samt GK i ett och två steg. I januari är skillnaden som störst mellan nuvarande metod och de övriga, vilka samtliga använder sig av vikter som varierar över året. Detta beror sannolikt på att kaffeförsäljningen i december skiljer sig något i karaktär från övriga månader (andra typer av kaffe säljs och mycket rea förekommer). Skälen till att de två multilaterala indexen skiljer sig så markant i nivå från de övriga under större delen av året är inte helt utredda ännu, men en preliminär hypotes är att det beror på att de storsäljande produkter som ökat i pris mellan februari och mars minskar i pris senare under året.



Figur 7: Jämförelse mellan statistik metod (fast korg), dynamisk metod (månadskedjning) och multilateral metod (Geary-Khamis index).

Som redan nämnts är detta en rapport från ett pågående arbete som kommer att utvecklas framöver. Det är dock redan nu uppenbart att de grupper som här använts för två-stegs-metoden inte är tillräckligt homogena, vilket kommer att beaktas i den fortsatta analysen. Det vore även intressant att inkludera effekten av gruppindelningar på den dynamiska metoden. I övrigt är den fortsatta analysen tänkt att i ett första skede inriktas på att hitta de faktorer i mikrodata som ger skillnader mellan olika indexmetoder. Övriga faktorer som skulle kunna utredas framöver är användandet av imputeringar inom ramen för den dynamiska metoden samt val av "fönsterlängd" för den multilaterala metoden. Nämnden inbjuds att komma med synpunkter på utformningen av det fortsatta arbetet.

Källor

Bilius, Å. och Tongur, C. (2017). *Empirisk jämförelse av fast varukorg visavi månadsmatchning för dagligvaror*. PM till möte med nämnden för konsumentprisindex, möte 3, 2017-10-19.

Chessa, A. (2016). *A new methodology for processing scanner data in the Dutch CPI*. EURONA — Eurostat Review on National Accounts and Macroeconomic Indicators.

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data. HICP Guideline tillgänglig via <https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf>.

ILO (2018). Consumer Price Index Manual (draft version), chapter 10. *Draft Chapter for comments*, tillgängligt via <https://www.imf.org/en/Data/Statistics/cpi-manual>.

Jonsson, F. och Jansen, E. (2015). *Användning av kassaregisterdata från Systembolaget för egen uträkning av index*. PM till möte med nämnden för konsumentprisindex, möte 252, 2015-05-18.

Norberg, A., m.fl. (2012). *Scanner data – comparability issues*. PM till möte med nämnden för konsumentprisindex, möte 245, 2012-04-19.

Sammar, M. (2011). *Kassaregisterdata i KPI*. PM till möte med nämnden för konsumentprisindex, möte 244, 2011-10-19.

Tongur, C. och Sandén, B. (2016). *Viktvaror från kassaregisterdata*. PM till möte med nämnden för konsumentprisindex, möte 1, 2016-10-17.