ESA/NUP/KP
Sofie Öhman
Can Tongur

PM till nämnden för KPI
Sammanträde nr 15
2022-10-13

Fortsatt arbete med kassaregisterdata

SCB arbetar vidare med att inhämta och utreda nya potentiella datakällor för Konsumentprisindex, särskilt kassaregisterdata som möjligt substitut till insamling på fält och internet. I promemorian redogörs för några relaterade frågor kring de återstående utmaningar som följer av den praktiska användningen av ytterligare och större datamängder.

1. Introduktion

Arbetet med att tillföra nya kassaregisterdata till KPI fortgår. SCB eftersträvar att fånga ytterligare datakällor, både för KPI och i bredare mening för statistikproduktionen. Under 2021 och 2022 har flera nya datakällor blivit tillgängliga för KPI varpå förutsättningar har skapats för återkoppling till nämnden genom denna promemoria. Från arbetet har väckts frågan om principerna för det fortsatta arbetet med nya datakällor.

2. Arbetet med datafångst

2.1 Nya datakällor från fler företag

Under 2021 och 2022 har SCB haft möten med ytterligare detaljhandelsföretag för att undersöka möjligheten att få in nya kassaregisterdata, till KPI och andra undersökningar på SCB. Flera företag har haft möjlighet att lämna denna typ av data och analyser har påbörjats för användning i KPI.

För några företag har förutsättningarna inte funnits att leverera den typ av kassaregisterdata som SCB efterfrågat – det har framkommit under möten med företagen att de har varierande tekniska och resursmässiga förutsättningar att framställa och leverera data. Den data SCB efterfrågar är inte alltid samlad på ett ställe varför de kan behöva kombinera hämtningar från olika databaser eller källor. Det är även vanligt förekommande att det krävs flera typer av kompetenser och ibland även externt IT-stöd för att

kunna ta fram den information SCB efterfrågar för just KPI:s syfte. Därför kan dialogen med företagen kring deras kassaregisterdata med återkopplingar innebära en process som tar tid.

2.2 Uppgiftslämnandet för nya datakällor

Även då detaljhandelsföretagen omfattas av uppgiftslämnarplikt står det inget i nuvarande föreskrift (SCB 2016) om att uppgifter ska lämnas genom just kassaregisterdata. I de förekommande fallen där företag lämnar uppgifter i form av kassaregisterdata har detta skett genom frivilliga överenskommelser.

Uppgifter från företagen från vilka SCB efterfrågar kassaregisterdata samlas annars in genom manuell webbinsamling, butiksbesök eller webbskrapning, vilket innebär att uppgiftslämnarbördan i praktiken tillkommer/ökar för de berörda företagen. Arbete pågår med att ta fram en ny föreskrift som innebär att uppgifterna ska lämnas till SCB i form av kassaregisterdata där det är tillämpligt.

3. Datamaterial

3.1 Struktur

Som ett resultat av en centraliserad dataavdelning har SCB strävat efter en generisk process vid inhämtande av nya kassaregisterdata enligt en för SCB mer homogen struktur kring insamlade variabler, för effektivitet och samanvändbarhet. Typiska variabler för denna upprättade struktur är bland andra (se även annex för fullständig specifikation):

- Transaktionsdatum
- Artikelnummer
- Omsättning (antal, kronor)
- Produktkategorisering (i olika nivåer)

Den homogena strukturen är framför allt anpassad efter förutsättningarna i data hos de detaljhandelsföretag som SCB haft kontakter med de två senaste åren. Erfarenheterna från befintliga datakällor i KPI har ibland visat att en sådan homogen struktur inte motsvaras av uppgiftslämnarnas olika förutsättningar.

För hittillsvarande implementeringar i KPI har kassaregisterdata haft varierande strukturer, från fall till fall, även om dessa i stort kan sammanfattas enligt den framtagna mallen, särskilt för vissa typer av produkter.

3.2 Klassificering

För de utredda nya datamängderna har förekommit ett omfattande spann i sortimenten med en variation mellan 10 000 och 50 000 unika artiklar. Det kan medföra ett betydande klassificeringsarbete för urval och implementering.

Det har för utredningen inte gjorts klassificeringar av de nya kassaregisterdata. Detta kan dock behöva betänkas i vilken utsträckning det är motiverbart i termer av arbetsintensiv kodning för nya datamängder.

3.3 Återkommande utredningsfrågor

Under utredningen av de nya kassaregisterdata har några aspekter återkommit. Dessa kan delas upp i två olika fall, dels generella aspekter som innefattar frågor rörande innehållet i data, dels övriga frågor som uppenbarats i jämförelserna med befintlig insamling av motsvarande prisuppgifter.

3.3.1 Generella aspekter

- Returer
    - När konsumenter returnerar köpta produkter innebär det en negativ omsättning/kvantitet och för kassaregister kan det medföra olika tidsperioder för köp/återköp. I fall där kvantiteter är låga kan detta ha en betydande effekt på det aktuella priset
- Decimalfel
    - Styckpriser som sporadiskt avviker med en faktor 10 mot övriga priser för samma artikel. Dessa misstänks vara s.k. potensfel (kommatecken som har hamnat fel)
- Motsägelsefulla kombinationer (omsättning mot kvantitet)
    - Positiv omsättning i kombination med negativ kvantitet, alternativt negativ omsättning i kombination med positiv kvantitet
- Interna variationer mellan kategorier
    - Samma artikelnummer kan förekomma i olika produktkategorier
    - Olika antal nivåer på kategorier
- Kategoriseringsolikheter mellan företag med liknande produkter
    - Olika typer av hierarkier och kategorier, asymmetrier även inom företag givet olika detaljeringsbehov för produkterna
- Relation moms/ej moms
    - I vissa fall kan momsbeloppen inte bestämmas baklänges till förväntat värde (25% eller 12%) i relationen inklusive/exklusive moms
- Diskrepanser vid olika uttag av samma period

    - I förekommande fall med flera typer av uttag, exempelvis månad respektive veckor för samma period har det återkommande noterats diskrepanser, vilket indikerar en betydelse av tidpunkten för uttaget
- Otydlighet i angivelsen av förpackningsstorlek/nettoinnehåll
    - Det kan vara oklarheter i enheten/storleken för vissa artiklar vilket kan orsaka problem vid byten

3.3.2 Jämförande aspekter

För en del av de nya kassaregisterdata görs nuvarande insamling till KPI genom webbskrapning, vilket har möjliggjort jämförelser mellan de två datakällorna. Jämförelsen begränsas till de fall där motsvarande försäljning funnits i kassaregisterdata, identifierat utifrån samma artikelnummer och även dag. Typiska aspekter är:

- olika variationsmönster
- olika indexutfall
- olika enheter.[1]

Därtill har tillgängligheten av två synkrona datakällor bidragit till större kunskaper kring de befintliga urvalens beskaffenhet inom webbskrapning, framför allt i termer av försäljningsvolymer vilket kan ligga till grund för kommande urval (redovisas inte mer specifikt här).

4. Användning av nya datakällor

4.1 Metoder för användning av kassaregisterdata

Jämförelserna av de nya datakällorna mot webbskrapade data gjordes på premissen att de webbskrapade produkterna hade identifierats via hemsidor. Analyserna är därför inte kompletta vad avser företagens hela sortiment eller för alternativa korgar annat än den befintliga.

[1] För ett antal produkter är det uppenbart att enhet skiljer sig mellan datamaterialen. Det är i sig inget problem då intressevariabeln är prisutveckling. Svårigheter kan dock uppstå vid produktersättningar när det inte går att härleda vilken enhet som avses i kassaregisterdata.

Det väcker åter frågan om effektiv användning av kassaregisterdata i termer av variablerna pris, kvantitet liksom klassificeringen av stora datamängder. SCB identifierade i början av arbetet med kassaregisterdata (Norberg, Sammar och Tongur 2011) följande fyra typiska användningsfall:

1) Ersätt enbart manuellt insamlade priser med kassaregisterdata
2) Använd kassaregisterdata som hjälpinformation
3) Använd allt kassaregisterdata för indexberäkningar
4) Använd kassaregisterdata för kvalitetsändamål så som granskning

En snarlik diskussion återfinns i nya KPI-manualen (§10.28 ILO 2020).

För kassaregisterdata har det skett omfattande metodutveckling genom så kallade multilaterala metoder, vilket även SCB utreder för dagligvaror inom ramen för ett s.k. Eurostat Grants-projekt. Fullskalig användning (alternativ 3 ovan) kan innebära att både pris och kvantitet används för indexberäkningar genom exempelvis multilaterala metoder eller vissa bilaterala metoder. Detta förutsätter ofta en form av gruppering av likvärdiga produkter, utöver en komplett och ständigt aktualiserad klassificering av datamaterialet.

4.2 Ytterligare aspekter att beakta

I analyserna av de nya datamängderna föreligger frågan om användning både för prisinsamlingen och för att kunna genomföra produktbyten under året – med tillhörande kvalitetsvärderingsfrågor. Med *byten* menas att produkter som inte längre kunnat påträffas i prisinsamling eller antas vara aktuella ersätts av nya.

Dagens utformning av urvalen, för betydande delar av KPI, beaktar förutsättningarna för kvalitetsvärderingar vid produktbyten eftersom detta är en resursintensiv process i fastkorgsansatsen. Detta ger en restriktion för mängden produkter som kan ingå i korgen givet befintlig indexmetod.

För att kunna genomföra värderingarna av kvalitetsegenskaper vid byten förutsätts information som datamängderna inte nödvändigtvis svarar för. I de fallen blir det fortfarande aktuellt att ha en ”klassisk” intervjuaransats via hemsidor. I andra fall kan det finnas uttömmande egenskapsinformation i datakällorna eller så kan detta fångas genom tredjeparts-sidor på internet, och förutsättningarna varierar mellan olika typer av produkter. Exempel på behov av mer uttömmande egenskapsinformation är vid kvalitetsvärderingar av hemelektronik (Eliasson m. fl. 2021) och kläder.

I föreliggande analyser har det framkommit varierande förutsättningar att använda kassaregisterdata, särskilt mot den nämnda komplexiteten i produktdefinitionerna. Det är inte alltid ett entydigt förhållande mellan vad en konsument uppfattar som en produkt (liksom vad som presenteras

på en hemsida) visavi vad som finns i företagens databaser. Exempelvis kan ett artikelnummer på en hemsida motsvaras av sammansättningen av flera artikelnummer i en databas. Vissa datamängder har därför snabbare kunnat implementeras i KPI när motsvarande komplexitet inte funnits.

4.3 Redan implementerade kassaregister

Kassaregisterdata som insamlingsmetod har tillämpats i över ett decennium i KPI. Idag beräknas KPI baserat på över 100 000 prisnoteringar från kassaregisterdata, utöver övriga datakällor. Kassaregisterdata används redan inom livsmedel, alkoholhaltiga drycker, läkemedel, tandvård, tågresor, charterresor, mäklartjänster och drivmedel. Från och med 2021 infördes kassaregisterdata även för hemelektronik som ersättning till den manuella prisinsamlingen.

I befintliga implementeringar av kassaregisterdata kan ibland olika datakällor ingå i samma beräkningsgrupp (elementäraggregat) och kan kräva helt olika produktionssystem/datahanteringar. En väg förbi detta är s.k. *mikroaggregat* (Ståhl 2021) vid utökad användning av kassaregisterdata för hantering av olika datakällor inom samma beräkningsgrupp.

Det finns tydliga skillnader i produktlivscyklerna exempelvis inom dagligvaror (där produkter uppvisar hög kontinuitet över tid) kontra hemelektronik och kläder som medför ett mer resurskrävande arbete för att upprätthålla urvalet. Utökad användning ändrar därmed också problembilden: dels uppkommer frågan om täckning/granularitet (Ståhl 2021) över tid per datakälla (homogeniteten), dels förutsätts metoder för kvalitetskorrigering beroende på vald indexmetod.

4.4 Alternativen till kassaregister

Från och med 2021 tillämpas automatisk webbskrapning i KPI från ett 20-tal hemsidor. I dagsläget fångas priser för nästan 2 000 produkterbjudanden per månad i flertal produktgrupper genom flera skrapningar i veckan varefter medelpris beräknas utifrån dessa.

Det finns två principiella former av webbskrapning (Eurostat 2020). Endera kan kompletta hämtningar göras av hela hemsidor, s.k. *bulk*, eller så kan specifika produkter på hemsidan hämtas, s.k. *target*. Båda former har tillämpats för prisinsamling till KPI men för närvarande görs enbart specifika hämtningar (*target)* av tekniska skäl. I och med att datamängden begränsas till enbart specifika hämtningar blir produktbytena desto mer begränsade eftersom ingen prishistorik fångas per automatik, vilket är fallet med *bulk*-hämtningar.

Det finns dock praktiska utmaningar med webbskrapning som SCB erfarit. Det är vanligt att hemsidor vidareutvecklas, vilket innebär att webbskrapningsverktyget snabbt kan behöva anpassas och medför ibland

snabba praktiska och tekniska omställningar. I sådana lägen blir resursbehoven för prisinsamlingen något oförutsägbara. Samtidigt innebär webbskrapning att kvantitetsunderlag inte finns att tillgå vilket betyder att produkturvalen hanteras i likhet med manuell insamling.

Utöver webbskrapning finns möjligheter att fortsätta med manuell insamling från nätet, vilket är en tekniskt okänslig metod. Därtill kommer även prislistor från internet genom så kallad API.

5. Summering

Det finns förutsättningar för ytterligare implementering av kassaregisterdata samtidigt som flertalet av de tillgängliga nya datakällorna innebär nya utmaningar i termer av klassificering och kvalitetsvärderingar. Insatserna behöver även värderas i förhållande till motsvarande täckning av marknaden varför nyttokvoten bör tänkas igenom avseende de möjliga vägarna fram genom antingen kassaregister, webbskrapning eller manuell insamling.

6. Frågor till nämnden

SCB inbjuder nämnden till diskussion om vägvalen för ytterligare datafångst till KPI, framför allt avseende vad nämnden anser vara av särskild betydelse i det fortsatta arbetet med nya datakällor till resterande delar av KPI.

Referenser

Eliasson, J., Hillström, E., Nordin, M., Ottosson, M. och Tongur, C. (2021). *Prisindex för mobiltelefoner och datorer med fast korg*. PM till nämnden för Konsumentprisindex, sammanträde 12, 2021. https://scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/prisindex-for-mobiltelefoner-och-datorer-med-fast-korg.pdf

Eurostat (2020). *Practical guidelines on web scraping for the HICP*. https://ec.europa.eu/eurostat/documents/272892/12032198/Guidelines-web-scraping-HICP-11-2020.pdf/

ILO (2020). *Consumer Price Index Manual*. https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_761444.pdf

Norberg, A., Sammar, M. och Tongur, C. (2011). *A study on scanner data in the Swedish CPI*. Room document, Ottawa Group Meeting 2011. Tillgänglig via www.ottawagroup.org.

SCB (2016). Statistiska centralbyråns föreskrifter om skyldighet att lämna uppgifter till statistisk avseende Konsumentprisindex. SCB-FS 2016-25.

Ståhl, O. (2021). *Aggregeringsprinciper inom COICOP 01 och 02.* PM till nämnden för Konsumentprisindex, sammanträde 12, 2021. https://scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/aggregeringsprinciper-inom-coicop-01-och-02.pdf

Annex

Leveransspecifikation för Kassaregisterdata för statistiska ändamål

I den här leveransspecifikationen beskrivs det dataleveransformat som SCB önskar när uppgiftslämnare tar ut information från kassaregister för statistikändamål. Nedan följer både den filformatsstruktur och de viktiga variabler som önskas ingå i underlaget.

En transaktionsdatafil skickas på veckovis basis och innehåller uppgifter motsvarande föregående kalenderveckas transaktioner. Vi ser gärna att leveranserna av denna fil sker i början av efterföljande vecka. Uppgifterna lämnas antigen via SFTP eller genom API-anrop.

SCB är endast intresserad av försäljning där moms för såld vara/tjänst betalats i Sverige. Leveransen ska representera respektive periods nettoförsäljning och därmed ska även produkter som returnerats inkluderas.

Varor/tjänster som sålts när företaget har agerat marketplace, dvs marknadsfört och sålt varor för en annan aktörs räkning, ska exkluderas från dataleveransen. Varor som skickats med som gåvor i samband med en transaktion och som inte kostat något extra för kunden skall exkluderas ur materialet. Notera att detta inte gäller specialerbjudanden så som ”ta 3 betala för 2”.

Utöver transaktionsdatafilen efterfrågar SCB en kompletterande egenskapsdatafil som vidare beskriver de sålda produkterna. Denna fil levereras månatligen.

Som komplement till denna leveransbeskrivning skickar SCB en exempelfil för att illustrera det efterfrågade materialet.

Transaktionsdatafil

Filer som skickas till SCB avseende kassaregisterdata är enligt kodning UTF-8, där data representeras som rader och kolumner med hjälp av en överenskommen fältavgränsare (förslagsvis semikolon). I de fall leverans sker genom SFTP så ska filen vara i antigen .txt eller .csv-format. Undvik att använda kommatecken som fältavgränsare eftersom

kommatecken ofta förekommer naturligt i vissa variabelvärden. Tecknet som används som separator får inte förekomma som ett tecken i variabelvärdena. Det går att lämna cellen tom när uppgiften saknas.

Transaktionerna redovisas per organisationsnummer, transaktionsdatum, transaktionstyp, artikelID (Artikelnummer eller EAN/GTIN). Transaktionerna redovisas även per butikstyp, kundtyp och försäljning/retur när det är aktuellt. Returer redovisas separat och aggregeras inte ihop med försäljningsdata. Det betyder att en produkt kan vara såld ett transaktionsdatum och redovisas som en retur vid ett senare transaktionsdatum. Om möjligt så innehar returer de transaktionstyper som de initiala köpen skett igenom.

Nedan presenteras ett önskemål om uppgifter som skall ingå i transaktionsdatafilen. Det finns ett antal uppgifter som behöver finnas med i underlaget för att SCB ska kunna nyttja leveransen. De finns listade i Tabell 1. Notera att minst en av uppgifterna Artikelnummer och EAN behöver förekomma och att cellen kan lämnas tom om något av numren saknas.

Utöver de obligatoriska uppgifterna så önskar SCB att transaktionsdatafilen även inkluderar uppgifter enligt Tabell 2. För SCB:s ändamål är det fördelaktigt om samtliga av dessa uppgifter inkluderas i filen. Om uppgiftslämnaren inte kan tillhandahålla samtliga uppgifter ser SCB gärna att exempelfilen fortfarande följs, men att de kolumner som ej innehåller variabelvärden lämnas tomma. Notera att vi trots indelningen av Tabell 1 och 2 önskar att de variabler som efterfrågas i båda tabellerna ingår i en och samma fil, se exempelfil som levereras separat. Inkludera även rubrikerna enligt specifikation i exempelfilen.

Tabell 1 Obligatoriska variabler i Transaktionsdatafil

Uppgift	Beskrivning	Format	Datatyp
Organisationsnummer	Organisationsnummer för företaget som genomför transaktionen	Organisationsnummer: • XXXXXX-XXXX	Text
Transaktionsdatum	Datum för när transaktionen genomförts	ÅÅÅÅ-MM-DD	Date
Transaktionstyp	Transaktion uppdelat per säljkanal Online = onlineförsäljning Offline = försäljning i butik Marketplace = försäljning där annan aktör sålt varan för företagets räkning Click and collect = försäljning där varan beställts online men hämtats ut i butik	Giltiga värden: • ”Online” • ”Offline” • ”Marketplace” • ”Click and collect”	Text

Uppgift	Beskrivning	Format	Datatyp
ArtNr	Unikt artikelnummer för den specifika produkten. Företagets egna artikelnummer som är skilt från EAN/GTIN.	1-50 tecken, Kan även lämnas tom	Text
EAN/GTIN	Artikelnummer enligt GS1-standarden	1-50 tecken, Kan även lämnas tom	Text
SaldaPerArtNr	Antalet sålda enheter per artikelnummer	-999999999999 - 999999999999	Heltal
TotalOmsPerArtNrInklMoms	Total omsättning per artikelnummer inklusive moms	-999999999999 - 999999999999	Decimaltal
TotalOmsPerArtNrExklMoms	Total omsättning per artikelnummer exklusive moms	-999999999999 - 999999999999	Decimaltal

Tabell 2 Övrig relevant information i transaktionsdatafil

Uppgift	Beskrivning	Format	Datatyp
Artikelbenamning	Artikelbenämningen för produkten	1-200 tecken, Kan även lämnas tom	Text
ExternLeverantor	Företagsnamn för den externa leverantören av produkten som sålts	1-50 tecken, Kan även lämnas tom	Text
ArtNrLeverantor	Den externa leverantörens artikelnummer för produkten	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNamn1	Produktens grupptillhörighet, grupp 1	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNummer1	Numret som motsvarar produktens grupptillhörighet, grupp 1	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNamn2	Produktens grupptillhörighet, grupp 2	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNummer2	Numret som motsvarar produktens grupptillhörighet, grupp 2	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNamn3	Produktens grupptillhörighet, grupp 3	1-50 tecken, Kan även lämnas tom	Text

ProduktgruppNummer3	Numret som motsvarar produktens grupptillhörighet, grupp 3	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNamn4	Produktens grupptillhörighet, grupp 4	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNummer4	Numret som motsvarar produktens grupptillhörighet, grupp 4	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNamn5	Produktens grupptillhörighet, grupp 5	1-50 tecken, Kan även lämnas tom	Text
ProduktgruppNummer5	Numret som motsvarar produktens grupptillhörighet, grupp 5	1-50 tecken, Kan även lämnas tom	Text
Butikstyp	Benämning för vilken butikstyp där transaktionen genomförts. Anges om företaget har flera typer av butiker med olika koncept	1-50 tecken, Kan även lämnas tom	Text
AntalPerPack	Antal per förpackning	-999999999999 - 999999999999, Kan även lämnas tom	Decimaltal
MattenhetPack	Förpackningens Måttenhet	1-50 tecken, Kan även lämnas tom	Text

Uppgift	Beskrivning	Format	Datatyp
Kundtyp	Transaktion uppdelat per kundtyp	Giltiga värden: • ”Företag/organisation” • ”Privatperson” , kan även lämnas tom	Text
Forsaljning/retur	Transaktion uppdelat per försäljning/retur	Giltiga värden: • ”Försäljning” • ”Retur” • ”Reklamation” , kan även lämnas tom	Text

Egenskapsdatafil

Utöver transaktionsdatafilen efterfrågar SCB kompletterande egenskapsdata som skickas på månatlig basis. Nedan listas uppgifterna som ska inkluderas i denna fil. Filen måste innehålla antingen EAN eller artikelnummer. SCB ser helst att även dessa filer följer ett semikolonseparerat textformat (Ex .txt eller .csv).

Uppgift	Beskrivning	Format	Datatyp
ArtNr	Unikt artikelnummer för den specifika produkten. Företagets egna artikelnummer som är skilt från EAN/GTIN.	1-50 tecken, Kan även lämnas tom	Text
EAN/GTIN	Artikelnummer enligt GS1-standarden	1-50 tecken, Kan även lämnas tom	Text
EgenskapsID	Identifierande kod för egenskapen	1-50 tecken, Kan även lämnas tom	Text
Egenskapsnamn	Namnet på egenskapen som beskrivs	1-50 tecken, Kan även lämnas tom	Text
Egenskapsvarde	Värdet för den angivna egenskapen	1-4000 tecken, Kan även lämnas tom	Text
Egenskapsenhet	Egenskapsvärdets måttenhet	1-50 tecken, Kan även lämnas tom	Text

Kontakt

Vid frågor kontakta SCB:s registerenhet

Mejl: registerdata@scb.se