

Folkbokföringsfelet

En metodrapport om skattning av
övertäckning, undertäckning och fel
folkbokföringsadress

Folkbokföringsfelet En

metodrapport om skattning av övertäckning, undertäckning och fel folkbokföringsadress

Producent SCB, Statistiska centralbyrån
Dataavdelningen
701 89 Örebro
010-479 40 00

Förfrågningar Karin Wegfors
+46 10 479 64 98
karin.wegfors@scb.se

Sara Westling
+46 10 479 68 26
sara.westling@scb.se

Du får kopiera och på annat sätt mångfaldiga innehållet.

Vi vill dock att du uppger källa på följande sätt:

Källa: SCB, Folkbokföringsfelet

En metodrapport om skattning av övertäckning, undertäckning och fel
folkbokföringsadress Dnr A2022/4407

Innehåll

Inledning	4
Termer och begrepp	4
Register och datakällor	5
Modell – Övertäckning	7
Misstänkt övertäckning.....	8
Inaktivitetsvariabel	9
Modellerad responsvariabel	9
Logistisk regression	10
Indikatorer	10
Skattning av övertäckning	16
Prediktion av övertäckning	16
Modell – Undertäckning	17
Modellen	17
Ingående komponenter	17
Modell – fel folkbokföringsadress	20
Modellen	20
Ingående komponenter	21
Skattat resultat av folkbokföringsfelet.....	24
Referenser.....	25

Inledning

SCB har tillsammans med Skatteverket utvecklat en metod för att systematiskt kunna följa upp kvaliteten i folkbokföringsregistret. Ett sätt att åstadkomma det är att med kvantitativa metoder uppskatta folkbokföringsfelet där folkbokföringsfelet är en avvikelse mellan vad en person borde ha för uppgifter och vad samma person faktiskt har för uppgifter.

Som ett led i arbetet har en metod utvecklats för att uppskatta folkbokföringsfelet i folkbokföringsregistret via SCB:s register, då utifrån tre kvalitetsaspekter:

- Övertäckning
- Undertäckning
- Mätfel – Fel folkbokföringsadress

Samarbetet startade 2018 i och med att Skatteverket fick ett uppdrag från regeringen om att utveckla en metod för att systematiskt kunna följa upp kvaliteten i folkbokföringsregistret. Uppdraget redovisades till Finansdepartementet i september 2018 i form av en rapport (Skatteverket, 2018). Metoden har sedan utvecklats ytterligare och numera görs beräkningar årligen av SCB. Eftersom rampopulation är densamma för Skatteverkets folkbokföringsregister och SCB:s Register över totalbefolkningen (RTB) och målpopulationen är densamma för de båda registren så bedöms samma modell kunna användas för att uppfylla både SCB:s och Skatteverkets behov av att uppskatta folkbokföringsfelet.

Termer och begrepp

Nedan definieras vissa centrala termer och begrepp.

Övertäckning – En täckningsbrist där rampopulationen innehåller personer som inte tillhör målpopulationen. I RTB finns samtliga personer som var folkbokförda i Sverige ett visst år. En del av dessa personer bor egentligen någon annanstans och borde inte vara folkbokförda i Sverige.

Undertäckning – En täckningsbrist där rampopulationen saknar personer som tillhör målpopulationen. Vissa personer som bor i Sverige och som inte är folkbokförda borde egentligen vara det enligt folkbokföringslagen.

Mätfel – En avvikelse mellan det efterfrågade sanna värdet och det värde som faktiskt observerats. Exempel på mätfel är folkbokförd på annan adress än där man borde vara folkbokförd om alla regler och lagar

efterlevts. Andra exempel kan vara egenskaper som kön, ålder eller civilstånd där det efterfrågade sanna värdet skiljer sig mot det registrerade.

Register och datakällor

Inom ramen för arbetet har flera olika register och datakällor identifierats som behövs för att ta fram skattningarna på folkbokföringsfelet och dessa presenteras nedan.

Registret över totalbefolkningen (RTB)

Det statistiska register som SCB ansvarar för som utgår från folkbokföringsregistret men som har bearbetats för att användas till produktion av statistik.

Inkomst- och taxeringsregistret (IoT)

Registret ligger till grund för den officiella inkomst- och skattestatistiken. Registret innehåller uppgifter från Skatteverket (taxeringsuppgifter, kontrolluppgifter från arbetsgivare, kontrolluppgifter från pensions- och försäkringsutbetalare, kontrolluppgifter över räntor och utdelningar samt standardiserade räkenskapsutdrag), Försäkringskassan, Centrala studiestödsnämnden, Statens tjänstepensionsverk, Pensionsmyndigheten, Försvarsmakten, Skolverket och Socialstyrelsen.

Universitets- och högskoleregistret

Registret ligger till grund för officiell högskolestatistik på grundnivå, avancerad nivå och forskarnivå. Uppgifterna samlas in från Sveriges högskolor och universitet.

Databasen för statistiska integrationsvariabler (STATIV)

Statistikdatabasens syfte är att utgöra underlag för beskrivning och analys avseende integration, segregation, jämställdhet och migration. Innehåller uppgifter från flera andra individregister som produceras av SCB samt data från Migrationsverket, Arbetsförmedlingen, Försäkringskassan och Socialstyrelsen.

Lägenhetsregistret

Registret används för den officiella bostads-, hushålls- och boendestatistiken. Uppgifterna kommer från Lantmäteriet som är den myndighet som förvaltar och ansvarar för registret.

Elever i grundskolan

Registret innehåller samtliga elever inskrivna i grundskola, förskoleklass eller fritidshem per den 15 oktober.

Registerbaserad aktivitetsstatistik

Uppgifterna beskriver en persons totala försörjning och anknytning till arbetsmarknaden.

Kontrolluppgiftsbaserad lönesummestatistik

Registret visar lönesummor, preliminär A-skatt och vilka sociala ersättningar som betalas ut. Uppgifterna baseras på kontrolluppgifter som hämtas från Skatteverket.

Registret för migrations- och asylstatistik

Registret innehåller uppgifter om beviljade uppehållstillstånd för utomnordiska immigranter. Uppgifterna hämtas från Migrationsverket.

Gränsregional statistik

Uppgifterna visar alla personer som haft sin huvudsakliga löneinkomst i annat nordiskt land under ett givet år.

Skatteverkets impulsdata i folkbokföringsdatabasen

Data som visar iakttagelser från andra myndigheter och privatpersoner om indikationer på fel i folkbokföringen.

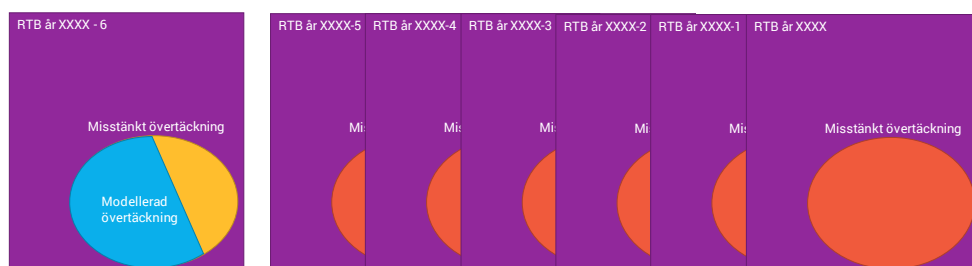
Modell – Övertäckning

Metoden för att skatta övertäckningen i RTB kan delas in i tre delar, vilka är:

- definitionen av den misstänkta övertäckningspopulationen
- den modellerade responsvariabeln
- den logistiska regressionen och förklaringsvariabler (som benämns indikatorer)

Utgångspunkten för att skatta övertäckningen för år XXXX är att definiera gruppen misstänkt övertäckning för år XXXX-6. Sedan följer man varje person framåt i tiden för att se om de lämnar ett avtryck i något register. Om personen lämnar ett avtryck i ett register klassificeras den som inte övertäckning, om personen försvinner från RTB klassificeras den som övertäckning och om personen är inaktiv varje år (6 år i rad) klassificeras den också som övertäckning. På det här sättet skapas en variabel som tar värdet 1 om personen antas tillhöra övertäckning och 0 om den antas tillhöra målpopulationen. Det är den här variabeln som benämns den modellerade responsvariabeln. Denna modellerade responsvariabel används sedan i en logistisk regression tillsammans med 24 förklaringsvariabler. De 24 förklaringsvariablerna består av indikatorer som tar värdet 1 eller 0. Indikatorerna är härledda utifrån registerdata där varje indikator har ett samband med den modellerade responsvariabeln. Den här modellen tilldelar varje person i den misstänkta populationen en sannolikhet att tillhöra övertäckningen. Det innebär att man får fram en modell baserat på data från år XXXX-6 som sedan går att applicera på misstänkta populationer framåt i tiden till år XXXX. För att skatta övertäckningens storlek summeras sannolikheterna för hela den misstänkta populationen.

Följande bild illustrerar de populationer som används i metoden där den första populationen är för den man modellerar responsvariabeln och de efterkommande åren är för vilket man applicerar modellen:



För varje misstänkt övertäckningspopulation har 24 indikatorer tagits fram vilket gör det möjligt att tillämpa den modell som härleds från år XXXX-6 framåt i tiden.

Misstänkt övertäckning

För att definiera den misstänkta övertäckningen letar man efter om en person lämnat avtryck i något register under samma år. De register som används är RTB, Inkomst- och taxeringsregistret (IoT) och databasen för statistiska integrationsvariabler (STATIV).

Om en person lämnat avtryck i något av dessa register klassificeras den inte som misstänkt övertäckning. Det dröjer drygt ett år innan registerdata finns tillgängligt för innevarande år, vilket innebär att man behöver vänta så länge innan man kan definiera den misstänkta övertäckningen.

RTB

För RTB gäller att om personen registrerats med någon av nedanstående demografiska händelser så kommer personen inte klassificeras som misstänkt övertäckning:

- Födelse
- Immigration
- Civilståndsändring
- Medborgarskapsbyte
- Inrikes flytt
- Emigration
- Dödsfall
- Skyddade personuppgifter

Det finns vissa undantag där en händelse leder till att man ändå klassificeras som misstänkt övertäckning. Undantagen är när personen ändrar civilstånd till änka/änkling/efterlevande partner samt när en person folkbokförs utan känd hemvist.

IoT

För IoT handlar det inte enbart om personens avtryck i registret utan man ser till personens familj. Om en person har ett sammanlagt värde för nedanstående variabler som är lika med eller överstiger ett basbelopp klassificeras den inte som misstänkt övertäckning:

- Studiemedel och studiehjälp (lån och bidrag) för familj
- Ekonomiskt bistånd för familj
- Introduktionsersättning till flyktingar för familj
- Arbetsinkomst för familj. Beräknas som summan av lön, företagarsinkomst, sjukpenning, föräldrapenning och ersättning i samband med militärtjänstgöring
- Sammanräknad förvärvsinkomst för familj. Beräknas som summan av inkomst av tjänst och inkomst av näringsverksamhet

Utöver ovanstående klassificeras även personer som enbart har garantipension som misstänkt övertäckning. Variabeln är härledd

utifrån att beloppet för garantipension är samma som beloppet för inkomst av tjänst.

STATIV

För STATIV gäller att om en person har ett värde för någon av nedanstående variabler klassificeras den inte som misstänkt övertäckning:

- Sysselsättningsstatus
- Studiedeltagande
- Inskrivnen på Arbetsförmedlingen
- Antal dagar inskriven på Arbetsförmedlingen under året
- Deltagande i svenska för invandrare (SFI)
- Studerande på gymnasiet
- Avgången från gymnasiet under året
- Grund för bosättning som pensionär/tillräckliga medel

Inaktivitetsvariabel

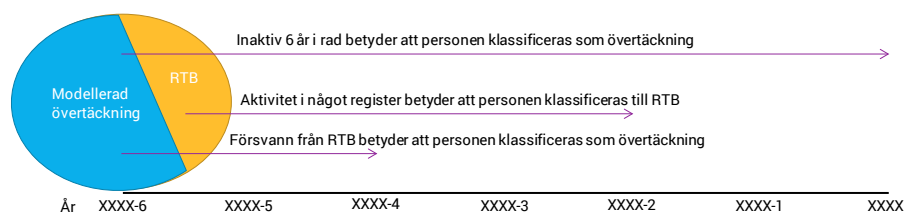
Inaktivitetsvariabeln är den variabel som används för att definiera den misstänkta övertäckningen och sätts till 1 om personen inte lämnat avtryck i RTB, inte lämnat avtryck i STATIV samt har ett sammanlagt värde av IoT-variablerna som understiger ett basbelopp. Utöver det sätts den till 1 om personen enbart har garantipension.

Modellerad responsvariabel

Den modellerade responsvariabeln definieras för år XXXX – 6 för samtliga personer i den misstänkta övertäckningen och tar värdet 1 eller 0 där 1 innebär att man tillhör övertäckningen och 0 innebär att man tillhör RTB. Samtliga personer för den misstänkta övertäckningen följs 6 år framåt i tiden till år XXXX.

- Om personen visar sig aktiv något av de efterföljande åren sätts den till 0
- Om personen försvinner från RTB (emigration eller dödsfall) sätts den till 1
- Om personen varit inaktiv 6 år i rad sätts den till 1

På det här sättet skapas en modellerad responsvariabel för hela den misstänkta övertäckningen för år XXXX – 6.



Logistisk regression

Logistisk regression kan ses som en statistisk metod för att härleda en statistisk modell som använder den logistiska funktionen för att modellera binära responsvariabler givet en uppsättning förklaringsvariabler. Den kan användas för att identifiera samband mellan responsvariabel och förklaringsvariabler.

Den statistiska modellen som härletts används för att skatta sannolikheten att tillhöra övertäckning givet en uppsättning av förklaringsvariabler, som här benämns indikatorer. Modellen som tagits fram går att skrivas med formeln:

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{24} \beta_i x_i)}}$$

som kan förklaras som sannolikheten att en person tillhör övertäckningen givet den statistiska modell som tagits fram som består av ett intercept och 24 förklaringsvariabler. Samtliga förklaringsvariabler, även kallat indikatorer, är binära variabler som tar värdet 1 eller 0.

Den logistiska regressionens parametrar, eller snarare koefficienter, har skattats med maximum likelihood-metoden där *Fishers scoring-algorithm* använts.

Indikatorer

I det här kapitlet beskrivs de indikatorer som skapas och sedan används i den logistiska regressionen. Samtliga indikatorer tar värdet 1 eller 0. Härledningen av indikatorerna utgår ifrån data från och med 2001 som startår fram till det senaste tillgängliga statistikåret för vilket samtliga register finns tillgängliga. Det är för denna tidsperiod (2001 – XXXX) som beräkningar görs. Varje indikator beräknas för varje enskilt år. Vissa indikatorer baseras på data från tidigare år medan andra indikatorer enbart baseras på ett år. Endast en indikator baseras på data både bakåt och framåt i tiden.

Eftersom det enbart används data tillbaka till 2001 går man miste om aktiviteter och inaktivitet tidigare år, vilket betyder att träffsäkerheten för indikatorerna i början av tidsperioden har sämre kvalitet än senare år. Men det påverkar inte skattningarna av övertäckningen för XXXX eftersom man utgår ifrån den misstänkta populationen 6 år tillbaka i tiden.

Indikator 1 – Gränspendling

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som går att definiera som gränspendlare (inkomst i annat nordiskt land men folkbokförd i Sverige).

Indikatorn har tagits fram eftersom det är rimligt att tro att registrerade gränspendlare faktiskt bor i Sverige och ska tillhöra RTB

Indikator 2 – Högskolepoäng

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som studerat mer än 10 högskolepoäng ett år.

Indikator 3 – Återkommande aktivitet i RTB

Den här indikatorn sätts till 1 om:

- Personen var inaktiv
- Personen har varit inaktiv minst tre gånger under tidsperioden
- Personen har varit aktiv i RTB minst tre gånger under tidsperioden samtidigt som ingen aktivitet i IoT eller STATIV registrerats

Denna indikator använder sig av data både framåt och bakåt i tiden. Det betyder att aktivitet/inaktivitet i början av en tidsperiod påverkar indikatorns resultat i slutet på tidsperioden. Även det omvända gäller där aktivitet/inaktivitet i slutet på tidsperioden påverkar indikatorns resultat i början av tidsperioden.

Indikatorn har tagits fram eftersom det är rimligt att tro att en person som flera gånger under en och samma tidsperiod registreras i RTB faktiskt finns i Sverige och därmed borde tillhöra RTB.

Indikator 4 – Aktiv RTB-familjemedlem

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som har minst en till i sin familj som är aktiv.

Indikatorn har tagits fram eftersom det är rimligt att tro att en person som är inaktiv men har aktiva familjemedlemmar förmodligen tillhör RTB även om personen själv inte har någon aktivitet.

Indikator 5 – På kommunen skriven

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som är folkbokförda på församlingen eller på kommunen.

Indikatorn har tagits fram eftersom ”på församlingen skriven” eller ”på kommuner skriven” tyder på att myndigheterna vet något om personen och att hen åtminstone har funnits i området men utan adress. Personer som står skrivna på församling eller kommun är dessutom personer som förmodligen inte lämnar så mycket avtryck i register trots att de faktiskt finns i Sverige.

Indikator 6 – Äldre utländska immigranter

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för personer som blir inaktiva efter att man som utländsk medborgare över 60 år immigrerat till Sverige. Samtliga efterföljande år av inaktivitet kommer ge träff på indikatorn.

Indikatorn har tagits fram eftersom det är rimligt att tro att äldre personer med utländskt medborgarskap som immigrerar till Sverige har tjänat ihop tillräckligt med pengar för att kunna bo i Sverige utan att synas i något register.

Indikator 7 – Ränteinkomst

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som hade ränteintäkter som var två prisbasbelopp eller större.

Indikatorn har tagits fram eftersom det är rimligt att personer som är inaktiva men har ränteinkomster med två prisbasbelopp eller större kan försörja sig själva och därmed borde tillhöra RTB.

Indikator 8 – Familjeinkomst

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer med en familj som har en inkomst (studiemedel, ekonomiskt bistånd, förvärvsinkomst, arbetsinkomst eller etableringsersättning) som är större än 0.

Indikatorn har tagits fram eftersom man kan misstänka att personer med familj som har en inkomst (även en låg inkomst) kan bo i Sverige utan att lämna avtryck i några register.

Indikator 9 – Anknytning till arbetsmarknaden

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som registrerats med en anknytning till arbetsmarknaden (helårsanställd, nyanställd, avgången, delårsanställd, kombinator eller företagare).

Indikatorn har tagits fram eftersom det är rimligt att tro att personer med en anknytning till arbetsmarknaden bor i Sverige.

Indikator 10 – Utländska högskolestudenter

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 ifall en person som invandrat, är utländsk medborgare och studerat mer än 10 högskolepoäng efterföljande år blir inaktiv.

En person som invandrat följs för hela tidsperioden och om personen studerar mer än 10 högskolepoäng ett år och året därpå blir inaktiv kommer denne att få träff på indikatorn. Om den däremot är aktiv mellan studier och inaktivitet nollställs indikatorn. Om personen är

inaktiv hela efterföljande period efter studier kommer samtliga år ge träff på indikatorn.

Indikatorn har tagits fram eftersom det finns personer som immigrerar till Sverige för att studera. Dom folkbokför sig under studietiden men flyttar sedan tillbaka till hemlandet utan att avregistrera sig. Syftet med indikatorn är att fånga dessa personer eftersom de troligtvis hör till övertäckningen.

Indikator 11 – Studier som grund för bosättning

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer med studier som grund för bosättning samtidigt som personen inte tagit några högskolepoäng det innevarande året.

Indikatorn har tagits fram eftersom det är rimligt att en person inte befinner sig i Sverige om den inte tar några högskolepoäng samtidigt som grunden för bosättning är studier. Den här indikatorn kan ses som ett komplement (med stort överlapp) till indikator 10 om utländska högskolestudenter eftersom denna även fångar de som saknar högskolepoäng innevarande år.

Indikator 12 – Immigration följt av inaktivitet

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för personer som blir inaktiva och saknar familj åren efter att man invandrat.

Indikatorn har tagits fram eftersom det finns flera som invandrar men sen utvandrar året därpå utan att avregistrera sig. Det kan handla om personer som flyttar vidare till andra länder inom EU.

Den här indikatorn har en brist som är att vissa personer är inaktiva 1-2 år efter invandring innan de kommer in i systemet och syns i de övriga registren. Men för att hantera det behöver man använda framåtblickande information vilket modellen inte stöder. Indikatorn bedöms dock vara tillräckligt bra för ändamålet.

Indikator 13 – Immigration följt av IoT och sedan inaktivitet

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för personer som immigrerat, varit aktiva i IoT året därefter och sedan blivit inaktiva. Samtliga efterföljande och sammanhängande år av inaktivitet sätts till 1.

Man utgår ifrån det året som personen invandrade och följer personen året därefter och ser om den hade en inkomst. Om personen hade en inkomst kommer den sedan få träffar på indikatorn för samtliga år där den blir inaktiv. Om personen blir aktiv något år så kommer de efterföljande åren inte ge någon träff även om personen är inaktiv.

Indikatorn har tagits fram eftersom personer som flyttar till Sverige för att jobba en kort tid kan sakna kunskap om eller incitament att man ska avregistrera sig.

Det finns en liten brist i indikatorn och det är att en person som immigrerat och sedan haft inkomst flera år i rad som sedan blir inaktiv kommer ge träff på indikatorn. Det finns naturliga skäl till att bli inaktiv i vissa fall och den här indikatorn ger förmodligen något större utslag än den borde. Men bedömningen är att det inte är ett stort problem och att den i det stora hela fångar in det beteende som man har som syfte att fånga in.

Indikator 14 – Utan känd hemvist

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som är registrerade som utan känd hemvist.

Indikatorn har tagits fram eftersom det är rimligt att tro att personer som är inaktiva och saknar känd hemvist inte finns i befolkningen och därmed är övertäckning.

Indikator 15 – Inkomst följt av nollinkomst och inaktivitet samtidigt

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för personer som haft inkomst och de efterföljande åren följs av nollinkomst och inaktivitet.

Om en person haft inkomst ett år där personen de efterföljande åren är inaktiv samtidigt som den har nollinkomst kommer det ge träff på indikatorn. Om personen blir aktiv efter nollinkomst så nollställs indikatorn och algoritmen letar på nytt efter nollinkomst och inaktivitet som kommer efter ett år av inkomst.

Indikatorn har tagits fram eftersom personer som plötsligt går till nollinkomst och blir inaktiva, från att ha varit aktiva med inkomst, rimligen borde höra till övertäckningen.

Indikator 16 – Utländsk medborgare som blir ensamstående

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för de åren en person med utländskt medborgarskap eller född utomlands blir inaktiv och ensamstående efter att ha varit aktiv med en familj större än 1 person.

Indikatorn har tagits fram eftersom det är rimligt att tro att personer som går från en familjesituation och aktiv till inaktiv och ensam förmodligen har glömt avregistrera sig.

Indikator 17 – Yngre familjemedlem kvar på samma adress

Denna indikator använder data bakåt i tiden till år 2001. Den sätts till 1 för en person som blir inaktiv efter att familje-ID ändrats till personens personnummer samtidigt som personen bor kvar på samma adress och

att det är minst 15 års åldersskillnad mellan personen som flyttade/dog och personen som bor kvar på samma adress.

Indikatorn har tagits fram eftersom personer som går från att vara yngst till äldst i en familj i samband med inaktivitet troligen är övertäckning.

Indikator 18 – Arbete som grund för bosättning

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer där grund för bosättning är arbete.

Den här indikatorn togs fram eftersom personer med arbete som grund för bosättning som är inaktiva misstänks utgöra övertäckning.

Indikator 19 – Inaktiv följt av död

Denna indikator använder sig av data bakåt i tiden och sätts till 1 för inaktiva personer som registrerats som död tidigare år.

Den här indikatorn togs fram eftersom det finns sena aviseringar för personer som dör och andra ev. fel i register som gör att en person kan bli kvar i dataunderlaget trots att personen är registrerad som död.

Indikator 20 – Endast garantipension

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer där inkomst av tjänst är större än noll men lika med garantipension samtidigt som personen har noll i transfereringar.

Den här indikatorn togs fram eftersom man misstänker att personer som är inaktiva och enbart har garantipension flyttat utomlands och därför är övertäckning.

Indikator 21 – Automatisk nolltaxerare

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som var automatiska nolltaxerare samt äldre än 25 år.

Den här indikatorn togs fram eftersom man misstänker att inaktiva personer som är automatiska nolltaxerare är övertäckning.

Indikator 22 – Utgående uppehållstillstånd

Denna indikator använder data bakåt i tiden och sätts till 1 för inaktiva personer som haft ett uppehållstillstånd som gått ut tidigare år. Data finns enbart från 2010 och framåt.

Den här indikatorn togs fram eftersom man misstänker att personer som är inaktiva med ett uppehållstillstånd som gått ut inte längre befinner sig i Sverige och därför är övertäckning.

Indikator 23 – Ej inskriven i grundskolan

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer mellan 6 – 15 år som inte är inskrivna i grundskolan.

Den här indikatorn togs fram eftersom man misstänker att barn i grundskoleålder som inte är inskriven i grundskolan flyttat utomlands.

Indikator 24 – Särskild postadress till utlandet

Denna indikator använder enbart data för ett år i taget och är därför inte beroende av data bakåt i tiden. Den sätts till 1 för inaktiva personer som har anmält en utländsk postadress till utlandet.

Den här indikatorn togs fram eftersom man misstänker att personer som har en adress i utlandet faktiskt bor i utlandet och därmed tillhör övertäckningen.

Skattning av övertäckning

För att skatta övertäckningen används den statistiska modell som härletts på populationerna för misstänkt övertäckning framåt i tiden till år XXXX. Sannolikheter skattas för varje person i hela gruppen misstänkt övertäckning och summeras sedan för respektive år. Summan av sannolikheterna är skattningen av övertäckningen.

Prediktion av övertäckning

Det är möjligt att prediktera om en person ska klassificeras som övertäckning eller inte. Summan av sannolikheterna i den misstänkta övertäckningen ger en skattning av övertäckningens storlek. Det antalet personer klassas som övertäckning. För att göra prediktionen sorteras populationen för misstänkt övertäckning efter sannolikhet att tillhöra övertäckning och ålder, i fallande ordning. Samtliga personer som är högst upp i sorteringen klassificeras som övertäckning. Anledningen till att ålder är med i sorteringen är för att flera personer har samma sannolikhet vid gränsen för prediktionen och erfarenheten säger att sannolikheten är högre att äldre personer tillhör övertäckningen.

Prediktionen av om en person är övertäckning eller inte används till skattningar av den totala befolkningssiffran som SCB tar fram till Eurostat.

Modell – Undertäckning

Undertäckning består av personer som inte är folkbokförda i Sverige men som borde vara det, om folkbokföringslagen efterlevts. Att skatta personerna via register är dock svårt eftersom dessa personer gör få eller inga avtryck i register. Valet föll därför på att utgå från en modell som är framtagen i syfte att skatta en total befolkningssiffra för Sverige (Bengtsson, 2015). Där utgjorde modellen för undertäckning en delkomponent vid beräkningen av den totala befolkningssiffran. Modellen har sedan justerats något för att passa syftet att skatta undertäckningen i folkbokföringsregistret och RTB.

Brister i rapporteringen av födselar och inflyttning resulterar i undertäckning. Undertäckning uppstår också om en person flyttar till Sverige och enligt lagar och förordningar ska folkbokföras men inte gör det. Det finns dock starka incitament för en person att folkbokföras eftersom många rättigheter som till exempel möjligheten att få olika typer av bidrag hänger ihop med att man är folkbokförd. Detta talar för att undertäckningen på grund av att personer inte folkbokför sig är liten.

Modellen

Modellen är en deterministisk komponentmodell där varje komponent modelleras för sig för att sedan summeras.

Modellen kan skrivas på formeln:

$$Y = \alpha + \beta + \gamma$$

Y = Skattat antal personer som ingår i gruppen undertäckning

α = Antal personer med sena aviseringar

β = Antalet utländska högskolestuderande som studerar i Sverige men inte är folkbokförda

γ = Antalet personer som har inkomst i Sverige men inte är folkbokförda

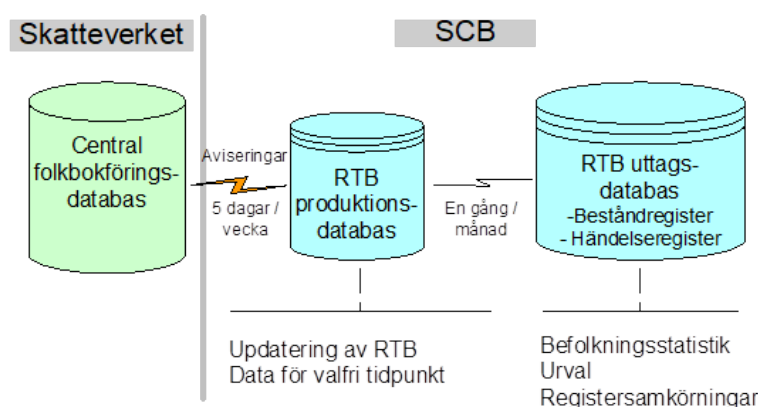
Ingående komponenter

Sena aviseringar – α

Figuren nedan visar hur uppgifterna från Skatteverket, så kallade aviseringar, uppdaterar RTB. Aviseringarna läggs in i RTB:s produktionsdatabas som uppdateras varje vardag. En gång i månaden tas information ur produktionsdatabasen för att skapa statistikregister som används för till exempel statistikproduktion, urvalsdragningar och

mikrodatautlämnanden. När statistikregistren är framställda är de ”låsta” och uppdateras inte senare med nya uppgifter. En avvägning mellan kvalitetsaspekterna aktualitet och täckning har gjort att man avvaktar en månad innan uttaget ur produktionsdatabasen görs. Aviseringar som inkommer senare än en månad kallas sena aviseringar.

Figur 1 Produktionsprocessen för RTB



Till undertäckningen i RTB räknas sena aviseringar rörande invandring och födselar. Antal sena aviseringar beräknas genom att summera alla aviseringar från Skatteverket avseende invandring och födselar som inkommit till SCB 1 februari t.o.m. 31 december aktuellt år och som avser tidigare år.

Högskolestuderande men inte folkbokförda – β

En del av undertäckningen består av ej folkbokförda studenter som studerar i Sverige mer än ett år. Dessa studenter borde enligt folkbokföringslagen vara folkbokförda i Sverige.

I högskoleregistret registreras alla högskolestudenter oavsett om de är folkbokförda eller inte. Genom att kontrollera vilka studenter som är inskrivna längre än ett år hittas de studenter som borde vara folkbokförda.

Antal utländska studenter beräknas genom att summera alla personer med tillfälliga personnummer som varit inskrivna i högskolan under höstterminen året före aktuellt år samt under både vår- och höstterminen under aktuellt år.

Inkomst i Sverige men inte folkbokförda - γ

I Inkomst- och taxeringsregistret (IoT) förekommer det personer med samordningsnummer. Ett samordningsnummer fungerar som en identitet för personer som inte är eller har varit folkbokförda, till exempel om man arbetat en kortare tid i Sverige. De som ingår i IoT är dels den grupp som valt att deklarerat med en vanlig kontrollblankett

(KU10) eller Kontrolluppgift – Sjöinkomst (KU16) från Skatteverket. Vidare ingår även uppgifter från bland annat Försäkringskassan och CSN för personer med samordningsnummer.

Personer med en disponibel inkomst över ett prisbasbelopp två år i rad anses bidra till undertäckningen. Valet av ett prisbasbelopp motiveras med att den gränsen används i flera andra sammanhang, exempelvis som gräns för inkomst då man inte har någon rimlig möjlighet att försörja sig.

Antal personer med inkomst i Sverige men som inte är folkbokförda beräknas genom att summera alla personer i IoT med samordningsnummer aktuellt år som även förekom året före och med en disponibel inkomst över ett prisbasbelopp båda åren.

Modell – fel folkbokföringsadress

Det förekommer att personer i Sverige är folkbokförda på fel adress. Det innebär att personer är folkbokförda på en adress där de inte borde vara folkbokförda enligt de lagar och regler som gäller.

Det har inte gått att hitta några tidigare registerbaserade modellansatser för att skatta antalet personer med fel folkbokföringsadress, vilket innebär att en ny metod och modell måste utvecklas. I det här kapitlet presenteras den modell som Skatteverket och SCB gemensamt landade i för att skatta antalet personer folkbokförda på fel adress.

Modellen

För att få en uppfattning om antalet personer med fel folkbokföringsadress har olika grupper av personer identifierats, som här benämns komponenter. En del av personerna i dessa komponenter misstänks ha fel folkbokföringsadress. Den utvecklade metoden är en deterministisk komponentmodell som liknar modellen som används vid skattning av undertäckningen. Varje komponent modelleras var för sig och summeras i slutändan ihop. Varje komponent kan skattas med flera mindre delmodeller där ett genomsnitt av dessa delmodeller beräknas för att få fram den enskilda komponentens skattning. Den här metoden har hämtat inspiration från Robert T Clemen (1989) som går igenom litteratur där man kombinerat resultatet från flera prognoser och sedan beräknat medelvärdet och funnit att det kan ge bättre skattningar.

Modellen kan skrivas på formeln:

$$Y = \frac{\sum_{i=1}^{N_1} \alpha_i}{N_1} + \frac{\sum_{i=1}^{N_2} \beta_i}{N_2} + \frac{\sum_{i=1}^{N_3} \gamma_i}{N_3} + \frac{\sum_{i=1}^{N_4} \mu_i}{N_4} + \frac{\sum_{i=1}^{N_5} \theta_i}{N_5}$$

Y = Skattat antal personer med fel adress

N_j = Antal modeller för den j:te komponenten

α_i = Den i:te modellen för att skatta antal studenter som pendlar orimligt långa sträckor

β_i = Den i:te modellen för att skatta antal personer som arbetspendlar orimligt långa sträckor

γ_i = Den i:te modellen för att skatta antal personer som bor på en orimligt liten boarea

μ_i = Den i:te modellen för att skatta antal personer folkbokförda i orimligt stora övriga hushåll

θ_i = Den i:te modellen för att skatta antal personer som efter impulser leder till en åtgärd om ändring av folkbokföringsadress

Komponenterna skattas så att de blir ömsesidigt uteslutande. Det görs genom att först beräkna antalet studenter med fel adress, där personer som även ingår i de andra komponenterna kan förekomma. Sedan beräknas antalet personer med långt pendlingsavstånd till arbetet där man selekterar bort de som även ingick bland studenterna. Beräkningarna görs i den ordning som visas i formeln ovan. Den sista komponenten, impulser, selekterar bort samtliga personer som förekommer i de fyra föregående komponenterna.

Ingående komponenter

I det här kapitlet beskrivs varje komponent mer ingående. Om skattningen för en komponent baseras på medelvärdet från flera mindre modeller så beskrivs de olika modellerna.

Modellens träffsäkerhet vilar på de antaganden som ligger till grund för respektive komponents modell och hur pass rimliga de är. Registervariablerna som används kan också vara behäftade med mätfel. Framtagningen av de gränser som används till de ingående komponenterna är subjektiv. Gränserna har tagits fram genom analys av percentilerna samtidigt som ett resonemang har förts kring variabelns sakinhåll utifrån syftet med det som ska skattas.

Studentpendling - α

En del studenter folkbokför sig inte på den adress där de faktiskt bor. Med hjälp av SCB:s register kan man skatta antalet studenter som är felaktigt folkbokförda genom att utgå från högskoleregistret och samköra med uppgifter från RTB.

Eftersom RTB visar befolkningen per den sista december varje år finns ett datum för när man är folkbokförd och det finns uppgifter om personens folkbokföringskommun. Den uppgiften kan samköras med högskoleregistret där uppgifter om studieort finns för varje termin. Även uppgifter om man läser på distans finns, och personendå går att selekteras bort.

Komponenten består av studenter som har studerat tre terminer i följd i samma kommun; vårterminen och höstterminen aktuellt år, samt vårterminen året efter. Studierna ska inte bedrivas på distans. Studenterna ska också vara folkbokförda i en annan kommun än studiekommunen 31 december aktuellt år. Det går inte att anta att alla dessa studenter är felaktigt folkbokförda eftersom det kan vara rimligt att bo i en kommun men studera i en annan, givet att pendlingsavståndet inte är alldeles för långt. Men det går att samköra

uppgifter om avstånd mellan kommunernas mittpunkter för att välja ut de personer som anses ha ett alldeles för långt pendlingsavstånd. Pendlingsavståndet delas in i fem grupper och sedan beräknas ett medelvärde av antal personer i grupperna:

- ≥ 40 km
- ≥ 70 km
- ≥ 100 km
- ≥ 130 km
- ≥ 150 km

Sättet att bryta ned en större grupp till mindre och mindre delar genom vissa regler är ett sätt att identifiera antalet studenter som är folkbokförda på fel adress. För varje nedbrytning antas att nedbrytningen blir mer och mer säker på att det är en felaktig folkbokföring. Samtidigt förloras information eftersom vissa studenter som selekterats bort längs vägen fortfarande kan ha fel adress. För att hantera det används medelvärdet av grupperna.

Arbetspendling – β

I STATIV finns uppgifter om vilket arbetsställe en person har. Om man arbetar i en kommun men är folkbokförd i en annan är det möjligt att man är felaktigt folkbokförd eftersom det kan vara så att man har sin egentliga hemvist i den kommun man arbetar.

Genom att mäta avståndet mellan folkbokföringskommunens och arbetsställekommunens mittpunkter går det att skatta antalet felaktigt folkbokförda. Det innebär antaganden om att det är orimligt att pendla över ett visst avstånd. Här antas att endast de personer som inte är gifta, registrerade partners eller sambo utgör den grupp som står för felaktig folkbokföring. Pendlingsavståndet delas in i fem grupper och sedan beräknas ett medelvärde av antal personer i grupperna:

- ≥ 400 km
- ≥ 500 km
- ≥ 600 km
- ≥ 700 km
- ≥ 800 km

Liten boarea – γ

Vissa hushåll bor mer trångt än andra och vad som är en rimlig boarea beror på flera faktorer. Resonemanget kring kopplingen mellan boarea och folkbokföringsfelet utgår ifrån att det finns en gräns för hur trångt man kan tänkas bo.

I den registerbaserade hushållsstatistiken från RTB finns ett ID-nummer för varje lägenhet. Det kan man samköra med lägenhetsregistret och kombinera uppgifter om boarea med antal personer per lägenhet.

I komponenten ingår personer som 31 december aktuellt år är folkbokförda i en lägenhet där boarean per person understiger 9 m². Boarean delas in i fem grupper och sedan beräknas ett medelvärde av antal personer i grupperna:

- < 9 m²
- < 8 m²
- < 7 m²
- < 6 m²
- < 5 m²

Stora övriga hushåll – μ

Vid en analys av vilka personer Skatteverket ändrat folkbokföring för i efterhand syntes ett mönster där hushållets storlek och om man tillhör ett så kallat övrigt hushåll spelar in. I övriga hushåll ingår bland annat hushåll där minst en person saknar en nära anhörig i hushållet.

Uppgifter om hushåll hämtas från den registerbaserade hushållsstatistiken. Där finns uppgifter om vilken lägenhet man är folkbokförd i, antal personer i samma lägenhet, hushållstyp (sammanboende med/utan barn, ensamboende med/utan barn och övriga hushåll med/utan barn) och hushållsställning (person i gift par/registrerat partnerskap/samboförhållande, ensamstående förälder, barn, övrig person).

Komponenten består av personer som 31 december aktuellt år är folkbokförda i en lägenhet där det finns minst 3 så kallade övriga personer, dvs. personer som inte har någon nära anhörig i samma hushåll. Storleken på hushållet delas in i fem grupper och sedan beräknas ett medelvärde av antal personer i grupperna:

- ≥ 4 personer i hushållet
- ≥ 6 personer i hushållet
- ≥ 8 personer i hushållet
- ≥ 10 personer i hushållet
- ≥ 12 personer i hushållet

Impulser – θ

Myndigheter och privatpersoner kan meddela Skatteverket om de misstänker att något är fel i folkbokföringen, så kallade impulser. Det kan till exempel handla om att en person felaktigt folkbokfört sig i en bostad där någon annan redan bor eller att två personer är folkbokförd i var sin bostad trots att de faktiskt bor ihop.

Komponenten består av personer som efter impulser om felaktig adress har fått sin folkbokföringsadress ändrad av Skatteverket året efter aktuellt år.

Skattat resultat av folkbokföringsfelet

Det totala folkbokföringsfelet skattas till drygt 214 000 personer 2019, vilket motsvarar 2,1 procent av befolkningen 31 december 2019. Nedan redovisas skattningarna för övertäckning, undertäckning och fel folkbokföringsadress 2013–2019.

Tabell 1. Skattning av folkbokföringsfelet 2013 - 2019

Antal personer, avrundat till närmsta 100-tal

Text	2013	2014	2015	2016	2017	2018	2019
Övertäckning	67 100	67 000	61 400	62 100	64 700	68 100	73 400
Undertäckning totalt	6 000	6 700	7 700	7 900	11 800	12 500	17 000
Sena aviseringar	1 300	1 400	1 900	1 300	3 200	1 300	4 100
Högskolestuderande men inte folkbokförda	1 100	1 100	1 000	1 100	1 100	1 100	1 100
Inkomst i Sverige men inte folkbokförda	3 700	4 200	4 800	5 600	7 500	10 100	11 800
Fel folkbokföringsadress totalt	115 200	115 500	113 100	122 000	119 000	117 400	123 900
Studentpendling	11 200	11 200	11 400	11 500	11 500	11 400	11 200
Arbetspendling	23 200	24 000	24 300	24 200	24 600	23 500	24 400
Liten boarea	35 100	36 200	36 300	40 100	40 900	39 800	38 900
Stora övriga hushåll	12 700	13 900	15 000	17 700	18 300	18 500	18 700
Impulser	33 000	30 200	26 000	28 400	23 600	24 200	30 800
Totalt	188 400	189 200	182 200	192 000	195 500	198 100	214 300

[Källa]

Referenser

Bengtsson, Tor (2015), "Resultatrapport Total befolkningssiffra", SCB BV/REG internrapport 2015-08-20

Clemen, Robert T. (1989), "Combining forecasts: A review and annotated bibliography", *International Journal of Forecasting* 5 (1989) 559-583

SCB (2015), "Övertäckning i Registret över totalbefolkningen – en registerstudie", *Befolkning och välfärd* 2015:1

Skatteverket (2018), Kvalitetsuppföljning i folkbokföringsregistret, tillgänglig på www.skatteverket.se.

SCB beskriver Sverige

Statistikmyndigheten SCB förser samhället med statistik för beslutsfattande, debatt och forskning. Vi gör det på uppdrag av regeringen, myndigheter, forskare och näringsliv. Vår statistik bidrar till en faktabaserad samhällsdebatt och väl underbyggda beslut