

Vem är kvar?

En fördjupad analys av ett slumpurval av samordningsnummer

En rapport från Skatteverket

4 augusti 2025

Rapportnummer: 2025:12

Rapportnummer: 2025:12
Diarienummer: 8-262337

Jon Olofsson
E-post: jon.olofsson@skatteverket.se
Telefon: 010-5723104

Skatteverket
Postadress: 205 30 Malmö
Telefon: 0771-567 567
Epost: skatteverket@skatteverket.se
www.skatteverket.se

Förord

Föreliggande rapport utgör ett underlag till *Nationell lägesbild över befolkningen*.¹

Syftet med denna rapport är att ge en statistisk beskrivning av de egenskaper som tydligast påverkar sannolikheten för att samordningsnummer skulle vilandeförklaras. Rapporten utforskar även hur en statistisk modell – så kallad maskininlärning (ML) – tränad på ett slumpmässigt urval skulle kunna användas för att identifiera vilka samordningsnummer som bör vilandeförklaras.

Rapporten har tagits fram vid Skatteverkets analys- och dataenhet av Jon Olofsson. Ett stort tack riktas Erik Gustavsson som läst och bidragit med idéer. Även tack till Lina Boberg på FFA som korrekturläst. Skatteverkets analys- och dataenhet svarar för innehåll och slutsatser.

Sundbyberg i augusti 2025

Henrik Hammar

Chef för analys- och dataenheten

¹ Regeringen har i regleringsbrevet för år 2025 gett Skatteverket i uppdrag att ta fram en ny nationell lägesbild som beskriver de personkretsar som finns och verkar i Sverige med eller utan tillstånd samt företeelser och risker kopplade till dessa personkretsar. Enligt uppdraget ska Skatteverket årligen ta fram en lägesbild över befolkningen, som ska kunna användas som underlag när regeringen och myndigheter prioriterar och inriktar framtida åtgärder. Se rapporten *Nationell lägesbild över befolkningen 2025* på Skatteverkets hemsida för den senaste framtagna lägesbilden.

Sammanfattning

Samordningsnummer är en identitetsbeteckning för personer som inte är folkbokförda i Sverige och som saknar personnummer. Ett samordningsnummer ska vilandeförklaras när det inte längre används, till exempel när en tillfällig tjänst eller utbildning avslutats och numrets innehavare återvänt till sitt hemland. När ett nummer vilandeförklaras förändras dess status och ska ses som en flagga för myndigheter att numrets information kan vara inaktuell. Det gör numret mindre användbart för individen samtidigt som det begränsar möjligheten till missbruk. Sedan 2021 är samordningsnumrens status tidsbegränsad och de vilandeförklaras efter 5 år om inget initiativ tas till förnyelse.

Samordningsnumrens tillförlitlighet har blivit allt viktigare i takt med att de fyller fler samhällsfunktioner. Mellan mars och juli 2024 granskade Skatteverket ett slumpmässigt urval på 1000 samordningsnummer, vilket visade att mer än hälften av dessa kunde vilandeförklaras. I och med att urvalet är slumpmässigt går det att generalisera analys baserat på detta urval till den fulla populationen av samordningsnummer.

Rapportens första syfte är att ge en statistisk beskrivning av de egenskaper som tydligast påverkar sannolikheten för att samordningsnummer skulle vilandeförklaras i det slumpmässiga urvalet. Analysen pekar på att samordningsnummer där Skatteverket är mindre säkra på personens identitet oftare är inaktiva och kan vilandeförklaras. Även samordningsnummer med en utländsk kontaktadress och som tillämpats i beskattningssyfte ökar sannolikheten för vilandeförklaring, medan nummer utfärdade av Migrationsverket och med nyligen inlämnade deklARATIONER har lägre sannolikhet. Nationalitet och kön tycks däremot inte spela någon större roll.

Rapporten utforskar även hur en statistisk modell – så kallad maskininlärning (ML) – tränad på detta urval skulle kunna användas för att identifiera vilka samordningsnummer som bör vilandeförklaras. Resultaten visar att en betydande andel nummer kan klassificeras som vilande, även om modellen ställs in på ett sätt som undviker att felaktigt klassificera aktiva nummer som inaktiva. Sammantaget pekar analysen på att Skatteverket skulle kunna effektivisera sitt arbete genom att kombinera traditionella slumpmässiga urval med denna typ av statistiska modeller. Genom att använda denna metod blir det enklare att hitta samordningsnummer som är inaktiva och kan vilandeförklaras.

De viktigaste resultaten från analysen är:

- Var och en av följande faktorer ökar sannolikhet för att ett samordningsnummer är inaktivt och kan vilandeförklaras: lägre identitetsnivå, högre ålder på numret eller att det historiskt använts för beskattningsändamål.
- Nummer där Migrationsverket är ingivare och nummer med en deklARATION *i närtid* har lägre sannolikhet att vilandeförklaras.
- Kontaktadressens karaktär spelar också en viss roll, medan nationalitet och kön inte har några tydliga effekter.
- ML-modellen indikerar att en stor andel nummer skulle kunna vilandeförklaras – ca 70 procent av samordningsnumren som tillhör den vuxna populationen. Även vid mer konservativa bedömningar kan en relativt hög andel vilandeförklaras (60 till 31 procent beroende på hur modellen specificeras).

Rekommendationer:

- Knyt perioden efter vilken samordningsnumren vilandeförklaras närmare syftet med samordningsnumret. Som ett led i detta bör man se över om samordningsnummer med lägre identitetsnivåer kan vilandeförklaras tidigare. Även samordningsnummer som skapats i beskattningssyfte skulle kunna vilandeförklaras tidigare.
- Slumpmässiga urval i kombination med AI-baserade metoder, såsom maskininlärning, utgör ett kraftfullt verktyg och bör i ökad utsträckning integreras i arbetet med att stärka Sveriges folkbokföring.

Innehållsförteckning

1	Inledning	5
2	Dataunderlag	7
3	En statistisk analys av vilka egenskaper som styr vilandeförklaringar i slumpurvalet	10
3.1	Metod	10
3.2	Resultat.....	11
4	Maskininlärning för att stärka Skatteverkets hantering av samordningsnummer	14
4.1	Modell	15
4.2	Utvärdering av modellens prestanda.....	16
4.3	Generalisering till populationen av samordningsnummer tillhörande vuxna.....	18
5	Sammanfattande slutsatser	20
6	Referenser	22

1 Inledning

Samordningsnummer är en identitetsbeteckning som tilldelas personer som inte är folkbokförda i Sverige och därmed saknar personnummer, men som ändå behöver kunna identifieras av svenska myndigheter – exempelvis vid beskattning, studier eller andra administrativa ärenden. Sedan 2023 används tre identitetsnivåer – *Styrket*, *Sannolik* och *Osäker* – som anger graden av säkerhet i identitetskontrollen bakom numret. Samordningsnummer möjliggör viktiga samhällskontakter, men skiljer sig från personnummer vad gäller rättslig status, tillförlitlighet och tillgång till rättigheter.

Till skillnad från personnummer är samordningsnummer tidsbegränsade och måste förnyas vart femte år för att förbli aktiva. Nummer som inte förnyas vilandeförklaras. Samordningsnummer ska även vilandeförklaras när de inte längre används, exempelvis efter avslutad utbildning, tjänstgöring eller annan tillfällig vistelse i landet.² Ett vilandeförklarat nummer signalerar att numret är inaktivt och att informationen kan vara inaktuell. Dessa samordningsnummer ska därför hanteras av myndigheter med ”viss försiktighet”.³

I takt med att samordningsnumrens funktion i samhället har breddats, har kraven på tillförlitliga register ökat. Ett korrekt och uppdaterat register är en viktig förutsättning för att minska risken för att identitetsnummer missbrukas, till exempel för brottsliga ändamål.

Både regering och Riksrevisionen har uppmärksammat dessa utmaningar. Regeringen har gett Skatteverket flera uppdrag för att stärka hanteringen av samordningsnummer och folkbokföring. Ett av dessa är uppdraget att årligen ta fram en *Nationell lägesbild över befolkningen*,⁴ med syfte att ge en samlad översikt av befolkningens sammansättning och av vilka som bor, vistas och verkar i Sverige. Lägesbilden ska beskriva relevanta personkretsar, belysa risker och fungera som underlag för insatser mot exempelvis felaktig folkbokföring och välfärdsbrott. Därtill har Skatteverket fått i uppdrag att säkerställa att information om folkbokföring och samordningsnummer används på ett ändamålsenligt sätt (Regeringen, 2024a), samt att utveckla det operativa samarbetet för att motverka identitetsmissbruk (Regeringen, 2024b). Även

² Detta gäller även när folkbokförda personer med personnummer utvandrar.

³ I vissa fall är det motiverat för myndigheter att använda vilandeförklarade samordningsnummer i sin handläggning. Till exempel pensionsmyndigheten skulle kunna använda vilandeförklarade samordningsnummer, för att sedan förnya numret inför en eventuell utbetalning för att se om personen fortfarande är i livet, samt samla in övriga nödvändiga uppgifter som krävs för en utbetalning.

⁴ Uppdraget att ta fram en ny *Nationell lägesbild över befolkningen* är inskrivet i Skatteverkets regleringsbrev för 2025. Den första lägesbilden publicerades under 2024 och omfattade tre personkretsar: folkbokförda, personer med samordningsnummer samt övriga relevanta personkretsar. Den senaste lägesbilden bygger vidare på detta arbete och publicerades i juni 2025 (Skatteverket, 2025).

Riksrevisionen har i en granskning betonat vikten av att myndigheter har tillräcklig kompetens och utbildning för att fastställa identitet i ärenden som rör folkbokföring och samordningsnummer (Riksrevisionen, 2024).

Mot denna bakgrund genomförde Skatteverkets folk- och fastighetsavdelning under perioden mars till juli 2024 en granskning av ett slumpmässigt urval om 1 000 samordningsnummer (Skatteverket, 2024). Ett slumpurval innebär att varje samordningsnummer i populationen valts ut med liknande sannolikhet, vilket gör det möjligt att dra slutsatser om hela beståndet baserat på analys av urvalet. Skatteverket använder slumpurval som metod vid olika typer av kvalitetskontroller och uppföljningar, eftersom det möjliggör statistiskt underbyggda bedömningar av förekomsten av fel.

Syftet med granskningen var att uppskatta andelen fel i beståndet och bedöma om samordningsnumren fortfarande var i bruk. Handläggare prövade om uppgifterna i Skatteverkets register var korrekta och om numren borde vilandeförklaras. Resultatet visade att över hälften av samordningsnumren var inaktiva och kunde vilandeförklaras.

Analysen som genomfördes i samband med slumpurvalet under 2024 lyfte särskilt fram hur olika identitetsbedömningar påverkar förekomsten av inaktiva samordningsnummer. Den första granskningen var emellertid inte heltäckande, vilket motiverar denna rapport som innehåller en mer omfattande analys av slumpurvalet, för att fördjupa förståelsen av vilka typer av samordningsnummer som bör vilandeförklaras. Med utgångspunkt i slumpurvalet kartlägger rapporten de statistiska egenskaper som kännetecknar vilandeförklarade samordningsnummer, med målet att förse beslutsfattare med ett kunskapsunderlag för framtida avvägningar kring tidpunkt och villkor för vilandeförklaring. Analysen visar bland annat att samordningsnummer där identiteten bedömts som *Osäker* eller *Sannolikt*, liksom nummer utfärdade i beskattningssyfte, oftare är inaktiva.

Rapporten undersöker även möjligheten att använda en enkel maskininlärningsalgoritm för att identifiera samordningsnummer som sannolikt är inaktiva.⁵ Modellen har tränats på det slumpmässiga urvalet och kan identifiera nummer i populationen av samordningsnummer tillhörande vuxna som uppvisar liknande egenskaper som de som vilandeförklarades i slumpurvalet. När modellen tillämpas på hela populationen av samordningsnummer tillhörande vuxna visar resultaten att en betydande andel

⁵ Med ”maskininlärning” avses här en statistisk modell som tränas på historiska data om samordningsnummer (bekräftat inaktiva eller aktiva nummer i slumpurvalet samt tillhörande attribut) för att automatiskt lära sig mönster som särskiljer inaktiva från aktiva nummer. Modellen använder sedan dessa inlärda mönster för att förutsäga sannolikheten att ett nytt samordningsnummer är inaktivt.

av de aktiva samordningsnumren skulle kunna klassificeras som vilande, även vid mer försiktiga bedömningar.

Denna rapport utgör därmed ett bidrag till Skatteverkets arbete med flera pågående regeringsuppdrag som syftar till att stärka folkbokföringen. Den fungerar dels som ett underlag till *Nationell lägesbild över befolkningen 2025* (Skatteverket, 2025) genom att belysa omfattningen och karaktären av den så kallade ”övertäckningen” – det vill säga samordningsnummer som felaktigt kvarstår inom folkbokföringen – dels som ett kunskapsbidrag till regeringsuppdragen om ändamålsenlig informationsanvändning och identitetskontroll (Regeringen, 2024a; 2024b). Det är dock viktigt att understryka att regeringsuppdragen har ett bredare anslag, medan denna rapport fokuserar på en avgränsad del av problematiken: hanteringen av inaktiva samordningsnummer. Inom ramen för regeringsuppdragen har man exempelvis uppmärksammat att samordningsnummer med styrkt identitet i vissa fall kan vara särskilt attraktiva att missbruka. Denna rapport behandlar inte den aspekten, utan avgränsar sig till frågan om hur inaktiva nummer kan identifieras och hanteras mer effektivt.

2 Dataunderlag

Det ursprungliga slumpurvalet drogs från en målpopulation bestående av den vuxna befolkningen med samordningsnummer som tilldelats efter den 18 juni 2021 – det datum då lagstiftning trädde i kraft som möjliggör vilandeförklaring av samordningsnummer. Urvalet gjordes den 31 januari 2024.

Tabellen nedan visar hur olika egenskaper hos samordningsnummer fördelar sig i detta slumpmässiga urval jämfört med en referenspopulation. Eftersom syftet med rapporten är att kunna göra uttalanden om den aktuella populationen, består denna referenspopulation av samordningsnummer som uppfyller samma kriterier som den ursprungliga målpopulationen, men som var aktuella vid årsskiftet 2024/2025. Det innebär att dessa samordningsnummer har existerat nästan ett år längre än de som ingick i den ursprungliga målpopulationen. Referenspopulationen gör det möjligt att se hur väl slumpurvalet representerar aktuella samordningsnummer.

Tabellen visar också egenskaper för hela den vuxna populationen med samordningsnummer, vilket möjliggör en bedömning av hur representativt slumpurvalet är i förhållande till samtliga samordningsnummer. Dataunderlaget i tabellen används sedan genomgående i resten av rapporten.

Analysen baseras på flera olika kategorier av variabler. För det första finns egenskaper direkt knutna till samordningsnumret: identitetsnivå, ingivare,

numrets ålder och kontaktadress. Dessa faktorer kan indikera i vilken grad numret är tillförlitligt eller hur länge det varit i bruk. För det andra förekommer mått på ekonomisk aktivitet, såsom om numret någon gång lämnat in en inkomstdeklaration (Ink 1:an) eller förekommit i en individuppgift (AGI).⁶ Slutligen finns variabler som är kopplade till personen bakom numret, till exempel medborgarskap och kön, vilka kan ge ytterligare ledtrådar om vad som driver att ett nummer vilandeförklaras.

Tabell 1. Genomsnitt för slumpurvalet, referenspopulationen och den fulla populationen av samordningsnummer

	Slumpurval	Referens- population	Hela vuxna populationen
Vilandeförklarade	0,57		
<i>Identitetsnivå</i>			
Styrkt	0,17	0,19	0,12
Sannolik	0,57	0,53	0,50
Osäker	0,26	0,27	0,37
<i>Ingivare</i>			
Skatteverket	0,55	0,52	0,29
Polis	0,16	0,17	0,10
Enskild Person	0,16	0,13	0,07
Migrationsverket	0,10	0,07	0,04
Annan/Okänd	0,03	0,11	0,51
<i>Numrets ålder</i>			
Mindre än 1 år	0,14	0,15	0,09
1 år	0,43	0,22	0,13
2 år	0,38	0,3	0,17
3 år eller äldre	0,05	0,33	0,61
<i>Kontaktadress</i>			
Svensk kontaktadress	0,46	0,39	0,33
Utländsk kontaktadress	0,37	0,42	0,26
Kontaktadress saknas	0,17	0,2	0,41
<i>Ink 1:an</i>			
Ink 1:an vid något tillfälle	0,55	0,51	0,53
Ink 1:an 2023	0,38	0,30	0,25
Ink 1:an 2022	0,28	0,23	0,19
Ink 1:an 2021	0,13	0,11	0,16
<i>Individuppgifter (IU)</i>			
IU vid något tillfälle	0,51	0,48	0,48

⁶ Individuppgifterna i arbetsgivardeklarationerna gör det möjligt att dokumentera ekonomisk aktivitet nära inpå tillfället för utvärdering. För slumpurvalet är tillfället för utvärdering den första april 2024 medan det för den fullständiga populationen utgörs av den sista december 2024. Förhoppningen är att dokumentera ekonomisk aktivitet så nära inpå utvärderingstillfället som möjligt. Variablerna kopplade till IU dokumenterar därför om det förekommit IU kvartalet innan utvärderingen, näst sista kvartalet innan utvärderingen, osv.

IU 1:a kvartalet innan utvärdering	0,17	0,11	0,10
IU 2:a kvartalet innan utvärdering	0,19	0,12	0,12
IU 3:e kvartalet innan utvärdering	0,25	0,14	0,12
IU 4:4 kvartalet innan utvärdering	0,23	0,13	0,12
	<i>Medborgarskap</i>		
Svensk	0	0,02	0,02
Internordisk	0,06	0,08	0,09
EU	0,41	0,42	0,41
Utom EU	0,41	0,33	0,29
Uppgift saknas	0,12	0,15	0,18
Kvinna	0,28	0,26	0,26
Antal observationer	999	234 849	423 415

Identitetsnivåerna har en likartad fördelning i slumpurvalet, målpopulationen och hela den vuxna populationen med samordningsnummer. De flesta nummer har en lägre identitetsnivå. I slumpurvalet har exempelvis 17 procent identitetsnivån *Styrkt*, drygt 50 procent har *Sannolik*, och 26 procent har *Osäker*.

Även variabler som speglar ekonomisk aktivitet har en liknande fördelning i de olika grupperna. Det gäller både indikatorer för huruvida samordningsnumret är kopplat till inkomstdeklarationer och till individuppgifter i arbetsgivardeklarationerna.

Vissa skillnader förekommer dock mellan slumpurvalet och de övriga populationerna. Exempelvis består slumpurvalet generellt av yngre samordningsnummer – både i jämförelse med referenspopulationen och med hela populationen av samordningsnummer. Detta är dock en naturlig följd av att slumpurvalet drogs tidigare än referenspopulationen. Även skillnaden gentemot hela populationen är att förvänta, då slumpurvalet fokuserar på en yngre del av samordningsnumren.

Fördelningen av ingivare (det vill säga den aktör som initierat ansökan om samordningsnummer) är likartad i slumpurvalet och målpopulationen, men skiljer sig från fördelningen i hela populationen. En bidragande orsak är att många nummer från 2021 och 2022 saknar uppgift om ingivare i analysdatabasen – och dessa nummer är underrepresenterade i slumpurvalet.

Sammanfattningsvis framstår slumpurvalet som representativt för referenspopulationen, trots att referenspopulationens samordningsnummer funnits under en längre tid. Det innebär att slutsatser som dras från

slumpurvalet kan generaliseras till alla samordningsnummer för vuxna som tilldelats efter den 18 juni 2021. Däremot finns vissa skillnader gentemot hela populationen av samordningsnummer.⁷

3 En statistisk analys av vilka egenskaper som styr vilandeförklaringar i slumpurvalet

I detta stycke undersöks slumpurvalet från 2024 för att identifiera vilka egenskaper som kännetecknade samordningsnumren som vilandeförklarades av handläggarna. För att möjliggöra detta tillämpas en probit-modell.

3.1 Metod

En probit-modell används för att förklara sannolikheten för att något ska inträffa, baserat på flera olika faktorer. Modellen uppskattar hur varje faktor, till exempel identitetsnivå och numrets ålder, påverkar sannolikheten för en viss händelse – i detta fall om ett samordningsnummer vilandeförklarades eller inte i slumpurvalet under 2024. Istället för att beräkna en exakt procentsats skapar probit-modellen en sannolikhetskurva som visar om olika faktorer ökar eller minskar sannolikheten för händelsen.

Eftersom en probit-modell tar hänsyn till flera variabler samtidigt kan den ge en mer nyanserad bild av hur olika faktorer samverkar och bidrar till att ett samordningsnummer vilandeförklaras. Det gör den ofta mer användbar än att till exempel enbart jämföra genomsnitt mellan olika grupper, eftersom det riskerar att dölja viktiga skillnader och orsaker. Probit-modellen fångar däremot upp detaljerna i hur faktorerna påverkar sannolikheten för händelsen.

Tabell 2 presenterar estimaten från probit-modellen. Varje kolumn motsvarar en specifik uppsättning förklaringsvariabler som syftar till att fånga sannolikheten för att ett samordningsnummer vilandeförklaras i Skatteverkets slumpurval. De redovisade estimaten ger en indikation på hur mycket varje variabel påverkar sannolikheten för vilandeförklaring, medan standardfelen (inom parentes) anger osäkerheten i skattningarna.⁸ Stjärnorna intill vissa estimat markerar vilken signifikansnivå som uppnåtts i t-testet.⁹

⁷ Notera att analysen begränsas något av att analysdata saknar uppgifter från Transportstyrelsens bilregister, trots att information om fordonsinnehav var en viktig komponent när handläggarna bedömde samordningsnumrens aktivitet i slumpurvalet.

⁸ Standardfelen som tillämpas i regressionen är robusta för heteroskedasticitet.

⁹ Noll-hypotesen är att det individuella estimatet är lika med 0.

Analysen omfattar variabler för identitetsnivå, ingivare, ålder på samordningsnumret, kontaktadressens karaktär, inkomstdeklarationer (Ink1:an), nationalitet samt kön.¹⁰

3.2 Resultat

Av resultaten framgår att identitetsnivå är en viktig förklaringsvariabel när samordningsnummer vilandeförklaras, då dessa estimat är signifikanta och positiva i samtliga specifikationer. Indikator-variabeln för identitetsnivå *Styrkt* har utelämnats från modellen och fungerar därmed som referenskategori. Resultatet betyder att samordningsnummer med identitetsnivå *Sannolik* och *Osäker* har en högre sannolikhet att vilandeförklaras jämfört med identitetsnivå *Styrkt*. Koefficienterna minskar något i storlek när modellen expanderar till att inkludera fler variabler, men förblir robusta och statistiskt signifikanta. När modellen utvärderas utifrån de förklarande variabelernas genomsnitt, har samordningsnummer med identitetsnivå *Sannolik* och *Osäker* nästan 50 procent högre sannolikhet att vilandeförklaras jämfört med nummer med identitetsnivå *Styrkt*. Med andra ord skulle ett samordningsnummer som ligger på genomsnittet i alla övriga avseenden ha en 50 procent högre chans att vilandeförklaras om det bytte identitetsnivå från *Styrkt* till *Osäker*.

Tabell 2. Estimat från Probit-regressionsmodell

	Vilande- förklarad	Vilande- förklarad	Vilande- förklarad	Vilande- förklarad
Id-nivå: Sannolik	1.10** (0.49)	0.98** (0.49)	1.24** (0.55)	0.88* (0.49)
Id-nivå: Osäker	0.94* (0.50)	0.88* (0.50)	1.13** (0.56)	0.74 (0.50)
Ingivare: Polis	0.26 (0.16)	-0.05 (0.42)	-0.35 (0.43)	-0.11 (0.43)
Ingivare: Enskild person	-0.22 (0.50)	-0.30 (0.50)	-0.15 (0.56)	-0.33 (0.51)
Ingivare: Migrationsverket	-1.16*** (0.16)	-1.10*** (0.20)	-1.31*** (0.22)	-1.06*** (0.21)
Ingivare: Annan	-0.50** (0.25)	-0.58** (0.29)	-0.81*** (0.29)	-0.57* (0.31)
Ålder: 1 år	0.27** (0.13)	0.25* (0.14)	0.42*** (0.15)	0.27* (0.14)
Ålder: 2 år	0.27** (0.13)	0.05 (0.16)	0.29* (0.17)	0.07 (0.16)
Ålder: 3 år	0.63*** (0.24)	0.30 (0.28)	0.55* (0.28)	0.33 (0.28)
Kontaktadress: utländsk		0.24**	0.02	0.28**

¹⁰ Rent formellt ställs den fullständiga modellen upp som $\Pr(\text{Vilandeförklarad}_i = 1 | X_i) = \Phi(\beta_0 + \beta_1 \text{IDnivå}_{\text{Styrkt},i} + \dots + \theta \text{Kvinna}_i)$ där Φ betecknar en standardnormalfördelning.

	(0.11)	(0.12)	(0.11)	
Kontaktadress: saknas	0.39 (0.40)	0.34 (0.41)	0.38 (0.40)	
Ink1: Vid något tillfälle	1.03*** (0.21)	0.64*** (0.24)	1.04*** (0.21)	
Ink1: 2023	-1.34*** (0.18)	-0.45** (0.21)	-1.34*** (0.18)	
Ink1: 2022	0.06 (0.14)	0.11 (0.15)	0.04 (0.14)	
Ink1: 2021	-0.28 (0.17)	-0.45** (0.19)	-0.30* (0.17)	
IU vid något tillfälle		0.29* (0.16)		
IU 1:a kvartalet		-0.39** (0.17)		
IU 2:a kvartalet		-0.31* (0.18)		
IU 3:e kvartalet		-0.22 (0.17)		
IU 4:4 kvartalet		-1.27*** (0.16)		
Nationalitet: Internordisk				-0.53 (0.82)
Nationalitet: EU				-0.11 (0.81)
Nationalitet: Utom EU				-0.20 (0.81)
Nationalitet: Uppgift saknas				0.08 (0.83)
Kvinna				0.03 (0.11)
Konstant	-0.80 (0.49)	-0.69 (0.50)	-0.75 (0.56)	-0.48 (0.97)

*p<0.1; **p<0.05; ***p<0.01

Variablerna baserade på ingivare visar blandade resultat. Här utgör indikatorvariabeln för *Skatteverket* referens för de övriga kategorierna. Samordningsnummer där ingivaren är *Migrationsverket* är konsekvent och starkt negativt associerade med sannolikheten för vilandeförklaring, och är statistiskt signifikant på 1 procent-nivån i alla specifikationer. Detta innebär att samordningsnummer där *Migrationsverket* är ingivare är mindre benägna att bli vilandeförklarade jämfört med samordningsnummer där *Skatteverket* är ingivare.¹¹ För ingivare som *Polisen* och *Enskild person* eller *Annan* är resultaten

¹¹ En möjlig förklaring är att Migrationsverket utfärdat många samordningsnummer till Ukrainare som flytt till Sverige i samband med Ryssland fullskaliga invasion. Dessa samordningsnummer var i regel i bruk och vilandeförklarades därför inte under arbetet med slumpurvalet.

mer osäkra och generellt sett inte statistiskt signifikanta, vilket kan tyda på att den typen av ingivare endast har en begränsad påverkan, eller att deras påverkan liknar den när *Skatteverket* är ingivare.

När det gäller samordningsnumrens ålder finns en tydlig tendens att äldre samordningsnummer är mer sannolika att bli vilandeförklarade. Här utgör indikatorvariabeln för nummer som är yngre än 1 år referens. Koefficienterna för dessa variabler är positiva och ofta statistiskt signifikanta. Speciellt tydligt är detta för nummer som är 3 år gamla eller äldre, även om signifikansen minskar något när ytterligare kontrollvariabler inkluderas. Förklaringen är att numrens ålder är korrelerade med benägenheten till att bedriva ekonomisk aktivitet, och denna benägenhet minskar över tid. När uppgifter från inkomstdeklarationerna inkluderas i modellen, försvinner den statistiska signifikansen för numrens ålder.

Samordningsnummer med utländsk kontaktadress har signifikant högre sannolikhet att bli vilandeförklarade jämfört med samordningsnummer med svensk kontaktadress (som utgör referensgrupp). Däremot saknas tydlig statistisk signifikans för nummer utan kontaktadress.

Uppgifter om inkomstdeklarationerna är också mycket relevanta. I modellen fångas dessa uppgifter av indikatorer som anger vilka år samordningsnumret lämnat in en inkomstdeklaration. Modellen inkluderar dessutom en variabel som visar om ett samordningsnummer någonsin lämnat in en inkomstdeklaration, benämnd *Ink1: Vid något tillfälle* i tabellen. Denna variabel har en starkt positiv och statistiskt signifikant koefficient, vilket innebär att samordningsnummer som har använts i beskattningsärenden har en högre sannolikhet att vilandeförklaras. Koefficientens storlek tyder på att denna egenskap är minst lika viktig som numrets identitetsnivå, om inte ännu viktigare.

Indikatorn för inkomstdeklaration år 2023 (*Ink1: 2023*) – det senaste året i förhållande till slumpurvalet – uppvisar däremot en starkt negativ koefficient. Ett samordningsnummer som lämnat in inkomstdeklaration år 2023 kommer att ha båda variablerna *Ink1: Vid något tillfälle* och *Ink1: 2023* påslagna samtidigt. Därmed är den totala effekten av att ha lämnat in en inkomstdeklaration 2023 summan av dessa två koefficienter. Detta innebär att ett samordningsnummer som lämnat en deklaration 2023 har en något lägre sannolikhet att vilandeförklaras jämfört med ett nummer som aldrig lämnat in någon deklaration.

Samordningsnummer med individuppgifter under det senaste året minskar sannolikheten för vilandeförklaring. Detta verkar gälla oberoende av när individuppgiften samlades in, men det är viktigt att komma ihåg att variablerna

som skildrar individuppgifterna enbart beskriver ekonomisk aktivitet från det senaste året och inte längre tillbaka i tiden som är fallet för inkomstdeklarationerna.

Slutligen är varken nationalitet eller kön statistiskt signifikanta i modellen. Dessa variabler tycks därför inte ha någon avgörande betydelse för sannolikheten att ett samordningsnummer blir vilandeförklarat. Resultatet kan tolkas som att individens bakgrund spelar mindre roll i beslutet om att vilandeförklara ett nummer. Istället framstår egenskaper relaterade till numret självt och hur personen bakom numret har agerat efter numrets tillkomst som mer avgörande.

Sammantaget visar resultaten att samordningsnumrets klassificering på id-nivå, typen av ingivare (speciellt Migrationsverket), åldern på numret samt hur samordningsnumret deklarerat historiskt är betydelsefulla faktorer som påverkar sannolikheten för att ett samordningsnummer blir vilandeförklarat. Kontaktadressens karaktär spelar också en viss roll, medan nationalitet och kön inte uppvisar några tydliga effekter i denna modell.

4 Maskininlärning för att stärka Skatteverkets hantering av samordningsnummer

Maskininlärning kan effektivisera och stärka Skatteverkets hantering av samordningsnummer, till exempel som här – i processen att identifiera nummer som kan vilandeförklaras. Genom att använda maskininlärningsmodeller baserade på historiska data kan Skatteverket automatiskt bedöma risken för att ett samordningsnummer är inaktivt och bör vilandeförklaras.

Eftersom urvalet som diskuterades i föregående stycken gjordes slumpmässigt är det representativt för hela populationen av samordningsnummer. En modell som tränas på detta underlag kan därför användas för att hjälpa Skatteverket att avgöra vilka nummer som bör vilandeförklaras.

Följande stycken undersöker potentialen för att implementera en maskininlärningsmodell och diskuterar de möjligheter och begränsningar som denna teknik medför i myndighetens arbete.

4.1 Modell

Vid valet av maskininlärningsmodell är det viktigt att ta hänsyn till att datamängden är begränsad. Av denna anledning utgår analysen från en LightGBM-modell, en metod som rekommenderas specifikt för begränsade dataunderlag.¹²

En LightGBM-modell är en typ av maskininlärningsmodell som kan användas för att lösa klassificeringsproblem, alltså när man vill förutsäga vilken kategori något tillhör, till exempel Ja/Nej, Aktiv/Inaktiv, eller som i detta fall – Vilande/Inte Vilande.

Modellen fungerar genom att bygga flera små och enkla ”beslutsträd” som steg för steg lär sig av tidigare exempel. Dessa träd samarbetar sedan för att göra en bra gissning. LightGBM-modeller är populära eftersom de är snabba, effektiva och ger pålitliga resultat, även vid begränsade datamängder.

Modellen tränas genom att kontinuerligt mäta modellens AUC-poäng¹³. Modellen tränas utifrån ett antal parametrar så som antal beslutsträd, inlärningssteg och maximalt antal blad per träd. Dessa parametrar påverkar modellens ”flexibilitet” och medför en speciell typ av avvägningar: till exempel ett lågt antal träd kan ge en svag modell, medan ett allt för högt antal träd kan leda till att modellen blir överanpassad (”overfitting”).¹⁴

Modellens förklarande variabler är inriktade på egenskaper som är direkt kopplade till samordningsnumret och dess användning, i stället för på personens bakgrund. Underlaget för modellen utgörs av identitetsnivå (Styrkt, Sannolik, Osäker), ingivare (t.ex. Skatteverket eller Migrationsverket), numrets ålder, uppgifter om numrets adress samt information från både inkomstdeklarationerna och individuppgifterna i arbetsgivardeklarationerna. Samtliga dessa variabler har beskrivits i föregående kapitel.

Dessa variabler återspeglar hur numret faktiskt används och hur det är registrerat i Skatteverkets system, vilket ger en mer direkt indikation på om samordningsnumret är aktivt eller riskerar att vara inaktivt. Nationalitet och

¹² Se exempelvis Data Cowboys (2021).

¹³ AUC står för *Area under curve* och visar hur väl modellen skiljer på två klasser (t.ex. ”inaktiv” och ”aktiv”). Ett högt AUC-värde (nära 1) innebär att modellen är bra på att skilja mellan klasserna, medan ett lägre värde (nära 0,5) betyder att modellen inte gör bättre ifrån sig än en slumpmässig gissning.

¹⁴ Modellens parametrar är framtagna genom så kallad ”gridsearch”, där olika kombinationer av hyperparametrar testas systematiskt för att identifiera den modell med högst prediktionsförmåga mätt i AUC. Den slutgiltiga modellen använder DART (Dropouts meet Multiple Additive Regression Trees) – en variant av gradient boosting som introducerar slumpmässiga bortfall av träd för att minska överanpassning. De optimerade parametrarna är ett inlärningssteg (*learning_rate*) på 0,09, 46 träd (*n_estimators*), maximalt 12 blad per träd (*num_leaves*) samt en regulariseringsparameter (*reg_alpha*) på 0,02.

kön, som snarare speglar individens bakgrund än numrets egenskaper, utslöts av två skäl: dels för att det i den tidigare analysen visat sig sakna förklaringskraft i bedömningen av vilandeförklaringar, dels för att minska risken för diskriminering och därmed göra modellen mer tillämpbar i praktiken.

4.2 Utvärdering av modellens prestanda

För att utvärdera modellens prestanda delades data slumpmässigt upp i två grupper där 85 procent av observationerna användes för att träna modellen och 15 procent för att testa hur väl modellen fungerar på ny data. Det är dock viktigt att återigen betona att urvalet är relativt litet, vilket medför en risk för att resultatet varierar mycket.

Efter att modellen tränats uppnåddes ett AUC-värde på 0,87 i träningsdata och 0,88 i testdata. Det tyder på att modellen har god förmåga att skilja mellan samordningsnummer som bör vilandeförklaras och de som inte bör det. Samtidigt finns viss förbättringspotential, då vissa nummer riskerar att klassificeras fel. I och med att AUC-värdet i testdata ligger nära det i träningsdata, är risken för att modellen övertränats (overfitting) låg.¹⁵

För att gräva djupare i modellens prestanda jämförs två olika klassificeringsstrategier. Den första antar att nummer där modellen förutser en sannolikhet över 0,5 kan vilandeförklaras. I den andra antas att nummer med en förutsedd sannolikhet över 0,8 kan vilandeförklaras. Den andra strategin är alltså mer konservativ och lägger större vikt vid att inte felaktigt klassificera ett aktivt nummer som vilande.

Vid ett tröskelvärde på 0,5 har modellen en noggrannhet (accuracy) på 84 procent. Det innebär att modellen gör rätt bedömning i 84 procent av fallen. Modellen är bättre på att förutse när ett nummer ska vilandeförklaras än när ett nummer ska förbli aktivt – F1-värde på 0,85 jämfört med 0,82.¹⁶

¹⁵ Eftersom dataunderlaget är relativt litet påverkas modellens exakta prestandamått av vilken slumpmässig delning (random state) som används vid träning och validering. De rapporterade värdena bör därför ses som representativa snarare än exakta.

¹⁶ F1-värde är ett mått som kombinerar två aspekter: 1) *precision* – andelen av de samordningsnummer som modellen klassade som vilande och som faktiskt var vilande, samt 2) *recall* – andelen av alla verkligt vilande samordningsnummer som modellen korrekt identifierade.

		Predikterade värden	
		Inte vilande	Vilande
Sanna värden	Inte vilande	56	18
	Vilande	6	70

Figur 1. Förvirringsmatris, $p > 0,5$

Förvirringsmatrisen ovan visar hur modellen presterar på testdata – där varje samordningsnummer har ett känt (verkligt) utfall. Matrisen jämför dessa sanna värden med modellens förutsägelser, det vill säga hur den klassificerar samordningsnummer som vilande eller inte vilande. De sanna positiva fallen återfinns i matrisens nedre högra ruta och representerar nummer som både i verkligheten och enligt modellen är vilande. Utav totalt 76 samordningsnummer som vilandeförklarades i testdata, återfinns modellen 70. De sanna negativa fallen finns i den övre vänstra rutan och visar samordningsnummer som korrekt förutsägs vara aktiva (alltså inte vilande). Här ger modellen korrekt förutsägelse i 56 av de totalt 74 samordningsnumren som finns i testdata.

I rutan för falska positiva (övre högra) syns de samordningsnummer som är aktiva men felaktigt förutsägs som vilande, medan falska negativa (nedre vänstra) tvärtom är vilande nummer som modellen inte lyckas upptäcka. Fördelningen mellan sanna och felaktiga klassificeringar illustrerar modellens tillförlitlighet och hjälper oss se vilken typ av fel som är vanligast. Detta är särskilt viktigt eftersom Skatteverket vill undvika att felaktigt vilandeförklara aktiva nummer, men samtidigt inte vill missa nummer som faktiskt borde betraktas som vilande.

Även om en modell totalt sett gör få fel, är det känsligt att felaktigt klassificera aktiva nummer som vilande. Andelen falska positiva bland de som predikteras

som vilande är 20 procent.¹⁷ Detta mått är centralt eftersom det fångar det fel Skatteverket vill minimera.

För att illustrera hur modellen kan justeras för att minska detta fel testas ett högre tröskelvärde (0,8) där numren klassificeras som vilande först om sannolikheten för detta är mycket hög.

Resultatet blir att noggrannheten (accuracy) faller till 74 procent, främst på grund av att fler samordningsnummer felaktigt klassificeras som aktiva (inte vilande). Detta illustreras i förvirringsmatrisen nedan, där 34 av de 76 vilande numren klassificerats som aktiva (inte vilande). Fördelen med det höjda tröskelvärdet är att antalet falska positiva minskar – endast 5 av de 74 aktiva samordningsnumren har felaktigt klassificerats som vilande. Andelen falska positiva bland de som predikteras som vilande uppgår till 11 procent.

		Predikterade värden	
		Inte vilande	Vilande
Sanna värden	Inte vilande	69	5
	Vilande	34	42

Figur 2. Förvirringsmatris, $p > 0,8$

4.3 Generalisering till populationen av samordningsnummer tillhörande vuxna

Detta avsnitt beskriver hur resultaten från ML-modellen kan användas för att uppskatta hur stor andel av de aktiva samordningsnumren som potentiellt kan vilandeförklaras. När modellens förutsägelser tillämpas på hela populationen – och inte enbart på det slumpmässiga urvalet – får vi en indikation på i vilken omfattning aktiva samordningsnummer i praktiken riskerar att vara inaktiva.

¹⁷ $1 - \text{Precision} = \frac{18}{18+70} \approx 0,20$.

Modellen kan dessutom användas som ett verktyg för riskbedömning genom att identifiera de ”kontrollvärda träffarna”, det vill säga de samordningsnummer som med hög sannolikhet är inaktiva. Dessa identifierade fall kan sedan prioriteras för en noggrannare granskning, vilket möjliggör en mer riktad och kostnadseffektiv resursanvändning. Avsnittets syfte är att illustrera hur väl modellen kan underlätta beslutsfattandet i större skala och därigenom bidra till en effektivare översyn av Skatteverkets samordningsnummer. Samtidigt är det viktigt att återigen påpeka att urvalets begränsade storlek kan medföra en viss osäkerhet kring modellens träffsäkerhet.

Vid årsskiftet 2024/2025 fanns det ungefär 423 000 samordningsnummer med aktiv status tillhörande en vuxen. Tabell 1 visar bakgrundsstatistik för denna population. Genom att strukturera uppgifterna på samma sätt som för slumpurvalet kan vi applicera modellen på hela populationen och därmed göra en förutsägelse om vilka nummer som kan vilandeförklaras.

Ett sätt att uppskatta hur många felaktiga samordningsnummer som finns är att summera sannolikheterna för alla nummer, vilket ger en ungefärlig beräkning av antalet fall som modellen klassar som positiva. En väl kalibrerad modell tenderar att ge en totalsumma som ligger nära antalet verkliga positiva fall.

När modellen används på hela populationen uppgår summan av sannolikheterna till cirka 301 000, motsvarande 71 procent av alla aktiva samordningsnummer tillhörande vuxna. Detta är en högre andel än vad som framkom i samband med slumpurvalet, där slutsatsen var att 57 procent skulle kunna vilandeförklaras. Denna skillnad kan delvis förklaras av att slumpurvalet inte är helt representativt för populationen. Till exempel saknar flera nummer en giltig adress i den fulla populationen och många nummer är äldre. Dessa egenskaper tenderar att öka sannolikheten för att modellen ska klassificera samordningsnumret som vilande.

Modellen kan göras mer restriktiv genom att höja tröskelvärdet för att vilandeförklara ett nummer. Tabellen nedan visar antalet samordningsnummer som klassas som vilande vid olika tröskelvärden. Som framgår fortsätter en stor andel av samordningsnumren att klassificeras som vilande även vid högre gränser – till exempel vid ett tröskelvärde på 0,9 klassificeras runt 35 procent av populationen som vilande.

Tabell 3. Predikterade vilandeförklaringar i den fulla populationen av aktiva samordningsnummer tillhörande vuxna¹⁸

Tröskelvärde	0.5	0.7	0.8	0.9
Antal som predikteras som vilande	320 139	295 471	253 433	129 583
Procent av den totala populationen	76	70	60	31
Procent falska positiva bland de som predikteras som vilande	20	16	11	11

5 Sammanfattande slutsatser

Analysen av det slumpmässiga urvalet visar att sannolikheten för att ett samordningsnummer är inaktivt påverkas av flera faktorer, bland annat identitetsnivå, utfärdandedatum, ingivare och deklarationshistorik. Särskilt tydlig påverkan föreligger hos nummer med låg identitetsnivå, äldre utfärdandedatum eller utländsk kontaktadress. Det är också tydligt att samordningsnummer oftare är inaktiva om innehavaren har deklarerat tidigare år men inte under det senaste.

Resultaten visar att samordningsnummer uppvisar stora variationer i aktivitetsgrad beroende på hur och varför de har utfärdats. Detta tyder på att samordningsnummer inte är en homogen grupp, trots att regelverket i dagsläget behandlar dem som om de vore det. Det finns därför skäl att överväga en mer differentierad reglering som bättre speglar detta förhållande.

Ett möjligt angreppssätt vore att tillämpa kortare giltighetstider för vissa grupper av nummer. Exempelvis skulle samordningsnummer med identitetsnivå *Osäker* eller *Sannolik* kunna vilandeförklaras tidigare än efter fem år. Det kan också övervägas om Skatteverket och Polisen i vissa fall bör ha möjlighet att utfärda nummer med kortare giltighetstid, särskilt när syftet är att i närtid möjliggöra skatteuppbörd eller lagföring av personer utan fullständig identitetsverifiering.

¹⁸ Notera att en version av denna tabell ingick i *Nationell lägesbild över befolkningen 2025* (Skatteverket, 2025). Siffrorna skiljer sig något från dem i lägesbilden, eftersom modellen har justerats för att öka dess noggrannhet.

Utformningen och genomförandet av sådana förändringar ligger utanför ramen för denna rapport. Däremot pekar resultaten på behovet av rättsliga överväganden och eventuellt informationsinsatser, med syfte att enskilda i större utsträckning själva begär vilandeförklaring.

Den andra delen av rapporten baseras på en statistisk analys med maskininlärning (ML). Modellen som tränats på det slumpmässiga urvalet indikerar att en betydande andel av de idag aktiva samordningsnumren sannolikt är inaktiva. Detta gäller även vid mer restriktiva modellinställningar som syftar till att undvika att felaktigt klassificera aktiva nummer som vilande.

Samtidigt är det viktigt att inte tolka modellens resultat som absoluta. Analysen är förenad med osäkerhet, bland annat på grund av begränsat urval och möjliga snedvridningar i förhållande till populationen. Trots detta ger resultaten ett relevant beslutsstöd för att prioritera ytterligare granskning av vissa nummer och validera modellens träffsäkerhet i praktiken.

Det bör också noteras att ett stort antal samordningsnummer kommer att vilandeförklaras automatiskt under 2026, när fem år passerat sedan det nya regelverket infördes 2021.

För att öka modellens precision kan ytterligare åtgärder övervägas, exempelvis att samla in ett större och mer representativt urval, låta modellen styra urvalet i nästa fas, eller att integrera fler datakällor. Det är samtidigt viktigt att komma ihåg att modellen inte tar hänsyn till juridiska och praktiska överväganden som krävs vid faktisk tillämpning.

Sammanfattningsvis visar analysen att AI- och ML-metoder kan fungera som ett kraftfullt komplement i hanteringen av samordningsnummer. Skatteverket har goda möjligheter att dra nytta av dessa tekniker för att ytterligare stärka kvaliteten inom svensk folkbokföring.

6 Referenser

Data Cowboys. (2021). *Which machine learning classifiers are best for small datasets*. Hämtad från <https://www.data-cowboys.com/blog/which-machine-learning-classifiers-are-best-for-small-datasets>.

Regeringen. (2024a). *Uppdrag att säkerställa att information från folkbokföringen och information om samordningsnummer används ändamålsenligt* (Regeringsbeslut Fi2024/02213, I:5). Finansdepartementet.

Regeringen. (2024b). *Uppdrag att utveckla det operativa samarbetet för att motverka identitetsmissbruk* (Regeringsbeslut Fi2024/02214, I:6). Finansdepartementet.

Riksrevisionen. (2024). *Vem där – fastställande av identitet vid statliga myndigheter* (RiR 2024:12).

Skatteverket. (2024). *Utvärdering av arbetet med slumpurval för samordningsnummer*. Intern rapport.

Skatteverket. (2025). *Nationell lägesbild över befolkningen 2025*. Hämtad från <https://www.skatteverket.se/download/18.6e1dd38d196873bc1e146e6/1750330290750/nationell-lagesbild-over-befolkningen-2025.pdf>

Postadress: 205 30 Malmö **Telefon:** 0771-567 567
skatteverket@skatteverket.se, www.skatteverket.se

