Personnummerering i Norge: Litt anvendt kodeteori og økonomi

Tor Helleseth[(1)] og Øyvind Ytrehus[(1,2)]
[(1)]Selmersenteret, Institutt for informatikk, Universitetet Bergen
[(2)]Simula Research Lab
Email: {torh,oyvind}@ii.uib.no

Abstrakt

Dagens fødselsnummerordning har begrenset kapasitet og varighet, og det er nødvendig å oppgradere ordningen. Vi beskriver de aktuelle alternativene for ny personidentifikator.

1 Motivasjon og historikk

Dagens fødselsnummersystem ble etablert i 1964 da oppbyggingen av et sentralt personregister startet. Alle som var bosatt i Norge på folketellingstidspunktet i 1960 fikk tildelt et eget nummer, og siden er alle nyfødte og bosatte fortløpende blitt tildelt fødselsnummer. Fødselsnummer tildeles i dag de som er bosatt i Norge i folkeregisterlovens forstand, de som er født i Norge, samt norske statsborgere som aldri har bodd i Norge men som skal ha norsk pass.

Fødselsnummeret er på 11 siffer og har denne oppbyggingen:

$$D_1D_2M_1M_2Å_1Å_2I_1I_2I_3K_1K_2 \tag{1}$$

der D: dato, M: måned, Å: år, I: individsiffer/personnummer, K: kontrollsiffer. I_3 viser kvinne (partall) eller mann (oddetall). I tillegg er løpenummerseriene $I_1I_2I_3$ tildelt i henhold til en vedtatt konvensjon som koder for fødsels*århundre*. Kontrollsifferne K_1 og K_2 beregnes ved en algoritme utviklet av Ernst Selmer [1]. Hensikten er å garantere at de mest vanlige feil ved registrering av fødselsnummer blir oppdaget.

Metoden for beregning av kontrollsifre (se Avsnitt 4) og metoden for å disponere nummerserier for koding av fødselsårhundre begrenser antall ubrukte og tildelbare fødselsnumre til ca. 413 pr. dag for perioden 2000-2039, og ca. 330 pr. dag for perioden 2040-2054. Dagens fødselsnummerordning er ikke utviklet for å ha kapasitet utover 2039 eller 2054, avhengig av vekst i folketallet. Estimater for vekst i folketallet indikerer imidlertid at kapasiteten i fødselsnummerordningen *kan* bli utilstrekkelig lenge før 2039.

Grunndatarapporten [2] studerer aspekter ved en omlegging av Folkeregisteret, og tar også opp problematikken rundt en videreføring av fødselsnummerordningen i fremtiden. I etterkant av [2] ble det opprettet et prosjekt i Skattedirektoratet for å komme med konkrete anbefalinger om en ny fødselsnummerordning, og som en konsekvens av dette ble Selmersenteret ved Universitetet i Bergen i januar 2012 forespurt om å foreslå, utrede,

Denne artikkelen ble presentert på konferansen NIK-2013; se http://www.nik.no/.

analysere og anbefale mulige alternativ for ny fødselsnummerordning. Dette arbeidet er dokumentert i [3, 4, 5] og oppsummeres i denne artikkelen.

Resten av denne artikkelen er strukturert på følgende måte: I Avsnitt 2 diskuterer vi kriterier for valg av personidentifikator. Avsnitt 3 har fokus på størrelse og aldersfordeling i en fremtidig populasjon. I Avsnitt 4 presenterer vi fire alternativer for en fremtidig personidentifikator. Fordeler og ulemper ved disse alternativene oppsummeres i Avsnitt 5.

2 Kriterier for valg av personidentifikator

En personidentifikator er en kort sekvens av tegn hentet fra et predefinert tegnsett eller alfabet, som skal brukes som et entydig "navn" for å identifisere en person. I denne artikkelen vil vi noen ganger bruke "identifikator" i betydningen en bestemt verdi av en slik tegnstreng, mens vi andre ganger bruker "identifikator" i betydningen "system av identifikatorer", dersom vi sammenligner slike systemer. Vi håper at det går fram av konteksten hvilken av disse betydningene vi refererer til. Dersom vi tror det er rom for tvetydighet, vil vi bruke begrepene "identifikatorverdi" og "identifikatorsystem".

I dette avsnittet vil vi diskutere forskjellige aspekt ved mulige identifikatorer, og i hvilken grad disse aspektene har betydning for valg av identifikator.

Informasjonsløs kontra informasjonsholdig identifikator

Et prinsipielt spørsmål er om identifikatoren i seg selv skal inneholde informasjon eller ikke. Dagens fødselsnummer inneholder eksplisitt informasjon om fødselsdato (Dag, måned, to siste sifre i årstall), og i tillegg kodet informasjon om kjønn og århundre.

Fødselsnummeret er utviklet med hensyn på bestemte forutsetninger om bruken. Én av disse er at registrering og bruk av identifikatoren ofte foregår offline, eller mer spesifikt uten en direkte kobling mot personregistre. Denne forutsetningen innebærer blant annet at dersom en ønsker å bruke rudimentær informasjon om individet for enkel lokal prosessering uten oppslag i registre, må denne informasjonen legges direkte inn i identifikatoren. Imidlertid er det vanlige bruksmønsteret i dag å sammenholde fødselsnummeret med data fra registre, som nesten alltid er tilgjengelig.

Hvordan bør informasjon kodes inn i en identifikator? Vi er av den oppfatning at dersom informasjon skal legges inn i identifikatoren, så bør denne informasjonen også kodes mest mulig tydelig, særlig for fødselsdato som er det viktigste informasjonselementet i mange sammenhenger. Konsekvensen av dette er: Fødselsdato bør legges inn i *klartekst*, for å gjøre den lettest mulig å lese, selv om det er mulig å finne mer kompakte representasjoner. Seks desimale sifre kan representere én million forskjellige verdier, men i hundre år er det (vanligvis!) 36576 forskjellige datoer på *DDMMÅÅ*-format. Det betyr at *prisen* for å velge et klartekst-format kontra en kompakt representasjon er maksimalt $6 - \log_{10}(36576) = 1{,}44$ desimalsymbol.

Kjønn bør kodes slik at det er tilnærmet like mange identifikatorer for menn og kvinner. Det koster i utgangspunktet én bit å kode for kjønn; det vil si at mengden av identifikatorer deles i to omtrent like store undermengder. Samtidig finnes det *omtrent*[1] like mange kvinner som menn for hver dag. Hvis det er nøyaktig like mange menn som kvinner for hver eneste dag, er det ingen kostnad forbundet med å kode for kjønn. Ettersom det vil være statistiske variasjoner fra dag til dag, og ettersom disse variasjonene vil være relativt større innenfor hver kjønnsgruppe enn i den samlete populasjonen, vil det likevel ha en viss kostnad å kode for kjønn.

[1] – men ikke eksakt, se Avsnitt 3

Århundre, i den grad det er nødvendig, bør kodes på en enkel måte. Før man bestemmer hvordan århundre skal kodes, må man bestemme et antall k av *gyldige* århundrer. Koding av århundre krever da minst $log_{10}(k)$ desimalsymbol. I praksis betyr det at hvis det for hver *DDMMÅÅ*-dato er tilgjengelig et antall C gyldige løpenumre, så vil dette tallet reduseres til (i gjennomsnitt) høyst C/k når det skal kodes for århundre. For eksempel er dagens fødselsnummerordning essensielt tenkt for bruk over to århundre, fra 1855 til 2054. C er lik 826 for dagens fødselsnummer, men på grunn av koding for århundre er det i gjennomsnitt ikke mer enn 413 nummer tilgjengelig for hver dag. Dagens fødselsnummerserier allokeres i henhold til en ad hoc-tabell. Alle nummer vil ikke bli utnyttet.

I noen land veier hensynet til datasikkerhet og personvern så tungt at man har valgt å ikke inkludere personinformasjon i identifikatoren, det vil si å bruke en informasjonløs identifikator. Dette er også hovedanbefalingen i Grunndatarapporten. En identifikator er fullstendig informasjonløs bare dersom man ut fra å observere identifikatoren ikke kan trekke ut noen informasjon om individet. Dette forutsetter også at hvert individ blir tildelt en tilfeldig ledig identifikator (det vil si at identifikatorene ikke blir tildelt i en fast rekkefølge, fordi dette ville gi utenforstående en omtrentlig informasjon om tildelingstidspunkt.)

Kapasitet

Ethvert system for personidentifikatorer med fast lengde og alfabet vil ha et endelig antall forskjellige identifikatorer med en gitt egenskap. Dette antallet vil vi kalle *kapasiteten* til systemet. For informasjonsløse identifikatorer vil kapasiteten rett og slett være det totale antall forskjellige identifikatorer.

For identifikatorer som inneholder eksplisitt informasjon, er det mer hensiktsmessig å uttrykke kapasitet med hensyn på hver spesifikk verdi av det eksplisitte informasjonsinnholdet. For eksempel bør man uttrykke kapasitet for dagens fødselsnummerordning i forhold til antall forskjellige tilgjengelige identifikatorverdier for hver verdi av tuppelet (*DDMMÅÅ*, kjønn, århundre).

Det er ikke et naturlig krav at kapasiteten skal være så stor som mulig, fordi dette vil komme i konflikt med krav til andre parametre for identifikatorer. Derimot er det viktig at kapasiteten er *stor nok* i forhold til de prognosene som er tilgjengelig for behov. Behov er diskutert i Avsnitt 3.

Feilkontroll

Registrering foregår ofte manuelt, gjerne ved håndskrift, ved verbal overføring, og eventuelt ved hjelp av sviktende korttids- og langtidshukommelse. Dette innebærer at det kan inntreffe feil. Disse feilene kan være av bagatellmessig karakter, men i noen anvendelser (for eksempel innen helsevesenet) kan feil ha katastrofale konsekvenser.

Som nevnt er dagens fødselsnummer utviklet med hensyn på en forutsetning om at registrering og bruk av identifikatoren ofte foregår offline, uten en direkte kobling mot personregistre, slik at kontroll av personopplysninger ikke er øyeblikkelig mulig eller tilgjengelig. I dag foregår registrering ofte, men ikke alltid, online. Med direkte tilgang til personregistre som vil gi supplerende personopplysninger og feilkontroll vil behovet for eksplisitt feilkontroll avta.

Ettersom online-registrering ikke alltid er mulig, vil vi likevel anbefale at man bruker en identifikator som inneholder kontrollsymboler, og som gjør det mulig å oppdage feil. I dagens fødselsnummer gir de to kontrollsymbolene mulighet til å oppdage alle enkle

substitusjonsfeil (det vil si feil i enkle symbol), alle kombinasjoner av feil i to av de siste fem symboler, og noen feil som ifølge statistisk materiale forekommer oftere enn andre (se [1]). Dagens fødselsnummer gir ikke grunnlag for å oppdage alle kombinasjoner av feil i to vilkårlige posisjoner, men Alternativ 2 og 4 (se Avsnitt 4) vil oppnå dette.

Kontrollsymboler gjør det i noen tilfeller mulig å utføre feilkorreksjon, dvs (i denne sammenhengen;) dersom et fødselsnummer er feilregistrert kan et dataprogram produsere en kort liste av sannsynlig korrekte fødselsnumre.

Som et eksempel skal vi studere en tenkt n-symbolers personidentifikator $p_1, p_2, \ldots, p_{n-1}, p_n$ med symboler som alle er hentet fra et alfabet A med $s(A)$ symboler. Uten tap av generalitet kan vi anta at alle p_i er numeriske verdier i mengden $\{0, 1, \ldots, s(A)-1\}$. Anta at $s(A) = q$, der q et primtall. Basert på symbolverdiene $p_1, p_2, \ldots, p_{n-1}$ skal vi lage et gyldig kontrollsiffer p_n slik at vi får et gyldig nummer $p_1, p_2, \ldots, p_{n-1}, p_n$. Vi velger på forhånd vekter $w_1, \ldots, w_n$ i mengden $\{1, 2, \ldots, q-1\}$ og beregner kontrollsifferet p_n ved å løse ligningen nedenfor slik at

$$S = w_1 \cdot p_1 + w_2 \cdot p_2 + \cdots + w_n \cdot p_n \tag{2}$$

er delelig med q. Når q er et primtall har dette alltid en løsning for p_n i mengden $\{0, 1, \ldots, q-1\}$. Vi lar $S \equiv i (\mathrm{mod}\ q)$ bety at S gir i til rest ved divisjon med q. Spesielt vil $S \equiv 0 (\mathrm{mod}\ q)$ bety at S er delelig med q.

Redundans innføres siden alle *gyldige* personidentifikatorer $p_1, p_2, \ldots, p_{n-1}, p_n$ har en verdi av S som er delelig med q, og derfor er bare $1/q$ av alle kombinasjoner av n symboler gyldige.

Prinsippet for feilkontroll er at alle feil som transformerer en gyldig identifikator til en tegnstreng der denne kontrollsummen *ikke* er delelig med q, vil oppdages ved ny beregning av kontrollsum.

Hva skjer om en enkel feil opptrer i posisjon nummer t, for eksempel at p_t endres til $p_t + e, e \neq 0$? Dette forandrer identifikatoren $p_1, p_2, \ldots, p_t, \ldots, p_{n-1}, p_n$ til $p_1, p_2, \ldots, p_t + e, \ldots, p_{n-1}, p_n$. Vi sjekker for feil ved å beregne sjekksummen

$$\begin{aligned} S' &= w_1 \cdot p_1 + w_2 \cdot p_2 + \cdots + w_t \cdot (p_t + e) + \cdots w_n \cdot p_n \\ &= S + w_t \cdot e \\ &\equiv w_t \cdot e (\mathrm{mod}\ q). \end{aligned}$$

Anta at q er et primtall og at ingen av vektene er delelige med q. Da vil alle enkle feil oppdages, fordi $S' = w_t \cdot e \neq 0 (\mathrm{mod}\ q)$.
Anta i tillegg at alle vektene er forskjellige. Da kan vi i tillegg oppdage alle *ombyttinger*: Dersom verdiene p_i og p_j i posisjonene i og j ved en feilregistrering blir byttet om, vil ny beregning av sjekksummenen gi

$$S' = S + (p_i - p_j) \cdot (w_j - w_i) \equiv (p_i - p_j) \cdot (w_j - w_i) (\mathrm{mod}\ q).$$

Ettersom både $p_i \neq p_j$ og $w_i \neq w_j$, er S' ulik null; dette indikerer feil.

Konklusjon om optimal bruk av ett kontrollsiffer: Dersom vi velger alfanumeriske tegn i personidentifikatoren fra et alfabet med høyst q symboler, der q er et primtall, kan vi ved hjelp av **ett kontrollsymbol** oppdage alle enkle feil og alle ombyttinger. Dette krever at alle vekter er ulik 0 og innbyrdes forskjellige, det vil si at den total lengden er mindre enn q.[2]

[2] I noen tilfeller kan vi øke lengden til å være lik q eller $q+1$. For å forenkle fremstillingen lar vi være å gå i detalj. I dagens fødselsnummer er forøvrig ikke alle vektene forskjellige. Derfor vil heller ikke alle ombyttinger oppdages.

Ved hjelp av $m(\geq 2)$ kontrollsymboler kan vi, ved riktig valg av kontrollsymboler, garantere å oppdage alle feilmønstre som berører m vilkårlig valgte posisjoner. I dette tilfellet kan vi også tillate noen vekter å være lik null. Se for eksempel [7].

Lengde og huskbarhet/brukervennlighet

Kapasitet kan lett oppnås ved å velge identifikatoren lang nok. I dag er det selvsagt ikke noe *teknisk* problem om identifikatoren er lang. Men en lang identifikator har åpenbare ulemper med hensyn på brukervennlighet, både med hensyn til det å huske egen og andres personidentifikator, og i forhold til sannsynlighet for feilregistrering.

Alfabet og alfabetstørrelse

Med begrepet *alfabet* mener vi mengden av symboler som er spesifikt tillatt i identifikatoren. I dagens fødselsnummer består for eksempel alfabetet av alle desimalsiffer, det vil si at alfabetet er {0,1,2,3,4,5,6,7,8,9}. Nedenfor vil vi se på alfabeter som består av desimalsifre i tillegg til en mengde av utvalgte bokstaver. Med "bokstav" mener vi (store) bokstaver fra "A" til og med "Z". Vi vil ikke ta med "Æ", "Ø" og "Å", fordi disse ikke er tilgjengelig på alle tastatur, og fordi disse bokstavene er grafisk like andre bokstaver. Tilsvarende bør man for eksempel unngå bokstavene "O" og "I" på grunn av grafisk likhet med sifrene "0" og "1".

Man kan tenke seg identifikatorer som har forskjellige alfabeter for de forskjellige posisjoner (for eksempel "desimaltall i de første ti posisjoner etterfulgt av én bokstav i posisjon 11"). Slike systemer brukes i noen land, og kan motiveres ut fra angivelige brukervennlighetshensyn, men det er likevel vanskelig å dokumentere en faktisk brukervennlighetsgevinst og slike systemer kommer relativt dårlig ut med hensyn til kapasitet. Derimot kan det argumenteres for identifikatorer som benytter bare et lite utvalg av symbolene i noen posisjoner, dersom disse posisjonene er tillagt et eksplisitt informasjonsinnhold, for eksempel de første seks posisjonene i dagens fødselsnummer.

Det å velge et alfabet med q symboler (for $q \geq 10$), blir derfor i vår sammenheng ensbetydende med å velge et alfabet som består av de ti desimalsifrene i mengden {0,1,2,3,4,5,6,7,8,9}, og $q-10$ bokstaver valgt fra mengden { "A", "B", ...,"Z" }. Det finnes svært mange måter å velge slike mengder på. For $q = 23$ finnes det for eksempel $\binom{26}{13} = 10400600$ måter å velge alfabetet på. Det er lett å argumentere for at mange av disse valgene vil være "dårlige", men vanskelig å påstå at ett bestemt alfabet er universelt bedre enn alle andre. I [4] har vi vist en måte å nærme oss dette problemet på, uten at vi tar med flere detaljer her.

Kompatibilitet og sameksistens med dagens identifikator

Det er ønskelig at dagens fødselsnummer skal bevares for de som har fått et tildelt. For å oppnå dette, må man konstruere en ny identifikator slik at den kan implementeres i fredelig sameksistens med den gamle. Det vil si (A) at for en gitt tegnsekvens må det gå klart frem hvilket identifikatorsystem den tilhører, og (B) at dette kravet må oppfylles selv om et begrenset antall feil inntreffer.

I tillegg er det ønskelig at den nye identifikatoren ikke får verdier som sammenfaller med brukte eller gyldige D-Nummer (se Avsnitt 3), som kan oppfattes som elementer i samme nummerserie som fødselsnumrene. D-nummer er på formen (1), der hvert symbol har samme betydning som i fødselsnummeret, bortsett fra at $D_1 = D_1 + 4$, det vil si at det legges 40 til dagverdien i dato.

Opplagte løsninger på sameksistensproblemet (i hvert fall versjon A) innebærer å bruke forskjellige lengder eller disjunkte alfabeter på eksisterende og ny identifikator. Det vil likevel være behov for å vurdere hvordan sameksistens fungerer i versjon (B) av problemet, det vil si når feil inntreffer.

En løsning på (B) kan være å beholde lengde og utvide alfabet, men i det nye identifikatorsystemet å bare tillate identifikatorer som er *tilstrekkelig forskjellige* fra alle identifikatorverdier i det eksisterende identifikatorsystem. Alternativ 2 er et eksempel på en ny identifikator som er konstruert med henblikk på slik sameksistens.

Utvidbarhet

På samme måte som dagens fødselsnummer er i ferd med å bli oppbrukt, vil en ny identifikator med endelig kapasitet også en gang i fremtiden komme i samme situasjon. Da er det en fordel om identifikatoren er valgt slik at det er lett å utvide og slik at den er kompatibel med tidligere identifikatorer.

Det å ha kontrollsymboler innbakt i identifikatoren gir også en fremtidig mulighet til å utvide kapasitet, ved at kontrollsymbolet en gang i fremtiden kan brukes til å bytte ut pålitelighet mot økt kapasitet.

Programvareendringer

Alle institusjoner som bruker dagens fødselsnummer og som har programvare som validerer denne identifikatoren ved å sjekke kontrollsymboler og datoformat eller som rett og slett henter ut informasjon fra identifikatoren, vil måtte oppdatere denne programvaren ved en endring av formatet. Det er vanskelig å kvantifisere den reelle kostnaden for slike oppdateringer, men det er rimelig å anta at denne kostnaden til en viss grad avhenger av i hvor stor grad ny identifikator er forskjellig fra den gamle.

Teknologi, datasikkerhet og personvern

I det halve århundret som har gått siden dagens fødselsnummerordning ble innført, har samfunnet gått gjennom en utvikling, både teknologisk og med hensyn på befolkningsvekst, som det var svært vanskelig å forutse på forhånd. På samme måte som det er vanskelig å gi pålitelige prognoser for framtidig befolkningsvekst, er det også vanskelig å spå hvordan teknologien vil påvirke rutiner for personregistrering.

Generelt kan man si at dersom det er ønskelig (av hensyn til datasikkerhet og personvern) å knytte for eksempel biometriske data til en personidentifikator, eller dersom man ønsker å ta i bruk identitetsbasert kryptografi, kan det være ønskelig å utvide rommet av tilgjengelige identifikatorer betydelig. Det kan bety at man ønsker en identifikator av større lengde enn det som er aktuelt nå, som benytter et stort alfabet. Varianter av Alternativene 2 eller 4 kan til en viss grad ivareta slike ønsker, men i praksis ville dette innebære en betydelig endring som ligger utenfor det som har vært vårt mandat.

3 Befolkningsmessige aspekter

Estimater for befolkningsvekst

I henhold til oppdrag har vi brukt estimater fra Statistisk Sentralbyrå [6], med særlig fokus på scenariet om *Høy nasjonal vekst*.

Populasjon for personidentifikator

Utlendinger som ikke innvandrer men som har et administrativt forhold på et visst nivå til den norske stat, får tildelt et D-nummer (se avsnittet om "Kompatibilitet...osv.") Vi har utredet en mulig ny personidentifikator også for denne gruppen [5], ettersom det er grunn til å frykte at kapasiteten på sikt vil være utilstrekkelig. Vi er også blitt bedt om, og har utredet [5], kapasitet i forhold til behov ved en fremtidig utvidelse av populasjon for personidentifikator, som da vil omfatte både de som i dag får fødselsnummer og de som i dag får D-nummer. Antallet personer (pr. år) som blir tildelt D-nummer har vokst med 8% årlig det siste tiåret. Ingen har ønsket å gi oss prognoser for fremtidig vekst for denne gruppen. I steden har vi forholdt oss til tre *scenarier*, nummerert I, II, og III. Alle forutsetter en årlig vekst på 8 % inntil antallet nye D-nummer pr. år flater ut ved henholdsvis 100.000, 200.000 og 500.000.

Når vi bare betrakter *nyfødte* i befolkningen gjelder det at antall fødsler definerer behovet for tilgjengelige personidentifikatorer for det året. Riktignok finnes det en viss ikke-uniformitet med hensyn til fordeling av fødselsdato over året (flest fødsler ni måneder etter jul!), men det er snakk om mindre variasjoner. Med innvandret befolkning og D-nummerpopulasjon forholder det seg annerledes: Behovet for tilgjengelige identifikatorer for en gitt dato *vokser* i takt med *fremtidige* hendelser som innvandring og D-nummertildeling. For å finne maksimum for dette fremtidige behovet må vi ta hensyn til både størrelse og *kjønns- og aldersfordeling* i fremtidig befolkning.

Nedenfor diskuterer vi en tilnærming til dette problemet. Spesifikt ser vi på fordeling i D-nummerpopulasjonen, men den samme problemstillingen gjelder innvandret befolkning, og det kan løses på samme måte.

La $\nu(T)$ være antall **nye** personer som blir tildelt D-nummer i år T, og la $\tau(T,F,S)$ være antall personer av kjønn S født år F som får D-nummer i år T. La videre $D_T(F,S)$ være antall personer av kjønn S født år F som er tildelt D-nummer siden ordningen startet (i 1964), observert i (utgangen av) år T. $D_T(F,S)$ er en ikke-avtagende funksjon av T, på grunn av relasjonen

$$D_T(F,S) = \sum_{t=1964}^{T} \tau(t,F,S) \tag{3}$$

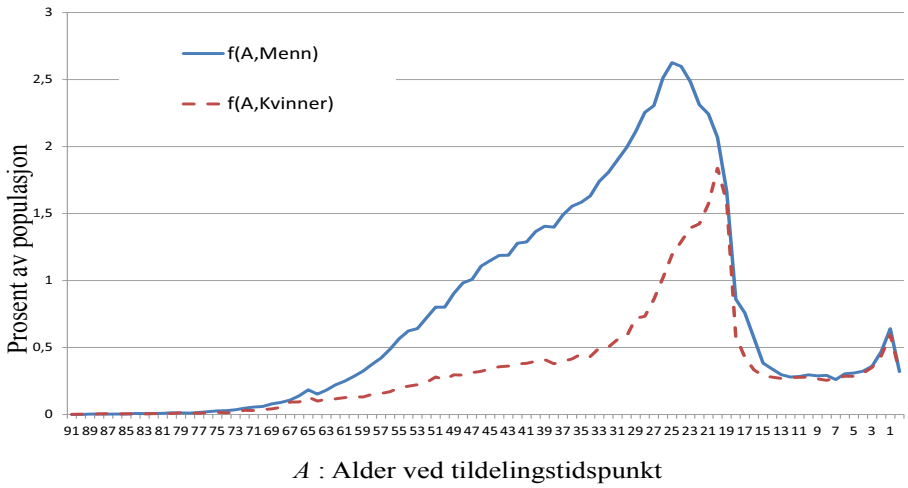
Vi kan anse systemet som sprengt dersom verdien av $D_T(F,S)$ noen gang, la oss si i år T, for et fødselsår F og for et kjønn S overstiger kapasiteten målt i D-nummer pr. kjønn pr. år. Derfor kan vi betrakte funksjonen

$$D(F,S) = \lim_{T\to\infty} D_T(F,S) \tag{4}$$

som den kritiske i forhold til kapasiteten. Både aldersfordeling og kjønnsfordeling ved tildelingstidspunkt for D-nummer er i praksis svært skjeve. La $f_T(A,S)$ være den relative frekvensen for alder A og kjønn S i tildelingsåret: Det vil si,

$$f_T(A,S) = \frac{\tau(T,T-A,S)}{\nu(T)} \tag{5}$$

I prinsippet kan fordelingen $f_T(A,S)$ variere tilfeldig fra år til år, og underliggende trender kan forandre fordelingen systematisk over tid. Statistikk viser imidlertid en forbløffende likhet i fordelingene for de tre siste årene med kjente data, det vil si at at $f_{2009}(A,S) \approx f_{2010}(A,S) \approx f_{2011}(A,S)$ for nesten alle verdier (A,S), og vi vil *anta* i videre analyse at $f_T(A,S)$ er en konstant fordeling $f(A,S)$ som er uavhengig av T og gitt ved



Figur 1: *Aldersfordeling $f(A,S)$ ved D-nummertildeling (gjennomsnitt 2009-2011)*

gjennomsnittet av $f_{2009}(A,S)$, $f_{2010}(A,S)$ og $f_{2011}(A,S)$. Denne gjennomsnittsfordelingen er vist i Figur 1.

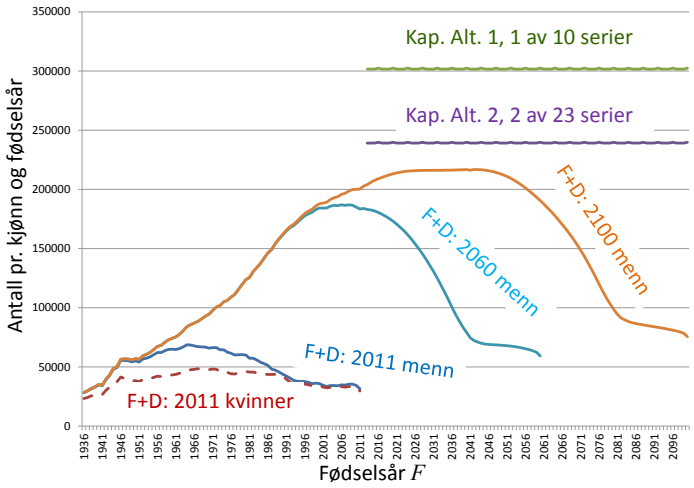
Betydningen av fordelingen $f_T(A,S)$ er at

$$\tau(T,F,S) = \nu(T) \cdot f_T(T-F,S), \tag{6}$$

og ettersom scenariene gir anslag for $\nu(T)$, kan vi på denne måten beregne et anslag for $\tau(T,F,S)$ og dermed, ved relasjonen (3), verdien av $D_T(F,S)$ for år T. Forutsetningene er altså at vi kan (A) anslå fremtidige verdier for $\nu(T)$, og (B) anta at alders- og kjønnsfordelingen $f_T(A,S)$ holder seg konstant. Forutsetning (A) er mest kritisk, og bestemmes av scenariene. Dersom årlig tilvekst i D-nummer, $\nu(T)$, vokser ut over kapasiteten, vil et (informasjonsbærende) personidentifikatorsystem kollapse. Dersom $\nu(T)$ stabiliserer seg ("flater ut"), kan vi observere at funksjonen $D_T(F,S)$ vil nærme seg sin endelige verdi for $T-S > 70$. Det kommer av at $f(A,S)$ er liten i tallverdi for $A > 70$. Dermed har vi at grenseverdien

$$D(F,S) \approx D_{F+70}(F,S) \tag{7}$$

Praktisk metode: I [5] diskuterer vi forskjellige scenarier (I, II og III) for forutsetning (A) ovenfor. For hvert scenario beregner vi et anslag for den asymptotiske verdien $D(F,S)$ anslått ved (7). Vi ser også på andre populasjoner, for eksempel den kombinerte fødselsnummer- og D-nummerpopulasjonen, med henblikk på en eventuell politisk beslutning om å tildele én personnummeridentifikator til alle i denne populasjonen, og for dette formålet definerer vi analogt alders- og kjønnsfordelingen $D_T^{(*)}(F,S)$ der * spesifiserer populasjonen. Et eksempel på resultat er vist i Figur 2, som viser (i) alders- og kjønnsfordeling i *hele den kombinerte fødselsnummer- og D-nummerpopulasjonen*, $D_T^{(\mathrm{F+D})}(F,S)$ for $T = 2011$ (tall fra Folkeregisteret), (ii) anslag for aldersfordeling $D_T^{(\mathrm{F+D})}(F,\mathrm{menn})$ for $T = 2060$ og $T = 2100$ under forutsetning om høy nasjonal vekst i folketall og Scenario II for D-nummertildeling, og (ii) kapasitet (pr. år) for de informasjonsholdige Alternativene 1 og 2, ved bruk av henholdsvis 1 (av 10 tilgjengelige) og 2 (av 23 tilgjengelige) nummerserier for å kode for fødselsårhundre. Koding for kjønn vil redusere kapasiteten, og vil kreve flere nummerserier.



Figur 2: *Fødselsårs- og kjønnsfordeling i (D+F)-populasjon,* Høy nasjonal vekst + Scenario II. *Se [5] for detaljer og andre scenarier.*

4 Alternativer for ny personidentifikator

Etter en vurdering av flere andre mulige identifikatorer endte vi opp med fire finalekandidater, nedenfor betegnet som Alternativ 1-4. Alternativ 1 og 2 inneholder samme informasjon som dagens fødselsnummer, mens Alternativ 3 og 4 er informasjonsløse. Alternativ 1 og 3 er desimale, mens Alternativ 2 og 4 er alfanumeriske, basert på et 23-symbolers alfabet.

Alternativ 0

Vi betegner dagens fødselsnummer som Alternativ 0, og tar det med som en basisreferanse.

Dagens fødselsnummer er som nevnt på formen gitt ved likning (1). Kontrollsifferne beregnes modulo 11, mens alle symboler er desimale, altså i alfabetet $A_{10} = 0,1,2,3,4,5,6,7,8,9$. K_1 blir bestemt ved at summen

$$S_1 = 3 \cdot D_1 + 7 \cdot D_2 + 6 \cdot M_1 + 1 \cdot M_2 + 8 \cdot \mathring{A}_1 + 9 \cdot \mathring{A}_2 + 4 \cdot I_1 + 5 \cdot I_2 + 2 \cdot I_3 + 1 \cdot K_1 \quad (8)$$

tvinges til å være lik $0 (\mathrm{mod}\ 11)$. Dette gir en ligning som for gitte verdier av $D_1, D_2, M_1, M_2, \mathring{A}_1, \mathring{A}_2, I_1, I_2, I_3$ gir en entydig løsning for K_1. I 1/11 av tilfellene er denne entydige løsningen $K_1 = 10$; i disse tilfellene blir løpenummeret $I_1 I_2 I_3$ forkastet. På samme måte blir K_2 bestemt ved at summen

$$S_2 = 5 \cdot D_1 + 4 \cdot D_2 + 3 \cdot M_1 + 2 \cdot M_2 + 7 \cdot \mathring{A}_1 + 6 \cdot \mathring{A}_2 + 5 \cdot I_1 + 4 \cdot I_2 + 3 \cdot I_3 + 2 \cdot K_1 + K_2 \quad (9)$$

settes lik $0 (\mathrm{mod}\ 11)$, og at ligningen løses med hensyn på K_2. Og på samme måte blir $1/11$ av de potensielle løpenumrene forkastet. Samlet betyr dette at bare $(\frac{10}{11})^2 = 82,6\%$ av alle potensielle løpenumre er gyldige, så for hver dato på formen *DDMMÅÅ* er det bare 826 gyldige valg av identifikatorverdier, som igjen er fordelt etter en vedtatt konvensjon utover århundrene 18xx, 19xx, og 20xx. Koden detekterer forøvrig alle enkle feil og noen mer spesielle feilmønstre [1], men ikke alle doble feil.

Alternativ 1

Alternativ 1 er en identifikator på formen (1) som dagens fødselsnummer, med den forskjellen at kontrollsifferet K_1 omgjøres til et nytt løpenummersiffer I_4. Dette øker kapasiteten, men svekker feilkontrollen. Kapasiteten for hver *DDMMÅÅ* er 9090, men av disse er en nummerserie på ca. 826 brukt opp (for praktiske formål) i dagens fødselsnummer, så gjenværende kapasitet er ca. 8264.

Utfordringen med Alternativ 1 er (a) å velge $I_1I_2I_3I_4$ på en slik måte at man ikke ved et uhell genererer en identifikator som er brukt før, og (b) å kode for århundre. (Koding av århundre skjer i dagens fødselsnummer gjennom bruk av første siffer I_1, som er knyttet til århundrene 18-, 19-, og 20- og som i den forstand er brukt opp. Koding av århundre gjennom valg av eksplisitte verdier for andre I_j er på samme måte ikke hensiktsmessig å gjennomføre.) Nedenfor følger en beskrivelse av hvordan disse utfordringene kan møtes.

Det som presist kjennetegner identifikatorverdier i eksisterende fødselsnummerserie, er at kontrollsymbolet K_1 er valgt slik at den tilhørende kontrollsummen S_1 er lik 0 modulo 11. En systematisk metode for å unngå *brukte* verdier er derfor i fremtiden, for Alternativ 1, å velge I_4 (tidligere kjent som K_1) slik at den tilhørende kontrollsummen S_1 er *ulik* 0 modulo 11.

For i tillegg å kunne kode for århundre på en systematisk måte, samt å beholde en viss feildeteksjonsevne, foreslår vi altså i første omgang å tildele verdier i en serie definert ved at $I_4(=K_1)$ settes slik at den tilhørende kontrollsummen S_1 er lik 1 modulo 11. Når denne serien er brukt opp kan en fortsette med å velge $I_4(=K_1)$ slik at S_1 er lik 2 modulo 11, og så videre. Hver av disse seriene har 826 identifikatorer pr. *DDMMÅÅ*; til sammen har vi derfor 8264 ubrukte identifikatorer for hver *DDMMÅÅ*. Dersom dagens langtidsprognoser for årlig befolkningsvekst er gyldig "for all fremtid" (noe som høres urimelig ut!), vil Alternativ 1 derfor kunne brukes for befolkningen (uten D-nummerpopulasjonen) i tusen år, og med implisitt koding for århundre.

Bruk av identifikatoren: Personer som har fått tildelt en identifikator etter dagens fødselsnummerordning, beholder denne. På et eller annet tidspunkt vil det ikke være flere tilgjengelige identifikatorer etter gammel ordning, og nye personer vil få tildelt et nummer fra serien med $S_1 = 1$.

Langsiktig og fremtidig opsjon. Hvis det på et fremtidig tidspunkt skulle være ønskelig å utvide kapasiteten ytterligere, er det mulig hvis det aksepteres en noe svakere feilkontroll. I_4 inngår i individnummeret og benyttes ikke som kontrollsiffer $(DDMMÅÅI_1I_2I_3I_4K_2)$. K_2 beregnes slik at den tilhørende kontrollsummen S_2 er lik 1 modulo 11. Når denne serien er brukt opp kan en fortsette med å velge K_2 slik at S_2 er lik 2 modulo 11, og så videre. En slik strategi vil vare lenge nok for mange praktiske formål, men vil svekke feilkontrollen betydelig.

Alternativ 2

Vi har foreslått [3] en alfanumerisk identifikator *DDMMÅÅAAAKK* som *Alternativ 2*, der de siste fem symbolene er fra et alfabet med $q = 23$ symboler. Med to kontrollsymboler kan vi garantere[3] at en *Alternativ 2*-identifikator som endres på vilkårlig måte i høyst to posisjoner, ikke blir forvekslet med en annen gyldig *Alternativ 2*-identifikatorverdi. *Som et tilleggskrav* setter vi at *minst tre av de fem symbolene AAAKK må være (ikkedesimale) bokstaver:* Dette tillater feildeteksjon i sameksistens med opprinnelig identifikator (eller mer presist, det vil si at feil i to posisjoner i et av dagens fødselsnumre vil ikke endre

[3]Kontrollsymbolene velges slik at Alternativ 2 blir en såkalt *Reed-Solomon*-kode. Se for eksempel [7].

det til en gyldig *Alternativ 2*-identifikatorverdi, og en feil som berører høyst to posisjoner i en *Alternativ 2*-identifikator vil ikke endre den til en gyldig identifikatorverdi i dagens system.)

Koding av århundre og kjønn kan gjøres på flere måter, for eksempel gjennom bruk av første bokstav.

Med disse begrensningene vil *Alternativ 2* ha tilgjengelig 7535 gyldige identifikatorer pr. *DDMMÅÅ* (som fremdeles må fordeles over fremtidige århundre etter behov.)

Alternativ 3

Alternativ 3 er en 11-sifret informasjonsløs identifikator på formen *IIIIIIIIIKK*, det vil si med ni løpenummersifre og to kontrollsifre, alle desimale. Kontrollsifrene beregnes på samme måte som i dagens fødselsnummer, og derfor er feildeteksjonsegenskapene grovt regnet omtrent som i dagens fødselsnummer. Dersom vi fjerner nummerserier som er brukt til eller kan forveksles med dagens fødselsnummer, D-nummer, helsenummer, samt nummer brukt av bankenes som "konstruerte kundenummer", gjenstår ca. 123.8 millioner ubrukte gyldige identifikatorverdier [5].

Alternativ 4

Vi har også foreslått et alfanumerisk informasjonsløst Alternativ 4, som er en 8-symbolers identifikator på formen *AAAAAAKK*, der de seks første symbolene er valgt fritt og de to siste er kontrollsymboler. Alle symbolene er valgt fra et alfabet med $q = 23$ symboler. Kontrollsymbolene er valgt i henhold til en Reed-Solomon-kode (se fotnote under Alternativ 2) og derfor kan vi oppdage alle kombinasjoner av inntil to feil. Dette alternativet har godt og vel 148 millioner tilgjengelige identifikatorverdier [3].

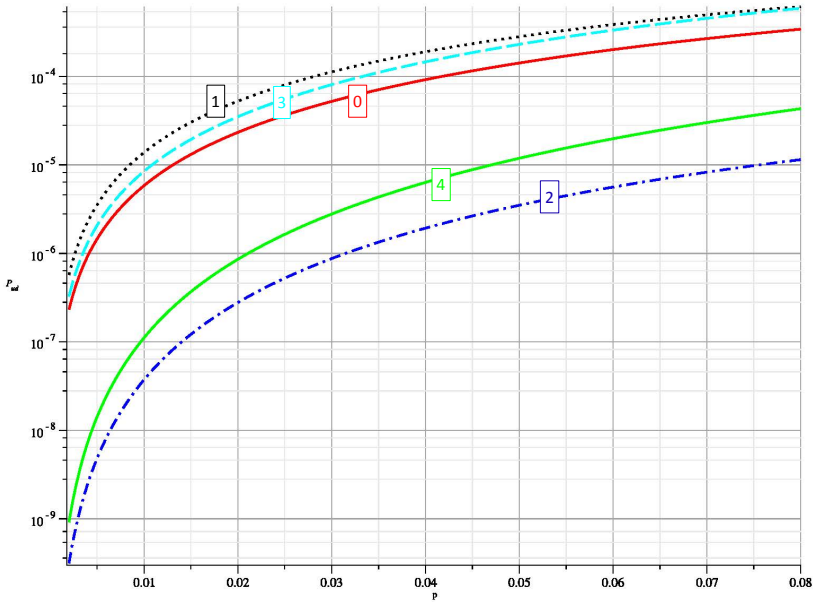
5 Diskusjon av alternativene

Alle alternativene (med unntak av Alternativ 0, dagens fødselsnummer) er valgt slik at de under rimelige forutsetninger vil oppfylle behovene for minst hundre år fremover, og antagelig mye lenger. Som en oppsummering diskuterer vi nå andre aspekter som er viktig med hensyn på valg av ny identifikator.

- Informasjonsløs kontra informasjonsholdig: Dette har prinsipielle sider. En informasjonsløs og alfanumerisk identifikator ville kanskje være det opplagte valget dersom vi skulle starte fra nytt i dag, men en prinsipiell konsekvens av et slikt valg ville nok være at hele populasjonen får tildelt ny identifikator, til en uberegnelig kostnad.

- Økonomi: I dagspressen har det fremkommet anslag fra én til et tosifret antall milliarder kroner for kostnaden ved en omlegging, avhengig av alternativ. Kvantifisering er vanskelig, men vi kan peke på at en informasjonsløs identifikator forandrer mange grunnleggende forutsetninger for eksisterende programvare, som i dag henter ut og bruker innbakt informasjon. Andre (og mindre omfattende?) behov for omprogrammering er knyttet til feilsjekking og gyldig alfabet. Vi antar at en rimelig rangering av alternativene, fra lavest til høyest pris, vil være 1, 2, 3, 4.

- Feilkontroll: Sannsynlighet for uoppdaget feil er vist i Figur 3, beregnet for en enkel *diskret minnefri kanal* der hvert enkelt symbol er feil med uavhengig sannsynlighet p. Se [3].



Figur 3: *Sannsynlighet for uoppdaget feil for Alternativ 0, 1, 2, 3 og 4. Se detaljer i [3].*

- Utvidbarhet: Kontrollsymboler kan i fremtiden omgjøres til løpenummersymboler. Derfor er Alternativ 2, 3, og 4 mer utvidbare enn Alternativ 1, og effekten er særlig stor for de to alfanumeriske Alternativene, 2 og 4.

I en helhetsvurdering er det som vanlig økonomien som teller.

Referanseliste

[1] Ernst S. Selmer, "Personnummerering i Norge: Litt anvendt tallteori og psykologi," *Nordisk Matematisk Tidsskrift,* 36-44, 1964.

[2] *Grunndatarapporten,* tilgjengelig på http://www.regjeringen.no.

[3] Tor Helleseth og Øyvind Ytrehus, "Rapport om egenskaper ved ny personidentifikator," Teknisk rapport for Skattedirektoratet, April 2012. Revidert januar 2013, 33 sider.

[4] Tor Helleseth og Øyvind Ytrehus, "Rapport om egenskaper ved ny personidentifikator: Del 2," Teknisk rapport for Skattedirektoratet, September 2012. Revidert januar 2013, 13 sider.

[5] Tor Helleseth og Øyvind Ytrehus, "Rapport om D-nummer og økt populasjon for personidentifikator," Teknisk rapport for Skattedirektoratet, Februar 2013, 38 sider.

[6] Statistisk sentralbyrå, http://www.ssb.no/folkfram/

[7] S. Lin and D. Costello, *Error Control Coding: Fundamentals and applications, 2nd Ed.*, Prentice-Hall, 2004.