Hovedoppgave i samfunnsøkonomi for cand.polit-graden

Kvantifisering av feilkilder i norsk befolkningsstatistikk

En pilotstudie i automatisk lenking av kirkebokinformasjon

Reidar P. Lavik

Mars 2007

Økonomisk institutt

Universitetet i Oslo

Forord

Jeg ønsker å takke Kåre Bævre ved Økonomisk institutt for godt samarbeid, god veiledning og for tilretteleggingen av kildedata for min hovedoppgave.

Bærum, 28. januar 2007
Reidar P. Lavik

Til min far
En stor takk til deg for alle intellektuelle bidrag og gode diskusjoner gjennom årenes løp.

Innholdsfortegnelse

1. Innledning

Kirkebøkenes informasjon om fødsler, dødsfall og giftermål er en av de viktigste kildene til tidlig norsk befolkningsstatistikk. Den offisielle nasjonale statistikken for disse demografiske begivenhetene er frem til år 1875 basert på prestenes årlige innrapportering av opptellinger fra kirkeboken. I de senere år har utstrakt digitalisering av kirkebøkene også gjort det stadig mer aktuelt å utarbeide revidert og mer omfattende statistikk ved mer detaljerte tabuleringer av selve kirkebøkene.

Kirkebøkene fra før år 1875 har flere mangler som informasjonskilde, særlig nevnes feil i angivelse av personalia, og sosiale skjevheter i hvem som ble registret i forskjellige studier. Jeg ønsket særlig å belyse problemet med under- og overrapportering av barn som dør unge eller som dødfødte, det er velkjent at disse ofte disse enten bare ble innført som døde eller som døpte. For å kvantifisere og klassifisere disse feilkildene er det nødvendig å lenke personer på tvers av kirkebøkenes dåps- og begravelseslister. En automatisert datalenking for et større utvalg av prestegjeld og sokn har behov for automatiserte metoder og algoritmer for lenking av nominative registreringer.

Utgangspunktet for denne oppgaven var derfor å måle denne formen for under- og overrapporteringen. Ved å velge et utgangspunkt i data for personer som døde unge var begrunnelsen å minimere problemet ved migrasjon, siden data i dåps- og begravelseslistene vil være ufullstendige ved at en person flytter inn eller ut i tidsrommet mellom dåp og begravelse. Siden det ikke foreligger noen automatisk maskinell[1] rekonstruksjon av persondata av større datamengder i Norge, kan oppgaven ses på som et pilotprosjekt for å gi noen svar på om norske

[1] Det har blitt gjort manuelle lenkinger og rekonstruksjoner av norske data, blant annet i Rendalen, Etne og Asker.

kirkeboksdata faktisk er gode nok for å benyttes til slik rekonstruksjon av persondata.

En rekonstruksjon av persondata med utgangspunkt i dåpsdata og begravelsesdata bygger på felles sammenliknbare felt. Problematikken rundt denne sammenlikning av data, ligger i det faktum at datasettene ikke har noen felles unike identifikatorer. En automatisk datalenking er en integrasjon av informasjonen fra to ulike datakilder. Data fra de to kildene antas å relatere til det samme individ i på en slik måte at det kan tolkes som en enkel post i datasettet. En slik datalenking, enten den er maskinell eller manuell, kan ikke under noen omstendighet gjenfinne 100 % av personene i en av listene, til det er tilstedeværelsen av feilkilder både i datakildene og i de standardiserte datakildene for høy.

For å kunne gjøre en analyse om hvorvidt datagrunnlaget faktisk kunne benyttes, var det behov for dataprogram for samordning av data og søk. Et dataprogram med en innebygget logikk for testing og klassifisering av mulige datalenker. I tillegg var det behov for en mulighet for modifikasjon og justering av beslutningsvariable og andre kriterier i prosessen. Et slikt dataprogram fantes ikke tilpasset norsk befolkningsstatistikk. Jeg valgte derfor å utvikle et dataprogrammet med logikk for filtrering av data og flere algoritmer for sammenlikning av de filtrete data.

Kildedata som kan benyttes i dataprogrammet er bearbeidete og standardiserte data fra 14 norske prestegjeld og sokn. Felt i datasettene som kan identifisere prestegjeldet eller soknet er kryptert og anonymisert med hensyn til hvilket prestegjeld data stammer fra. Standardiseringen av de opprinnelige kildedata er foretatt av Kåre Bævre (Økonomisk institutt, UiO), han har utviklet og videreutviklet et sett med algoritmer og metoder for sortering, standardisering og bearbeiding av kildedata. Jeg valgte å ta utgangspunkt i data fra begravelseslistene og lenke sammen med dåpslistene, siden disse innholdt færre data år for år.

Datalenking ble introdusert som metodikk innen statistisk befolkningsanalyse av Dunn (1946) og videreutviklet for bruk i datamaskinbasert sammenlikning av Newcombe (1959). Det skilles i teorien mellom deterministisk- og probabilistisk datalenking som metoder for datalenking. Hovedteoremet og det teoretiske rammeverket rundt probabilistisk datalenking stammer fra Fellegi og Sunter (1969). Dette rammeverket danner en bakgrunn for min oppgave, uten at selve modellen kan benyttes aktivt i min datamodell for norske data.

Oppgaven behandler i kapittel 2 en generell oversikt over hovedkilden til mitt datagrunnlag, kirkeboken, som sammen med folketellinger er det viktigste grunnlaget innen slektsforskning. I kapittel 3 gir jeg kort oversikt over de standardiserte datakildene, som danner grunnlaget for min datamodell og automatiske datalenking. Kapittel 4 tar seg datalenking som en metode for å lenke sammen data fra to datakilder. Videre ser jeg på ulike strategier for å minimere feilkilder og informasjon.

De viktigste implikasjonene av den teoretiske modellen til Fellegi og Sunter vises i kapittel 5. Datamodellen, prosessene og programlogikken gjennomgås sammen med et eksempel i kapittel 6. Resultatene i fra den empiriske analysen og tre tester for validering av mine algoritmer presenteres i kapittel 7 sammen med en oppsummering av potensielle feilkilder. Problematikken med feilkilder i datakildene og i datamodellen diskuteres i kapittel 8. Oppgaven avsluttes i kapittel 9 med en noen avsluttende bemerkninger og oppsummering av de viktigste konklusjonene. Tabeller og figurer som kun omtales i oppgaven er lagt under appendiks.

2. Kirkeboken

Kirkebøker er sammen med folketellinger den viktigste kilden innen slektsforsking (RHD, 2006a). Kirkeboken eller ministerialbok føres av soknepresten. Den inneholder som regel alle kirkelige handlinger (ministerialia), som dåp, konfirmasjon, vigsel og begravelse, innenfor et angitt tidsrom.

Kirkebøkene skal etter Kgl.res. 10. juni 1837 etter et gitt tidsrom avleveres fra den lokale kirken til Arkivverket (Riksarkivet og Statsarkivet). Dette gjaldt alle papirdokumenter og protokoller som var blitt så gamle at de ikke lenger var til nytte ved embetsførselen, men som i antikvarisk, statistisk eller historisk henseende måtte anses å være av noen interesse. Denne avleveringen foregikk etappevis fram til 1950-årene.

2.1 Bakgrunn

I de sørlige deler av Europa finner man kirkebøker helt tilbake til 1300- og 1400-tallet, og i Danmark begynte man føring av kirkebøker allerede på slutten av 1500-tallet. I Norge er den eldste bevarte kirkeboken fra Andebu med start av innføring i år 1623.

Under Kirkeritualet av 1685 ble det bestemt at det skulle føres en oversikt over dåp, giftermål og gravferd i hvert sokn. Denne bestemmelsen ble fulgt opp ved lov i år 1687, som trådde i kraft året etter. Det var etter dette påbudet om føring av kirkebøker, at flertallet av prestegjeldene kom i gang på 1700-tallet. Siden det ikke var satt en felles standard for oppsett av datafelt varierte dette fra sokn til sokn. Det var opp til den enkelte prest å dele boken eller bøkene mellom de kirkelige handlingene som dåp, begravelse og giftermål, og bevare en viss kronologisk føring. Denne klare mangelen av uniformitet i føringen av kirkeboken medførte at

kirkebøkene fra denne tiden har klare begrensninger som gode kilder (Thorvaldsen, 1996).

I 1812 kom det et reskript om at kirkebøker skulle føres på et ferdig trykt skjema med rubrikker med datafelter[2]. I tillegg til dette forsøket om standardisering av innskrivningen av de kirkelige handlingene, skulle bøkene nå føres in duplo, som medførte at både klokker og prest skulle føre ned den kirkelige handlingen i hver sin bok. Reskriptet i 1820 understrekte dette ønsket om en klarere uniformitet i føringen av kirkebøkene, og medførte noen små endringer i skjemautformingen og det faktum at klokkerene var ikke lenger forpliktet til å føre in duplo, men de fortsatte allikevel mange steder.

Det tok tid å innføre nye standarder, og mange prester valgte å skrive ut den gamle kirkeboken istedenfor å ta utgiften med å kjøpe ny. Ved midten av 1800-tallet vendte de fleste prestene tilbake til kirkebøker med trykt skjema. Det var på denne tiden blitt enklere å få kjøpt trykte ark hos en del større papirforhandlere som hadde spesialisert seg på framstilling av forretningsbøker og protokoller.

Standarden fra 1820 ble i all hovedsak stående til de nye kirkebokskjemaene fra 1877 kom etter en Kgl.res. 13. juli i 1877. Dette var mye i bakgrunn av et press fra statistikerne i Indredepartementet med ønske om å benytte det statistiske datamaterialet. Disse skjemaene er med kun små endringer fremdeles gjeldene (RHD, 2006b).

[2] For oversikt over hvilke felter som var påkrevd ved registrering, se tabell 24 og tabell 25 i appendiks.

2.2 Digitalisering

Registreringssentral for historiske data (RHD) har siden 1981 arbeidet med å registrere gamle kirkebøker digitalt. Formålet for RHD er:

> *"å gjøre historisk kildemateriale, som folketellinger og kirkebøker lettere tilgjengelig ved at man dataregistrerer originalmaterialet og produserer alfabetiske registre i bokform" (RHD, 2006b)*

Mangelen av en felles standard for registrering og utveksling av data, gjorde at det i Norge på slutten av 1980-tallet ble utarbeidet en felles registrerings- og utvekslingsstandard, kalt Histform, dette var et resultat av et initiativ fra Landslaget for lokalhistorie. Histform standarden angir en felles post- og feltstruktur som skal benyttes når man registrerer data fra kirkebøker. Histform har i de senere år blitt videreutviklet og forbedret (Kyrre) i et samarbeidsprosjekt mellom DIS-Norge og RHD.

> *"Registreringsstandarden består av en fastlagt post- og feltstruktur for datafilene for hver av de ulike kildetypene (1865, 1870, 1875, 1885, 1891, 1900 og 1910), og ett felles sett med anbefalte registreringsinstrukser for disse tellingene. Disse standardene tar hensyn til kildetrohet og effektivitet i registreringsarbeidet. Utvekslingsstandarden består av en fastlagt post- og feltstruktur for hver kildetype, ikke nødvendigvis identisk med den tilsvarende i registreringsstandarden, samt regler for visse dataverdiers utseende. Utvekslingsstandardene er mer generelle og skal gjøre det enklere anskaffe data sammen med ulike typer programvare som bl.a. er i bruk blant norske historikere. Standarden vil også gjøre det lettere utnytte ulike datasett i komparativ forskning. Som grunnlag for standarden er det benyttet en mengde kopier av de originale folketellingskildene, Statistisk Sentralbyrås instrukser for folketellingsarbeidet da det ble utført, og dokumentasjon av hvordan folketellingsmaterialet siden er blitt registrert, lagret og*

viderebehandlet med datamaskin ved de forskjellige institusjoner her i landet." (RHD, 2006b)

I Norge er mye av den lokale slektsforskningen og arbeid med gamle kilder samlet rundt aktivitetene til DIS-Norge[3]. Den digitale og manuelle transkriberingen krever noen grunnleggende regler for at data kan benyttes som et utvekslingsformat mellom forskjellige digitale verktøy.

2.3 Dataformater i registreringen.

Samarbeidet mellom DIS-Norge og Arkivverket har medført utarbeidelsen av noen grunnleggende regler og føringer for transkriberingen av kirkeboksdata. Hovedpunktene er kildetrohet og metoder for arbeid med etternavn og datoer (DIS-Norge, 2006).

Kildetrohet. Kildetroheten har som nevnt vært en av komiteens fremste ledestjerner, men det anbefales likevel å løse opp på dette prinsippet i de situasjoner der følgende tre forutsetninger er oppfylt (i det minste 1+2 eller 1+3):

1) Brudd på kildetroheten synes ikke å redusere opplysningenes informasjonsverdi for noe formål.

2) Brudd på kildetroheten høyner den maskinlesbare kildeversjonens brukskvalitet og -potensial.

3) Brudd på kildetroheten øker registreringshastigheten vesentlig. Som antydet vil det alltid være et spørsmål om skjønn når man skal avgjøre om kildetroheten i visse faste situasjoner bør fravikes eller ikke. Komiteen har valgt å anbefale brudd bl.a. i følgende

[3] DIS-Norge (Databehandling i Slektsforskning) er Norges største slektsforskerforening med ca. 7500 medlemmer per 2005, og har 19 *lokallag* som med noen få unntak dekker hvert sitt fylke.

situasjoner: Endelsene i alle patronymikon-lignende etternavn forkortes under registreringen til "s." eller "d.". Unntatt er navn som ender på "søn", "zen" eller de svenske "son" og "dotter", som registreres fullt ut. Forkortelsene sparer skrivearbeid, samtidig som det blir enkelt å skille navneformene maskinelt under konverteringen.

Etternavn. Personenes oppgitte etternavn fordeler seg på to vesensforskjellige typer, nemlig patronymika (farsnavn) og slektsnavn (familienavn). For enkelte anvendelser av dataene vil det være hensiktsmessig å kunne behandle de to typene hver for seg, f eks når man vil knytte forbindelser mellom personer og deres fedre i et materiale med historiske persondata. En nærliggende løsning er å fordele de to typene etternavn på hvert sitt felt i alle formatene, men navn som ender på "-sen" skaper problemer, fordi de kan tilhøre begge typer. En eventuell oppsplitting vil være svært komplisert, og vil aldri kunne bli 100 % historisk korrekt, enten den utføres manuelt under registreringen eller helt eller delvis maskinelt i konverteringsfasen. Komiteen har bestemt seg for å beholde etternavnene samlet i ett felt i registreringsformatet, og ved overgangen til utvekslingsformatet splitte dem opp maskinelt i patronymika og andre etternavn/slektsnavn utelukkende på grunnlag av navnets endelse ("-sen", "-datter" m fl). Denne løsningen er ikke fullgod, men den er vurdert å være bedre enn ikke å foreta noen oppsplitting i det hele tatt (RHD, 2006b).

Datoer. I folketellingene er personenes fødselsdato i 1910-tellingen den eneste fullstendige datoen som er systematisk oppgitt. Men fordi man her ønsker samme instrukser som for de senere kirkeboksformatene, har formen på datoene vært ivrig diskutert likevel. Bruk av tre separate felt for dag, måned og år har vært vurdert, sammen med ulike ett-felts former med og uten punktum mellom tallene. Valget falt til slutt på formen 'dd.mm.åååå' i registreringsformatene, for å oppnå naturlig

registreringsrekkefølge og best mulig lesbarhet. I utvekslingsformatet ble formen 'åååå.mm.dd' valgt, først og fremst for å oppnå enkel sortering. (jfr. de felles registreringsinstruksene.)

Hovedregel: Fødselsår registreres med fire siffer. Hele fødselsdatoer registreres med åtte siffer og to punktum, dvs. på formen 'dd.mm.åååå'. Presiseringer, unntak og eksempler:

a) Dersom kun de to siste sifrene i fødselsåret til en person er oppgitt, skal fødselsåret allikevel registreres fullt ut med fire siffer. (De to første sifrene, som oftest ”18”, kan imidlertid gjerne forhåndsutfylles eller påføres automatisk av registreringsprogrammet.)

b) Det er ikke nødvendig å registrere dagnummer eller månedsnummer mellom 1 og 9 med en ledende null for å oppnå to siffer. En dato på formen ”d.m.åååå” er følgelig fullt ut akseptabel.

c) Dersom kun måneden, men ikke dagen, i en dato er oppgitt, registreres dagen som ”0” eller ”00”, f eks ”00.07.1885”. Det samme gjelder dersom dagen, men ikke måneden, er oppgitt, f eks ”22.0.1885”. Det siste vil forekomme meget sjelden.

d) Fødselsåret eller -datoen er oppgitt for så å si alle personer i de nevnte tellingene. I de få tilfellene der verken år eller dato er oppgitt, registreres mangel-symbolet '!!' alene i feltet, for å vise at opplysningen ikke er uteglemt under registreringen.

e) Ved bruk av usikkerhetstegn i tilknytning til dagen eller måneden i en dato, må usikre dag- eller månedsnummer under 10 registreres med ledende null for å oppnå entydighet, f eks '03??.5.1844'. Andre eksempler på bruk av usikkerhetstegn: '18??' (de to siste sifrene i

årstallet usikre), '188??' (siste siffer i årstallet usikkert), '1885??' (hele årstallet usikkert), '??.07.1885' (dagen usikker), '??2.07.1885' (første siffer i dagnummeret usikkert), '22.0??.1885' (andre siffer i månedsnummeret usikkert), '22.07??.1885' (måneden usikker), '22.07.1885??' (årstallet usikkert) eller '22??.07??.1885??' (hele datoen usikker).

f) Dersom en dato er korrigert ved overstrykning, registreres dette f.eks slik '11.3 %8%.1894' (månedsnummeret 8 er strøket over og erstattet med 3).

2.4 Standardiseringssystemer

For å minimere feil ved registrering og transkribering av digitaliserte data er det utarbeidet metoder for å standardisere navneformater. I Norge benyttes blant annet systemene Foneq og Fondef. Foneq er et standardiseringssystem og Fondef et formal språk for spesifisering av transkripsjonsregler i navnestandardisering (Nygaard, 1992). I Norge benyttes standard 4G og Kyrre som retningslinjer for transkribering.

Internasjonalt benyttes forskjellige standarder, Soundex er en fonetisk algoritme som indekserer navn fra engelsk uttale. Særlig på New Zealand, Australia og Canada har forskningsmiljøene, innenfor helsestatistikk kommet langt innen utvikling av teoretiske rammeverk og automatiserte metoder for lenking av persondata.

2.5 Kvalitativ vurdering av informasjonskilden

Data fra kirkeboken har vært igjennom tre prosesser før den blir benyttet som data i min oppgave, det kan ha oppstått feil eller mangler i alle eller noen av prosessene. Det kan være mangler ved selve føringen av

kirkeboken, det kan ha skjedd feil ved overføringen av data fra kirkeboken eller det kan være feil og mangler ved de standardiserte datasettene.

2.5.1 Mangler ved føring av kirkeboken

Mangel av en uniform standard for føring av kirkebøkene preger datamaterialet fra før 1820, det er stor forskjell både i antall bevarte bøker og datamaterial i ulike sokn. Det var helt opp til den lokale presten hvor mye personalia han ønsket å registrere, dette kunne medføre en underregistrering av ministerialdata. Særlig kan man se i perioden før 1820 at det er en underregistrering av data i begravelser for kvinner og spedbarn og i noen tilfeller helt mangel av innføringer. Underrapportering kan også være et betydelig problem, særlig med tanke på spedbarn, dødfødte og de som dør i løpet av den første leveuken.

Et annet problem er typologiske feil ved føring av data i kirkeboken, særlig gjelder dette for fornavn og farsnavn. En person kunne bli døpt som Andres og gravlagt som Anders. Typologiske feil gjør seg også gjeldene ved innskrivning av siffer. Sifferene 1, 7 og 9 kan lett byttes om ved registrering av dager, måneder og år.

På 1700- og 1800-tallet var de kirkelige handlinger, som dåp og begravelse, knyttet sammen med utgifter. Det kan godt tenkes at fattigfolk begravet dødfødte uten dåp og at en begravelse kunne skje eller bli registrert lang tid etter fødselen. Det også var en underrapportering i begge registre av eldre mennesker uten lokal familie.

Migrasjon, flytting mellom ulike sokn, gjør seg utslag i ikke lenkbare poster i dåpstabell eller i begravelsestabell. Ved begravelser og dåp for personer som ikke er bosatt og registrert i prestegjeldet eller soknet, vil det også resultere i en registrering som ikke kan lenkes.

2.5.2 Overføring av data fra kirkeboken

Overføring av data fra kirkebøkene skjer manuelt ved transkribering (avskrift), skanning av original kilden, skanning av mikrofilm, ocr-behandling og dataregistrering med et egnet dataverktøy i en database. Transkriberingsfeil vil forekomme ved alle manuelle former for avlesning eller inntasting av data. I Norge benyttes dataverktøy som BD87, Augustus og Augustus2 i registreringsarbeidet. Benyttelse av dataprogrammer med ulike standarder kan det medføre tap av data og/eller feil ved overføring data. En stor del av overførte data har blitt lagret i tabellform i databaseverktøyet Access og noe i regneark i Excel.

Kirkebøkene var ført for hånd og de eldste kirkebøkene var ført i gotisk skrift. I mange tilfeller kan de være vanskelig å tolke og transkribere for personer uten den rette kompetanse, dette har antageligvis økt graden av feilføringer i overføringen av data fra kirkeboken.

2.5.3 Behandling av overførte data

De overførte datasettene har videre blitt grovrenset manuelt for klare mangler og feil. Den viktigste standardiseringen av datakildene blir allikevel gjort maskinelt. Ulike metoder og algoritmer benyttes i denne prosessen, og de bearbeidete utdata kan innholde forskjellige følgefeil som følge av prosesseringen og standardiseringen. Hovedgrunnen for standardiseringen av datakildene er å redusere andelen feil som kan komme av innskrivingen eller transkriberingen.

3. Kildedata

Datakildene for min oppgave er to datasett som er ferdig standardisert og bearbeidet med begravelsesdata og fødselsdata fra flere sokn/prestegjeld. Jeg vil se nærmere på struktur, registrering, standardisering og feilkilder under omtalen av poststrukturen punkt 3.4. Datasettene er bygd opp med

felt og datastruktur etter retningslinjene gitt i Histform/Kyrre.

Datakildene er ikke endret etter registrering, lagring og maskinell bearbeiding. Dataene er også blitt kryptert og anonymisert slik at det ikke skal være mulig å se hvilket sokn/prestegjeld posten i datasettet har sin opprinnelse i. Den geografiske dimensjonen har i min oppgave vært uinteressant, derfor har ikke den geografiske tiknytning med hensyn til prestegjeld vært et poeng.

3.1 Organisering

Registreringen av data i kirkeboken ved dåp og ved begravelse var forskjellig. **Tabell 1**[4] viser hvilken standard datastruktur som var påkrevd ved registrering av dåp og begravelse, men som nevnt innledningsvis ble det ikke alltid gjennomført. Av tabellen ser man at det er kun fornavn og kjønn som er felles i datastrukturen i de to registreringene.

Tabell 1

Feltstruktur ved registrering av dåp og begravelse, gjeldende fra år 1821.

Dåp	**Begravelse**
Fornavn	Fornavn
	Patronymikon
Far	
Mor	
Sokn	Sokn
Dåpsdato	
	Dødsdato
	Gravlagt dato
	Fødselsdato
Kjønn	Kjønn
	Alder

[4] Tabell 1 er et utdrag av de tabellformater for de standardiserte skjemaer som skulle benyttes ved kirkelige handlinger. Tabell 24 og tabell 25 i vedlegget viser alle tilgjengelige felter.

3.2 Informasjon i dåpsdata og begravelsesdata

Mine kildedata har vært igjennom en maskinell bearbeiding hvor fornavn og farsnavn (patronymikon) har blitt standardisert gjennom ulike mønstre for gjenkjenning navnestammer og patronymikon. Fødselsdato har blitt standardisert blant annet etter graden av usikkerhet i den registrerte verdien for å minimere problemene med feil i transkriberingen.

I tillegg til de navnefeltene som benyttes, beregnes fødselsdato ut fra alder og dødsdato i begravelsesdata for estimering av fødselsdato. Særlig alder er et usikkert felt, med tanke på at presten vanligvis benyttet heltall for å beskrive alderen til voksne personer. I tillegg til den kalkulerte fødselsdato er det angitt en pluss og minus bredde for fødselsdatoen både for den estimerte fødselsdato fra begravelsesdata og i de tilfeller det er usikkerhet rundt fødselsdato i dåpsdata. Alle datoverdier er angitt som egne felt i mine standardiserte kildedata, basert blant annet på usikkerhet, med et felt for året, et felt for måneden og et felt for dagen. Eksempelvis kan data i datofeltet være registrert som ”2.mar. 1849” eller ”2/3/1849”, disse verdiene blir videre standardisert som verdien ”02” i feltet for dag, verdien ”03” i feltet for måned og verdien ”1849” for fødselsåret. For videre bruk i min datamodell blir feltene sammensatt til ”02.02.1848” etter standarden i Histform.

3.3 Harmonisering av inndata

Eneste endring i forhold til de standardiserte data i datasettene jeg har foretatt ligger i presentasjonen av navnefelter og datoformatene. Jeg har samlet dato for dag, måned og år i et felt[5] med datoformatet spesifisert i Histform som standard, dd.mm.åååå, data i feltene med navneinformasjon har kun blitt ryddet for ledende og sluttende mellomrom.

[5] En oversikt over kombinasjoner av data i fødselsdatofeltet vises i tabell 29 i appendiks.

3.4 Poststruktur

De standardiserte datasettene har en poststruktur[6] med 24 datafelt i både dåps- og begravelsesdatasettet. Datamodellen har en enklere poststruktur som inneholder feltene fornavn, farnavn, estimert fødselsdato og kjønn fra begge tabeller, i tillegg til identifikatorer og geografiske felt.

3.4.1 Fornavn- og farsnavnsinformasjon

Fornavnet er et primærfelt, siden dette feltet var påkrevd ved registrering i både dåpslisten og begravelseslisten. Fullt navn på far er registrert i dåpslisten mens det i begravelseslisten er standardisert fra patronymikon.

3.4.1.1 Registrering

I begravelseslisten er det registret fornavn og fullt etternavn, og for data i dåpslisten er det kun fornavnet til barnet registrert og fulle navn til begge foreldre.

3.4.1.2 Standardisering og estimering

For navnedata i begravelsesdatasettet har fornavn og farnavn blitt standardisert etter metoder med bakgrunn i Histform. En eventuell forskjellig bruk av suffiksene –d, -datter, dottir og -dtr for patronymikon hos kvinner har blitt standardisert til det samme. Dette gjelder også for forskjellig bruk av –son, -s og –søn. Ved det nyere –sen oppstår også problemer med kjønnsangivelse og underrepresentasjon av menn uten patronym, siden det er usikkert om det er fars fornavn eller familienavn. Standardiseringen tar heller ikke hensyn til gårdsnavn.

3.4.1.3 Generelle problemer

Et vanlig problem ved navneinformasjonen er bruken av flere navn både i fornavn og farsnavn. En person kan være døpt og registrert med flere

[6] Dette er et utdrag av hele feltstrukturen i de standardiserte datasettene jeg hadde som datakilde. Oversikt over alle feltene i dåpstabellen finnes i tabell 27 og alle tilgjengelige felt for begravelsestabellen finnes i tabell 26 i appendiks.

fornavn under dåpen, mens presten kun registrerer et av dem etter begravelsen. Sannsynligheten for ortografiske feil og transkriberingsfeil øker selvfølgelig også ved flere registrerte navn. Vanlige problemer (Fure, 2000) som endring i skrivemåten, feilregistrering, feil i transkribering og feil i kodingen er heller ikke uvanlig. Et annet problem er dobbeltføring i kirkeboken hvor den ene posten i tillegg er feilskrevet.

3.4.2 Datoinformasjon

Data i datasettet for dåp må antas å være de mest nøyaktige, siden informasjon om dato for fødsel er angitt i de fleste tilfeller. Særlig i perioden etter 1877[7] synker andelen med ufullstendige data. I de tilfeller hvor fødselsinformasjonen er ukjent vil de aller fleste være født innenfor et ni måneders intervall forutfor for dåpsdagen[8]. Ved begravelse registrerte presten dato for død, begravelse og alder. Ved bakgrunn i disse data innholder datasettet for begravelse en estimert fødselsdato.

3.4.2.1 Registrering

For registreringen av dåpshandlingen var dato for selve dåpsseremonien den dato som var påkrevd registrert. Begravelsesdato og alder var påkrevd for begravelser, etter 1877 ble det også krav om fødselsdato.

3.4.2.2 Standardisering og estimering

Alder og begravelsesdato finnes i de fleste poster, fødselsdato er derimot ofte mangelfull. Hvis fødselsdato er mangelfull, estimeres det et intervall for fødselsdato. Hvis bare dagen er ukjent så settes intervallet på pluss og minus en måned. Er både måned og dag ukjent vil intervallet angis som pluss minus et år fra en estimert dato på grunnlag av dødsdato fratrukket alder. Alderen kan eksempelvis være skrevet ”7 år og 4 uker” eller ”7,5 år”, standardiseringen tar hensyn til dette og reflekterer forskjellen i estimert øvre og nedre grense for fødselsdato.

[7] Tabell 25 i appendiks viser hvilke felt som var påkrevd etter endringene i 1877
[8] Dåpen skulle være senest åtte dager etter fødsel. I reskriptet fra 1812 ble dette erstattet med ni måneder.

3.4.2.3 Generelle problemer

Det var ikke uvanlig at presten kastet jord på uker og måneder etter gravleggelsen for så å føre det i kirkeboken. Den estimerte datoen for fødselsdato kan derfor sies å være svært usikker i mange tilfeller. Klare typografiske feil i dato eller manglende informasjon i datakilden blir merket under standardiseringen, denne merkingen får den videre implikasjon at alderspennet blir bredere og resulterer i flere mulige like gode treff.

4. Datalenking

Datalenking er en integrasjon av informasjon fra to ulike datakilder. Data fra de to kildene antas å relatere til det samme individ i på en slik måte at det kan tolkes som en enkel post i datasettet.

Data er vanligvis lenket gjennom felles identifikatorer/nøkkelfelt (f.eks. fødselsnummer), når en unik identifikator er tilstede i begge datasett er datalenkingen en relativt enkel operasjon. Utfordringene oppstår derfor i tilfeller hvor det ikke er felles identifikatorer tilstede eller når disse identifikatorene er ufullstendige og derfor ikke troverdige.

4.1 Bakgrunn

En av pionerene innenfor den begynnende defineringen av datalenking som fag, var H. L. Dunn ved United States National Bureau of Statistics.

> *"Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Record linkage is the name of the process of assembling the pages of this Book into a volume" (Dunn, 1946)*

Den canadiske genetikeren Howard Newcombe startet med sine kollegaer allerede i 1959 arbeidet med datamaskinbasert sammenlikning av uensartet data fra forskjellige medisinske dataregistre. De ble tidlig klare over fordelen med bruk av datamaskin ved automatisering av store mengder med data (Newcombe, 1959).

> *"Computerized record linkage has the advantages of quality control, speed, consistency, reproducibility of results, and the ability to handle large volumes of data." (Newcombe, 1988)*

Med datalenking, norsk oversettelse av det engelske ordet record linkage, menes lenking av data fra en eller flere datakilder som representerer en entitet. Det skilles mellom deterministisk- og probabilistisk datalenking som metoder for datalenking.

4.1.1 Deterministisk datalenking

Deterministisk datalenking eller eksakt lenking, er hvor man ser på ett eller flere perfekte par av datafelt mellom ulike datasett. Et eksempel kan være et det eksisterer fullt fødselsnummer for posten i hvert datasett. Typografiske feil av fødselsnummer i et av datasettene vil medføre at noen positive treff vil forsvinne eller mangle blant resultatene.

Deterministisk eller eksakt lenking ikke mulig for min datamodell, siden det ikke eksisterer en unik identifikator i datasettet, for nyere norske data vil det være mulig å benytte fullt fødselsnummer.

4.1.2 Probabilistisk datalenking

Probabilistisk datalenking er hvor man, i mangel unike sammenfallende identifikatorer, benytter informasjon fra en større mengde datafelt.

Probabilistisk datalenking, bruk av sannsynlighet for om to datasett referer til samme individ. Kalkulerer (vekting) for sannsynligheten for om den lenkingen er et treff (Christen, 2002).

Vektingen kan enten være verdispesifikk eller ikke-verdispesifikk (generell), ikke-verdispesifikk gis vektverdi som f.eks 2,0 ved treff og -1,5 ved ikke treff i fødselsdato. Ved verdispesifikk kalkuleres oddsen/frekvensen raten for hver verdi på grunnlag av: vekt = log2 * (utfall av lenkede par/utfall av ikke lenkede par)

4.2 Strategier ved lenking av data

Felles felt i mine standardiserte datasett for dåp og begravelse, består av de standardiserte fornavn, farsnavn og kalkulert fødselsår i begravelsesdatasettet.

Utfallet av min datalenking kan ha to utfall, enten er det et positivt treff eller så er det et negativt treff. Hvis det derimot finnes flere personer med de samme feltene og jeg velger feil person vil dette kunne karakteriseres som en falsk positiv. Den samme kan skje ved å forkaste en person med typologiske feil i sine datafelt, da vil dette treffet kunne karakteriseres som et negativt treff, men falsk negativ. Denne strategien vises i **tabell 2** med to typer utfall, hvor hvert utfall har to muligheter.

Tabell 2

Mulige utfall for probabilistisk datalenking

Utfall	**Mulighet**
Positivt treff	Sanne positiver
	Falske positiver
Negativt treff	Sanne negativer
	Falske negativer

Tabell 2 kan også presenteres i form som **tabell 3** for kvantifisering og kalkulering av følsomhet for treffene, spesifisering av ikke-treffene og positiv predikert verdi (PPV). Kombinasjonen av poster fra de to datasettene klassifiseres da enten som et antall av treff eller ikke-treff kombinasjoner (Blakely og Salmond, 2002). Videre kan paret bli lenket sammen eller ikke-lenkes. Mulighetene for lenken blir da enten at det er en sann positiv ved treff og en falsk positiv ved ikke-treff. Ved ikke-lenking vil et treff karakteriseres som en falsk negativ og et ikke-treff som en sann negativ.

Tabell 3

Kalkulering av følsomhet, spesifisering og PPV

	Treff	**Ikke-treff**
Lenket	**a** Sanne postiver	**b** Falske positiver
Ikke-lenket	**c** Falske negativer	**d** Sanne negativer

Videre kan man beregne følsomheten for treffene, definert som:

$$\text{følsomhet} = a / (a + c)$$

og en spesifisering av ikke-treffene definert som;

$$\text{spesifisering} = d / (b + d)$$

5. Teoretisk modell

Siden det er så mange muligheter for feil er det viktig at algoritmen gir de sannsynlige treffene probabilistisk. Utforming og design av algoritmen er derfor en signifikant komponent i oppgaven.

Det teoretiske rammeverket for datalenking er utviklet fra Fellegi og Sunter (1969), som viste at det er mulig å definere optimale regler som reduserer antallet falske lenker. I tillegg viste teorien teststatistikk for evaluering av feilrater og spesifisering av antagelser, som var nødvendig for estimering av treff sannsynligheter som benyttes til å kalkulere test statistikken. Videreutvikling og justeringer av dette rammeverket finnes beskrevet hos en mengde forfattere innen emnet.

Alle disse modellene antar at alle datapar fra to datasett enten refererer til et individ eller at ikke treff refererer til to ulike personer; optimal treff krever at hvert individ sammenliknes med hvert mulig treff.

For å redusere antallet mulige sammenlikninger har det blitt utviklet nye teorier, som introduserer begrepet ”blokkeringsfaktorer”, som kjønn og fødselsdato for å begrense sammenlikningene til individer med like faktorer. I tillegg til disse teoriene er det også utviklet modeller som Baysian models (Baysian networks), HMM modeller (Hidden Markov Modell) for standardisering og duplikat modeller for å fjerne duplikater og falske positiver.

5.1 Fellegi og Sunter modellen for datalenking.[9]

En sammenligning for å beslutte om paret er:

1) Lenke A_1
2) Mulig lenke A_2
3) Ikke-lenke A_3

Mulige feilkilder blir da om A_1 er feil beslutning og det ikke er treff eller ved A_2 og at paret faktisk er et treff

Gitt to populasjoner A og B med a og b som notasjon for elementer i disse. Vi antar at noen elementer er felles for A og B,

$$A \times B = \{ a,b);a \in A,\ b \in B$$

er unionen av to atskilte (disjunkte) datasett:

(1) $$M = \{ a,b);a = b,\ a \in A,\ b \in B$$

og

(2) $$U = \{ a,b);a \neq b,\ a \in A,\ b \in B$$

Som vi kan kalle lenket og ikke-lenket

Hver enhet i populasjonen har et sett egenskaper, f.eks fornavn, farsnavn, kjønn, fødselsdato osv. Anta videre at vi ut fra populasjonen genererer to datasett med poster med hvert sitt sett med egenskaper. Denne datasett

[9] Dette er et utdrag av fra: A Theory for record Linkage, Ivan P. Fellegi and Alan B. Sunter, Journal of the American statistical association, Vol.64, No. 328 (Dec., 1969), 1183-1210

genereringen gir også rom for ufullstendige poster og feil fra f.eks feilrapportering, ingen rapportering, feil koding, feil tasting, transkribering osv med det resultat at et ikke-lenket par av A og B kan rapporteres som treff enten pga feil eller mangel av egenskaper i likhet med et lenket A og B kan feilrapporteres som ikke-treff.

Vi bruker $\alpha(a)$ og $\beta(b)$ som notasjon for poster som korresponderer med datasett A og B. Vi antar også at de tilfeldige utvalgene A_3 og B_3 er valgt fra A og B. Vi ser ikke bort fra den sannsynligheten at $A_3 = A$ og $B_3 = B$. De to gitte filene L_A og L_B er et resultat av en datasett generering fra A_3 og B_3 For enkelhets skyld dropper vi fotnotasjonen s

Første steg i datalenking prosessen blir å prøve og sammenlikne postene i datasettene, og kode sammenlikningen med "fornavn er likt", "farnavn er ulikt", "Fødseldato mangler" osv

Sammenlikningsrom (comparison spaces).
Vi kan formelt definere sammenlikningsvektoren som en vektor funksjon for postene $\alpha(a)$ og $\beta(b)$

$$\textbf{(3)} \qquad \gamma[\alpha(a), \beta(b)] = \{ \gamma^1[\alpha(a), \beta(b)], \dots , \gamma^K[\alpha(a), \beta(b)] \}$$

med γ som en funksjon av A $\times$ B Vi kan skrive $\gamma(a, b)$ eller $\gamma(\alpha, \beta)$ eller simpelthen γ for vårt bruk. Settet av alle mulige funn av γ kalles et sammenlikningsrom, Γ.

I prosessen med lenking observerer vi $\gamma(a, b)$ og ønsker å avgjøre om paret (a, b) er et treff $\gamma(a, b) \in M$ (vi kaller dette treffet for en positiv lenke med notasjon A_1) eller om $(a, b) \in U$ er et treff dvs et treff for ikke-par (vi kaller dette treffet for en positiv ikke-lenke med notasjon A_3). Det kan være tilfeller, uansett definert feilnivå (som definert under), hvor vi

kan avgjøre om paret er en positiv lenke/ikke-lenke, vi samler disse tilfellene som mulige lenker (med notasjon A_2)

Lenkeregel (linkage rule).

En lenkeregel kan nå defineres som en kobling fra Γ, sammenlikningsrommet, til et sett av tilfeldige beslutningsfunksjoner $D = \{ d(\gamma) \}$ hvor;

(4) $$d(\gamma) = \{P(A_1|\gamma), P(A_2|\gamma), P(A_3|\gamma)\}; \quad \gamma \in \Gamma$$

og

(5) $$\sum_{i=1}^{3} \mathrm{P(Ai} \mid \gamma) = 1$$

Mao korresponderende til hver observert verdi av γ, setter lenkeregelen sannsynlighetene for hver av de tre mulige variantene. For noen eller alle de mulige verdiene av γ som beslutningsfunksjonen kan være en utartet tilfeldig variabel, den kan tilordne en av variantene med sannsynlighet lik 1.

Vi må ta hensyn til nivåene av feil tilordnet en lenkeregel. Vi antar forløpig at et par av poster $[\alpha(a), \beta(b)]$ blir valgt for sammenlikning til den samme sannsynlighetsprosess fra $L_A \times L_B$ (dette er det samme som å velge et tilfeldig par av felter (a, b) fra $A \times B$, på bakgrunn av konstruksjonen av L_A og L_B). Den resulterende sammenlikningsvektor $\gamma[\alpha(a), \beta(b)]$ er en tilfeldig variabel. Vi bruker notasjonen for den betingete sannsynlighet til γ gitt at $(a, b) \in M$ fra $m(\gamma)$ som gir

(6) $$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in M \}$$
$$= \sum_{(a,b)\in M} \mathrm{P}\{\gamma[\alpha(a), \beta(b)]\} \cdot \mathrm{P[(a, b) \mid M]}$$

På samme måte kan den betingete sannsynlighet for γ, gitt at $(a, b) \in U$ fra $u(\gamma)$ skrives som

(7) $$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in U\} = \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot \mathrm{P}[(\mathrm{a}, \mathrm{b}) \mid \mathrm{U}]$$

Det er to typer av feil som kan oppstå i en lenkeregel. Den første inntreffer når et ikke-treff i sammenlikningen blir lenket og har sannsynligheten

(8) $$P(A_1|U) = \sum_{\gamma\in\Gamma} \mathrm{u}(\gamma \cdot \mathrm{P}(\mathrm{A}_1 \mid \gamma)$$

Den andre inntreffer når et treff i sammenlikningen er ikke-lenket og har sannsynligheten

(9) $$P(A_3|M) = \sum_{\gamma\in\Gamma} \mathrm{m}(\gamma \cdot \mathrm{P}(\mathrm{A}_3 \mid \gamma)$$

En lenkeregel for rommet Γ kan sies å være en lenkeregel for nivåene μ, λ ($0 < \mu < 1$ og $0 < \lambda < 1$), og betegnet av $L(\mu, \lambda, \Gamma)$ hvis

(10) $$P(A_3|M) = \mu$$

og

(11) $$P(A_3|M) = \lambda$$

I gruppen av lenkeregler for Γ som tilfredsstiller ***(10)*** og ***(11)*** vil lenkeregel $L(\mu, \lambda, \Gamma)$ være den optimale lenkeregel hvis relasjonen:

(12) $$P(A_2|L \leq P(A_2|L')$$

Holder for hver $L'(\mu, \lambda, \Gamma)$ i gruppen av lenkeregler

I forklaringen av definisjonen må det bemerkes at den optimale lenkeregel maksimerer sannsynlighetene positive anordninger av sammenlikningene for det satte feilnivå i ***(10)*** og ***(11)*** eller mao, det minimerer sannsynligheten for å anordne sammenlikningsfeil. Dette ser ut til å være en fornuftig tilnærming siden A_2 vil kreve stor manuell lenking, alternativt hvis sannsynligheten for A_2 ikke er liten, vil lenke prosessen være et tvilsomt verktøy.

Det er ikke vanskelig å se at for spesielle kombinasjoner av μ og λ i gruppen av lenkeregler som tilfredsstiller ***(10)*** og ***(11)*** vil være tomme. Følgelig vil de kombinasjoner av μ og λ som er mulige for å tilfredsstiller likningene ***(10)*** og ***(11)*** simultant med et sett D med beslutningsfunksjoner definert av ***(4)*** og ***(5)***. På dette punktet er det tilstrekkelig å bemerke at et par med felter μ og λ vil være tillatelig bare hvis en eller begge av verdiene er tilstrekkelige store som medfører at feilnivåene reduseres.

Det fundamentale teoremet.

Først defineres lenkeregel L_0 for Γ. Vi starter med å definere den unike sorteringen av det gitte settet av mulige realiseringer av γ.

Hvis noen verdier av γ er slik at både $m(\gamma)$ og $u(\gamma)$ er lik null, da er den betingete sannsynlighet for realisering av den verdien av γ lik null, og derfor trenger den ikke å tillegges Γ. Vi tilegner nå en vilkårlig rekkefølge for alle γ hvor $m(\gamma) > 0$ men $u(\gamma) = 0$.

Så sorterer vi de gjenværende γ på en slik måte at den korresponderende sekvensen av

$$m(\gamma) / u(\gamma)$$

er monotont synkende. Når verdien av $m(\gamma) / u(\gamma)$ er den samme for mer enn en γ sorterer vi de γ vilkårlig.

Vi indekserer det ordnede settet $\{\gamma\}$ med fotnotasjonen i; $(i = 1, 2, \ldots, N_\Gamma)$; og skriver

$$u_i = u(\gamma_i);\ m_i = m(\gamma_i).$$

La være et tillatelig par av feilnivåer og velg n og n' slik at

(13) $$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

(14) $$\sum_{i=n'}^{N\Gamma} m_i \geq \lambda > \sum_{i=n'+1}^{N\Gamma} m_i$$

Hvor N_Γ er antallet punkter i Γ.

La oss anta for nå at når *(13) og (14)* er oppfylt at vi har $1 < n < n' - 1 < N_\Gamma$. Dette vil sikre at nivåene (μ, λ) er tillatelige. La $L_0(\mu, \lambda, \Gamma)$ være den lenkeregelen definert som følger: observert en sammenlikningsvektor γ, velg A_1 (positiv lenke) hvis $i < n - 1$, velg A_2 hvis $n < i < n' - 1$ og velg A_3 hvis (positiv ikke-lenke) $i > n' + 1$. Når $i = n$ eller $i = n'$ vil et vilkårlig valg være nødvendig for å oppnå de eksakte feilnivåene μ og λ, formelt kan dette vises som,

(15)
$$d(\gamma_i)=\begin{cases}(1,0,0) & i\le n-1 \quad (a)\\ (P_\mu,1-P_\mu,0) & i=n \quad (b)\\ (0,1,0) & n<i\le n'-1 \quad (c)\\ (0,1-P_\lambda,P_\lambda & i=n' \quad (d)\\ (0,0,1) & i\ge n'+1 \quad (e)\end{cases}$$

hvor P_μ og P_λ defineres som løsninger til funksjonene

(16)
$$u_n\cdot P_\mu=\mu-\sum_{i=1}^{n=1}u_i$$

(17)
$$m_{n'}\cdot P_\lambda=\lambda-\sum_{i=n'+1}^{N\Gamma}m_i$$

Teorem:
La $L_0(\mu, \lambda, \Gamma)$ være den lenkeregelen definert av ***(15)***. Da er *L* den beste lenkeregelen til for nivåene.

Denne teorien kan også vises gjennom klassisk teori for hypotese testing.

To logiske konsekvenser av teorien kan vises.

Logisk konsekvens 1: Hvis

$$\mu=\sum_{i=1}^{n}u_i,\quad \lambda=\sum_{i=n}^{N\Gamma}m_i,\quad n<n',$$

da blir den beste lenkeregel $L_0(\mu, \lambda, \Gamma)$ for nivåene (μ, λ)

(18)
$$d(\gamma_i)=\begin{cases}(1,0,0) & hvis \quad 1\le i\le n\\ (0,1,0) & hvis \quad n<i<n'\\ (0,0,1) & hvis \quad n'\le i\le N_\Gamma\end{cases}$$

hvis vi definerer

$$T_\mu = m(\gamma_h) / m(\gamma_h)$$

$$T_\lambda = m(\gamma_{h'}) / m(\gamma_{h'})$$

dette medfører at lenkeregel ***(18)*** kan skrives som

(19)
$$d(\gamma) = \begin{cases} (1, 0, 0) & hvis \quad T_\mu \le m(\gamma)/u(\gamma) \\ (0, 1, 0) & hvis \quad T_\lambda < m(\gamma)/u(\gamma) < T_u \\ (0, 0, 1) & hvis \quad m(\gamma)/u(\gamma) \le T_\lambda \end{cases}$$

Logisk konsekvens 2: La T_μ *og* T_λ være et hvilket som helst positivt tall slik at

$$T_\mu > T_\lambda$$

Det eksisterer et par med feilnivåer (μ, λ) korresponderende til T_μ *og* T_λ slik at lenkeregelen ***(19)*** er den beste for disse nivåene. Nivåene (μ, λ) er gitt fra:

(20)
$$\mu = \sum_{\gamma \in \Gamma\mu} u(\gamma)$$

(21)
$$\lambda = \sum_{\gamma \in \Gamma\lambda} m(\gamma)$$

hvor

(22)
$$\Gamma_\mu = \{\gamma : T_\mu \le m(\gamma) / u(\gamma)\}$$

(23)
$$\Gamma_\lambda \{\gamma : m(\gamma) / u(\gamma) \le T_\lambda\}$$

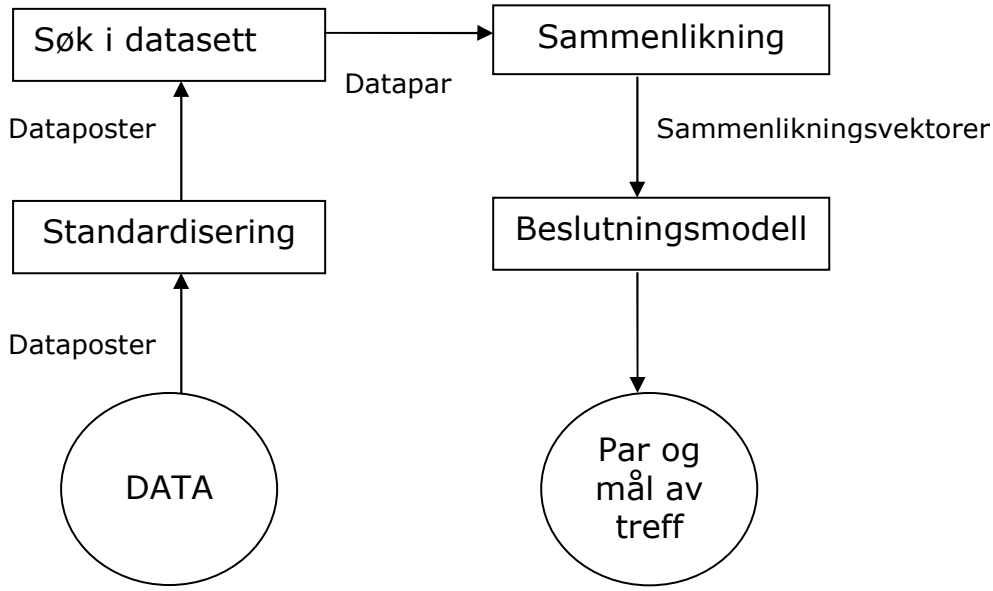
I mange verktøy vil man være villige til å tolerere feilnivå på et høyt nok nivå for å utelukke forekomsten av A_2. I dette tilfellet velger vi n og n' eller alternativt T_μ og T_λ slik at det midtre sett av γ i **(18)** og **(19)** blir tomme. Mao hver *(a, b)* blir allokert enten til *M* eller *U*.

Teorien om allokering av observasjoner kan anses som et spesialtilfelle i denne teorien.

6. Prosesser, dataflyt og programlogikk.

Prosessene i det konstruerte dataprogrammet kan beskrives som vist i **figur 1**. Datakilden er ordnet i to tabeller med digitaliserte poster med dåpsdata og begravelsesdata fra kirkebøkene. Disse datapostene har så blitt standardisert med hensyn på fornavn, farsnavn og fødselsdato. Algoritmen i dataprogrammet tar utgangspunkt i en og en post i begravelsesdatasettet og søker igjennom alle poster i dåpsdatasettet, og danner datapar når feltene korresponderer. Alle dataparene gis en vekting etter hvor nær obeservasjonen er[10]. Etter vektingen sammenliknes så alle treffene for hver post i begravelsesdatasettet etter gitte kriterier fra beslutningsmodellen integrert i algoritmen. Det beregnes en sannsynlighet med bakgrunn i vektene på hvor sikker treffet er, sannsynlighetsmodellen er veldig grov, og samsvarer på ingen måter med Fellegi og Sunter modellen fra kapittel 5. Ambisjonen var å benytte denne sannsynligheten for å fjerne lenker under en viss sannsynlighet.

Figur 1
Prosessoversikt og modell benyttet i oppgaven.

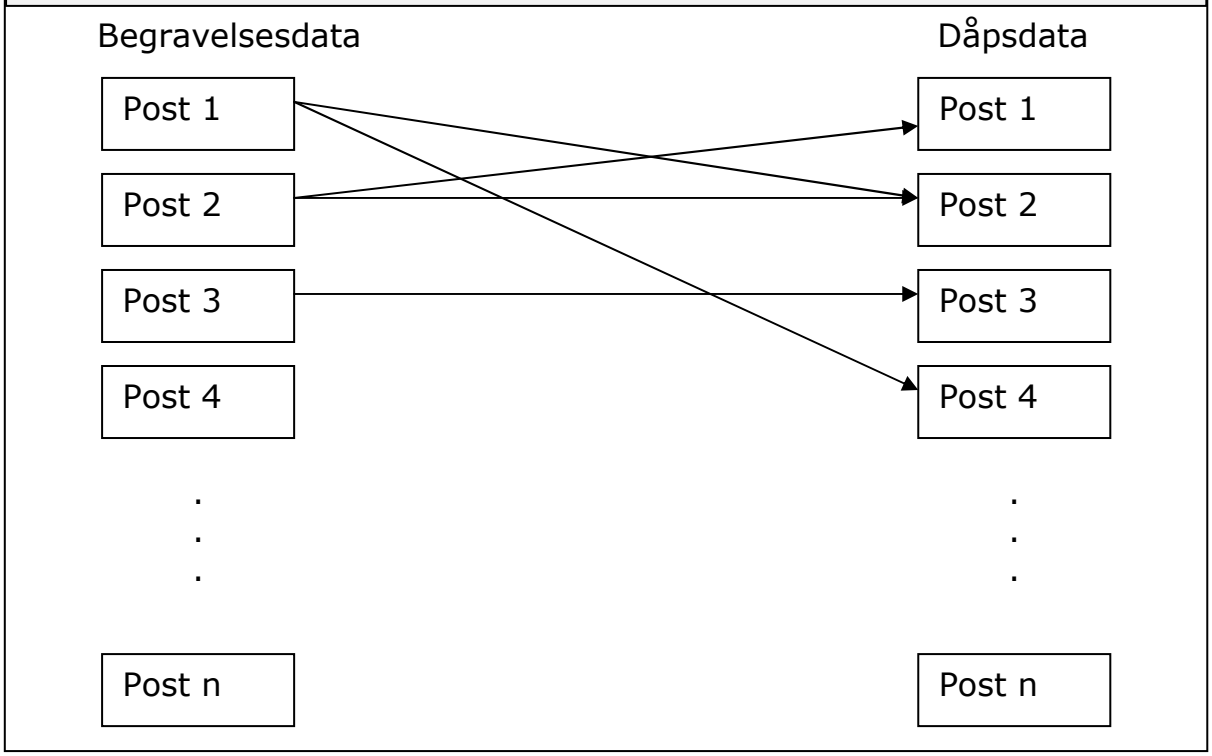


[10] For oversikt over hvilken vekt de ulike treffene er gitt, se tabell 31 i appendiks.

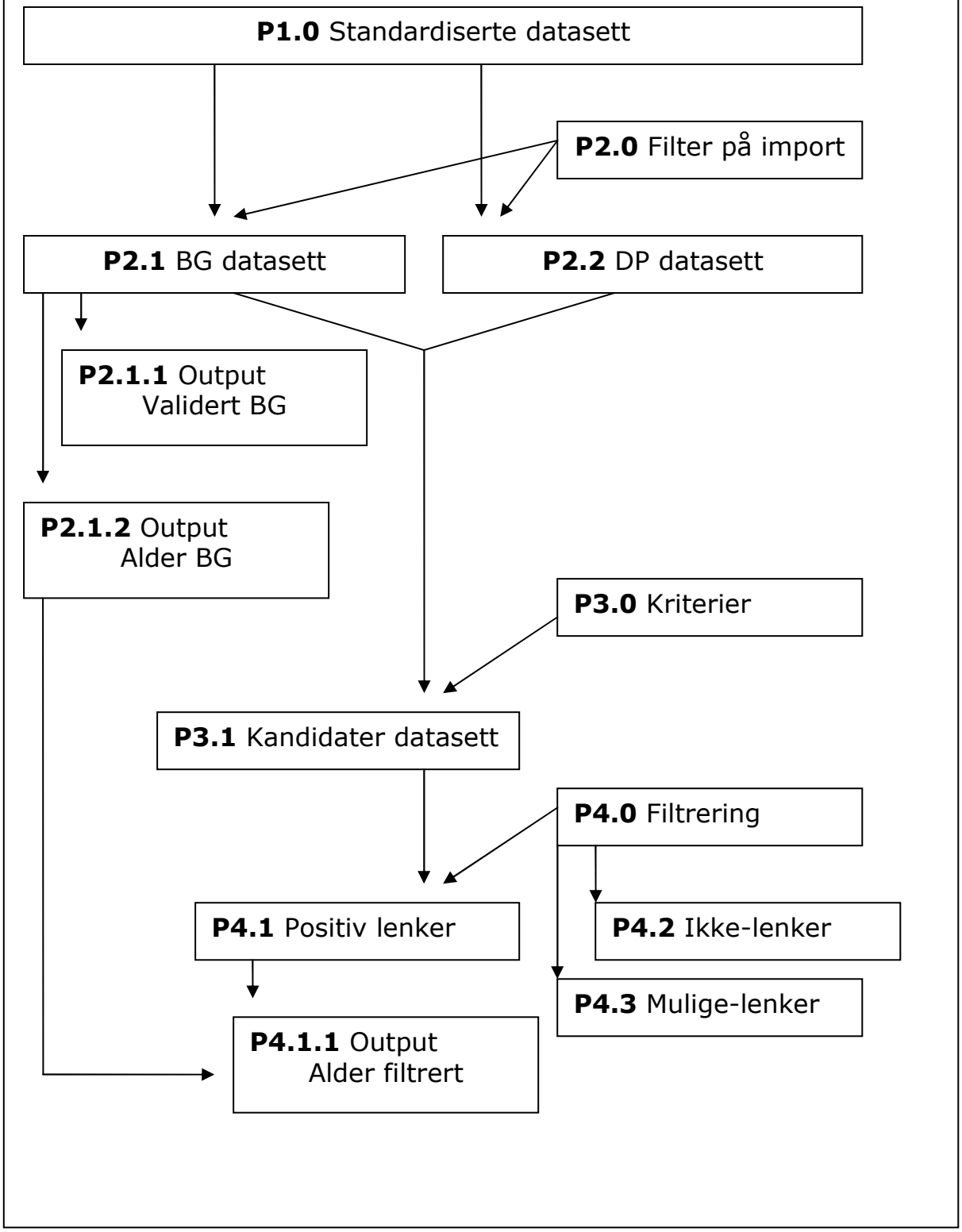
Jeg velger å ta utgangspunkt i begravelsesdatasettet siden postene angir en estimert fødselsdato, som gir meg mulighetene for filtrerte søk mot gitte år i dåpsdata. Dåpslistene angir ingen begravelsesdato, slik at et søk måtte inkludert hele gruppen i datasettet med begravelser.

Dannelsen av datapar kan illustreres som vist i **figur 2**. Post nummer en til post nummer n i datasettet med begravelsesdata sammenliknes en etter en med alle poster i datasettet for dåpsdata. Post nummer en i datasettet med begravelsesdata kan eksempelvis da i etterkant være lenket mot post to og post fire, post to mot post en og mot post to. Vektingen av graden av likhet mellom feltene vil avgjøre hvilket datapar som besluttes å være den beste kandidaten. Post nummer tre har kun etablert et datapar mot post nummer tre i dåpsdata, hvis det er treff for flere av feltene og derfor høy vekting, vil dette paret besluttes å være den beste kandidaten. Post nummer fire har ikke blitt lenket mot noen poster i dåpsdata, og blir av den grunn blokkert ut i den videre prosessen.

Figur 2
Mulige lenker for postene i datasettene



Figur 3
Dataflyt og programlogikk. Begravelsesdata (BG) og dåpsdata (DP).



Forklaring av dataflyten og programlogikk.

Programmet er programmert i VB.Net og inneholder over 1000 kodelinjer og flere skjermbilder for de forskjellige utsnittene av datasettene. Datasettene er lagret i tabeller i en MS Access relasjonsdatabase.

P1.0 Standardiserte datasett.

Datasettene er standardiserte med utgangspunkt i en datakilde med anonymiserte data fra 14 prestegjeld/sokn i Norge, som er et utvalg av alle prestegjeld i datakilden.

P2.0 Filtrering av datasettene.

Det er mulig å søke i (filtrere) datasettene etter følgende kriterier; fødselsdato med startår og sluttår, nummer på prestegjeld og alder. Dette gjør det mulig å se på forskjellige utsnitt, som for eksempel alder under 5 år eller de personer som er født mellom år 1820 og år 1830.

Fødselsår kan velges mellom verdiene fra og med år 1700 og til og med år 1900, gitt at prestegjeldet har data fra perioden. Brukeren kan velge alder i intervallet fra 0 til 100 år. Start- og sluttår velges også for dåpsdata, men gitt øvre og nedre grense for begravelsesdata med usikkerhet rundt fødseldatoen, bør de velges minus ett år for startår og pluss ett år for sluttår. Feltene fødselsmåned (BRTMNTH), fødselsdag (BRTHDAY) og fødselsår (BRTHYEAR) settes sammen til feltet fødselsdato standardisert (FDATO_S) med punktum som separator, feltypen char (tekstfelt) beholdes siden mange felt med ukjente dager og måneder benytter 99 som flagg for feil i kildedata.

I tillegg til den standardiserte og kalkulerte fødselsdato, settes et pluss/minus intervall siden denne datoen er såpass usikker. Ved høy usikkerhet settes det opptil pluss/minus et år og i de tilfeller man er helt sikre på datoen sette dato intervallet til pluss/minus 0 dager. For å angi en øvre og nedre akseptabel grense benyttes to kalkulerte felt

sammenslått av feltene fødselsdag-lav (BRTHDL), fødselsmåned-lav (BRTHML) og fødselsår-lav (BRTHYL). Disse feltene slås sammen til feltet fødselsdato-lav (FDATO_L). Tilsvarende for øvre grense, benyttes feltene BRTHDU, BRTHMU og BRTHYU som slås sammen til nytt felt fødselsdato_øvre (FDATO_U) med punktum som separator, siden det standardiserte feltet alltid innholder verdier, kan det nye feltet endres til datetime (datoformat) datatype, for bruk i datosøk.

Ingen data i de standardiserte datasettene forkastes. Dette medfører at det heller ikke sjekkes for dubletter i inndata, verken i begravelsesdatasettet eller i dåpsdatasettet. Feltet kjønn er tatt med for en eventuell utvidelse av datamodellen bak algoritmen for utvelgelse av datalenker. Prosessen videre er også ment å være en test på modellens evne til datalenking, og avsluttes med en oppsummering av resultatene.

P2.1 Begravelsesdatasett

Datasettet med begravelsesdata inneholder feltene for bruk i lenkeprosessen, med tillegg av feltet UID som benyttes som unik identifikator av posten. Feltene alder (AGE_DAYS) og prestegjeld (MUNICIP) benyttes som oppslagsfelter for filtrering i prosess **P2.0** **Figur 4** viser et utsnitt av 458 filtrerte poster fra begravelsesdatasettet med personer født fra og med år 1840 til og med år 1849.

Figur 4.

Utsnitt av 458 poster fra begravelsesdatasettet, født fra 1840 tom 1849

Dataoutput

Datasett filter | Ny SQL lenking | Kandidater | Datalogg | Innverdier/Konstanter

Antall poster i bg_enc datasett: 458

UID	FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	AGE_DA	FDATO_S	FDATO_L	FDATO_U
11572061027	(null)	(null)	(null)	(null)	0	30	99.02.1840	04.01.1840	03.02.1840
11572061046	ANE	AN	ANDREAS	ANDRES	1	30	99.05.1840	08.04.1840	08.05.1840
11572061051	(null)	(null)	(null)	(null)	8	8	22.07.1840	22.07.1840	22.07.1840
11572061060	LAURITZ	LORES	(null)	(null)	0	6	24.10.1840	24.10.1840	24.10.1840
11572061066	KAREN	KOREN	HALVOR	HOLVOR	1	240	99.05.1840	22.04.1840	22.05.1840
11572061077	FREDERICH	FREDREK	(null)	(null)	0	180	02.10.1840	02.10.1840	02.10.1840
11572061078	HANS PETTER JULIUS	HONS PETER JOLOS	(null)	(null)	0	330	09.05.1840	09.05.1840	09.05.1840
11572061081	HELMINE CHRISTINE	HELMEN KRESTEN	IVER	EVER	1	455	99.01.1840	30.12.1839	29.01.1840
11572061087	JOHANNE	JOHON	BAARD	BOR	1	365	99.99.1840	05.06.1839	05.06.1840
11572061090	OLAVA	OLOVE	OUGEN	OGEN	1	545	99.01.1840	23.12.1839	22.01.1840
11572061094	CHRISTOPHER	KRESTOVER	(null)	(null)	0	180	99.03.1841	05.02.1841	07.03.1841

P2.1.1 Validerte data fra begravelsesdatasettet.

For å vurdere validiteten til de standardiserte data som benyttes videre i prosessen valideres data ut fra følgende kriterier:

- Eksistens av sammenliknbare data i feltet med fornavn
- Eksistens av sammenliknbare data i feltet for kjønn
- Eksistens av sammenliknbare data i feltet for fødselsdato

Fornavn og kjønn er to felt det i høy grad gjenfinnes gode data i, mens i feltene farsnavn og fødselsdato er det en høyere variasjon mellom årene. **Figur 6** viser en validering av begravelsesdata fra perioden 1840 til og med 1849. Standardiserte felter som fornavn og farsnavn følger av om det er faktisk observerte data.

Figur 5.

Filtrerte data fra perioden 1840-1849 med antall lesbare data i feltet

Kildedata vurdering:

År	Fornavn	Forn_s	Farnavn	Farn_s	Sex	Fdato	%Snitt/Sum
1840	96,6% - (56)	96,6% - (56)	41,4% - (24)	41,4% - (24)	98,3% - (57)	25,9% - (15)	66,7% - (58)
1841	100,0% - (46)	100,0% - (46)	19,6% - (9)	19,6% - (9)	100,0% - (46)	28,3% - (13)	61,2% - (46)
1842	100,0% - (56)	100,0% - (56)	50,0% - (28)	50,0% - (28)	100,0% - (56)	21,4% - (12)	70,2% - (56)
1843	100,0% - (35)	100,0% - (35)	57,1% - (20)	57,1% - (20)	100,0% - (35)	22,9% - (8)	72,9% - (35)
1844	100,0% - (48)	100,0% - (48)	45,8% - (22)	45,8% - (22)	100,0% - (48)	14,6% - (7)	67,7% - (48)
1845	100,0% - (51)	100,0% - (51)	45,1% - (23)	45,1% - (23)	100,0% - (51)	29,4% - (15)	69,9% - (51)
1846	100,0% - (44)	100,0% - (44)	45,5% - (20)	45,5% - (20)	100,0% - (44)	38,6% - (17)	71,6% - (44)
1847	100,0% - (37)	100,0% - (37)	64,9% - (24)	64,9% - (24)	100,0% - (37)	43,2% - (16)	78,8% - (37)
1848	100,0% - (36)	100,0% - (36)	52,8% - (19)	52,8% - (19)	100,0% - (36)	38,9% - (14)	74,1% - (36)
1849	100,0% - (47)	100,0% - (47)	55,3% - (26)	55,3% - (26)	100,0% - (47)	23,4% - (11)	72,3% - (47)

Antall observerte dataposter varierer fra 37 til 58 innenfor tidsintervallet. Karakteristisk for disse filtrerte data er den høye andelen av ufullstendige data i feltene for farnavn og fødselsdato. Dette kan være et utslag av mange –sen navn i originalkilden. I året 1844 kan kun 7 av 48 data i feltet fødselsdato tolkes, det er så lavt som 14,6 % av totalen. For hele perioden ligger gjennomsnittet av alle tolkbare data i året mellom 61,2 % og 78,8 %

P2.1.2 Aldersgruppering av utsnittet.

Vises i sammenheng med positive lenker i **P4.1.1**

P3.0 Kriterier.

Det er utarbeidet en beslutningsmodell med forskjellige kriterier for positive lenker[11], standardverdien for eksakte treff i fornavn, farsnavn og fødselsdato gir en sannsynlighet for positiv-lenke på 99 %. Resultatene for treff samles i feltet treff. Verdiene som benyttes er, 2 for positiv-lenke, 1 for mulig-lenke og 0 for ikke-lenke. En lenke som har treff for alle felt vil vises med kombinasjonen *"2, 2, 2, 2, 0, 2",* det femte feltet kjønn vil alltid gi 0 som treff, siden det ikke benyttes i kalkuleringen av sannsynligheten. Et treff hvor standardisert fornavn er identisk og fødselsdato ligger innenfor datointervallet vil gi følgende kombinasjon *"1, 1, 2, 2, 0, 1, 1".*

Algoritmen for søk etter mulige lenker ville ha gjort rundt $2.025.000.000^3$ kalkulasjoner hvis alle tilgjengelige data skulle benyttes. Avhengig av type datamaskin ville dette tatt fra 15 minutter til 1 time, filtrering av inndata med hensyn på tid, har derfor vært et argument under testing av algoritmene.

Figur 6.
Utsnitt av 649 poster med mulige lenker.

Dataoutput

Datasett filter | Ny SQL lenking | Kandidater | Datalogg | Innverdier/Konstanter

Antall filtrerte kandidater: 649

UID	FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	FDATO_S	FDATO_L	FDATO_U	TREFF	PRO	AGE_DAYS
33001007287	JOHAN	JOHON	LARS	LORS	0	25.12.1845	25.12.1845	25.12.1845	2,2,2,2,0,2	0,95	48
33001007226	JENS	JENS	HANS	HONS	0	25.03.1845	25.03.1845	25.03.1845	2,2,2,2,0,2	0,95	354
33001007218	JULIUS	JOLOS	JOHANNES	JOHONES	0	08.01.1845	08.01.1845	08.01.1845	2,2,2,2,0,2	0,95	597
33001007268	KJØNIG	KJENEK	GUNDER	GONDER	0	17.09.1845	17.09.1845	17.09.1845	2,2,2,2,0,2	0,95	430
33001007348	HANS SYVER	HONS SEVER	JOHAN	JOHON	0	05.10.1846	05.10.1846	05.10.1846	2,2,2,2,0,2	0,95	60
33001007356	JOHN	JON	NICOLAI	NEKOLOJ	0	01.11.1846	01.11.1846	01.11.1846	2,2,2,2,0,2	0,95	113
33001007377	PEDER	PETER	HANS	HONS	0	05.04.1847	05.04.1847	05.04.1847	2,2,2,2,0,2	0,95	144
33001007323	MORITS	MORES	JOHANNES	JOHONES	0	05.07.1846	05.07.1846	05.07.1846	2,2,2,2,0,2	0,95	660
33001007354	LARS	LORS	CHRISTOPHER	KRESTOVER	0	28.10.1846	28.10.1846	28.10.1846	2,2,2,2,0,2	0,95	702
33001007052	ALEXANDER	ALEKSONDER	CHRISTEN	KRESTEN	0	06.10.1842	06.10.1842	06.10.1842	2,2,2,2,0,2	0,95	2192
33001007043	SØREN	SEREN	OLE	OLE	0	23.07.1842	23.07.1842	23.07.1842	2,2,2,2,0,2	0,95	2321
33001007428	ELLING	ELENG	VILLADS	VELOS	0	25.11.1847	25.11.1847	25.11.1847	2,2,2,2,0,2	0,95	455
33001007464	HANS ERNST	HONS ERNS	ARNE	ARNE	0	29.04.1848	29.04.1848	29.04.1848	2,2,2,2,0,2	0,95	300
33001007404	HANS	HONS	AUGEN	OGEN	0	29.07.1847	29.07.1847	29.07.1847	2,2,2,2,0,2	0,95	635
33001007510	VILLADS	VELOS	OLAVES	OLOS	0	12.02.1849	12.02.1849	12.02.1849	2,2,2,2,0,2	0,95	104
33001007458	CHRISTOPHER	KRESTOVER	OLE	OLE	0	10.06.1848	10.06.1848	10.06.1848	2,2,2,2,0,2	0,95	556
33001007193	LARS	LORS	NILS	NELS	0	08.09.1844	08.09.1844	08.09.1844	2,2,2,2,0,2	0,95	1933
33001007405	HANS	HONS	CHRISTEN	KRESTEN	0	01.08.1847	01.08.1847	01.08.1847	2,2,2,2,0,2	0,95	910
33001007412	HANS	HONS	VILLADS	VELOS	0	06.05.1847	06.05.1847	06.05.1847	2,2,2,2,0,2	0,95	1220

[11] Se oversikt over kriteriene med tilhørende sannsynligheter i tabell 30 i appendiks.

I **figur 6** ses et utsnitt av 649 mulige lenker sortert på sannsynlighet, vist i figuren med feltet pro, siden feltet kjønn er utelatt vil den maksimale sannsynligheten være gitt fra treffverdiene *"2, 2, 2, 2, 0, 2"* noe som medfører at total sannsynlighet er lik $0{,}99^5$ som avrundes til 95 % i programmet.

P3.1 Filtrering av datasettene for mulige lenker (kandidater).
Resultatene av algoritmen tilordnet forskjellige kriterier vises i utsnittet i **figur 6**. Algoritmen har tatt utgangspunkt i de 458 postene i begravelsesdatasettet, og sjekket hvert felt mot hver av de 981 postene i tilsvarende datasett for dåpsdata. Dette resulterte i 649 mulige datalenker, hvor en post i dåpsdatasettet kan danne datapar med flere poster i begravelsesdatasettet og en post i begravelsesdatasettet kan danne datapar med flere poster i dåpsdatasettet[12]. Det kan også finnes dublettpar som ikke filtreres bort.

P4.0 Filter på mulige lenker.
De 884 postene fra dåpsdatasettet representerer alle mulige kandidater for en positiv datalenke. Algoritmen som filtrerer ut de kandidatene med størst sannsynlighet har følgende kriterier:

- Hvis kun en kandidat, trekk den ut fra datasettet
- Hvis flere kandidater, trekk den ut med høyest sannsynlighet og legg de andre tilbake
- Hvis flere kandidater med like høy sannsynlighet, trekk ingen ut og legg alle tilbake
- Hvis det er ingen kandidater, legg tilbake

Hver post i begravelsesdatasettet som inngår i et datapar av de 649 postene evalueres mot hver av de andre parene hvor posten inngår.

[12] Denne sammenhengen er vist i figur 2.

Alle kandidatene som trekkes ut fra de mulige kandidatene samles i et datasett med positive lenker. Datasettet med mulige kandidater er fortsatt klassifisert som mulige kandidater siden finnes kandidater med like høy sannsynlighet hvor begge ble lagt tilbake. En utvidelse av algoritmen vil med høy sannsynlighet trekke ut flere fra datasettet med mulige datalenker.

P4.1 Positive lenker.

Datasettet innholder treff på 649 datapar med mulige datalenker. Av disse dataparene har algoritmen med bakgrunn i de gitte kriteriene trukket ut 187 datapar og tolket disse som positive treff. **Figur 7** viser et utsnitt av disse 187 postene sortert etter fornavn. Under kolonnen for sannsynlighet (PROB) kan man se at sannsynlighetsintervallet spenner seg fra 1 % til 95 %, siden et av kriteriene var hvis det var kun en kandidat med skulle denne trekkes ut fra utvalget av mulige kandidater. Blant disse 187 positive treffene vil det nok finnes noen dubletter som ikke har blitt filtrert ut, slik at det totale antallet er nok noe lavere.

Figur 7.

Utsnitt av de 187 positive treff fra hele datasettet med datapar.

Dataoutput

Datasett filter | Ny SQL lenking | Kandidater | Datalogg | Innverdier/Konstanter

Antall ener kandidater: 187

FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	FDATO_S	FDATO_L	FDATO_U	TREFF	PROB
ALEXANDER	ALEKSONDER	CHRISTEN	KRESTEN	0	06.10.1842	06.10.1842	06.10.1842	2,2,2,2,0,2	0,95
ALEXANDER	ALEKSONDER	A.	E	0	07.05.1841	07.05.1841	07.05.1841	2,2,1,1,0,1	0,34
ANDERS	ANDRS	ANDREAS	ANDRES	0	06.05.1842	06.05.1842	06.05.1842	2,2,1,1,0,2	0,48
ANDERS	ANDRS	OLE	OLE	0	24.05.1849	24.05.1849	24.05.1849	2,2,2,2,0,2	0,95
ANDERS	ANDRS	AXEL	AKSEL	0	30.08.1840	30.08.1840	30.08.1840	2,2,1,1,0,1	0,34
ANDREA	ANDRE	OLE	OLE	1	27.11.1843	27.11.1843	27.11.1843	2,2,2,2,0,1	0,67
ANDREA	ANDRE	AMUND	AMON	1	09.03.1848	09.03.1848	09.03.1848	0,0,2,2,0,1	0,01
ANDREA	ANDRE	OUGEN	OGEN	1	23.04.1841	23.04.1841	23.04.1841	2,2,1,1,0,1	0,34
ANDREAS	ANDRES	SIMON	SEMEN	0	10.04.1842	10.04.1842	10.04.1842	1,2,1,2,0,1	0,34
ANDREAS	ANDRES	MIKAEL	MEKEL	0	16.11.1848	16.11.1848	16.11.1848	2,2,1,1,0,2	0,48
ANDREAS	ANDRES	TORKILD	TORKEL	0	13.03.1846	13.03.1846	13.03.1846	2,2,1,1,0,1	0,34
ANDREAS	ANDRES	SIMON	SEMEN	0	10.04.1842	10.04.1842	10.04.1842	0,0,2,2,0,1	0,01
ANE MARIE	AN MORE	ARNE	ARNE	1	04.03.1841	04.03.1841	04.03.1841	1,2,1,2,0,1	0,34
ANE THORINE	AN TOREN	OUGEN	OGEN	1	13.03.1842	13.03.1842	13.03.1842	2,2,2,2,0,1	0,67
ANNE	AN	DAVID	DOVE	1	29.12.1848	29.12.1848	29.12.1848	2,2,2,2,0,1	0,67
ANNE HELENE	AN HELEN	HANS	HONS	1	30.09.1845	30.09.1845	30.09.1845	2,2,1,1,0,1	0,34
ANNE HELLENE	AN HELEN	JOHN	JON	1	06.01.1845	06.01.1845	06.01.1845	1,2,1,2,0,1	0,34
ANNE MARIA	AN MORE	NILS	NELS	1	07.11.1849	07.11.1849	07.11.1849	1,2,1,2,0,1	0,34

P4.1.1 Aldersgruppering av positive lenker.

Hvis man deler opp de positive lenkene i forskjellige alderintervall kan man vurdere datagrunnlaget og treffprosenten intervallet. **Figur 8** viser 6 aldersintervallene, totalen for året med et snitt av verdiene. Særlig for aldersgruppen av personer yngre enn ett år er datagrunnlaget svakt, og tydelig underrapportering av personer under 1 måned. Antall positive lenker ligger gjennomsnitt mellom 28,6 % og 62,7 % i perioden.

Figur 8.

Andel positive lenker for forskjellige aldergrupper i begravelsesdatasettet.

Dataoutput

Datasett filter | Ny SQL lenking | Kandidater | Datalogg | Innverdier/Konstanter

Kandidat data

Aldersgruppering: BG - (treff)

År	%<1mnd	%1-6mnd	%6-12mnd	%1-2år	%2-5år	%>5år	Antall
1839	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(1)INF%	0-(1)INF%
1840	4-(0),0%	0-(0)NaN%	4-(4)100,0%	4-(3)75,0%	5-(2)40,0%	41-(10)24,4%	58-(19)32,8%
1841	0-(0)NaN%	1-(0),0%	3-(3)100,0%	1-(1)100,0%	6-(4)66,7%	35-(12)34,3%	46-(20)43,5%
1842	0-(0)NaN%	5-(1)20,0%	8-(2)25,0%	6-(2)33,3%	2-(1)50,0%	35-(11)31,4%	56-(17)30,4%
1843	0-(0)NaN%	1-(0),0%	3-(0),0%	3-(1)33,3%	3-(2)66,7%	25-(7)28,0%	35-(10)28,6%
1844	0-(0)NaN%	1-(0),0%	1-(0),0%	2-(1)50,0%	9-(3)33,3%	35-(12)34,3%	48-(16)33,3%
1845	4-(3)75,0%	3-(3)100,0%	3-(1)33,3%	3-(2)66,7%	9-(4)44,4%	29-(19)65,5%	51-(32)62,7%
1846	0-(0)NaN%	5-(4)80,0%	2-(2)100,0%	4-(3)75,0%	4-(1)25,0%	29-(14)48,3%	44-(24)54,5%
1847	0-(0)NaN%	3-(1)33,3%	1-(0),0%	7-(4)57,1%	5-(3)60,0%	21-(7)33,3%	37-(15)40,5%
1848	1-(0),0%	4-(1)25,0%	1-(1)100,0%	2-(2)100,0%	2-(2)100,0%	26-(9)34,6%	36-(15)41,7%
1849	3-(1)33,3%	3-(1)33,3%	6-(2)33,3%	1-(1)100,0%	7-(1)14,3%	27-(12)44,4%	47-(18)38,3%
1850	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%	0-(0)NaN%

P4.2 Ikke-lenker

Det ble opprettet 649 mulige datapar, mange av disse postene utgjorde kryssreferanser. Det vil si at en post i begravelsesdatasettet traff mot flere felter i flere ulike dåpsposter.

I gjennomsnitt ble det dannet nesten 4 datapar per post i begravelsesdatasettet som inngikk i et datapar. **Figur 9** viser et utsnitt av de 284 postene fra dåpsdatasettet som ble klassifisert som ikke positive lenker. Siden andre datapar ble trukket ut med høyre sannsynlighet, er det en fellesnevner for dette utnittet at sannsynlighetene (PROB feltet) ligger mellom 1 % og 34 %. En post med et vanlig fornavn eller farnavn, for eksempel *Hans*, danner mange mulige datapar, med kun treff i kombinasjonen fornavn og fødselsdatointervall, vil sannsynligheten for et positivt treff være lavt ved treff i kun to felt.

Figur 9.

Utsnitt av 284 ikke positive lenker fra begravelsesdata.

Antall ikke funnet kandidater: 284

	FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	FDATO_S	FDATO_L	FDATO_U	TREFF	PROB
▶	OLE	OLE	HANS	HONS	0	19.02.1839	19.02.1839	19.02.1839	0,0,2,2,0,1	0,01
	ANDERS	ANDRS	HANS	HONS	0	06.06.1839	06.06.1839	06.06.1839	0,0,2,2,0,1	0,01
	OLINE MARIE	OLEN MORE	LARS	LORS	1	31.05.1839	31.05.1839	31.05.1839	0,0,2,2,0,1	0,01
	JOHANNE	JOHON	HANS	HONS	1	25.04.1839	25.04.1839	25.04.1839	0,0,2,2,0,1	0,01
	JENS	JENS	PAUL	POL	0	13.07.1839	13.07.1839	13.07.1839	2,2,1,1,0,1	0,34
	ANE HELENE	AN HELEN	HANS	HONS	1	01.08.1839	01.08.1839	01.08.1839	0,0,2,2,0,1	0,01
	CHRISTINE M	KRESTEN M	HANS	HONS	1	05.10.1839	05.10.1839	05.10.1839	0,0,2,2,0,1	0,01
	JOHAN EDVAR	JOHON EDV	HANS	HONS	0	24.10.1839	24.10.1839	24.10.1839	0,0,2,2,0,1	0,01
	GEORG	JOR	LARS	LORS	0	15.12.1839	15.12.1839	15.12.1839	0,0,2,2,0,1	0,01
	HANS	HONS	NICOLAI	NEKOLO	0	07.01.1840	07.01.1840	07.01.1840	2,2,1,1,0,1	0,34
	HANS	HONS	NICOLAI	NEKOLO	0	07.01.1840	07.01.1840	07.01.1840	2,2,1,1,0,1	0,34
	HANS	HONS	NILS	NELS	0	25.01.1840	25.01.1840	25.01.1840	2,2,1,1,0,1	0,34
	HANS	HONS	NILS	NELS	0	25.01.1840	25.01.1840	25.01.1840	2,2,1,1,0,1	0,34
	ELEN MARIE	ELEN MORE	HANS	HONS	1	18.04.1840	18.04.1840	18.04.1840	0,0,2,2,0,1	0,01
	ELEN MARIE	ELEN MORE	HANS	HONS	1	18.04.1840	18.04.1840	18.04.1840	0,0,2,2,0,1	0,01
	ELEN MARIE	ELEN MORE	HANS	HONS	1	18.04.1840	18.04.1840	18.04.1840	0,0,2,2,0,1	0,01
	ELEN MARIE	ELEN MORE	HANS	HONS	1	18.04.1840	18.04.1840	18.04.1840	0,0,2,2,0,1	0,01
	ELEN MARIE	ELEN MORE	HANS	HONS	1	18.04.1840	18.04.1840	18.04.1840	0,0,2,2,0,1	0,01

Av utsnittet kan det ved visuell tolking antas at noen er dubletter som gir lik sannsynlighet og derfor ikke tolkes av modellen, for eksempel *Edvard* uten kjent farnavn og mulig født i år 1842.

Figur 10.

En post som ikke ble lenket i begravelsesdatasettet

EDVARD	EDVOR	(null)	(null)	0	730	99.99.1842	29.05.1841	29.05.1842

Alle poster i **figur 11** er mulige kandidater for denne ikke-lenkbare posten i **figur 10** fra begravelsesdata. Alle kandidatene er like sannsynlige treff for *Edvard*, siden de treffer på fornavn, farnavn og er innenfor det estimerte fødselsdatointervallet. Siden alle kandidatene gir den samme sannsynlighet forkastes dem etter kriteriene gitt under punktet P4.0

Figur 11.

Kandidater fra dåpsdatasettet som er mulige treff med *Edvard*.

FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	FDATO_S	FDATO_L	FDATO_U
EDVARD	EDVOR	OLE	OLE	0	12.06.1841	12.06.1841	12.06.1841
EDVARD	EDVOR	SIMON	SEMEN	0	11.12.1841	11.12.1841	11.12.1841
EDVARD	EDVOR	JENS	JENS	0	28.12.1841	28.12.1841	28.12.1841
EDVARD	EDVOR	ANDERS	ANDRS	0	29.04.1842	29.04.1842	29.04.1842

P4.3 Mulige-lenker.

Av de totalt 458 postene i begravelsesdatasettet klarte programmet å finne 187 positive treff. De gjenværende 271 postene (**figur12**) av det opprinnelige filtrerte begravelsesdatasettet har en særlig høy frekvens av manglende data for farsnavn, dette medfører mange mulige datapar med lavere sannsynlighet for positiv lenking.

Figur 12.

Utsnitt av de 271 ikke lenkbare postene i begravelsesdatasettet.

Antall igjen i BG datasett: 271

	FORNAVN	FORN_S	FARNAVN	FARN_S	SEX	AGE_DAYS	FDATO_S	FDATO_L	FDATO_U
▶	(null)	(null)	(null)	(null)	0	30	99.02.1840	04.01.1840	03.02.1840
	ANE	AN	ANDREAS	ANDRES	1	30	99.05.1840	08.04.1840	08.05.1840
	(null)	(null)	(null)	(null)	8	8	22.07.1840	22.07.1840	22.07.1840
	LAURITZ	LORES	(null)	(null)	0	6	24.10.1840	24.10.1840	24.10.1840
	OLAVA	OLOVE	OUGEN	OGEN	1	545	99.01.1840	23.12.1839	22.01.1840
	PEDER	PETER	(null)	(null)	0	90	99.11.1841	27.10.1841	26.11.1841
	MARTINUS	MOTENOS	(null)	(null)	0	90	99.01.1842	18.12.1841	17.01.1842
	JOHANNES	JOHONES	(null)	(null)	0	90	99.01.1842	26.12.1841	25.01.1842
	HELENE	HELEN	ELLINGS	ELENGS	1	180	99.02.1842	10.01.1842	09.02.1842
	HANS JØRG	HONS JORG	(null)	(null)	0	90	99.06.1842	18.05.1842	17.06.1842

Tabell 4 oppsummerer resultatene av gjennomgangen for dataflyten og programlogikken. Av de 458 postene fra begravelsesdata som ble sammenliknet med 884 poster fra dåpsdata, ble det dannet 649 datapar. Av disse dataparene ble 187 klassifisert som positive lenker. Det var igjen 271 poster som ikke var lenkbare, dette utgjør 59 % av de originale data. Det var 284 poster igjen med mulige lenker blant dåpsdata. Jeg ønsker videre å benytte noen tester for å prøve å besvare hvorfor ikke flere data ble gjenfunnet.

Tabell 4.

Oversikt over antall poster gjennom dataflyten.

Variabel	**Begravelsesfil**	**Dåpsfil**
Antall poster	458	884
Antall datapar	649	
Positive lenker	187 (41 %)	
Ikke lenkbare	271 (59 %)	284

7. Tolkning av empiriske resultater

Jeg vil i dette kapitelet gjennomgå tre forskjellige tester med resultater. Testene er utført for å prøve og validere datalenkingen gjort av mine algoritmer, et annet ønske er å klassifisere og kvantifisere de mulige feilkildene for at mulige faktiske positive lenker blir forkastet av programmet. Siden det ikke er sannsynlig å forvente og gjenfinne 100 % av personene i begravelsesdatasettet i dåpsdatasettet, ønsker jeg ved testene å vise hvor høy treffsannsynlighet man kan forvente å oppnå ved å kombinere den maskinelle datalenkingen med en manuell datalenking. Den mest presise metoden for å sjekke om datamodellen og algoritmene klarer å lenke sammen og velge ut de beste dataparene er ved manuell observasjon av resultatene.

Jeg velger et lite utvalg fra begravelsesdatasettet filtrert på lav alder fra kun et fødselsår. Et lite utvalg gjør det mulig å analysere data og klassifisere feilkilden. Ved å velge personer som hadde en alder ved begravelse på under 5 år, antar jeg at usikkerheten rundt migrasjon reduseres og at sannsynligheten for rekonstruksjon av data øker.
Selv med et lite utvalg av hele datasettet vil belyse de generelle forhold.

7.1 Validitetstest 1 av modellen

Kildedata i test 1 er hentet fra prestegjeld 1 med personer fra begravelsesdata som er 5 år eller yngre og estimert født i år 1849. Det filtrerte datasettet innholder 20 poster i begravelsesdatasettet og 215 personer i dåpsdatasettet med tidsintervall pluss/minus et år.

En tolkning av inndata i **tabell 5** viser at det er ufullstendige data for farsnavn og fødselsdato. Av de 12 personene har 55 % ufullstendig fødselsinformasjon og 55 % har ufullstendige farsnavn.

Tabell 5.

Test 1: Prosentandel av full feltinformasjon for de 20 personer i datasettet for begravelse

Fornavn	Forn_s	Farnavn	Farn_s	Sex	Fdato	%Snitt/Sum
100,0% - (20)	100,0% - (20)	55,0% - (11)	55,0% - (11)	100,0% - (20)	55,0% - (11)	77,5% - (20)

Det filtrerte datasettet innholder 28 mulige kandidater og etter bearbeiding viser programmet 13 positive lenker for de 20 postene i begravelsesdatasettet.

Oppgaven videre blir å sjekke manuelt hvorfor de gjenværende 7 (35 %) postene, vist i **tabell 6**, ikke ble lenket av programmet.

Tabell 6.

Test 1: Gjenværende 7 personer fra begravelsesdata som ikke ble gjenfunnet i dåpsdata

Fornavn	Forn_s	Farnavn	Farn_s	Alder	Fdato_S	Fdato_L	Fdato_U
ELEN MARIA	*ELEN MORE*	*(null)*	*(null)*	*35*	*99.02.1849*	*31.01.1849*	*07.02.1849*
LOVISE LAURENTSE	*LOESE LORENSE*	*(null)*	*(null)*	*90*	*99.01.1849*	*16.12.1848*	*15.01.1849*
OLAVES	*OLOS*	*OLE*	*OLE*	*21*	*04.04.1849*	*28.03.1849*	*04.04.1849*
CHRISTEN EMILIUS	*KRESTEN EMELOS*	*(null)*	*(null)*	*239*	*05.10.1849*	*05.10.1849*	*05.10.1849*
HANS HENRIK	*HONS HENREK*	*(null)*	*(null)*	*850*	*99.06.1849*	*01.06.1849*	*30.06.1849*
ANNE LAURENTIA	*AN LORENTE*	*(null)*	*(null)*	*1000*	*99.11.1849*	*13.10.1849*	*12.11.1849*
BERTE MARIA	*BERTE MORE*	*(null)*	*(null)*	*1095*	*99.99.1849*	*20.09.1848*	*20.09.1849*

Tabell 7 viser de 2 personene som ikke kan gjenfinnes blant personene i dåpsdatasettet. Ingen personer finnes med liknende fornavn, og siden farnavnet er ukjent kan det heller benyttes for å angi som et entydig treff.

Tabell 7.

Test 1: To personer fra begravelsesdatasettet som ikke gjenfunnet ved manuelt søk.

Fornavn	fornavn_s	farnavn	farnavn_s	Fdato_s	fdato_l	fdato_u
LOVISE LAURENTSE	*LOESE LORENSE*	*(null)*	*(null)*	*99.01.1849*	*16.12.1848*	*15.01.1849*
HANS HENRIK	*HONS HENREK*	*(null)*	*(null)*	*99.06.1849*	*01.06.1849*	*30.06.1849*

Av de resterende 5 personene kan 3 av personene i dåpsdatasettet tolkes som positive lenker som vist i **tabell 8**[13]. Årsaken til at personene ikke ble fanget opp av programmet var fordi fornavnet var forskjellig skrevet og fordi farsnavnet manglet helt. Ved et tillegg i beslutningsmodellen for algoritmen med sjekk for standardisert fornavn og større bredde i intervallet for fødselsdato ville disse personene vært tolket som positive lenker.

Tabell 8.

Test 1: Positive lenker ved manuell sjekk av fornavn.

Fornavn	fornavn_s	farnavn	farnavn_s	Fdato_s	fdato_l	fdato_u
CHRISTEN EMILIUS	*KRESTEN EMELOS*	*(null)*	*(null)*	*05.10.1849*	*05.10.1849*	*05.10.1849*
KRISTEN EMILIUS	KRESTEN EMELOS	ANDERS	ANDRS	05.10.1849	05.10.1849	05.10.1849
ANNE LAURENTIA	*AN LORENTE*	*(null)*	*(null)*	*99.11.1849*	*13.10.1849*	*12.11.1849*
ANNA LAURENTIA	AN LORENTE	P.F.	P F	01.10.1849	01.10.1849	01.10.1849
BERTE MARIA	*BERTE MORE*	*(null)*	*(null)*	*99.99.1849*	*20.09.1848*	*20.09.1849*
BERTE MARIE	BERTE MORE	PAUL	POL	14.09.1849	14.09.1849	14.09.1849

De 2 siste personene har så mange kandidater at programmet ikke kunne rangere mellom dem. Ved manuell sjekk av **tabell 9** ser jeg at det er 2 kandidater som kan klassifiseres som mulige lenker. For personer uten registret farnavn, vil alle personer med samme fornavn kunne antas å

[13] Tabellen er en blanding av data fra de to datasettene, raden i kursiv med grå bakgrunn viser posten fra begravelsesdatasettet, direkte under med hvit bakgrunn vises posten i dåpsdatasettet som er mest sannsynlig. Dette formatet vil bli fulgt for tabellene som viser testresultater.

danne datapar. I tabellen er personen *Elen Maria* ikke registrert med farnavn og det finnes 6 personer i dåpsdatasettet med dette navnet.

Tabell 9.

Test 1: Mulige kandidater til postene *elen maria* og *olaves* fra begravelsesdatasettet.

Fornavn	fornavn_s	Farnavn	farnavn_s	Fdato_s	fdato_l	fdato_u
ELEN MARIA	*ELEN MORE*	*(null)*	*(null)*	*99.02.1849*	*31.01.1849*	*07.02.1849*
ELEN MARIA	ELEN MORE	PEDER	PETER	13.02.1849	13.02.1849	13.02.1849
ELEN MARIA	ELEN MORE	HANS	HONS	14.03.1849	14.03.1849	14.03.1849
ELEN MARIA	ELEN MORE	NILS	NELS	10.12.1849	10.12.1849	10.12.1849
ELEN MARIA	ELEN MORE	OLE	OLE	15.03.1850	15.03.1850	15.03.1850
ELEN MARIA	ELEN MORE	JOHANNES	JOHONES	20.09.1850	20.09.1850	20.09.1850
ELEN MARIA	ELEN MORE	ANDREAS	ANDRES	18.11.1850	18.11.1850	18.11.1850
OLAVES	*OLOS*	*OLE*	*OLE*	*04.04.1849*	*28.03.1849*	*04.04.1849*
OLAVES	OLOS	OLE	OLE	11.03.1849	11.03.1849	11.03.1849
OLAVES	OLOS	OLE	OLE	22.07.1849	22.07.1849	22.07.1849
OLAVES	OLOS	OLE	OLE	24.12.1849	24.12.1849	24.12.1849
OLAVES	OLOS	OLE	OLE	25.03.1850	25.03.1850	25.03.1850

Programmet har ingen logikk for rangering av like treff, i tilfeller med mange nesten identiske treff må de derfor klassifiseres som mulige lenker, siden man heller ikke kan fastslå at de er positive ikke-lenker. **Tabell 10** viser at hele 90 % ble gjenfunnet manuelt.

Tabell 10.

Test 1: Resultater av test 1.

Positive lenker gjenfunnet maskinelt	13
Positive lenker	3
Ikke funnet	2
Mulige lenker	2
Gjenfunnet av programmet	65 %
Gjenfunnet manuelt	90 %

7.2 Validitetstest 2 av modellen

Testen har blitt gjennomført mot prestegjeld 1 for personer yngre enn 5 år født i år 1830. Antall personer i begravelsesdatasettet er 17 personer, og dåpsdatasettet inneholder 215 personer fra år 1829 til og med år 1831.

Tabell 11 viser en tolkning av inndata hvor det er ufullstendige data for farnavn og fødselsdato. Av de 17 personene har 100 % ufullstendig fødselsinformasjon og 41 % har ufullstendig informasjon om farnavn.

Tabell 11.

Test2: Graden av full info. for personer yngre enn 5 år født i år 1830.

Fornavn	Forn_s	Farnavn	Farn_s	Fdato	%Snitt/Sum
100,0% (17)	100,0% (17)	58,8% (10)	58,8% (10)	0% (0)	69,6% (17)

Programmet finner 9 (53 %) positive lenker for de 17 personer begravelsesdatasettet. **Tabell 12** viser de 8 personene fra begravelsesdatasettet som ikke ble gjenfunnet. En manuell sjekk kan vise hvorfor de gjenværende 8 (47 %) ikke ble lenket av programmet.

Tabell 12.

Test2: 8 personer fra begravelsesdata som ikke ble gjenfunnet

Fornavn	Forn_s	Farnavn	Farn_s	Sex	Alder	Fdato_S	Fdato_L	Fdato_U
MARTIN THORVALD	*MOTEN TORVOL*	*ANDREAS*	*ANDRES*	*0*	*30*	*99.02.1830*	*09.01.1830*	*08.02.1830*
FREDERICH	*FREDREK*	*(null)*	*(null)*	*0*	*635*	*99.11.1830*	*28.10.1830*	*27.11.1830*
JENS	*JENS*	*(null)*	*(null)*	*0*	*820*	*99.11.1830*	*18.10.1830*	*17.11.1830*
KISTINE	*KJTEN*	*ANDERS*	*ANDRS*	*1*	*1275*	*99.02.1830*	*25.01.1830*	*24.02.1830*
INGER HELENE	*ENGER HELEN*	*(null)*	*(null)*	*1*	*1825*	*99.99.1830*	*23.01.1829*	*23.01.1830*
HANS	*HONS*	*(null)*	*(null)*	*0*	*1640*	*99.11.1830*	*27.10.1830*	*26.11.1830*
OLENE MARIE	*OLEN MORE*	*(null)*	*(null)*	*1*	*1825*	*99.99.1830*	*07.06.1829*	*07.06.1830*
NILS	*NELS*	*OUGEN*	*OGEN*	*0*	*1730*	*99.10.1830*	*08.09.1830*	*08.10.1830*

Tabell 13 viser de 2 personene som ikke kan gjenfinnes blant postene i dåpsdatasettet. Ved en manuell sjekk finner jeg ingen personer med denne kombinasjonen av fornavn og farnavn i dåpsdatasettet. Søket mitt med ulike kombinasjoner med fødseldato gir heller ingen gode kandidater.

Tabell 13.

Test 2: To personer fra begravelsesdatasettet som ikke gjenfunnet ved manuelt søk.

Fornavn	fornavn_s	Farnavn	farnavn_s	Fdato_s	fdato_l	Fdato_u
MARTIN THORVALD	*MOTEN TORVOL*	*ANDREAS*	*ANDRES*	*99.02.1830*	*09.01.1830*	*08.02.1830*
NILS	*NELS*	*OUGEN*	*OGEN*	*99.10.1830*	*08.09.1830*	*08.10.1830*

Tabell 14 viser tre personer som har flere kandidater fra dåpsdatasettet, men ingen av kandidatene tolkes som sterke nok for å klassifisere som mulige lenker, selv om datoinformasjonen ligger innenfor det estimerte intervallet.

Den estimerte fødselsdatoen settes strengt i dåpsdata hvis informasjonen i kildedata tolkes som god. Dette kan ses i den andre posten med *Olene Marie*, som har et dobbelt fornavn og det kan antas at både *Olene* og *Marie* er gode kandidater. Fødselsdato informasjonen til både *Olene* og *Marie* tolkes som god, og de har derfor ikke noe estimert intervall. På dette grunnlaget forkastes begge kandidatene.

For *Hans* er problemet at det ikke er informasjon om farnavn og fødselsdag angivelsen mangler. Siden Hans er et vanlig navn finnes det mange alternativer, men ingen av dem er sterke nok til å klassifisere som en positiv lenke.

Disse tolkes derfor for ubestemmelige, siden ingen informasjon er god nok til å avgjøre den beste kandidat.

Tabell 14.

Test 2: 3 personer fra begravelsesdata med flere mulige kandidater i dåpsdata

Fornavn	fornavn_s	Farnavn	farnavn_s	Fdato_s	fdato_l	Fdato_u
KISTINE	*KJTEN*	*ANDERS*	*ANDRS*	*99.02.1830*	*25.01.1830*	*24.02.1830*
KISTINE	KJTEN	OLE	OLE	30.05.1829	30.05.1829	30.05.1829
KISTINE	KJTEN	HANS	HONS	24.01.1830	24.01.1830	24.01.1830
OLENE MARIE	*OLEN MORE*	*(null)*	*(null)*	*99.99.1830*	*07.06.1829*	*07.06.1830*
OLENE	OLEN	JOHANNES	JOHONES	22.10.1830	22.10.1830	22.10.1830
MARIE	MORE	INGEBRICT	EMBRET	05.05.1830	05.05.1830	05.05.1830
HANS	*HONS*	*(null)*	*(null)*	*99.11.1830*	*27.10.1830*	*26.11.1830*
HANS	HONS	MONS	MONS	06.03.1830	06.03.1830	06.03.1830
HANS	HONS	GUNDRO	GONRO	25.12.1830	25.12.1830	25.12.1830
HANS	HONS	PEDER	PETER	02.04.1831	02.04.1831	02.04.1831
HANS	HONS	LARS	LORS	30.04.1831	30.04.1831	30.04.1831
HANS	HONS	SVEND	SVEN	04.06.1831	04.06.1831	04.06.1831
HANS	HONS	CHRISTEN	RESTEN	30.11.1831	30.11.1831	30.11.1831

De resterende tre personene i **tabell 15** mangler i tillegg til farnavn informasjon kandidater innenfor estimert fødselsdatointervall, slik at de klassifiseres som ikke funnet.

Tabell 15.

Test 2: Kandidater fra begravelsesdata med flere ubestemmelige kandidater

Fornavn	Fornavn_s	Farnavn	farnavn_s	Fdato_s	fdato_l	Fdato_u
FREDERICH	*FREDREK*	*(null)*	*(null)*	*99.11.1830*	*28.10.1830*	*27.11.1830*
FREDERICH	FREDREK	HANS	HONS	22.04.1829	22.04.1829	22.04.1829
FREDERICH	FREDREK	LARS	LORS	14.10.1830	14.10.1830	14.10.1830
JENS	*JENS*	*(null)*	*(null)*	*99.11.1830*	*18.10.1830*	*17.11.1830*
JENS	JENS	PEER	PETER	06.02.1829	06.02.1829	06.02.1829
JENS	JENS	ANDERS	ANDRS	19.10.1829	19.10.1829	19.10.1829
JENS	JENS	JOHANNES	JOHONES	03.12.1830	03.12.1830	03.12.1830
JENS	JENS	JØRGEN	JORGEN	10.10.1831	10.10.1831	10.10.1831
INGER HELENE	*ENGER HELEN*	*(null)*	*(null)*	*99.99.1830*	*23.01.1829*	*23.01.1830*
INGER HELENE	ENGER HELEN	NILS	NELS	02.04.1830	02.04.1830	02.04.1830
INGER HELENE	ENGER HELEN	KJØNIG	KJENEK	12.06.1831	12.06.1831	12.06.1831

Ved oppsummering av test 2, kan vi se av **tabell 16** at programmet fant alle de positive lenkene. Det ble ikke funnet flere positive lenker ved en manuell kontroll. Ved den manuelle kontrollen ble ytterligere fem datapar klassifisert som mulige, men med en høy grad av usikkerhet. Totalt lenket programmet på 53 % av personene i begravelsesdatasettet.

Tabell 16. Resultater av test 2.	
Positive lenker gjenfunnet maskinelt	9
Positive lenker funnet manuelt	0
Ikke funnet	5
Mulige lenker	3
Gjenfunnet av programmet	53 %
Gjenfunnet manuelt	62 %

7.3 Validitetstest 3 av modellen

Motivasjonen for denne testen er å sjekke om data etter 1890 med mer fullstendig informasjon har en høyere grad av gjennfinnelsesprosent enn data fra år 1830 og år 1849. Testen har blitt gjennomført mot prestegjeld 1 i år 1890 til og med år 1897 for personer yngre enn 5 år. Antall personer i begravelsesdatasettet som er født i dette tidsrommet er 71 personer, og dåpsdatasettet inneholder 796 personer fra år 1889 til og med år 1898. **Tabell 17** viser en tolkning av inndata viser at det er ufullstendige data for farsnavn og fødselsdato. Av de 71 personene har 100 % fullstendig fødselsinformasjon og for navnefeltene er små variasjoner i feltinformasjonen foruten farsnavn for 1897 hvor 43 % har ufullstendig informasjon om farsnavn. Feltinformasjonen er mer fullstendig i disse årene, noe som kommer av nye regler for føring av kirkeboken[14].

[14] Oversikt over påkrevde felt vises i tabell 24 og tabell 25 i appendiks.

Tabell 17.

Test3: For år 1890 til år 1897 med personer yngre enn 5 år.

År	Fornavn	Forn_s	Farnavn	Farn_s	Fdato	snitt/sum
1890	100% (14)	100% (14)	92,9% (13)	92,9% (13)	100% (14)	100% (14)
1891	100% (7)	100% (7)	100% (7)	100% (7)	100% (7)	100% (7)
1892	100% (10)	100% (10)	90,0% (9)	90,0% (9)	100% (10)	100% (10)
1893	88,9% (8)	88,9% (8)	77,8% (7)	77,8% (7)	100% (9)	100% (9)
1894	90,9% (10)	90,9%(10)	81,8% (9)	81,8% (9)	100% (11)	100% (11)
1895	100% (9)	100% (9)	77,8% (7)	77,8% (7)	100% (9)	100,0% (9)
1896	75,0% (3)	75,0% (3)	75,0% (3)	75,0% (3)	100% (4)	100,0% (4)
1897	85,7% (6)	85,7% (6)	57,1% (4)	57,1% (4)	100% (7)	100,0% (7)

I **tabell 18** vises de 12 personene som ikke ble gjenfunnet av de 71 personene i begravelsesdatasettet. Av disse igjen er det 4 personer uten noen navneinformasjon, noe som gjør det umulig å klassifisere dem som positive lenker. De siste 8 skal jeg teste manuelt.

Tabell 18.

Test3: 12 personer fra begravelsesdata som ikke ble gjenfunnet

Fornavn	Forn_s	Farnavn	Farn_s	Sex	Alder	Fdato_S	Fdato_L	Fdato_U
JOHAN	*JOHON*	*AMUND*	*AMON*	*0*	*95*	*20.03.1890*	*20.03.1890*	*20.03.1890*
ANNE GURINE HEDVIG	*AN GOREN HEDVEK*	*LARS*	*LORS*	*1*	*22*	*25.10.1890*	*25.10.1890*	*25.10.1890*
PEDER VARTINIUS	*PETER VORTENOS*	*GUSTAV*	*GOSTOV*	*0*	*140*	*24.11.1890*	*24.11.1890*	*24.11.1890*
LARS	*LORS*	*HANS*	*HONS*	*0*	*18*	*20.03.1893*	*20.03.1893*	*20.03.1893*
ALF ARTHUR	*AL ARTOR*	*MARTIN*	*MOTEN*	*0*	*110*	*04.07.1894*	*04.07.1894*	*04.07.1894*
CARLOTTA	*SJOLOTE*	*(null)*	*(null)*	*1*	*756*	*04.07.1895*	*04.07.1895*	*04.07.1895*
CLARA LOVISE	*KLORE LOESE*	*(null)*	*(null)*	*1*	*530*	*20.09.1897*	*20.09.1897*	*20.09.1897*
JENS	*JENS*	*AUGUST*	*OGOS*	*0*	*1445*	*16.01.1897*	*16.01.1897*	*16.01.1897*
(null)	*(null)*	*(null)*	*(null)*	*1*	*21*	*20.03.1893*	*20.03.1893*	*20.03.1893*
(null)	*(null)*	*(null)*	*(null)*	*1*	*320*	*12.01.1894*	*12.01.1894*	*12.01.1894*
(null)	*(null)*	*(null)*	*(null)*	*0*	*12*	*05.04.1896*	*05.04.1896*	*05.04.1896*
(null)	*(null)*	*(null)*	*(null)*	*1*	*29*	*09.07.1897*	*09.07.1897*	*09.07.1897*

Ved en manuell sjekk av dåpsdatasettet finner jeg 4 positive lenker til data i begravelsesdatasettet. I **tabell 19** kan man se at personen *Anne* med høy grad av sannsynlighet kan antas å være personen *Anne Gurine Hedvig*, siden også farnavn og fødselsdato stemmer overens. Det samme er gjeldene for personene *Peder* og *Klara*. For disse personene kommer også transkriberingsfeil inn i bildet, og for den siste personen *Lars* er det problemet med dobbeltregistrering som gjelder, siden personen er registrert to ganger i dåpsdatasettet. De resterende 8 personene lar seg ikke gjenfinne ved manuell sjekk av dåpsdatasettet.

Tabell 19.

Test 3: Fire positive lenker gjenfunnet manuelt.

Fornavn	Fornavn_s	Farnavn	farnavn_s	Fdato_s	fdato_l	Fdato_u
ANNE GURINE HEDVIG	*AN GOREN HEDVEK*	*LARS*	*LORS*	*25.10.1890*	*25.10.1890*	*25.10.1890*
ANNE	AN	LARS MARENTIUS	LORS MORENSOS	25.10.1890	25.10.1890	25.10.1890
PEDER VARTINIUS	*PETER VORTENOS*	*GUSTAV*	*GOSTOV*	*24.11.1890*	*24.11.1890*	*24.11.1890*
PEDER MARTINIUS	PETER MOTENOS	OLE GUSTAV	OLE GOSTOV	24.11.1890	24.11.1890	24.11.1890
LARS	*LORS*	*HANS*	*HONS*	*20.03.1893*	*20.03.1893*	*20.03.1893*
(null)	(null)	HANS	HONS	20.03.1893	20.03.1893	20.03.1893
(null)	(null)	HANS	HONS	20.03.1893	20.03.1893	20.03.1893
CLARA LOVISE	*KLORE LOESE*	*(null)*	*(null)*	*20.09.1897*	*20.09.1897*	*20.09.1897*
KLARA LOVISE	KLORE LOESE	LAURITS	LORES	20.09.1897	20.09.1897	20.09.1897

Resultatene for datalenkingen av 71 personer i alderen under 5 år i perioden år 1890 til år 1897, vises i **tabell 20**. Programmet klarte å finne 83 % av de positive lenkene, og hele 89 % ble gjenfunnet etter den manuelle gjennomgangen. Hovedgrunnen for denne høye graden gjenfinning ligger i det faktum at datagrunnlaget var meget godt i denne perioden, med full fødselsdag informasjon for alle personene.

Tabell 20. Resultater av test 3. Totalt 71 personer i begravelsesdatasettet.	
Positive lenker gjenfunnet maskinelt	59
Positive lenker funnet manuelt	4
Ikke funnet	8
Mulige lenker	0
Gjenfunnet av maskinelt	83 %
Gjenfunnet manuelt og maskinelt	89 %

Oppsummeringen av resultatene av mine tre tester vises i **tabell 21**, 67 % ble lenket av programmet mens rundt 80 % av personene i begravelsesdatasettet ble gjenfunnet ved en kombinasjon av maskinell og manuell datalenking. Et resultat jeg mener er meget positivt og godt. Variasjonene tilsier at jo mindre ufullstendig informasjon av data i feltene jo bedre er datagrunnlaget og desto høyere vil sannsynligheten være for mange positive treff i datalenkingen.

Tabell 21. Oppsummering av resultatene for testene.				
Testnummer	**1**	**2**	**3**	**Gj.snitt**
Positive lenker gjenfunnet maskinelt	13	9	59	
Positive lenker funnet manuelt	3	0	4	
Ikke funnet	2	5	8	
Mulige lenker	2	3	0	
Gjenfunnet av maskinelt	65 %	53 %	83 %	**67 %**
Gjenfunnet manuelt og maskinelt	90 %	62 %	89 %	**80 %**

Et videre interessant spørsmål er derfor å klassifisere og kvantifisere hvilke feilkilder som påvirker dette resultatet. Dette ønsker jeg å se videre på i neste del.

7.4 Manuell test av feilkilder.

Algoritmene til mitt dataprogram i kapittel 6 fant rundt 40 % positive lenker fra begravelsesdatasettet. Jeg ønsker videre å analysere de dataparene som ikke ble klassifisert som negative lenker. Jeg antar at et flertall av personene lar seg lenke, og gitt den lave treffprosenten på 41 % stiller jeg spørsmålet:

Hvilke feilkilder gjør seg gjeldene når enkelte poster ikke blir klassifisert som positive lenker?

Jeg har manuelt klassifisert hver av de data i begravelsesdatasettet fra testene som ikke ble lenket maskinelt. De har videre blitt klassifisert og gruppert i syv forskjellige grupper av feilkilder. Ved flere av postene var det alternative feilkilder, det kan eksempelvis være en kombinasjon av flere fornavn og et konservativt estimat for fødselsdato. Antall funnet feil overstiger derfor antallet poster med det doble, siden rangering av feilkildene ikke har vært mulig.

Data som kunne ha vært tolket som treff hvis de estimerte datoverdier hadde vært mindre konservative er lagt i gruppe 1. Personer som ikke kan gjenfinnes manuelt i dåpsdatasettet i gruppe 2. Hvis det finnes flere like gode kandidater i dåpsdatasettet for en post i begravelsesdatasettet summeres disse i gruppe 3. Flere navn i fornavn og/eller farsnavn er gruppert i gruppe 4. Hvis data i datakilden er feil stavet enten med årsak i ortografiske feil i innskrivingen eller etter transkriberingen er disse kategorisert i gruppe 5. Gruppe 6 innholder de lenkene som kunne ha vært lenker hvis algoritmene hadde vært mindre konservative, det vil si hvis det hadde vært sjekket for flere alternative kombinasjoner av felt. Dublettverdier vil vises som like gode kandidater men er trukket ut i en egen gruppe 7, dubletter behøver ikke være identiske felt for felt, det kan forkomme feilføring i en av dem.

Oversikten i **tabell 22** viser min klassifisering av de ca 20 % gjenstående postene i begravelsesdatasettet i ulike kategorier. Summen og prosentandelen indikerer hvordan de fordeler seg ut fra totalen av gjenstående poster. Oppsummerer jeg de manuelt observerte feilkildene fra test 1-3 og i en ytterligere test 4[15], viser det seg at gruppe 1 til og med gruppe 3 er inkludert i 75 % av de klassifiserte feilene. Både fødseldatointervallet og problemet med rangering kan utbedres slik at andelen i disse gruppene reduseres.

Summen av verdier i gruppe 2 er bekrefter antagelsen at en andel av personene flytter til og fra et prestegjeld i løpet av et liv sammen med at det kan være en underrapportering av dødfødte som kun blir registrert i et av registrene. De resterende feilkildene er feil som ved videreutvikling kan korrigeres slik at summen vil nærme seg null, men aldri elimineres siden det alltid vil finnes transkriberingsfeil og ortografiske feil ved en manuell føring som ikke kan avsløres selv ved manuelt ettersyn.

Tabell 22.

Observerte feilkilder ved manuell sjekk av ikke-lenkete personer

Gr.	**Feilkilde**	**1**	**2**	**3**	**4**	**Sum**
1.	For smalt fødselsdato intervall	4	3		9	**26 %**
2.	Ikke funnet i dåpsdata	2	4	8	2	**26 %**
3.	Flere like kandidater	3	3		8	**23 %**
4.	Flere fornavn			2	3	**8 %**
5.	Transkribering/ortografi				4	**7 %**
6.	Programfeil	2			2	**7 %**
7.	Dubletter			1	1	**3 %**
	Totalt	61				**100 %**

[15] Test 4 er utført etter samme prinsipper som test 1-3. Utvalget er fra prestegjeld 3 med personer i begravelsesdatasettet født i år 1849. Av de 37 postene fra år 1849 fant algoritmen 19 positive lenker og 18 ikke-lenket. Testen er ikke presentert av plasshensyn.

8. Betydning av feil og usikkerhet.

Feil og usikkerhet i den automatiske datalenkingen kan som vist i tabell 22 ha stor betydning for andelen av positive lenker så vel for antallet forkastete mulige lenker. Tabellen viser 7 ulike grupper med feilkilder, som hver for seg reduserer graden av treff. Jeg vil videre splitte opp gruppene i to kategorier, en for mulige feilkilder i datakildene og en annen for mulige feilkilder i min datamodell.

8.1 Feil og usikkerhet i datakildene.

Det kan ikke under noen omstendighet forventes å gjenfinne 100 % av postene i begravelsesdatasettet selv for alderintervallet 5 år eller yngre som har blitt benyttet i testene. Til det er det for mange faktorer som reduserer graden av lenking, en utenforliggende faktor som migrasjon gjør at begravelses- eller fødselsdata kun finnes i den ene listen. Feil i ortografi og transkribering gir følgefeil fra kildedata, og dubletter forekommer i begge datasett. Registreringen av dødfødte barn er usikker både med hensyn til om de er registrert i et eller begge registrene og hvilke personalia som har blitt registrert.

Tabell 23 viser feilkilder som årsaker en reduksjon i sannsynligheten for å lenke alle i personene fra begravelsesdatasettet. Sannsynlighetsintervallene er egendefinerte verdier som bare kan tolkes som forsiktige anslag.

Gitt disse anslagsvise verdiene vil det mest pessimistiske estimatet ha en forventet treffprosent rundt 65 %, og med det mest optimistiske ha en forventet treffprosent rundt 85 % med automatisert datalenking. Det er usikkert hvor stor reduksjon av feilkildene man klarer med forbedringer i rutinene. Jeg deler videre opp feilkildene i de primært gitte feilkildene og de sekundært gitte, som kan endres i standardiseringen eller min datamodell.

Tabell 23. Feilkilder som er årsaker til reduksjon i graden av datalenking.	
Primær feilkilder	**Sannsynlighet**
Migrasjon	*3 % - 5 %*
Manglende registrering av dødfødte i original data	*2 % - 5 %*
Manglende informasjon om dødfødte i datasettene	*2 % - 5 %*
Feil og mangler i datakildene	*3 % - 5 %*
Reduksjon: min- og maksverdi	*10 % - 20 %*
Sekundær feilkilder	**Sannsynlighet**
Dubletter i begravelsesdata	*2 % – 5 %*
Feil og mangler i standardiseringen	*2 % - 5 %*
Dubletter i dåpsdata	*2 % – 5 %*
Reduksjon: min- og maksverdi	*6 % - 15 %*
Total reduksjon: min- og maksverdi	**16 % - 35 %**

Hva kan man forvente at den best mulige lenkeprosenten kan være fra originaldata?

Migrasjon er forventet å være lavere for unge barn, og for barn som dør svært unge kan man se bort fra denne effekten. Videre må det antas at den øker opp til en viss alder, ved å velge utsnitt fra datasettene filtrert på personer i det laveste aldersgruppen vil migrasjon synke mot minimumsverdien. Den estimerte sannsynlighetsverdien vil derfor strekt avhenge av hvilken aldersgruppe vi betrakter. For årene etter 1877[16] reduseres sannsynligheten for feil og mangler i datakildene betraktelig. Det vil alltid være større eller mindre grad av migrasjon i en populasjon, avhengig av alder, bosted og tidsperiode. Hvis man i fremtiden får digitalisert alle landets prestegjeld og sokn, kan feilkilden reduseres ved at man kan lenke på tvers av de geografiske grensene.

[16] Kravene om hvilke datafelt som skulle registreres øker etter ny poststruktur i 1877. For begravelsesdata ble det nå påkrevd å registrere fødselsdato og fars navn, noe som gjør at kildedata blir mer komplett i utgangspunktet siden fødselsdato ikke behøver å estimeres fra alder.

Problematikken rundt under- og overrapportering av barn som dør unge eller er dødfødte blir mer komplisert, i mine datasett var et flertall registrert uten andre felt enn fødselsdato og estimert fødselsdato. En tilnærming til den problematikken gjennom forbedret algoritme og standardisering, vil medføre helt ny metodikk og være en tung prosess å gjennomføre. Ved feil og mangler i originaldata vil disse opptre som følgefeil under standardiseringen og i datalenkingen. Det kan utarbeides automatiseringsmetoder slik at usikre poster og felter filtreres ut før en videre manuell sjekk.

Den best mulige lenkeprosenten gitt de primære feilkildene vil anslagsvis være i intervallet 80 % til 90 %.

Hvor nært kan vi komme med utbedringer i de sekundært bestemte feilkildene?

Dublettverdier forekommer både begravelsesdatasettet og dåpsdatasettet, hvor postene er helt identiske felt for felt. Disse kan fjernes ved enkle metoder i standardiseringen, noe mer avansert blir det når dublettverdien i tillegg har ortografiske feil og derfor ikke er entydig identisk. I slike tilfeller må det søkes på nærliggende poster i datasettet, siden det må antas at registreringene sammenfaller i samme tidsrom i kirkeboken som ble ført kronologisk.

Graden av feil og mangler i standardiseringen, varierer i fra periode til periode og fra prestegjeld til prestegjeld. En videreutvikling av metodene og algoritmene i standardiseringsprogrammet til Kåre Bævre, vil kunne redusere påvirkningen av feil og mangler i originaldata. Ved mer effektive og veldefinerte metoder kan denne sekundærgruppen feilkilder reduseres mot 0 %.

8.2 Feil og usikkerhet i datamodellen

Det er i all hovedsak to typer feil man kan gjøre i klassifiseringen av et lenket par, parene kan enten være falske negativer eller de kan være falske positiver. Min beslutningsmodell har vært konservativ i den forstand at algoritmen kun klassifiserer enkelt treff og den beste av mange som positive lenker. Gjennomgangen av eksemplene i kapittel 7 viser at det er flere usikkerhetsmomenter i datamodellen.

Det er påvist ved manuell gjennomgang av data i fødselsdata at det finnes personer som er registret ved flere anledninger, disse dublettverdiene opptrer som identiske eller på grunn av ortografiske feil som nesten identiske, det er ingen test for dublettverdier før behandlingen av data i modellen. Hvis dublettverdiene gir et positivt treff og er den mest sannsynlige datalenken, vil ikke algoritmen kunne velge hvilken av de to som er det beste treffet. Begge blir derfor lagt tilbake i datasettet med mulige lenker.

Et annet usikkerhetsmoment er om treffet er en falsk positiv lenke, denne usikkerheten er høyere når det er kun et datapar for posten og den velges uavhengig av sannsynlighet. I modellen tolkes treff helt ned til 1 % som et positivt treff, man må jo anta at usikkerheten øker jo lavere sannsynligheten blir. Et tiltak for å sikre bedre kvalitet på klassifiseringen av positive datalenker burde være å legge inn en minimumsverdi for sannsynligheten.

Hvis det er to positive treff, må jo en være falsk positiv, gitt at de ikke er dublettverdier. Skal begge forkastes hvis ikke en av dem inngår som beste treff i en annen datalenke? I min modell ble begge lagt tilbake siden jeg ikke med stor grad av sikkerhet kunne rangere den ene foran den andre. En bedre løsning hadde vært flere mulige felt å sammenlikne, problemet er at det ikke finnes flere felles felt. En mulighet for å kunne skille bedre mellom kandidatene hadde vært å dele opp datointervallet og

gitt en høyere sannsynlighet jo nærmere man kom den estimerte fødselsdatoen. Dette ville gitt forskjellige verdier som hadde gjort det mulig å rangere mellom de ulike treffene. Feltet fornavn inneholder ofte flere enn et fornavn, også her kunne man ha splittet opp informasjonen i flere felt. Dette kunne ha medført flere potensielle kandidater og man kunne ha differensiert den ikke verdi-spesifikke verdien man gav ved treff, eksempelvis verdien 2 ved alle fornavn og verdien 1 ved kun ett av dem.

Feltet kjønn er ikke benyttet som sammenlikningsfelt, selv om det er et av de to felles feltene mellom dåpsdata og begravelsesdata. Kjønnet er kjent for flertallet av de standardiserte postene, hvis dette feltet skal benyttes for å skille mellom treff med lik sannsynlighet, betinger det at kjønnet er satt feil eller at navnet er standardisert feil på bakgrunn i feil innskrivning av kjønn.

Enda en mulighet kunne ha vært å rangere feltene, det er ikke integrert en robust logikk for en rangering av felt, feltene er likestilt og likeverdige i datamodellen. Det er tre felt som inngår i prosessen frem til den mest sannsynlige positive lenken. Hvis to treff er like sannsynlige, eksempelvis har den ene lenken treff på fornavn og fødselsdato, mens den andre har treff på farsnavn og fødselsdato. Det kunne vært mulig å si at den mest sannsynlige av dem to er den med treff i fornavn siden dette feltet viser ved tester (**figur 12**) å innholde færrest feil i data.

Selve standardiseringsprosessen som genererte mine datakilder kan også være kilde til både unøyaktigheter og usikkerhetsmomenter. Standardisering har blitt gjort for navnefeltene og for fødselsdatointervallet, i tillegg har det skjedd en kalkulasjon for å estimere fødselsdato. Standardisering av navn er en tung prosess, det kunne sikkert ha vært utviklet en mer komplisert algoritme for norske navn, særlig gjelder dette for innskrivningsfeil av vokaler som gir en bokstavsforskjell i standardisert navn som min datamodell ikke klarer å

tolke, men som lett kan ses manuelt. Det estimerte intervallet for fødselsår burde hatt et større spenn, ved manuell sjekk av ikke-lenker var en stor del feiltolket. Et større spenn i kombinasjon med økende grad av sannsynlighet jo nærmere fødselsdag den estimerte datoen lå, hadde gitt flere lenker og gjort det enklere å rangere.

I utgangspunktet var det sannsynlig at antallet positive lenker skulle være høyest blant unge personer. Denne antagelsen fordi det er liten migrasjon mellom ulike prestegjeld og fordi det skulle være mer nøyaktig aldersinformasjon og fødselsinformasjon. Den høye andelen av lave treff for aldergrupper under 5 år, særlig for dødfødte og personer under 1 år viser at dette var en feil antagelse. Det viste seg ved forskjellige utsnitt av brukbare data for aldersgruppene at det var mest usammenliknbare data i disse gruppene. Dette var også tilfellet hvis jeg trakk ut dødfødte og personer under en måned.

Kriteriene i beslutningsmodellen som jeg benyttet i kapittel 6 og sjekket for 12 mulige kombinasjoner av felt. Antallet kombinasjoner er i realiteten over 20, og for hver ny kombinasjon øker antallet mulige datapar. Hvis jeg for eksempel øker søket av kombinasjoner på felt fra 12 til 15 for datasettene fra kapittel 6, vil antallet mulige datalenker stige fra 649 til 2235. Antallet positive lenker øker fra 187 til 257, noe som tilsvarer 56 % gjenfunnet maskinelt mot de opprinnelige 41 %. Antallet datapar for hver gjenfunnet post i begravelsesdatasettet har økt fra 3,5 til 8,5 i gjennomsnitt. Hoveddelen av disse nye positive lenkene vil ha en sannsynlighet under 20 % og antageligvis være falske positive, uten at jeg har gjort en manuell test.

9. Avsluttende bemerkninger

Motivasjonen med denne oppgaven var å kvantifisere og klassifisere mulige feilkilder i norsk befolkningsstatistikk gjennom en konstruert datamodell for automatisk lenking av data basert på standardiserte data fra de norske kirkebøkene. Et utalt ønske innledningsvis var å måle graden av under- og overrapportering blant barn som dør unge, særlig for gruppen av dødfødte, ved å konstruere algoritmer for lenking av personer på tvers av kirkebøkene. Konstruksjon og testing av metodene og algoritmene i datamodellen viste seg å være et svært omfattende arbeid. Det primære ble undertiden å teste om datagrunnlaget kunne benyttes for automatisk datalenking og rekonstruksjon av persondata.

Om det standardiserte datagrunnlaget kan jeg fastslå at det er store variasjoner i antallet fullstendige datafelt fra prestegjeld til prestegjeld. I enkelte prestegjeld var datagrunnlaget meget godt, med nærmere 100 % fullstendige og tolkbare data i enkelte årsintervaller. Datagrunnlaget var mer fullstendig etter år 1820 enn før og enda bedre etter år 1880, noe som har sammenheng med pålegg om bedre registrering av de kirkelige handlinger. Jeg har også observert at datagrunnlaget varierer i de forskjellige alderssegmentene, hvor datagrunnlaget er dårligst blant de yngste, da i særdeleshet blant dødfødte og barn som dør under de første leveukene. Noe jeg mener at bekrefter det at praksisen for registrering av dødfødte og svært unge barn har vært mangelfull og inkonsekvent med hensyn til om presten benyttet dåps- eller begravelseslisten.

Datamodellen, som ble utviklet i forbindelse med dette arbeidet, klarte en grovsortering og lenking av datapar. Ved en videreutvikling bør algoritmene spesifiseres mer grundig med tanke på hvilke kriterier det skal søkes etter og differensieres mer på vekting av feltene. Dette vil føre til en klarere rangering mellom dataparene som igjen ville forenkle prosessen med utvelgelse av de positive lenkene. En utvidet datamodell

bør klare å gjenfinne opptil 90 % i gjennomsnitt av de data i begravelsesdatasettet som faktisk lar seg gjenfinne. Kriteriene har vært konservative både for algoritmene under standardisering av mine datasett og for lenkingen i min datamodell. En årsak for at denne tilnærmingen ble valgt, var for å kunne klassifisere feilkildene ved automatisk lenking. Hvis kriteriene hadde vært satt mindre konservativt ville også antallet positive datalenker økt og antallet gjenværende for manuell observasjon blitt redusert. Den manuelle gjennomgangen og identifiseringen av feilkilder viste at de fleste kategorier av feilkilder kan reduseres ved en videreutvikling av algoritmene.

Datamodellen, som ble utviklet i forbindelse med denne oppgaven er generisk i den forstand at datamaterialet kan utvides til å innholde data fra flere prestegjeld og sokn. En automatisk lenking på tvers av disse vil derfor også være mulig i en konstruksjon et større befolkningsregister. Arbeidet, som ble utført i forbindelse med denne oppgaven bør derfor kunne benyttes videre både for rekonstruksjon i enkelte prestegjeld og for rekonstruksjon i større målestokk.

Litteraturliste

Backer, Julie E. (1947): "Population Statistics and Population Registration in Norway". Part 1: The vital Statistics of Norway: an historical Review, *Population Studies* 1, Nr 2, 212-226.

Blakely, Tony and Clare Salmond (2002): "Probabilistic Record Linkage and a Method to calculate the positive predictive Value", *International Journal of Epidemiology* 31, 1246-1252.

Christen, Peter (2002), "Parallel Techniques for High-Performance Record Linkage (Data Matching)",
http://datamining.anu.edu.au/talks/2002/dcs2002.pdf
(Lastet ned: 12.10.2006)

DIS-Norge (2006), "Kirkeboksregistrering som nasjonal dugnad!",
http://www.disnorge.no/kildereg/kirkeboker/dugnad.html
(Lesedato: 12.10.2006)

Dunn, H. L. (1946): "Record Linkage." *American Journal of Public Health* 36, 1412–1416.

Dyrvik, Ståle (1983): Ei innføring i metodane: historisk demografi. Universitetsforlaget, Bergen.

Ersland, Geir Atle, Edgar Hovland og Ståle Dyrvik (1997): *Festskrift til Historisk institutts 40-års jubileum*. Universitetet i Bergen, Historisk institutt, Skrifter Nr 2.

Fellegi, Ivan P. and Alan B. Sunter (1969): "A Theory for Record Linkage", *Journal of the American Statistical Association* 64, Nr. 328, 1183-1210.

Fure, Eli (2000): "Interactive Record Linkage: The cumulative Construction of Life", *Demographic Research 3,* http://www.demographic-research.org/volumes/vol3/11, (Lastet ned: 12.10.2006)

Newcombe, H. B., Kennedy, J. M., Axford, S. J. and A.P.James (1959): "Automatic Linkage of Vital Records." *Science* 130:954–959.

Newcombe, H. B. (1988): Handbook of Record Linkage Methods for Health and Statistical Studies, Administration and Business. Oxford University Press, Oxford.

Nygaard, Lars (1992): "Name Standardization in Record Linking: an improved algorithmic Strategy", History & Computing 4, Nr.2.

RHD (2006a), "Norsk standard for registrering og utveksling av nominative folketellingsdata", http://www.rhd.uit.no/histform/innledn.html, (Lesedato: 12.10.2006)

RHD (2006b), "Om RHD", http://www.rhd.uit.no/info/rhd.html, (Lesedato: 12.10.2006)

Schofield, Roger (1992): "Automatic family reconstruction", *Historical Methods* 25, Nr 2, 75-79.

Thorvaldsen, Gunnar (1996), Håndbok i registrering og bruk av historiske persondata. TANO/Aschehoug, Oslo.

Riksarkivet (2006), "Kirkebøker og kirkebokføring", http://www.riksarkivet.no/arkivverket/kilder/ofte/kirkebok.html (Lesedato: 12.10.2006)

Appendiks

[17]**Tabell 24.** Poststruktur i kirkebøker ved fødsel fra år 1812 til år 1877		
1812	**1820**	**1877**
Fødselsdato Navn Dåpsdato Foreldra: Navn, stand, yrke og bosted Faddere: Navn, stand og bosted Merknader	Fødselsdato Dåpsdato Navn Ekte eller uekte født Foreldra: Navn, borgelig stilling og oppholdssted Faddere Eventuell hjemmedåp	Fødselsdato Dåpsdato Hjemmedåp Navn Foreldra: Fødselsår Faddere: navn og borgerlig stilling Ekte eller uekte født Merknader

Tabell 25. Poststruktur i kirkebøker ved begravelser fra år 1812 til år 1877		
1812	**1820**	**1877**
Dødsdato Gravleggingsdato Navn Stand, yrke og Oppholdssted Alder Merknader	Dødsdato Gravleggingsdato Navn og stand Alder Oppholdssted Eventuelt dødsfall av smittsom sjukdom eller ved ulykke	Dødsdato Gravleggingsdato Jordfestingsdato Navn, borgerlig stilling (næringsvei) og ekteskapelig status Ektemannens navn og stilling (gifte koner), fars navn og stilling Fødselsdato Fødestad og Bosted Dødsårsak

[17] Historisk demografi, Ei innføring i metodane, Ståle Dyrvik, Universitetsforlaget, Oslo, 1983

Tabell 27.

Feltstruktur i den standardiserte dåpstabellen

FELTNAVN	BESKRIVELSE
UID	Universell ID-nummer
INTID	ID-nummer i filen
FILEID	ID-nummer for rådatafil
MUNIZIP	Prestegjeldnummer anonymisert
BRTHYEAR	Beste estimat for fødselsdato-år
BRTHMNTH	Beste estimat for fødselsdato-mnd
BRTHDAY	Beste estimat for fødselsdato-dag
BRTHYUP	Øvre grense fødselsdato-år
BRTHMUP	Øvre grense fødselsdato-mnd
BRTHDUP	Øvre grense fødselsdato-dag
BRTHYLO	Nedre grense fødselsdato-år
BRTHMLO	Nedre grense fødselsdato-mnd
BRTHDLO	Nedre grense fødselsdato-dag
FORNAVN	Nedtegnet fornavn
FORN_S	Standardisert fornavn
FARNAVN	Nedtegnet farsnavn
FARN_S	Standardisert farsnavn
GAARNAVN	Gårdsnavn
GNAVN_S GID	Gårdsid-ikke implementert
SEX	Kjønn
STILLBORN	Dødfødt ja/nei

Tabell 28.

Oversikt over prestegjeld i standardiserte datasett.

Sokn	begravelsesposter	dåpsposter	Kodet fornavn
1	299		Nei
2	9414	13445	Nei
3	5017		Nei
4	559		Nei
5	6544		Nei
6	7642	14929	Ja
7	2690	5045	Ja
8	946	1845	Nei
9	3846	11135	Nei
10	3340	6859	Ja
11	9501	12277	Ja
12	1209		Nei
13	9317		Nei
14	5211		Nei
Antall: 14	**Antall: 65535**	**Antall: 65535**	

Tabell 29.

Sannsynlighetsverdier ved datalenking av begravelsesdata.

Feltnavn	Treff	Mulig treff	Ikke-treff
Fornavn	0,99	0,70	0,10
Fornavn_s	0,99	0,70	0,10
Farnavn	0,99	0,70	0,10
Farnavn_s	0,99	0,70	0,10
Fdato	0,99	0,70	0,10

Tabell 30.

Klassifisering av fødselsdato, kombinasjoner dag, måned og år.

	BG data			DP data			
	Dag	**Mnd**	**År**	**Dag**	**Mnd**	**År**	**Verdi**
	XX	XX	XXXX	XX	XX	XXXX	0
	99	XX	XXXX	XX	XX	XXXX	1
	XX	XX	XXXX	99	XX	XXXX	1
	99	XX	XXXX	99	XX	XXXX	2
	99	99	XXXX	XX	XX	XXXX	2
	XX	XX	XXXX	99	99	XXXX	2
	XX	XX	XXXX	XX	99	XXXX	?
	99	99	XXXX	XX	99	XXXX	?
	XX	99	XXXX	XX	99	XXXX	?
	99	99	XXXX	99	99	XXXX	?

Tabell 31.

Klassifisering av attributter/felt for datalenking[18]

Nr	Fornavn	Forn_s	Farnavn	Farn_s	Dato	Dato2	TREFF	Max %
0	X	X	X	X	X		2,2,2,2,0,2	0,95
1	X	X	X	X		X	2,2,2,2,0,1	0,67
2	X	X			X		2,2,1,1,0,2	0,47
3	X	X				X	2,2,1,1,0,1	0,34
4			X	X	X		1,1,2,2,0,2	0,47
5			X	X		X	1,1,2,2,0,1	0,34
6		X		X	X		1,2,1,2,0,2	0,47
7		X		X		X	1,2,1,2,0,1	0,34
8			X	X	X		0,0,2,2,0,2	0,01
9			X	X		X	0,0,2,2,0,1	0,21
10	X	X			X		2,2,0,0,0,2	0,01
11	X	X				X	2,2,0,0,0,1	0,005
12	X	X	X	X			2,2,2,2,0,0	0,10
13	X	X					2,2,1,1,0,0	0,05
14			X	X			1,1,2,2,0,0	0,05
15		X		X			1,2,1,2,0,0	0,05

[18] Antallet komb. er over 20, algoritmen i modellen benytter 15 av disse. Ved ukjent farnavn gis verdien 1.

Figur 12.

Filtrerte poster fra begravelsesdatasettet for prestegjeld 1

År	Fornavn	Forn_s	Farnavn	Farn_s	Sex	Fdato	%Snitt/Sum
1800	98,4% - (60)	98,4% - (60)	24,6% - (15)	24,6% - (15)	100,0% - (61)	91,8% - (56)	73,0% - (61)
1801	100,0% - (55)	100,0% - (55)	34,5% - (19)	34,5% - (19)	100,0% - (55)	90,9% - (50)	76,7% - (55)
1802	100,0% - (58)	100,0% - (58)	43,1% - (25)	43,1% - (25)	100,0% - (58)	87,9% - (51)	79,0% - (58)
1803	100,0% - (56)	100,0% - (56)	37,5% - (21)	37,5% - (21)	100,0% - (56)	89,3% - (50)	77,4% - (56)
1804	98,5% - (65)	98,5% - (65)	37,9% - (25)	37,9% - (25)	100,0% - (66)	97,0% - (64)	78,3% - (66)
1805	100,0% - (38)	100,0% - (38)	34,2% - (13)	34,2% - (13)	100,0% - (38)	92,1% - (35)	76,8% - (38)
1806	100,0% - (50)	100,0% - (50)	46,0% - (23)	46,0% - (23)	100,0% - (50)	94,0% - (47)	81,0% - (50)
1807	100,0% - (52)	100,0% - (52)	57,7% - (30)	57,7% - (30)	100,0% - (52)	96,2% - (50)	85,3% - (52)
1808	100,0% - (41)	100,0% - (41)	46,3% - (19)	46,3% - (19)	100,0% - (41)	92,7% - (38)	80,9% - (41)
1809	100,0% - (34)	100,0% - (34)	50,0% - (17)	50,0% - (17)	100,0% - (34)	79,4% - (27)	79,9% - (34)
1810	100,0% - (47)	100,0% - (47)	36,2% - (17)	36,2% - (17)	100,0% - (47)	95,7% - (45)	78,0% - (47)
1811	98,2% - (55)	98,2% - (55)	50,0% - (28)	50,0% - (28)	100,0% - (56)	92,9% - (52)	81,5% - (56)
1812	98,2% - (56)	98,2% - (56)	42,1% - (24)	42,1% - (24)	100,0% - (57)	86,0% - (49)	77,8% - (57)
1813	98,2% - (56)	98,2% - (56)	45,6% - (26)	45,6% - (26)	100,0% - (57)	93,0% - (53)	80,1% - (57)
1814	100,0% - (67)	100,0% - (67)	49,3% - (33)	49,3% - (33)	100,0% - (67)	88,1% - (59)	81,1% - (67)
1815	100,0% - (53)	100,0% - (53)	60,4% - (32)	60,4% - (32)	100,0% - (53)	81,1% - (43)	83,6% - (53)
1816	100,0% - (78)	100,0% - (78)	55,1% - (43)	55,1% - (43)	100,0% - (78)	87,2% - (68)	82,9% - (78)
1817	98,7% - (78)	98,7% - (78)	40,5% - (32)	40,5% - (32)	100,0% - (79)	87,3% - (69)	77,6% - (79)
1818	98,3% - (57)	98,3% - (57)	56,9% - (33)	56,9% - (33)	100,0% - (58)	89,7% - (52)	83,3% - (58)
1819	100,0% - (57)	100,0% - (57)	43,9% - (25)	43,9% - (25)	100,0% - (57)	93,0% - (53)	80,1% - (57)
1820	100,0% - (66)	100,0% - (66)	43,9% - (29)	43,9% - (29)	100,0% - (66)	92,4% - (61)	80,1% - (66)
1821	100,0% - (78)	100,0% - (78)	43,6% - (34)	43,6% - (34)	100,0% - (78)	89,7% - (70)	79,5% - (78)
1822	100,0% - (54)	100,0% - (54)	46,3% - (25)	46,3% - (25)	100,0% - (54)	92,6% - (50)	80,9% - (54)
1823	100,0% - (71)	100,0% - (71)	49,3% - (35)	49,3% - (35)	100,0% - (71)	81,7% - (58)	80,0% - (71)
1824	100,0% - (68)	100,0% - (68)	52,9% - (36)	52,9% - (36)	100,0% - (68)	95,6% - (65)	83,6% - (68)
1825	100,0% - (82)	100,0% - (82)	57,3% - (47)	57,3% - (47)	100,0% - (82)	80,5% - (66)	82,5% - (82)
1826	100,0% - (71)	100,0% - (71)	59,2% - (42)	59,2% - (42)	100,0% - (71)	88,7% - (63)	84,5% - (71)
1827	100,0% - (57)	100,0% - (57)	54,4% - (31)	54,4% - (31)	100,0% - (57)	89,5% - (51)	83,0% - (57)
1828	100,0% - (68)	100,0% - (68)	67,6% - (46)	67,6% - (46)	100,0% - (68)	83,8% - (57)	86,5% - (68)