Tidsskrift for samfunnsforskning

UNIVERSITETSFORLAGET

FORSKNINGSKOMMENTAR
VITENSKAPELIG PUBLIKASJON

Årgang 60, nr. 4-2019, s. 398–408
ISSN online: 1504-291X
DOI: https://doi.org/1018261/issn.1504-291X-2019-04-04

Microdata.no: Ny teknologi gir forskere umiddelbar tilgang til norske registerdata

Microdata.no: New technology allows instant access to Norwegian register data

Jannike Gottschalk Ballo
Pilotbruker av microdata.no, stipendiat, Arbeidsforskningsinstituttet ved OsloMet
jannba@oslomet.no

Innledning

Bruken av registerdata i forskning har de siste tiårene vært stigende, særlig innen velferds- og arbeidsforskning. Samtidig har det blitt vanskeligere og dyrere å få utlevert registerdata til forskning. Med ambisjon om å forenkle forskernes tilgang på registerdata har Norsk senter for forskningsdata (NSD) og Statistisk sentralbyrå (SSB) utviklet og lansert den nettbaserte dataportalen og analyseverktøyet microdata.no. Hensikten med denne kommentaren er å belyse hvordan microdata.no som teknologisk nyvinning evner å fylle et skrikende behov for raskere og billigere tilgang til registerdata. Tjenesten har både fordeler og ulemper, men utgjør i sum et stort potensial for registerdataforskningen i Norge. For at løsningen skal kunne slå ut i full blomst, kreves det imidlertid utstrakt bruk i forskning og økte bevilgninger til videre utvikling av tjenesten. Denne kommentaren baserer seg på mine egne erfaringer fra bruk av microdata.no som pilotbruker og forsker.

Norge har sammen med de andre skandinaviske landene svært gode registerdata. Staten registrerer opplysninger om personer, virksomheter og foretak i stor skala, noe som danner grunnlag for akkumulert kunnskap om velferd, arbeid, næringsliv og befolkningens atferd og sammensetning over tid. I registerdata ligger derfor også svar på mange av samfunnets fremtidige utfordringer og løsninger så fremt de gjøres tilgjengelige for forskning på effekter og årsakssammenhenger. Forskere har tatt til orde for at de gode registrene i Norge utgjør et komparativt fortrinn i internasjonal forskningssammenheng. Det er derfor viktig at kvaliteten på dataene vedlikeholdes, og at tilgjengeliggjøring prioriteres (Bratsberg, Røed & Raaum, 2012).

«La oss få på plass verdens beste system for registerdata», sa professor og sosiolog Knud Knudsen for snart ti år siden og siktet blant annet til tilgjengeligheten for forskning (Jakobsen, 2010). Likevel har tilgangen til utlån av registerdata snarere blitt dyrere, mer tidkrevende og mindre transparent (Fløtten, Helgøy, Skule & Storsul, 2017; Jakobsen, 2014). I 2017 var gjennomsnittlig leveringstid for såkalte mikrodata fra SSB 7–14 måneder, avhen-

gig av fakturabeløp. SSB tar ikke betalt for data, men de tar betalt for tiden de bruker på å tilrettelegge datasettene (SSB, 2018). Konsekvensene av lang ventetid og høy kostnad er at studenter og stipendiater sjelden har mulighet til å bruke registerdata i sin forskning. Ansatte forskere rapporterer på sin side om problemer med å komme i mål med prosjekter innen fristene (Fløtten et al., 2017). En gruppe sykefraværsforskere måtte ta til takke med data fra 2007 til en artikkel publisert i 2014 fordi prisen på å få oppdatert dataene var for høy (Jakobsen, 2014). Praksisen med å låne ut kopier av registerdata er i det hele tatt lite samfunnsøkonomisk. Slik reglene er i dag, er det ikke mulig å søke om utlån av registerdata før man har fått finansiering til prosjekter, og har man først fått finansiering til et prosjekt, begynner pengene å løpe uavhengig av om dataene er på plass eller ei. Tradisjonelle utlån av registerdata er altså forbundet med en rekke svakheter, som kan oppsummeres i tre hovedpunkter:

For det første tar SSB seg betalt mange ganger for den samme jobben – nemlig tilrettelegging og avidentifisering av data. Dette er administrativ flytting av informasjon som må repeteres for hver enkelt datasøknad. Å flytte data er en prosess andre bransjer har automatisert for lenge siden fordi det er lite lønnsomt å sette manuelle ressurser til å gjøre noe en maskin kan gjøre både raskere og bedre.

For det andre innebærer *kopier* av data en betydelig kvalitetssvekking. Med en gang registerdata blir til kopier, fjernes koblingen til kilden og samtidig muligheten for kontinuerlig oppdatering. Samtidig innebærer manuell tilrettelegging en risiko for feil som kan være vanskelig å oppdage for forskeren. Utvalg kan trekkes feil, variabler kan kodes feil og matriser forskyves. I tillegg fjernes muligheten for å etterprøve forskerens eget arbeid med dataene, noe som er en alvorlig konsekvens for forskningen (Camerer et al., 2018; Kvittingen, 2018). Kopierte datafiler overlevert til forskere kan bearbeides uten at det føres logg over hva som skjer med dataene. Til tross for databehandlingsplaner og de beste intensjoner om å holde orden kan det være en utfordring å opprettholde tilstrekkelig dokumentasjon. Dette gjelder særlig når mange forskere jobber sammen i felles prosjekter.

For det tredje er utlånte registerdata avidentifiserte, men personer vil likevel være *indirekte identifiserbare via personopplysninger* (eksempelvis bosted, yrke, fødselsår, kjønn, inntekt, sivilstatus). Det eneste som beskytter personvernet til de som forskes på, er forskerens etiske forpliktelser og løfte om taushet.

Det nettbaserte analysemiljøet microdata.no er et forsøk på å løse disse utfordringene – nemlig å redusere behovet for gjentakende bearbeiding, sikre høy datakvalitet og sporing av forskeraktivitet og samtidig ivareta personvernet. Ambisjonen er å erstatte en betydelig andel av alle utlån av registerdata og i tillegg stimulere til økt bruk av registerdata i samfunnsforskningen.

Hva er microdata.no?

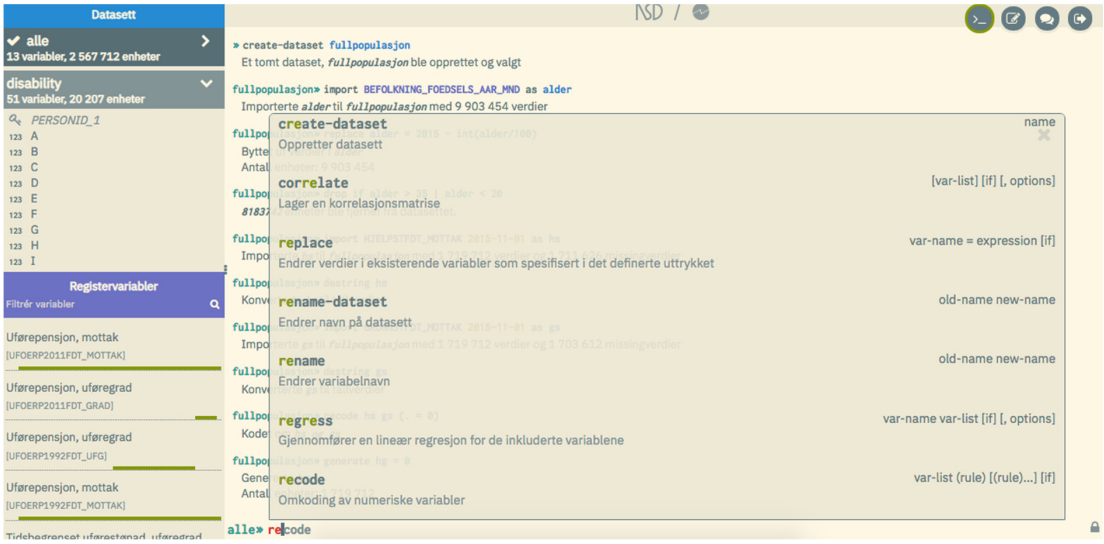
Microdata.no er en plattform for analyse av registerdata med innebygd personvern. Den innebygde anonymiteten øker sikkerheten for de som blir forsket på, mens integreringen av variabler, metadata og analyseverktøy i en og samme applikasjon gir forskerne umiddelbar tilgang til å analysere dataene. Løsningen, som fikk Datatilsynets pris for «Innebygd personvern i praksis 2018» (Datatilsynet, 2019), er verdens første til å fullt ut beskytte personvernet og samtidig kunne gi direkte tilgang til analyse av rådata. Det er med andre ord store forskningsmessige og samfunnsøkonomiske potensial som ligger i denne løsningen.

Microdata.no er utviklet av Norsk senter for forskningsdata (NSD) og SSB med finansiering fra Norges forskningsråd. Tjenesten er tilgjengelig for forskere og studenter ved god-

kjente forskningsinstitusjoner med innlogging via ID-porten. Det er kreves ingen form for søknad eller innmelding om behandling av personopplysninger. Ideen er å gi analysetilgang til SSBs registerdata uten at dataene forlater SSB. På denne måten vil dataene som brukes i forskning, alltid være oppdaterte, og ressurser som tidligere gikk til tilrettelegging og utlevering, kan reduseres betydelig. Microdata.no tilbyr ferdigformaterte registerdata tilrettelagt for statistisk analyse. Man slipper med andre ord unna den tidkrevende bearbeidingsprosessen som ellers må til når man får utlevert registerdatafiler, noe som både sparer ressurser og senker terskelen for å komme i gang med registerdata for første gang.

Microdata.no skiller seg fra analyse av registerdata i konvensjonelle analyseprogrammer blant annet ved at alle tilgjengelige variabler – la oss kalle dem *databasen* – er en integrert del av analyseverktøyet. Databasen inneholder samtlige registervariabler som SSB har tilrettelagt for bruk i microdata.no og omfatter både forløpsdata og statusdata om befolkning, utdanning, inntekt, arbeidsmarked og trygd. Etter hvert som flere variabler blir klargjort, legges disse inn i databasen fortløpende. For å kunne gjøre analyser på variablene i databasen må forskere bygge *egne datasett*. Dette gjøres ved å opprette ett eller flere tomme datasett for så å importere variabel for variabel for et gitt tidspunkt eller tidsrom fra databasen. All videre omkoding av variabler gjøres deretter på egne datasett. Det er altså ikke mulig å endre på dataene i databasen. Samtidig er ikke egne datasett å anse som kopier løsrevet fra databasen, ettersom enhver endring eller retting SSB gjør av originaldata i databasen, vil gjenspeiles umiddelbart i egne datasett.

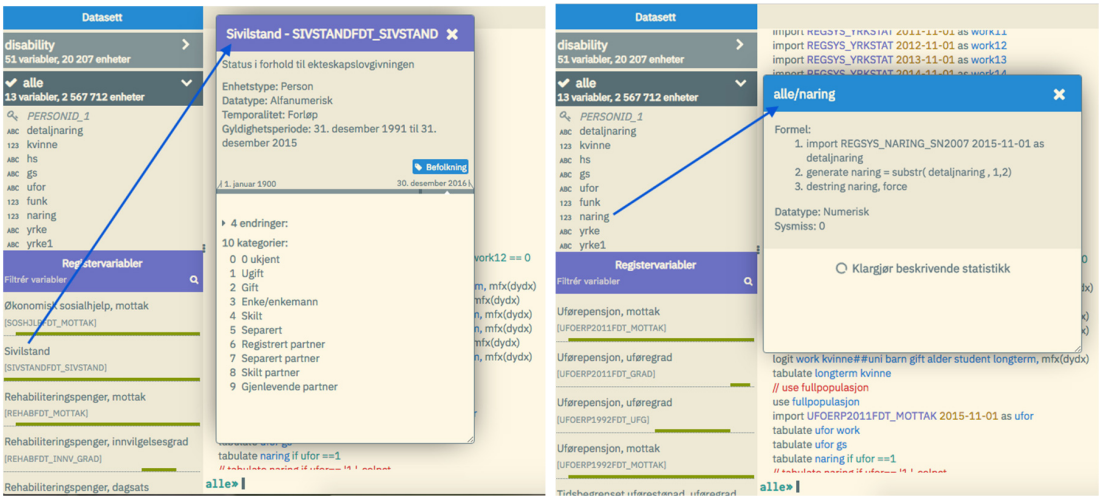
Microdata.no har et grensesnitt og kommandospråk som minner om analyseprogrammet Stata. Hjelpetekster i form av pop-up gjør kommandospråket brukervennlig, og terskelen for nybegynnere relativt lav. Figur 1 viser en skjermdump fra microdata.no. I kolonnen til venstre ser man listen over egne datasett øverst og den søkbare databasen med registervariabler nederst. Kommandolinjen samt vinduet med hjelpetekster som dukker opp når man begynner å skrive kommandoer, vises nederst i bildet.



Figur 1. Microdata.no analysemiljø

All informasjon og datadokumentasjon knyttet til variablene, såkalte metadata, er fullt integrert i grensesnittet ved hjelp av små pop-up-vinduer (se figur 2 til venstre), noe som gjør det enkelt å holde oversikt over hvilke data man jobber med, og reduserer risikoen for feilkoding av variabler. Alle trinn i omkodingen av en variabel lagres i tillegg som endringshistorikk i variabelens metadatavindu (se figur 2 til høyre). Analyseverktøyene omfatter ulike

kommandoer for deskriptiv statistikk, lineær regresjon, paneldataregresjon med faste og tilfeldige effekter, logistisk og multinomisk logistisk regresjon og tilhørende marginaleffekter samt ulike grafiske fremstillinger.



Figur 2. Skjermdump metadata og endringshistorikk

Datasett kan organiseres enten som tverrsnitt eller med panelstruktur, og det er mulig å gjøre koblinger til foreldre med tilhørende personopplysninger. All kommandohistorikk lagres fortløpende og kan hentes frem igjen når som helst. Den integrerte skripteditoren åpner for at forskere kan bygge egne skript med kommentarer, slik brukere av statistikkprogrammer som Stata, SAS og R er kjent med. Gjennom utveksling av kommandohistorikken kan analyser enkelt deles mellom kolleger. Nettopp deling og gjenbruk av skript er langt enklere i microdata.no enn ved bruk av datasett utlevert på tradisjonelt vis. Sistnevnte data har ofte noe ulik innretning fra prosjekt til prosjekt, som gjør at man som regel må bygge skript fra bunnen av hver gang. I microdata.no er rådata og oppsett likt for alle. Microdata.no er åpen for alle forskere ved godkjente institusjoner. Det er derfor heller ikke nødvendig å søke om ekstra tillatelser eller hente inn taushetserklæringer når man vil samarbeide med andre forskere.

Personvernhensynet og winsorisering

Microdata.no kan gi brukere tilgang til en hel registerdatabase via ID-porten fordi teknologien ivaretar personvernet. Hovedideen i sikringen av personvernet handler om å forhindre at brukere får mulighet til å visuelt inspisere rådataene. Dette betyr at i motsetning til ved behandling av registerdata i konvensjonelle statistikkprogrammer får man i microdata.no ikke tilgang til rader og kolonner av individdata. Rådatamatrisene holdes skjult for forskerne. Videre sørger microdata.no for at alle aggregerte data er avidentifisert, og at all bruk av systemet lagres i fem år. Det vil si at forskernes aktivitet i analysemiljøet *ikke* er anonym, og at mistenkelig aktivitet, som forsøk på identifisering av enkeltpersoner, kan oppdages og spores.

De som det forskes på, skal imidlertid være anonyme, og for å sikre at de forblir anonyme, er det implementert en rekke tekniske tiltak. Disse er det nødvendig å vite om når man bruker microdata.no fordi det påvirker alle statistikker og analyser. Tiltakene omfatter foreløpig 1) begrensning på minste populasjonsstørrelse og tilfeldig utvalgstrekking; 2) støylegging av deskriptiv statistikk; 3) anonymisering av spredningsplott; og 4) winsorisering. De tre første er relativt enkle å forstå og skaper håndterbare utfordringer for analyse og

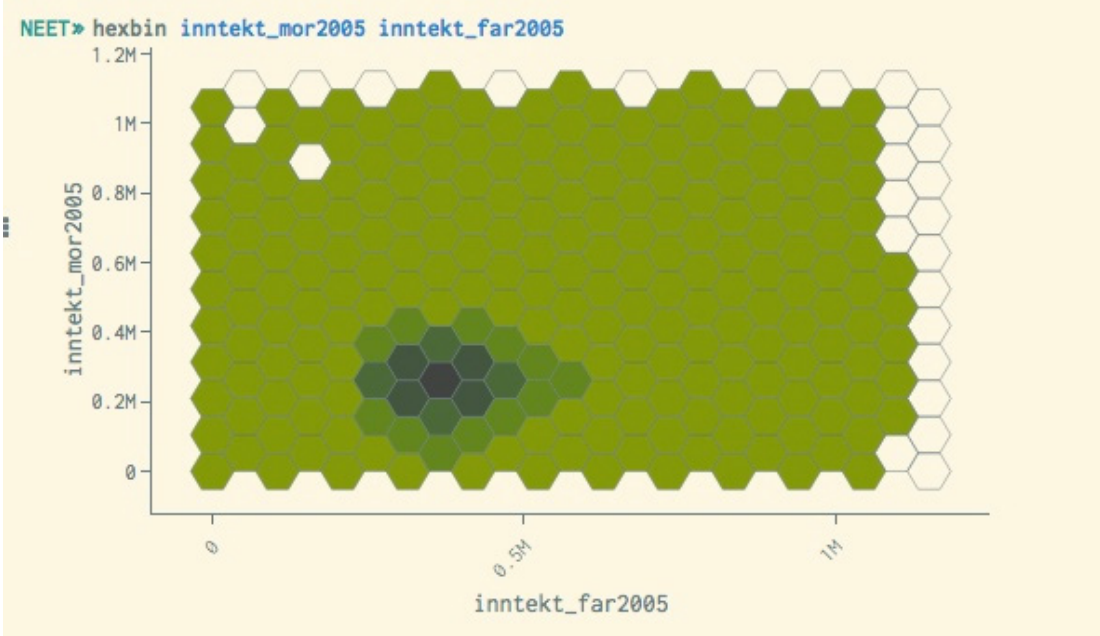
fortolkning av resultater. Det siste tiltaket kan være problematisk og krever en mer detaljert drøfting.

Det første tiltaket handler om *minste tillatte populasjonsstørrelse* i egne datasett, som er 1000 individer. Forsøk på å opprette datasett med færre enheter stoppes av den såkalte avsløringskontrollen. Videre er det ikke mulig å trekke et tilfeldig utvalg fra en populasjon, selv om utvalget består av 1000 enheter eller mer. Datasett kan kun begrenses ved å selektere på variabler. Argumentet er at det ikke skal være nødvendig å trekke tilfeldige utvalg når fullpopulasjonsdata er tilgjengelig, og at serverne skal være sterke nok til å raskt estimere modeller, selv om man har datasett med flere millioner observasjoner. Foreløpig erfaring fra microdata.no tilsier imidlertid at arbeidskapasiteten *har* begrensninger, og at muligheten for å jobbe med mindre utvalg i oppbyggingen av modeller ville vært svært tidsbesparende. Det er heller ingen åpenbare grunner til at tilfeldig utvalgstrekking er i strid med verken personvernhensyn eller etterprøvbarhet, som fremheves som to grunnleggende prinsipper i microdata.no.

Det andre tiltaket for å beskytte anonymiteten til enkeltindivider er *støylegging av deskriptive tabeller*. Samtlige opptellinger i frekvenstabeller støylegges med inntil +/- 5. Frekvenser under fem observasjoner vises aldri, men kamufleres med enten 0 eller 5. Støyen er stokastisk, men konstant. Det vil si at dersom man ber om gjentatte frekvenstabeller fra samme populasjon, vil resultatet av støyleggingen alltid være det samme. Støyen er imidlertid synlig ved at tabellene ikke blir additive. Antall observasjoner i frekvenstabeller adderer seg sjelden eksakt opp til antall observasjoner i marginalcellene. Se NSD og SSB (2018, s. 95–99) for utledning av formlene som støyleggingen hviler på. Det er verdt å merke seg at støyen legges til tabeller, men endrer ikke rådata. Derfor vil ikke statistiske analyser (utover deskriptive tabeller) påvirkes. Effekten av støyleggingen vil naturlig nok være stor i deskriptive statistikker med få enheter og nærme seg ubetydelig etter hvert som antallet enheter øker.

Det tredje tiltaket for å sikre personvernet er å *erstatte spredningsplott med hexbinplott*. Spredningsplott er en graf som viser sammenhengen mellom to variabler (langs x- og y-aksen), hvor verdier er representert med punkter. I et hexbinplott (se Figur 3) deles i stedet området mellom aksene inn i sekskanter med fargevariasjon etter antall enheter per sekskant. Selve hensikten med hexbinplottet er å redusere detaljnivået ved å kamuflere enkeltobservasjoner, men samtidig tilby en visualisering av eventuelle mønster i dataene. Her er det imidlertid viktig å være oppmerksom på at winsorisering, som er det siste tiltaket, vil påvirke plottet (NSD & SSB, 2018).

Winsorisering er altså det fjerde tiltaket som har til hensikt å forhindre identifisering av ekstreme observasjoner. Microdata.no benytter 2 prosent-winsorisering, som betyr at de én prosent høyeste verdiene og de én prosent laveste verdiene på alle numeriske variabler omkodes til henholdsvis 99-prosentilen og 1-prosentilen. Winsorisering virker anonymiserende på ekstreme observasjoner som tilhører halene til fordelingen til en variabel. Typiske personer som anonymiseres med winsorisering, er kjente personer med svært høy inntekt og formue i små kommuner. Til tross for anonymisering via støylegging (tiltak 1 ovenfor) vil forskere kunne produsere en tabell som i tillegg til en rekke andre variabler angir gjennomsnittsformue (eller sum formue) for en relativt liten populasjon. Uten winsorisering vil det av tabellen være lett å identifisere i hvilken celle kommunens rikeste befinner seg, og derfor også enkelt å lese av personens verdier på de andre variablene i tabellen. Med winsorisering er ikke lenger dette mulig. Siden winsoriseringen faktisk endrer rådataene ved import av variabler og bruk av kommandoene *keep if* og *drop if*, vil dette påvirke alle statistikker, analyser og grafer.



Figur 3. Eksempel på hexbinplott. Fars inntekt mot mors inntekt i 2005

Winsorisering ble først introdusert av Dixon (1960) som en teknikk for å gjøre statistiske modeller mer robuste mot utliggere (Dixon & Yuen, 1974; Kokic & Bell, 1994; Pan, Sarkar & Mudholkar, 2009; Rivest, 1994). Det er imidlertid sjeldnere at teknikken brukes som en «one size fits all»-løsning av anonymiseringshensyn med robusthet som en mulig positiv bivirkning. Ifølge SSB er det i mange land, inkludert Eurostat-data, vanlig å anonymisere mikrodata fra utvalgsundersøkelser ved topp- og bunnkoding av numeriske variabler slik at ekstreme verdier kamufleres. Det gjelder imidlertid på statiske datasett i motsetning til microdata.no, som er en dynamisk database hvor winsoriseringen justeres etter hvilken populasjon forskeren har valgt å analysere. Det finnes lite litteratur som tar for seg hvordan winsorisering virker inn på regresjoner, men det er mulig å komme med noen generelle bemerkninger.

Ved normalfordelte variabler vil endringen være relativt uproblematisk for analysen: Gjennomsnitt, standardavvik og estimater blir mer robuste for utliggere. Medianer og kvartiler påvirkes ikke. Når numeriske variabler er skjevfordelte, er det mindre åpenbart hvilken effekt winsoriseringen har for robusthet fordi forskere ikke selv kan inspisere hvor lange halene er, og hvor mye informasjon som går tapt. Histogrammer gir en indikasjon på variablenes normalfordeling og kan være et nyttig redskap når man skal vurdere winsoriseringens betydning for analysen. Dette kan fungere bra i bivariate analyser, men i multivariate analyser er det nesten umulig å forutsi konsekvensene av winsoriseringen ettersom flere mekanismer virker sammen.

For eksempel vil winsoriseringen av variabler som er svært skjevfordelte, kunne introdusere skjevhet i analysen. Det vil si en type skjevhet som innebærer at resultatene med winsorisering avviker betydelig fra resultatene man ville fått uten denne anonymiseringsmetoden. Typiske tilfeller hvor denne formen for skjevhet kan oppstå, er studier av små grupper som avviker betydelig fra referansegruppen på relevante variabler. Winsoriseringen vil da potensielt fjerne informasjon fra enten venstre eller høyre hale, som er viktig for analysen fordi observasjoner tilhørende gruppen man er interessert i, kan befinne seg her. I ekstreme tilfeller vil hele gruppen bli spist opp av winsoriseringen. John Tukey, som

var en av de første til å skrive om robust estimering, advarte mot overbehandling av utliggere og minnet om at «the distribution relevant to statistical practice is that of the values actually provided and not of the values which ought to have been provided» (Tukey, 1960, s. 457). Med andre ord er det viktig å huske på at også ekstreme observasjoner faktisk tilhører utvalget eller populasjonen man er interessert i, og at det kan skape problemer om man fjerner eller omkoder disse etter en fast regel. Kvaliteten på variablene[1] vil imidlertid spille en viktig rolle; utliggere som skyldes registreringsfeil, vil være viktigere å få winsorisert bort enn ekstreme observasjoner som skyldes faktiske avvik.

I det hele tatt er det vanskelig å vurdere en eventuell skjevhet i analysen fordi forskere ikke selv har kontroll på prosessen. Det er ikke mulig å gjøre analyser med og uten winsorisering for sammenligning, og det er umulig å inspisere ekstreme observasjoner manuelt for å avdekke årsaken til avvikene. For å gi brukere av microdata.no en bedre forståelse av hvordan winsorisering påvirker regresjoner, vil det være nyttig å simulere effekten, for eksempel ved reproduksjon av tidligere registerdatastudier med og uten 2 prosent-winsorisering. Dette vil gi innblikk i typiske tilfeller hvor winsoriseringen skaper skjevhet, og tilfeller hvor metoden skaper mer robusthet. Likevel er det uklart i hvilken grad winsorisering vil påvirke muligheter for publisering av forskningsresultater. Man kan spørre seg om topptidsskriftene vil akseptere winsoriserte data med potensielle skjevheter som forskere ikke kan gjøre fullstendig rede for. Simulering og drøfting av winsoriseringens effekt på registerdataanalyser vil kunne skape større innsikt og bedre vurderingsgrunnlag for forskere og tidsskriftredaksjoner.

Dagens winsoriseringsmodell er valgt fordi den effektivt anonymiserer ekstreme observasjoner og er enkel å implementere uten å legge til grunn teoretiske forventninger om variablenes fordeling. I de fleste tilfeller vil winsoriseringen gjøre analysene mer robuste, men forskere skal være spesielt på vakt når hensikten er å studere marginaliserte grupper som forventes å avvike fra referansegruppen.

Begrenset fleksibilitet og residualdiagnostikk

Siden microdata.no er en analyseløsning som er laget for å gi rask tilgang til individdata uten å svekke personvernet, har fleksibiliteten i verktøyet måttet tåle visse begrensninger. Den kanskje største ulempen er at forskere er prisgitt det integrerte analyseverktøyet, uten mulighet for å benytte konvensjonelle statistikkprogrammer og analysepakker. Denne ulempen skyldes i første rekke ikke mangel på kompatibilitet, men at anonymiseringsmodellen til microdata.no skal sikre konfidensialitet gjennom hele analyseprosessen. Å slippe til eksternt programmerte verktøy utgjør for stor risiko for personvernet. Alle analyseteknikker må derfor programmeres og kontrolleres av utviklerne bak microdata.no. I teorien er det ingen begrensninger for hva som kan implementeres, men dette krever naturlig nok kunnskap og ressurser for metodeutvikling. I tillegg utgjør den skjulte programmeringen en betydelig risiko for feil. I motsetning til programmer som Stata og R, hvor koden er tilgjengelig og kan inspiseres av brukerne, utelukker microdata.no ekstern kontroll av kildekoden. Hvilke rutiner microdata.no har for kontroll med eget analyseverktøy, er derfor et relevant spørsmål.

En annen ulempe i microdata.no er brannmuren mellom forskerne og datamatrisen. Alle som har jobbet med kvantitativ analyse, vet at å inspisere datamatrisen er en nyttig

1. Det kan være store kvalitetsforskjeller i de ulike registervariablene. For eksempel er det kjent at skattemeldingsdata inneholder svært lite feil, mens NAV-data inneholder ganske mye feil.

metode for å gjøre seg kjent med datasettet og for å sjekke at bearbeiding og omkoding har hatt ønsket resultat. Denne muligheten har man ikke i microdata.no og kan i starten gi en opplevelse av å arbeide i blinde. De integrerte metadataene veier imidlertid opp, og ved hyppig bruk av deskriptiv statistikk blir man raskt trygg på datasettet og variablene.

Grafisk visualisering av resultater er en videre begrensning med microdata.no. Grafene som tilbys er strengt kontrollerte og kan ikke redigeres. Dette skyldes hensynet til anonymiseringen fordi grafer som kan skaleres fritt, vil stå i fare for å identifisere enkeltobservasjoner. Grafisk fremstilling av resultater er imidlertid et svært viktig formidlingsverktøy som i de fleste tilfeller krever individuell tilpasning fremfor standardformatering. Her ligger det et betydelig forbedringspotensial hva gjelder muligheter for redigering av de grafiske fremstillingene som tilbys i dag, og utvalget av grafer bør utvides.

Til slutt er det verdt å påpeke en ganske alvorlig mangel som børe være enkel å utbedre uten at det går på bekostning av personvernet. Foreløpig er det ikke mulig å produsere variabler for predikerte verdier og residualer i etterkant av modeller. Det er klart at man ikke kan tillate inspeksjon av residualer enkeltvis, men det gjør man sjelden med registerdata uansett på grunn av mengden observasjoner. Snarere er det interessant å sjekke struktur og mønstre i residualene for å utelukke homoskedastisitet og skjevhet på grunn av utelatte variabler (Baum 2006, s. 124). Residualdiagnostikk kan også bidra til å avdekke utliggere som winsoriseringen ikke har fanget opp. En teknisk løsning for analyse av residualer bør kunne implementeres uten at personvernet kompromitteres.

Til tross for begrensninger knyttet til analysemetoder og grafer, vil tids- og kostnadshensyn føre til at microdata.no i svært mange tilfeller likevel vil være det foretrukne valget når registerdata skal analyseres. Allerede nå er det fullt mulig å gjennomføre longitudinelle analyser som, variabelutvalget tatt i betraktning, åpner for rike og unike muligheter for norsk samfunnsforskning. I situasjoner hvor det likevel er ønskelig og nødvendig å søke om utlån av mikrodata fra SSB – det kan hende at analyseteknikken man ønsker å bruke ikke støttes av microdata.no, eller at man ønsker ytterligere variabler som foreløpig ikke er tilgjengelige – vil man ved hjelp av microdata.no kunne skrive en mer spisset og bedre begrunnet søknad om datautlån, som både vil være rimeligere og raskere å behandle.

Potensialet for samfunnsforskningen

Microdata.no ble lansert i mai 2018 og har ennå ikke fått ordentlig fotfeste blant registerdataforskere. Tjenesten er avhengig av at forskere tar den i bruk og melder tilbake ønsker og behov for at den skal videreutvikles på en fruktbar måte. Dersom målet om å erstatte en betydelig del av tradisjonelle utlån av forskningsdata skal oppnås, er det viktig at verktøy for analysemetoder utvikles fortløpende og i takt med relevant forskning. Etter hvert som flere forskere og studenter tar tjenesten i bruk, vil det i tillegg være behov for forbedring av ytelsen: Det er helt avgjørende at man ikke støter på kapasitetsproblemer ved kjøring av modeller på store datasett. Utviklingsarbeid, investeringer og vedlikehold krever ressurser og bevilgninger, men vil over tid være langt rimeligere enn å fortsette praksisen med utlån av data enkeltvis.

Utover økt datatilfanget, flere analysemetoder og bedre ytelse innebærer microdata.no et stort forskningspotensial i muligheten for å koble på forskernes egne data. Dette er en idé som er lansert, men ikke ennå realisert. Å slå sammen spørreundersøkelsesdata med registerdata er en metode som har blitt stadig vanligere de seneste årene (se for eksempel Fevang, Røed, Raaum & Zhang, 2004). Fordelen er at man kan motvirke undersøkelsestretthet blant informanter ved å droppe alle spørsmål om opplysninger registerdata kan gi svar på, og

konsentrere undersøkelsen om spørsmål bare respondentene selv kan gi svar på. I 2012 ble denne sammenkoblingen av opplysninger betegnet som «et uutnyttet forskningspotensial» (Bratsberg et al., 2012). I dag kan den neppe betegnes som uutnyttet, men det er fortsatt vanskelig og tidkrevende å søke om de tillatelser og tilganger som en slik sammenkobling av opplysninger krever. Derfor vil det i mange tilfeller være praktisk umulig å gjennomføre en kobling til registerdata. Microdata.no kan imidlertid løse de praktiske og personvernmessige utfordringene og vil kunne revolusjonere måten man driver kvantitativ samfunnsforskning på i Norge. Forutsetningen er at det kommer på plass en teknologisk løsning som muliggjør at forskere kan laste opp egne data og koble på registerdata. I tillegg bør det implementeres en strømlinjeformet prosess for søknad om koblingstillatelse ettersom SSB ikke har behandlingshjemmel for data som brukere eier selv.

Formålet med spørreundersøkelser er som regel å hente inn informasjon om vurderinger, holdninger, subjektive oppfatninger og handlingsbegrunnelser. Dette er opplysninger man ikke finner i administrative registre. Derimot inneholder registre svært presis informasjon om personers faktiske historie knyttet til bosted, familieforhold, utdanning, arbeid og inntekt. Sammen utgjør register- og utvalgsdata interessante virkelighetsbeskrivelser som danner grunnlag for analyser av årsakssammenhenger man ikke kunne gjennomført på dataene hver for seg (Bratsberg et al., 2012). I tillegg til muligheten for å laste opp forskernes egne data bør det vurderes om microdata.no også kan inneholde de utvalgsdata som NSD allerede har i sin databeholdning, med tilsvarende kobling til registerdata. Dette gjelder blant annet undersøkelser som arbeidskraftundersøkelsen (AKU), integreringsbarometeret, levekårsundersøkelsen, norsk medborgerpanel, ungdata og valgundersøkelsen.

Tilgang for internasjonale forskere er et utviklingstrinn som er planlagt. Dette er viktig ikke bare for å gi utenlandske forskere tilgang til norske data, men for etterprøvbarheten når norske forskere publiserer forskning basert på microdata.no internasjonalt. Ved å oppgi kommandoskriptet, eventuelt bare en kode, ved innsending av artikkelmanus vil fagfeller kunne gå inn og kontrollere og reprodusere de eksakte resultatene som danner grunnlaget for artikkelen.

Forbedret reproduserbarhet og delbarhet har i det hele tatt forbedringspotensial. For eksempel kan man tenke seg en løsning hvor forskere får kontakt med hverandre og kan kommunisere internt i analysemiljøet. På den måten kan skript deles internt uten å gå veien om klipp og lim i e-post, eller flere forskere kan tillates å jobbe i samme skript samtidig. Dette vil effektivisere arbeidsflyten og øke graden av transparens og etterprøvbarhet i forskningen. Her er det verdt å minne om at all bruk, enhver kommando, lagres i fem år. De som forskes på, er anonyme, men forskerne er ikke det, og all behandling av data er sporbar.

Som allerede nevnt vil ikke microdata.no kunne dekke alle behov rundt registerdataforskning. Selv med sofistikert teknologisk utvikling vil det være tilfeller hvor innsyn i datamatrisene er nødvendig. Det er derfor viktig at utviklingen av microdata.no ikke går på bekostning av en forbedret tilgjengelighet av rådata i de tilfellene hvor det er nødvendig. Microdata.no kan imidlertid være en del av løsningen for å effektivisere slik tilgang. Man kan tenke seg en egen type innlogging via ID-porten hvor forskere får tilgang til ferdig bearbeidede datasett som de får *fullt innsyn* i, og som er begrenset til variabler de har søkt tillatelse til å behandle. På den måten får man utnyttet fordelene med microdata.no, nemlig tilgang til kilden (ikke kopier), ferdig tilrettelagte data, sporing av aktivitet og etterprøvbarhet. Flere andre land, deriblant Sverige og Danmark, har allerede lignende fjerntilgangløsninger.

Konklusjon

Microdata.no imøtegår krav norske samfunnsforskere har stilt i mange år, nemlig å få enklere og billigere tilgang til norske registerdata. Samtidig som forskning basert på registerdata har blitt stadig mer etterspurt, har det på grunn av stigende priser og lang saksbehandlingstid blitt vanskeligere for forskere å få tilgang til registerdata. Microdata.no utgjør et alternativ til utlån av data ved å integrere registerdata, analyseprogram og personvern i én enkelt portal. Tjenesten er unik i verdenssammenheng og vil bidra til innsparing av betydelige kostnader, ressurser og tid samt stimulere til økt bruk av registerdata i forskning. Til tross for disse fordelene har løsningen noen ulemper, som begrenset utvalg av analysemetoder, begrenset fleksibilitet knyttet til grafisk visualisering og residualdiagnostikk samt standard-winsorisering av datasett. For å bøte på dagens begrensninger og for at tjenesten skal oppnå sitt fulle potensial, er utstrakt bruk i forskning og økte bevilgninger en forutsetning.

Takk

Takk til Ørnulf Risnes i NSD og Johan Heldal i SSB for at de har gitt meg innsikt i sine refleksjoner rundt utviklingen av metodiske og tekniske løsninger i microdata.no.

Referanser

Bratsberg, B., Røed, K. & Raaum, O. (2012). Gjør registerdata AKU overflødig? *Økonomiske analyser,* (5), 46–52.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637.

Datatilsynet. (2019). *Microdata.no vant pris.* Hentet 19. mars 2019 fra https://www.datatilsynet.no/aktuelt/aktuelle-nyheter-20192/microdata.no-vant-pris/

Dixon, W.J. (1960). Simplified estimation form censored normal samples. *The Annals of Mathematical Statistics, 31*(2), 385–391.

Dixon, W.J. & Yuen, K.K. (1974). Trimming and winsorization: A review. *Statistische Hefte, 15*(2), 157–170. DOI: https://doi.org/10.1007/bf02922904

Fevang, E., Røed, K., Raaum, O. & Zhang, T. (2004). Undersysselsatte i Norge: Hvem, hvorfor og hvor lenge. *Frisch-rapport,* (7), 2004.

Fløtten, T., Helgøy, I., Skule, S. & Storsul, T. (2017, 30. oktober). Norges gull må være tilgjengelig for forskning. *Aftenposten*. Hentet fra https://www.aftenposten.no/meninger/debatt/i/79Ln8/Norges-gull-ma-vare-tilgjengelig-for-forskning--Tone-Flotten_-Ingrid-Helgoy_-Sveinung-Skule-og-Tanja-Storsul

Jakobsen, S.E. (2010, 18. juli). Samfunnsvitere må bruke tall. *forskning.no*. Hentet fra https://forskning.no/okonomi-samfunnskunnskap-sosiologi/2010/06/samfunnsvitere-ma-bruke-tall

Jakobsen, S.E. (2014, 9. april). Pris-hopp på offentlige opplysninger. *forskning.no*. Hentet fra https://forskning.no/forskningsfinansiering-forskningspolitikk-forskningspriser/2014/04/pris-hopp-pa-offentlige

Kokic, P. & Bell, P. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of official statistics, 10*, 419–419.

Kvittingen, I. (2018, 10. september). Krise i forskningen: Klarer fortsatt ikke å bekrefte andres studier. *forskning.no*. Hentet fra https://forskning.no/samfunnskunnskap-om-forskning/krise-i-forskningen-klarer-fortsatt-ikke-a-bekrefte-andres-studier/1231743

NSD & SSB. (2018, Juni 2018). *Brukermanual for analysesystemet microdata.no*. Hentet fra https://microdata.no/brukermanual.pdf

Pan, J., Sarkar, I. & Mudholkar, G.S. (2009). Robust winsorized regression using bootstrap approach AU – Srivastava, Deo Kumar. *Communications in Statistics – Simulation and Computation, 39*(1), 45–67. DOI: https://doi.org/10.1080/03610910903308423

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika, 81*(2), 373–383.

SSB. (2018, 30. oktober). *Data til forskning – steg for steg.* Hentet fra https://www.ssb.no/omssb/tjenester-og-verktoy/data-til-forskning#Hva_koster_det_og_hvor_lang_tid_tar_det

Tukey, J. (1960). A survey of sampling from contaminated distributions. I I. Olkin (red.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (s. 448–485). Stanford, CA: Stanford University Press.