Kandidatuppsats

Statistiska institutionen

Bachelor thesis, Department of Statistics

Nr 2014:14

Röjandekontroll med syntetiska dataset

Tillämpning och teori

Disclosure Control with Synthetic datasets

Application and Theory

Alexander Astlind och Joakim Bonilla

Självständigt arbete 15 högskolepoäng inom Statistik III, HT2014

Handledare: Lars Lyberg

Sammanfattning

I uppsatsen avser vi ge läsaren en översikt om hur syntetiska dataset kan användas som röjandekontrollmetod i statistikproduktion. För att uppnå detta mål kommer vi presentera en teoretisk genomgång där metoderna för detta stegvis förklaras. Vi kommer även ge en demonstration på ett praktiskt tillvägagångsätt i SAS.

Nyckelord: Multipel imputering, Nyckelvariabler, Mikrodata.

Förord

Vi vill framförallt tacka vår handledare Lars Lyberg som gav oss värdefull vägledning och uppmuntran.

Innehållsförteckning

1 Introduktion ... 7
1.1 Disposition ... 7
1.2 Det svenska statistiska systemet ... 7
1.3 Röjande och röjandekontroll ... 8
1.4 Nyckelvariabler ... 8
1.5 Röjningskontrollmetoder ... 9
2 Introduktion till imputering ... 10
2.1 Bortfallsmekanismer ... 10
2.2 Introduktion till enkel imputering ... 11
2.3 Multipel imputering ... 12
2.3.1 Joint Modeling ... 13
2.3.2 Fully conditional specification ... 13
2.3.3 Frekventistisk multipel imputering ... 14
2.4 Syntetiska dataset ... 15
2.5 Historisk utveckling av syntetiska dataset för röjanderiskkontroll ... 16
2.6 Exempel på syntetiska dataset för röjningskontroll ... 16
3 Multipel imputering för röjanderiskkontroll ... 17
3.1 Helt syntetiska dataset ... 17
3.1.1 Inferens för univariata estimatorer ... 18
3.1.2 Multivariata estimatorer för hypotesprövning ... 19
3.2 Delvis syntetiska dataset ... 20
3.2.1 Inferens för univariata estimatorer ... 20
3.2.2 Multivariata estimatorer ... 21
3.3 Bedömning av röjningsrisk för Syntetiska dataset ... 22
3.3.1 Bedömning av röjningsrisk för delvis syntetiska dataset ... 23
3.4 Data användbarhet ... 24
4 Egen syntetisering av mikrodata ... 24
4.1 Om undersökningen ... 24
4.2 Ett delvis syntetiskt dataset ... 25
4.2.1 Modellen ... 25
4.2.2 Utvärdering av användbarhet ... 26
4.2.3 Utvärdering av röjningsrisk ... 29

5 Slutsatser efter teoretisk översikt 31
5.1 Jämförelse mellan delvis syntetiska och helt syntetiska dataset 31
5.1.1 Uppgiftlämnarperspektiv 31
5.1.2 Användarperspektiv 31
5.1.3 Statistikproducentens perspektiv 32
6 Slutsatser efter implementering i SAS 32
7 Jämförelse med andra röjningsmetoder 33
Litteraturförteckning 34
Bilaga A 37

1 Introduktion

När nationella statistikbyråer publicerar data krävs ansträngningar för att skydda uppgiftslämnarnas identitet från att röjas. De kraven är oftast reglerade i respektive lands lagstiftning. I Sverige utgör offentlighets- och sekretesslagen ett skydd som reglerar att uppgifterna som lämnas till SCB enbart används till forskning eller statistisk produktion (SCB 2014). Ett annat skäl än de rent juridiska är att statistikbyråer är måna om att bygga upp ett förtroende bland uppgiftslämnarna så att de ska känna att de kan svara på frågor av känslig karaktär utan att riskera identifiering. Tidigare har det enbart varit statistikproducenten som har haft tillgång till ursprungsdata på individnivå (dvs. *mikrodata*) och presenterat denna i tabulerad form. Men i och med Internets framfart och persondatorer med allt högre kapacitet så ökar efterfrågan på mikrodata. Det är idag inte ovanligt att bland annat forskare vill kunna genomföra egna analyser direkt från mikrodata. Detta ställer krav på de som tillhandahåller data att ingen uppgiftslämnare kan identifieras utifrån sina svar.

1.1 Disposition

I uppsatsen kommer vi att gå igenom metoder som skyddar uppgiftslämnare från att identifieras även när deras svar används till att skapa mikrodatafiler som sedan publiceras. Främst kommer vi inrikta oss på metoder som helt eller delvis ersätter originaldatasetet med fabricerade variabelvärden. I kapitel 3 presenteras en teoretisk översikt om dessa ämnen. Detta följs av ett förslag på praktisk implementering i SAS i kapitel 4. I bilaga A bifogas SAS-koder.

Innan dess ges en introduktion till hur identifikationsskydd i svensk offentlig statistikproduktion behandlas idag samt en definiering av relevanta nyckelbegrepp. Detta behandlas i kapitel 1.

I kapitel 2 förklaras metoder för multipel imputering, metoder som sedan används till konstruera dataseten i nästkommande kapitel. Dessförinnan ges en kort introduktion till andra imputeringsmetoder.

Slutsatser från den teoretiska översikten samt den praktiska implementeringen presenteras i kapitel 5, 6 och 7.

1.2 Det svenska statistiska systemet

I Sverige produceras den officiella statistiken av 27 olika myndigheter. Dessa myndigheter beslutar själva om statistikens innehåll och omfattning inom sitt respektive ansvarsområde, det är möjligt för regeringen att fatta beslut om vad för sorts statistik som myndigheten ska producera (SCB 2011). Bestämmande av vilka statistiska mått som ska redovisas sker i samråd med viktiga användare och kraven från EU. Kvalitetskraven på den officiella statistiken för EU-länderna finns specificerade i Eurostats kvalitetsramverk. SCB har det övergripande ansvaret för att samordna den statliga statistikproduktionen.

Samtliga statistikproducerande myndigheter i Sverige har en skyldighet att beakta behovet av skydd för fysiska och juridiska intressen. Vidare måste de myndigheter som producerar statistik även följa de bestämmelser om sekretess som finns i offentlighets- och sekretesslagen (2009:400) och personuppgiftslagens (1998:204) regler om skydd mot kränkning av den personliga integriteten. När statistikproducerande myndigheter tillhandahåller statistiskt material till utomstående får inget material som kan möjliggöra röjande förekomma i data (SCB 2011). Alla data som tillhandahålls för allmänheten i Sverige publiceras i tabulerad form. SCB hävdar att efterfrågan på avidentifierad mikrodata, så kallade ”*public use files*”, inte är tillräckligt stor i Sverige och att publicering av sådan data kan riskera att skapa oro bland uppgiftslämnarna.

1.3 Röjande och röjandekontroll

Med *röjande* menas att en *angripare* har lyckats identifiera en eller fler objekt utifrån släppt data för att på så sett komma över ytterligare information (Hundepool et.al 2012 s. 247). Syftet med ett röjande skiljer sig ofta åt mellan angriparna. Det kan t.ex. handla om angripare som vill uppdatera sina egna databaser eller angripare som enbart är ute efter information om ett specifikt objekt (Elliott & Dale 1999). Med angripare menas en dataanvändare som försöker länka ihop en individ med ett objekt i mikrodata eller hitta specifika attribut för denna i populationen från aggregerade data (Hundepool et.al 2012 s.250).

Angriparnas tillvägagångssätt skiljer sig ofta åt, detta gör det svårt för dataproducenter att upprätthålla ett fullgott skydd. Störst risk för röjande uppstår om angriparen har god kännedom om objekts bakgrundsinformation samt tillgång till mikrodata då denna kan utföra omfattande datamatchningar (SCB).

Trots att ett röjande inte alltid leder till att angriparen får tag i korrekta uppgifter så kan det få konsekvenser för den röjda individen. Det är därför viktigt att inte enbart skydda de känsliga variablerna genom olika manipulationsmetoder utan även de variabler som kan tänkas hjälpa angriparen att röja individer. Vidare är det svårt att definiera vad som är känsligt då man måste hänsyn till att denna definition kan ändras över tid (SCB).

Med *röjandekontroll* menas de åtgärder dataproducenter använder för att skydda data. Behovet av röjandekontroll kan ses både utifrån uppgiftslämnarna och från statistikproducentens perspektiv. För uppgiftslämnarna är det naturligtvis viktigt att känslig data inte kommer till illasinnade angripare och/eller allmänhetens kännedom. Från statistikproducenterna sida är det angeläget att bygga upp ett långsiktigt förtroende gentemot uppgiftslämnare så att risken för röjande inte påverkar deras vilja att delta i framtida undersökningar (SCB). På senare år har mer resurser lagts på att hitta nya sett att minska den upplevda röjanderisken för att minska bortfallet (Hundepool et.al 2012 s.3). Det är den upplevda risken för röjning som kan få en individ eller organisation från att avstå från deltagande i en undersökning (Biemer & Lyberg 2003 s.104).

1.4 Nyckelvariabler

En *nyckelvariabel* är en sådan variabel som förekommer i två olika dataset och därmed kan användas för att länka ihop uppgifter om en individ mellan dataseten (Hundepool

et.al 2012 s. 250). Nyckelvariabler kan klassas utifrån två dimensioner, *nyckeltyp* och *nyckelkvalitet*. Nyckeltyp kan ses som ett mått på hur tillgänglig nyckeln är för allmänheten och för potentiella angripare. Det syftar alltså på var och hur enkelt det är att få tag på nämnda nyckel. Dessa nyckeltyper kan i sin tur delas in i tre huvudgrupper (Elliot & Dale 1999).

1) Offentlig information
   Detta är information som finns tillgänglig för allmänheten, exempelvis uppgifter ur telefonkatalogen, taxeringskalendern och så vidare. Denna nyckel är enklast för angripare att få tag på.
2) Personlig information
   Detta är sådan information som angriparen kan erhålla genom personlig kontakt med individer i undersökningen. Desto längre relationen pågår mellan parterna desto fler och mer detaljerade nycklar kan angriparen erhålla. Exempel på sådana nycklar är ålder, civilstånd, och yrkeskategori.
3) Organisatoriska databaser
   Den tredje typen av information syftar på databaser som kommersiella företag eller myndigheter byggt upp. För att angriparen ska kunna använda denna nyckel krävs alltså tillgång till externa databaser.

Med nyckelkvalitet menas hur användbar informationen i nyckeln är för angriparens försök att röja individer ur ett dataset. Om en viss typ av svar i undersökningen sticker ut som särskilt ovanliga blir bakomliggande uppgiftslämnare lättare att identifiera, förutsatt att angriparen har kännedom om detta. Vet till exempel en angripare att företag A är det enda företaget med mer än 1000 anställda i region Z så kan denna enkelt identifiera företaget om både antal anställda och region presenteras i ett dataset. Vidare beror nyckelkvaliteten på hur noggrant variabler presenteras. Att länka ihop information som finns i två separata dataset underlättas om svaren har *kodats* på en alltför detaljrik nivå. Antag att en angripare har kännedom om en uppgiftslämnares exakta inkomst, sannolikheten att erhålla unika eller nästan unika träffar i datasetet ökar ju mer precist denna variabel presenteras. Kodning är en process för att klassificera öppna svarskategorier till förbestämda kategorier (Biemer & Lyberg 2003 s. 217).

Även nyckelns stabilitet påverkar kvaliteten. Stabila variabler, så som födelsedatum, förenklar matchningar med tidigare utgivna dataset samt förlorar inte sitt värde över tid. (Elliot & Dale 1999). Mer föränderliga variabler, som t.ex. inkomst, kan däremot tappa sitt värde över tid men är fortfarande viktiga att skydda.

1.5 Röjningskontrollmetoder

Genom åren har en rad olika röjningskontrollmetoder arbetats fram för att stärka skyddet mot angripare. Det finns metoder som används för tabulerad data och andra som appliceras på mikrodata. Fokus häri kommer att läggas på de metoder som används för röjningskontroll av mikrodata.

Förgrovning

En av de enklaste röjningskontrollmetoderna för mikrodata är *förgrovning* som bygger på att minska detaljnivån för variablerna. Antag att en uppgiftslämnare har uppgett hela sitt födelsedatum, då kan statistikproducenten välja att endast publicera födelseåret för att försvåra identifiering (SCB).

Lägga till störningar i datasetet

För att förhindra identifiering kan de observerade värdena utsättas för störningar. En vanlig störning är att addera eller multiplicera originalvärdena med en slumpvariabel som följer en känd fördelning, till exempel normalfördelningen.

Dolda värden

En individs värde på någon eller några av variablerna kan döljas om statistikproducenten anser att individens värden är sådana att de kan användas för identifiering.

Vi kommer längre fram i uppsatsen gå igenom hur syntetiska dataset kan användas som röjanderiskmetod. Detta är en metod förutsätter att imputeringar görs i datasetet. Vi kommer därför i nästkommande kapitel behandla imputering.

2 Introduktion till imputering

Med *imputering* menas metoder som ersätter saknade värden i dataset (Biemer & Lyberg 2003 s. 81). Oftast genomförs dessa metoder för att korrigera för *partiellt bortfall*, dvs. när uppgiftslämnare felaktigt har utelämnat att svara på en fråga (Biemer & Lyberg 2003 s. 80). Den ursprungliga anledningen till att skatta dessa saknade värden är att skapa kompletta dataset vilket underlättar matematiska beräkningar (Biemer & Lyberg 2003 s. 114). Idag används metoderna dessutom till att höja kvaliteten på dataset, där kvalitet oftast syftar på högre precision vid diverse estimatskattningar (Israëls et. al. 2011 s. 4-5).

2.1 Bortfallsmekanismer

Bortfallsmekanismer syftar på orsaken till varför data saknas. Det är nödvändigt att innan diverse imputeringsmetoder appliceras på data först undersöka detta. Antag till exempel att en delpopulation i en undersökning med en viss egenskap systematiskt väljer att inte svara på en fråga rörande denna egenskap. Om detta inte modelleras för i imputeringsmetoden så är risken stor att ett bias introduceras i datamaterialet för den variabeln. Dessa bortfallsmekanismer introducerades först av Rubin och presenteras nedan (Rubin 1976):

Missing completely at random (MCAR)

Vid MCAR beror förekomsten av saknade värden för en variabel inte av det specifikt saknade värdet, värdet för andra variabler eller bortfallsmönstret för andra variabler (Pigott 2001 s. 355). Det vill säga förekomsten av saknade värden är helt slumpmässigt. Om vi antar att r innebär besvarad fråga. Ges sannolikheten för att en fråga besvaras av (Missingdata 2014a):

$$Pr(r|Y_{obs}, Y_{miss}) = \Pr(r), \qquad (\mathbf{2.1})$$

Där Y_{obs} är observerade värden och Y_{miss} är saknade värden.

Missing at random (MAR)

Saknade värden för en variabel förekommer slumpmässigt men är betingade av värdet eller bortfallet för de andra observerade variablerna (Pigott 2001 s. 355). Bortfallet beror alltså inte av de saknade värdena eller andra omätbara variabler(Missingdata 2014b). Förekomsten av saknade värden är möjlig att modellera genom att kontrollera de faktorer som betingar bortfallet genom modellering. Sannolikheten för att uppgiftslämnare har svarat ges av (enligt tidigare definition av storheterna) (Missingdata 2014b):

$$Pr(r|Y_{obs}, Y_{miss}) = \Pr(r|Y_{obs}), \qquad (\mathbf{2.2})$$

Missing not at random (NMAR)

Förekomsten av saknade värden är systematiskt beroende av okända faktorer (Pigott 2001 s.356) eller de saknade värdena och bortfallet är därför omöjligt att modellera utan att få information om de saknade värdena.

2.2 Introduktion till enkel imputering

Med *enkel imputering* avses här metoder som ersätter ett värde med endast **ett** modellerat/skattat värde. I nästa kapitel (2.3) redogörs för *multipel imputering* som är metoder är som skattar **flera** värden för varje ersatt. Det är metoderna för multipel imputering som sedan används i vidare beskrivna röjandekontrollmetoder.

Nedan följer några exempel på metoder för enkel imputering:

Deduktiv imputering

Deduktiv imputering skiljer sig från de övriga i det avseendet att den inte skattar värden. Istället beräknas värdet med hjälp av logiska slutsatser (Israëls et. al. 2011 s. 13).

Lått oss exemplifiera det med att en uppgiftslämnare uppger att denna är under 18 år men missat att fylla i om denna innehar körkort. Statistikproducenten kan då med logiskt slutledning anta att uppgiftslämnaren inte innehar körkort och därmed imputera det korrekta svaret.

Imputering av medelvärde

Ett enkelt, men inta alltid statistiskt försvarbart, sett att imputera är att använda sig av medelvärdet av variabeln ifråga som skattning för det saknade värdet. Metoden

rekommenderas enbart om det finns ett förhållandevis litet antal saknade värden (Israëls et. al. 2011 s. 18). Anledningen är att metoden systematiskt underskattar variansen vilket poängteras med variansformeln nedan,

$$\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad (\mathbf{2.3})$$

(om skattade värden för y_i är lika med $\bar{y}$ så bidrar det inte till någon förändring i täljaren).

Regressionsimputering

Detta är som namnet antyder en metod där de saknade värdena skattas med hjälp av en regressionsmodell. Då metoden använder sig av ytterligare information, dvs. andra variabler än y_i, så ges generellt sett bättre skattningar än när enbart medelvärdet används. Men om varje enskilt skattat värde blir ”för bra” så uppstår samma problem som tidigare, variansen underskattas. Därför introduceras ofta en slumpmässig felterm in i modellen (Israëls et. al. 2011 s. 26).

Hot deck

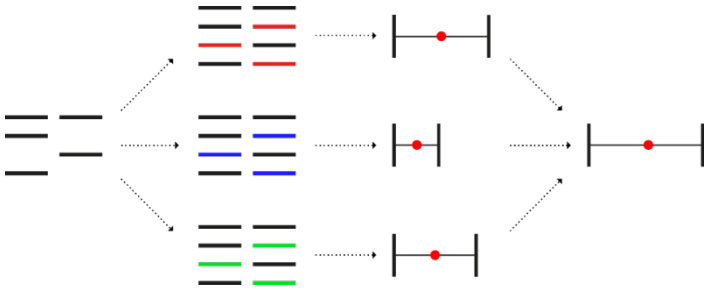
I den här metoden ersätts de saknade värdarna från erhållna värden från andra observationer i datasetet, så kallade *givare*. Valet av givare sker slumpmässigt, men vanligtvis har givare vars övriga egenskaper matchar väl med *mottagaren* (individen med det saknade värdet) en högre sannolikhet att bli vald (Israëls et. al. 2011 s. 33).

2.3 Multipel imputering

Förutom deduktiv imputering så har alla tidigare nämnda metoder en gemensam nackdel; oavsett hur väl en modell anpassas för att skatta bortfall så går det inte med säkerhet hävda att det imputerade värdet är lika med det ”korrekta” värdet (dvs. det värde som uppgiftslämnaren skulle fyllt i). För att spegla denna osäkerhet som enkel imputering generellt sett innebär föreslår Rubin en metod som skattar **flera** värden för varje saknat (Rubin 1978). Det innebär att metoden genererar flera dataset, ett för varje runda av skattade värden. Varje dataset behandlas sedan som om de vore kompletta och det är variationen mellan dessa dataset som reflekterar osäkerheten (Rubin 1978).

Idéerna bakom multipel imputering illustreras bäst genom ett förenklat exempel. Anta att vi har ett dataset med enbart två variabler, Y_1 där vi har observerat alla variabelvärden och Y_2 där några värden saknas. Låt Y_S beteckna de enheter där data saknas. För vissa enheter kommer data för både Y_1 och Y_2 ha observerats. Sambandet mellan Y_1 och Y_2 undersöks då för att skatta fördelningen för $P(Y_S|Y_1)$. Genom att simulera $m \geq 2$ antal värden för varje saknat från denna fördelning erhålls de skattade värdena $\hat{Y}_S$.

Därefter analyseras varje dataset för sig som om de vore komplett. Resultatet av dessa analyser slås sedan samman (poolas) för att ge giltiga inferenser. Figuren 2.1 nedan ger en överskådlig blick över hela processen.



Figur 2.1. Schema över imputeringsprocessen. Vi utgår från ett dataset med saknade värden, i steg ett imputeras värden för dessa i m stycken dataset. I steg två så analyseras dessa separat som om de vore kompletta och i steg tre kombineras resultatet från steg två.

Multipel imputering är däremot inte en enskild metod. Det finns flera varianter på hur de saknade värdena skattas. Vi skiljer här mellan *Joint Modeling* (JM) och *Fully Conditional Specifikation* (FCS).

2.3.1 Joint Modeling

JM innebär att de skattade värdena dras från en simultanfördelning, dvs. $P(\hat{Y}_S|Y_O)$ där Y_O är de observerade värdena. Dock går det generellt sett inte att specificera denna direkt utan $P(\hat{Y}_S|Y_O)$ skrivs om till:

$$P(\hat{Y}_S|Y_O) = \int P(\hat{Y}_S, \theta|Y_O)d\theta = \int P(\hat{Y}_S|Y_O, \theta)P(\theta|Y_O)\, d\theta\,, \qquad (\mathbf{2.8})$$

Utifrån denna ekvation så genereras imputeringarna i två steg. Först dras slumpmässiga parametervärden för θ utifrån de observerade data posteriori fördelning $P(Y_S, \theta|Y_O)$. Sedan används dessa θ för att generera Y_S från sannolikhetsfördelningen, dvs. $P(Y_S|Y_O, \theta)$. Det första steget, då θ skattas, brukar kräva *Markov Chain Monte Carlo* (MCMC) simuleringar då data sällan följer standardfördelningar. Kort sammanfattat är MCMC simuleringar ett sett att genom matematiska iterationer (omberäkningar) skatta en bakomliggande sannolikhetsfördelning för att möjliggöra slumpmässiga dragningar från denna (Geyer).

2.3.2 Fully conditional specification

FCS-modellering undviker problemet med att dra värden från krångliga simultan-fördelningar med många variabler genom att istället dra värden flera gånger från

univariata fördelningar som är lättare att härleda (Drechsler 2011, s. 15). En regressionsmodell som är lämplig för den specifika variabeln som ska imputeras skapas för alla variabler i datasetet, på det sättet behandlas varje variabel separat. Kontinuerliga variabler kan imputeras med vanliga regressionsmodeller, binära variabler kan imputeras med logit-modeller.

Denna metod är iterativ, för att starta iterationen krävs att startvärden imputeras. Denna första imputering kan ske genom enklare imputeringsmetoder som *mean imputation* eller genom *regressionsimputering*. Värdena behöver endast vara rimliga eftersom de sedan kommer att uppdateras för varje genomförd iteration.

Processen kan belysas genom att Låta $X = X_1, X_2, \dots, X_k$ vara ett antal variabler, där varje variabel $X_j = X_j^{obs}, X_j^{miss}$ är partiellt observerad, och $j = 1, \dots, k$. Vidare låt t vara en iterationsräknare. Det första steget är att imputera variabel X_1 betingad på övriga variabler i datasetet $X_2, \dots., X_k$. Sedan kommer X_1 att användas som en prediktor för nästa variabel X_2. Processen fortsätter iterativt för varje variabel tills önskad konvergens har uppnåtts och varje imputerad variabel används som en ny prediktor för imputeringar av värden för nästa variabel. Matematiskt kan processen beskrivas på följande sett om vi använder oss av tidigare introducerade storheter (Van Buuren & Oudshoorn 1999):

$$\text{För } X_1\text{: dra imputeringarna } X_1^{t+1} \text{ från } P(X_1|X_2^t, X_3^t, \dots, X_k^t)$$

$$\text{För } X_2\text{: dra imputeringarna } X_2^{t+1} \text{ från } P(X_2|X_1^{t+1}, X_3^t, \dots, X_k^t)$$

$$:$$

$$\text{För } X_k\text{: dra imputeringarna } X_k^{t+1} \text{ från } P(X_k|X_1^{t+1}, X_3^{t+1}, \dots, X_k^t)$$

Med denna metod dras nya värden för parametrarna direkt från a posteriori fördelningen för observerad data. Det behövs alltså inga MCMC tekniker för att få nya värden från den kompletta a posteriori fördelningen. (Drechsler 2011,s.15)

2.3.3 Frekventistisk multipel imputering

Nationella statistikbyråer följer ett frekventistiskt ramverk vilket innebär att dessa organisationer i huvudsak inte använder bayesianska imputeringsmetoder (Björnstad 2005). I ett försök att komma till rätta med några av de svårigheter nationella statistikbyråer ställs inför i samband med imputering presenterade Björnstad metoder för att korrekt uppskatta variansen vid imputering med icke-bayesianska metoder. Vi introducerar först Rubins kombinationsmetoder (Rubin 1996) för att göra inferens från dataset som har genererats med multipel imputering. På så sett illustreras lättare de problem som uppstår när inferens ska göras vid icke bayesiansk imputering. Låt q vara parametern som ska skattas, vidare är $q^{(i)}$ ett punktestimat i ett dataset i där $i = 1, \dots, m$ och $u^{(i)}$ är variansen för den parameter som skattas.

Punktestimatet ges av:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m\,, \qquad (\mathbf{2.4})$$

I formel (2.5) beräknas variansen mellan de imputerade dataseten och i (2.6) beräknas variansen inom de imputerade dataseten. Slutligen kombineras dessa i (2.7) som ger den slutgiltiga variansen för estimatet.

$$b_m = \sum_{i=1}^{m} \left(q^{(i)} - \bar{q}_m\right)^2 / \,(m-1), \qquad (\mathbf{2.5})$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m, \qquad (\mathbf{2.6})$$

$$T_f = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m, \qquad (\mathbf{2.7})$$

Vid icke-bayesiansk imputering kommer variansen att bli liten och inte spegla den osäkerhet som finns i själva imputeringsprocessen (Björnstad 2005), eftersom imputeringar faktiskt inte är observerad data. Björnstad föreslog att variansen mellan de imputerade dataseten(2.5) skulle ges en större vikt. Denna vikt är ett $(k = 1)$ för bayesianska imputeringsmetoder, vilket ges av den första ettan i formel (2.7). I de flesta fall som presenterades gavs vikten k för formel (2.7) som en funktion av svarsfrekvensen (Björnstad 2005). Det finns ingen generell formel för att uppskatta k vid icke-bayesiansk inputering utan formlerna för att uppskatta denna vikt förändras beroende på vilken imputeringsmetod som används, vilket kan ses som en nackdel.

I fortsättningen av denna uppsats kommer endast multipel imputering enligt ett bayesianskt ramverk att behandlas. Detta är en följd av att litteraturen kring syntetisering av dataset som röjningskontrollmetod, uteslutande använder sig av bayesiansk imputering.

2.4 Syntetiska dataset

Syntetiska dataset innehåller data som inte är direkt observerat utan istället är skapat genom modeller, i detta sammanhang genom multipel imputering. I röjanderisk-sammanhang är tanken är att släppa data som har generats från modellerna för att försvåra identifiering (Hundepool et.al 2012:258)

2.5 Historisk utveckling av syntetiska dataset för röjanderisk-kontroll

I 1993 års andra nummer av *Journal of Official Statistics* skrev Rubin och Little varsin artikel om metoder för att öka röjanderiskkontrollen med syntetiska dataset baserade på multipel imputering.

Rubin menar i sin diskussionsartikel ”Statistical Disclosure Limitation” (Rubin 1993) att det bästa sättet är att syntetisera hela datasetet. Detta uppnås genom att behandla alla individer som tillhör urvalsramen, men inte urvalet, som bortfall och imputera värden för dessa med multipel imputering. Det genereras då lika många syntetiska dataset som antalet imputeringar. Ett visst antal av dessa väljs sedan slumpmässigt ut för att publiceras till allmänheten. Fördelen med Rubins metod är att dataseten enbart innehåller fabricerad data vilket minskar risken för att enskilda individer kan identifieras.

Littles metod, beskriven i ” Statistical Analysis of Masked Data” (Little 1993), går istället ut på att enbart fabricera värden för känsliga variabler och därmed skapa så kallade delvis syntetiska dataset. Känsliga variabler kan här både innebära känsliga variabler för individen (t.ex. inkomst, mm.) och variabler som underlättar identifiering av en specifik individ i datasetet (nyckelvariabler). Då delvis syntetiska dataset innehåller faktiska uppgifter om individer är dessa oftast av högre statistisk kvalitet än fullt syntetiska. Det innebär dock en högre risk för att enskilda individer kan identifieras.

Ett decennium senare presenterades för första gången de korrekta metoderna för att erhålla korrekta inferenser från syntetiska dataset. I (Raghunathan, et al 2003) presenterades dessa för helt syntetiska dataset medan (Reiter 2003) visade det motsvarande för delvis syntetiska.

Ett år senare, dvs. 2004, publicerades metoder för hur helt syntetiska dataset skapas simultant med multipel imputering för bortfall (Reiter 2004).

Metoder för bedömning av röjanderisker i delvis syntetiska dataset presenterades i två artiklar från 2008 och 2009 (Drechsler, Reiter 2008; Reiter, Mitra 2009).

2.6 Exempel på syntetiska dataset för röjningskontroll

Hitintills har enbart några få statistikproducenter valt att använda syntetisering av dataset som röjandekontrollmetod och då det förekommit så är det främst av det delvis syntetiska slaget (Drechsler 2011, s.53). Nedan listas några exempel på undersökningar:

Survey of Consumer Finances (SCF) från 1995 är det första exemplet på en sådan undersökning (Drechsler 2011, s. 8). SCF är en undersökning som fokuserar på de amerikanska hushållens finanser. Det är en treårig undersökning som genomförs av Board of Governors of the Federal Reserve System i samarbete med Statistics of Income Division (SOI) av Internal Revenue Service (IRS). Därav innehåller den känsliga variabler av privatekonomisk karaktär, de har därför valt att syntetisera alla variabler som i undersökningen uttrycks i dollar innan publikation (Kennickell, 1998).

Ett annat exempel där delvis syntetiserad data används finns att finna på sajten ”*On the Map*” (http://onthemap.ces.census.gov/). Denna sida förmedlar pendlingsmönster för invånare över hela USA med data från den totalundersökning U.S. Census Bureau utförde 2010 (Drechsler 2011, s.9). Det är invånarnas väg från hemmet som syntetiserats och sedan publicerats. Mer information om hur syntetiseringen genomfördes presenteras av bland annat Machanavajjhala (Machanavajjhala, et. al. 2008)

Det i särklass största syntetiseringsprojektet är produktionen av publika mikrodatafiler innehållande uppgifter från Survey of Income and Program Participation (SIPP) där uppgiftslämnarna är länkade med uppgifter från Social Security Administration (SSA). Dessa mikrodata filar heter i sin tur SIPP Synthetic Beta (SSB) och är skapade av U.S. Census Bureau. Det är också U.S. Census Bureau som genomfört SIPP medan IRS samlar in uppgifterna till SSA. Förutom ”kön” och ”civilstånd” syntetiserades hela datasetet och gjordes tillgängligt för bland annat forskare efter att tillstånd givits. U.S. Census Bureau erbjuder dessutom forskare att validera sina analysresultat mot en SIPP Gold Standard File (GSF) som är en oredigerad version av datasetet. Ytterligare information om projektet går att läsa i ”The Creation and Use of the SIPP Synthetic Beta” (Abowd et. al. 2013)

3 Multipel imputering för röjanderiskkontroll

3.1 Helt syntetiska dataset

Rubin föreslog som tidigare nämnt att hela dataset bör syntetiseras för att förbättra röjanderiskkontrollen. Syntetiseringsprocessen inleds genom att behandla de enheter som inte valts ut i urvalet men tillhör målpopulationen som bortfall vilket ger kompletta dataset. Därefter imputeras multipla värden för bortfallet baserat på observerad data från den genomförda undersökningen. Imputeringarna genereras från en bayesiansk a posteriori fördelning. Detta ger m helt syntetiska dataset som nu är kompletta. Det andra steget är dra enkla slumpmässiga urval från dessa m dataset publiceras sedan (Reiter 2002).

Låt oss illustrera detta med ett exempel. Vi har information Y_{ink} om en population med N objekt utifrån ett urval n. Vi har även viss kompletterande information X om hela populationen (genom t.ex. administrativ data). Det vi saknar är information om de utanför urvalet Y_{exl}. Vi antar även för enkelhetens skull att det inte förekommer något partiellt bortfall i Y_{ink}. Det vill säga uppgiftslämnarna har besvarat alla frågor i undersökningen eller total vägrat och inte besvarat några frågor alls.

Det syntetiska datasetet skapas nu i två led. Först skapas m syntetiska dataset genom att oberoende dra Y_{exl} från a posteriori sannolikhetsfördelningen $p(Y_{exl}|X, Y_{ink})$. Om det publicerade datasetet enbart ska innehålla fabricerade data så dras alla värden för N från den här fördelningen. I nästa steg dras enkla slumpmässiga urval från de m skapade dataseten. Detta görs för att minska storleken på datamaterialet som annars kan bli för stort för att ge ut (Drechsler 2011 s. 39).

3.1.1 Inferens för univariata estimatorer

Skattningsmetoder för syntetiska dataset exemplifieras lättast om vi tänker oss att Q är den okända parametern som vi vill skatta. Den okända parametern Q skulle kunna vara ett medelvärde för någon av variablerna i data eller en korrelationskoefficient mellan två variabler. Vid analysen av de syntetiska dataseten erhålls punktestimat som betecknas $q^{(i)}$ och variansestimatet för varje enskilt dataset $u^{(i)}$ för $i = 1, \dots, m$, där m är antalet dataset. För att spegla osäkerheten i de syntetiska dataseten beräknas b_m Skattningen av Q fås genom att beräkna $\bar{q}_m$, b_m och $\bar{u}_m$ med hjälp av formlerna nedan:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \qquad (\mathbf{3.1})$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2 \Big/ (m-1), \qquad (\mathbf{3.2})$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m, \qquad (\mathbf{3.3})$$

Sedan sätts (3.2) och (3.3) in i formeln för variansen:

$$T_f = \left(1 + \frac{1}{m}\right) b_m - \bar{u}_m, \qquad (\mathbf{3.4})$$

Slutligen kombineras (3.1) och (4.4) för att bilda konfidensintervall. I fall n är stort kan en t-fördelning användas för skattningar av Q. Frihetsgraderna ges av $v_f = (m-1)(1 - \bar{u}_m / ((1 + m^{-1)}) b_m))^2$. Nackdelen med variansformeln (5.4) är att den kan anta negativa värden för modesta m (Reiter 2002). Reiter föreslog som en lösning på problemet att beräkna variansen med formel (5.5) istället:

$$\overset{*}{T_f} = max(0, T_f) + \delta \left(\frac{n_{syn}}{n} \bar{u}_m\right), \qquad (\mathbf{3.5})$$

där $\delta = 1$ om $T_f < 0$ och $\delta = 0$ annars.

Beteckningen n_{syn} är antalet obeservationer i det publicerade datasetet som har dragits ur den syntetiserade populationen. I praktiken innebär formel 5.5 att variansen ges av den genomsnittliga variansen av parametern som skall skattas givet att det syntetiska datasetet innehåller lika många observationer som det ursprungliga datasetet.

3.1.2 Multivariata estimatorer för hypotesprövning

Vid multivariat hypotestestning prövas hypotesen $\boldsymbol{Q} = \boldsymbol{Q_0}$ för en multivariat estimator med k komponenter. För att kunna genomföra testet beräknas $\overline{\boldsymbol{Q}}_m$,$\boldsymbol{B}_m$ och $\overline{\boldsymbol{U}}_m$ som är de multivariata motsvarigheterna till $\bar{q}_m$, b_m och $\bar{u}_m$, som introducerades i föregående avsnitt. Walds test är ett av hypotestesten som kan användas för multivarata estimatorer. Testet ges av denna formel:

$$S_h \ = (\overline{\boldsymbol{Q}}_m - \boldsymbol{Q}_0\ )^T\, \overline{\boldsymbol{U}}_m{}^{-1}(\overline{\boldsymbol{Q}}\ _m\ -\ \boldsymbol{Q}_0\ )/(k(r_h - 1)), \qquad (\mathbf{3.6})$$

r_h ges av $r_h = (1 + 1/m)trace(\boldsymbol{B}_m\, \boldsymbol{U}m^{-\mathbf{1}}\,)/k,$ $\qquad (\mathbf{3.7})$

Teststatistikan S_h jämförs mot en F-fördelning,F_{k,v_h} där frihetsgraderna ges av:

$v_h = 4 + (t - 4)(1 - (1 - 2/t)/r_h)^2$, och t fås av $t = k(m - 1)$, $(\mathbf{3.8})$

Om **Q** består av många komponenter k kan det vara besvärligt att använda sig av $\overline{\boldsymbol{U}}_m$. Ett likelihood-kvottest bör då istället användas för hypotestester av syntetiska dataset. Metoden introducerades av Reiter (2005). Om vi antar att ψ är en vektor med parametrar som analytikern önskar att testa, låt $\psi^{(i)}$ vara maximum likelihood skattningen av ψ som har beräknats från $D^{(i)}$. $D^{(i)}$ är det ”ionde” imputerade datasetet, som går från $i = 1, \dots, m$. Hypotesen ställs upp som $\boldsymbol{Q}(\psi) = \boldsymbol{Q}_0$. I detta fall är $\boldsymbol{Q}(\psi)$ en funktion av ψ med k dimensioner. $\psi_0^{(i)}$ är maximum likelihood skattningen som har beräknats från datasetet $D^{(i)}$, estimatet sätts sedan in i testet $\boldsymbol{Q}(\psi) = \boldsymbol{Q}_0$.

$$L^{(i)} = 2\log f\left(D^{(i)}\middle|\psi^{(i)}\right) - 2\log f\left(D^{(i)}\middle|\psi_0^{(i)}\right), \qquad (\mathbf{3.9})$$

$$\bar{L} = \sum_{i=1}^{m} L^{(i)}/m, \qquad (\mathbf{3.10})$$

$$\bar{\psi} = \sum_{i=1}^{m} \psi^{(i)}/m, \qquad (\mathbf{3.11})$$

$$\bar{\psi}_0 = \sum_{i=1}^{m} \psi_0^{(i)}/m, \qquad (\mathbf{3.12})$$

$$\bar{L}_0 = (1/m)\sum_{i=1}^{m}(2\log f\left(D^{(i)}\middle|\bar{\psi}\right) - 2\log f\left(D^{(i)}\middle|\bar{\psi}_0\right)), \qquad (\mathbf{3.13})$$

$$\hat{S}_h = \bar{L}_0/\left(k(\hat{r}_h - 1)\right), \qquad (\mathbf{3.14})$$

Frihetsgraderna ges av $\hat{r}_h = ((m+1)/t)(\bar{L} - \bar{L}_0)$, $\quad (\mathbf{3.15})$

Teststatistikan $\hat{S}_h$ jämförs mot en $F_{k,\hat{v}_h}$ fördelning där $\hat{v}_h$ ges av

$$\hat{v}_h = 4 + (t-4)(1-(1-2/t)/rf)^2 \text{och } t = k(m-1), \qquad (\mathbf{3.16})$$

Den översta formeln (3.9) visar tillvägagångssättet för att beräkna Likelihood-kvottest statistikan för ett av de imputerade dataseten. Medelvärdet av likelihood-kvottestest statiskorna fås genom att dividera summan av varje enskilt värde med antalet dataset, vilket visas i (3.10). På liknande sett fås båda medelvärdet för vektorerna ψ och ψ_0. Medelvärdet för likelihood-kvottest statistikan vid ψ och ψ_0 fås med hjälp av (3.13). Likelihood- kvottest statistikan ges slutligen av formel (3.14).

3.2 Delvis syntetiska dataset

Så som namnet antyder så ersätts enbart en del av variablerna i delvis syntetiska dataset. Bedömningen av vilka variabler som bör syntetiseras görs av statistikproducenten. Det kan dels handla om att ersätta känsliga variabler och dels om att ersätta nyckelvariabler. Det är enklast att beskriva själva processen matematiskt, beteckningar och formler är tagna ur Drechsler (2011 s. 53-54).

Låt de individer som i datasetet valts ut till att få deras variabelvärden ersatta betecknas som $Z_j = 1$ och övriga till $Z_j = 0$. Indexet j representerar här individerna som deltar i undersökningen och går från $j = 1, \dots, n$.

Låt vidare $Z = (Z_1, \dots, Z_s)$ där indexet s nu avser antalet variabler i datamaterialet.

Låt $Y = \left(Y_{ers}, Y_{org}\right)$ beteckna insamlade variabelvärden, där Y_{ers} är de värden som blivit ersatta och Y_{org} vara originalvärden (oförändrade). Vidare låt $Y_{ers}^{(i)}$ vara de imputerade värdena för Y i det syntetiska datasetet i, där i går från $i = 1, \dots, m$.

Varje sådant $Y_{ers}^{(i)}$ skapas genom dragningar från a posteriori sannolikhetsfördelningen $f\left(Y_{ers}^{(i)} \middle| Y, Z\right)$.

Denna process upprepas nu m gånger, vilket skapar $D^{(i)} = \left(Y_{org}, Y_{ers}^{(i)}\right)$ för $i = 1, \dots, m$, där $\mathbf{D} = \left\{D^{(1)}, \dots, D^{(m)}\right\}$ släpps offentligt.

3.2.1 Inferens för univariata estimatorer

För att få giltiga estimatorer går det att använda de matematiska reglerna introducerade av (Reiter 2003), dessa presenteras nedan.

Låt Q vara ett estimat, t.ex. en populations medelvärde eller varians. Antag att en analytiker estimerar Q med en punktskattning q och variansen av q med estimatorn u. För enkelhetens skull antag att det inte förekommer något bortfall i det observerade datasetet. Låt $q^{(i)}$ och $u^{(i)}$ vara värdena av q och u i det syntetiska datasetet $D^{(i)}$ för

$i = 1, \dots, m$. Analytikern beräknar $q^{(i)}$ och $u^{(i)}$ genom att behandla varje $D^{(i)}$ som riktig data. Följande formler behövs då för inferenser av skalären Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \qquad (\mathbf{3.17})$$

$$b_m = \sum_{i=1}^{m} \left(q^{(i)} - \bar{q}_m\right)^2 / \, (m-1), \qquad (\mathbf{3.18})$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m, \qquad (\mathbf{3.19})$$

$\bar{q}_m$ används som estimat för Q och, $T_d = \; b_m/m + \; \bar{u}_m$,

för att estimera variansen av $\bar{q}_m$ som här inte kan bli negativ och således behövs ingen korrektion för delvis syntetiska dataset.

När n är stort kan inferenser för skalären Q baseras på t-fördelningen med $\boldsymbol{v}_d = (m-1)(1 + \bar{u}_m/(b_m/m))^2$ frihetsgrader.

3.2.2 Multivariata estimatorer

De multivariata estimatorerna för delvis syntetiska dataset är baserade på samma princip som de för helt syntetiska. Låt $\bar{\boldsymbol{Q}}_m$,$\boldsymbol{B}_m$ och $\bar{\boldsymbol{U}}_m$vara de multivariata motsvarigheterna till $\bar{q}_m$, b_m och $\bar{u}_m$. Vidare låt hypotesen som ska testas ställas upp som $\boldsymbol{Q} = \boldsymbol{Q_0}$ där $\boldsymbol{Q}$ är en multivariat parameter med k komponenter. Walds test statistika ges av:

$$S_d \; = (\bar{\boldsymbol{Q}}_m - \boldsymbol{Q}_0\,)^T \, \bar{\boldsymbol{U}}_m{}^{-1} (\bar{\boldsymbol{Q}}_{\,m} - \; \boldsymbol{Q}_0\,)/(k(1 + r_d)), \qquad (\mathbf{3.20})$$

Där r_d ges av $r_d = (1/m) trace(\boldsymbol{B}_m \, \boldsymbol{U} m^{-\mathbf{1}}\,)/k$, $\qquad (\mathbf{3.21})$

S_d jämförs sedan mot en F-fördelning, F_{k,v_d} där frihetsgraderna fås av:

$v_d = 4 + (t-4)(1 + (1 - 2/t)/r_d)^2$, och t fås av $t = k(m-1)$, $(\mathbf{3.22})$

Likt fallet för helt syntetiska dataset kan användandet av $\bar{\boldsymbol{U}}_m$ vara besvärligt om $\boldsymbol{Q}$ består av många komponenter, k. Det går då istället att använda ett likelihood-kvottest för delvis syntetiska dataset. Låt $\boldsymbol{\psi}$ beteckna en vektor av parametrar som ska testas, och $\boldsymbol{\psi}^{(i)}$ vara maximum likelihood estimatet av $\boldsymbol{\psi}$ beräknat från ”ionde” syntetiska dataset $D^{(i)}$ där i går från 1,…, m. Hypotesen ställs då upp som $\boldsymbol{Q}(\boldsymbol{\psi}) = \boldsymbol{Q}_0$, där $\boldsymbol{Q}(\boldsymbol{\psi})$ är en funktion av $\boldsymbol{\psi}$ med k dimensioner.

Vidare låt $\boldsymbol{\psi}_0^{(i)}$vara maximum likelihood estimatet av $\boldsymbol{\psi}$ som fås från $D^{(i)}$. Detta estimat används sedan i hypotesen. Nedanstående formler används för att utföra diverse hypotestest.

$$L^{(i)} = 2\log f\left(D^{(i)}\middle|\psi^{(i)}\right) - 2\log f\left(D^{(i)}\middle|\psi_0^{(i)}\right), \qquad (\mathbf{3.23})$$

$$\bar{L} = \sum_{i=1}^{m} L^{(i)}/m, \qquad (\mathbf{3.24})$$

$$\psi = \sum_{i=1}^{m} \psi^{(i)}/m, \qquad (\mathbf{3.25})$$

$$\bar{\psi}_0 = \sum_{i=1}^{m} \psi_0^{(i)}/m, \qquad (\mathbf{3.26})$$

$$\bar{L}_0 = (1/m)\sum_{i=1}^{m}(2\log f\left(D^{(i)}\middle|\bar{\psi}\right) - 2\log f\left(D^{(i)}\middle|\bar{\psi}_0\right)), \qquad (\mathbf{3.27})$$

$$\hat{S}_d = \bar{L}_0/\left(k(1+\hat{r}_d)\right), \text{där } \hat{r}_d \text{ ges av } \hat{r}_d = (1/t)(\bar{L} - \bar{L}_0), \qquad (\mathbf{3.28})$$

och t återigen ges av $t = k(m-1)$,

Den första formeln i stycket ovan (3.23) visar beräkningarna för en log-likelihood-kvot-test-statistika som är associerad med ett dataset $D^{(i)}$. Eftersom information från alla dataset används i hypotestestet så beräknas medelvärdet för $L^{(i)}$, $\psi^{(i)}$, $\psi_0^{(i)}$, detta visas i formel (3.24) till och med (3.26). Medelvärdet av log-likelihood-kvottest-statistikorna behövs också och visas i formel (3.27).

Liklihood-kvottest statisikan ges slutligen i formel 5.28. $\hat{S}_d$ jämförs sedan mot en $F_{k,\hat{v}_d}$ – fördelning. Frihetsgraderna fås genom att modifiera formel 5.22 till 5.30, lägg märkte till att det är skattningarna från 5.28 som används.

$$\hat{v}_d = 4 + (t-4)(1 + (1 - 2/t)/\hat{r}_d)^2, \qquad (\mathbf{3.29})$$

3.3 Bedömning av röjningsrisk för Syntetiska dataset

Generellt sett är röjningsrisken för syntetiska dataset liten (Drechsler 2011, s. 42). Enskilda objekt med ett variabelvärde som särskiljer sig från resten av populationen kan

dock ha en större röjningsrisk. Ett vanligt förekommande exempel finns att finna i företagsundersökningar. I dessa är det vanligt att de större företagen har en högre sannolikhet att väljas i urvalet eftersom de har en sådan betydande roll, dock är dessa mindre till antalet vilket gör att de sticker ut. Om det syntetiska dataset ska spegla verkligheten måste då några företag även i denna sticka ut, men eftersom de större företagen är så få till antalet så är det svårt att imputera värden som skiljer sig allt för mycket mot de ursprungliga utan att förvränga variabelsamband och parameterestimat betydligt.

För undersökningssituationer som rör privatpersoner är det främst den upplevda röjningsrisken som måste beaktas. Det kan förekomma imputerade värden som ligger väldigt nära de värden som respondenten uppgav, vilket kan få uppgiftslämnaren att tro att datainsamlaren har publicerat ursprungsdata. Detta kan även ske omvänt, nämligen att en inkräktare tror sig ha hittat en individs korrekta data när de imputerade värdena ligger nära de verkliga värdena.

Teorin kring bedömning av röjningsrisk för syntetiska dataaset är inte lika utvecklad som den för delvis syntetiska dataset. Det finns några formella metoder som har lanserats för att bedöma röjningsrisken för syntetiska dataset men som inte är applicerbara i vårt illustrationsexempel som redovisas i avsnitt 4. För fördjupning kring dessa metoder rekommenderas läsning av Abowd och Vilhuber (2008).

3.3.1 Bedömning av röjningsrisk för delvis syntetiska dataset

För delvis syntetiska dataset är röjningsrisken något högre jämfört med risken för helt syntetiska dataset (Drechsler 2011, s. 56). Röjningsrisken för ett delvis syntetiskt dataset stiger ytterligare om angriparen vet att en viss individ har deltagit i undersökningen, eftersom observerade värden blir kvar i det publicerade datasetet. Det är därför viktigt att noggranna kontroller görs innan allmänheten får tillgång till partiellt syntetiska dataset. Röjningskontrollen för partiellt syntetiska dataset är minst lika viktig som att säkerställa att användbarheten för data är hög. Endast om data uppfyller statistikproducentens krav på röjningskontroll och dataanvändbarheten bör datasetet göras tillgängligt för allmänheten (Drechsler 2011, s. 56).

En metod som vanligtvis används för att bedöma röjningsrisken i delvis syntetiserad mikrodata är så kallad Distance Matching (DM). När metoden tillämpas jämförs det ursprungliga datasetet mot det syntetiserade. Detta sker genom att välja ett objekt i det ursprungliga datasetet och sedan succesivt jämföra det mot alla objekt i det syntetiserade datasetet genom att beräkna avståndet mellan variabelvärdena för respektive objekt. Avståndet beräknas med olika varianter av euklidiskt avstånd. De variabler som inte är syntetiserade används som en form av nyckelvariabler som gör det möjligt att dela in objekten i båda dataseten för att på så sätt utesluta fall som inte matchar på dessa kända variabler (Abowd et.al.2013).

3.4 Data användbarhet

Hitintills har vi främst berört dels hur skapandet av syntetiska dataset utförs och dels hur bedömningar av röjanderisker går till. I nästkommande stycken berör vi metoder för att utvärdera användbarheten för syntetiska dataset. Med användbarhet menas här hur väl de syntetiska dataseten behåller datakvalitet. Metoderna gäller för både helt och delvis syntetiska dataset.

Det finns idag två typer av metoder som kvantifierar användbarheten av syntetiska dataset (Drechsler 2011 s. 41-42). De första metoderna jämför breda skillnader mellan original- och syntetiserad data. De går i korthet ut på att kvantifiera något statistiskt avstånd mellan datasetens fördelningar. När avståndet mellan fördelningarna ökar så minskar generellt sett även datakvaliteten.

De andra metoderna jämför skillnader i specifika modeller eller estimat mellan original- och syntetiserad data. Ett enkelt sätt är att använda sig av intervallöverlappning (Karr et al. 2006). Först beräknas ett 95 % konfidensintervall för det relevanta estimatet från det syntetiska datasetet $(L_s, Ö_s)$, där L_s och $Ö_s$ står för lägre respektive övre intervallgräns. Därefter utförs motsvarande beräkningar för motsvarande estimat på originaldatasetet, $(L_0, Ö_0)$. Sedan beräknas ett intervall som utgör sträckan där de tidigare intervallen överlappar varandra, $(L_k, Ö_k)$. Till sist beräknas användbarhetsmåttet som:

$$I = \frac{Ö_k - L_k}{2(Ö_0 - L_0)} + \frac{Ö_k - L_k}{2(Ö_s - L_s)}, \qquad (\mathbf{3.30})$$

Vid hög användbarhet så erhålls ett I värde nära 1 och vid lågt ett värde nära 0. För att bedöma hela dataset så beräknas ett medelvärde av I för de variabler som ingår. Fördelen med den här metoden, jämfört med att enbart jämföra punktestimat, är att den även tar hänsyn till standardavvikelser.

4 Egen syntetisering av mikrodata

Från världsbankens hemsida (http://microdata.worldbank.org) går det få tag i en mängd olika data med tillhörande mikrodataset. Därifrån valde att syntetisera mikrodata för undersökningen *Azerbaijan - Financial Literacy Survey* från 2009 utförd av Azerbaijan Micro-finance Association (AMFA). Denna nedladdades 2014-12-29 och har referens-nummer AZE_2009_FLS_v01_M. Tidigare försökte vi även syntetisera data från Ugandas census från 2012 men då väldigt få variabler i denna korrelerar med varandra var detta svårt, dock gav det oss värdefulla erfarenheter.

4.1 Om undersökningen

Azerbaijan - Financial Literacy genomfördes som tidigare nämnt 2009. Den innehåller 1207 observationer från hushåll och individer med 154 variabler av framförallt privat-ekonomisk karaktär. Syftet med undersökningen var att ta reda på hur Azerbajdzjans invånare hushåller med sina pengar och hur medvetna de är om de olika finansiella

tjänster som erbjuds av bland annat banker. Uppgiftslämnarna valdes slumpmässigt ut ur 16 städer, 2 bosättningar och 37 byar i proportion i hur den totala befolkningen bor. I mikrodatasetet delades dessa in i ”Capital”, ”Other urban” och ”Village”. Åldern på uppgiftslämnarna sträckte sig i undersökningen mellan 18 och 82 år.

4.2 Ett delvis syntetiskt dataset

Det finns flera datorprogram som är utvecklade specifikt för röjanderiskkontroll. Ett program som kan användas för att syntetisera data är IVEware. Det är ett program utvecklat av forskare vid University of Michigan, där bland annat Raghunathan ingår. Fördelen är att dataseten kan skapas med några enkla knapptryckningar, det är alltså inte nödvändigt att lära sig något nytt programmeringsspråk.

Programmet går att ladda hem gratis på http://www.isr.umich.edu/src/smp/ive/.

Vi valde dock att hålla oss till SAS. Det första vi gjorde där var att undersöka data-materialet för att välja ut en lämplig variabel att syntetisera. Valet föll på ”*Genomsnittlig inkomst från jobb*”. Detta eftersom det dels kan uppfattas som en känslig fråga och dels eftersom den är numerisk variabel vilket underlättar jämförelser med originaldatasetet. Det är även en variabel som korrelerar med en mängd andra variabler vilket är en nödvändig förutsättning för att skapa bra modeller.

4.2.1 Modellen

Variablerna som inkluderades i modellen var:

”Genomsnittlig inkomst från jobb”
”Utvärdera ditt finansiella kunnande på en femgradig skala”
”Kön”
”Utbildning”
”Antal personer som bor i hushållet”
”Hushållets totala inkomst”
”Hushållets totala utgifter”

Vidare använde vi också variabeln som delar upp uppgiftslämnarna på huruvida de bor i huvudstaden, en stad eller en by för att skapa en modell för varje imputeringsmetod. Dessa sammanställdes sedan innan vi utvärderade resultatet.

Då de tre kontinuerliga variablerna (jobbinkomst, hushållsinkomst, hushållsspenderande) alla var snedfördelade så transformerade vi dessa genom att beräkna kubikroten, en metod som Drechsler rekommenderar (Drechsler 2010 s.21). Vi rensade även datasetet från partiellt bortfall eftersom vi ville fokusera på syntetiseringen.

Även individer utan jobb rensades bort eftersom de inte svarat på frågan. Alternativet skulle varit att koda om deras svar till noll men vi ville visa hur man kan gå tillväga om

frågor innehåller överhoppningssteg. Det slutgiltiga datamaterialet som vi syntetiserade innehåll totalt 592 observationer.

Sedan planerade vi för hur syntetiseringsprocessen skulle genomföras och framförallt hur SAS-koden skulle byggas.

Till och börja med så duplicerade vi alla observationer för att sedan radera hälften av variabelvärdena för genomsnittlig jobbinkomst. På så vis införde vi ett simulerat partiellt bortfall. Detta bortfall skattades sedan med hjälp av multipel imputering. Vi valde här att använda oss av en FCS (se stycke 4.4.2) teknik med ett minimum på 20 iterationer och med 20 imputeringar vardera för att skatta varje enskilt ”saknat” värde. I modellen infogade vi två restriktioner. Det första sattes så att enbart positiva variabelvärden skapades. Det andra avrundade de skattade värdena till närmaste heltal (närmaste kubikrot). Detta för att erhålla skattade värden utryckte i heltal.

Efter att dessa skattade värden imputerats raderade vi de ursprungliga värdena från datasetet och fortsatte istället med att utvärdera resultatet. I bilaga 1 så bifogar vi hela SAS koden med kommentarer.

4.2.2 Utvärdering av användbarhet

Vi valde att jämföra originaldatasetet mot enbart det första av de totalt 20 syntetiska dataseten för att underlätta utskrifter från SAS.

Jämförelsen mellan originaldata och ett av de syntetiska visar att vi lyckades rätt väl med imputeringen:

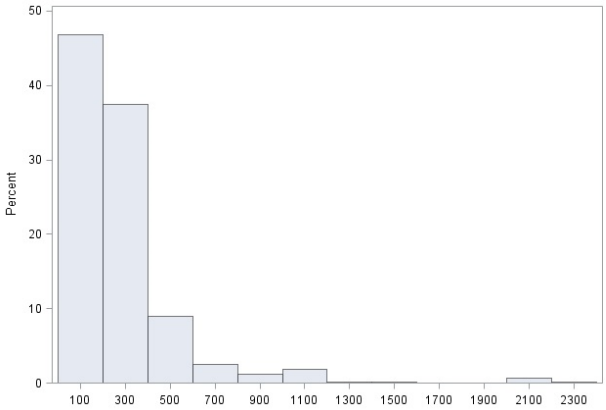
Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	260.73818	240.46173	281.01462
Std Deviation	251.19695	237.65738	266.38478
Variance	63100	56481	70961

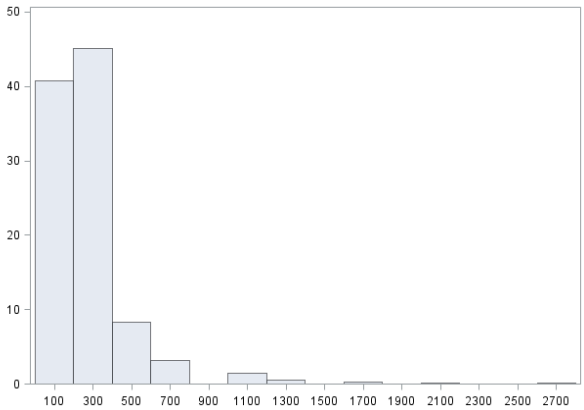
Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	263.73311	243.43632	284.02990
Std Deviation	251.44894	237.89580	266.65202
Variance	63227	56594	71103

Tabell 4.1 och 4.2: Medelvärde, standardavvikelse, varians och konfidensintervall för genomsnittlig inkomst för originaldatasetet och det syntetiserade.

Tabellerna ovan visar konfidensintervall för det skattade medelvärdet av den utvalda variabeln. Tabellen till vänster visar originaldatasetet och den till höger visar motsvarande värden för det syntetiska. Vi beräknade I värdet enligt beräkningarna presenterade i formel 5.31 och erhöll 0.93. Detta visar på en god överlappning av intervallen.

Vi fortsatte sedan med att jämföra fördelningen för variabeln mellan de båda dataseten. Detta illustreras nedan med histogram. Återigen visas resultatet för originaldatasetet till vänster och motsvarande för det syntetiska till höger i graferna.





Graf 4.1 och 4.2: Histogram över genomsnittlig inkomst för originaldatasetet och det syntetiserade.

Som graferna visar så stämmer fördelningarna relativt bra överens, även om det syntetiska tenderar att överskatta inkomsterna något.

Vi undersökte även hur väl det syntetiska datasetet har lyckats bibehålla förhållandet till övriga variabler. Detta gjorde vi genom att anpassas vanliga OLS regressioner. Som vanligt visas resultaten för de vanliga till vänster och för de syntetiska till höger.

Parameter Estimates					
Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	286.16649	32.75082	8.74	<.0001
Egen bedömning av finansiellt kunnande	1	0.42295	0.23714	1.78	0.0750
Kön	1	-111.55497	14.25894	-7.82	<.0001
Arbetssektor	1	3.67411	3.25290	1.13	0.2592
Utbildning	1	-1.73058	5.16996	-0.33	0.7379
Antal personer i hushållet	1	-27.67746	3.98360	-6.95	<.0001
Hushållets genomsnittliga inkomster	1	0.33150	0.03872	8.56	<.0001
Hushållets genomsnittliga utgifter	1	0.22897	0.04598	4.98	<.0001

Parameter Estimates					
Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	280.72111	35.82764	7.84	<.0001
Egen bedömning av finansiellt kunnande	1	0.49335	0.25942	1.90	0.0577
Kön	1	-112.23670	15.59851	-7.20	<.0001
Arbetssektor	1	2.12559	3.55850	0.60	0.5505
Utbildning	1	-3.54397	5.65566	-0.63	0.5312
Antal personer i hushållet	1	-22.11983	4.35785	-5.08	<.0001
Hushållets genomsnittliga inkomster	1	0.26864	0.04236	6.34	<.0001
Hushållets genomsnittliga utgifter	1	0.27252	0.05029	5.42	<.0001

Tabell 4.3 och 4.4: Parameterestimat för predatorer till genomsnittlig inkomst för original och syntetiserad data.

Tabell 4.3 och 4.4 visar resultatet av OLS regressionen med de 7 variabler som inkluderades i multipel imputerings-modellen. De visar att det syntetiska datasetet förhållandevis väl bibehåller datastrukturen i mot de övriga variablerna. Det korrigerade R-värdena för regressionerna var dock något lägre för det syntetiska datasetet med 0.5895 mot 0.6558 baserat på originaldata. Något förvånande visade det sig att det syntetiska datasetet även behåller sitt förhållande för variabler utanför

imputeringsmodellen. Nedan visar en regression som anpassats med variabeln ålder som predikator.

Tabell 4.5 och 4.6: Parameterestimat där ålder har anpassats i en enkel regression mot genomsnittlig inkomst för både original och syntetiserad data.

Parameter Estimates					
Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	340.73711	39.05525	8.72	<.0001
Ålder	1	-1.93541	0.91145	-2.12	0.0341

Parameter Estimates					
Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	338.94267	39.11204	8.67	<.0001
Ålder	1	-1.81954	0.91278	-1.99	0.0467

Då skillnaden mellan dataseten är små i förhållande till variabler även utanför imputeringsmodellen så visar det på att metoden är robust.

För att testa modellen ytterligare så upprepade vi proceduren 10 gånger men med olika slumptal i SAS-koden. Vi erhöll då följande värden på punktestimat och konfidens intervall.

Försök	Punktestimat	Varians	Lägre 95% KI gräns	Övre 95% KI gräns	*I* värde
1	264	63227	243	284	0.93
2	259	67108	238	280	0.96
3	263	62817	243	283	0.94
4	258	55968	239	277	0.94
5	252	52969	233	270	0.77
6	256	59442	237	276	0.90
7	266	61492	246	286	0.86
8	264	58520	244	283	0.94
9	263	60388	244	283	0.94
10	269	70281	248	291	0.79
Medelvärde	262	61221.2	242	282	

Tabell 4.7: Beskrivande statistik för olika syntetiska dataset.

Resultatet visar att metoden inte introducerar något bias. Däremot förekommer en variation men då dataproducenten har ”facit” från undersökningsresultatet kan denna välja ut de dataset som bäst beskriver originalet.

Om vi antar att dataproducenten publicerar alla de 10 dataseten i det här exemplet skulle variansen bli negativ om man beräknar den enligt formel 3.4. Däremot ger formel 3.5

värdet 61221.2 (dvs. medelvärdet av varianserna) vilket skulle ge ett 95 % konfidens-intervall mellan 242 till 282 vilket ger ett sammantaget I värde på 0.96.

En nackdel med modellen är att ingen hänsyn till det logiska sambandet ”Genomsnittlig inkomst från jobb” < ”Hushållets totala inkomst” togs. I cirka 16 % av observationerna bröts detta samband. I ett försök att modellera för detta så transformerade vi variablerna som Drechsler föreslår (Drechsler 2010 s.23). Vi började med att dividera ”Genomsnittlig inkomst från jobb” med ”Hushållets totala inkomst”. Denna kvot logit transformerades efter att värden som var lika med 1 avrundades till 0.999999999. Sedan upprepade vi den tidigare processen med undantaget att imputeringsmodellen inte längre avrundar till heltal (eftersom det skulle innebära en alldeles för stor förgrovning).

Tyvärr gav det här förfaringssättet upphov till att grovt överskatta den genomsnittliga inkomsten vilket visas i tabell 4.8 nedanför. Vidare ändrades inte 259 av observationerna någonting alls. Därför övergav vi denna modell.

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	319.15412	297.02773	341.28050
Std Deviation	269.18893	254.42949	285.77995
Variance	72463	64734	81670

Tabell 4.8: Parameterestimat och konfidensintervall för syntetiserad genomsnittlig inkomst då hänsyn tagits till logiska restriktioner.

4.2.3 Utvärdering av röjningsrisk

Det kan föreligga en röjningsrisk i syntetiska dataset om modellen har anpassats så bra att de enskilda syntetiska variabelvärdena ligger väldigt nära respektive originalvärde, speciellt om dessa sticker ur datasetet som ovanliga. Därför undersökte vi om de 5 högsta värdena i respektive dataset kom från ”samma” uppgiftslämnare. Det visade sig att 2 stycken av de ursprungliga uppgiftslämnarna kom med i 5-i topp listan, dock inte på samma placering eller med ett variabelvärde som ligger i närheten av det ursprungliga.

Den genomsnittliga relativa förändringen mellan de ursprungliga och syntetiserade variabelvärdena beräknas för varje syntetiserat dataset genom:

$$\sum_{i=1}^{n} \left| \left(\frac{Y_{org,i}}{Y_{synt,i}} - 1 \right) \right| \Big/ n, \qquad (\mathbf{4.1})$$

Där Y_{synt} betecknar syntetiserade värden, Y_{org} de ursprungliga och n är antalet observationer i datasetet. Detta mått är dock beroende av vilken storhet variablerna beräknas på och fungerar således enbart för att jämföra olika metoder applicerade på samma dataset och variabler.

I vårt syntetiska datasätt är den genomsnittliga relativa förändringen 60.9 % vilket ändå visar på en relativ hög förändring av de enskilda originalvärdena.

5 Slutsatser efter teoretisk översikt

5.1 Jämförelse mellan delvis syntetiska och helt syntetiska dataset

Vid en jämförelse mellan de båda metoderna kommer vi bedöma de i vår mening tre viktigaste aspekterna av röjanderiskkontroll. Den första aspekten omfattar ett uppgiftlämnarperspektiv, dvs. hur väl skyddar metoderna uppgiftslämnarnas integritet både från ett faktiskt och upplevt synsätt. Den andra aspekten rymmer frågor utifrån ett användarperspektiv. Här ser vi över hur väl de syntetiska dataseten bibehåller data-strukturer och således möjliggör för vidare analyser. Vi kommer här även göra en bedömning över användarvänligheten, behövs till exempel speciella program för analyser och hur mycket förkunskap krävs från användaren. Den sista aspekten tar upp vilka ansträngningar metoderna kräver från statistikproducenterna, hur enkla är dessa att implementera och hur applicerbara är de på komplicerade datamaterial.

5.1.1 Uppgiftlämnarperspektiv

Då helt syntetiska dataset enbart innehåller imputerade uppgifter från fabricerade individer utanför urvalet så blir identifikationen av faktiska individer nästintill omöjligt (Drechsler 2011, s. 62). Om angriparen däremot inte är intresserad av en specifik individ utan snarare av en mindre grupps egenskaper så är skyddet sämre eftersom de då kan erhålla skattningar som kan ligga nära det riktiga värdet. Desto bättre modell som ligger till grund för imputeringarna desto mer precisa skattningar kommer angriparen att erhålla. Ta till exempel en angripare som är intresserade av en grupp storföretags genomsnittliga omsättning. Eftersom dessa företag troligtvis särskiljer sig ur datamaterialet så blir denna information relativt enkel för en angripare att skatta utifrån ett helt syntetisk dataset (Reiter 2008).

Delvis syntetiska dataset kan inte erbjuda ett lika fullgott skydd i identifieringssamanhang eftersom det innehåller faktiska uppgifter från faktiska uppgiftslämnare (Drechsler s. 62-63).

5.1.2 Användarperspektiv

När statistikproducenter samlar in data används ofta komplicerade urvaldesigner med olika vikter och stratifieringar, något analytikern då måste ta hänsyn till i vidare analyser. Helt syntetiska dataset har den fördelen att de kan genereras utifrån andra enklare designer vilket underlättar för analytikern (Reiter 2008). Helt syntetiska kan i vissa fall även ge inferenser av högre kvalitet än analyser gjorda på urspungsdata. Det beror på att extern bakgrundsinformation från till exempel register kan användas i den bakomliggande modellen för det syntetiska datasetet (Reiter 2008). Vidare, eftersom alla data är fabricerad, är det möjligt att i undersökningar av geografisk karaktär även bifoga platsangivelser i mikrodatasetet vilket möjliggör analyser mellan mindre områden (Reiter 2008). Kvaliteten för de helt syntetiska dataseten beror däremot fullt på hur väl dataproducenten lyckas beskriva olika variabelsamband vilket en enskild analytiker ofta har svårt att bedöma (Drechsler 2011 s. 62).

Generellt sett kommer delvis syntetiska dataset behålla datastrukturer från urspungmaterialet bättre eftersom det enbart är vissa variabelvärden som imputeras.

5.1.3 Statistikproducentens perspektiv

Ur statistikproducentens synvinkel ställer de olika metoderna olika krav, beroende bland annat på hur det insamlade datamaterialet ser ut. Vid helt syntetiskt dataset är det svårt att skapa modeller som lyckas bevara simultanfördelningen för alla variabler, speciellt om det rör sig om komplexa undersökningar (Reiter 2008).

Vid skapandet av delvis syntetiska dataset är de bakomliggande modellerna enklare eftersom de inte omfattar lika många variabler. Men på grund av den generellt sett lägre säkerheten krävs det mer omfattande test av röjanderisker (Drechsler 2011, s.56). De kräver också att statistikproducenten utvärderar vilka variabler och för vilka uppgiftslämnare data ska syntetiseras (Reiter 2008).

Alla röjandekontrollmetoder kan inte appliceras på alla typer av data och få klarar av att dessutom behålla både variablers logiska restriktioner och deras simultanfördelning jämte varandra. Metoderna för syntetiska dataset är flexibla i den meningen att de går att applicera på en mängd olika datamaterial samtidigt som dess inbördes egenskaper bevaras:

- De kan baseras på datamaterial som består av både kontinuerliga och kategoriska variabler.
- De kan även skapas med hänsyn till variablers logiska restriktioner, om till exempel variabel Y alltid bör vara större än variabel Z så går detta att definiera i modellen.
- I undersökningar är det vanligt att frågeformulär innehåller instruktioner som anger att uppgiftslämnaren ska hoppa över delar av frågor (vanligtvis underkategorier till frågor). Genom att skapa fullt syntetiska dataset genom en sekventiell specifikationsmodell så går det att bevara dessa så kallade ”överhoppsmönster” i datasetet.

6 Slutsatser efter implementering i SAS

Att skapa ett delvis syntetiskt dataset i SAS är både enkelt och svårt samtidigt. Det hela beror på hur datasetet ser ut och hur många variabler som är nödvändiga att syntetisera. Vi har visat att det är relativt enkelt att skapa ett dataset där enbart en variabel, som dessutom korrelerar med flera andra variabler, ska syntetiseras. Vi skapade ett dataset som behöll sin tidigare struktur, vilket visades med ett högt I värde och liknande regressionskoefficienter som originaldatasetet. Även så kallade överhoppningssteg mellan frågor kan i SAS enkelt hanteras genom att dela upp datamaterialet i flera delar. Det var dock svårare att samtidigt lyckas att behålla variablers logiska förhållanden sinsemellan.

7 Jämförelse med andra röjningsmetoder

Valet av röjningsriskmetod för ett dataset bör baseras på vilken typ av data som det ska appliceras på. Syntetiska dataset passar sämre om få variabler korrelerar eftersom det då blir svårt att skapa bra imputeringsmodeller. I dessa fall fungerar metoder som t.ex. förgrovning, top- och bottomkodning och liknande bättre. Däremot är syntetiska dataset flexibla i den meningen att de kan skapas för både kontinuerliga och kategoriska variabler vilket är en egenskap många andra röjningsriskmetoder saknar. Om dataproducenten dessutom redan är bekant med multipel imputering från editeringsprocessen så är steget till att använda det även i röjningskontrollsprocessen inte stort. En nackdel med syntetiska dataset är annars att de är svårare att applicera än enklare metoder.

Den stora fördelen med syntetiska dataset är ändå att datasetets simultanfördelning till stor del behålls, ofta även till variabler utanför modellen. Många andra röjningskontroll-metoder behåller endast enklare univariata egenskaper. Eftersom syftet med publika dataset är att externa användare ska ha möjlighet att genomföra egna analyser direkt på mikrodata så väger denna aspekt tungt.

Röjningsriskkontroll är ofta en kompromiss mellan dataanvändbarhet och uppgifts-lämnarskydd, dessa egenskaper begränsas i sin tur av dataproducentens resurser i form av pengar, tid och kunskap. Syntetiska dataset klarar kompromissen bättre än många andra metoder. Anledningen till varför det inte används mer frekvent torde därför vara att det fortfarande är för resurskrävande.

Litteraturförteckning

Abowd, J. M. & Benedetto, B. & Stinson, M. H. (2013) The Creation and Use of the SIPP Synthetic Beta. *U.S. Census Bureau.*

http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf pp. 1-5, Hämtad 2015-01-03

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, New Jersey. Pp. 80-81,104,114,117,217

Björnstad, J.F. (2005) Non-Bayesian Multiple Imputation. *Journal of Official Statistics,* 23(4). pp. 433-452

Dale, A. & Elliot, (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics,* 14. pp 6-11

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. *Privacy in Statistical Databases*, pp. 227– 238. New York: Springer

Drechsler, J. (2010) Generating Multiply Imputed Synthetic Datasets: Theory and Implementation. *Dissertation*. Otto-Friedrich-Universitat Bamberg. pp.21,23

Drechsler, J. (2011) Synthetic datasets for statistical disclosure control. Springer. New York.pp,8-9,15,39-42,53-56,62-63

Hundepool, A. (2012). Statistical disclosure control. Chichester, West Sussex, United Kingdom: John Wiley & Sons Inc. pp.3,247-250,258

Geyer, C. J. Introduction to Markov Chain Monte Carlo. http://www.mcmchandbook.net/HandbookChapter1.pdf. Hämtad 2014-12-22

Israëls, A. & Kuyvenhoven, L. & Van der Laan, J. & Pannekoek, J. & Schulte Nordholt, E. (2011). *Imputation - Statistical Methods (201112),* Statistics Netherlands, The Hague/Heerlen. pp. 13-39

Karr, A. F. & Kohnen, C. N. & Oganian, A. & Reiter, J. P. & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60(3). pp. 224–232.

Kennickell, A. B. (2008) Multiple Imputation in the Survey of Consumer Finances. *Prepared for the August 1998 Joint Statistical Meetings, Dallas, TX*. pp. 11

Little, R. J. A. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics,* 18(4). pp. 407-426

Machanavajjhala, A. & Kifer, D. & Abowd, J.M. & Gehrke, J.& Vilhuber, L. (2008). Privacy: Theory Meets Practice on the Map, *International Conference on Data Engineering (ICDE),* pp. 277-286.

Missingdata (2014a)

http://missing-data.lshtm.ac.uk/index.php?option=com_content&view=article&id=75:missing-completely-at-random-mcar&catid=40:missingness-mechanisms&Itemid=96, Hämtad 2014-12-15

Missingdata (2014b)

http://missing-data.lshtm.ac.uk/index.php?option=com_content&view=article&id=76%3Amissing-at-random-mar&catid=40%3Amissingness-mechanisms&Itemid=96, Hämtad 2014-12-15

Pigott, T. D. (2001) A Review of Methods for Missing Data. *Educational Research and Evaluation,* 7(4). pp. 355-356

Raghunathan, T. E. & Reiter, J. P. & and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19(1), 1–16.

Reiter, J. P. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics,* 18(4). pp. 531–543

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* 29. 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. pp. 1-20.

Reiter, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Royal Statistical Society*, 168(1). pp. 185–205

Reiter, J. P. (2008). Protecting data confidentiality in public release datasets: Approaches based on multiple imputation, *The Imputation Bulletin*, 8(2). pp. 3-7

Reiter, J. P. & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* 1. pp. 99–110.

Rubin, D.B. (1976). Inference and Missing Data, *Biometrika*. 63(3) .pp.581-592.

Rubin, D.B. (1993). Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2), pp. 461-468

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434). pp. 473-489

SCB. *Statistisk röjandekontroll -av tabeller, databaser och kartor*. pp. 8-10

SCB. (2011). *Sveriges officiella statistik – årlig rapport för 2011*, SCB-tryck Örebro. pp. 7-10

SCB. (2014). http://www.scb.se/lamnauppgifter. hämtad 2014-11-25

Van Buuren, S. & Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*, TNO Prevention and Health, Leiden pp. 1-19

Bilaga A

```
*importerar dataset;
proc import datafile="E:\Dokument\Skola\Stockholms Universitet\HT
2014\Uppsats\Azerbaijan\AZE_2009_FLS_v01_M_STATA" out=AZE dbms = dta
replace;
run;

*namnger variablerna
(då ett par av uppgiftslämnarna angivit hushållets genomsnittliga
inkomst och utgifter i ett intervall så beräknas även genomsnittet av
dessa intervall);
data AZE;
set AZE;
q_67 = (q_63_inc_min+q_63_inc_max)/2;
q_68 = (q_63_exp_min+q_63_exp_max)/2;
LABEL     q_62 ="Inkomst från jobb"
          q_21 ="Egen bedömning av finansiellt kunnande"
          q_56 ="Kön"
          q_60 ="Utbildning"
          q_61 ="Arbetssektor"
          q_64 ="Antal personer i hushållet"
          q_67 ="Hushållets genomsnittliga inkomster"
          q_68 ="Hushållets genomsnittliga utgifter"
          q_58 ="Ålder"
   Sity_village="Geografisk indelning";
run;

*tilldelar observationerna identifikationsnummer.

data AZE;
set AZE;
IDNew=_n_;
run;

*ger en överblick av bortfallsmönstret;

Proc MI data=AZE nimpute=0;
run;


*då det partiella bortfallet är negligerbart tas dessa observationer
bort,
även uppgiftslämnare som ej tillhör arbetskraften tas bort

q_61 är arbetskategori (är kodat från 1-9 där 7 är studenter, 8 är
arbetssökande, 9 är ej arbetssökande utanför arbetskraften)
och q_62 är inkomstvariabeln (koden 99 anger att uppgiftslämnaren
vägrat svara på frågan);

data aze_rensad;
 set aze;
 if q_61 > 7 then delete;
 if q_62 = 99 then delete;
 if q_62 = . then delete;
run;
```

```
*undersöker om de numeriska variablerna är snedfördelade, vilket de
alla till viss del är;

proc univariate data=aze_rensad cibasic;
   var q_62 q_67 q_68;
   histogram q_62 q_67 q_68;
run;

*beräknar därför kvadratroten ur dessa;

Data aze_rensad;
set aze_rensad;
CRTq_62 = (q_62)**(1/3);
CRTq_67 = (q_67)**(1/3);
CRTq_68 = (q_68)**(1/3);
run;


*delar upp observationerna på geografiskt område;


data BY;
 set aze_rensad;
 if Sity_village = 1 then delete;
 if Sity_village = 2 then delete;
run;

data STAD;
 set aze_rensad;
 if Sity_village = 1 then delete;
 if Sity_village = 3 then delete;
run;

data BAKU;
 set aze_rensad;
 if Sity_village = 2 then delete;
 if Sity_village = 3 then delete;
run;


*(1) vi börjar här med de observationerna som bor i byar;


*duplicerar alla observationer;
data BYIMP;
set BY;
do SYNT = 1 to 2 by 1;
output BYIMP;
end;
run;

*samt tar bort variabelvärdet ”inkomst” för hälften;
data BYIMP;
 set BYIMP;
 if SYNT = 1 then do; CRTq_62=.; end;
run;
```

```
*dessa skattas sedan men hjälp av multipel imputering,
nimpute anger antalet dataset som ska skapas,
round anger att de imputerade värdena ska avrundas,
minmaxiter anger det maximala antalet iterationer,
class anger vilka variabler i modellen som är kategoriska,
fcs anger att en ”fully conditional specifikation” modell ska
användas,
inbiter anger det minimala antalet iterationer,
;

Proc MI data=BYIMP out=BYIMP nimpute=20 round=1 seed=14614 minimum=0
minmaxiter=500;
class q_21 q_56 q_60  q_61;
fcs nbiter=20
logistic(q_21 q_56 q_60  q_61)
reg(CRTq_67 CRTq_68 q_64);
var CRTq_62  CRTq_67 CRTq_68  q_21 q_56 q_60  q_61 q_64;
run;

*sedan raderas originaldata, enbart syntetiskt kvar;

 data BYIMP;
 set BYIMP;
 if SYNT > 1 then delete;
run;

*transformerar tillbaka inkomstvariabeln;
Data BYIMP;
set BYIMP ;
syntinkomst = (CRTq_62)**3;
run;


*(2)upprepar samma sak med de som bor i en stad;
data STADIMP;
set STAD;
do SYNT = 1 to 2 by 1;
output STADIMP;
end;
run;

data STADIMP;
 set STADIMP;
 if SYNT = 1 then do; CRTq_62=.; end;
run;

Proc MI data=STADIMP out=STADIMP nimpute=20 round=1 seed=99870
minimum=0 minmaxiter=500 ;
class q_21 q_56 q_60  q_61;
fcs nbiter=20
logistic(q_21 q_56 q_60  q_61)
reg(CRTq_67 CRTq_68 q_64);
var CRTq_62  CRTq_67 CRTq_68  q_21 q_56 q_60  q_61 q_64;
run;
```

```
 data STADIMP;
 set STADIMP;
 if SYNT > 1 then delete;
run;



Data STADIMP;
set STADIMP ;
syntinkomst = (CRTq_62)**3;
run;

proc univariate data=STADIMP cibasic;
   var syntinkomst;
   histogram syntinkomst;
run;
proc univariate data=STAD cibasic;
   var q_62;
   histogram q_62;
run;

*(3)upprepar samma sak för de som bor i Baku;
data BAKUIMP;
set BAKU;
do SYNT = 1 to 2 by 1;
output BAKUIMP;
end;
run;

data BAKUIMP;
 set BAKUIMP;
 if SYNT = 1 then do; CRTq_62=.; end;
run;

Proc   MI   data=BAKUIMP   out=BAKUIMP   nimpute=20   round=1   seed=22547
minimum=0 minmaxiter=500 ;
class q_21 q_56 q_60  q_61;
fcs nbiter=20
logistic(q_21 q_56 q_60  q_61)
reg(CRTq_67 CRTq_68 q_64);
var CRTq_62 CRTq_67 CRTq_68  q_21 q_56 q_60  q_61 q_64;
run;

data BAKUIMP;
 set BAKUIMP;
 if SYNT > 1 then delete;
run;

Data BAKUIMP;
set BAKUIMP ;
syntinkomst = (CRTq_62)**3;
run;

*sammanställer de tre (1)(2)(3)dataseten;
data AZEIMP;
set byimp bakuimp stadimp;
run;
```

```
proc univariate data=AZEIMP cibasic;
   var syntinkomst;
   histogram syntinkomst;
run;

proc univariate data=AZE_rensad
cibasic;
   var q_62;
   histogram q_62;
run;

*skiljer ut ett imputerat dataset;
data m_ettAZEIMP;
set AZEIMP;
if _imputation_ > 1 then delete;
run;

*ritar upp histogram och KI;
proc univariate data=m_ettAZEIMP
cibasic;
   var syntinkomst;
   histogram syntinkomst;
run;

*jämför olika regressionsmodeller;
proc reg data=m_ettAZEIMP;
   model syntinkomst=q_58;
run;

proc reg data=aze_rensad;
   model q_62=q_58;
run;

*jämför korrelationen;
proc corr data=Aze_rensad;
var Sity_village q_21 q_56 q_58 q_60 q_61 q_62 q_67 q_68 q_64;
run;
proc corr data=m_ettAZEIMP;
var Sity_village q_21 q_56 q_58 q_60 q_61 syntinkomst q_67 q_68 q_64;
run;



*************************
     ANVÄNDA SLUMPTAL

BY         STAD       BAKU
1 3564644 54657894 41056460
2 564654  1278712  897981
3 1270981 21657    123132
4 652324  570211   879012
5 947722  5770246  5605712
6 14614   99870    22547
7 878972  54067    587801
8 2125789 24707    0084567
9 089798  897087   5714405
10 100047 0879     87987
*************************;
```