NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 05 2017

ISSN: 1500-4066

March 2017

Discussion paper

Tre grupper skatteytere i søkelyset: Har de ulike kjennetegn?

BY

Jonas Andersson AND **Jostein Lillestøl**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Tre grupper skatteytere i søkelyset: Har de ulike kjennetegn?

av[1]

Jonas Andersson, Norges handelshøyskole
Jostein Lillestøl, Norges handelshøyskole

Sammendrag

Denne rapporten analyserer data for tre grupper av skatteytere som har vært i skatteetatens søkelys: Skatteytere som etter skatteamnesti har meldt seg frivillig med opplysninger om tidligere uoppgitt skattbar inntekt eller formue i utlandet, skatteytere der det foreligger automatisk kontrolloppgave utland (AKU) og skatteytere som er ilagt tilleggsskatt. Et knippe potensielle kjennetegn er valgt ut, og gruppene er sammenlignet innbyrdes og med en tilsvarende referansegruppe av vanlige skatteytere, med sikte på å avdekke forskjeller mellom gruppene. Tre ulike analysemetoder er prøvd ut: Enkel kategorisering, korrespondanseanalyse og klassifikasjonstrær. Disse er av eksplorativ karakter og egnet for grafisk presentasjon av resultatene. En sammenligning av resultatene, samt ulike fordeler og ulemper ved de tre metodene blir diskutert i forhold til det aktuelle ambisjonsnivå: Finne de kjennetegn som skiller gruppene eller også etablere regler til nytte for klassifisering av individer med ukjent tilhørighet.

Summary

This report analyzes data for three groups of taxpayers in scrutiny of the tax authorities: Taxpayers who, after a tax amnesty, have voluntary come forward with previous unreported taxable income or wealth abroad, taxpayers where there exist automatic abroad control scheme and taxpayers who have been assigned penalty tax. A number of potential key characteristics are selected, and the groups are compared against one another and with a control group of ordinary taxpayers, with the objective to uncover differences between the groups. Three different methods of analysis are used: Simple categorization, correspondence analysis and classification trees. They are of explorative nature and suitable for graphical presentation of results. A comparison of results, and some advantages and disadvantages of the three methods are discussed, in relation to the ambition level: Find the characteristics that separates the groups, or more, establish rules for classification of individuals with unknown membership.

[1] Denne rapport er utarbeidet som del av prosjekt ved Norsk Senter for Skatteforskning (NoCeT). Vi takker Bård Støve for bidrag med R-koding.

1 Innledning

Denne rapport er en oppfølger til SNF-Rapport 10/2012: «Kjennetegnanalyser av skatteytere som unndrar skatt ved å skjule formuer og inntekter i utlandet» (forf. Jonas Andersson, Jostein Lillestøl, Bård Støve). Der studerte man private skatteytere kjennetegnet ved at de i perioden 2009-2010 meldte seg frivillig med opplysninger om skatteunndragelser, dvs. skattbar inntekt eller formue i utlandet som ikke var rapportert i tidligere års selvangivelser. Dette var som følge av den norske skatteamnesti-ordningen, der skatteyter må betale den skatten som ordinært hadde blitt beregnet på formuen/inntekten, men slapp tilleggsskatt og/eller anmeldelse. Datamaterialet besto av anonymiserte opplysninger, i alt over 500 variable, fra selvangivelsene og enkelte andre kilder for de 577 «frivillig rettede» som til da hadde meldt seg, sammen med de samme opplysningene for en referansegruppe av «ordinære» skatteytere. Tilgjengelig som referansegruppe var 300 000 private skatteytere, dvs omtrent 10 % av skatteytermassen. Datagrunnlaget for denne rapporten er utvidet, ved at flere frivillig rettede er kommet til fra den etterfølgende perioden 2011-2013, slik at materialet nå omfatter 998, dvs. omlag 1000 slike. I tillegg er kommet to nye grupper som er interessante i forbindelse med unndragelse eller mulig unndragelse. Disse er omlag 5500 skatteytere der det foreligger automatisk kontrolloppgave utland (AKU) og omlag 30 000 skatteytere som er ilagt tilleggsskatt. Samtidig er noen andre opplysninger om hver skatteyter gjort tilgjengelig, mens mange uten interesse er fjernet. Dette utvidede material gjør det mulig å sammenligne de tre gruppene innbyrdes og opp mot referansegruppen. Dataene er anonymisert slik:

- Fødselsnummer er erstattet med et anonymt løpenummer
- Kommunenummer er erstattet med klassifikasjon iht. sentralitetskode
- Anonymisert statsborgerskap eller fødeland dersom der er 5 eller færre fra landet
- Alle beløp er avrundet til nærmeste 1000 kroner
- Skatteytere blant de 5 med høyest inntekt eller formue i ett eller flere år er fjernet

Våre dataanalyser er, dersom ikke annet er sagt, utført på registreringene fra selvangivelsen for året 2012. Dette innebærer en del forhold som en bør være oppmerksom på ved tolkning av resultatene. Skatteytere i gruppen av de frivillig rettede er kommet til etter hvert dette året og i årene før, og tallene er derfor ikke nødvendigvis representative for situasjonen på det tidspunkt skatteyter (eller arvinger) fant det for godt å melde seg til skattemyndighetene.

I likhet med Andersson et. al. (2012, s. 11) baserer vi valget av variabler på tidligere studier, spesielt verdt å notere er Clotfelter (1983) and Feinstein (1991).

Notatet er i tre hovedkapitler: Beskrivende statistikk (kapittel2), korrespondanseanalyse (kapittel 3) og klassifikasjonstrær (kapittel 4). For alle tre delene gjelder at resultatene i størst mulig grad blir presentert grafisk, der forskjeller mellom de fire gruppene data trer fram.

Kapittel 2 om beskrivende statistikk består i stor utstrekning av tabellering av utvalgte kategoriske «forklarende» variable, med de relative hyppigheter for variabelkategoriene innen hver av de fire gruppene, og der resultatene presenteres i form av sammenstilte søylediagrammer. Denne form for enkel tabellering kan skjule viktig informasjon, ved at det kan være spesielle kombinasjoner av forklarende variabler som gir de mest påfallende forskjeller mellom gruppene. Et steg videre vil kunne være å lage to-veis hyppighetstabeller for utvalgte par av kategorivariable for hver av de fire gruppene, og så foreta sammenligninger. I praksis kunne dette gjøres grafisk med romlige sammenstilte søyler. Med mange aktuelle forklarende variable blir det mange tabeller og grafer, og det er lett å miste oversikten, spesielt dersom det er mange kategorier for en eller flere variabler. For rent eksplorative formål kan såkalt korrespondanseanalyse være et alternativ eller et supplement, med beregninger som er velegnet for grafisk presentasjon og fortolkning. Dette er tema i kapittel 3. Her blir et knippe kategoriske variable studert i sammenheng, og hoveddimensjonene i datamaterialet avdekket. Variabelkategoriene og gruppekategoriene representeres ved skårer som kan tegnes inn i et «kart», der retning og avstand gir uttrykk for assosiasjoner. For å bruke korrespondanseanalyse må de numeriske variable knyttet til inntekt, formue og gjeld kategoriseres (ingen/lav, middels, høy).

Klassifikasjonstre (CART) er tema i kapittel 4. Som navnet sier tar metoden sikte på klassifikasjon, i vårt tilfelle å skille en spesiell gruppe ut fra datamaterialet på grunnlag av de tilgjengelige kategoriske og numeriske «forklarende» variable. Dette skjer ved en trinnvis binær splitting av observasjonsenhetene i grupper, der de spesielle enhetene fremstår klarere i den ene gruppen enn den andre. Dette organiseres i en trestruktur, som fremstår som beslutningsregler for å lokalisere undergrupper der de spesielle er i overvekt eller i hvert fall sterkt representert. Et klassifikasjonstre tillater blanding av numeriske og kategoriske forklarende variable, og er lett å tolke både grafisk og numerisk. I dette kapitlet tar vi etter tur for oss hver av de tre spesielle gruppene (frivillig retting, AKU, tilleggsskatt), holdt opp mot en referansegruppe. Med dette får en avdekket om det er de samme eller ulike variabler som sannsynliggjør hver av de spesielle gruppene.

For vårt formål er det ikke nødvendig å benytte alle 300 000 tilgjengelige referansepersoner. Det er heller ikke hensiktsmessig, av datatekniske årsaker og et ønske om å reservere en del av materialet for bekreftende analyser. Ved studiet av hver av de tre spesielle gruppene har vi valgt ut like mange referansepersoner, og lagt til 100 000 referansepersoner, slik at vi samlet har en populasjon der de spesielle utgjør et mindretall (dog ikke i henhold til deres andel av alle norske personlige skatteytere).

2 Beskrivende statistikk

Betrakt utvalgte variable for de fire gruppene, herunder spesielt sosio-demografiske variable og økonomiske variable knyttet til inntekt, formue og gjeld.

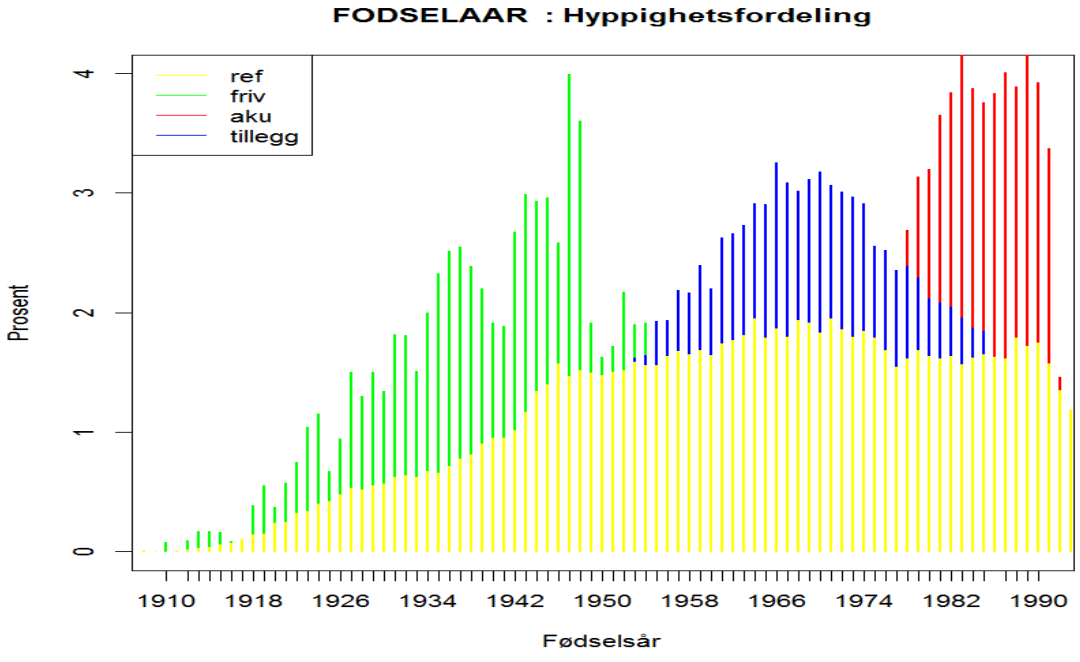
Variabler i databasen er av to typer:

Kategorivariabler: FODSELAAR, SIVILSTAND, SKM_GRUPPE, PERSON_KODE + andre koder

Numeriske variable: PERSON_INNTEKT_LONN, NETTO_FORM_STAT, GJELD_IB_OG_UB osv.

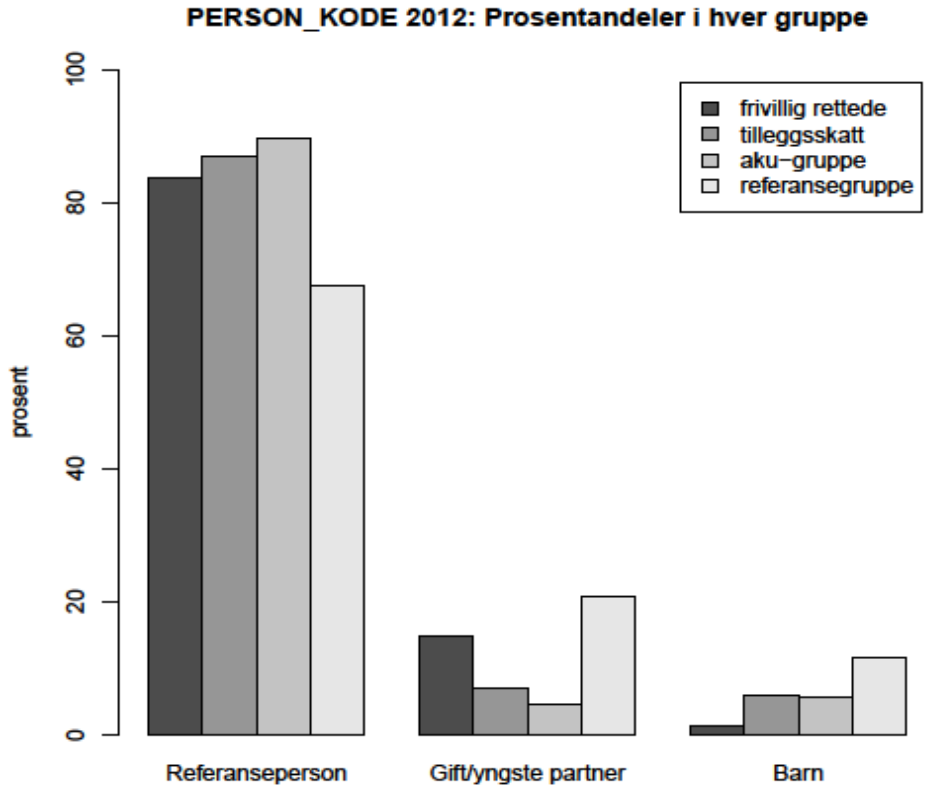
Her betyr IB innenbygds og UB utenbygds. Databasen inneholder flere ulike inntekts- og formue- og gjeldsvariable, som refererer seg direkte til selvangivelsens poster. Eksempelvis svarer IB_ALM_INNT_E_SERF til post 3.6. Alminnelig inntekt etter særfradrag innenbygds, som er grunnlaget for beregning av kommune-, fylkes- og fellesskatt. Der det foreligger alternative variable, spiller det for vårt formål liten rolle hvilken som velges.

Figur 1 viser hyppighetsfordelingen for de fire gruppene for variabelen FODSELAAR. Vi ser tydelige forskjeller mellom de fire gruppene. Mens referansegruppen (gul) fordeler seg over et bredt spekter av aldre, finner vi at de frivillig rettede (grønn) har sitt tyngdepunkt i de eldre aldersklasser, og AKU-gruppen (rød) sitt tyngdepunkt i de yngre aldersklasser, mens de med tilleggsskatt (blå) er i midten. At det er flere eldre blant de frivillig rettede er ikke overraskende, siden vi her trolig finner et betydelig antall avdøde, med uoppgitte utenlandsformuer, som er blitt ordnet opp i av arvinger.



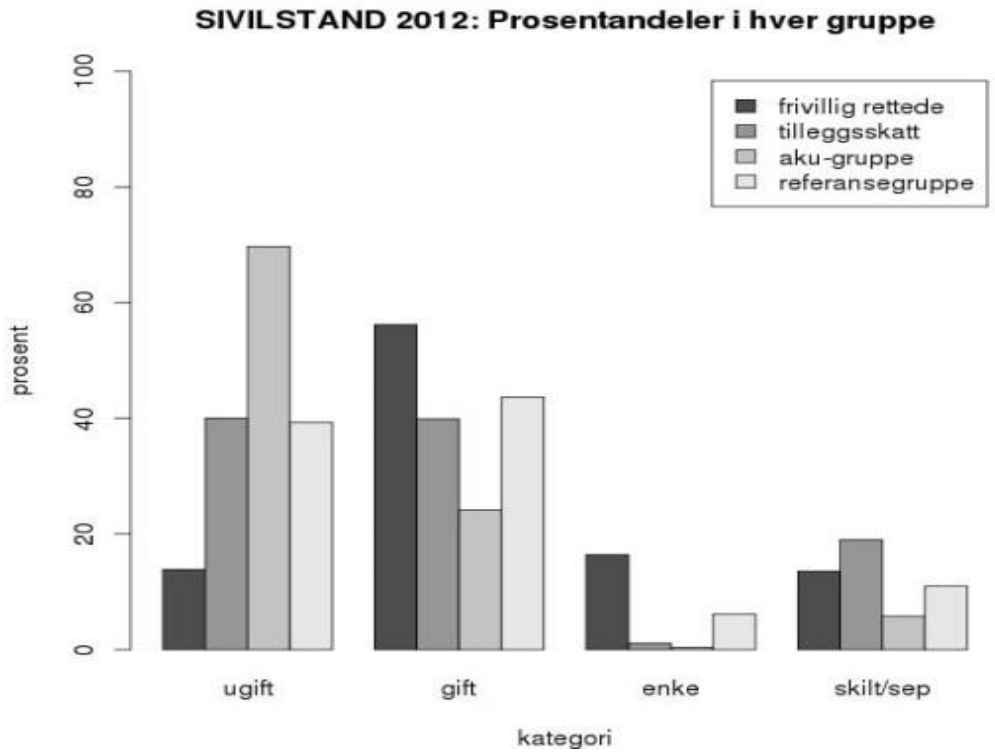
Figur 1: Fordeling av fødselsår for de fire gruppene.

Figur 2 viser hyppighetsfordelingen for kategorivariabelen PERSONKODE, det vil si hvorvidt ligningen dreier seg om referanseperson, gift/yngre partner eller barn til en referanseperson. Vi ser at i de tre gruppene (frivillig retting, AKU, tillegg) dreier seg oftere om referansepersonen enn i kontrollgruppen. Mens prosentandelen av registrerte referansepersoner i kontrollgruppen er om lag 70%, er den over 80% i de tre andre gruppene, og høyest for AKU-gruppen med om lag 90%. Figur 3 viser hyppighetsfordelingen for den beslektede variabelen SIVILSTAND. Vi ser at blant de frivillig rettede er det minst andel ugifte og høyest andel gifte.[2] Det er i AKU-gruppen vi finner den høyeste andelen ugifte. For øvrig merker vi oss at det i den frivillig rettede gruppen er en betydelig andel enker.



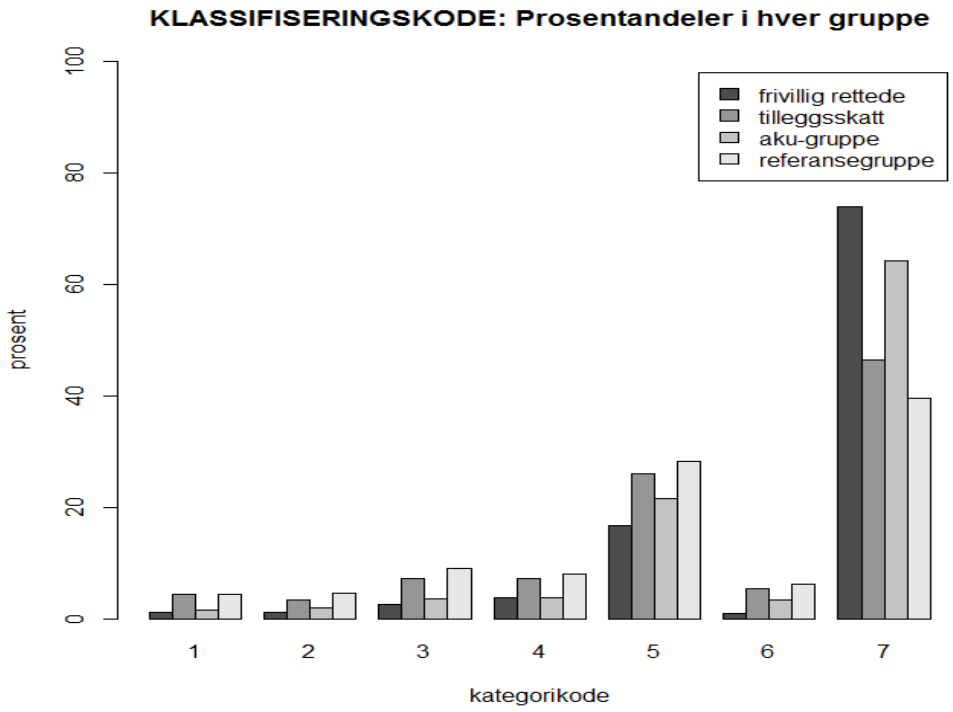
Figur 2: Fordeling av personkode for de fire gruppene.

[2] SIVILSTAND har partnerskap (evt. gjenlevende eller avbrutt) som egne kategorier (totalt 9 kategorier). Disse andelene er små og er her kategorisert sammen tilsvarende ekteskapsgruppe (gift, enke, skilt/separert).



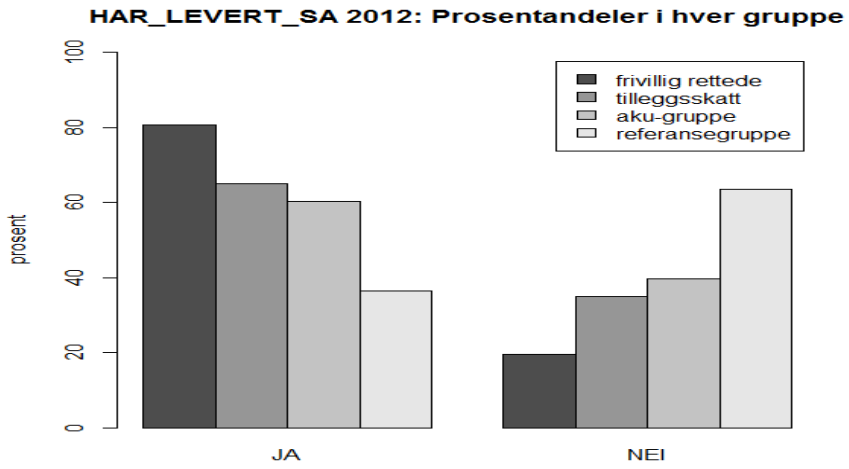
Figur 3: Fordeling av sivilstand for de fire gruppene.

Forskjeller mellom gruppene med omsyn til type aktivitet, angitt ved KLASSIFISERINGSKODE, er vist i Figur 4. Vi ser at den største andelen i gruppen frivillig rettede er sentral tjenesteyting. Her finner vi også en stor andel i AKU-gruppen. En annen variabel SENTRALITETSKODE grupperer sammen til de fire kategoriene (usentral, mindre sentral, noe sentral, sentral), og gir et tilsvarende bilde at den største andelen blant de frivillig rettede hører til sentralt, men det er ikke store forskjeller mellom gruppene.

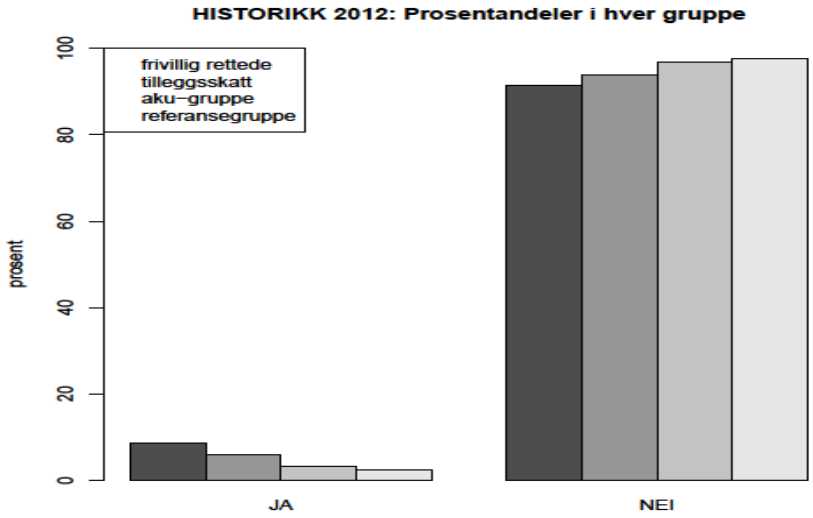


Figur 4: Fordeling av klassifiseringskode for de fire gruppene. 1=primærnæring, 2=landbruk og industri, 3=industri, 4=tjenesteyting + industri (mindre sentralt), 5=tjenesteyting + industri (sentralt), 6=tjenesteyting (mindre sentralt), 7=tjenesteyting (sentralt).

Figur 5 viser om skatteyter har levert selvangivelse (JA) eller ikke (NEI). Vi ser at det er størst andel for de frivillig rettede (80%) og minst i referansegruppen (under 40%). Det er vel ikke oppsiktsvekkende, etter at skatteyter er kommet i søkelyset på et tidspunkt og dermed må/bør passe seg. Figur 6 viser om skatteyter er registrert med en historikk (JA) eller ikke (NEI). Her er det ikke store forskjeller mellom gruppene, men vi registrerer at de frivillig rettede ligger høyest med omtrent 9%, mens kontrollgruppen har 3% med en historikk.

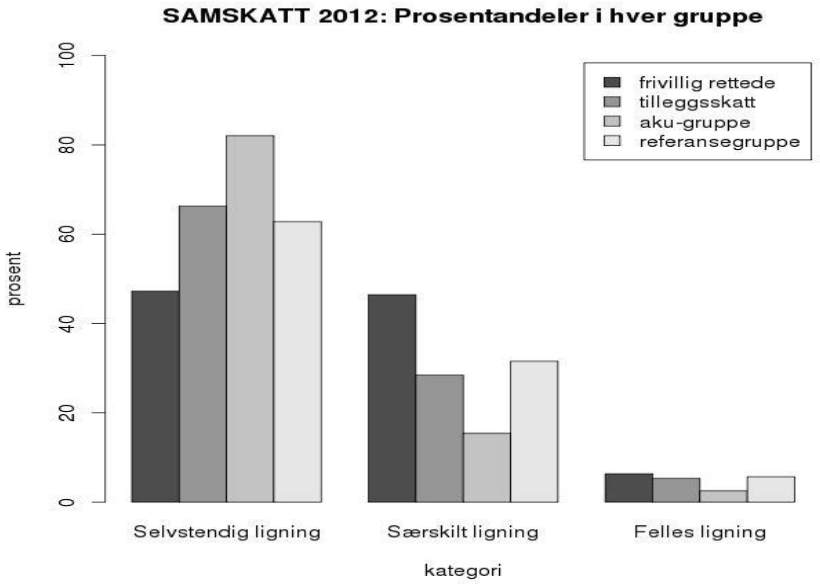


Figur 5: Andel som levert selvangivelse for hver av de fire gruppene.



Figur 6: Andel med historikk for hver av de fire gruppene.

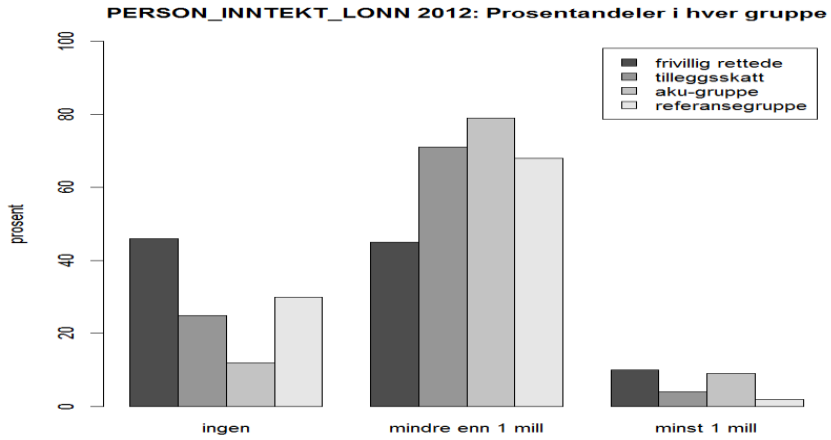
Variabelen SAMSKATT i Figur 7 angir ligningstype (Selvstendig/enslig, Særskilt, Felles). Vi ser at de frivillig rettede fordeler seg med omtrent like stor andel i gruppen med selvstendig (enslig) og særskilt ligning, mens det i AKU-gruppen er langt større andel i selvstendig kategorien.



Figur 7: Andel selvstendig og fellesligning for hver av de fire gruppene.

Vi skal så se nærmere på de økonomiske variablene inntekt/lønn, formue og gjeld.

Den numeriske variabelen PERSON_INNTEKT_LONN er gruppert i tre kategorier, og andelene i hver kategori for hver av de fire gruppene (frivillig rettede, tilleggsskatt, AKU-gruppe, referansegruppe) er gitt i Figur 8. For alle fire grupper gjelder at inntekt/lønn for den store majoritet er mindre enn 500 000, og at den for de frivillig rettede er enda lavere. Det er ellers ikke store forskjeller mellom de fire gruppene, unntatt for null-inntekter og svært lave inntekter.



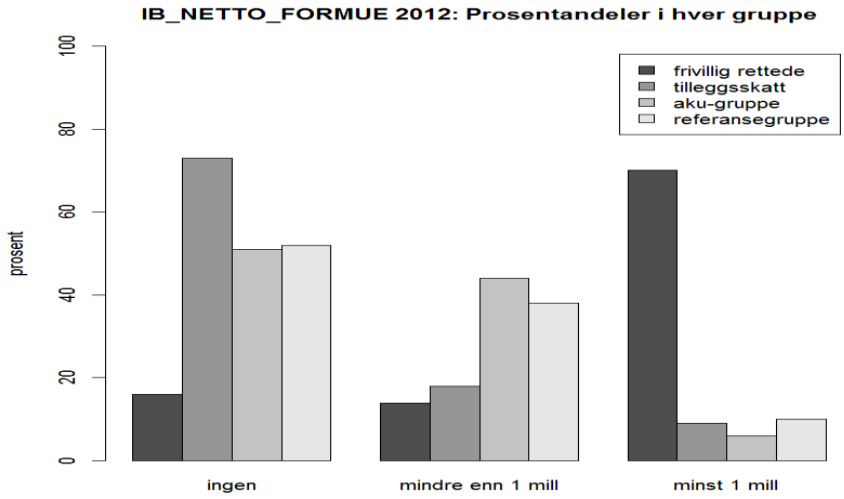
Figur 8: Andel med 0, <1 mill og >1mill NOK per år i lønnsinntekt for hver av de fire gruppene.

Forskjellene er illustrert i Tabell 1. Vi ser at hele 46% av skatteyterne blant de frivillig rettede er uten inntekt/lønn.

Inntekt/Lønn	Frivillig rettet	Tillegg	AKU	Referanse
=0	46%	25%	12%	30%
<100 000	56%	38%	27%	43%
< 250 000	62%	53%	43%	55%
< 500 000	75%	80%	73%	84%
< 1 000 000	90%	96%	91%	98%

Tabell 1: Lønnsfordeling for de fire gruppene.

Figur 9 viser prosentandeler i hver av de fire gruppene for variabelen IB_NETTO_FORMUE etter gruppering i tre kategorier. Vi ser at gruppen med frivillig rettede har betydelig større andel med større netto formuer, mens blant de med tilleggsskatt er det en stor andel med ingen rapportert formue. Her skiller AKU-gruppen seg lite fra referansegruppen.



Figur 9: Fordeling av innenbygds nettoformue for de fire gruppene.

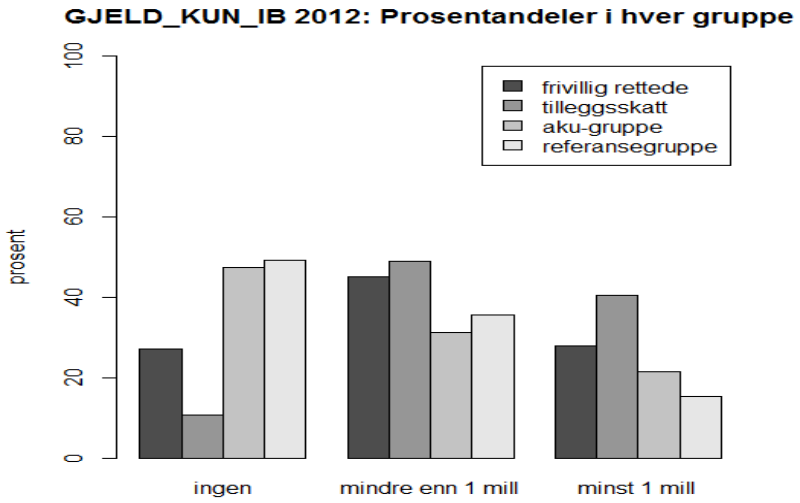
Flere detaljer kan leses ut av følgende tabell (basert på en den alternative formue-variabelen NETTO_FORM_STAT).

Formue	Frivillig	Tilleggskatt	AKU-gruppe	Referanse
=0	16%	72%	50%	52%
>1 mill	72%	10%	6%	10%
>5 mill	36%	3%	1%	0.85%
>10 mill	21%	1.5%	0.8%	0.3%

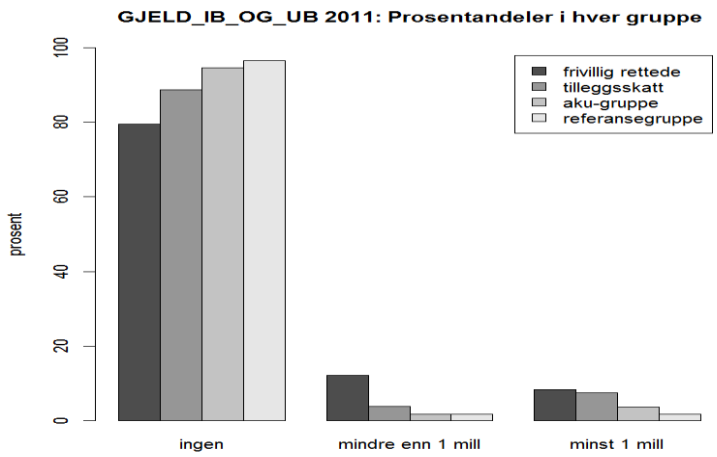
Tabell 2: Fordeling av formue for de fire gruppene.

Her har 6% blant de frivillig rettede formue over 50 mill. og 3% formue over 100 mill.kr.

Figur 10 og Figur 11 viser litt om gjeldsforholdene. For gjeld til Norge (IB) alene i Figur 10 ser vi at de fire gruppene er det omlag like store andeler som har gjeld mindre enn 1 mill. kroner. For de frivillig rettede og AKU-gruppen er det det større andel med gjeld minst 1 mill. kroner, og tilsvarende mindre andel uten gjeld. For variabelen som registrerer gjeld for de som har både innenlandsk og utenlandsk gjeld (IB og UB) er det ikke store forskjeller mellom gruppene. Det store flertall har ingen registrert gjeld som gjør at de faller i denne kategorien. For de de fire gruppene ser vi at andelen med stor gjeld er størst for de frivillig rettede og de med tilleggsskatt, og minst for referansegruppen.



Figur 10: Fordeling av innenlands gjeld for de fire gruppene.



Figur 11: Fordeling av all gjeld for de fire gruppene.

En kan spørre seg hvilke variabler (om noen) som er egnet til å skille ut en spesiell gruppe, frivillig retting, tilleggsskatt eller AKU, fra referansegruppen. Kunnskap om dette kan være nyttig for Skatteetaten ved prioritering av ressurser til kontroller. Dette er et klassifikasjonsproblem, der visse kategorier for variablene, mer enn andre kategorier, gjør en bestemt gruppetilhørighet mer sannsynlig. Merk at ovenstående beskrivende analyse på sett og vis besvarer det motsatte spørsmål: For gitt gruppetilhørighet, hvor forskjellig er forekomsten av kategoriene for hver av variablene?

En sammenligning av strukturen i hvert av de sammenstilte søylediagrammene ovenfor, kan gi en viss pekepinn på hvilke variabler som kan være nyttige fremfor andre til klassifikasjon av gruppetilhørighet. Det er imidlertid ikke lett. En forholdsvis enkel alternativ mulighet er å beregne et empirisk mål for samvariasjon (også kalt assosiasjon) i hver av de to-veis hyppighetstabellene, der en sammenligner hver av de spesielle gruppene med referansegruppen. For hver variabel blir det tre 2×k tabeller, der k er antall kategorier for den aktuelle variabel. Cramer's V er et slikt mål for samvariasjon, basert på kjikvadrat-observatoren for testing av uavhengighet i to-veis tabeller. Definisjonen er (for tilfellet med 2×k-tabeller)

$$V = \sqrt{\frac{\chi^2}{n}}$$

der n er det totale antall observasjoner i tabellen. V er et tall mellom 0 og 1, slik at liten verdi svarer til liten samvariasjon og høy verdi til stor samvariasjon. Samvariasjon betyr her avvik fra uavhengighet, at hyppighetsfordelingen i de to grupper er forskjellige, og at en dermed har mulighet for å utnytte den informasjon som ligger i observert kategori for variabelen for et individ med ukjent gruppetilhørighet til å klassifisere dette til den spesielle gruppen eller ikke.

Følgende rettesnor blir ofte brukt for å karakterisere størrelsen på samvariasjonen:

Neglisjerbar:	0.01 – 0.09
Liten:	0.10 – 0.29*
Moderat:	0.30 – 0.49**
Stor:	0.50 – 0.69***
Svært stor:	0.70 – 0.99****

Tabell 3 gir Cramer's V for de kategoriske variablene og de kategoriserte numeriske variable ovenfor, for hver av de tre spesielle gruppene holdt opp mot referansegruppen. I tillegg kommer variabelen SKM_GRUPPE (skjemagruppe) som vil komme inn i senere kapittel. I tabellen er ikke-neglisjerbar samvariasjon markert med * for liten samvariasjon og ** for moderat samvariasjon. Sterkere samvariasjon enn dette, markerte med ytterligere stjerner, forekommer ikke.

Variabel \ Gruppe → ↓	Frivillig rettede n_1=998	Tilleggsskatt n_2=29400	AKU n_3=5500
PERSON_KODE	0.038	0.181*	0.106*
SIVILSTAND	0.060	0.136*	0.136*
KLASSIFISERINGSKODE	0.069	0.061	0.112*
HAR_LEVERT_SA	0.087	0.236*	0.105*
HISTORIKK_KODE	0.014	0.069	0.033
SAMSKATT_KODE	0.091	0.351**	0.122*
PERSON_INNTEKT_LONN	0.069	0.071	0.131*
IB_NETTO_FORMUE	0.189*	0.178*	0.037
GJELD_KUN_IB	0.046	0.345**	0.037
GJELD_IB_OG_UB	0.036	0.161*	0.081
SKM_GRUPPE	0.039	0.078	0.326**
	Referansegruppe n_0=100 000+		

Tabell 3: Mål for samvariasjon Cramer's V

Vi vil bruke skjemaet ovenfor til å fortolke nytteverdien av de enkelte variable for å kunne si noe om en bestemt spesiell gruppetilhørighet. Det mest påfallende i Tabell 3 er mangelen på ikke-neglisjerbare variable (pekere) for gruppen Frivillig rettede. Her er det bare IB_NETTO_FORMUE som byr seg fram. Mer lovende er gruppen Tilleggsskatt, der hele syv av de ti variabler byr seg fram, og to av variablene med V-verdi over 0.30, som indikerer moderat samvariasjon, nemlig SAMSKATT_KODE og GJELD_KUN_IB. For AKU-gruppen er det seks av de ti variablene som byr seg fram, men ingen med mer enn lav samvariasjon. En variabel synes å være neglisjerbar som peker uansett gruppen, nemlig HISTORIKK_KODE.

Det er interessant å se om de holdepunkter vi får her harmonerer med de sammenstilte søylediagrammene ovenfor. Eksempelvis, se på de to variablene med høyest V for gruppen Tilleggsskatt. Vi ser at for GJELD_KUN_IB vil ingen gjeld typisk kunne «frita» skatteyteren for å tilhøre gruppen med tilleggsskatt, mens betydelig gjeld ikke gjør det. Når det gjelder SAMSKATT_KODE er det vanskelig å forstå hvorfor den skulle bli pekt ut klarere enn andre variabler.

Dersom formålet er å peke ut variable med potensiale for prioritering av kontrollinnsats, er det en rekke forhold en bør ta i betraktning ved eventuell bruk av ovenstående analyse basert på samvariasjonsmål.

(i) Ved å ta for seg variablene en for en kan en overse at visse kategori-kombinasjoner knyttet til flere variabler kan være gode pekere, kanskje bedre. Forøvrig må vi innse at, selv om vi får utpekt variabler som kan være lovende for klassifikasjonsformål, så gis ikke klassifikasjonsregelen uten videre analyse. Dette belyses i senere kapittel.

(ii) Den underliggende kjikvadrat-test gir statistisk signifikans for alle kombinasjonene (variabel, gruppe) i Tabell 3. Med så mange observasjoner som vi har er dette ikke uventet, siden en da fanger opp selv det minste avvik fra uavhengighet. Dette kan derfor ikke legges til grunn for vurdering av potensialet. Cramer's V med sin n i nevner under rottegnet gir den rimelige nedskalering av kjikvadrat-verdien.

(iii) Cramer's V frigjør oss samtidig fra at det er vesensforskjellige antall individer i de fire gruppene. Merk at $n= n_0 + n_i$ for i=1,2,3 for de tre gruppene. Denne marginalisering kan tilsløre det potensiale de enkelte variable har for å predikere gruppetilhørighet. Det er klart at det i en populasjon med ukjent andeler av spesielle grupper iblandet de vanlige, vil individer i en liten spesiell gruppe ha mindre sjanse for å bli pekt ut enn individer fra en større gruppe, selv styrken på de utvalgte pekere er den samme. Gruppen Frivillig rettede representerer her den største utfordringen på to måter: Den har svake pekere og er liten i forhold til populasjonen av vanlige skatteytere.

(iv) Et samvariasjonsmål som Cramer's V er symmetrisk, dvs. tar ikke omsyn til hvilken retning en eventuell prediksjon skal gjøres. Et beskrivende mål for to-veis tabeller kalt Lambda (Goodman & Kruskal) tar omsyn til retningen. Dette er definert som den relative reduksjon i antall prediksjonsfeil uten bruk av den gitte variabel til å gjøre bruk av den. Implisitt her ligger at prediksjonen er den antatt mest sannsynlige kategori, her gruppe. Resultatet vil derfor avhenge av hvor stor referansegruppen er. Bruk av Lambda, eller andre mål med retning med tilsvarende egenskap, vil derfor være uegnet i vår kontekst, da gruppen av «vanlige» skatteytere er så stor at «vanlig» typisk vil være det mest sannsynlige uansett hva de utvalgte pekere ellers sier.

3 Korrespondanseanalyse

I utgangspunktet er korrespondanseanalyse en dekomposisjon av en to-veis hyppighetstabell for kategoriske data, der resultatet er egnet for grafisk presentasjon som punkter i Euklidske rom av lav dimensjon. Multippel korrespondanseanalyse (MCA) er en utvidelse til mer enn to kategoriske variable (fler-veis tabeller), og kan betraktes som et motstykke til prinsipalkomponentanalyse for numeriske data. I dette tilfellet analyseres den såkalte indikatormatrisen, der kolonnene utgjør blokker av 0-1 variable, en blokk for hver kategorisk variabel, med egne kolonner for hver kategori innen hver blokk. Radene i indikatormatrisen er individene.[3] I situasjoner der også numeriske variable inngår, kan disse med fordel kategoriseres. Selv om dette gir en ordinal kategorisering, blir dette ikke tatt omsyn til i algoritmen, da en ikke kan anta at en eventuell avhengighet med andre variable er ordinal.

Som resultat av dekomposisjonen får vi at både variabelkategorier og individer kan plasseres inn i to-dimensjonale «kart», som kan gi grunnlag for fortolkninger. Sentralt i algoritmen som leder til løsninger er søken etter forskjeller, enten det er forskjeller i hyppighet for kategoriene for den enkelte variabel eller avvik fra uavhengighet mellom variabler. De funne avvik blir representert numerisk med skårer i ortogonale dimensjoner, med fallende grad av betydning (rangert etter egenverdier). Plott i de to mest betydningsfulle dimensjonene er som regel tilstrekkelig, men muligheten for å studere flere dimensjoner er til stede.

Vi kan typisk lage tre ulike MCA-kart: Variabelkart, kategorikart og individkart. Alle tre gir et kart, der de to aksene representerer to hoveddimensjoner i datamaterialet. Tolkning av de ulike kartene er i hovedsak som følger:

Individkart: Nærhet mellom to individer i plottet betyr at de tenderer mot å tilhøre samme kategori på de kategoriske variabler som inngår i MCA-spesifikasjonen.

Kategorikart: Nærhet mellom kategorier for ulike variable betyr at disse tenderer mot å være felles for mange individer, mens nærhet mellom kategorier for samme variable betyr at grupper av individer i disse kategoriene ligner hverandre (og kategoriene kan med fordel slås sammen).

Variabelkart: Her vises den korrelasjon hver variabel har med hver av de to hoveddimensjonene av datamatrisen.

Vi vil her illustrere metoden i en situasjon, der vi som utgangspunkt tar ti utvalgte variable fra

[3] Merk at hver kategori har sin kolonne, og derfor vil en rad alltid ha en 1-er og resten 0-er i hver blokk. Alle linjesummer er derfor lik antall variabler.

vår database, beskrevet i Tabell 4.[4] Disse omfatter syv kategoriske variable, hvorav en er gruppetilhørighet (Frivillig retting, AKU-gruppe, Tilleggsskatt, Kontrollgruppe), med tillegg av tre variabler som representerer henholdsvis formue, inntekt og gjeld, her kategorisert i tre kategorier (0, opptil 1mill. NOK, over 1 mill. NOK).[5]

Variabel	Kortnavn	Kategorier
SKM_GRUPPE	SKM	10-Full skatteplikt, bosatt 13-Utvandret 14-Midlertidig opphold 20-Stedbunden skatteplikt, bosatt 30-Stedbunden skatteplikt, ikke bosatt 40-Dødsbo 70-Diplomat + noen utelatte
PERSON_KODE	PERS	1=Referanseperson, 2=Gift/yngste partner, 3=barn
HISTORIKK_KODE	HIST	J=Ja, N=Nei
HAR_LEVERT_SA	LEV	J=JA, N=Nei
SENTRALITETSKODE	SENT	0=Usentralt, 1=Mindre sentralt, 2=Noe sentralt, 3=Sentralt
SAMSKATT_KODE	SAM	1=Enslig, 2=Særskilt, 3=Felles
IB_NETTO_FORMUE	FORMUE	0, 1=opptil 1mill, 2= over 1 mill
IB_ALM_INNT_E_SERF	INNTEKT	0, 1=opptil 1mill, 2= over 1 mill
GJELD_IB_OG_UB	GJELD	0, 1=opptil 1mill, 2= over 1 mill
Gruppetilhørighet	GRUPPE	FRIV, AKU, TILL, KONT

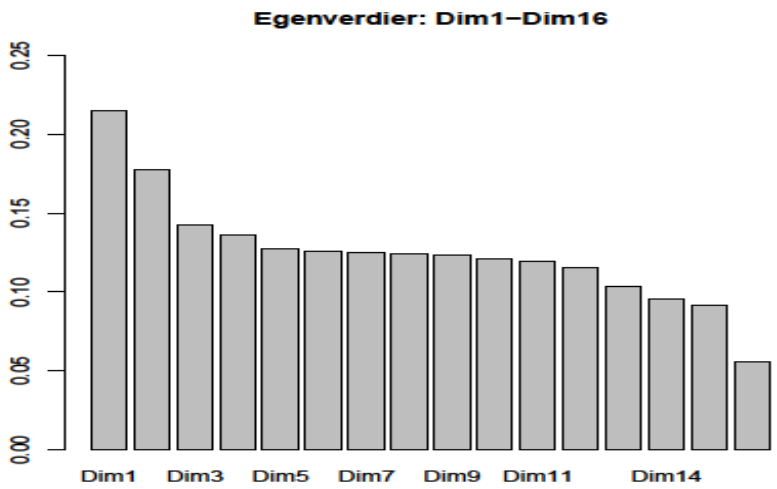
Tabell 4: Variabelbeskrivelser.

Dersom man ønsker det, kan en eller flere variabler pekes ut som supplerende. En slik blir ikke med i beregningen av løsningen, men dens kategorier kan skårberegnes etterpå, og plottes inn i kartene som supplerende punkter i kartet. Supplerende variable innlemmes ofte for å bidra til fortolkningen av hoveddimensjonene, men kan også innlemmes for å avdekke hvordan ulike grupper plasserer seg i et eksisterende kart, laget uten kjennskap til denne gruppetilhørigheten. I vårt tilfelle er det aktuelle å la GRUPPE være supplerende variabel.

Datamaterialet for vår analyse omfatter antall observasjoner av størrelsesorden (før fradrag av manglende registrering for enkelte variable): Frivillig rettede (1000), AKU-gruppe (5500), Tilleggsskatt (30 000), Kontrollgruppe (100 000).

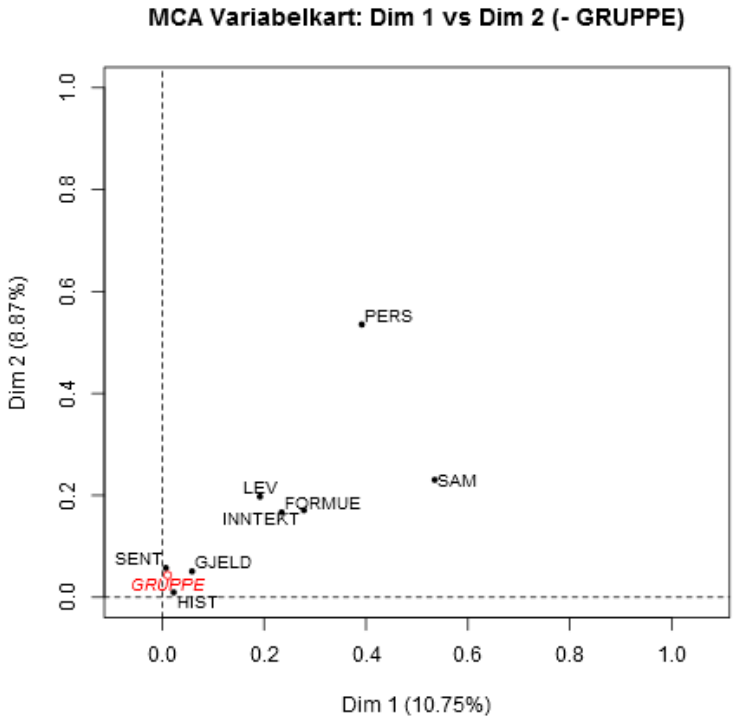
Nedenfor følger MCA-analyse utført med R-programvaren FactoMineR. Siden kategorikartet lett blir altfor detaljert og uoversiktlig har vi i utgangspunktet utelatt variabelen SKM, uten at dette vil spille noen rolle for det store bildet. GRUPPE er spesifisert som supplerende variabel.

[4] Disse variable er tidligere funnet å ha en viss prediktiv verdi ved CART-analyse for gruppen Frivillig retting.
[5] En annen mulighet er å studere hver av de tre gruppene (Frivillig, AKU, Tillegg) mot en like stor kontrollgruppe. En må da ha en egen variabel for hver av disse gruppene (binær med verdier f.eks. 0/1 eller FALSE/TRUE).

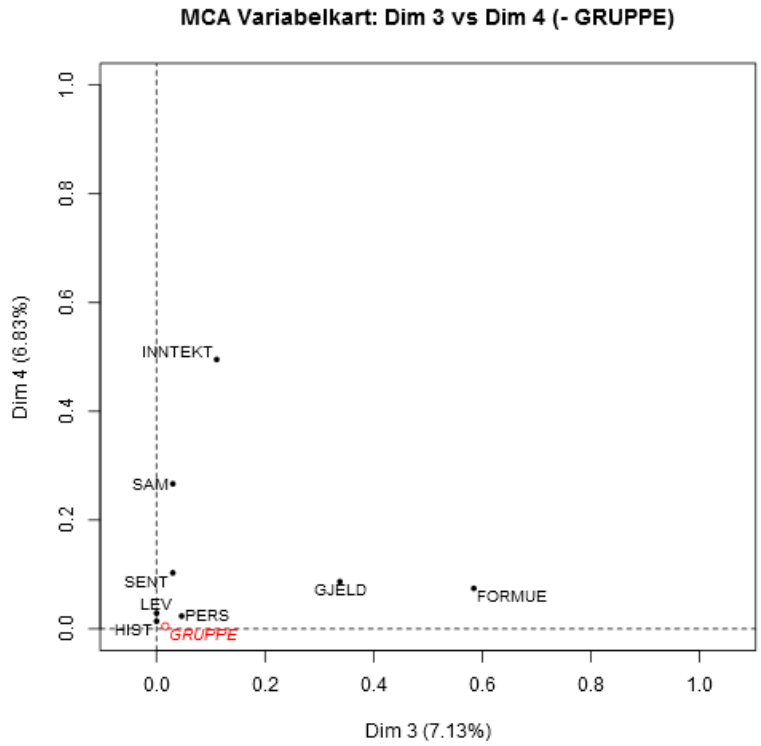


Figur 12: Egenverdiplott fra korrespondanseanalyse.

Egenverdiplottet i Figur 12 viser to framtredende dimensjoner Dim 1 og Dim 2, og det er trolig ikke mer nyttig informasjon å hente utover disse. Variabelkartet i Figur 13 viser Dim 1 vs Dim 2, mens vi for illustrasjonens skyld også tar med variabelkartet for Dim 3 vs Dim4 i Figur 14.

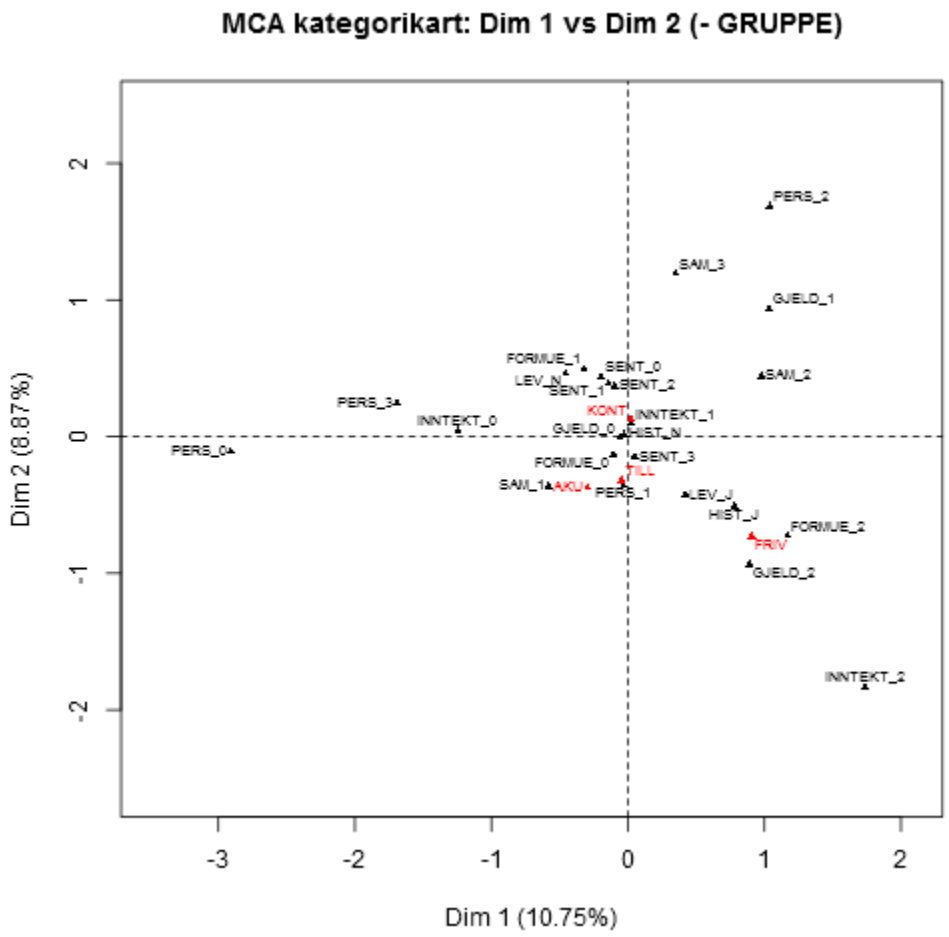


Figur 13: Variabelkart fra korrespondanseanalyse.



Figur 14: Variabelkart fra korrespondanseanalyse.

På variabelkartet for Dim 1 vs Dim 2 i Figur 13 ser vi at SAM korrelerer mest med Dim 1, mens PERS korrelerer bra med både Dim 1 og Dim 2. I noen grad gjør LEV, FORMUE og INNTEKT også det, mens for GJELD er korrelasjonen svak, og for SENT og HIST ubetydelig. På variabelkartet for Dim 3 vs Dim 4 i Figur 14, ser vi at Dim 3 representer GJELD og FORMUE, mens Dim 4 representerer i hovedsak INNTEKT. Vi ser at HIST ikke framtrer i noen av disse dimensjonene. At den supplerende variabelen GRUPPE plasserer seg med ubetydelig korrelasjon med alle fire dimensjonene er ikke overaskende, siden punktet representerer et «gjennomsnitt» av individene i hele tallmaterialet, uten at variabelen selv har vært bestemmende for dimensjonene. Håpet er imidlertid at de fire kategoriene for GRUPPE vil framtre i et kategorikart. Her følger i Figur 15 kategorikartet for Dim 1 vs Dim 2.



Figur 15: Kategorikart fra korrespondanseanalyse.

Vi husker at kategorier nær hverandre har en tendens til å opptre sammen. Vi kan

(i) ta utgangspunkt i en variabelkategori (sort), og spørre hvilken av de fire gruppene (rød) som er nærmest

(ii) ta utgangspunkt i en gruppekategori (rød), og spørre hvilke variabelkategorier (sort) som ligger rimelig nær og i hvilken retning

Med det siste utgangspunktet ser vi følgende i de respektive nære omgivelser:

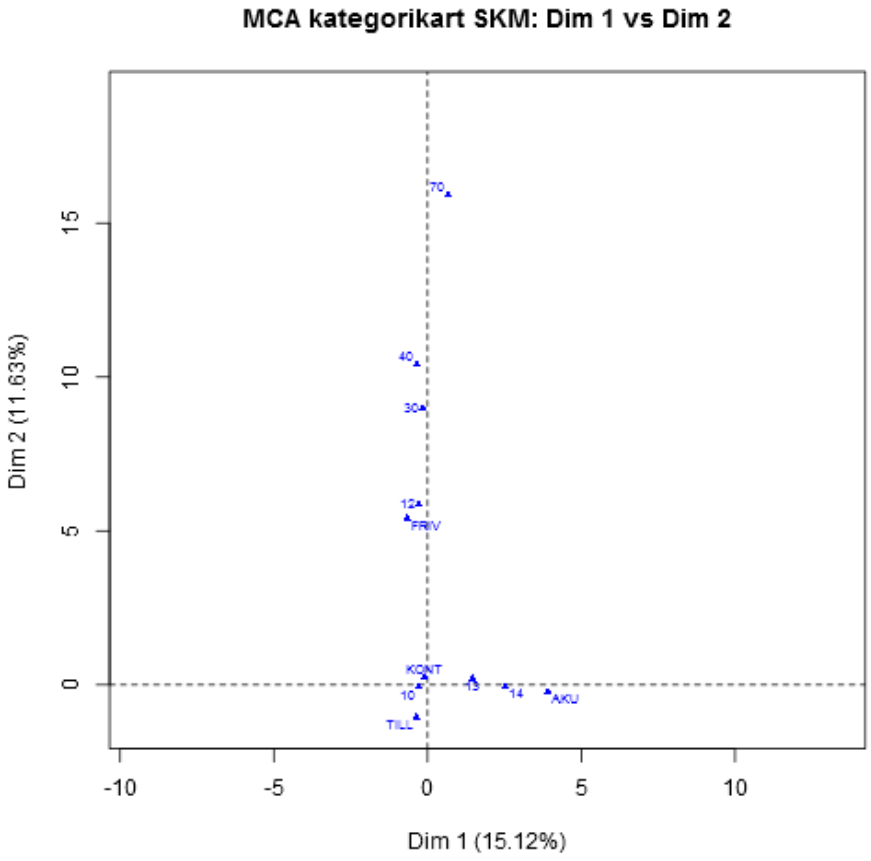
KONT:	INNTEKT_1,FORMUE_1,SENT_012
FRIV:	INNTEKT_2,FORMUE_2,GJELD_2,HIST_J
AKU/TILL:	FORMUE_0, SAM_1, PERS_1, SENT_3

Dette betyr at referansegruppen blir mest assosiert med middels inntekt (1) og middels formue

(1), samt sentralitetskode 0,1 og 2, i hvert fall i større grad enn de tre andre gruppene. Gruppen med frivillig retting blir mest assosiert med høy inntekt (2), høy formue (2), høy gjeld (2) og med en historikk (J), i hvert fall i større grad enn de øvrige gruppene. AKU-gruppen (automatisk kontroll utland) og tilleggsskatt er vanskelig å skille innbyrdes (i de to dimensjonene), men assosieres sterkest med ingen formue (0), enslig skatt (1), personkode (1) og sentralitetskode (3). For øvrig merker vi oss at de frivillig rettede ligger lengst vekk fra kontrollgruppen, og kan derfor betraktes å ha minst til felles med denne.

Noen alternative spesifikasjoner: Dersom en isteden spesifiserer GRUPPE som ordinær variabel, blir kategorikartet for de første to dimensjonene i hovedsak som ovenfor, men med de tre kategoriene AKU, TILL og FRIV forskjøvet nedover, slik at FRIV nå kommer tett på INNTEKT_2. Dersom en i tillegg spesifiserer SKM (skjemagruppe), er det ingen sterke assosiasjoner med noen av de tre gruppekategoriene utenom kontrollgruppen. Den skjemagruppe som sterkest skiller (AKU, TILL, FRIV) fra KONT er kode 70 (Diplomat), og dernest kode 14 (Midlertidig opphold), som skiller (AKU, TILL) fra KONT.

Det er fristende å tolke nærhet mellom kategorier for to variabler som grad av evne til å predikere kategori for den ene variabel ut fra kjent kategori for den andre (forklarende) variabel. Her er det mulighet for å trå feil. Vi vil her problematisere to forhold: (i) I vårt tilfelle er vi spesielt interessert å predikere GRUPPE, og datamateriale har et antall i hver kategori som ikke reflekterer andelene i populasjonen (for FRIV var om lag 1000 alt som fantes). (ii) Enkelte forklarende variable har kategorier med få eller ingen observasjoner. Det gjelder for eksempel SKM (skjemagruppe). Hvilken rolle dette spiller for løsningen er ikke lett å overskue fullt ut. Det første problemet ble søkt løst ved å la GRUPPE være supplerende variabel, som dermed ikke bestemte selve løsningen. Det andre problemet kunne vi forsøkt løse ved å utelate kategorier med få individer. I våre data er SKM et problem, og variabelen ble utelatt helt. For å belyse begge problemene med et eksempel, ser vi på kategorikartet for en analyse med de to variable GRUPPE og SKM (Figur 16).



Figur 16: Kategorikart. 10=full skatteplikt (bosatt), 13=utvandrat, 14=midlertidig opphold, 20=stedbunden skatteplikt (bosatt), 30=stedbunden skatteplikt (ikke bosatt), 40=dødsbo, 70=diplomat. NB: Noen kategorier er utelatte her.

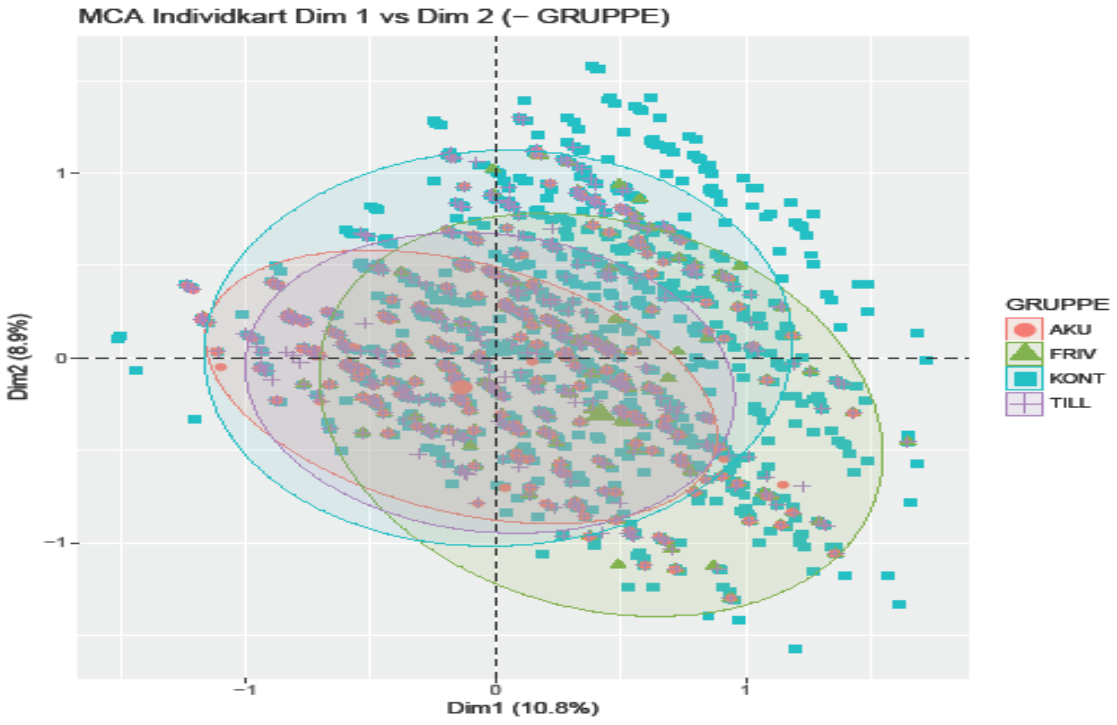
I dette kartet er kontrollgruppe-kategorien KONT nær krysset i aksekorset. Her finner vi også, ikke uventet, SKM-kategorien med kode 10 (full skatteplikt bosatt). Vi ser at den første aksen Dim 1 bidrar til å skille AKU fra de øvrige KONT, TILL og FRIV, og at den andre aksen Dim 2 bidrar til å skille FRIV fra KONT, TILL og AKU. Vi ser at kategorien TILL (tilleggsskatt) er nær KONT på nedsiden, uten SKM-kategorier som skiller klart fra kontrollgruppen. SKM-kategori 13 (utvandret) og 14 (midlertidig opphold) kan tilsynelatende bidra til å «peke ut» individer som AKU. Vurdert ut fra avstandene i kartet er det ingen andre SKM-kategorier som er bedre, men avstanden til KONT og TILL er om lag like lang, så individet kan like gjerne høre til der. Vi ser i Dim 2 at FRIV er assosiert med kategoriene 30 (stedbunden skatteplikt, ikke bosatt), 40 (dødsbo) og 70 (diplomat), og siden disse punktene ligger over, peker de mer i retning av FRIV enn til (KONT, AKU, TILL). Merk imidlertid at avstanden fra FRIV til SKM-kategorien 10 for skatteytere flest ikke er så mye lengre. Hvor misvisende dette kan være, ser vi fra hyppighetstabellen selv (Tabell 5):

	10	13	14	30	40	70
AKU	1909	592	2903	2	1	1
FRIV	940	15	8	4	24	1
KONT	84858	2375	7014	89	488	8
TILL	26720	533	1253	4	10	0

Tabell 5: Frekvenser for SKM-kategori for de fire gruppene.

Her ser vi at det forholdsvis få individer med kodene 30, 40 og bare ti individer for kode 70 (diplomater), hvorav åtte i kontrollgruppen, og en i hver av gruppene AKU og FRIV. Det kan da synes rart at 70 assosieres så sterkt med FRIV og ikke AKU. Et enda mer påfallende resultat får vi dersom vi tilfeldigvis hadde fått med oss et individ i kontrollgruppen med kode 12 (Barn under 13 år). Kartet vil da se ut som ovenfor, med punktet 12 tett på FRIV. Grunnen til disse merkverdighetene er de store forskjellene i de marginale antall, både horisontalt og vertikalt i tabellen.

Et individkart for de opprinnelige spesifikasjonen (uten SKM) er gitt i Figur 17. For å få en informativ graf plottes kun et representativt utvalg av individene, og like mange fra hver gruppe. Vi ser at de fire populasjonene overlapper i stor grad, og at det er gruppen med de frivillig rettede som adskiller seg mest fra kontrollgruppen.



Figur 17: Individkart.

4 Klassifikasjonstrær (CART)

Vi ønsker her å finne et mindre antall variabler som best mulig skiller hver av de tre gruppene (friv, tillegg, AKU) fra kontrollgruppen, dvs. tre parvise sammenligninger. I tillegg kan man se på muligheten for å skille de fire gruppene under ett. Vi vil benytte CART-metoden («Classification and regression trees») implementert i programpakken *rpart* i programsystemet R. Til forskjell fra CART-analysene i SNF-Rapport 10/2012, der formålet var å kartlegge muligheten for korrekt klassifisering i en «nål i høystakk» kontekst, vil vi her sette om lag like store grupper opp mot hverandre, dvs. sette hver av de tre gruppene (friv, tillegg, AKU) opp mot en like stor kontrollgruppe, og vi ønsker i første omgang begrense antall variable fortrinnsvis til de numeriske variable knyttet til inntekt, formue og gjeld, samt noen sentrale (og forståelige) kategoriske variabler.

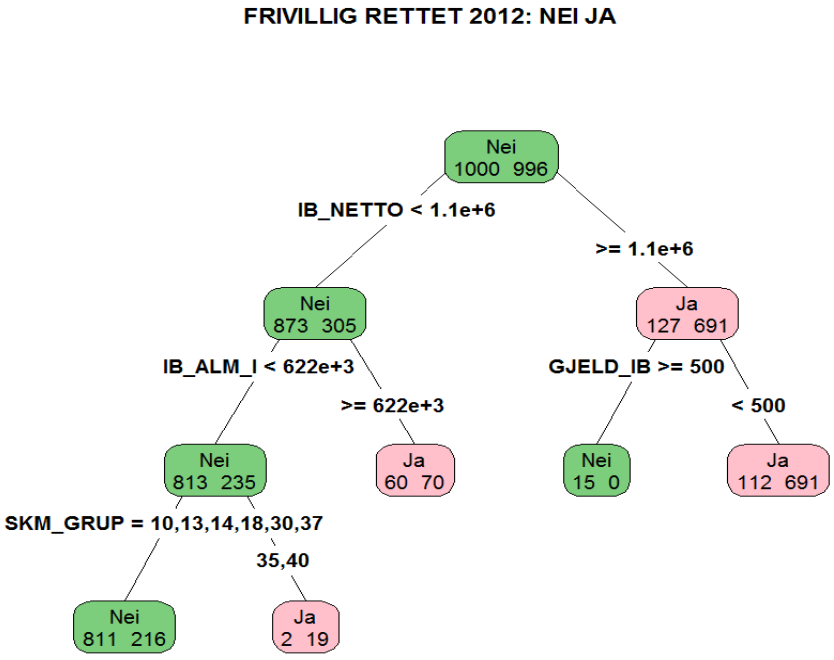
Frivillig rettede vs. Kontrollgruppe

Datamaterialet for 2012 omfattet 998 frivillig rettede skatteytere. Disse settes opp mot en om lag like stor kontrollgruppe, bestående av 1000 tilfeldig utvalgte slike skatteytere fra databasen. Ut fra en rekke testkjøringer med ulike utvalg av kategoriske og numeriske variable, der man trinnvis eliminerer variabler som ikke ser ut til å ha noen klassifiserende evne, sitter vi igjen med følgende aktuelle variable:

Kategoriske variable: Primært SKM_GRUPPE, PERSON_KODE og sekundært HISTORIKK_KODE, HAR_LEVERT_SA, SENTRALITETSKODE, SAMSKATT_KODE

Numeriske variable: IB_NETTO_FORMUE, IB_ALM_INNT_E_SERF, GJELD_IB_OG_UB

Kjøring med disse tre numeriske variable og kategorivariabelen SKM_GRUPPE ga treet i Figur 18, som involverer alle disse fire variabler.



Figur 18: Klassifikasjonstre for gruppen Frivillig rettet.

Den omsnudde treet i Figur 18 tolkes slik: Hver forgrening i treet representerer ny informasjon, og hver boks i treet besvarer spørsmålet er skatteyteren i gruppen frivillig rettet (NEI/JA) ut fra den informasjon som nå er tilgjengelig. Starter vi ved roten vet vi bare at det er 1000 kontrollgruppen og 996 frivillig rettede, idet to av de opprinnelige 998 mangler registrering på minst en de de involverte variabler i treet og er utelatt. Svaret NEI i denne boksen betyr kun at de 1000 i kontrollgruppen utgjør flertallet av de totalt 1996 involverte skatteytere. Dette innebærer at dersom vi skal tippe gruppetilhørighet for en tilfeldig skatteyter i boksen, gir NEI større sjanse for å tippe riktig enn JA. Den variabel av de fire spesifiserte som alene er best egnet til å separere de to gruppene viser seg å være IB_NETTO_FORMUE, med et skillepunkt på 1.1 mill. kroner, og slik at skatteytere med formue mindre enn dette går ned til venstre boks, med 1178 skatteytere, der 873 er skatteytere fra kontrollgruppen (svar NEI) og 305 frivillig rettede (svar JA). Kunnskapen om liten formue har altså gitt oss holdepunkt for at en tilfeldig skatteyter i den boksen ikke hører med blant de frivillig rettede. Hele boksen blir da markert med NEI og farget grønn. Skatteytere med formue større enn 1.1 mill. kroner går ned til høyre boks, med 818 skatteytere, der 127 er skatteytere fra kontrollgruppen (svar NEI) og 691 frivillig rettede (svar JA). Her er majoriteten de frivillig rettede, og svaret på det overordnede spørsmål

er JA. [6] På neste nivå i treet har vi to forgreninger, og vi ser at den variabel som i tillegg til formuen skiller de to gruppene ikke er den samme for de med små og de med store formuer. For de med små formuer, er det inntektsvariabelen IB_ALM_INNT_E_SERF med skillepunkt for inntekter på mindre enn eller større enn 622 000 kroner, og for de med høy formue er det gjeldsvariabelen GJELD_IB_OG_UB, med skillepunkt for gjeld større enn eller mindre enn 500 kroner (i praksis uten gjeld). Følger vi hver av grenene mot høyre som gir JA-flertall, ser vi at blant de med lav formue og høy inntekt har vi 130 skatteytere, hvorav et lite flertall på 70 tilhører de frivillig rettede, mens blant de med høy formue og lav/uten gjeld har vi 803 skatteytere, hvorav et stort flertall på 691 tilhører de frivillig rettede. Merk at her er det (høy) gjeld som «frikjenner». For skatteytere med lav formue og lav inntekt oppnås oppsplitting enda ett nivå, og da med kategorivariabelen skjemagruppe SKM_GRUPPE, der kodetall (35, 40) tilordnes JA-gruppen, og 19 av de 21 skatteyterne klassifisert i henhold treet virkelig tilhører de frivillig rettede. Kodetall 40 representerer dødsbo, og det er ikke uventet at dette trer fram i treet på denne måten.

De tre røde sluttboksene (nodene) i treet markert JA, gir oss klassifikasjonsregler for antatt tilhørighet til JA-gruppen, (de frivillig rettede) med utgangspunkt i en populasjon av omtrent like mange av hvert slag. Tallene i de enkelte sluttnodene gir da et visst inntrykk av sjansen for å klassifisere riktig en tilfeldig skatteyter. Prosentandelen korrekte JA i de tre røde sluttboksene med JA-klassifikasjon er henholdsvis 90%, 54%, 86%, dvs. vi i to av situasjonene har forholdsvis høy sjanse for å klassifisere korrekt, mens i den tredje gjør vi det ikke stort bedre enn 50%, slik sjansen er uten mer informasjon om skatteyter.

I praksis står en imidlertid overfor en populasjon med ordinære skatteytere av en langt høyere størrelsesorden enn de frivillig rettede, og en har et «nål i høystakken» problem. Sjansen for å predikere korrekt er dermed liten, men beslutningsregler av dette slaget kan likevel være nyttig i (automatiserte) prioriteringer av kontrollinnsats. Variabelutvelgelsen med tilhørende klassifikasjonsregler er her etablert med utgangspunkt i en balansert populasjon. Det er argumenter for og imot dette. Eksplisitte sannsynlighetsvurderinger er likevel mulig, ved å bringe inn apriori-sannsynligheter eller foreta kryssvalidering med resampling-metoder.

Vår spesifikasjon var (IB_NETTO_FORMUE, IB_ALM_INNT_E_SERF, GJELD_IB_OG_UB, SKM_GRUPPE). Ulike tillegg av kategorivariable er forsøkt. Dersom en legger til PERSON_KODE, kommer den ikke med, og treet blir som ovenfor. Legger en til de sekundære kategorivariablene ovenfor, forsvinner både inntektsvariabelen og skjemagruppe. Den gitte spesifikasjon er derfor vår foretrukne.

[6] Det betyr ikke at dette er det beste tips stilt overfor hele populasjonen av skatteytere med ukjent gruppetilhørighet, under hensyntagen til denne informasjonen, siden andelen ordinære skatteytere der er mye større, og det derfor er stor apriori sannsynlighet for at en tilfeldig skatteyter tilhører denne gruppen.

Det kan ha interesse å sammenligne med de funn vi gjorde i kapittel 2 og kapittel 3. I kapittel 2 ble IB_NETTO_FORMUE plukket ut som den mest lovende variabelen. Dette er nettopp vår første splitt-variabel her. Cramer's V har imidlertid ikke fanget opp potensialet som sekundært ligger i gjelds- og inntektsvariablene, og tertiært i skjemagruppe. Vi har derfor fått demonstrert at enkle bivariate analyser er utilstrekkelig. I kapittel 3 (Figur 15) ser vi at frivillig retting (FRIV) er sterkt assosiert med (GJELD_2, FORMUE_2, INNTEKT_2), dvs. øverste kategori på hver av de økonomiske variabler, men også assosiert med kategorien HIST_J, dvs. har en historie. Treet i Figur 18 omfatter nettopp de tre økonomiske variablene, men beslutningsregelen avviker fra å peke ut de i øverste kategori til gruppen frivillig rettede. Dette kan synes noe overraskende, men det kan synes som om korrespondanseanalyse ikke er like velegnet til å fange opp betingede strukturer på flere nivåer.

Det er også av interesse å sammenligne resultatene her med funnene i den tidligere SNF rapport Nr 10/12, som var basert på et noe mindre tallmateriale. Der fokuserte vi på klassifikasjonsevne i en kontekst der den spesielle gruppen var liten i forhold til en mye større kontrollgruppe. I dette kapitlet er utgangspunktet to like store grupper, og målsettingen av mer eksplorativ natur. Resultatene er derfor ikke fullt ut sammenlignbare, men også i den tidligere rapport kom formue ut som den primære skillevariabelen, etterfulgt av inntekt, mens gjeld ikke pekte seg ut. Blant de kategoriske variable inngikk også der skjemagruppe.

Et alternativ til klassifikasjonstre (CART) som også plukker ut de essensielle variable, er logistisk regresjon etter LASSO-kriteriet (Least Absolute Shrinkage and Selection Operator). Dette innebærer at man, i forhold til vanlig logistisk regresjon, straffes for å ha med forklarende variable uten særlig betydning, noe som medfører at disse får regresjonskoeffisienten satt til null. En analyse basert på R-programmet *Lasso2* endte opp med følgende variable med regresjonskoeffisienter forskjellig fra null: IB_NETTO_FORMUE, IB_ALM_INNT_E_SERF, GJELD_IB_OG_UB, HAR_LEVERT_SA. Løsningen vil imidlertid avhenge av valgt nedskalering av de numeriske variable, noe som typisk er nødvendig for å sikre rask konvergens av algoritmen. Her er brukt de opprinnelige verdier dividert med 1000. Valget av kriterium og skalering kan gjøres ved kryssvalidering, ut fra hva som gir best prediksjon.[7]

En logistisk modell forutsetter at log-odds er lineær i de forklarende variabler. Dette er urealistisk for de numeriske variable inntekt, formue og gjeld. Mer realisme kan oppnås ved å log-transformere disse. Siden null er forkommende verdi må en legge til et ubetydelig positiv verdi før log-transformering. Det mest rimelige er egentlig å kategorisere de numeriske variable, hver i to eller flere ordinale kategorier, der null godt kan være egen kategori. Valget av kategorisering blir da delvis ad hoc, og det kan være mer tilfredsstillende med en metode som CART, som har kategoriseringen innbakt.

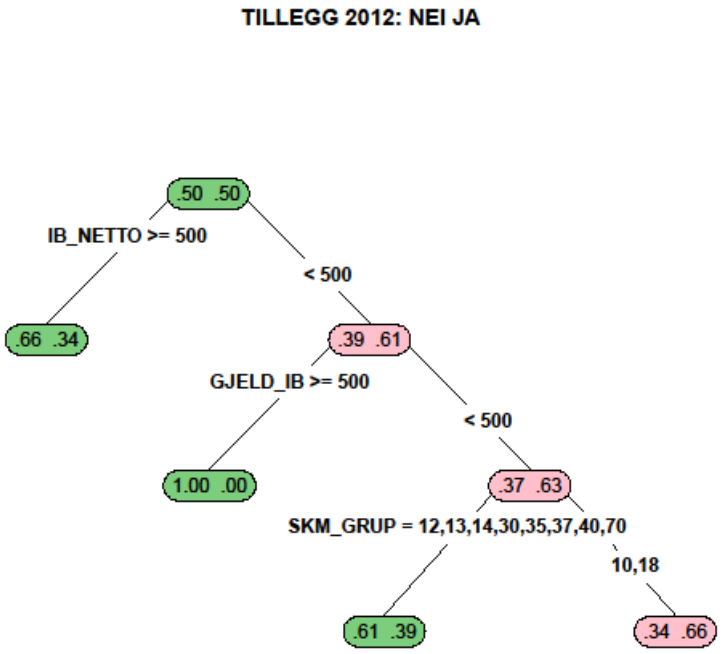
[7] Samme analyse basert på R-programmet *glmnet* med kryssvalidering ga til dels andre resultater.

Tilleggsskattgruppe vs. Kontrollgruppe

For sammenligningens skyld ser vi på tilsvarende klassifikasjonstre for gruppen med tilleggsskatt, med de tre numeriske variablene og de to primære kategorivariablene, dvs

(IB_NETTO_FORMUE, IB_ALM_INNT_E_SERF, GJELD_IB_OG_UB, SKM_GRUPPE, PERSON_KODE).

I databasen for 2012 var det om lag 29400 skatteytere med tilleggsskatt (og med registrering på alle fem variable), og det ble valgt ut et tilsvarende antall fra referansegruppen. Det etablerte tre er vist i Figur 19, der vi istedenfor antall i hver boks angir andelen i hver kategori (NEI, JA).



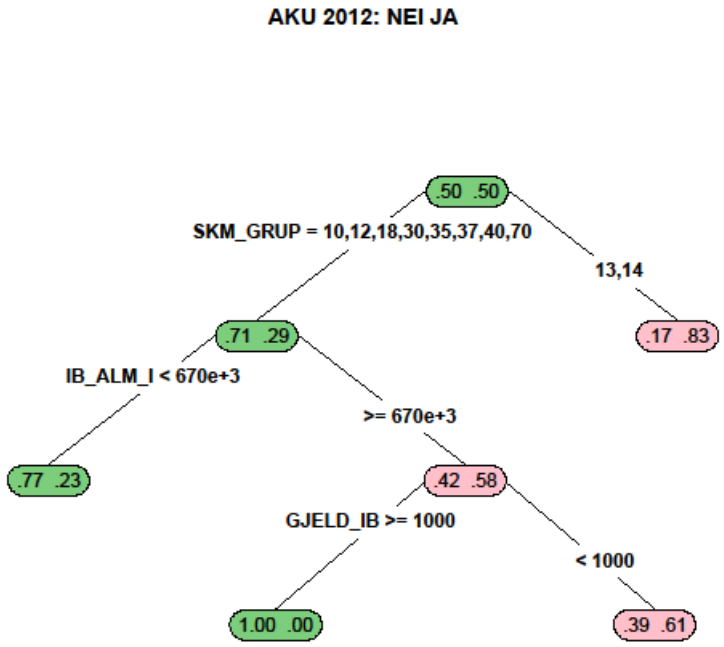
Figur 19: Klassifikasjonstre for tilleggskattgruppen.

Vi ser at den beste kategoriseringen på første nivå også her er formue. Her blir bare de med liten/uten formue kategorisert ytterligere, og da med omsyn til gjeld, der bare de med liten/uten gjeld blir kategorisert igjen, og da med omsyn til skjemagruppe. Inntektsvariabelen og personkode har tydeligvis ingen klassifikasjonsevne utover de variable som nå er med. Treet gir derfor holdepunkter for at det er i gruppen av skatteytere (i praksis) uten formue og gjeld, og som er i kategoriene (10,18) for skjemagruppe at vi har best sjanse for å finne de med tilleggsskatt. Denne gruppen (i den nederste røde boksen) består av 30199 skatteytere ut av 58842, hvorav 19841 i virkeligheten hadde tilleggsskatt. Det er denne andelen 19841/30199=0.66 som fremkommer i denne boksen. Brukt som klassifikasjonsregel har vi altså

oppnådd at sjansen for korrekt klassifisering i en populasjon av like mange av hvert slag, er økt fra 50% til 66%, altså ikke særlig mye i forhold til uten skatteyterinformasjonen, og lavere enn det vi oppnådde for gruppen av frivillig rettede.

AKU-gruppe vs. Kontrollgruppe

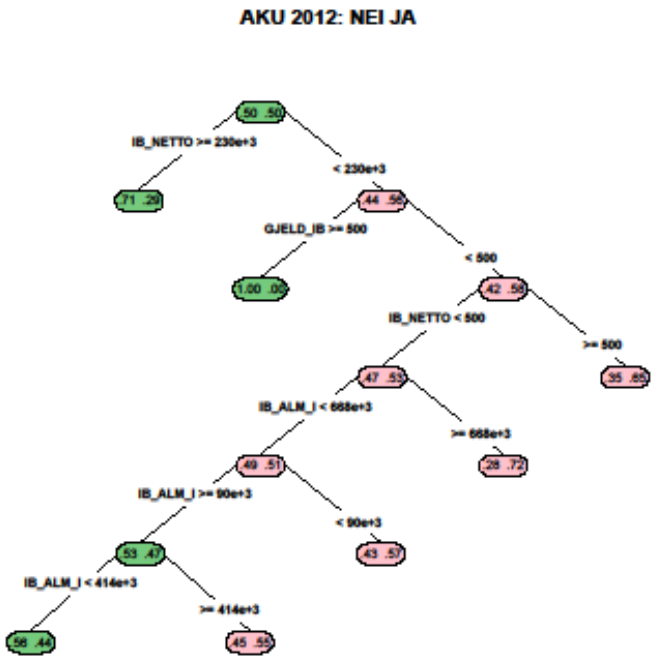
For AKU-gruppen med om lag 5500 skatteytere i databasen gir samme spesifikasjon treet i Figur 20.



Figur 20: Klassifikasjonstre for AKU-gruppen.

Her er skjemagruppe den variabel som alene skiller best, der kategoriene (13,14) gir indikasjon på tilhørighet til AKU-gruppen. For de øvrige har en, for de med høy inntekt (over 670 000 kr.) og lav/uten gjeld, en overvekt av skatteytere i AKU-gruppen. Igjen gir de oppnådde klassifikasjonsregler ikke spesielt sterke indikasjoner på korrekt gruppetilhørighet, men noe bedre enn for gruppen med tilleggsskatt. Det er skjemagruppe som bidrar mest, siden vi i den røde JA-sluttnoden for kodetall (13,14) finner 4237 skatteytere, hvorav 3525 tilhører AKU-gruppen, slik at 3525/4237=0.83, mens vi i den andre røde JA-sluttnoden har 1120 skatteytere, hvorav 685 tilhører AKU-gruppen, slik at 685/1120=0.61.

En sammenligning av trærne for de tre gruppene kan nå belyse om de har ulike kjennetegn. Vi ser at formue inngår for frivillig rettede og tilleggsskatt, mens gjeld og skjemagruppe inngår i alle tre gruppene frivillig rettede, tilleggsskatt og AKU. På den annen side ser vi at inntekt bare inngår for de frivillig rettede. Strukturen for klassifisering er nokså ulik i de tre gruppene. For AKU «forstyrres» bildet ved at skjemagruppe er første splitt, og rollen til de numeriske variable er betinget gitt denne. Utelater vi SKM_GRUPPE og spesifiserer kun de tre numeriske variable får vi treet i Figur 21.



Figur 21: Klassifikasjonstre for AKU-gruppen for tre numeriske variable

Her ser vi at alle tre numeriske variable er representert, med formue som første splitt, slik tilfellet var for de to andre gruppene. Klassifikasjonsregelen ble her noe mer komplisert. Eksempelvis peker den røde AKU-boksen lengst til høyre ut skatteytere med moderat positiv formue uten gjeld, idet < 500 typisk betyr null og dermed kan navngi ≥ 500 som positiv. Vi ser forøvrig at samme variabel kan forekomme flere ganger i en slik betinget struktur med binære splitt, og at dette muliggjør at det er verdier i midtområde som peker mot en bestemt gruppe.

Dersom en i praksis vil bruke en regel for klassifikasjon basert på historiske data til å klassifisere nye skatteytere med ukjent gruppetilhørighet, må en selvsagt vurdere om regelen gir god mening skattefaglig sett. Om så ikke er tilfelle, kan de etablerte trær likevel gi grunnlag for faglige diskusjoner som kan være nyttige ved prioriteringer av ressursbruk innen etaten.

5. Diskusjon

Denne rapporten omfatter statistiske analyser av et datamateriale bestående av tre grupper skatteytere som har vært i Skatteetatens søkelys holt opp mot en referansegruppe av antatt ordinære skatteytere. Tre ulike metoder av eksplorativ natur er prøvd ut: Enkel kategorisering, korrespondanseanalyse og klassifikasjonstrær. De tre metodene egner seg alle til grafisk presentasjon, men er ellers av ulik natur.

Kategorisering er begrenset til en «forklarende» variabel om gangen, kan presentere gruppeforskjeller på en iøynefallende måte i form av sammenstilte søylediagrammer. Disse kan gi verdifull kunnskap i kombinasjon med generell apriori kunnskap i skatteetaten om skatteytere. Dette grunnlag er imidlertid utilstrekkelig for å avgjøre hvilke variabler som skiller best. Supplerende samvariasjonsmål (Cramer's V) kan gi en pekepinn på dette. Dersom ambisjonsnivået er klassifikasjon, gir denne angrepsmåten likevel et utilstrekkelig grunnlag, både fordi slike regler ikke er lett å avlede ut fra grafene, og fordi det er en en-variabelmetode, og ikke tar omsyn til kombinasjoner av variabler.

Korrespondanseanalyse er i utgangspunktet en fler-variabel metode som kan håndtere et stort antall variable samtidig. Med variabelkart har vi fått visualisert hvilke variabler som utgjør hoveddimensjonene i datamaterialet (her inntekt og formue/gjeld). Med kategorikart har vi fått visualisert i hvilken utstrekning de enkelte variabelkategoriene kan assosieres med gruppetilhørighet. Her har vi i en figur fått mye informasjon som kan kombineres med apriori kunnskap. Kartene gir pekepinn på hvilke variabler som har betydning og i noen utstrekning klassifikasjonsregler, dvs. hvilke kategorikombinasjoner av flere variabler som peker mot en bestemt av de fire gruppene.

Klassifikasjonstrær er også en fler-variabel metode, og direkte rettet mot å etablere beslutningsregler. Dette skjer ved en sekvensiell prosess, der en ender opp med en forholdsvis enkel struktur, der variabler uten betydning er fjernet. Denne presenteres enkelt i form av et tre, som er forholdsvis enkelt å fortolke. Metoden innebærer at enkelte variabler som har tilsynelatende liten prediktiv verdi alene, kan få prediktiv verdi betinget gitt andre variabler. Det innebærer også at enkelte variabler som kan ha en viss prediktiv verdi alene, ikke er med i treet, fordi de samvarierer sterkt med en eller flere variabler som er med. Klassifikasjonstrær har den åpenbare fordel at de gir beslutningsregler som tar eksplisitt omsyn til betingede strukturer.

Vi har i denne rapporten hatt et spørsmål i bakhodet: Hvilke variabler har størst betydning for å skille mellom gruppene? I klassifikasjonskonteksten kan en spørre om det er fruktbart å rangere de enkelte variablene etter prediktiv verdi. En mulighet er å se hva en mister dersom en variabel var utilgjengelig. Det er kombinasjonen av variabler som gir det beste resultat, men det kan godt hende at en annen kombinasjon av variabler gir et nesten like godt resultat. Det har vist seg at en kan forbedre klassifikasjonsevnen ved å kombinere mange ulike trær (såkalt «skog av trær»). Et mulig mål for betydningen av de enkelte variabler kan da være å telle opp hvor ofte hver variabel forekommer som «gren eller kvist» i skogen. En skog kan imidlertid stå i veien for læring: «En ser ikke trærne for bare skog!».

Referanser

Andersson, J., Lillestøl, J. og Støve, B. (2012). Kjennetegnsanalyser av skatteytere som unndrar skatt ved å skjule formuer og inntekter i utlandet. *SNF Rapport* Nr 10/12.

Clotfelter, C.T. (1983). Tax Evasion and Tax Rates: An Analysis of Individual Returns. *The Review of Economics and Statistics* 65 (3), 363-373.

Feinstein, J.S. (1991). An Econometric Analysis of Income Tax Evasion and its Detection. *RAND Journal of Economics* 22 (1), 14-35.