HVA VET VI OM DEM SOM SKJULER INNTEKT OG FORMUE I SKATTEPARADIS? F

JONAS ANDERSSON er professor i bedriftsøkonomisk analyse ved Institutt for foretaksøkonomi ved Norges Handelshøyskole (NHH) og medarbeider ved Senter for skatteforskning ved NHH. Andersson har arbeidet med utvikling og anvendelse av statistiske metoder innenfor økonomifaget, herunder spesielt tidsrekkeanalyse.

JOSTEIN LILLESTØL er professor i statistikk ved Institutt for foretaksøkonomi ved Norges Handelshøyskole (NHH) og medarbeider ved Senter for skatteforskning ved NHH. Lillestøl har tidligere arbeidet med teoretiske og anvendte problemstillinger på mange felt, herunder revisjon, risikoanalyse og kvalitetsstyring.

BÅRD STØVE er førsteamanuensis i statistikk ved Matematisk institutt ved Universitetet i Bergen og assosiert medarbeider ved Senter for Skatteforskning ved Norges Handelshøyskole (NHH). Han var tidligere ansatt i en postdoktor-stilling ved NHH. Støve har arbeidet med utvikling av ikke-parametriske økonometriske metoder, samt anvendelser av slike metoder på ulike praktiske problemstillinger.

GUTTORM SCHJELDERUP er professor i samfunnsøkonomi og bedriftsøkonomisk analyse ved Institutt for foretaksøkonomi ved Norges Handelshøyskole (NHH) og leder Senter for skatteforskning ved NHH. Schjelderup har ledet flere offentlige utvalg og hans viktigste forskningsfelt omhandler skatteparadiser, skattlegging av flernasjonale foretak og utformingen av skattesystemer i en åpen økonomi.

SAMMENDRAG

I denne artikkelen presenteres resultater fra et prosjekt utført for SNF (Samfunns- og Næringslivsforskning) på oppdrag av Skattedirektoratet (SKD). Hensikten med prosjektet var å finne kjennetegn ved personlige skattytere som har unndratt skatt gjennom å skjule formuer og/eller inntekt i utlandet/skatteparadiser, og dermed potensielt kunne brukes ved prioriteringer av kontrollarbeid. Tilgjengelig for analysen var ligningsdata for skatteytere som hadde meldt seg frivillig som følge av skatteamnestiordningen, holdt opp mot tilsvarende data for en kontrollgruppe av regulære skatteytere. Sentralt i prosjektet har vært å vurdere en rekke tilgjengelige analysemetoder med omsyn til deres styrke og svakheter. I artikkelen presenteres først konteksten for problemstillingen og en kort litteraturomtale. Deretter beskrives datagrunnlaget, og valget mellom ulike statistiske metoder ut fra ulike valgkriterier, herunder klassifikasjonsevne og egnethet for implementering. Varianter av logistisk regresjon og klassifikasjonstrær (C&RT) utmerket seg som de mest lovende metodene, og det er relativt komplekse modeller med mange kjennetegn som viser seg å ha best klassifiseringsevne, herunder spesielt såkalte ensemblemetoder. Artikkelen problematiserer også hva som skal menes med de enkelte kjennetegns «viktighet» og bruken av resultatene på nye data.

1. INNLEDNING

Et av de viktigste gjennombruddene i økonomisk forskning er innsikten om at skjult informasjon svekker markeders virkemåte og slik sett bidrar til lavere vekst. Skjult informasjon skaper også private insitamenter til å begå handlinger som fra samfunnets synspunkt er skadelige på en rekke områder.[1] Finanskrisen har ført til større oppmerksomhet om skatteparadisene og det faktum at de bidrar til informasjonsasymmetrier på en rekke viktige samfunnsområder. I Norge har fokus spesielt vært rettet mot det forhold at skatteparadisene tilbyr hemmelighold som gjør det mulig for skatteytere å skjule inntekt og formue.

Erfaringer på tvers av land viser at når skatteyter selv skal rapportere inntekt og formue, øker omfanget av skatteunndragelse.[2] Slik innsikt har ført til ordninger i mange land hvor tredjeparter rapporterer inntekt på vegne av skatteyter. I Norge oppgir for eksempel arbeidsgiver lønn og trekker skatt til skattemyndighetene. Likeledes rapporterer banker i Norge innskudd og renter til skattemyndighetene. I EU forsøker man å få til det samme på tvers av land gjennom det såkalte Sparedirektivet.

I Norge har skattedirektoratet hatt fokus på skatteparadisene. Skattedirektør Svein Kristensen skrev i en kronikk i Dagens Næringsliv at Skattedirektoratet mener nordmenn bare på bankkontoer i skatteparadiser har gjemt vekk rundt 200 milliarder kroner.[3] Dette tallet baserer seg på ulike prosjekter som Skatteetaten jobber med. Bare i den såkalte LTG-saken, hvor en ansatt i Liechtenstein-banken LGT stjal en CD med navn på utenlandske innskytere, har norske skattemyndigheter avslørt skatteunndragelse på om lag 523 millioner kroner.

Som et ledd i arbeidet mot skjulte formuer og inntekt i skatteparadiser har Skatteetaten innført en skatteamnestiordning. Vilkåret for å komme inn under denne er at endringen av egen ligning ikke kommer som en konsekvens av kontrolltiltak fra skattemyndighetenes side, eller at skattemyndighetene har fått opplysninger fra andre om skjult formue og inntekt.[4] Bare de siste tre årene har denne ordningen ført til at skjulte formuer til en verdi av over to milliarder kroner er blitt rapportert til skattemyndighetene. Det meste av dette er penger som har vært skjult i utlandet, spesielt i Sveits og Luxemburg. I et tiårsperspektiv har ordningen ført til en formuesøkning på rundt 20 milliarder kroner i Norge.

I dette arbeidet har vi forsøkt å finne kjennetegn ved personlige skatteytere som har unndratt skatt gjennom å skjule formuer og/eller inntekt i utlandet/skatteparadiser. Slike kjennetegn kan for eksempel være av demografisk, økonomisk eller sosioøkonomisk karakter. Hensikten har vært å finne ut om det er mulig å gruppere skatteyterne ut fra risikofaktorer, slik at Skatteetaten kan sette inn riktig virkemiddel mot de ulike risikogruppene. Vårt utgangspunkt har vært at ved å basere seg på databaserte statistiske modeller ved utvelgelsen av kontrollobjekter kan formodentlig evnen til å avsløre unndratt skatt kunne økes, og omfanget av utførte kontroller uten avsløringer kunne reduseres.

I første del av artikkelen ser vi på hvilken statistisk modell som er best egnet til å analysere et utvalg av skatteytere som vi vet faller i to grupper, en ærlig gruppe og en unndragelsesgruppe. Et hovedfunn er at metoden klassifikasjonstre gir best resultat ut fra en rekke kriterier som vi redegjør for under. I del to av analysen benytter vi den tilpassede klassifikasjonstremodellen på et nytt utvalg av skatteytere som vi ikke vet om er ærlige eller unndrar. Hensikten er å klassifisere dette nye utvalget av skatteytere slik at de som blir klassifisert som skatteunndragere kan prioriteres for kontroll.

Flere empiriske studier fra ulike land har undersøkt sosioøkonomiske og andre kjennetegn for å predikere skatteunndragelse samt benyttet spørreundersøkelser for å avdekke holdninger rundt skatt/selvangivelse. Se for eksempel Lee og Carley (2009), Collins mfl. (1992), Clotfelter (1983) og Webley mfl. (2001). Noen av kjennetegnene som merker seg ut i disse studiene, er kjønn, alder, utdanning, inntekt, profesjonell støtte ved selvangivelse, sosiale nettverk og hvorvidt skatteyter er

1. For forskning knyttet til denne type problemstillinger mottok professorene George A. Akerlof, A. Michael Spence og Joseph E. Stiglitz nobelprisen i økonomi i 2001.
2. Se for eksempel Kleven mfl. (2011).
3. Kronikk i Dagens Næringsliv 20.5.2010.
4. Kvalifiserer man for skatteamnesti, slipper man å betale tilleggsskatt, som ellers kan utgjøre inntil 60 prosent. Skatten man skulle betalt i tillegg til renter, må imidlertid betales. Skatt og renter kan beregnes ti år tilbake i tid.

selvstendig næringsdrivende. I våre analyser vil vi forvente at noen av disse funnene vil bli bekreftet, og forhåpentligvis vil vi kunne avdekke andre kjennetegn eller kombinasjoner av kjennetegn som kan benyttes for å øke oppdagelsessannsynligheten for unndragelse.

2. METODE OG DATA

Vi har evaluert tilgjengelige analysemetoder ut fra litteraturstudier og utprøving på anonymiserte data som Skatteetaten har stilt til rådighet. Tilgjengelig for dataanalysen var 577 skatteytere som har unndratt skatt, dette omfatter de som til da hadde meldt seg frivillig gjennom skatteamnestiordningen, kalt frivillig retting-gruppen. I tillegg var det tilgjengelig et langt større tilfeldig utvalg av personlige skatteytere, kalt kontrollgruppen. Data omfattet i alt 570 variabler fra ulike felter i den enkelte skatteyters ligning fra årene 2007 og 2008, og også data fra andre databaser som Skatteetaten har tilgjengelig.

I utgangspunktet er det flere problemer knyttet til den konteksten resultatene skal brukes i, og til selve datagrunnlaget. For det første vil i praksis antall skatteundragere være lite i forhold til antall lovlydige, slik at vi har et slags «nål i høystakken»-problem. For det andre utgjør gruppen av de frivillig rettede ikke et tilfeldig utvalg fra en større populasjon av slike. Eksempelvis kan det være at personer i en bestemt livsfase ønsker å rydde opp før møtet med sin Gud, eller at det er arvinger som griper inn. For det tredje vil det nok også fins noen uærlige i kontrollgruppen, men disse utgjør forhåpentligvis en såpass liten andel at de ikke innvirker på estimeringen av modellen i særlig grad. For det fjerde kan vi stille spørsmålet om de resultatene vi finner, og de tilpassede modellene, har prediktiv verdi for behandlingen av kommende års skatteytere. Har de fleste som burde rettes der ute, kommet frem, slik at sjansen for å finne flere er drastisk redusert? Har de som er igjen, og de nye som eventuelt kommer til, andre og vesensforskjellige karakteristika?

Et spørsmål som også melder seg, er hvor stor kontrollgruppe en bør velge i analysefasen. For modellestimeringen er det ønskelig at de to gruppene er av samme størrelsesorden, mens for beregning av realistiske sannsynligheter for korrekt klassifisering er det ønskelig at datamaterialet omfatter de to gruppene, unndragere og ærlige, i samme forhold som ved ligningsoppgjøret. Siden gruppen frivillige er liten, har det vært nødvendig å inngå et kompromiss der en ikke har altfor stor ubalanse mellom gruppene, men likevel ivaretar «nål i høystakk»-konteksten. Prosjektet hadde tilgjengelig en kontrollgruppe på 300 000 skatteytere, men man fant det hensiktsmessig å foreta analyser med 10 000 i kontrollgruppen, som er stort nok til å estimere en brukbar modell selv med mange variabler, og som i noen grad reflekterer «nål i høystakk»-problematikken. En regner med at de samme modellene er gode og de samme variablene er de viktige også i den aller videste kontekst, og at det kun er sannsynlighetene for korrekt klassifikasjon som blir redusert. I den første grovvurdering av metodene er det imidlertid brukt en kontrollgruppe på 1000 individer. Det er nok for læring og formidling av metodenes egenskaper.

Ved prioritering av metode er følgende vurderingskriterier brukt:

- klassifikasjonsevne under «vanlige» og «spesielle» forhold
- tilgjengelighet med brukervennlig programvare
- forståelig modell med lett tolkbare utskrifter
- modell og programvare kan ta omsyn til ulike typer data og mange variabler
- grad av automatisering
- fleksibelt mht. muligheter (f.eks. legge til subjektiv kunnskap)
- evne til å skape innsikt og grunnlag for kommunikasjon
- mulighet for å utnytte eksisterende SKD-kompetanse

Det er mange og til dels svært ulike analysemetoder som kan være aktuelle og er vurdert. Disse faller i følgende hovedkategorier:

a. klassisk diskriminantanalyse
b. kategorisk regresjon: logit, probit (uten og med LASSO)
c. klassifikasjonstre: CART, CHAID, QUEST
d. ikke-parametriske metoder: k nearest neighbor, kernel density estimation
e. *Multiple Adaptive Regression Splines* (MARS)
f. nevrale nettverk (*Artificial Neural Nets*, ANN): *backpropagation*-metode osv.
g. *Support Vector Machines* (SVM)
h. genetisk programmering

TABELL 1 Variabler i tentativ analyse.

FRIV	= 1 hvis frivillig retting, 0 ellers
KOMM	Formueskatt til kommunen 2008
STAT	Formueskatt til staten 2008
UB	Antall utenbygdskommuner du betaler skatt til
F0807	Endring i formue fra år 2007 til år 2008
ALMIB	Alminnelig inntekt innenbygds
SKJERM	Skjermingsfradrag aksjer
HIST	Historikk kode, 1 hvis tidligere blitt skatteberegnet
SENT	Sentralitetskode, 3 hvis bor sentralt
LØNN	Personinntekt lønn
ALDER	Alder
TOPPSKATT	Beløp
KJONN	Kjønn
OVER	Skatt – forskuddstrekk

TABELL 2 Klassifikasjonsresultat.

	PRED = 0	PRED = 1	Radsum
FRIV = 0	880	77	957
FRIV = 1	90	485	575
Kolsum	970	562	1532

Disse metodene, som hver for seg fins i ulike varianter, er i varierende grad knyttet til statistiske modeller og statistisk teori, mest de to første. De siste er knyttet til maskinell læring og såkalt kunstig intelligens (AI). I praksis har man forsøkt å kombinere begge perspektiver under termen Knowledge Discovery in Databases (KDD).[5] Det fins også bayesianske varianter av disse metodene, der subjektive oppfatninger kan komme til uttrykk, samt enkelte hybrider av metodene, blant annet av statistisk tre og logistisk regresjon. Det fins også muligheter for å kombinere klassifikasjoner fra flere modeller, av den same type eller ulike typer, såkalt *model averaging*. Det er et generelt prinsipp, men vi finner dette først og fremst i forbindelse med maskinlæring under betegnelsen *ensemble meta-algorithms*. Formålet er å oppnå bedre stabilitet, redusere avvik, forbedre klassifikasjonsevne og unngå såkalt *overfitting*. Eksempler på slike metoder er såkalt *boosting* og *bagging* (kort for *bootstrap aggregating*).

Vi har tentativt forsøkt metodene a–g i ulike varianter med et lite utvalg av variabler av ulik karakter (dikotome, kategoriske, numeriske) og fra ulike deler av selvangivelsen. Variablene er gitt i tabell 1.

3. RESULTATER

Tentative analyser ble utført på en rekke av metodene basert på de 577 frivillig rettede (FRIV = 1), med en kontrollgruppe på 1000 skatteytere (FRIV = 0). På grunn av manglende registreringer på enkelte av disse variablene ble antallet som inngår, henholdsvis 575 og 957, altså totalt 1532. Mange av metodene klarer med god treffsikkerhet å klassifisere data i riktig gruppe. Metodene gir i varierende grad innsikt i hvilke kjennetegn som har betydning for klassifikasjonen, og noen har karakter av en «sort boks» som klassifiserer uten å gi særlig innsikt i hva som er avgjørende for at et individ havner i den ene eller andre gruppen.

Som illustrasjon viser vi i tabell 2 klassifiseringsresultatet med et klassifikasjonstre av typen CART.

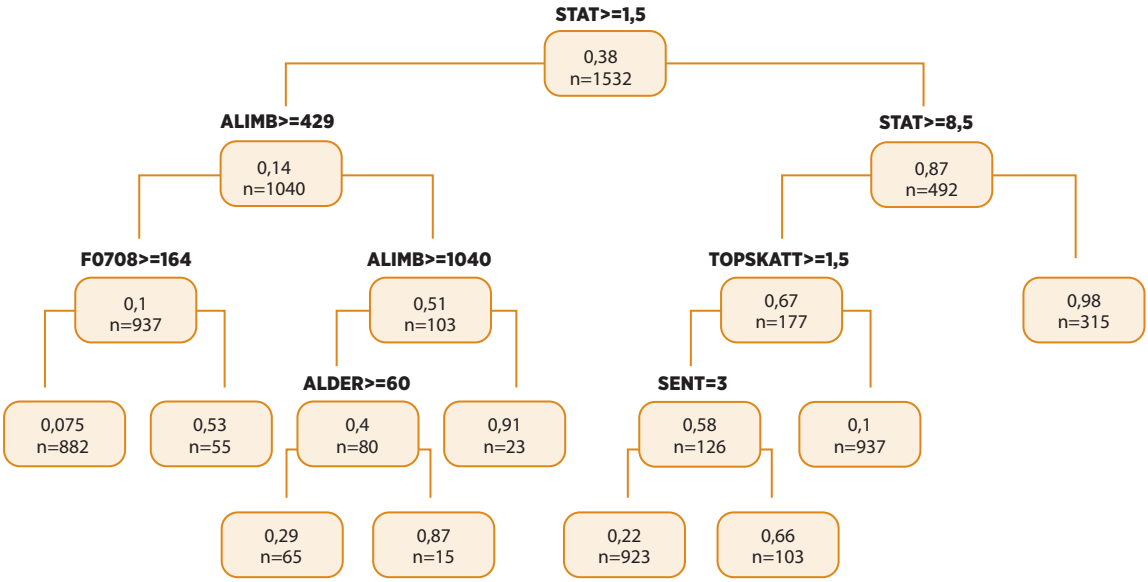
Vi ser av tabell 2 at 485 av de 575 frivillig rettede ble klassifisert korrekt. Logistisk regresjon klassifiserte 460 korrekt, altså noe dårligere, mens den såkalte *boosted tree*-metoden, som er en vekting av mange trær, klassifiserte hele 514 av de 575 korrekt.

Nå er dette klassifikasjoner av de data som er grunnlaget for modellen, såkalte *within sample*-klassifikasjoner, mens målet er å anvende en etablert modell på nye tilfeller. Det er en viss risiko for såkalt overtilpasning, slik at tallene i tabellen er for optimistiske. Analysene ble derfor gjentatt med *out of sample*-perspektivet, med klassifisering av nye observasjoner. Dette er problematisk når den frivillig rettede gruppen er så liten, men kan omgås ved såkalt kryssvalidering.

Det viste seg å være betydelig variasjon mellom metodene, både i treffsikkerhet og i hvilke variabler som pekes ut. Dette gjelde både for *within sample*- og *out of sample*-sammenligninger. Ut fra en samlet vurdering av de tentative analysene og de øvrige kriteriene, der klassifikasjonstre (CART) kom svært godt ut og logistisk regresjon kom rimelig godt ut, valgte man å gå i dybden på disse to metodene. Logistisk regresjon var fortsatt

5. Den interesserte finner mer om metodene i Hastie, Tibshirani og Friedman (2009).

FIGUR 1 Klassifikasjonstre (CART).



aktuelt fordi dette er en såpass velkjent metode, og om den ikke kommer ut best, så er den et godt utgangspunkt å diskutere egenskaper opp imot.

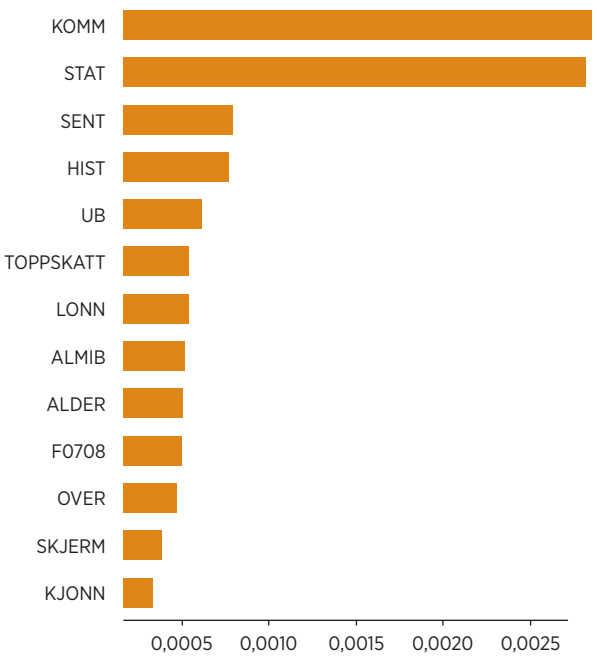
Vi vil senere bruke eksemplet ovenfor til å forklare mer detaljert hva et klassifikasjonstre er. De mer grundige og realistiske analysene er som nevnt tidligere utført på en større kontrollgruppe (10 000) og et bredere spekter av variabler. Det ble tidlig klart at det var kombinasjoner av mange kjennetegn som ga de beste resultatene. Klassifikasjonstre håndterer dette greit, mens for logistisk regresjon kan mange variabler være et problem, spesielt når variablene er sterkt korrelerte. Dette reiser spørsmålet om det er hensiktsmessig å lage indekser som fanger opp ulike aspekter ved den enkelte skatteyter, for eksempel basert på en faktoranalyse. En slik analyse av alle de numeriske variablene fra 2008 ga seks forholdsvis klare faktorer knyttet til henholdsvis inntekt, formue, næring, pensjon, jordbruk/skogbruk/fiske og underskudd, i denne rekkefølge med omsyn til å forklare samvariasjonsstukturen i datamaterialet. Ulike vurderinger knyttet til dette pekte i retning av klassifikasjonstre som den mest lovende metoden av de to.

Hva er så et klassifikasjonstre? Jo, en trinnvis splitting av observasjonsenhetene i grupper, der enhetene med unndragelse fremstår klarere i den ene gruppen enn de(n) andre. Dette organiseres som en trestruktur og gir oss beslutningsregler for å lokalisere undergrupper der enheter med unndragelse er i overvekt eller i hvert fall sterkt representert. Reglene etableres ved å splitte de forklarende variablenes verdiområde i henhold til kriterier, som i noen grad kan styres av bruker. Blant fordelene ved et klassifikasjonstre er:

- tillater mange observasjoner og mange variabler
- tillater blanding av numeriske og kategoriske forklarende variabler
- tillater ikke-lineære sammenhenger og manglende observasjoner
- lett å tolke for bruker, både grafisk og numerisk

Dersom den avhengige variabelen er numerisk istedenfor kategorisk, kalles det et regresjonstre, og begge går under fellesbetegnelsen statistisk tre, der en implementering er CART (*classification and regression trees*). I eksemplet vist i tabell 1 og 2 ser treet ut som i figur 1 (laget med pakkene rpart og DMwR i statistikkprogrammet R, se R Development Core Team (2012), Therneau og Atkinson med R-tilpasning av Ripley (2012) og Torgo (2010)).

FIGUR 2 Variablenes «viktighet».



Treet i Figur 1 leses fra toppnoden til de terminale nodene i bunnen, der en kan følge den trinnvise oppsplittingen. Ved hver forgrening står den variabelen det splittes etter, og splitteregelen som et utsagn, der Ja-svar går til høyre og Nei-svar til venstre. I hver boks står det hvor mange personer n som er igjen på dette stadium i splittingen, samt den estimerte sannsynligheten for at en person i denne boksen tilhører FRIV = 1-gruppen (unndragere). Vi ser at de n = 1532 observasjonene først blir splittet ved variabelen STAT, som er formueskatt til Staten. Det optimale splittpunktet er beregnet til å være cirka 1 500 000 kroner. Følger vi den høyre hovedgren med de n = 492 som er over dette beløp, får vi en ny splitt j STAT med splittpunkt ca 8 500 000 kroner. Det er n = 315 individer over dette beløp, og disse splittes ikke ytterligere og klassifiseres til gruppen FRIV = 1 (unndragere). Følger vi den andre grenen med de n = 177 individene med moderat formueskatt til staten, ser vi at vi får splitt i TOPPSKATT, med splittpunkt cirka 1 500 000 kroner, der n = 51 med høyere toppskatt en opp i en terminal node, og klassifiseres til FRIV = 1. For de n = 126 med lavere toppskatt har vi ny splitt i SENT, der de n = 1103 individene med sentralitetskode 3 (sentralt) faller i en egen gruppe og klassifiseres til FRIV = 1, mens de n = 23 øvrige klassifiseres til FRIV = 0 (ærlig). Følger vi den venstre hovedgrenen med de n = 1040 med lav formueskatt til staten, ser vi at de fleste som klassifiseres til FRIV = 0 (ærlig), er i siste node helt til venstre, som omfatter de n = 882 individer som også har lav alminnelig inntekt og liten endring i formuen fra 2007 til 2008.

Slike trær ender altså opp med en struktur der variabler med best prediktiv verdi inngår, og noen variabler utelates, i henhold til kriterier som i noen grad kan styres av brukeren. Variabler som isolert sett har prediktiv verdi, men er sterkt korrelert med en annen variabel av stor betydning som er med, har liten sjanse for å komme med. Vi ser at vi ved å følge splittingen fra toppnoden til de terminale noder som svarer til FRIV = 1 (røde bokser), får en beslutningsregel som kan brukes til å klassifisere nye individer, men de treffsannsynligheter som er angitt, er selvsagt helt urealistiske i situasjonen der unndragere utgjør en svært liten andel.

Ved å spesifisere grupper av interessante variabler kan en lære en hel del av slike trær. Med svært mange variabler kan nok strukturen bli uoversiktlig. Som tidligere nevnt ser det ut til at modeller med mange variabler vil kunne gi de beste klassifikasjonene. Istedenfor svært store modeller er det mer hensiktsmessig å ha flere moderat store modeller og så utnytte disse i kombinasjon. En implementering av dette er såkalte *boosting trees*, som er en vekting av mange ulike trær. På denne måte kan man få med flere forklarende variabler, også noen som har liten forklaringsgrad isolert sett, men som i kombinasjon kan bety noe. I tillegg er man ikke så sårbar for manglende observasjoner som ved modeller der alle variabler spesifiseres for en og samme modell. En ulempe kan være at det ikke utkrystalliseres en enkel, lett forståelig beslutningsregel, men at prediksjonen langt på vei skjer som en «sort boks».

Det kan være ønskelig å ha en rangert liste over de variabler som har betydning for klassifikasjonen, ikke minst i situasjoner der metoden er en «sort boks». Aktuell programvare gir oss typisk dette, basert på en «viktighetsskår». Det er ulik praksis med omsyn til valg av slik skår, om det er betydningen isolert sett eller i kombinasjon med andre variabler man ønsker å få fram. Eksempelvis kan godt en variabel som overhodet ikke er med i treet, tildeles stor betydning. Det er

bare det at den er sterkt korrelert med en variabel som er med i treet. Figur 2 viser en graf for vårt eksempel, laget i R-programmet rpart, som illustrerer nettopp dette. Her er to variabler pekt ut å ha større betydning enn de øvrige, nemlig formueskatt til kommune og formueskatt til stat. Den første av disse er imidlertid ikke med i treet i figur 1. Vi ser også at toppskatt og alminnelig inntekt hjemkommune, som er med i treet, har lavere viktighetsskår enn flere variabler som ikke er med.

Av ovenstående diskusjoner skjønner vi at en må være svært forsiktig med utsagn om hva som betyr noe og ikke betyr noe. Det kan også være en fordel at slik informasjon ikke trer så klart frem at det medfører mistolkninger og i verste fall tilpasninger hos skatteplanleggere. Vi våger likevel å si at blant de kjennetegn som konsekvent dukker opp med de forskjellige metodene, kan nevnes at skatteyter med noen eller flere av egenskapene veldig høy eller null inntekt, høy formue, bor i sentrale strøk, er menn, er gamle eller har en stor økning i formue fra 2007 til 2008, generelt er mer hyppige blant de som har levert frivillig retting enn blant de øvrige.

4. NOEN AVSLUTTENDE MERKNADER

Vi konkluderer med at bruk av denne typen analyser kan gjøre det mulig for Skattedirektoratet å foreta en god kontrollobjektutvelgelse. Den mest treffsikre metoden blant de som er undersøkt, er såkalte *boosting trees*, en metode som utnytter større deler av datamaterialet, også variabler som alene har liten betydning. Metoden er brukervennlig og gir et godt grunnlag for grafisk kommunikasjon. Det kan tilføyes at Skatteetaten allerede har noe erfaring med denne metoden på andre felter. Analysen, som er av en eksplorativ karakter, i motsetning til en analyse som baseres på økonomisk teori, kan, i tillegg til å predikere risiko for at en skatteyter unndrar skatt, brukes til å generere hypoteser som startpunkt for videre studier. Egenskaper som da kan være spesielt interessante å granske mer i detalj, er de som konsekvent dukket opp i våre analyser. I et eventuelt fremtidig arbeid på dette felt vil det være interessant å få analysert data om de skatteytere som Skatteetaten har avslørt gjennom ordinært kontrollarbeid. Dette fordi vi har et problem med skjevt utvalg i frivillig rettinggruppen, det vil si siden disse skatteytere har meldt seg frivillig gjennom skatteamnestiordningen, er de ikke nødvendigvis representative for den generelle gruppen av skatteunndragere. Vi må derfor være forsiktig med å trekke slutninger om kjennetegn hos de skatteyterne som unndrar, men som ikke har meldt seg, basert på de resultatene vi har funnet om dem som har meldt seg frivillig. Det er imidlertid ikke noe i veien for å bruke de kjennetegn som vi funnet her, som hypoteser å teste på nye data. M

Takk til Anders Berset, Torhild Henriksen og Jarle Møen for verdifulle innspill.

REFERANSER

Collins, J.H., V.C. Milliron og D.R. Toy. 1992. Determinants of Tax Compliance: A Contingency Approach. *Journal of the American Taxation Association, 14*: 1–29.

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

Hastie, T., R. Tibshirani og J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. (2 utg.). Stanford, CA: Springer.

Kleven, H., M. Knudsen, C. Thustrup Kreiner, S. Pedersen og S. Saez. 2011. Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark. *Econometrica, 79*: 651–692.

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–00051–07–0, http://www.R-project.org/.

Therneau, T.M. and Atkinson, B... R port by Ripley, B. (2012). *rpart: Recursive Partitioning*. R package version 3.1–52. http://CRAN.R-project.org/package=rpart

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Webley, P., M. Cole og O-P. Eidjar. 2001. The prediction of self-reported and hypothetical tax-evasion: Evidence from England, France and Norway. *Journal of Economic Psychology, 22*: 141–155.