10/12

Rapport
Report

Kjennetegnsanalyser av skattytere som unndrar skatt ved å skjule formuer og inntekter i utlandet

Jonas Andersson
Jostein Lillestøl
Bård Støve

SNF
Et selskap i NHH-miljøet

SAMFUNNS- OG NÆRINGSLIVSFORSKNING AS

Institute for Research in Economics and Business Administration

SNF
Samfunns- og næringslivsforskning AS

- er et selskap i NHH-miljøet med oppgave å initiere, organisere og utføre eksternfinansiert forskning. Norges Handelshøyskole, Universitetet i Bergen og Stiftelsen SNF er aksjonærer. Virksomheten drives med basis i egen stab og fagmiljøene ved NHH og Institutt for økonomi (UiB).

SNF er Norges største og tyngste forskningsmiljø innen anvendt økonomisk-administrativ forskning, og har gode samarbeidsrelasjoner til andre forskningsmiljøer i Norge og utlandet. SNF utfører forskning og forskningsbaserte utredninger for sentrale beslutningstakere i privat og offentlig sektor. Forskningen organiseres i programmer og prosjekter av langsiktig og mer kortsiktig karakter. Alle publikasjoner er offentlig tilgjengelig.

SNF
Institute for Research in Economics and Business Administration

- is a company within the NHH group. Its objective is to initiate, organize and conduct externally financed research. The company shareholders are the Norwegian School of Economics and Business Administration (NHH), the University of Bergen (UiB) and the SNF Foundation. Research is carried out by SNF´s own staff as well as faculty members at NHH and the Department of Economics at UiB.

SNF is Norway´s largest and leading research environment within applied economic administrative research. It has excellent working relations with other research environments in Norway as well as abroad. SNF conducts research and prepares research-based reports for major decision-makers both in the private and the public sector. Research is organized in programmes and projects on a long-term as well as a short-term basis. All our publications are publicly available.

SNF RAPPORT NR 10/12

Kjennetegnsanalyser av skattytere som unndrar skatt ved å skjule formuer og inntekter i utlandet

av

Jonas Andersson
Jostein Lillestøl
Bård Støve

SNF prosjekt nr. 6567
"Skatt som satsingsområde – et forprosjekt"

Prosjektet er finansiert av Skattedirektoratet

SAMFUNNS- OG NÆRINGSLIVSFORSKNING AS
BERGEN, SEPTEMBER 2012

ISBN 978-82-491-0805-3 Trykt versjon
ISBN 978-82-491-0806-0 Elektronisk versjon
ISSN 0803-4036

Innholdsfortegnelse

Summary in English

This project was carried out in the period 2011-2012 by the Institute for research in Economics and Business Administration (SNF) sponsored by the Norwegian Tax Administration.

The goal of this project has been to find characteristics of individual tax payers that have evaded taxes by hiding wealth and/or income overseas / in tax havens. We have considered demographic, economic or socio-economic variables in an effort to classify tax payers according to different risk factors. The aim of this classification has been to enable the Norwegian tax administration (here referred to as the "tax administration") to prioritize their efforts to detect tax evaders. By using appropriate methods for selection of which tax payers to audit, the detection rate of evasion can be increased and audit of honest tax payers be reduced.

An important part to the study has been to evaluate existing methods for pattern recognition. In the report several methods are presented together with their strengths and weaknesses. Methods that are implemented in the statistical software Statistica have been given special attention since this software is used by the tax administration. A number of different classes of methods were tentatively studied and evaluated according to their classification ability and supplementary criteria related to practical use. Logistic regression and classification trees (C&RT) turned out to be the most promising methods. These methods were studied more thoroughly, trying out different variants like LASSO and boosting trees.

The pattern recognition methods are applied to data obtained from the tax administration. Before the authors of this report obtained the data, measures were taken to ensure anonymity of the tax payers. The data set consists of a group of 300 000 private tax payers (the control group), for which no indication of tax evasion exists, and a group of about 600 individuals that have used the opportunity to correct previously incomplete reports of income and wealth (the amnesty group). The corrections have been made voluntarily within an amnesty program[1], i.e. the filers have not been punished for their previous behavior (but have of course been forced to pay the debt owed to the tax administration). Data for around 570 variables have been obtained from tax returns and other registers that the tax administration has access to.

In the analysis statistical models are fitted to different selections of the data. The goal of all the methods is to classify the individual tax payer as either honest or as an evader. Since there are so few tax payers in the amnesty group in relation to the control group, the problem somewhat resembles that of finding a needle in a haystack. This problem is discussed in the report. Many of the methods can nevertheless with decent precision, classify the individuals to the correct group. In the table below we report the error rate (i.e. the amount of misclassification) for an analysis with 1000 individuals in the control group. The comparison should be read as a relative comparison between the methods, and

[1]The amnesty program was initiated by the tax authority in 2009, and in total over 700 individuals has reported hidden wealth or income. In the last three years between three and four billion NOK has been reported as hidden wealth. The money has mainly been hidden in Switzerland or Luxemburg.

not as the obtainable error rates in the context of all taxpayers. In this respect, the tree-based methods stand out as the best.

Results from one analysis

	In-sample		Out-of-sample	
Model	**Falsely predicted as evader**	**Falsely predicted as honest**	**Falsely predicted as evader**	**Falsely predicted as honest**
Logistic regression	3 %	20 %	3 %	31 %
Non-parametric log. reg.	2 %	21 %	3 %	26 %
Logistic reg. LASSO	2 %	30 %	1 %	33 %
C&RT	9 %	12 %	12 %	20 %
Boosted tree	2 %	8 %	4 %	19 %

Although there is some evidence that certain particular characteristics contain predictive power of which group an individual belong to, the main result is that it is methods using relatively complex combinations of them, like boosted trees, that are most useful. However, some variables seem to be unavoidable in order to produce good predictions. Examples of such variables are income, wealth, gender, age and whether a tax payer lives in a densely populated area. Tax payers with one or several of the following features; very high or zero income, high wealth, being a man, being old, living in a densely populated area are more frequently observed among those in the amnesty group.

For practical use, one would have to apply a fitted model to a list of random tax payers, and the model will then classify these tax payers as either honest or possible tax evader. The tax payers classified in the evader group should then be audited, and based on this study we will expect that the detection rate of evasion will be increased compared to random auditing of tax payers.

We suggest studying other data, e.g. data on tax payers that have been caught in regular audits, in order to see whether the results in the present study are possible to generalize to situations where the tax payers have not voluntarily corrected previous underreporting.

Sammendrag

Målet med prosjektet har vært å finne kjennetegn ved personlige skattytere 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 med analysene er finne ut om det er mulig å gruppere skattyterne ut fra risikofaktorer slik at Skatteetaten kan sette inn riktig virkemiddel mot de ulike risikogruppene. Ved bruk av adekvate statistiske metoder for bruk til kontrollobjektutvelgelsen kan avsløring av unndratt skatt økes og antall kontroller uten avsløringer reduseres.

En viktig del av arbeidet har bestått i å evaluere tilgjengelige analysemetoder. I rapporten presenteres flere analysemetoder relativt grundig og vi gjennomgår styrker og svakheter ved de ulike metodikkene. Vi har fokusert på metoder som i hovedsak er tilgjengelig i programvaren Statistica, som er et analyseverktøy Skatteetaten benytter internt. En rekke ulike klasser av metoder er forsøksvis studert og evaluert ut fra deres klassifikasjonsevne og supplerende kriterier for praktisk bruk. Logistisk regresjon og klassifikasjonstrær (C&RT) viste seg å være de mest lovende metodene. Disse metodene ble undersøkt mer grundig, der ulike varianter som LASSO og "boosting trees" ble utprøvet.

De gjennomgåtte metodene for kjennetegnsanalyse er illustrert på anonymiserte data mottatt fra Skatteetaten. Datasettet består av et tilfeldig utvalg av 300 000 personlige skatteytere, kalt kontrollgruppen og ett utvalg av rundt 600 skatteytere som har unndratt skatt, kalt frivillig retting gruppen. Dette er skatteytere som har meldt seg frivillig gjennom skatteamnestiordningen. I overkant av 570 ulike variable på den enkelte skatteyter er mottatt. Dette er data fra ulike felter fra den enkelte skatteyters ligning og data fra andre databaser som Skatteetaten har tilgjengelig.

I analysene tilpasses en statistisk modell til ulike utvalg av data både fra kontrollgruppen og frivillig retting gruppen. Alle disse modellene har som mål å klassifisere den enkelte skatteyter som enten ærlig (kontrollgruppen) eller skatteunndrager (frivillig retting gruppen). Siden det er så få skatteytere i frivillig retting gruppen relativt til kontrollgruppen kan en passende beskrivelse av problemet være 'å lete etter nålen i høystakken'. Dette problem diskuteres i rapporten. Imidlertid klarer mange av metodene med relativt stor grad av treffsikkerhet å klassifisere de mottatte data i riktig gruppe. I tabellen nedenfor rapporteres feilratene for en analyse av et utvalg med 1000 individer i kontrollgruppen. Sammenligningen skal ses som en relativ sammenligning mellom metodene, og ikke som realistiske feilrater i en analyse av alle skatteytere. I denne kontekst fremstår metodene som baseres på statistiske trær som de beste.

Resultater fra en analyse

	In-sample		Out-of-sample	
Model	**Feilaktig klassifisert som unndrager**	**Feilaktig klassifisert som ærlig**	**Feilaktig klassifisert som unndrager**	**Feilaktig klassifisert som ærlig**
Logistisk regresjon	3 %	20 %	3 %	31 %
Ikke-parametrisk log. reg.	2 %	21 %	3 %	26 %
Logistisk reg. LASSO	2 %	30 %	1 %	33 %
C&RT	9 %	12 %	12 %	20 %
Boosted tree	2 %	8 %	4 %	19 %

Riktignok viser resultatene totalt sett at det er visse kjennetegn som er forskjellig for de to gruppene, og disse kjennetegnene kan benyttes til en gruppering av skatteytere. Men som regel vil det være metoder med kombinasjoner av mange kjennetegn, slik som "boosted trees", som gir de beste resultatene, dvs. det er relativt komplekse modeller med mange kjennetegn som gir en mest riktig klassifisering. Det er imidlertid noen variabler som ser ut til å være uunnværlige for å gjøre en god klassifikasjon. Eksempler på disse er inntekt, formue, kjønn, alder og hvorvidt en skatteyter bor i sentrale strøk. Skatteytere med en eller flere av følgende karakteristikk; meget høy eller null inntekt, stor formue, er mann, er gammel, bor i sentrale strøk har en relativt høyere frekvens i gruppen som levert frivillig retting.

Vi konkluderer med at basert på resultatene fremkommet i rapporten, kan bruk av de mer komplekse modellene gjøre det mulig for Skatteetaten å foreta en bedre kontrollobjektutvelgelse. Dette må skje i to steg; først tilpasse en modell til et utvalg av skatteytere som vi vet faller i to grupper, en ærlig gruppe og en unndrager-gruppe. Deretter kan denne modellen benyttes på et nytt utvalg av skatteytere som vi ikke vet er ærlig eller unndrar. Modellen vil så klassifisere dette nye utvalget av skatteytere, og de skatteytere som blir klassifisert som unndragere bør prioriteres for kontroll.

Vi oppfordrer imidlertid til å gjennomføre ytterligere studier for å få bekreftet disse funnene. Da ville det spesielt vært interessant å studere de personlige skatteytere som har blitt avslørt gjennom vanlige kontroller i Skatteetaten, og ikke de som har meldt seg frivillig gjennom skatteamnestiordningen. Da det kan tenkes at de skatteytere som har meldt seg frivillig gjennom amnestiordningen har andre kjennetegn enn de skattytere som unndrar skatt eller formue og som ikke melder seg frivillig.

1. Innledning

I dette kapittelet definerer vi kort prosjektet og dets mål, samt hvordan prosjektet er gjennomført. Rapporten er videre organisert som følger: I kapittel 2 presenteres kort deler av den akademiske litteraturen på området, i kapittel 3 presenteres kort hvilket datagrunnlag som er mottatt fra Skatteetaten, i kapittel 4 rapporteres resultatene fra noen enkle innledende beskrivende analyser. Kapittel 5 er en generell metodeoversikt, som avslutningsvis peker ut de mest aktuelle metodene. Dette kapittelet kan hoppes over for de som ønsker å gå direkte til resultatene. I kapittel 6 gjøres de første klassifiseringsanalyser med flere metoder, som munner ut i to prioriterte metoder. Kapittel 7 og 8 presenterer de prioriterte metodene for seg, der målet er å presentere metodene og utfordringene ved disse. I kapittel 9 presenteres de analysene som tar sikte på å gjøre det best mulig med alle variable tilgjengelig. Kapittel 10 presenterer våre konklusjoner og betraktninger rundt videre muligheter.

Forskergruppen fra SNF har bestått av førsteamanuensis Bård Støve, professor Jonas Andersson og professor Jostein Lillestøl. Som en del av prosjektet har forskergruppen også deltatt med foredrag/innlegg på en rekke seminarer og workshops arrangert av Skatteetaten. [2]

1.1 Prosjektdefinisjon

Norske skatteytere skjuler store verdier i utlandet/skatteparadiser[3]. Skattedirektoratet (SKD) ønsker derfor å undersøke hvilke kjennetegn som karakteriserer slike skatteytere. Dette for å sikre en mer målrettet bruk av virkemidler, og øke oppdagelsessannsynligheten for unndragelse. Dette prosjektet er et første steg for å etablere et forskningssamarbeid mellom SNF og Skattedirektoratet innenfor kjennetegnsanalyser. Videre er dette temaet også interessant i et akademisk perspektiv.

Målet med prosjektet er å finne kjennetegn ved personlige skattytere 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 med analysene er finne ut om det er mulig å gruppere skattyterne ut fra risikofaktorer slik at Skatteetaten kan sette inn riktig virkemiddel mot de ulike risikogruppene. Ved å utvikle statistiske modeller for bruk til kontrollobjektutvelgelsen kan avsløring av unndratt skatt økes og antall kontroller uten avsløringer reduseres.

For å avdekke skatteunndragelser via skatteparadiser, har Skatteetaten igangsatt flere prosjekter der det mest sentrale tiltaket er kontroll av utenlandske betalingskort brukt i Norge (betalingskortprosjektet). I tillegg ber stadig flere skattytere om såkalt skatte-amnesti/frivillig retting, slik at de kan oppgi tidligere skjulte utenlandsformuer til skattemyndighetene uten å måtte betale tilleggsskatt.

[2] Vi takker Anders Berset, Torhild Henriksen, Jarle Møen og Guttorm Schjelderup for verdifulle innspill.

[3] Skatteetatens egne beregninger anslår pengeplasseringer i utlandet tilhørende norske personlige skatteytere og virksomheter på om lag 50 – 130 MRD NOK (referanse – presentasjon ved NHH 21. juni 2009). Clotfelter (1983) anslår videre at 20-58 % av amerikanske skattebetalere unndrar skatt (gjelder alle former for unndragelse)

Betalingskortprosjektet og frivillig retting genererer en del informasjon som er benyttet som innputt i kjennetegnsanalysene. Det er imidlertid utfordringer knyttet til å utnytte datagrunnlaget fra disse to tiltakene i forskningssammenheng. I dette prosjektet gjennomføres derfor en begrenset pilotstudie for å se i hvilken grad det er mulig å gjennomføre mer omfattende kjennetegnsanalyser på dette grunnlaget.

Ved siden av å prøve å finne kjennetegn som nevnt over, vil vi også gjennomgå og diskutere styrker og svakheter ved de ulike metoder for kjennetegnsanalyser. Flere av metodene vil så benyttes på de data som forskergruppen har mottatt.

Vi forstår at dersom denne pilotstudien gir positive resultater/funn, vil Skatteetaten arbeide for at det etableres et hovedprosjekt for å analysere et større datasett, for eksempel også undersøke karakteristika for selskaper.

Oppdragsbeskrivelsen var:

1. Studere våre to populasjoner vs. kontrollgruppen
2. Vurdere hvilke variable som initielt bør være en del av kjennetegnsanalysen, basert på pkt. 1 og litteraturstudier.
3. Gi en oversikt og gjennomgang av de ulike metoder som eksisterer
4. Benytte en rekke metodikker innen kjennetegnsanalyse (klassifikasjon), samt vurdere styrker og svakheter av dem

Det er to hovedaspekter ved oppdraget:

- Finne forklaringsvariable: Blinke ut spesielle variable (kjennetegn) som ser ut til å være bærere av informasjon om de individer som har en tendens til skatteunndragelse, og som kan skille dem fra de øvrige. Dette for å kunne koble sammen den empiriske analyse med de erfaringer som skattedirektoratet har fra før. Disse variablene bør gi god mening saklogisk.

- Predikere skatteunndragere: Etablere en modell basert på et sett av forklaringsvariable som kan sannsynliggjøre at et individ har, for Skatteetaten, skjult inntekt og formue. Aktuelle variable er de man har funnet ovenfor, men også andre variable er aktuelle, som ikke i samme grad kan tolkes saklogisk. Formålet er å kunne skille skatteunndragere fra de øvrige (klassifisere) best mulig, og i beste fall kunne beregne sannsynligheter for unndragelse på grunnlag av gitte variabelverdier på grunnlag av modellen.

1.2 Muligheter for generalisering

Den empiriske delen av studien er en såkalt observasjonsstudie, med de begrensninger dette innebærer i forhold til et kontrollert eksperiment. De omtrent 600 individene som har levert frivillig retting av sin selvangivelse er ikke utvalgte i henhold til noen tilfeldig prosess men har, hvilket navnet på tiltaket tilsier, frivillig valgt å levere en korrigering av sin tidligere selvangivelse. Dette har to konsekvenser for vårt arbeid med å finne forskjeller i mønster mellom individer som levert en frivillig retting og de som ikke gjort det. Den første konsekvensen er at det vanskeliggjør en evaluering av om en variabel forklarer sannsynligheten for at et individ tilhører de som bør gjøre frivillig retting. For å

gjøre en slik evaluering må man ha kontroll på hvor stor sannsynligheten er for at et individ kommer med i utvalget, hvilket vi altså ikke har i dette tilfellet. Dette innebærer ikke at det er umulig å trekke konklusjoner fra datamaterialet, men bare at påstander av typen "variabelen er statistisk signifikant" må unngås eller gjøres med stor forsiktighet. Prediksjoner av sannsynlighet til gruppetilhørighet er imidlertid ikke hemmet på samme måte av dette problem.

Den andre konsekvensen er at det spesielle utvalget gjør en eventuell generalisering til fremtidlige situasjoner vanskeligere enn det ville være i en eksperimentsituasjon. Dette ettersom man ved et eksperiment kjenner hvilken populasjon man har tatt utvalget fra. I den foreliggende studien er dette mer spekulativt, da vi har en miks av samtlige individer som har levert frivillig retting og et tilfeldig utvalg fra alle skattebetalere.

I tilknytning til dette vil vi poengtere at begrepet "skatteunndrager" og "person som har levert frivillig retting" brukes synonymt i fortsettelsen av denne rapporten. Dette reflekterer et annet aspekt av utvalgsproblematikken, nemlig at vi ikke observerer alle som unndrar skatt men kun en delmengde av disse. Antagelsen at de individer som ikke har levert frivillig retting alle er "ærlige" er sikkert ikke korrekt, noe som betyr at kontrollgruppen blir kontaminert, uvisst i hvilken grad. Vi modellerer altså i realiteten sannsynligheten for innlevering av frivillig retting og ikke sannsynligheten for skatteunndragelse. På den andre siden finnes her en mulighet for å fingranske de individer i kontrollgruppen som de brukte metodene predikerer å tilhøre frivillig rettingsgruppen.

Det finnes også et tidsaspekt her som ikke er med i studien. Er individer som har levert frivillig retting i år 2010 systematisk forskjellige fra de som gjorde det år 2008? Det kan være slik at bedømmelsen disse gjorde om sannsynligheten å bli oppdaget er blitt forandret over tid. Dette er ytterligere en heterogenitet i datamaterialet som kan være et problem fra et metodisk ståsted, men som på den andre siden også kan gi en mulighet til å predikere fremtidlig adferd.

Videre er det et problem for analysene at vi ikke har fått opplysninger om når en skatteyter har levert en frivillig retting. Det som er interessant er nemlig verdien på kjennetegnene året før skatteyteren leverte frivillig retting. Vi har i studien ikke tatt hensyn til dette aspektet, og dette vil kunne gi opphav til en feilkilde i analysene. Imidlertid vil selvsagt noen kjennetegn være konstante over tid, for eksempel kjønn.

Konklusjonen her er at studier som denne kan brukes til å generere hypoteser (idéer) til hva slags individer som unndrar inntekter og formue i sin selvangivelse, men ikke til å teste hypoteser. For å gjøre slike tester må data samles inn på en mer kontrollert måte slik at utvalgsannsynligheten er kjent på forhånd. De dataene vi har analysert var ikke primært innsamlet for vår analyse, men som et biprodukt av muligheten til frivillig retting.

1.3 Programvare

Det finnes mange ulike programmerpakker for denne type analyser. Avhengig av hvem som skal gjøre analysene fins ulike kriterier som er viktige ved et slikt valg. Vi mener at det her gjelder å primært tenkte på følgende spørsmål:

Skal pakken primært brukes i ”produksjonen”, det vil si for å gjøre prediksjoner av sannsynligheten for skatteunndragelse i praksis? Eller er det viktig å kunne utvikle analysene med nye aspekter i form av andre modeller og/eller variabler?

I det første tilfellet er det viktig med et brukergrensesnitt som er lett å bruke. I det andre tilfellet er det viktig at et slikt brukergrensesnitt ikke står i veien for utvikling av analysene. For SKD's vedkommende ligger sannsynligvis hensikten et sted mellom disse to ytterligheter. Vi har derfor brukt to ulike programmer, Statistica og R.

Statistica er, slik vi ser det, et godt kompromiss mellom de to hensiktene og brukes allerede av SKD. Det er mulig å utvikle nye metoder også i Statistica med en variant av Visual Basic som er tilgjengelig i programmet, men fordelen med dette program er dets elegante brukergrensesnitt. Når det gjelder fleksibilitet og helt nye metoder har vi imidlertid en preferanse for R, som er en såkalt open-source programpakke. Man har derved alltid adgang til den siste versjonen av programmet og alle de nyeste metodene som finnes i det. Et voksende antall statistikere bruker dette program og legger ut R-kode til sine nyutviklede metoder, hvilket gjør at SKD i fremtiden raskt kan få adgang til nye mønstergjenkjenningmetoder, hvis man tilegner seg kunnskap i å bruke R.

I tillegg er det et fullverdig programmeringsspråk der fremtidlige nye aspekter av det foreliggende problemet kan inkorporeres lettere, mener vi, enn i Statistica. Det skal imidlertid poengteres at det tar noe lengre tid i å lære seg å bruke R effektivt i sammenligning med Statistica.

1.4 Gjennomføring

Forskergruppen har hatt jevnlige (video)møter med SKD underveis i prosjektet, i tillegg har forskergruppen deltatt på flere workshops/seminarer arrangert av Skatteetaten/SKD.

Hvilke metoder det er aktuelle å teste ut datagrunnlaget på, har vært gjenstand for diskusjon. Det vises for øvrig til SNFs spesifikasjon av oppdraget.

Analysene er i hovedsak utført i Statistica, men noen analyser er også utført i R. Vi har valgt å bruke variabelnavn slik de fremkommer i Skatteetatens datafiler, selv om de bare til en viss grad er selvforklarende. Likedan har vi brukt utskrifter slik de fremkommer i programvaren, selv om disse ikke alltid er det mest estetisk tiltalende. Grunnen er gjenkjenningseffekten hos brukene i Skatteetaten.

2. Eksisterende kunnskap

I dette kapitlet gir vi en kort oversikt av den akademiske litteraturen om skatteunndragelse. Den første teoretiske modellen som analyserte skatteunndragelser ble utviklet av Allingham & Sandmo (1972), og baserer seg på Beckers (1968) teori, som sier at rasjonelle individer vrir sin adferd i retning av handlinger som gir høy avkastning. Senere teoretisk forskning er i hovedsak justering og utvidelser av Allingham & Sandmos modell. Vi viser til Skatteunndragelsesutvalget (NOU 2009:4) for ytterligere referanser.

Flere empiriske studier fra ulike land har undersøkt sosio-økonomiske og andre kjennetegn for å predikere skatteunndragelse, samt benyttet spørreundersøkelser for å avdekke holdninger rundt skatt/selvangivelse. Se for eksempel Lee & Carley (2009), Collins et al (1992), Clotfelter (1983) og Webley et al (2001). Kort oppsummert er funnene som følger:

- Kjønn: menn unndrar oftere enn kvinner
- Alder: det er mindre sannsynlig at eldre unndrar skatt (grunnet økt risikoaversjon)
- Utdanning: reduserer tilbøyeligheten for unndragelse
- Inntekt: lav inntekt og veldig høy inntekt gir større sannsynlighet for unndragelse (grunnet økt mulighet for svart inntekt eller grunnet høy kapitalinntekt)
- Profesjonell støtte (revisor/jurist etc.) ved selvangivelse (kompleksitet) kan gi unndragelser (kreative fradrag mv.)
- Sosialt nettverk: om nettverket (kollega/bransje mv.) unndrar, høyere sannsynlighet for unndragelse
- Selvstendig næringsdrivende: større muligheter for unndragelser (svart arbeid mv.)

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

En rekke andre studier behandler skatteunndragelse, bl.a. Feinstein (1991) tar i sin empiriske studie på en finurlig måte høyde for at ikke alle skattesnytere blir oppdaget. Dette gjøres gjennom å modellere oppdagelsessannsynlighet som en funksjon av revisorenes kompetanse. En dansk studie av Kleven et al (2011) viser med et naturlig eksperiment at unndragsannsynligheten for tredjehåndsrapporterte, i forskjell til selvrapporterte, oppgaver omtrent er null. I tillegg viser de at trussel om revisjon har en betydlig effekt på hvordan selvrapporteringen gjøres. Engström & Holmlunds (2009) studie viser at selvstendig næringsdrivende sannsynligvis i viss utstrekning bruker de større muligheter til skatteunndragelse som disse har. Det er også flere studier som går på mer rendyrkede sosiologiske og psykologiske forklaringer. Frey & Feld (2002) kommer frem til at skatteetatens adferd mot skatteytere har noe å si på tendensen til skatteunndragelse. Videre viser Torgler (2002), med data på tvers av land, at høy grad av demokratiske rettigheter gir en mindre tendens til skatteunndragelse.

3. Datagrunnlag

Fra Skatteetaten har SNF mottatt tre datasett:

1. Kontrollgruppen bestående av 10 % av skatteytermassen (dvs. 300 000 private skatteytere)
2. Frivillig retting (601 private skatteytere, hvorav 24 manglet vesentlig informasjon, slik at 577 er relevante)
3. Betalingskortsaker (6 private skatteytere)

For skatteyterne i hvert av de tre datasettene, har vi mottatt i overkant av 570 variable. Variablene omfatter informasjon fra selvangivelsen, samt andre kjennetegn. Listen over alle variablene er i et appendiks som utelates i dette offentlig tilgjengelige dokumentet. Alle data er anonymisert. Vi har fått dataene i to leveranser, der siste leveranse inkluderer flere variable og noen flere frivillig retting saker. Dataene er hentet både fra Skatteetatens interne databaser, men også andre databaser. For en del individer er det manglende data for ulike variable, men så lenge variasjonen er tilstrekkelig i disse variablene utgjør dette ikke et stort problem for analysene. Se imidlertid kommentarer til enkelte metoder vedrørende dette momentet.

Imidlertid er det også en rekke variable som nesten ikke inneholder noe informasjon, for eksempel er noen kun registrert med verdi 0 på alle individer. Disse variablene er derfor utelatt fra analysene.

Da betalingskortsakene utgjør kun 6 observasjoner, egner dette ikke seg til videre analyser. Vi anbefaler imidlertid å se på disse dataene om flere observasjoner blir tilgjengelig i fremtiden.

Kontrollgruppen representerer et tilfeldig utvalg av private skattytere i Norge. For analysene vil vi anta at disse skatteyterne er lovlydige, men med all sannsynlighet vil det også blant disse være skatteytere som unndrar skatt. Dette vil vi imidlertid ikke ta hensyn til i de senere analyser i dette notatet.

Frivillig retting gruppen består av skatteytere som frivillig har kommet med opplysninger om skatteunndragelser. I Norge er det et "løpende" skatteamnesti (kun èn gang for hver skatteyter) jfr. ligningsloven § 10-3 nr 2 bokstav c. Skattyterne som får amnesti må betale den skatten som ordinært hadde blitt beregnet på formuen/inntekten, men vedkommende slipper tilleggskatt og/eller anmeldelse. Satsingen på å avdekke skjulte midler i skatteparadiser startet for alvor opp i 2009, og de siste årene har ordningen vært populær. Totalt har ca 700 skatteytere meldt seg siden 2009, og for dette prosjektet har vi mottatt flesteparten av disse. Imidlertid er det viktig å bemerke at vi har dermed med et skjevt utvalg å gjøre, i og med at disse skatteytere som har meldt seg frivillig, ikke nødvendigvis er representative for gruppen av skatteunndragere. Vi må derfor være forsiktig med å trekke slutninger om de skatteyterne som unndrar, men som ikke har meldt seg, basert på de resultatene vi finner om de som har meldt seg frivillig. Det vil derfor være interessant å få analysert data om de skatteytere som Skatteetaten har avslørt gjennom ordinært kontrollarbeid i et fremtidig arbeid.

Det har vært et visst arbeid med å klargjøre dataene for analysene. Vi har allerede vært inne på at det var nødvendig å utelate noen variable, fordi disse inneholdt kun 0 verdier for alle skatteytere. Videre har det vært nødvendig å få hver enkelt av variablene definert med riktig dataformat (kategori, kontinuerlig mv.). Dette er spesielt viktig ved bruk av Statistica, da flere av funksjonene ikke vil kjøre

om programmet oppdager at en bestemt type datakolonne har et dataformat som ikke samsvarer med hva Statistica forventer.

For oversiktens skyld definerer vi under de ulike typer data vi arbeider med:

- Numeriske variable: med verdier på en (ubegrenset) måleskala
  - f.eks. antall og beløp
- Kategorivariable: med kodetall for et endelig antall mulige kategorier
  - f.eks. statsborgerskap
- Dikotome variable (indikator eller dummyvariabel): to grupper (gjerne kodet hhv. 0 og 1)
  - f.eks. kjønn

4. Frivilling retting: Innledende analyser

I dette kapittelet rapporteres noen enkle deskriptive analyser for noen av variablene og kjennetegnene til hver av gruppene frivillig retting og kontroll. Formålet er begrenset, bare å gi en første innsikt i forskjeller mellom de to gruppene. Videre ser vi på noen problemer knyttet til variabelutvelgelse, spesielt når variable har høy korrelasjon (høy samvariasjon).

4.1 Deskriptiv statistikk

Tabell 1 og 2 gir beskrivende mål for syv utvalgte variable for hhv. kontrollgruppen og frivillig retting gruppen. De beskrivende målene er hhv. gjennomsnitt, median, minimum, maksimum og standardavvik.

Tabell 1: Deskriptiv statistikk for kontrollgruppen

Variabel	Mean	Median	Minimum	Maximum	Std.Dev
FODSELAAR	1962	1964	1901	1991	18
PERSON_INNTEKT_LONN	233983	192000	0	13 500000	268454
NETTO_FORM_STAT	369013	0	0	2183692000	6095855
TOPPSKATT	4638	0	0	1562000	20228
ANT_BILER	1	1	1	91	1
ANT_LEDE_NEST_MEDL	2	1	1	447	4
ANT_KJOP_SALG	18	2	1	3811	87

Tabell 2: Deskriptiv statistikk for frivillig retting gruppen

Variabel	Mean	Median	Minimum	Maximum	Std.Dev
FODSELAAR	1945	1944	1911	1988	14
PERSON_INNTEKT_LONN	341454	20000	0	9648000	720521
NETTO_FORM_STAT	19560622	2890000	0	1623673000	88750172
TOPPSKATT	29419	0	0	1099000	82251
ANT_BILER	1	1	1	7	1
ANT_LEDE_NEST_MEDL	4	2	1	52	7
ANT_KJOP_SALG	17	2	1	508	54

Fra tabellene ser vi at gjennomsnittsalderen, gjennomsnittsinntekten, gjennomsnittsformuen og gjennomsnittlig toppskatt er betydelig høyere for frivillig retting gruppen enn for kontrollgruppen, mens det ikke er påfallende store forskjeller for de øvrige tre variable.

Vi kan se i mer detalj hvordan de to gruppene skiller seg fra hverandre ved å sammenligne frekvenstabeller, som for de numeriske variable er basert på grupperte tall. Tabellene 3 og 4 viser i

mer detalj aldersfordelingen i de to gruppene. Vi ser klart at det er flere eldre i frivillig retting gruppen sammenlignet med kontrollgruppen.

Tabell 3: Fordeling av alder i kontrollgruppen

Alder	Andel (%)	Kumulativ andel (%)
FØDSELAAR<=1935	9.5	9.5
FØDSELAAR<=1945	9.5	19.1
FØDSELAAR<=1955	15.2	34.3
FØDSELAAR<=1965	17.4	51.7
FØDSELAAR<=1975	19.4	71.1
FØDSELAAR<=1985	17.3	88.3
FØDSELAAR>1985	10.0	98.3

Tabell 4: Fordeling av alder i frivillig retting gruppen

Alder	Andel (%)	Kumulativ andel (%)
FØDSELAAR<=1935	23.3	23.3
FØDSELAAR<=1945	27.3	50.6
FØDSELAAR<=1955	22.5	73.1
FØDSELAAR<=1965	13.8	86.9
FØDSELAAR<=1975	8.2	95.1
FØDSELAAR<=1985	2	97.1
FØDSELAAR>1985	0.2	97.3

Fra tabell 5 ser vi at i kontrollgruppen er det omtrent like mange menn som kvinner, mens i frivillig retting gruppen er mennene klart i flertall. Vi noterer også at vi mangler data om dette på 17 individer i frivillig retting gruppen.

Tabell 5: Kjønnsfordeling

Kjønn	Andel i kontrollgruppen (%)	Andel i friv. rett. gruppen (%)
Menn	48.9	60.2
Kvinner	51.1	36.9

Tabell 6 viser fordelingen i bomønsteret til skatteyterene i de to gruppene. Her er kode 0 i spredte strøk, mens kode 3 er mest sentrale strøk. Vi ser at det er klart flere bosatt i de sentrale strøk i frivillig retting gruppen sammenlignet med kontrollgruppen.

Tabell 6: Bomønster

SENTRALITETSKODE	Andel i kontrollgruppen (%)	Andel i friv. rett. gruppen (%)
0	9.9	1.7
1	6.4	1.5
2	17.0	4.8
3	65.6	88.0

Tabell 7 viser inntektsfordelingen til våre to grupper. Det er tydelig at det er flere som ikke har personinntekt i frivillig retting gruppen, samtidig som det er flere som har høye lønninger (over 750 000) i denne gruppen, sammenlignet med kontrollgruppen.

Tabell 7: Inntektsfordeling

Inntekt	Andel i kontrollgruppen (%)	Andel i friv. rett. gruppen (%)
PERSON_INNTEKT_LONN<=0	27.1	40.9
PERSON_INNTEKT_LONN<=150000	19.1	16.6
PERSON_INNTEKT_LONN<=250000	9.9	4.8
PERSON_INNTEKT_LONN<=350000	14.2	3.0
PERSON_INNTEKT_LONN<=450000	13.9	6.0
PERSON_INNTEKT_LONN<=750000	12.6	11.5
PERSON_INNTEKT_LONN >750000	3.3	13.1

Tabell 8 viser formuesfordelingen til de to gruppene. Vi ser at det er klart en mye større andel som har høy formue i frivillig retting gruppen sammenlignet med kontrollgruppen.

Tabell 8: Formuesfordeling

Formue	Andel i kontrollgruppen (%)	Andel i friv. rett. gruppen (%)
NETTO_FORM_STAT<=0	52.3	12.1
NETTO_FORM_STAT<=250000	23.4	4.3
NETTO_FORM_STAT<=500000	8.6	2.3
NETTO_FORM_STAT<=1000000	8.1	9.2
NETTO_FORM_STAT<=2000000	4.7	11.0
NETTO_FORM_STAT>2000000	2.8	57.1

Basert kun på disse enkle beskrivende analyser ser vi at vi har fått bekreftet flere av de allerede kjente kjennetegnene fra litteraturen, bl.a. at det er de med svært høy inntekt og svært lav inntekt (spesielt 0 inntekt) som har meldt seg frivillig (har vært skatteunndragere).

For nærmere å karakterisere forskjellene mellom våre to grupper, kontrollgruppen og frivillig retting gruppen, har vi beregnet absolutte t-kvoter og såkalt Hellingeravstand for respektivt de numeriske og

de kategoriske variablene. En stor verdi skal tolkes som at det er stor forskjell i fordelingen av en variabel mellom de to gruppene.

t-kvoten er beregnet fra formelen

$$t = \frac{\bar{x}_K - \bar{x}_F}{\sqrt{\frac{s_K^2}{n_K} + \frac{s_F^2}{n_F}}}$$

der $\bar{x}$, s og n med indeks K (kontroll) og F (frivillig retting) representerer gjennomsnitt, standardavvik og antall observasjoner i de to gruppene. Hellingeravstand for de kategoriske variablene beregnes som

$$H = \sqrt{1 - \sum_i \sqrt{\hat{p}_i^K \hat{p}_i^F}}$$

der $\hat{p}_i^K$ og $\hat{p}_i^F$ er de observerte andelene i kategori i for den aktuelle kategoriske variabelen for kontrollgruppen og frivillig retting gruppen.

Disse beregningene for hver variabel kan rapporteres i to tabeller, en for num,eriske variable og en for kategorivariable. Her kan man lete etter variable med store verdier, som dermed indikerer store forskjeller mellom gruppene. Tabellene kan hensiktsmessig organiseres etter fallende verdier på det bergende mål. Slike tabeller er ikke med i dette offentlig tilgengelige dokument, men vi noterer her at mange av de variablene som er nevnt ovenfor i dette kapittelet forekommer høyt opp på listen i disse tabellene. Slike tabeller har imidlertid noe begrenset verdi som grunnlag for å velge ut variable til de aktuelle klassifikasjonsmodellene.

4.2 Variabelutvelgelse

Blant variablene finnes flere naturlige grupper: Formuesvariable, inntektsvariable, pensjonsvariable osv. Variable innen hver gruppe vil kunne ha samme informasjonsinnhold som andre variable innen gruppen. Dette vil typisk medføre at variablene er sterkt korrelerte, såkalt kolinearitet. Dette vil kunne ha ulike konsekvenser, alt etter hvilken analysemetode man har valgt. For kategorisk regresjon med flere variable fra samme gruppe vil det kunne medføre ustabilitet, og økt usikkerhet ved kategorisering og prediksjon. Man må derfor ha dette for øye ved utvikling av regresjonsmodellen, enten ved forutgående diagnoser eller en systematisk prosess for variabelutvelgelse. Slike fins i form av såkalt trinnvis regresjon, som fins i ulike varianter: forlengs inkludering, baklengs ekskludering og i kombinasjon, såkalt full trinnvis regresjon. Dette innebærer at en ny variabel innlemmes/fjernes bare hvis den tilfører noe nytt i forhold til de variable som er i modellen fra før. En slik prosess kan i ulik grad gjøres brukerstyrt. For metoder av typen statistisk tre, som vi senere skal komme til, vil en slik trinnvis utvelgelse være en sentral del av selve metoden, dvs. at dersom en ny variabel kommer til fra en gruppe av variable, er denne på sett og vis den beste representanten for gruppen, og øvrige variable i gruppen kommer typisk bare med senere dersom de(n) tilfører noe mer, etter at andre enkeltvariable eller variable fra andre grupper er kommet med. På dette vis omgår man her kolinearitetsproblemet.

4.3 Faktoranalyse og indekser

I situasjoner der man har naturlige grupper av variable, eller at ulike variable kan sies å reflektere en felles underliggende tilbøyelighet eller styrende adferdsfaktor, kan det være tjenlig å prøve å oppsummere dette i en konstruert variabel, f.eks bruke gjennomsnittverdien eller en veid sum av variablene i gruppen. På dette vis kan i noen tilfeller få med seg informasjonsinnholdet i gruppen som helhet, uten å måtte velge ut en enkelt variabel som representant for denne. En slik konstruert variabel kaller vi en *indeks.* En måte å avdekke grupper av variable og eventuelt underliggende (latente) adferdsvariable på, er såkalt *faktoranalyse,* som også kan gi grunnlag for etablering av indekser.

Hovedideen i faktoranalyse er at en forestiller seg at hver variabel kan uttykkes som en lineær kombinasjon av en rekke underliggende uobserverbare og uavhengige faktorer, såkalte *latente variable*. Dataanalytisk innebærer det at vi i prinsippet kan reprodusere sentrale trekk ved datamaterialet med færre variable, helst så få som mulig. En beregnet faktorløsning gir, for hver variabel, de såkalte *faktorladningene* på hver av faktorene, som gir uttrykk for hvor mye variabelen avhenger av akkurat den faktoren. En faktorløsning er ikke entydig, og man får ofte en løsning som er lettere å tolke ved såkalt rotasjon. Analysen bidrar også til å finne ut hvor mange underliggende (latente) faktorer det er rimelig å ta i betraktning. Det fins en rekke ulike algoritmer for faktoranalyse, med ulike beregningskriterier og form på løsning (urotert, rotert etter ulike kriterier).

Det er gjennomført faktoranalyser med alle 2008 numeriske variable, der hovedinntrykket er oppsummert i følgende tabell mht. aktuelle faktorer (med tolkning) og hvilke variable som lader på disse:

Tabell 9: Faktoranalyse

Faktor 1 Inntekt	Faktor 2 Formue	Faktor 3 Næring	Faktor 4 Pensjon	Faktor 5 JSF	Faktor 6 Underskudd
IB_ALM_INNT_E_SERF IB_ALM_INNT_F_SERF PERSON_INNTEKT_LONN T_GRL_LONN T_AVG_LONN TOPPSKATTEGRUNNLAG TOPPSKATT LONN_111A	IB_NETTO_FORMUE SUM_SKATT_AVGIFT NETT_FORMUE_STAT SKATT_FORM_KOMM SKATT_FORM_STAT	IB_P_INNT_NER_U_REF SUM_IB_PERSONINNT_ALLE_TYPER T_GRL_NERING T_AVG_NERING	PENSJON_INNTEKT T_GRL_PENSJON T_AVG_PENSJON ALDERSPENSJON_217	IB_P_INNT_JSF_U_REF T_GRL_JSF T_AVG_JSF	NEG_ALM_INNT_KUN_IB UDEKKET_UNDERSKUDD

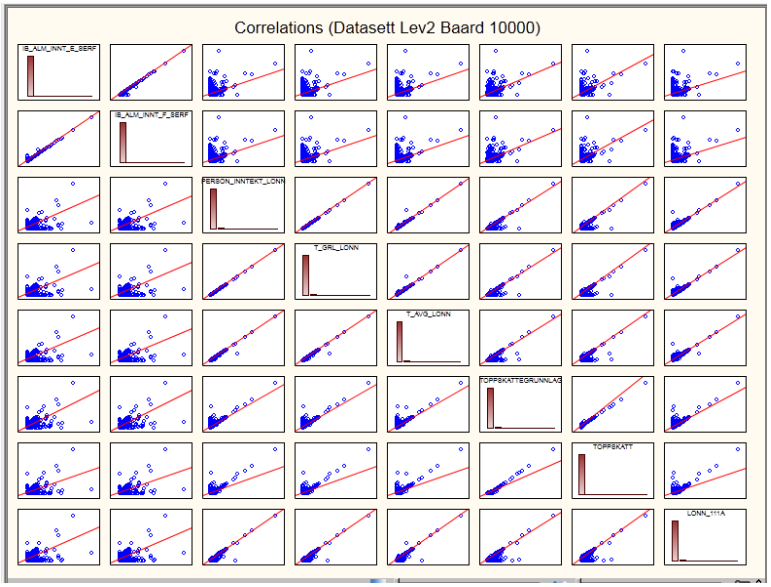
*Spesifikasjon Principal component extractio, Varimax normalized rotation, missing substituted by means, 6 factors.

Legger vi på to faktorer til får vi:
Faktor 7: RESTSKATT_OVERSKYTENDE, AVSRENTER_KREDIT
Faktor 8: IB_GODTJ_NERING
men disse forklarer såpass lite at de ikke er verd å vie særlig oppmerksomhet.

Tabell 10 og grafen nedenfor viser hvor sterkt korrelasjonen er mellom de 8 variablene som er dominerende i Faktor 1. Vi ser at for enkelte variable er korrelasjonen så høy med en enn flere andre at den ikke har noen tilleggsverdi og like gjerne kan fjernes.

Tabell 10: Korrelasjon i Faktor 1

Variable	Means	Std.Dev.	IB_ALM_INNT_E_SERF	IB_ALM_INNT_F_SERF	PERSON_INNTEKT_LONN	T_GRL_LONN	T_AVG_LONN	TOPPSKATTEGRUNNLAG	TOPPSKATT	LONN_111A
	Correlations (Datasett Lev2 Baard 10000 korrigert2) Marked correlations are significant at p < ,05000 N=9712 (Casewise deletion of missing data)									
IB_ALM_INNT_E_SERF	252042,1	404366,1	1,000000	0,995999	0,568591	0,568591	0,567740	0,683104	0,638642	0,551920
IB_ALM_INNT_F_SERF	258575,0	405943,0	0,995999	1,000000	0,557930	0,557930	0,557060	0,684852	0,642748	0,540221
PERSON_INNTEKT_LONN	249589,7	291476,9	0,568591	0,557930	1,000000	1,000000	0,994990	0,883834	0,751950	0,970800
T_GRL_LONN	249589,7	291476,9	0,568591	0,557930	1,000000	1,000000	0,994990	0,883834	0,751950	0,970800
T_AVG_LONN	19266,5	22650,4	0,567740	0,557060	0,994990	0,994990	1,000000	0,873316	0,732945	0,966844
TOPPSKATTEGRUNNLAG	317888,8	276237,8	0,683104	0,684852	0,883834	0,883834	0,873316	1,000000	0,877092	0,855869
TOPPSKATT	5113,5	23904,3	0,638642	0,642748	0,751950	0,751950	0,732945	0,877092	1,000000	0,713760
LONN_111A	230642,7	277193,7	0,551920	0,540221	0,970800	0,970800	0,966844	0,855869	0,713760	1,000000

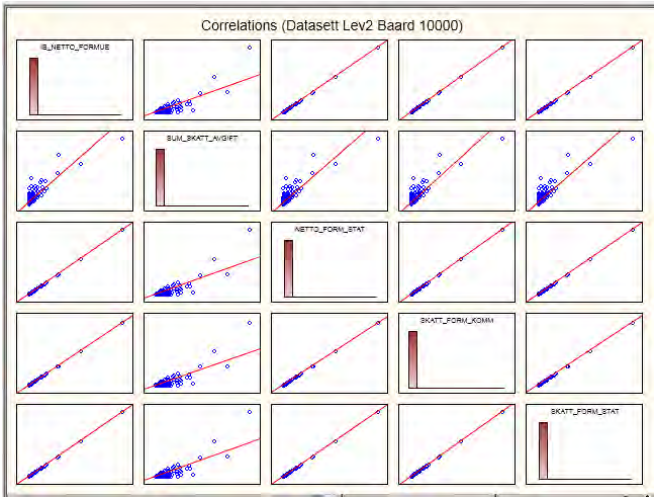


Figur 1: Korrrelasjon mellom variable i Faktor 1

Tilsvarende følger tabell og graf for korrelasjonene for de variable som er dominerende i Faktor 2. Her er alle formuesvariablene så sterkt korrelerte at det er nokså meningsløst å bruke mer enn en, eller evt. et veid gjennomsnitt. SUM_SKATT_AVGIFT er også høyt korrelert med de øvrige, men har muligens en selvstendig rolle.

Tabell 11: Korrelasjon i Faktor 2

Variable	Means	Std.Dev.	IB_NETTO_FORMUE	SUM_SKATT_AVGIFT	NETTO_FORM_STAT	SKATT_FORM_KOMM	SKATT_FORM_STAT
Correlations (Datasett Lev2 Baard 10000) Marked correlations are significant at p < ,05000 N=10577 (Casewise deletion of missing data)							
IB_NETTO_FORMUE	1293413	21116173	1,000000	0,820431	0,999976	0,999965	0,999935
SUM_SKATT_AVGIFT	91101	264478	0,820431	1,000000	0,821376	0,819486	0,820111
NETTO_FORM_STAT	1315312	21211772	0,999976	0,821376	1,000000	0,999933	0,999946
SKATT_FORM_KOMM	8256	147676	0,999965	0,819486	0,999933	1,000000	0,999976
SKATT_FORM_STAT	4701	84739	0,999935	0,820111	0,999946	0,999976	1,000000



Figur 2: Korrelasjon mellom variable i Faktor 2

Merknad. Dersom en har to eller flere variable som er sterkt korrelerte, og har tanke om å fjerne en (eller flere) av dem, bør en ha sideblikk til hvilke(n) variabel som det er størst risiko for å mangle observasjon fra.

5. Kjennetegnsanalyse og klassifikasjon: Generelt

I dette kapitlet gis en generell introduksjon om klassifisering av objekter i grupper basert på observerbare kjennetegn, samt en kort oversikt over ulike metoder for kjennetegnsanalyse og klassifikasjon som kan håndtere problemstillinger av typen:

En kategorisk variabel mislighet (1) eller ikke (0) skal forklares ved en rekke kategoriske og numeriske variable (målevariable) med formål å kunne best mulig klassifisere nye cases der gruppetilhørighet (1) eller (0) ikke er kjent.

Denne metodegjennomgangen er ikke nødvendig for forståelsen av de følgende kapitler med konkrete anvendelser på det foreliggende datamateriale, og leseren kan gjerne gå til det avsluttende avsnitt i dette kapitlet, som gir en begrunnelse for våre anbefalinger til Skattedirektoratet ved valget av metode, en begrunnelse som også gjelder for det foreliggende datamateriale med frivillig retting, som ikke behøver å bety mislighet. En nærmere presentasjon av de metoder som anbefales kommer i senere kapitler.

5.1 Genrerelle betraktninger

Først den generelle introduksjon til det å klassifisere objekter i ulike grupper basert på kjennetegn: Vi betrakter en populasjon av objekter eller individer som faller i en av to (evt. flere) grupper, som krever ulik handling. På beslutningstidspunktet er gruppetilhørigheten ukjent, men det foreligger observerbare karakteristika som muligens kan indikere gruppetilhørigheten for hvert objekt eller individ. Denne problemstilling forkommer i mange ulike sammenhenger, men vi vil her se på skatterevisjon. Eksempler fra dette felt er

1. Karakteristika som kan indikere unndragelse av skatt
    - regnskapstall, organisasjonsstruktur etc
2. Karakteristika som kan indikere evne til å betale skatterestanser

I begge eksempler tenker vi oss at de observerbare karakteristika skal brukes til å fatte en beslutning, som her typisk vil være hvor man skal sette inn ressursene. I det første eksemplet foreligger i prinsippet gruppetilhørigheten ved beslutningstidspunktet, men kan bare avdekkes gjennom omfattende granskninger, hvis om mulig (retrospektivt perspektiv). I det andre eksemplet vil gruppetilhørigheten først materialisere seg på et senere tidspunkt (prospektivt perspektiv). Det er teoretiske nyanser mellom de to perspektivene, som ikke gjør stor forskjell i praksis, i hvert fall hvis vi begrenser oss til enkle klassifikasjonsproblemet. Imidlertid åpner det prospektive perspektiv for beslutninger som påvirker den senere gruppetilhørighet, eksempelvis ved tilrettelegging av betalingsplan ut fra de observerte karakteristika.

Innen rammen ovenfor kan vi ha ulike kontekster, eksempelvis

- Liten populasjon – store beløp – kompliserte forhold – mange karakteristika
    - eksempelvis mistanke om unndragelser knyttet til skatteparadis
- Stor populasjon – moderate beløp – få karakteristika

I de neste kapitlene vil vi illustrere de ulike metoder på de data vi har mottatt fra Skatteetaten. De fleste metoder tar utgangspunkt i en binær variabel Y som indikerer skatteunndragelse (1) (dvs. observasjoner fra frivillig retting gruppen) og ikke skatteunndragelse (0) (dvs. observasjoner fra kontrollgruppen), og en samling mulige forklarende variable X_1, X_2 , X_3 ,…,X_p , dvs. forklaringsvariablene (eller kombinasjoner / transformasjoner av variablene) vi har mottatt for begge grupper.

Våre data omfatter et mindre antall individer (cases) med frivillig retting (friv=1) og et stort antall cases (friv=0) som kontrollgruppe. De tilgjengelige cases med frivillig retting er alle man til nå har fått blant alle aktuelle skatteytere, og det kan være flere individer med samme uønskede adferd som ikke har kommet frem enda eller forblir uoppdaget. Visse observerbare karakteristika er formodentlig felles for individene med frivillig retting, og formodentlig også med individer med tilsvarende ”svin på skogen”, som ikke er kommet frem. Siktemålet er at disse karakteristika kan komme til nytte også i en situasjon der gruppetilhørigheten er ukjent, som tilfellet er foran behandlingen av et nytt skatteoppgjør, der en skal prioritere kontrollarbeidet.

De skatteobjekter som gir grunnlag for retting kan betraktes som ”nåler i en høystakk”, med de konsekvenser det har for å finne de enkelte nåler, uten å måtte rote i store deler av høystakken. En mer treffende allegori er å lete etter spesielle uønskede høystrå i en høystakk, høystrå som ligner de øvrige i de fleste henseende, men skiller seg på visse karakteristika, og der vi i utgangspunktet ikke vet hvilke karakteristika som skiller best.

For å avdekke de variable som kan ha prediktiv verdi for frivillig retting, er det ikke hensiktsmessig med altfor stor kontrollgruppe. Det er dels på grunn av begrensninger i programvare mht. tidsbruk i håndteringen av store datamengder, og at de ”signaler” de enkelte variable gir er så svake at man ikke blir oppmerksom på dem, med mindre man finjusterer måleinstrumentet. En håper likevel at ”nål i høystakk” problematikken blir tilstrekkelig eksponert, og de variable som avtegner seg i en mindre høystakk, også er de som er relevante i den store.

For noen av metodene har vi derfor benyttet et mindre utvalg av kontrollgruppen (vanligvis 10 000 observasjoner), dette vil i hovedsak ikke ha stor betydning for resultatene og konklusjonene vi trekker. Vi vil klargjøre for hver enkelt analyse hvordan vi har gått fram. Vi vil antyde hvilke forklaringsvariable som kan benyttes til prediksjon av skatteunndragelse, samt hvilke metoder som ser ut til å fungere best og presentere styrker og svakheter ved hver av dem. For den ikke-tekniske leser, er det mulig å hoppe over de neste avsnittene, og gå direkte til det kapittelet, hvor vi oppsummerer våre funn.

Et sett av data (”treningsdata”) blir først brukt til å tilpasse modellen for den aktuelle metode til dataene, og så vurdere hvor godt metoden klassifiserer disse dataene (”within sample”). Dette gir et godt grunnlag for valg mellom metoder, men den virkelige testen er hvor godt den etablerte klassifikasjonsregel fungerer på nye data (”testdata”, ”out of sample”). Her er vi stilt overfor et problem, idet antall frivillig rettede i forhold til kontrollgruppen er svært lite. Ideelt sett skulle treningsdata og testdata ikke overlappe. For å få et tilstrekkelig informasjonsgrunnlag til å etablere en brukbar klassifikasjonsregel ser vi oss nødt til å ta med alle de frivillig rettede i treningsdatasettet, og bare skifte ut kontrolldataene. En mulighet for å omgå dette er såkalt kryssvalidering, men dette er holdt utenfor i denne rapport.

5.2 Metodeoversikt

Det er mange og til dels svært ulike analysemetoder som kan være aktuelle, og vi vil gi en kort omtale av følgende hovedkategorier:

a. Klassisk diskriminantanalyse
b. Kategorisk regresjons: logit, probit
c. Statistisk tre – (”recursive partitioning”): CART, CHAID, QUEST etc.
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 etc.
g. Support Vector Machines (SVM)
h. Genetisk programmering

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 "kunstig intelligens" (AI). Mange har kjempet innbitt for ”sin metode”, men etter hvert har man sett en tilnærming, der ulike leire har gjensidig påvirket hverandre. I praksis har man forsøkt å kombinere begge perspektiver under termen Knowledge Discovery in Databases (KDD). Det fins også Bayesianske varianter av disse metodene, der subjektive oppfatninger kan komme til uttrykk, samt enkelte hybrider av metodene, bl.a. av statistisk tre og logistisk regresjon.

De nevnte metoder finnes som kommersiell programvare, og også som shareware eller freeware, som R. De velrenommerte store programvareleverandører, som SAS og Statistica har i dag innlemmet de mest velprøvde metodene i sitt tilbud, mens nisjeprodukter fins også for nyere maskinlæringsmetoder. Det fins nok flere andre metoder enn ovennevnte som fortjener en egne kategorier, bl.a. former for induktiv inferens. Når disse ikke er tatt med i denne vurdering, er det i hovedsak fordi de ikke har funnet samme utbredelse i praksis, og ikke er like tilgjengelig i programvare.

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”. Eksemler på slike metoder er såkalt ”boosting” og ”bagging” (bootstrap aggregating)

En kort omtale av metodene a-h følger:

Diskriminantanalyse består i å splitte et mangedimensjonalt rom av forklarende variable i to, dvs. i de deler som best mulig fanger opp henholdsvis mislighet og ikke mislighet. Dette skjer ved å legge et hyperplan gjennom rommet, uttrykt ved en lineær funksjon av de forklarende variable. Denne bestemmes ut fra erfaringsdata og kan brukes til å klassifisere nye observasjoner, alt ettersom hvilken side av hyperplanet den nye observasjonen ligger. Planet kan parallellforskyves for å ivareta ønsket om en annen vektlegging av risikoen for de to typer feilklassifikasjon enn likevekting. Alternativt kan prosedyren uttrykkes ved to lineære klassifikasjonsfunksjoner, slik at den nye observasjonen klassifiseres i den kategori som oppnår størst verdi. Klassisk diskriminantanalyse forutsetter numeriske forklarende variable og helst såkalt multinormalitet. I situasjoner med potensielt mange forklarende variable må man typisk gå gjennom en prosess med variabelutvelgelse før den endelige analyse. Den

håndterer ikke lett manglende observasjoner og er ikke prediktiv av samme natur som konkurrerende regresjonsmetoder av typen logit/probit. Det fins modifikasjoner av klassisk diskriminantanalyse som tar omsyn til ovennevnte, bl.a. er det mulig å få kvantifisert sannsynligheter for mislighold for gitte karakteristika. Likevel er diskriminantanalyse etterhånden kommet i bakgrunnen som aktuell metode for denne typen problemer av prediktiv natur.

Kategorisk regresjon av typen logit, probit dreier seg i hovedsak om å modellere sannsynligheten for mislighet som funksjon av forklarende variable, som kan være såvel numeriske som kategoriske. Ut fra erfaringsdata estimeres regresjonssammenhengen, og denne brukes til å klassifisere nye observasjoner. En kan da bruke beregnede sannsynligheter for mislighet for gitte forklarende variable til å treffe beslutning om å gi gripe inn eller prioritere oppfølging. De ulike variablene gis en vekt gjennom sine regresjonskoeffisienter, men det kan likevel være vanskelig å tolke den rolle hver variabel har for klassifikasjon av mislighold, dels på grunn av mulig samvariasjon mellom de forklarende variable. I situasjoner med potensielt mange forklarende variable må man typisk gå gjennom en prosess med variabelutvelgelse før den endelige analyse. Det kan også være vanskelig å implementere ikke-lineariteter og samt manglende observasjoner. Denne type analyse er den som er mest vanlig i standard statistisk programvare, men ikke alle håndterer manglende observasjoner. Merk at vanlig regresjonsanalyse anvendt på 0-1 variable (mislighold, ikke mislighold) vil innebære en teoretisk inkonsistens, ved at de estimerte sannsynligheter ikke nødvendigvis ligger mellom null og en. Siden standard regresjon er velkjent, bør man ikke utelukke dette, siden prediksjoner av gruppetilhørighet likevel ikke behøver å bli dårlige.

Statistisk tre (klassifikasjonstre/regresjonstre) er trinnvis splitting av observasjonsenhetene i grupper, der enhetene med mislighet fremstår klarere i den ene gruppen enn de(n) andre. Dette organiseres som en trestruktur, som fremstår som beslutningsregler for å lokalisere undergrupper der enheter med mislighet 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 statistisk tre er:

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

Mange andre klassifikasjonsmetoder er i utgangspunktet ikke i stand til å møte alle disse krav, men deres forkjempere har gjort anstrengelser for å bøte på dette. På den annen side har enkelte kritikere av statistisk tre pekt på faren for ”overfitting”, men det er en risiko som også er til stede ved mange maskinlæringsmetoder. Det fins mange ulike varianter av algoritmer for statistisk beslutningstre (CART, CHAID, QUEST etc.). I situasjoner med et stort antall forklarende variable, kan det være nødvendig med en variabelutvelgelse på forhånd, men videreutviklede varianter fins, der dette ikke er nødvendig, såkalt ”boosted trees” og ”random forest”.

Ikke-parametriske metoder omfatter diverse metoder som ikke er forankret i parametriske statistiske modeller. ”k nearest neigbor” metoden består i å representere observasjonene i et mangedimensjonalt rom, og splitte rommet iht. mislighet ved bruk av (generaliserte) avstandsmål. Nye observasjoner kan så klassifiseres ut fra sin avstand. ”Kernel density estimation” omfatter metoder for å estimere sammenhenger, der en i utgangspunktet har få spesifikke antakelser om disse. Dette krever forholdsvis flere observasjoner enn metoder basert på parametriske modeller, og gir ikke i samme

grad tolkningsmuligheter.

Multiple adaptive regression splines (MARS) er en teknikk for såkalt ”data mining” med utgangspunkt i regresjon og såkalte ”splines”, som er funksjoner i stand til å fange opp kompliserte sammenhenger. Metoden håndterer også greit manglende observasjoner, og korrigerer for spesifiserte forklarende variable som viser seg irrelevante. Evne til å håndtere ikke-linearitet er felles med en del andre data mining metoder, så som ANN og SVM. Fordelen med MARS kan være at resultatet er lettere tolkbart, med nærmere tilknytning til klassiske statistiske regresjonsmodeller.

Nevrale nettverk (”Artificial Neural Nets”, ANN) er metoder knyttet til maskinlæring og har i flere sammenhenger kunnet vise til noe bedre prediksjonsevne enn tradisjonelle statistiske metoder. En tenker seg da overføring av informasjon fra forklarende variable ”på kryss og tvers" gjennom flere lag ”layers”. I en viss forstand er dette en generalisering av tradisjonell regresjon, som bare bruker et lag. Ulempen er at den prediksjonsmodell som etableres blir en teknisk "sort boks", med usynlig og vanskelig tolkbart innhold. Det er riktignok mulig å kjøre nettverk med på nytt ved å fjerne variable etter tur, og måle endringer som ved vanlig regresjon, med det er tungvint og lite brukervennlig. Med sine mange ”frihetsgrader” er en betydelig fare for ”overfitting”, dvs. at modellen tilpasser seg spesielle særegenheter i data, uten prediktiv verdi. Å unngå dette kan stille betydelige krav til brukeren. Den suksess slike metoder har på enkelte felt ligger trolig i dens evne til å fange opp eventuelle ikke-lineære sammenhenger. Nevrale nettverk fikk raskt en viss utbredelse i mange miljøer, men det er vel blitt noe mer nøkternhet til metoden etter hvert.

Support Vector Machines (SVM) er som ANN basert på teorier for maskinlæring, og har vunnet en viss popularitet, bla. fordi metoden i sammenligning med ANN har noe kortere regnetid og kanskje også en lettere forståelig teori. Standard SVM algoritmen bruker data med kjent gruppetilhørighet til å bygge en modell der nye data med ukjent gruppetilhørighet blir tilordnet en av gruppene. Dette skjer ved å representere hvert case som ”punkter” i et uendeligdimensjonalt ”egenskapsrom”, der forskjeller mellom gruppene fremtrer klarest mulig og er lettere å analysere enn i det opprinnelige endeligdimesjonale datarommet. Tanken bak dette er at ikke-lineære sammenhenger har bedre mulighet for å fremtre i det alternative rommet. SVM bruker såkalte kjernetransformasjoner fra det opprinnelige datarommet. SVM håndterer imidlertid ikke interaksjoner mellom de forklarende variable like effektivt som ANN. Utskrifter fra SVM er ikke like tolkbare som en del andre metoder.

Genetisk programmering er en form for maskinlæring basert på analogier fra tilpasning og overlevelse for organismer i naturen, der de organismer som er mest tilpasningsdyktige til de gitte omgivelser overlever og reproduserer seg. Det enkelte individs egenskaper representeres her ved aritmetiske og logiske relasjoner. Individene overfører de gunstige egenskapene til neste generasjon, som typisk tilpasses bedre og bedre til omgivelsene for hver generasjon. I denne sammenheng utnyttes de genetiske begrepene rekombinasjon, seleksjon og mutasjon. Her vil sammenhenger og relasjoner som ”står seg godt” ettersom de ”får brynt seg” på ulike deler av tallmaterialet overleve. Metoden har i likhet med nevrale nettverk evne til å fange opp ikke-lineære sammenhenger, men er i mindre grad enn slike utprøvd i praksis.

Det er gjennom årene foretatt en del systematiske sammenligninger av ulike metoder, uten at man i dag kan gi en generell konklusjon mht. deres relative fortrinn. Dette henger sammen med flere forhold. Noen studier sammenligner få metoder av gangen, noen studier er ikke fullt ut upartiske,

noen er basert på et datagrunnlag som ikke nødvendigvis er relevant for andre (norske) forhold. Noen studier bruker forholdsvis snevre vurderingskriterier, der kun risiko for klassifikasjonsfeil, og ikke de sider som går på utvikling og praktisk bruk som beslutningsstøttesystem. De sammenligninger mellom metoder som foreligger i dag, går i hovedsak på regresjon (logit, probit), statistisk tre (CART eller lignende) og nevrale nettverk (ANN). Andre maskinlæringsmetoder, som genetisk programmering, er i mindre grad prøvd ut i praksis på bredt grunnlag, og typisk ikke kommet med i de sammenlignende studier som er allment tilgjengelige. Et område der man i vitenskapelig litteratur har holdt flere av ovennevnte metoder opp mot hverandre, er vurdering av konkursrisiko som grunnlag for etablering av systemer for kredittverdighet. Også her er synet delt, og vi finner støtte både for statistisk tre og logistisk regresjon, og i noen grad også maskinlæring.

5.3 Vurderingskriterier

Vurderingskriterier blir her diskutert ut fra den allmenne problemstilling (mislighet) skissert innledningsvis i dette kapitlet, som også kan gjøres gjeldende for den problemstilling og datamateriale (frivillig retting) som ligger til grunn for denne rapporten.

Blant de kriterier som er aktuelle i tillegg til klassifikasjonsevne er:

- Prediksjonsevne 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 variable
- 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

Vektleggingen av disse avhenger i noen grad av formålet med analysen. Formålet kan dreie seg om kun klassifikasjon, dvs. en ja/nei beslutning for inngripen eller prioritering av nærmere gransking. I noen tilfeller er man interessert i mer, der eksplisitte estimerte sannsynligheter for mislighet ut fra gitte karakteristika hos klienten komme til nytte. Her er det kategorisk regresjon som direkte går på dette, mens noe i denne retning er også oppnåelig ved enkelte av de andre metodene.

Med tolkbarhet menes at resultatet kan presenteres i en form som er meningsfylt for bruker og kan danne utgangspunkt for egne tilleggsvurderinger av så vel klienten som systemets egen funksjonalitet. Blant metoder med høy prediksjonsevne, bør en derfor ikke nødvendigvis velge den med den høyeste, men også vurdere de øvrige kriterier. Dersom man legger stor vekt på tolkbarhet og andre sider ved brukervennligheten, vil både noen av de klassiske statistiske metodene og noen av de nyere metodene for maskinlæring kunne komme til kort. De siste fordi at den såkalte kunnskapen forblir i "den sorte boksen" og ikke stimulerer brukeren.

I forbindelse med utvikling og implementering har en spørsmålet om hvilke variable som skal inngå. Ved flere av metodene, bl.a. diskriminantanalyse og regresjon, er det ønskelig å redusere antall forklarende variable. Dette kan gi bedre prediksjoner fordi det blir færre parametere å estimere og en

unngår eventuell kolinearitet. Reduksjon av variable er ikke en entydig formell prosess, skjønn og hensyn til tolkbar modell kan spille inn her. Klassisk diskriminantanalyse ser ut til å ha kommet til kort ut fra flere av de nevnte vurderingskriterier, og de ikke-parametriske metodene møter heller ikke dette krav. Det samme gjelder nok maskinlæringsmetodene.

Blant metodene a-h som er listet i forrige avsnitt kommer statistisk tre (klassifikasjonstre) etter vår oppfatning godt ut på alle de nevnte vurderingskriteriene, mens kategorisk regresjon også synes å være et alternativ som bør utprøves. Disse to metodene er på et vis komplementære, og det faktum at det fins en hybrid av disse to, gjør også kombinasjonen interessant. Alle de listede metodene er tilgjengelig i statistisk programvare som SAS, Statistica og R. Skattedirektoratet er allerede brukere av Statistica, og det bør også tillegges vekt at SKD har intern kompetanse som har gjort bruk av statistisk tre og kategorisk regresjon på beslektede problemstillinger.

Senere kapitler tar sikte på å dokumentere utprøving av de mest aktuelle metodene, i hovedsak kategorisk regresjon og klassifikasjonstre. De fleste av de øvrige er også forsøkt, uten at de kunne oppvise spesielle fortrinn.

6. Frivillig retting: Innledende klassifikasjoner

For å få et første innsyn til klassifiseringspotensialet til de mest aktuelle metodene tar vi for oss gruppen av frivillig retting sammen med en referansegruppe med 1000 tilfeldig valgte individer fra hele kontrollgruppen. Med dette er de to gruppene omtrent av samme størrelsesorden (som langt på vei hadde vært ønskelig), men langt fra den "nål i høystakk" problematikk vi har i praksis, der de frivillig rettede, og forhåpentligvis også unndragere, utgjør en svært lite antall i forhold til antall utenom. Dette har neppe særlig mye å si for metodenes relative klassifikasjonsevne. Målet er her bare å kunne gi støtte til de generelle vurderingene av valg av metode som er gitt i forrige kapittel. De to mest aktuelle metodene var logistisk regresjon og klassifikasjonstre. For å sammenligne og prioritere tar vi med noen ulike varianter som blir kort forklart i de etterfølgende kapitler, samt vanlig lineær regresjon, som kan gi brukbare prediksjoner, selv om en slik modell er teoretisk inkonsistent. Flere metoder er utprøvd på samme vis (bl.a. nearest neighbor og nevrale nett), men tas ikke med her.

Sammenlikningen gjøres på to måter, "in-sample" og "out-of-sample". I disse begreper skal "sample", eller stikkprøven, betraktes som den del av datasettet som brukes til å tilpasse modellen, for eksempel en logistisk regresjonsmodell. Med "in-sample" forstås da at evalueringen av prediksjonene utføres på samme individer som modellen er tilpasset til. Med "out-of-sample" mener vi at vi til evalueringen bruker individer som ikke inngår i stikkprøven. Den siste varianten er på mange måter mer i tråd med den situasjon som Skatteetaten i realiteten står overfor når man skal predikere klassetilhørighet hos et individ. En modell er da tilpasset og denne brukes for å predikere klassetilhørighet for et nytt individ som ikke var med i stikkprøven.

6.1 "In-sample" klassifikasjon

Vi vil predikere FRIV (=0 eller 1) på grunnlag av et mindre utvalg av aktuelle variable av ulik karakter som gitt i tabell nedenfor. Vi vil først se hvordan den modell som ligger til grunn tilpasser seg de utvalgte data ("within sample"), og deretter hvordan den klassifikasjonsregel som modeller gir klassifiserer nye data ("out of sample"). For hver metode brukes den såkalte "default" klassifikasjonen, dvs. i de tilfeller der sannsynligheter blir anslått, blir individet klassifisert til den mest sannsynlige gruppen. Siden det er to typer feilklassifiseringer med ulikt alvor:

1. Klassifisere skatteyter som frivillig retter (FRIV=1) når denne ikke er det (FRIV=0)
2. Klassifisere skatteytere som ikke frivillig retter (FRIV=0) når den er frivillig retter (FRIV=1)

Alle aktuelle metoder har mulighet for å vektlegge ulik grad av alvor, samt også apriori oppfatninger om gruppetilhørigheten. Dette er foreløpig holdt utenfor.

Tabell 12: Variabler benyttet i analysen

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
NACE	NACE kode arbeid
ALMIB	Alminnelig inntekt innenbygds
SKJERM	Skjermingsfradrag aksjer
HIST	Historikk kode, 1 hvis tidligere blitt skatteberegnet
SENT	Sentralitetskode, 3 hvis bor sentralt
LONN	Personinntekt lonn
ALDER	
TOPPSKATT	Beløp
KJONN	
OVER	Skatt – forskuddstrekk

Etter å ha tatt vekk observasjoner der alle variablene manglet verdier ble antall observasjoner 1532.

Vi starter med standard lineær regresjon der vi forklarer FRIV med alle variablene, i tabellen. Dette ga følgende tilpasning til data:

Tabell 13: Lineær regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	884	73	957
FRIV=1	143	432	575
Kolsum	1027	505	1532

Av de 957 individene som ikke har levert frivillig retting blir 884 klassifisert som FRIV=1. Vi vet selvfølgelig ikke hvorvidt de øvrige 73 er feilklassifisert av regresjonsmodellen eller om de bør granskes mer nøye. La oss for denne illustrasjons skyld anta at de er feilklassifiserte. Vi antar dermed at alle individer i gruppen FRIV=0 er "rene".

Da det kan være grunn til å tro at de numeriske forklarende variablene til FRIV ikke har et lineært forhold, vil vi også prøve en ikke-parametrisk regresjon. Tabell 14 viser at flere individer ble korrekt klassifiserte, både i frivillig gruppen og kontrollgruppen.

Tabell 14: Ikke-parametrisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	914	43	957
FRIV=1	121	454	575
Kolsum	1035	497	1532

Vi undersøker også mulighetene med regresjon med såkalt LASSO, som er en metode som fjerner variable som fjerner mindre aktuelle variable fra regresjon ved å stte de tilhørende regresjonskoeffisientene lik null.

Tabell 15: Lineær regresjon med LASSO

	PRED=0	PRED=1	Radsum
FRIV=0	866	91	957
FRIV=1	185	390	575
Kolsum	1051	481	1532

Som vi ser i tabell 15 er dette tilsynelatende et tilbakeskritt. Klassifiseringen er dårligere i begge kategoriene av FRIV. Vi kommer tilbaks til dette nedenfor, der vi ser på out-of-sample prediksjoner.

Vi går nå videre og ser på en klassifisering basert på logistisk regresjon, som er den mestr vanlige formen for kategorisk regresjon. Denne metode tar høyde for at sannsynligheter er mellom null og en. Det finnes imidlertid ikke noen garanti at vi får en bedre approksimasjon enn den lineære modellen ga i dette spesifikke eksempel. Resultatet ser vi i tabell 16.

Tabell 16: Logistisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	930	27	957
FRIV=1	115	460	575
Kolsum	1045	487	1532

Prediksjonene er blitt uniformt bedre enn med lineær regresjon. Vi kan i tabell 16 se at flere individer i begge gruppene har blitt klassifisert korrekt. Også for logistisk regresjon kan vi ta høyde for ikke-lineæere forhold og restriksjoner på parameterverdiene (LASSO). Resultatene for dette vises i tabell 17 og 18.

Tabell 17: Ikke-parametrisk logistisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	935	22	957
FRIV=1	120	455	575
Kolsum	1055	477	1532

Ikke-parametrisk logistisk regresjon klassifiserer omtrent like godt som vanlig (parametrisk) logistisk regresjon. Vi får grunn til å komme tilbaks også til dette i neste avsnitt.

Tabell 18: Logistisk regresjon med LASSO

	PRED=0	PRED=1	Radsum
FRIV=0	937	20	957
FRIV=1	173	402	575
Kolsum	1110	422	1532

Logistisk regresjon med LASSO gir her en klassifisering som er betydelig dårligere enn med vanlig logistisk regresjon.

La oss gå over til det som vi har pekt ut som det mest aktuelle alternativ til logistisk regresjon, nemlig klassifikasjonstre. For denne metoden er resultatet gitt i Tabell 19.

Tabell 19: Klassifikasjonstre

	PRED=0	PRED=1	Radsum
FRIV=0	872	85	957
FRIV=1	67	508	575
Kolsum	939	593	1532

Av den gruppen som vi er mest interessert i, de som har gjort frivillig retting (FRIV=1), er færre individer feilklassifiserte enn for de andre metodene. For kontrollgruppen er dette imidlertid ikke fallet. Her må vi nevne en avveiing som må gjøres når man skal bruke en slik modell for å bestemme hvorvidt et individ bør sjekkes manuelt og de ressurser dette medfører: Vi vil ha en stor sannsynlighet for å finne et individ som har skjult inntekter eller formue, men vi vil ikke at sannsynligheten til et individ som ikke har skjult noe skal sjekkes manuelt skal være for stor. I dette perspektiv er klassifikasjonsstreet et tilbakeskritt når det gjelder å ikke finne "falske positive", men et skritt fremover når det gjelder styrke å finne "korrekte positive".

Tabell 20: "Boosted" regresjonstre

	PRED=0	PRED=1	Radsum
FRIV=0	934	23	957
FRIV=1	47	528	575
Kolsum	981	551	1532

"Boosted" regressjonstrær bruker mange ulike trær og på denne måte flere forklarende variabler. Det er ikke åpenbart at denne metoden skal gi bedre prediksjoner, da man involverer flere "svake" forklarende variabler. I dette tilfellet ser vi imidlertid at vi forbedrer prediksjonen på begge relevante aspekter. Vi får færre falske negative og flere korrekte positive prediksjoner enn de andre metodene som vi har brukt i dette eksemplet.

6.2 "Out-of-sample" klassifikasjon

Analysene i forrige avsnitt var gjort "in-sample", det vil si en modell var tilpasset data og evalueringene av klassifiseringen ble gjort på de samme dataene. Her kan det åpenbart være et problem med såkalt overtilpassning. Av denne grunn skal vi i dette avsnittet gjennomføre analysene en gang til, men nå ta høyde for dette problem. Vi gjør dette på følgende måte: Vi velger tilfeldig ut 1000 av de 1532 observasjonene for å tilpasse modellene, de øvrige 532 brukes etterpå for å gjøre de samme sammenligningene som gjordes i det forrige avsnittet. Resultatene er gitt i tabellene 21-28.

Tabell 21: Lineær regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	307	34	341
FRIV=1	59	132	191
Kolsum	366	166	532

Tabell 22: Ikke-parametrisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	327	14	341
FRIV=1	51	140	191
Kolsum	378	154	532

Tabell 23: Lineær regresjon med LASSO

	PRED=0	PRED=1	Radsum
FRIV=0	301	40	341
FRIV=1	58	133	191
Kolsum	359	173	532

Tabell 24: Logistisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	332	9	341
FRIV=1	56	135	191
Kolsum	359	173	532

Tabell 25: Ikke-parametrisk logistisk regresjon

	PRED=0	PRED=1	Radsum
FRIV=0	332	9	341
FRIV=1	49	142	191
Kolsum	381	151	532

Tabell 26: Logistisk regresjon med LASSO

	PRED=0	PRED=1	Radsum
FRIV=0	338	3	341
FRIV=1	63	128	191
Kolsum	401	131	532

Tabell 27: Klassifikasjonstre

	PRED=0	PRED=1	Radsum
FRIV=0	300	41	341
FRIV=1	39	152	191
Kolsum	339	193	532

Tabell 28: "Boosted" Klassifikasjonstre

	PRED=0	PRED=1	Radsum
FRIV=0	326	15	341
FRIV=1	36	155	191
Kolsum	362	170	532

Mønsteret er omtrent det samme som for "in-sample" prediksjonene. De korrekte positive øker på samme måte når man går fra lineær regresjon til logistisk regresjon til klassifikasjonstre og til slutt "boosted" klassifikasjontrær. De falske positive er litt færre for logistisk regresjon. De ikke-parametriske metodene ser ut til å gi bedre prediksjoner enn de parametriske (vanlig lineær og logistisk regresjon) også out-of-sample. Dette var forventet in-sample da vi på denne måte har mulighet å få en bedre datatilpasning. Det var imidlertid ikke forventet out-of-sample da problemet med overtilpassning ikke er tilstede. Når det gjelder LASSO, som fungerer som automatisk metode å rense vekk variabler som ikke ser ut til å forklare noe i den avhengige variabelen, hadde vi forhåpning om at dette skulle forbedre prediksjonene out-of-sample, men det slo ikke til.

Disse mønster bekreftes ved gjentak med 1000 andre tilfeldig valgte observasjoner til modelltilpasningen (og 532 andre til evalueringen).

I de påfølgende kapitler vil vi gjennomgå de to metodene for kjennetegnsanalyse og klassifikasjon som ble utpekt på grunnlag av metodepresentasjonen og diskusjonen av vurderingskriteriene i kapittel 5, og den inneledende analyse i dette kapitlet, nemlig kategorisk regresjon og klassifikasjonstre. Vi vil gi en kort introduksjon til disse metodene, og presentere utfordringene og vurdere styrker og svakheter ved metodene brukt på de mottatte data. I første omgang anvender vi dataene til å illustrere metodene, og bruker da ofte variable som senere ikke nødvendigvis å være spesielt gode for det aktuelle klassifikasjonsformål, så framt de illustrerer en utfordring eller at statistisk poeng. Blant annet vil vi illustrer "nål i høystak" problematikken i forbindelse med klassifikasjonstre. Først avslutningsvis presenterer vi de modeller som ser ut til å gjøre det spesielt godt i prediktiv forstand. Det viser seg at dette ofte er modeller der de enkelte variablenes betydning ikke kommer så klart frem, og som derfor ikke i samme grad fremmer læringen.

7. Kjennetegnsanalyser: Kategorisk regresjon

I dette kapittelet tar vi for oss to metoder for kategorisk regresjon: logistisk regresjon og ikke-parametrisk additiv logistisk regresjon. Her illustreres metodens egenskaper og utfordringer på utvalgte deler av tallmaterialet, heller enn å kaste inn alle variable og finne den beste modell av dette slaget.

7.1 Logistisk regresjon

Den mest vanlige og mest velegnede formen for kategorisk regresjon er såkalt logistisk regresjon. Logistisk regresjon tar utgangspunkt i en statistisk modell, der sannsynligheten for skatteunndragelse P(Y=1) uttrykkes som funksjon av de forklarende variable. Mer konkret antas log-odds å være en lineær funksjon av de forklarende variable, der disse kan være så vel numeriske som kategoriske, dvs.

$$L = \ln(\frac{P(Y=1)}{P(Y=0)}) = \alpha + \beta_1 X_1 + ... + \beta_p X_p$$

Dette innebærer at for beregnet log-odds L er $P(Y=1) = e^L / (1+e^L)$ og $P(Y=0) = 1/(1+e^L)$. Ut fra erfaringsdata estimeres regresjonssammenhengen, og denne kan brukes til å klassifisere nye observasjoner. I og med at man har en modell, kan statistisk signifikans av variable vurderes, og prediksjonsevne vurderes. I tilfellet der mange forklarende variable er tilgjengelig, må man først finne et mindre antall variable som best mulig egner seg for prediksjon. Dette krever en viss statistisk kompetanse, og kan ikke automatiseres. Under følger en punktvis oppsummering av logistisk regresjon:

- Fanger opp ”globale” egenskaper ved data
- Gir formel for sannsynligheter for gruppetilhørighet
- Håndterer best lineære og ”glatte” datastrukturer
- Litt fleksibilitet mht ikke-linearitet og interaksjon (ved transformasjoner)
- Kan interpolere ikke-observerte verdier
- Mindre grad automatisert
- Gir grunnlag for innsikt og læring
- Demonstrert godt resultat i mange kontekster og er en av de mest brukte metodikkene for en binær responsvariabel

Under følger et eksempel på logistisk regresjon anvendt på våre data i Statistica, der logistisk regresjon gjennomføres ved å velge ’Generalized Linear/Nonlinear (GLZ) Models’, og deretter ’logit’. For å illustrere metoden inkluderer vi bare seks lett forståelige forklaringsvariable: fødselsår, personlig inntekt, formue, endring i lønn fra 2007 til 2008 og endring i formue fra 2007 til 2008. Disse vet vi fra den innledende analysen kan ha en forklarende effekt. En estimert modell basert på 10 000 observasjoner vises i Tabell 29, og der man isteden har modellert sannsynligheten for at en skatteyter er med i kontrollgruppen (dvs. FRIV = 0).

Tabell 29: De estimerte koeffisienter til en logistisk regresjon

Effect	FRIV - Parameter estimates (Datasett1END.sta) Distribution : BINOMIAL, Link function: LOGIT Modeled probability that FRIV = 0							
	Level of Effect	Column	Estimate	Standard Error	Wald Stat.	Lower CL 95, %	Upper CL 95, %	p
Intercept		1	-83,92	6,96	145,42	-97,5550	-70,2768	0,000000
FODSELAAR		2	4,50E-02	3,58E-03	157,90	3,80E-02	5,20E-02	0,000000
PERSON_INNTEKT_LONN		3	-1,56E-06	1,70E-07	83,61	-1,89E-06	-1,22E-06	0,000000
NETTO_FORM_STAT		4	-5,88E-07	3,08E-08	364,86	-6,48E-07	-5,28E-07	0,000000
LONN_08_07		5	-5,83E-08	4,28E-07	0,02	-8,98E-07	7,81E-07	0,891796
FORMUE_08_07		6	4,52E-07	4,82E-08	87,94	3,58E-07	5,47E-07	0,000000
KJONN	K	7	0,1608	0,0573	7,87	0,0484	0,2732	0,005038
Scale			1,0000	0,0000		1,0000	1,0000	

Vi ser fra de lave p-verdiene at alle forklaringsvariablene unntatt endring i lønn fra 2007 til 2008 er statistisk signifikante. Tolkningen av koeffisientene er som følger: Vi ser at personlig inntekt og formue har negativ koeffisient. Det betyr at økt personlig inntekt og økt netto formue gir mindre sannsynlighet for at skatteyteren er med i kontrollgruppen, og større sannsynlighet for frivillig retting gruppen. For fødselsår er det estimert en positiv koeffisient, som betyr at dess yngre skatteyteren er, dess større sannsynlighet er det at skatteyteren er med i kontrollgruppen. En positiv formuesendring ser ut til å gi økt sjanse for tilhørighet til kontrollgruppen, og kvinner har større sjanse for å tilhøre kontrollgruppen enn menn.

I tillegg til tabellen ovenfor gir programvaren også klassifikasjonstabeller, som i forrige kapittel, og mulighet for å kunne gjøre mer omfattende "within sample" og "out of sample" vurderinger. Vi kan også for hvert case få listet opp de estimerte sannsynligheter for å tilhøre hver av de to gruppene (FRIV=0, FRIV=1).

Det faktum at variablene er signifikante betyr ikke at de nødvendigvis er gode til å klassifisere. Med et stort tallmateriale vil man typisk oppnå statistisk signifikans selv ved små forskjeller mellom grupper. Et spørsmål som dukker opp er hvorvidt den antatt lineære sammenhengen mellom log-odds og inntekts- og formuesvariablene er gyldig i hele variasjonsområdet, og tilsvarende for alder. Her kan det foreligge mulighet for og nødvendighet av å transformere eller kategorisere enkelte av variablene.

I fortsettelsen av dette kapitelet vil vi utdype logistisk regresjon med de variable som lå til grunn for analysen i foregående kapittel, med tillegg av noen kategoriske forklarende variabler. Med dette får vi vist hvordan slike håndteres med logistisk regresjon, også for den ikke-parametrisk versjonen, og samtidig også vist noen av de praktiske problemer kategorivariable representerer, spesielt hvis de er mange. I utprøvingen av de ulike metodene har vi erfart at det som regel er bedre å ha et balansert datasett ved "in sample" identifiseringen av modellen, fremfor å bruke en stor kontrollgruppe sammen med den frivillig rettede gruppen. Vi har derfor valgt å bruke 1000 tilfeldig utvalgte observasjoner fra kontrollgruppen for å estimere modellen, for etterpå å evaluere prediksjonene på de omtrent 10 000 observasjonene som er brukt i det ovenstående.

Vi tilpasset en slik modell til våre data og presenterer estimatene av koeffisientene i tabell 30. Her er (UB2, NACETRUE, HISTN, KJONNM) dikotome 0-1 variable avledet av (UB, NACE, HIST, KJONN) med navn som angir hva som er 1-gruppen. Eksempelvis er UB2=1 dersom skatteyter betaler skatt til minst to kommuner. SENT1, SENT2, SENT3 er også dikotome variable som representerer hver av kategoriene 1, 2 og 3 for variabelen SENTRALITETSKODE, slik at kode=0 svarer til basiskategorien når disse tre variablene er null.

Tabell 30: En estimert logistisk regresjonsmodell

Variabel	Estimate	Std. Error	z-value	p-value
(Intercept)	-3.91551004	0.72024840	-5.43633280	0.00000005
KOMM	0.00014362	0.00012228	1.17453290	0.24018163
STAT	0.00008230	0.00021328	0.38588420	0.69958243
UB2	0.96553437	0.19668032	4.90915590	0.00000091
F0807	-0.00000055	0.00000015	-3.58017230	0.00034337
NACETRUE	0.24677896	0.25250838	0.97731000	0.32841572
ALMIB	-0.00000032	0.00000063	-0.50477840	0.61371451
SKJERM	0.00006656	0.00003632	1.83240230	0.06689150
HISTN	-2.02737813	0.29255565	-6.92988870	0.00000000
SENT1	0.24006733	0.75327438	0.31869840	0.74995524
SENT2	0.48866787	0.64069655	0.76271340	0.44563431
SENT3	1.96315389	0.55946100	3.50900940	0.00044978
LONN	0.00000025	0.00000061	0.41165910	0.68058931
ALDER	0.02924981	0.00614010	4.76373890	0.00000190
TOPPSKATT	0.00002462	0.00000816	3.01814080	0.00254331
KJONNM	0.39891395	0.18632233	2.14098840	0.03227497
OVER	0.00000314	0.00000100	3.13563920	0.00171480

Tolkningen av koeffisientene er som følger, og merk at vi her har modellert sannsynligheten for at en skatteyter er med i frivilliggruppen (dvs. FRIV = 1). For alder har vi estimert en positiv koeffisient, det betyr at dess eldre skatteyteren er, dess større sannsynlighet er det at skatteyteren er med i frivilliggruppen. For formuevariablene er det estimert positive koeffisienter, dvs. dess høyere formue, dess større sannsynlighet er det for at skatteyteren er med i frivillig retting gruppen. De er imidlertid ikke statistisk signifikante. Her har vi muligens eksempel på to variabler som inneholder stort sett samme informasjon. Hvis en av dem tas vekk så blir den andre signifikant. Det er imidlertid en grunn til at vi beholder den i tabellen ovenfor. På tross av at de ikke er signifikante vil variabelvalgsmetoden som illustreres i avsnitt 7.2 velge dem ut. Vi kan notere at koeffisienten til LONN er positiv (0.00000025) men ikke signifikant. En mulig forklaring til det kan være at vi observerer en overrepresentasjon i frivilliggruppen både av skatteytere med høy lønn og skatteytere med lav lønn. Videre ser vi at hvis UB>2, det vil si skatteytere med 2 eller flere utenbygdskommuner som de betaler skatt til, har økt sannsynlighet for at FRIV=1. Individer med HIST=0, det vil si de som ikke tidligere blitt skatteberegnet, er det mindre sannsynlighet at vi finner i frivilliggruppen. Tilsvarende tolkninger er mulig for de resterende koeffisienter.

Tabell 31: En estimert logistisk regresjonsmodell

	PRED=0	PRED=1	Radsum
FRIV=0	9203	325	9528
FRIV=1	150	425	575
Kolsum	9353	750	10103

Tabell 31 viser hvor godt modellen klassifiserer de 10103 skatteytene. Vi ser at 425 av de 575 (74%) av de i frivilliggruppen blir korrekt klassifiserte.

Det er selvfølgelig ikke en helt rettferdig konklusjon å si at modellen er så god på å fange opp alle skatteytenes gruppetillhørighet ettersom vi brukte alle i frivilliggruppen for å tilpasse modellen. For å se mer på dette tok vi vekk 100 fra frivilliggruppen da vi estimerte modellene. Disse bruktes etterpå for å evaluere prediksjoner. Av disse 100 klarte modellen å identifisere 64 korrekt.

Vi estimerte også modellen med alle 10103 skatteytene i delsamplet og fant, noe paradoksalt, at modellen er bedre på å klassifisere de 10103 observasjonen da vi kun bruker 1000 i kontrollgruppen. Dette kan ha sin forklaring i at det er viktig å ha en balansert fordeling mellom kontrollgruppen og frivilliggruppen. Vi vet ikke hva som i slikt fall er en optimal fordeling men dette er noe som kan være verdt å studere mer.

Ikke (Semi-) parametrisk logistisk regresjon

Vi kan også bruke en semi-parametrisk logistisk regresjonsmodell, der vi tillater at de numeriske variablene påvirker sannsynligheten for FRIV=1 på en ikke-lineær måte. En slik modell kan skrives

$$\ln\left(\frac{P(FRIV=1)}{P(FRIV=0)}\right) = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

der avhengigheten av de forklarende variable er en sum av ukjente funksjoner, som tallmaterialet selv søker å estimere. Vi kan i tillegg legge til dikotome (0-1) variabler som representerer de kategoriske variablene. Disse kommer da i tillegg til høyresiden i formelen ovenfor, og påvirker sannsynligheten for at FRIV=1 på en lineær måte. Vi har brukt de samme variablene som i avsnitt 7.1. Den estimerte modellen presenteres i tabell 32 og tabell 33.

Tabell 32: En estimert ikke-parametrisk logistisk regresjonsmodell: kategoriske variabler

Variable	Estimate	Std. Error	z-value	p-value
(Intercept)	8.96275640	1.36032920	6.58866740	0.00000000
UB2	0.82564270	0.19992190	4.12982650	0.00003630
NACETRUE	0.30724730	0.25530720	1.20344140	0.22880553
HISTN	-1.94337850	0.29941060	-6.49068140	0.00000000
SENT1	0.20180540	0.75269570	0.26811030	0.78861445
SENT2	0.49572060	0.63123160	0.78532290	0.43226428
SENT3	1.96619600	0.54899490	3.58144660	0.00034170
KJONNM	0.39690630	0.18837780	2.10697000	0.03512018

Tabell 33: En estimert ikke-parametrisk logistisk regresjonsmodell: numeriske variabler

Variable	Chi.sq	p-value
s(KOMM)	1.5606977	0.21156321
s(STAT)	0.1121046	0.73776081
s(F0807)	12.6956598	0.00036651
s(ALMIB)	0.6376571	0.42459553
s(SKJERM)	2.840163	0.09195377
s(LONN)	0.1554899	0.69374671
s(ALDER)	28.9783887	0.00001934
s(TOPPSKATT)	7.0349555	0.02942099
s(OVER)	11.7638467	0.00060457

Tabell 32 med de kategoriske variablene skal leses på samme måte som en vanlig regresjonsmodell der z-verdien representerer en test for at en koeffisient er lik null. I tabell 33 finner vi de variabler som vi tillot å påvirke sannsynligheten på en ikke-lineær måte. Der presenteres en såkalt scoretest for hvorvidt tilpassningen blir signifikant dårligere hvis man utelater en variabel. Det er ikke noen kvalitativ forskjell på en vanlig logistisk regresjon og den ikke-parametriske i dette fall når det gjelder hvilke variabler som skal være med i modellen. Vi ser også på hvor godt modellen klarer å klassifisere skatteyterne av de 10103. Resultatet presenteres i tabell 34.

Tabell 34: En estimert ikke-parametrisk logistisk regresjonsmodell

	PRED=0	PRED=1	Radsum
FRIV=0	9211	317	9528
FRIV=1	149	426	575
Kolsum	9360	743	10103

Prediksjonene er marginalt bedre enn de for vanlig logistisk regresjon. For å sammenligne vil vi også for denne modell se hvor god den er å fange opp de med FRIV=1 hvis de ikke vært med i estimeringen av modellen. Vi fulgte her samme prosedyre som for vanlig logistisk regresjon og fant at av de 100 med FRIV=1 klarte modellen av å finne 64 stykker, nettopp det samme resultatet som for vanlig logistisk regresjon.

7.2 Variabelutvelgelse

For logistisk regresjonsanalyse med (altfor) mange potensielle forklarende variable, fins det flere muligheter for mer eller mindre automatisk variabelutvelgelse. Blant disse er de tre trinnvise metodene: Baklengs eksklusjon, forlengs inklusjon, og full trinnvis regresjon, en variabel som er kommet til tidlig kan kastes ut igjen dersom den er overflødig i lys av de nye som er kommet til. Målet er å ende opp med et mindre antall variable som forklarer omtrent like godt som langt flere. Dette gir færre parametre å estimere, og dermed mindre estimeringsusikkerhet, noe som i sin tur kan ventes å gi mindre prediksjonsusikkerhet. Spesielt søker dette å unngå at sterkt korrelerte variable skal inngå regresjonen, noe som inflaterer estimerings- og prediksjonsusikkerheten. Et eksempel på dette så vi i avsnitt 7.1 der formue-variablene tilsynelatende ikke så ut til å forklare noe når de begge var med i modellen. Et alternativ her er såkalte ”shrinkage” metoder. Det er imidlertid flere problemer knyttet til slike halvautomatiske utvelgelsesmetoder, som vi ikke skal komme inn på her. Vi har ikke prioritert å

undersøke disse metodene nærmere, med ett unntak, såkalt LASSO (Least Absolute Shrinkage and Selection operator). Dette er en metode som setter restriksjoner på hvor store koeffisientene får være. Dette kan gjøres av ulike grunner, bl.a. fjerne variable som ikke forklarer mye, og unngå sterk korrelasjon mellom forklarende variabler. Vi har forsøkt LASSO som innfører restriksjonen slik at mange koeffisienter blir null, og hadde visse forhåpninger til denne, som foreløpig ikke er innfridd (se Kapittel 6). Denne forhåpning har sin grunn i at man på denne måte får vekk estimeringsusikkerhet. La oss se på dette for samme variabelsett som vi brukte på logistisk og ikke-parametrisk logistisk regresjon. Her er, for enkelhet skyld kun de numeriske variablene brukt.

Tabell 35: Sammenligning av koeffisienter i en logistisk regresjon med og uten LASSO

Variable	**Logistic**	**LASSO**
(Intercept)	-3.822096e+00	-3.173364e+00
KOMM	-5.098592e-06	1.534353e-05
STAT	3.940516e-04	2.794104e-04
F0807	-7.351229e-07	-3.211199e-07
ALMIB	1.085725e-06	1.381229e-06
SKJERM	8.949406e-05	2.317148e-05
LONN	5.522605e-08	0.000000e+00
ALDER	2.972154e-02	2.085975e-02
TOPPSKATT	1.676018e-05	4.420858e-06
OVER	3.392946e-06	0.000000e+00

Vi ser her at variablene LONN og OVER får verdiene null i LASSO-regresjonen. Modellen gir oss altså et hint om hvilke variabler som ikke er viktige. Notere også at variabelen OVER er valgt vekk til tross for at den var signifikant i den vanlige logistiske regresjonen. På den annen side har LASSO ikke valgt vekk KOMM og STAT som ikke var signifikante, sannsynligvis forklart av at de er meget sterkt korreleret. Disse aspekter bør studeres mer da de kan gi veiledning om spørsmålet om hvordan vi velger ut en kombinasjon av, kanskje svake, prediktorer og iblant til og med velger vekk tilsynelatende sterke prediktorer. Når det gjelder prediksjoner var vår erfaring med LASSO imidlertid i hovedsak at prediksjonene kun var marginalt bedre å klassifisere observasjoner som tilhører kontrollgruppen, men gjorde det på den annen siden en vesentlig dårligere jobb med frivillig gruppen. Vi skal imidlertid poengtere at restriksjonene på parametrene kan bestemmes med hjelp av data, noe vi hittil ikke har forsøkt. Her kan det være et potensiale for forbedringer.

8. Kjennetegnsanalyser: Klassifikasjonstre (CART)

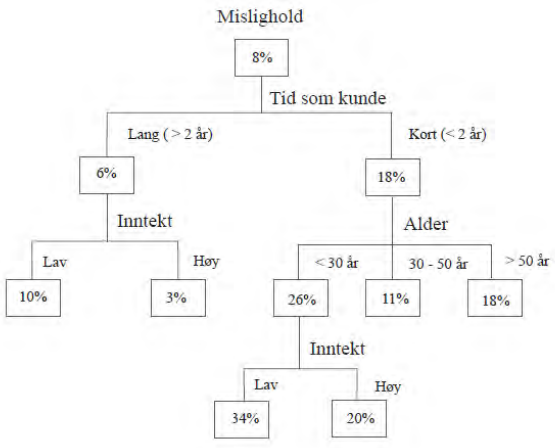
Dette kapitlet tar for seg klassifikasjonstre, som er den type statistisk tre som er aktuelt når det tale om gruppetilhørighet. Vi gir først en generell innledning, og deretter våre erfaringer basert på vårt tallmateriale. Hensikten er i første omgang å presentere metoden og utfordringene, heller enn å finne fram til den beste modellen. Siden den metoden vi ender opp med å anbefale kommer i dette kapitlet er fremstillingen noe mer omfattende, og følges opp av et eget kapittel mer de endelige analyser.

8.1 Hva er et klassifikasjonstre?

Klassifikasjonstre tar sikte på å trinnvis splitte populasjonen i homogene grupper etter visse regler, og organisere dette i en trestruktur med binære forgreninger ved såkalte noder ("recursive partitioning"). Toppnoden svarer til den binære avhengig variable (f.eks mislighold eller ikke), og ved hver ny node velges den variabel som best splitter de gjenværende objektene i to (evt. flere) grupper, slik at objektene i hver gruppe blir mest mulig "homogene". Splittingen stopper ved en terminal node når ingen variabel bidrar til å gjøre forskjellen mellom de to grupper i en splitt vesentlig eller når det ikke er nok observasjoner igjen på grenen til at man med rimelighet kan avgjøre fortsatt struktur, iht egnede kriterier. Metoden tillater ulike datastrukturer, og for numeriske variable X skjer splitten typisk ut fra om $X \leq c$ eller $X > c$, der c er konstant bestemt av data. Det anses fordelaktig å splitte i to framfor flere, idet dette gir best mulighet til å oppdage ikke-lineære strukturer, og utelukker ikke at etterfølgende splitt av samme variable gir det samme som splitt i flere enn to grener av gangen.

Eksempel: Mislighold av banklån

Til grunn er et datasett bestående av et større antall lånekunder, der det er registrert ulike personopplysninger som alder, inntekt og tid som kunde, samt om lånet er misligholdt eller ikke. De forklarende variable kan enten være skalavariable eller kategorivariable, men vi ser at i treet er disse redusert til et lite antall kategorier. I dette treet er mislighold en dikotom responsvariabel og en ser at 8% av kundene har misligholdt lånet. Treet nedenfor presenterer undergrupper som i ulik grad omfatter misligholdere. Vi ser at blant de som har vært kunde mindre enn 2 år er det 18% misligholdere, mens blant de med mer enn 2 år som kunde er det bare 6% misligholdere. En gruppe med et stort antall misligholdere er låntakere under 30 år med lav inntekt og kort kundeforhold.



Figur 3: Illustrasjon av klassifikasjonstre

Merk at et statistiske tre blir ofte kalt beslutningstre, en betegnelse som også er brukt i en noe annen sammenheng, og er unngått her. Under følger en punktvis liste over det som er karakteristisk ved denne metoden:

- Kan fange opp ”lokale”, spesielle og mindre åpenbare trekk ved data
- Håndtere også ikke-lineære sammenhenger og komplekse datastrukturer
- Håndterer store datamengder og mange variable
- Håndterer ”outliers” og manglende observasjoner
- Betydelig grad av automatisering
- Gir ikke formel med mulighet for å interpolere
- Kommuniserer godt via figurer
- Gir grunnlag for innsikt og læring
- Demonstrert godt resultat i mange kontekster

Hybrid analyse: En hybrid av statistisk og logistisk regresjon er foreslått, der man først lager et statistisk tre, der alle observasjonene er med og er tilordnet en eller annen av de terminale nodene. Man definerer så en kategorisk variabel som representerer hvilken terminal node en observasjon tilhører, i praksis representert ved binære variable, en for hver terminal node, som indikerer om observasjonen tilhører denne terminale noden eller ikke. Disse variablene brukes så i en logistisk regresjon, der det er mulig også å ha andre variable med. Det kan være variable som fanger opp globale effekter som er felles over alle noder, som typisk er svake og ikke oppdages i utviklingen av treet, men som likevel kan være betydningsfulle, og som man er oppmerksom på fra tidligere studier. Innenfor den oppsatte logistiske modell kan man så teste om slike globale effekter er tilstede. Mulighetene for å oppdage slike er nå også forbedret, siden eventuelle sterkere ikke-lineære effekter og interaksjoner er korrigert for med de kategoriske variablene som representerer de terminale nodene. Det er også mulig å innlemme variable som allerede fins i det statistiske tre. Merk at vi nå har en modell som tillater beregning av (estimerte) sannsynligheter for mislighold, og at tilleggsvariablene gir anledning til differensiering innen hver terminal node.

Merknad. Logistisk regresjon er mer sårbar for ”missing value” problematikk enn statistisk tre. Det spiller også inn for den hybride metode, men å droppe alle ”missing” observasjoner i den avsluttende logistiske regresjon, vil ikke ha samme mulige konsekvens som for ren logistisk regresjon. Det statistiske tre vil som regel gi bra resultat uansett.

Et viktig spørsmål i forbindelse med statistiske trær er når en skal stoppe å splitte. I prinsippet kunne man fortsette inntil alle sluttnodene kun besto av cases av de ene slaget (FRIV=1 eller FRIV=0). Det er både upraktisk og ikke spesielt lurt. Prøver en ”å presse sitronen” ekstra, oppstår lett såkalt overtilpasning (”overfitting”), der men har fått en løsning som i for høy grad har tilpasset seg særegenheter i tallmaterialet, og som ikke har prediktiv verdi for nye data. Det kan bli løst på ulikt vis. Sentralt i algoritmene er stoppekriterier og beskjæring (såkalt ”pruning”). Man kan for eksempel velge å fortsette å splitte inntil antallet eller andelen av cases i en sluttnode er under et visst nivå. Valg av størrelsen på tre kan også skje ved såkalt kryssvalidering (”cross validation”), som kan skje på ulikt vis. Den enkle muligheten å dele data i to datasett læringsdata og testdata, er mindre aktuelt her, siden vi har så få cases i den ene gruppen (FRIV=1). Alternativene er da såkalt V-fold cross validation eller ”global cross-validation”, der man i begge tilfeller gjentar byggingen av treet, men trekker ut en andel

som bruker som testdata, eksempelvis med V=3, hvor man splitter datasettet tilfeldig i tre like store deler, og bygger tre trær, der hvert deldatasett spiller rollen som testdata etter tur.

Statistiske trær tillater spesifikasjon av apriori sannsynligheter for kategoriene, og gir derfor mulighet beregne aposteriori sannsynligheter, gitt de observerbare karakteristika.

Andre utvidelser av statistisk tre

I litteraturen fins flere forslag til utvidelser av statistiske trær, som kan komme til nytte dersom et tre alene ikke strekker til. Det gjelder for eksempel dersom vi har mange variable, der hver og en ikke har særlig informasjonsverdi, men samlet sett kan ha det. Vi vil her gi en kort omtale av metoder der flere (mange) trær inngår, nemlig såkalte ”Boosting Trees” og ”Random Forests”.

”Boosting trees”: Består av en sekvens av enkle binære trær, der hvert nytt tre er basert på data som er en modifikasjon av dataene i det foregående tre, for eksempel det som er uforklart (”residualene”). Den endelige beslutningsregel kan da være en veid sum av de reglene som hvert tre gir, og der vektene er bestemt av data selv. Problemet med overtilpasning søkes løst ved såkalt ”stochastic gradient boosting”, der hvert nytt tre i sekvensen er basert på et tilfeldig utvalg av casene, og således ikke så lett tilpasser seg særegenheter uten prediktiv verdi. ”Boosting trees” har vist seg å være en god læringsalgoritme, også der de enkelte variable hver seg gir svake signaler. En ulempe kan være at det ikke utkrystalliseres en enkel lett forståelig beslutningsregel, men at prediksjonen langt på vei skjer som en ”black box”, selv om tilgjengelige algoritmer også kan peker ut variablenes (relative) betydning.

”Random Forests”: Metoden er i hovedsak som følger: Gitt N cases og M variable og velg hvor mange variable m man ønsker at beslutningsregelen skal være basert på. Lag et treningsdatasett på n cases, ved å velge disse tilfeldig (med tilbakelegging) fra de N tilgjengelige, der en kan bruke de gjenværende som testdata. Gruppetilhørigheten av disse kan så predikeres og et mål for prediksjonsfeilen beregnes. Deretter bygges et tre som følger: Ved hver node i treet, velges tilfeldig m av variablene, og beste splitt basert på disse bestemmes. Dette fortsetter til treet er ”utvokst”. Dette gjentas til at man har fått et ”ensemble” med trær. For et nytt case (fra test datasettet) bruker tar man et av trærne og noterer seg i hvilken sluttnode caset ender opp i, og hvilken gruppetilhørighet som da blir signalisert (1 eller 0). Dette gjentas for alle trær i ensemblet, og gruppetilhørigheten blir så predikert ut fra hva som forekommer mest (1 eller 0). ”Random Forests” har vist seg å være en god læringsalgoritme, som tillater store datamengder med mange variable, og er i liten grad hemmet av manglende observasjoner. En ulempe kan også her være at det ikke utkrystalliseres en enkel beslutningsregel, men at prediksjonen langt på vei skjer som en ”black box” (selv om tilgjengelige algoritmer også kan peker ut variablenes (relative) betydning.

Det fins en rekke varianter av statistisk tre tilgjengelig i programvare. Blant disse er Classification and Regression Trees (CART=C&RT) og CHi-squared Automatic Interaction Detector (CHAID), som i hovedsak kun skiller seg fra hverandre i valg av splitte-kritererier, uten at man kan si at den ene metoden generelt er bedre enn den andre. I Statistica finner vi i Data Mining modulen valgene: C&RT, CHAID, I-Trees, Boosted Trees, Random Forests. Her er I-Trees en interaktiv opsjon som dekker både C&RT og CHAID, mens de to siste valgene gir mulighet for å kombinere klassifikasjon fra flere trær, som hver seg ikke gir særlig god separasjonsevne. Tilsvarende muligheter finnes også som gratis nedlatbar programvare i det statistiske språket R.

8.2 Erfaringer med CART-analyser

Vi vil her rapportere våre erfaringer med C&RT i Statistica, med sideblikk til alternativene nevnt ovenfor. Vi tar i første rekke sikte på å illustrere metoden og dens potensiale, og vil ikke legge vekt på at enkelte av de variable som inngår kan vise seg unyttige for Skatteetaten i sine prioriteringer.

Vi tar også sikte på å illustrere noen av de problemer som er knyttet til at gruppen av frivillig rettede er svært liten i forhold populasjonen av alle skatteytere, ved først å betrakte en situasjon der de frivillig rettede (FRIV=1) utgjør om lag 500 av 100 000 (0.5%). Vi har likevel et "nåler i en høystakk" problem, med de konsekvenser det har for virkelig å finne de individer som er samme kategori, når gruppetilhørigheten ikke er kjent. Vi vil så se på hva som skjer dersom de frivillig rettede utgjør om lag 500 av 10 000 (5%). Vi gjør dette uten å bry oss om disse andelene er realistiske, eller om alle funne variable er relevante for praksis. Om sjansene for å finne (FRIV=1) er lav, vil likevel kunnskap om variable som potensielt har prediktiv verdi, kunne hjelpe ved prioriteringen av kontrollaktiviteter i skatteetaten. Formålet er her altså i hovedsak å belyse metoden, og samtidig gi en viss innsikt i hvilke variable som kan ha verdi eller er uten verdi for oss i det videre arbeid. Vi vil i første omgang begrense oss til bakgrunnsvariablene og variablene fra 2008 selvangivelsen, slik de fremkom i datafilen fra leveranse 1 (som også inneholdt 2007 tallene).

8.2.1 CART (FRIV=1, 0.5 %, "default").

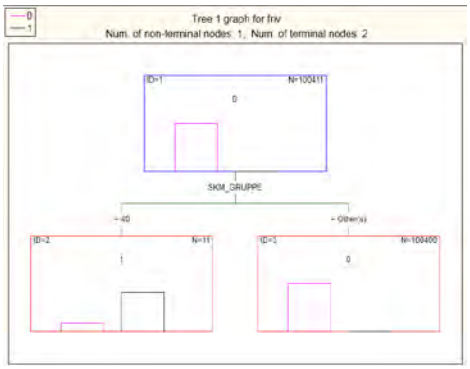
Vi vil finne variable som kan ha prediktiv verdi for frivillig retting (FRIV=1) eller ikke (FRIV=0). Med svært mange variable er det hensiktmessig å innlemme disse gruppevis, ikke minst fordi manglende observasjon vil redusere antall tilgjengelige cases. Det viser seg også at ved å spesifisere alle 2008-variable fikk vi overhodet ingen splitt med "default" valget.

Først ble alle 11 kategorivariable fra SENTRALITETSKODE til FORSINKELSES_AVGIFT_KODE spesifisert, herunder SKM_GRUPPPE og HISTORIKK_KODE (KJØNN og SIVILSTAND ble holdt utenfor). Her ble kun SKM_GRUPPE (skjemagruppe) plukket ut, og der casene med kodetall 40 ble pekt ut til å omfatte mange med frivillig retting. Dette var lovende, men ettersom SKM_GRUPPE=40 betegner avdød, har denne variabelen likevel ikke nyttig i prediktiv forstand. Det kan likevel være instruktivt å presentere utskriften, da de illustrerer noe av potensialet og utfordringene med en "lett forståelig" variabel.

Utskrift 1: Fra Statistica

SENTRALITETSKODE
KLASSIFISERINGSKODE
SKM_GRUPPE
PERSON_KODE
HISTORIKK_KODE
PSA_KODE_FP
PSA_RETTE_KODE
SAMSKATT_KODE
SA_FRITAK_KODE
HAR_LEVERT_SA
FORSINKELSES_ASVGIFT_KODE

N=100411 SKM_GRUPPE	friv 0,	friv 1,	Row Totals
10,	92670	543	93213
13,	1213	2	1215
14,	5126	19	5145
18,	429	0	429
30,	369	2	371
28,	9	0	9
70,	14	2	16
40,	2	9	11
20,	1	0	1
12,	1	0	1
All Grps	99834	577	100411

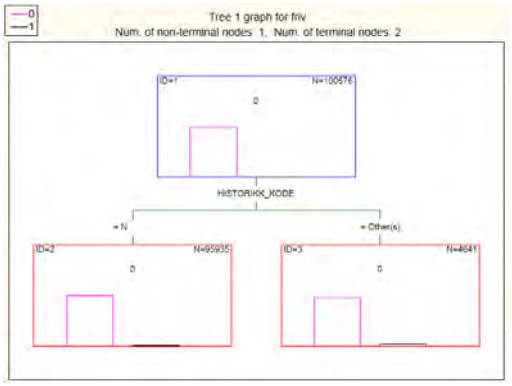


Vi ser av tabellen at blant de N=100 411 case er det 577 med frivillig retting. Langt de fleste av disse, i alt 543, er SKM_GRUPPE=10, men de utgjør bare 0.5% av denne store gruppen 93 213 caes. For SKM_GRUPPE=70 (diplomater) med kun 16 cases ser vi at 2 har frivillig retting, noe som utgjør 14%.

SKM_GRUPPE=40 (avdød) peker seg ut med 11 cases, hvorav hele 9, dvs. 82%, har frivillig retting. Det er dette som er fanget opp i figuren til høyre. Det er ingen annen gruppe som kan oppvise tilsvarende "oppdagingsevne", og heller ingen annen av de spesifiserte kategoriske variable som kan oppnå tilsvarende, eller såpass mye at de er verd å supplere treet med ytterligere splitting. Dette kan endres ved å myke opp innlemmelseskriteriet. Figuren viser splittingen av de N=100411 cases mht variabelen SKM_GRUPPE, hvorav de aller fleste ikke hadde noen frivillig retting, illustrert med høy rød søyle i øverste del av figuren, med en nesten usynlig sort søyle ved siden. Deretter blir SKM_GRUPPE valgt ut som de enste variabelen som klarer å skille (godt) mellom casene med frivillig retting/ikke (FRIV=1, FRIV=0), og at det er kodetall=40 som gjør forskjellen. I venstre boks finner vi de 11 avdøde, hvorav det var frivillig retting for en langt de fleste (høy sort søyle). I den høyre boksen finner vi de resterende casene med de øvrige skjemagruppene, der vi igjen har en høy rød søyle. Også her er det et stort antall med frivillig retting blant de mer enn 100 000 casene, men deres andel er så liten at det ikke er synlig i grafen. Merk at tallet inne i figuren (0 eller 1) angir hvilken gruppe som er i flertall.

Vi prøvde også å spesifisere variable en av gangen, og den blant de øvrige kategorivariable i den første gruppen som viste tegn "til liv" var HISTORIKK_KODE (Tidligere skatteberegnet N=Nei, J=Ja) med følgende resultat:

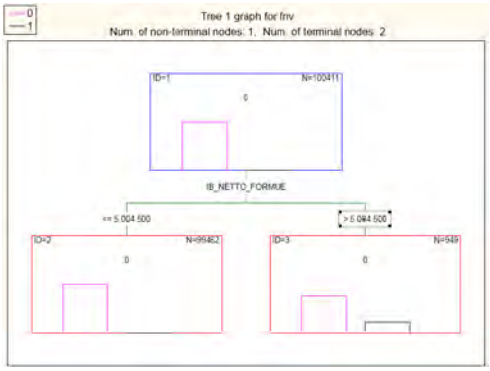
Utskrift 2: Fra Statistica



N=100576 HISTORIKK_KODE	friv 0,	friv 1,	Row Totals
N	95539	396	95935
J	4460	181	4641
All Grps	99999	577	100576

Vi ser at vi får forholdsvis større andel med frivillig retting for Ja-gruppen, men andelen er liten, bare 181/4460=4%. Dette illustrerer muligheten til å bruke en (eller flere) variable til prioritering av innsats, selv om en ikke får isolert de enkelte med frivillig retting spesielt godt.

Neste steg i den innledende eksplorative analysen var å holde på variabelen SKM_GRUPPE og innlemme de etterfølgende 12 IB-variablene, som alle er numeriske variable. Vi fikk da splitt mhp. den ene variablen IB_NETTO_FORMUE, som vist i figur 5 nedenfor.

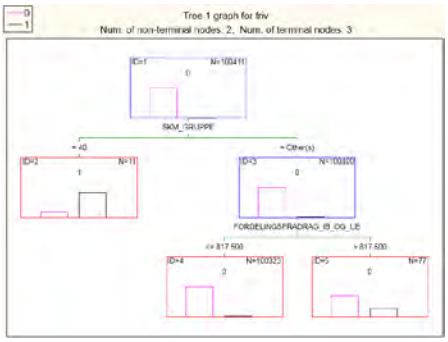


Figur 4: Klassifikasjonstre

Vi ser at programmet har valgt for oss å kategorisere i to kategorier, de med formue under og over 5 004 500. I den siste kategorien er 949 cases, hvorav ca. 25% har frivilllig retting.

Merknad. Vi fikk ikke denne splitt når IB-variablene ble spesifisert uten SKM_GRUPPE og heller ingen splitt når IB_NETTOFORMUE ble spesifisert alene!

Vi forsøkte så SKM_GRUPPE (og HISTORIKK) i kombinasjon med noen av de andre numeriske variablene, og fikk:

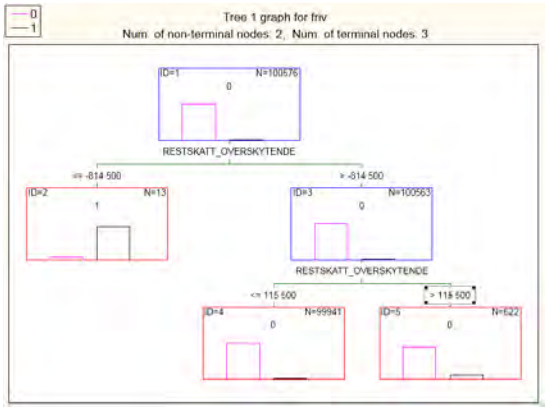


Figur 5: Klassifikasjonstre

Her ser vi at vi først får splitt mhp SKM_GRUPPE=40 eller =Other, som tidligere, og at vi i den siste gruppen får en ytterligere splitt mhp FORDELINGSFRADRAG_IB_OG_UB. Blant de med fradrag over 817 500 finner vi en betydelig andel med frivillig retting.

Merknad. Ved spesifikasjon av IB_NETTOFORMUE, kom denne ut alene, med samme graf som ovenfor! Det samme gjaldt dersom vi la til de øvrige IB-variablene.

Vi forsøkte så flere av de andre numeriske variablene, alene eller i kombinasjon. Av interesse er RESTSKATT_OVERSKYTENDE alene, som ga resultatet:



Figur 6: Klassifikasjonstre

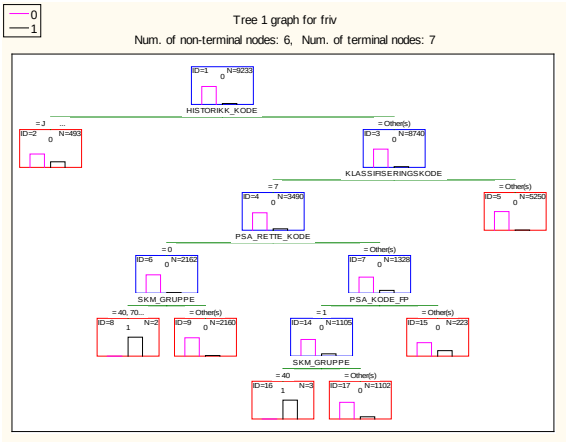
Her ser vi at vi først får en splitt for verdier under og over verdien -814 500, og der vi i den første gruppen har en svært stor andel med frivillig retting. Den siste gruppen blir igjen splittet i under og over 115 500. I den siste gruppen er det en ikke ubetydelig andel med frivillig retting. Dette er et eksempel på at for enkelte variable kan både høy og lav verdi indikere frivillig retting. Dette vil en lett kunne overse dersom man kun holdt seg til lineære (regresjons-)modeller.

Med dette har vi illustrert metoden for søk etter variable og tilhørende beslutningsregel i en "nål i høystakk" situasjon. Vi har sett at antall variable som plukkes ut er lite, og beslutningsreglene enkle når default-kriteriene brukes. Spesifiserer vi mer romslige innkluderingskrav, vil vi selvfølgelig få mer kompliserte strukturer, men med økt risiko for falske eller unyttige forslag.

Merknad. Ved spesifikasjon av mange potensielle variable, kan en risikere at mange cases med manglende observasjon på en eller flere av variablene blir fjernet, noe som kan gi tilsynelatende merkelige forskjeller ved endring av spesifikasjonene.

8.2.2 CART (FRIV=1, 5%, "default")

Vi skal nå se hvordan metoden fortoner seg dersom de frivillig rettede utgjør en betydelig større andel. Vi har latt kontrollgruppen bestå av 10 000 cases istedenfor 100 000 (her er brukt de 10 000 første i datafilen). Resultat ved spesifikasjon av de 11 kategorivariablene (fra SENTRALITETSKODE til FORSINKELSES_AVGIFT_KODE) ble:



Figur 7: Klassifikasjonstre

Vi ser at vi nå fikk et mer omfattende beslutningstre, der vi kan lete etter de med frivillig retting i boksene ved endene av treet (terminale noder) med høye sorte søyler, eventuelt ikke neglisjerbare slike. Vi ser at første splitt er mhp HISTORIKK_KODE, der kode=J har et betydelig antall frivillig rettede. Blant de med kode=Other følger splitting mhp KLASSIFISERINGSKODE, der vi får skilt ut kode=7 (sentrale tjenesteytingskommuner) som mest interessant. For disse følger splitt mhp PSA_RETTEKODE, der kode=0 gir splitt mhp SKM_GRUPPE, der vi finner utpekt to frivillig rettede med kode=40,70, mens PSA_RETTEKODE=1 gir splitt mhp PSA_KODE_FP, der kode=1 med over 100 cases igjen gir splitt mhp SKM_GRUPPE, der kode=40 peker ut tre cases, mens det fortsatt er et betydelig antall frivillig rettede for de øvrige kodene som ikke blir isolert. For PSA_KODE_FP kode=Other gjenstår over 200 cases, hvorav en betydelig andel er frivillig rettede. Dette treet illustrer en del gode poenger: Vi fikk først en splitt som straks pekte ut en gruppe med betydelig andel frivillig rettede, mens den andre gruppen har en undergruppe (sentrale tjenesteytende kommuner) med en mindre andel frivillig rettede, der ytterligere variable har prediktiv verdi, mens slike variable ikke har (tistrekkelig) prediktiv verdi i de andre typer kommuner.

Merknad. Hvis vi i tillegg spesifiserer KJØNN og SIVILSTAND, stoppet høyre forgrening i figuren med PSA_RETTE_KODE=Other, med fortsatt et betydelig antall frivillig rettede skjult, mens PSA_RETTE_KODE=0 fikk splitt mhp SIVILSTAND, der kode=8 (=skilt partner) pekte ut to frivillig rettede, og kode=Other fikk splitt mhp SKM_GRUPPE, der kode=40, 70 pekte ut nye to.

Programmet klassifiserer automatisk (”by default”) til FRIV=0 og 1 iht. majoritet i den terminale boks. Det er vist i tabellen nedenfor. I dette tilfellet var det bare 5 tilfeller med klassifikasjon til FRIV=1 alle korrekte, mens det ligger 453 frivillig rettede blant de som blir klassifisert til FRIV=0. (At vi har kun 9609 cases totalt skyldes de manglende observasjoner på flere av variablene.)

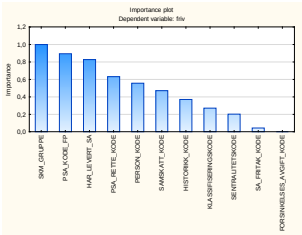
Tabell 36: Klassifikasjonsmatrise fra Statistica

	Classification matrix 1 (Datasett Lev2 Baard 10000 I Dependent variable: FRIV Options: Categorical response, Analysis sample			
	Observed	Predicted 0	Predicted 1	Row Total
Number	0	9151		9151
Column Percentage		95.28%	0.00%	
Row Percentage		100.00%	0.00%	
Total Percentage		95.23%	0.00%	95.23%
Number	1	453	5	458
Column Percentage		4.72%	100.00%	
Row Percentage		98.91%	1.09%	
Total Percentage		4.71%	.5%	4.77%
Count	All Groups	9604	5	9609
Total Percent		99.95%	.5%	

Vi kan også få en tabell som gir variablenes prediktive betydning eller en tilsvarende graf, se utskrift 3.

Utskrift 3: Fra Statistica

	Predictor importance 1 (Datasett Lev2 Dependent variable: FRIV Options: Categorical response, Tree r	
	Variable rank	Importance
PSA_KODE_FP	100	1,000000
SKM_GRUPPE	91	0,906082
HAR_LEVERT_SA	90	0,897113
PERSON_KODE	71	0,706219
PSA_RETTE_KODE	65	0,648104
SAMSKATT_KODE	47	0,471488
HISTORIKK_KODE	45	0,451826
KLASSIFISERINGSKODE	28	0,275792
SENTRALITETSKODE	23	0,227945
SA_FRITAK_KODE	3	0,030732
FORSINKELSES_AVGIFT_KODE	0	0,000000



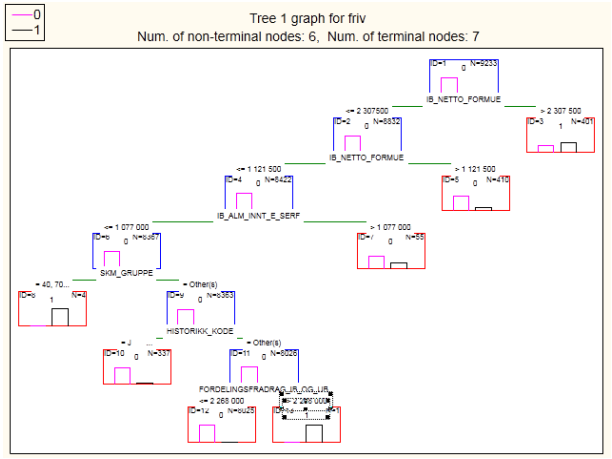
Vi ser at SKM_GRUPPE kommer ut på topp, etterfulgt av PSA_KODE. Noe overraskende ser vi at enkelte variable som er ikke kommet med i treet (f.eks HAR_LEVERT_SA) ligger høyere på listen enn variable som var med i treet. Dette forklares med at de kan være gode substitutter for en eller flere av de variable som er med i treet, og derfor er tatt med i det rapporterte mål for "prediktiv betydning". Uten dette vil vi eksempelvis kunne risikere at vi aldri ble oppmerksom på en variabel som ikke var best i noen splitt, men eksempelvis var nest best i alle splitt. Enkelte andre programmer rapporter prediktiv betydning uten å ta omsyn til dette. Vi ser også at vi samtidig har fått holdepunkter for å se helt bort fra de to siste variablene på listen. Fjerner vi disse, får vi samme treet (men en liten økning i antall cases).

Det er av en viss interesse å foreta denne type sammenligninger av variablenes betydning, både mellom ulike metoder, og valg mellom ulike opsjoner for den enkelte metode, og om mulige utvidelser av antall cases med frivillig retting kan spille noen rolle.

En sammenligning med hvordan "Boosted Trees" metoden ser på variablenes betydning er foretatt. Det viser seg at bildet er nokså ulikt. Her kommer HISTORIKK_KODE ut på topp etterfulgt av de øvrige variable med nokså lik vekt, unntatt de samme to siste. Dette kan virke noe forvirrende, men vi har i hvert fall holdepunkter for å se helt bort fra disse to variablene.

Hvilke av de terminale nodene som skal følges opp først vil kunne avhenge av omstendighetene, ikke minst hvor mange cases det er i angjeldende boks. At vi nå har avdekket et mer komplisert mønster, henger sammen med at andelen frivillig rettede er større, og at det dermed er lettere å finne slike.

Vi går videre med nye variable: Som ovenfor legger vi til IB-variablene, utover de kategorivariablene som er avdekket (det ser ikke ut til å spille noen rolle om vi her tar med alle). Dette gir følgende bilde:

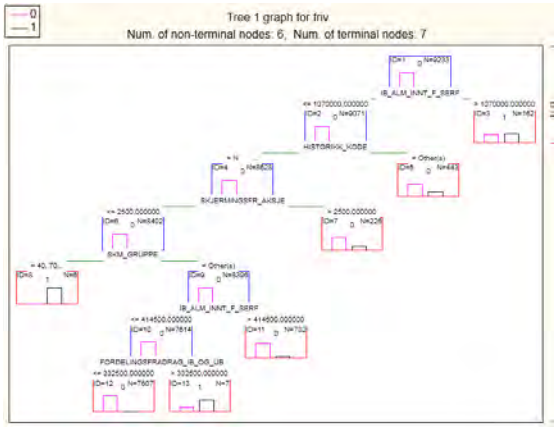


Figur 8: Klassifikasjonstre

Vi ser at flere av kategorivariablene ovenfor nå er borte, noe som kan henge sammen med at de er bærere av noe av den samme informasjon som noen av de tilkomne numeriske variable. Nå splittes først IB_NETTO_FORMUE, der høy formue betyr overvekt av frivillig endring, og er interessant uten ytterligere informasjon. Den neste splitt er også mhp denne variabelen, som sett i sammenheng innebærer at variabelen er splittet i tre kategorier. Her blir det en vurdering om også den midtre kategorien er interessant uten ytterligere informasjon. For den lave kategorien er neste splitt mhp IB_ ALM_INNT, der høy indikerer noe frivillig retting. For de med lav alminnelig inntekt, kan

SKM_GRUPPE=40 og 70 være av interesse. For de andre skjemagruppene, vil HISTORIKK=Other og høyt fordelingsfradrag tilsynelatende være av interesse. Her har vi imidlertid bare ett case tilbake, så noen generell konklusjon her er ikke tilrådelig.

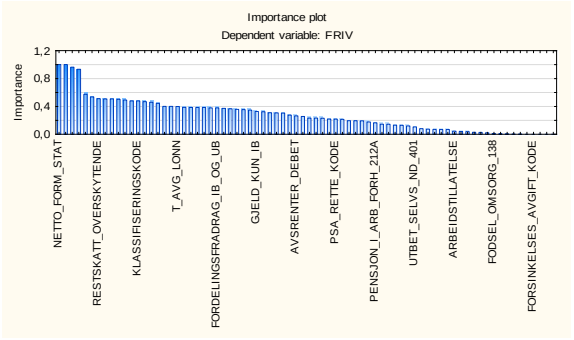
Vi har fortsatt med å legge nye variable til de som har vist seg "levedyktige" ovenfor. Ved å legge til de etterfølgende kategorivariable fra 2008 fikk vi bare at STATSBORGERSKAP erstattet IB_ALM_INNTEKT, og at det var kode=USA som erstattet høy inntekt, og med et betydelig antall frivillige rettinger. Beholder vi inntektsvariabelen, og legger isteden til flere numeriske variable (fra NEG_ALM_INNTEKT t.o.m AVSRENTER_DEBET), får vi i hovedsak at NETTO_FORM_STAT erstatter IB_NETTO_FORMUE, der de med høy og middels er direkte interessante, mens de med lav formue er har en tilsvarende struktur som ovenfor, der FORDELINGFRADRAG går ut og RESTSKATT_OVERSKYTENDE kommer inn i en kombinasjon med de øvrige kjente variable, men som leder til en boks med bare ett case. Dette gir grunnlag for igjen å returnere til variablene i grafen ovenfor, med sikte på å legge til variable som i større grad komplementer disse. Forsøker vi nye numeriske variable t.o.m NÆRINGSOVERSKUDD, kommer SKJERMINGSFR_AKSJE inn, og vi har en struktur som ikke avviker mye fra den i grafen ovenfor, men har flere cases i de terminale nodene. Med faglig innsikt i hva disse variablene representerer, kan det være av interesse å holde de to grafene opp mot hverandre:



Figur 9: Klassifikasjonstre

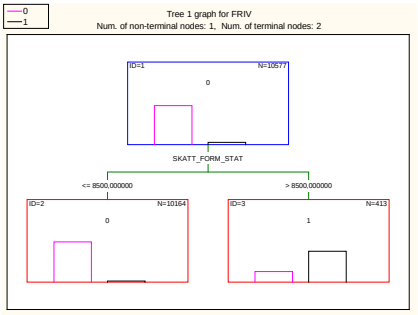
Vi går videre med modifikasjon av den siste spesifikasjonen som hadde fire kategorivariable og fire numeriske variable. Fremforhandlet utenlandsformue og inntekt bidrar ikke til noe utover dette. Hva så med T-variablene? Av disse ser T_GRL_PENSJON ut som kan ha en viss verdi (men denne indikerer muligens frivillig retting av andre grunner enn de vi er på jakt etter). Toppskattvariablene ga tilsynelatende ikke noe nytt. Spesifikasjon av de øvrige numeriske 2008-variable frem til 07-variablene ga et langt mer omfattende tre, der flere pensjonsvariable dukket opp. Mønsteret var imidlertid komplisert, men kunne kanskje lede til noe mer fruktbart ved å utelate enkelte av variablene som ikke har særlig praktisk betydning (f.eks bare leder til terminal node med ett enkelt case).

På dette stadium er det utført en rekke alternative analyser, herunder "Boosted Trees" og "Random Forests". For disse to gjelder at vi ikke i samme grad rammes av manglende observasjoner ved spesifisering av mange variable. En "Boosted tree"- analyse med alle kategori- og numeriske 08-variable, ga fire variable som betydelig viktigere enn de andre.



Figur 10: Variabel viktighet

Disse var NETTO_FORMUE_STAT, SKATT_FORMUE_STAT; SKATT_FORMUE_KOMM og IB_NETTO_FORMUE. For de øvrige variable var betydningen noe lavere og jevnt fallende, uten en naturlig gruppe av nest beste variable. Det som da kommer ut er en splitt mhp. en av disse variablene, SKATT_FORMUE_STAT:



Figur 11: Klassifikasjonstre

Det er ikke overraskende at alle disse fire er like viktige, og at det som kommer ut er en splitt mhp en av dem. Dette reflekter selvsagt det faktum at de er sterkt korrelerte, og at en av dem er nok til å representere alle. Denne form for Importance-plott gir altså ikke noe godt grunnlag for å vurdere den partielle virkningen av å legge til en variabel.

8.2.3 Feature selection and variable screening

Statistica har dette som valg på Data Mining menyen. Det er noe uklart hva "feature selection" egentlig lover oss, og etter hvilke underliggende kriterium rangeringen av variable skjer, og om kovariasjon tas omsyn til. Man har imidlertid valget mellom å spesifisere hvor mange prediktorer man vil ha (default=10), eller at man får alle prediktorer med P-verdi mindre enn et spesifisert nivå (default=0.01), men det ser ut til kun å styre mengden av output og ikke hvor komplisert modell man vil ha. Ved spesifikasjon av alle variable før 07-variablene, og utskrift av alle, får vi en tabell med variablene ordnet etter kjikvadratverdier, med en tilhørende P-verdi. Her er utsnitt av toppen og bunnen i tabellen:

Utskrift 4: Beste prediktorer

	Best predictors for categoric	
	Chi-square	p-value
NETTO_FORM_STAT	2730,652	0,000000
IB_NETTO_FORMUE	2708,192	0,000000
SKATT_FORM_KOMM	2613,764	0,000000
SKATT_FORM_STAT	2609,790	0,000000
SUM_SKATT_AVGIFT	1983,087	0,000000
IB_ALM_INNT_F_SERF	1539,916	0,000000
IB_ALM_INNT_E_SERF	1485,534	0,000000
ANTALL_UB	1396,597	0,000000
SKJERMINGSFR_AKSJE	937,801	0,000000
HISTORIKK_KODE	819,618	0,000000
T_GRL_PENSJON	633,511	0,000000
PENSJON_INNTEKT	633,511	0,000000
NEG_ALM_INNT_KUN_IB	624,712	0,000000
FORDELINGSFRADRAG_IB_OG_UB	624,712	0,000000
AVSRENTER_DEBET	624,671	0,000000
TOPPSKATT	607,016	0,000000
T_AVG_PENSJON	607,009	0,000000
TOPPSKATTEGRUNNLAG	562,366	0,000000
IB_GODTGJ_NER	21,960	0,004990
ANT_BILER	19,114	0,007838
RESTSKATT_OVERSKYTENDE	17,333	0,000172
ANV_TOLVDEL	15,895	0,319848
SYKEPENGER_SELVS_ND_405	11,968	0,002519
T_GRL_JSF	11,107	0,085136
T_AVG_JSF	11,107	0,085136
AFP_227	5,250	0,730513
DAGPENGER_147	5,249	0,730617
IB_P_INNT_JSF_U_REF	4,375	0,626074
FODSEL_OMSORG_138	3,777	0,876637
SA_FRITAK_KODE	2,846	0,241030
LONN_KONK_123	1,367	0,849979
TRYGDEGR_KODE	1,086	0,581044
TILLEGGSSKATT	0,559	0,755999
TOPPSKATTEGR_KORR	0,058	0,971508

Vi merker oss at P=0.0000, som betyr klar statistisk signifikans, går svært langt ned i tabellen. Denne tabellen samsvarer med tilsvarende tabell for Boosted tree når det gjelder topp og bunn, men mange innbyrdes forskjeller i midten, for de marginalt betydningsfulle variablene.

8.2.4 Klassifikasjonsevne

Som tidligere nevnt tallfester CART (og dens alternativer) også klassifikasjonsevnen på ulikt vis, både ved hyppighetene for korrekt og feil klassifisering, og som opsjon også beregnede kostnader etter spesifikasjon av kostnader ved de to typer feilklassifisering. Vi har hittil holdt dette utenfor, både fordi de absolutte tall avhenger av størrelsen på kontrollgruppen, som her er nokså vilkårlig, og valgt ut fra andre vurderinger enn at den i størrelse representerer populasjonen (alle skatteytere). Det kan imidlertid ha en viss interesse å se litt nærmere på forskjeller mellom ulike metoder og utvalg av variable for samme kontrollgruppe.

Nedenfor viser vi to slike tabeller. Først en for en tidligere attraktiv CART-spesifikasjon med fire kategorivariable og fire numeriske variable, og deretter en "Boosted tree" spesifikasjon med alle variable 2008-variablene.

Utskrift 5: Klassifikasjonsmatriser fra CART (til venstre) og "Boosted trees" (til høyre)

	Classification matrix 1 (Datasett Lev2 Baard 10000 k Dependent variable: FRIV Options: Categorical response, Analysis sample			
	Observed	Predicted 0	Predicted 1	Row Total
Number	0	9443	125	9568
Column Percentage		97.80%	33.69%	
Row Percentage		98.69%	1.31%	
Total Percentage		94.19%	1.25%	95.43%
Number	1	212	246	458
Column Percentage		2.20%	66.31%	
Row Percentage		46.29%	53.71%	
Total Percentage		2.11%	2.45%	4.57%
Count	All Groups	9655	371	10026
Total Percent		96.30%	3.70%	

	Classification matrix (Datasett Lev2 Baard 10000 kor Response: FRIV Analysis sample;Number of trees: 179			
	Observed	Predicted 0	Predicted 1	Row Total
Number	0	9281	703	9984
Column Percentage		99.15%	58.68%	
Row Percentage		92.96%	7.04%	
Total Percentage		87.90%	6.66%	94.55%
Number	1	80	495	575
Column Percentage		0.85%	41.32%	
Row Percentage		13.91%	86.09%	
Total Percentage		0.76%	4.69%	5.45%
Count	All Groups	9361	1198	10559
Total Percent		88.65%	11.35%	

Vi ser at CART klassifiserer 53.7% av de frivillig rettede korrekt, mens "Boosted tree" hele 86.1%. Samtidig klassifiserer CART 86.7% av kontrollgruppen korrekt, mens "Boosted tree" greier hele 93.0%. "Boosted tree" gjør det tilsynelatende langt bedre. En må imidlertid være oppmerksom på at dette er en "within sample" vurdering, og en reell "out of sample" vurdering med et nytt testdatasett, vil typisk gi lavere korrekte prosenter. Igjen gjøres oppmerksom på at disse tall er knyttet til den valgte størrelsen på kontrollgruppen (her 10 000). Med større kontrollgruppe selvsagt prosenten korrekt klassifiserte blant de frivillig rettede bli lavere, men fortsatt høy i kontrollgruppen, simpelthen fordi

langt de fleste hører hjemme der (og det er liten risiko ved å tippe at en tilfeldig person tilhører denne gruppen). Våre analyser ga tilsynelatende en indikasjon på "Boosted tree" ikke fungerer like godt "out of sample", i hvert fall ved spesifikasjon av mange variabler. Det må her skytes inn at det bare er "Boosted tree" som kan gi en slik indikasjon, uten å ha et helt nytt testdatasett, og at indikasjonen egentlig rammer begge metoder. Vi har altså en situasjon der innlemmelse av flere variable kan gi bedre "within sample" tilpasning, men muligens dårligere "out of sample" egenskaper. Dette er en situasjon som ikke er ukjent ellers.

Vi har også forsøkt CART på alternative programvareplatformer, herunder R (gratis) og XLSTAT (utbredt rimelig tilleggsprodukt til Excel). Vi vil kort rapportere våre erfaringer.

8.2.5 Representative trær

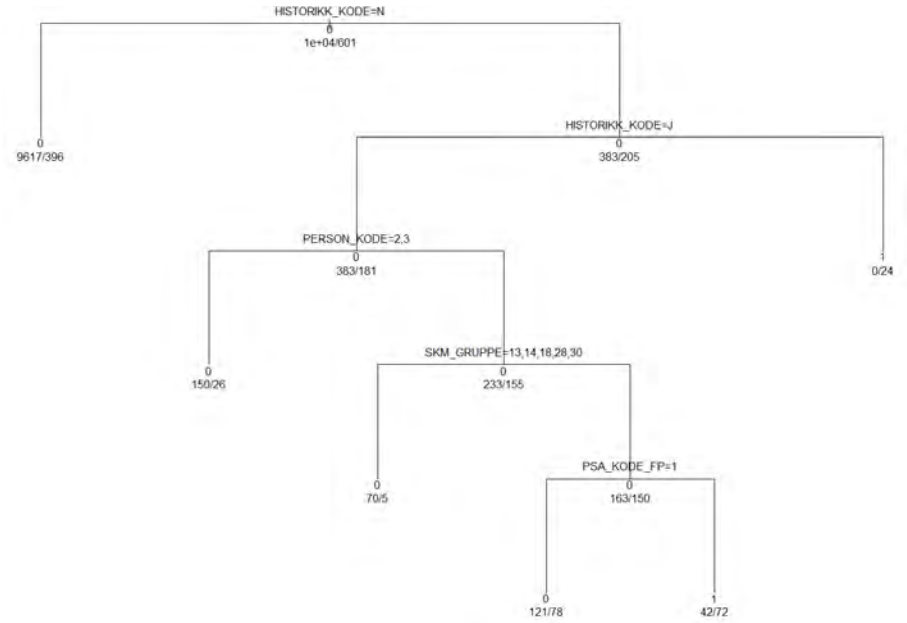
Spørsmålet om hvilke variable som har størst betydning for klassifikasjonsevne er ikke lett, og forsøk på rangering kan lett bli misvisende. Blant annet er det ikke uvanlig at trær med ulik trestruktur (topologi) og ulike variable gir omtrent samme klassifikasjonsevne. Dette kan oppleves i praksis ved at to analyseprogrammer kan produsere svært ulike trær, uten at dette behøver å bety at det er noe feil eller at et program er bedre enn et annet. Splittingen skjer som en trinnvis prosess med et mer eller mindre "nærsynt" splittekriterium på hvert trinn, og med ulikt valg av kriterium, som hver for seg er like rimelige, kan en ikke vente at topologien i det ferdige treet blir den samme. Vi har sett at en klassifikasjonsmetode som kombinerer mange trær kan bety at svært mange variable er med på å bestemme gruppetilhørigheten. Bildet kompliseres ved at samme variable, eller en sterkt korrelert variabel, kan forekomme flere ganger i ulike trær, og i ulike kombinasjoner med andre variable. Dette gjør at vi "ser ikke skogen for bare trær".

Det kan oppfattes som en ulempe at de beste klassifikasjonsmetodene har karakter av "en sort boks", som ikke gir direkte grunnlag for vurdering av topologien. Det kan være ønskelig å ikke bare ha et tallmessig mål på den enkelte variables betydning, men også ha innsikt i hvordan variablene henger sammen. Dette kan bidra til å skape innsikt og diskusjon. Kan hende er det ikke slik at formålet med analysen ikke er å lage en beslutningsregel til slavisk bruk, men at formålet heller er beslutningsstøtte sammen med annen informasjon.

Et alternativ til å rangere variablene basert på et mer eller mindre velvalgt kriterium, er å forsøke å lage et knippe representative trær. Dette er ikke så håpløst som det kan høres ut, idet mange av de trærne som inngår typisk vil skille seg fra hverandre i et lite antall noder, mens andre kan tilsynelatende ha ulik struktur, men likevel gi nokså like klassifikasjoner. En mulighet er å definere et avstandsmål mellom par av trær som anvendes på skogen av trær. Man foretar så en klyngeanalyse, for eksempel basert på såkalt multidimensjonal skalering (MDS). Dette gir oss et "kart" der hvert tre er representert som et punkt, og det typisk vil være klynger av trær, som er like hverandre etter det valgte avstandsmål. Man kan så velge ut et representativt tre fra hver klynge, enten det mest sentrale eller det som er gitt størst vekt. Denne metode beskrives i detalj i Chipman, George og McCulloch (1998).

8.2.6 R-rpart: Recursive partitioning and regression trees

I situasjonen med N=10 000+577+24 observasjoner (de 24 med manglende 2007 verdier) ga spesifikasjonen av de 11 første kategorivariablene følgende tre:



Figur 12: Klassifikasjonstre fra R

Her markerer koden til splittevariablen de de kategorier som sendes videre til venstre gren. Symbolet ved de enkelte noder angir prediksjonen FRIV=0 eller 1 med de . Treet er noe enklere enn treet fra Statistica, men samsvarer i hovedsak med dette. Her er riktignok PERSON_KODE retningsgivende (kode=2,3 fjerner ektefelle/partner og barn) istedenfor KLASSIFISERINGSKODE og PSA_RETTE_KODE. Den terminale noden i bunnen med FRIV=1 svarer som før til SKM_GRUPPR (kode=40, 70), som var avdøde og diplomater og PSA_KODE_FP (kode=0), som er PSA uten endring. Den terminale noden til høyre svarer til de 23 observasjonene uten registrerte 2007-verdier, som alle tilhørte gruppen av de frivillig rettede. At vi har fått en terminal node ved hver splitt, beror på en tilfeldighet.

8.2.7 XLSTAT: Classification and regression trees

I XLSTAT har man valget mellom CHAID, C&RT og QUEST. Av brukerstøtten kan det se ut som den generelle anbefalingen favoriserer CHAID. Vi har likevel brukt tiden på C&RT for å kunne sammenligne direkte med tidligere resultater. Man har valget mellom to splittemetode: Twoing og Gini. Begge ga betydelige forskjeller fra våre funn med Statistica. I situasjonen med N=10 000+577 ga spesifikasjonen av de 11 første kategorivariablene følgende (med opsjonen Twoint):

Første splitt mhp KLASSIFISERINGKODE (1,3,6) vs (2,4,5,7), som i sin tur ble splittet i henholdsvis PERSONKODE (1,2) vs 3 og SENTRALITETSKODE (0,2) vs (1,3). Deretter kom nye splitt der de samme tre variable opptrer i ulike kombinasjoner i et mønster som virket uryddig og forvirrende, og der de oppståtte grupper i de enkelte splitt oftest ikke har noe lett forklarlig fellestrekk. I Statistica var HISTORIKK_KODE den første splitt, og det er påfallende at KLASSIFISERINGSKODE er den eneste variabelen som er felles. Det var ingen klare bokser, der FRIV=1 gjorde seg spesielt sterkt gjeldende i de terminale noder. Dette kan skyldes en ”default” spesifikasjon med. en satt nedre grense antall

cases som tillates i en terminal node, men hvis dette reduseres blir det svært mange nivåer og et enda mer uryddig bilde. Dette kan muligens kompenseres med alternative parametervalg, noe som krever betydelig innsikt i "foredlingen" av treet. Med denne erfaringen kan produktet ikke anbefales.

9. Kjennetegnsanalyse: Hovedresultater

I dette kapittelet presenterer vi noen omfattende analyser som tar sikte på å gjøre klassifikasjonen av de skatteytere vi har mottatt data over, så god som mulig. Vi vil nå benytte mange flere variable enn analysene i de tidligere kapitler. Imidlertid benytter vi også nå kun 10 000 observasjoner fra kontrollgruppen, men dette vil ikke ha mye å si for hovedkonklusjonene.

Basert på vurderingene i de tidligere kapitler rundt hvilke type modeller som ser ut til å fungere best for klassifisering, samt ønske om å teste en annen type modell (nevral nettverk), har vi endt opp med å bygge tre modeller; en C&RT (regresjonstre), en nevral nettverksmodell og en boosted tree modell. I Statistica finnes det også et verktøy 'Data Miner Recipes', som vi har valgt å benytte for disse analysene.

9.1 'Data Mining Recipe'

I dette avsnittet presenterer vi resultater av analyser som er fremkommet ved hjelp av Statistica's verktøy 'Data Miner Recipes'. Dette er et verktøy som skal kunne hjelpe til i alle steg i et datamining-prosjekt, fra klargjøring av data til modellbygging og evaluering til klassifisering av nye observasjoner.

Vi har brukt dette verktøyet på 10 000 observasjoner fra kontrollgruppen og alle observasjoner fra frivillig retting gruppen. Vi har nå benyttet de fleste forklarende variable, unntatt variablene for 2005-2007, se listene i appendiks for en oversikt. Vi har imidlertid inkludert endringsvariablene for lønn og formue for alle år (dvs. endring i lønn og formue fra 2006-2005 og opp til 2008-2007). Disse har vi selv beregnet ved å ta differanser, for eksempel er variabelen FORMUE_06_05 beregnet ved å trekke NETTO_FORM_STAT05 fra NETTO_FORM_STAT06. Tilsvarende er de andre endringsvariablene beregnet.

Som tidligere er målet å bygge en modell som kan klassifisere observasjonene våre i to grupper, kontrollgruppe eller frivillig retting gruppe, på bakgrunn av de ulike forklaringsvariable (kjennetegn) som er tilgjengelig. Vi vil evaluere hvor gode modellene er til å klassifisere de mottatte observasjoner ved in-sample prediksjon, dvs. hvor godt modellen klassifiserer dataene som er brukt for å tilpasse modellen. På bakgrunn av disse resultatene vil vi vurdere om slike modeller kan benyttes til å forbedre kontrollobjektutvelgelsen av personlige skatteytere.

Ved hjelp av 'Data Miner Recipes' har vi bygget tre ulike modeller; en C&RT, en nevral nettverksmodell og en boosted trees modell. To av disse modellene, C&RT og boosted trees, er valgt på bakgrunn av vurderingene i de tidligere kapitler, og den siste modellen, nevralt nettverk, er inkludert for sammenligningens skyld. For de metodiske detaljene rundt hver av disse modellene viser vi til kapittel 5.

Resultatene fra prediksjonene in-sample for de tilpassede modeller er vist i tabell 36 under. Vi ser at alle disse modellene klarer å klassifisere observasjonene våre til gruppene (frivillig retting og kontroll) med stor grad av nøyaktighet, for eksempel har boosted trees kun en feilprosent på i underkant av 2 %. Denne forbedringen av klassfisering in-sample i forhold til de tidligere modellene vi har undersøkt

skyldes selvsagt at vi nå har benyttet mange flere forklarende variable i modelltilpassingen. Vi gjentar igjen at en out-of-sample vurdering med et nytt testdatasett vil vanligvis gi høyere feilprosent. På bakgrunn av disse lave feilprosentene synes det allerede nå at bruk av slike analyser kan gjøre det mulig å foreta en god kontrollobjektutvelgelse.

Tabell 37: In-sample prediksjonsgrad fra de tre modeller

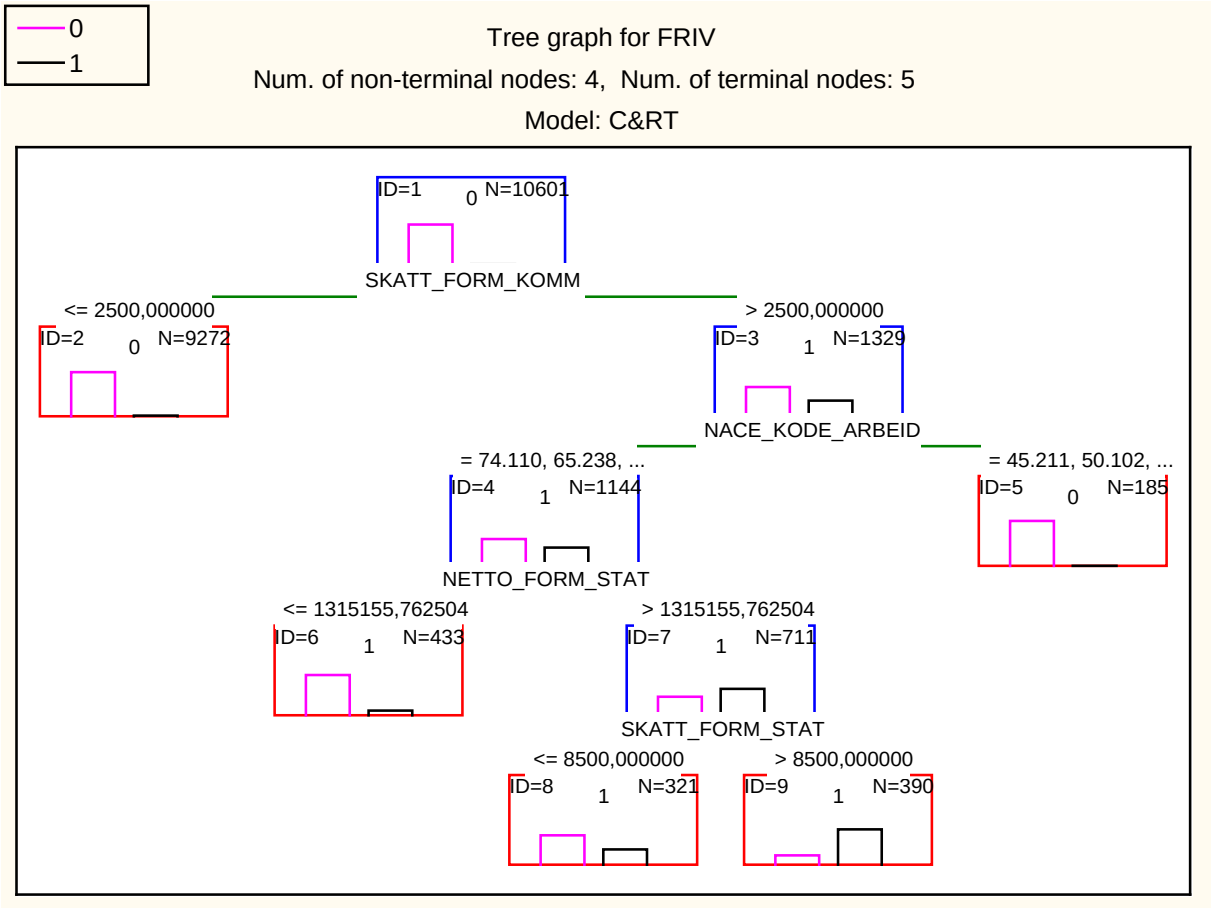
	1	2	3
Model selected for deployment	3		
Model Evaluation Summary	ID	Name	Error rate (%)
	3	Boosted trees	1,91
	4	Neural network	2,94
	1	C&RT	7,73
Table	Step options		
	Date and time	03.10.2011 15:32:39	

Videre vises mer detaljerte resultater for hver enkelt modell. Se tabell 38 for in-sample prediksjonsevne for klassfikasjonstreet. Her ser vi at modellen predikerer 9319 skatteytere av 10 000 fra kontrollgruppen korrekt, og 463 skatteytere av 601 korrekt (egentlig 575, siden noen av de 601 ikke inneholder noe data for alle kjennetegn) for frivillig retting gruppen. Totalt sett gir dette en feilprosent på 7,73 (som vi også ser fra tabell 37).

I figur 13 vises splittene i klassifikasjonstreet som er tilpasset. Vi ser at første splitt skjer på variabelen SKATT_FORM_KOMM. Fra listen over de viktigste forklaringsvariable (kjennetegn) i tabell 39, ser vi at det er variable vi tidligere har funnet som gode prediktorer, som synes viktigst. Dvs. at formue og endring i formue fra år til år, alder, næringen skatteyter arbeider i er blant de viktigste. Denne listen over viktige kjennetegn harmonerer også godt med listene i Appendiks (se slutten av kapittel 4.1. for mer informasjon om listene i Appendiks).

Tabell 38: In-sample prediksjonsevne for klassifikasjonstreet

	Summary Frequency Table (Summary_of_Deployment_(Datasett1END)) Table: FRIV(2) x Model-1-Prediction(2)			
	FRIV	Model-1-Prediction 0	Model-1-Prediction 1	Row Totals
Count	0	9319	681	10000
Column Percent		98,54%	59,53%	
Row Percent		93,19%	6,81%	
Total Percent		87,91%	6,42%	94,33%
Count	1	138	463	601
Column Percent		1,46%	40,47%	
Row Percent		22,96%	77,04%	
Total Percent		1,30%	4,37%	5,67%
Count	All Grps	9457	1144	10601
Total Percent		89,21%	10,79%	



Figur 13: Klassifikasjonstreet

Tabell 39: Liste over variabel viktighet i C&RT (klassfikasjonstreet) (tabellen fortsetter på neste side)

	Predictor importance (Datasett1END) Response: FRIV Model: C&RT	
	Variable Rank	Importance
SKATT_FORM_KOMM	100	0,995625
SKATT_FORM_STAT	100	1,000000
NETTO_FORM_STAT	99	0,985491
IB_NETTO_FORMUE	96	0,960451
SUM_SKATT_AVGIFT	45	0,452584
ANTALL_UB	44	0,438997
FORMUE_06_05	43	0,431701
NACE_KODE_ARBEID	41	0,411044
NAERING_HOVEDGR_BESK	37	0,367784
FODSELAAR	35	0,351466
FORMUE_07_06	35	0,345959
NACE_KODE_NAERING	35	0,350433
IB_ALM_INNT_F_SERF	34	0,342197
IB_ALM_INNT_E_SERF	33	0,329946
SKJERMINGSFR_AKSJE	31	0,306779
HAR_LEVERT_SA	31	0,308440
FORMUE_08_07	29	0,291853
HOVEDGR_BESK_ARBEID	26	0,261487
RESTSKATT_OVERSKYTENDE	25	0,246132
HISTORIKK_KODE	25	0,245020
NAERING_HOVEDOMR_BESK	23	0,229633
FORDELINGSFRADRAG_IB_OG_UB	20	0,201017
NEG_ALM_INNT_KUN_IB	20	0,201017
PENSJON_INNTEKT	20	0,197316
T_GRL_PENSJON	20	0,197316
T_AVG_PENSJON	19	0,192564
TOPPSKATT	19	0,194540
TOPPSKATTEGRUNNLAG	17	0,170572
SUM_LISTEPRIS_BILER	17	0,172040
PENSJON_I_U_ARB_211	16	0,155125
ALDERSPENSJON_217	16	0,163040
GJELD_IB_OG_UB	14	0,137874
GJELD_KUN_IB	13	0,130816
FOEDE_LAND	13	0,133441
AVSRENTER_DEBET	11	0,108373
PSA_RETTE_BESK	11	0,114767
STATSBORGERSKAP	9	0,087991
PERSON_INNTEKT_LONN	8	0,077217
T_GRL_LONN	8	0,077217
LONN_111A	8	0,077631

	Predictor importance (Datasett1END) Response: FRIV Model: C&RT	
	Variable Rank	Importance
HOVEDOMR BESK ARBEID	8	0,076443
SUM IB PERSONINNT ALLE TYPER	7	0,069145
AVSRENTER KREDIT	7	0,073430
T GRL NERING	7	0,071576
T AVG LONN	7	0,073759
LONN 08 07	7	0,068651
KORR ALM INNTEKT	6	0,063035
FREMFOR UTENL INNT	6	0,058358
T AVG NERING	6	0,061505
PENSJON I ARB FORH 212A	6	0,056757
ANT BILER	5	0,051339
ANTALL ARB GIV	5	0,048170
LONN 07 06	5	0,051628
LONN 06 05	5	0,052991
SAMSKATT KODE	5	0,051816
IB GODTGJ JSF	4	0,041776
IB GODTGJ NER	4	0,042291
TILLEGGSSKATT	4	0,041061
T GRL JSF	4	0,039545
T GRL FAMB	4	0,041071
T AVG JSF	4	0,039545
T AVG FAMB	4	0,041071
TOPPSKATTEGR KORR	3	0,032857
ANV TOLVDEL	3	0,033825
UTBET SELVS ND 401	3	0,034581
SYKEPENGER SELVS ND 405	3	0,025295
KJONN	3	0,031240
ARBEIDSTILLATELSE	3	0,026708
LONN KONK 123	2	0,022190
FODSEL OMSORG 138	2	0,019134
DAGPENGER 147	2	0,017159
ANDRE TRYGDEYTELSER	2	0,015961
AFP 227	2	0,019437
ANTALL MSYKKEL	1	0,005266
SUM LISTEPRIS MSYKKEL	0	0,004794
SA FRITAK KODE	0	0,004217

I tabell 40 er in-sample klassifikasjonsevne for boosted trees gitt. Her ser vi at kun 51 av 10 000 observasjoner i kontrollgruppen blir feilklassifisert, mens det tilsvarende tallet for frivillig retting gruppen er 152. Med andre ord har faktisk klassifikasjonstreet færre feilklassfiseringer for frivillig retting gruppen, men totalt sett er boosted trees best med en in-sample feilprosent på kun 1,91 % (se også tabell 37).

Tabell 40: In-sample prediksjonsevne for boosted tree

	Summary Frequency Table (Prediction) Table: FRIV(2) x Model-3-Prediction(2)			
	FRIV	Model-3-Prediction 0	Model-3-Prediction 1	Row Totals
Count	0	9949	51	10000
Column Percent		98,50%	10,20%	
Row Percent		99,49%	0,51%	
Total Percent		93,85%	0,48%	94,33%
Count	1	152	449	601
Column Percent		1,50%	89,80%	
Row Percent		25,29%	74,71%	
Total Percent		1,43%	4,24%	5,67%
Count	All Grps	10101	500	10601
Total Percent		95,28%	4,72%	

Tilsvarende er resultatene for det nevrale nettverket gitt i tabell 41 under. Her feilklassifiseres hele 220 av frivillig retting gruppen, men kun 92 av skatteyterne i kontrollgruppen feilklassifiseres. Totalt sett konkluderer vi med at boosted trees kan være en godt egnet metode for klassifisering av skatteytere, og at denne metoden også kan benyttes for kontrollobjektutvelgelse.

Tabell 41: In-sample prediksjonsevne for nevralt nettverk

	Summary Frequency Table (Prediction) Table: FRIV(2) x Model-4-Prediction(2)			
	FRIV	Model-4-Prediction 0	Model-4-Prediction 1	Row Totals
Count	0	9908	92	10000
Column Percent		97,83%	19,45%	
Row Percent		99,08%	0,92%	
Total Percent		93,46%	0,87%	94,33%
Count	1	220	381	601
Column Percent		2,17%	80,55%	
Row Percent		36,61%	63,39%	
Total Percent		2,08%	3,59%	5,67%
Count	All Grps	10128	473	10601
Total Percent		95,54%	4,46%	

Etter disse modelltilpasningene (i 'Data Miner Recipe' kan andre typer modeller også selvsagt tilpasses), kan vi nå anvende den beste modellen på nye data (såkalt 'deployment'). Dette vil være av interesse for å klassifisere nye skatteytere i en ærlig gruppe og en (mulig) unndragergruppe. Vi har benyttet alle våre mottatte kontrolldata på 300 000 skatteytere for en anvendelse av boosted trees modellen tilpasset ovenfor. De aller fleste av disse blir klassifisert som ærlige skattytere, mens noen tusen klassifiseres i den andre gruppen. Dvs. modellen er nå benyttet til å fremskaffe en prioritert liste av skatteytere for kontroller, vi har med andre ord benyttet modellen til kontrollobjektutvelgelse.

Fra disse analysene ved hjelp av 'Data Miner Recipes' konkluderer vi at modellene klarer å klassifisere observasjonene in-sample med stor grad av nøyaktighet. Ved å anvende disse modellene på nye skatteytere, vil modellene kunne gi en liste av skattytere som bør prioriteres i forhold til kontroller. På bakgrunn av resultatene våre, tror vi at slike lister fremkommet fra klassifisering vha. modeller, kan avdekke flere tilfeller av skatteunndragelse enn hva kontroller av ett tilfeldig utvalg av skatteytere vil avdekke.

Rent praktisk må slik bruk av modeller for kontrollobjektutvelgelse altså skje i to steg; først må det tilpasses en modell til et utvalg av skatteytere som vi vet klassetilhørigheten til, en ærlig gruppe og en unndrager-gruppe (slik vi har gjort i dette kapittelet). Deretter kan denne modellen benyttes på et nytt utvalg av skatteytere som vi ikke vet er ærlig eller unndrar (såkalt 'deployment' i 'Data Mining Recipes' i Statistica). Modellen vil så kunne klassifisere dette nye utvalget av skatteytere, og de skatteytere som blir klassifisert som unndragere bør prioriteres for kontroll. Imidlertid kan det selvsagt være ærlige skatteytere som blir klassifisert som unndragere, man bør derfor være forsiktig med å utpeke enkeltindivider som skatteunndrager før en nærmere kontroll er utført.

10. Konklusjoner og forslag til videre arbeid

Norske skatteytere skjuler store inntekter og formuer i utlandet, og Skattedirektoratet ønsker derfor å undersøke hvilke kjennetegn som karakteriserer slike skatteytere. Dette for å sikre en mer målrettet bruk av virkemidler internt, dvs. forbedre kontrollobjektutvelgelse og øke oppdagelsessannsynligheten for unndragelse.

I denne rapporten har vi gitt en oversikt av de ulike statistiske metoder som eksisterer for slike kjennetegnsanalyser, samt vurdert styrker og svakheter ved de ulike metodikkene. I alle slike analyser er målet å klassifisere et individ i en av to mulige grupper, ærlig skatteyter eller skatteunndrager. En klassifisering av skatteytere kan selvsagt også gjøres til flere grupper, men det har ikke vært tema for denne rapporten.

Basert på data fra Skattedirektoratet, der vi har mottatt et datasett bestående av et tilfeldig utvalg av 300 000 personlige skatteytere (kontrollgruppen) og et datasett av personlige skatteytere (601 individer) som har benyttet seg av skatteamnestiordningen og innrømmet skatteunndragelse (frivillig retting gruppen), har vi illustrert hvordan de ulike metodikkene kan benyttes. Siden det er så få skatteytere i frivillig retting gruppen relativt til kontrollgruppen kan en passende beskrivelse av problemet være 'å lete etter nålen i høystakken'.

I rapporten peker vi på noen av utfordringene ved slike kjennetegnsanalyser. Blant annet er variabelseleksjon, dvs. finne hvilke variable (kjennetegn) som skal inkluderes i en modell, en utfordring. Problemet oppstår her fordi man har svært mange variable tilgjengelig (i denne analysen i overkant av 570), og flere av disse kan samvariere sterkt. Dette gir opphav til multikolinearitet, som kan påvirke en analyse negativt. Videre er neste utfordring modellseleksjon; siden det finnes en rekke ulike statistiske modeller for kjennetegnsanalyser, er valget av hvilken modell som skal benyttes for analysen sentral. Et statistisk vurderingskriterie er å måle hvor godt modellen klassifiserer treningsdataene, dvs. de data som er benyttet for å tilpasse modellen, såkalt in-sample klassifisering. Imidlertid kan det oppstå et problem med såkalt overtilpasning ("overfitting"), dvs. at modellen tilpasser seg den tilfeldige variasjonen i datamaterialet istedenfor den underliggende sammenhengen vi er på jakt etter. Derfor er ofte et bedre vurderingskriterie out-of-sample klassifisering. Her brukes et testdatasett der vi kjenner gruppetilhørigheten til individene, men der datasettet ikke er brukt til modelltilpasningen. Et eksempel på dette er gitt i kapittel 6. Andre vurderingskriterier ift. valg av modell, kan være ønske om en lett forståelig modell der kjennetegnene som er plukket ut, enkelt kan identifiseres. Noen av de undersøkte metodene er nemlig i større grad "sorte bokser", der brukeren ikke med enkelthet kan identifisere de viktigste kjennetegn. Med store datasett er det ofte også for enkeltindivider data som mangler. Dette er det såkalt "missing value" problem, som vi beskriver nærmere i bl.a. kapittel 8. Vi gjør oppmerksom på at i datasettene vi har mottatt har det vært betydelig innslag av manglende data for flere av variablene for mange individer.

Det er videre to hovedaspekter ved rapporten; det første aspektet er å finne kjennetegn som ser ut til å være bærere av informasjon om de individer som har en tendens til skatteunndragelse, og som kan skille dem fra de øvrige. Det andre aspektet er å etablere en modell basert på et sett av forklaringsvariable (kjennetegn) for å predikere / sannsynliggjøre at et individ er en skatteunndrager.

Basert på tidligere litteratur har vi først undersøkt våre to grupper ved hjelp av enkle frekvenstabeller og beskrivende statistikk (kontrollgruppen og frivillig retting gruppen), og avdekket at flere av de klassiske kjennetegn ved en skatteunndrager også kan benyttes som kjennetegn for de skatteunndragerne vi har å gjøre med i dette prosjektet. De klassiske kjennetegnene er bl.a. kjønn, alder, formue og inntekt. En kort oppsummering av våre to grupper indikerer at menn unndrar mer enn kvinner, eldre unndrar mer enn unge, en person med meget høy formue unndrar mer enn en med mindre formue, en person med meget høy eller ingen inntekt unndrar mer enn en med inntekt nær gjennomsnittet, samt at en som lever i sentrale strøk unndrar mer enn en som lever i mindre sentrale strøk (se kapittel 4). Imidlertid blir det for enkelt å bare benytte seg av disse kjennetegnene for å gjøre en fornuftig klassifisering / predikering av en skatteyter. Vi har derfor benyttet flere av de presenterte metodene for å gjøre en best mulig klassifisering. Det viser seg at mange av metodene med relativt stor grad av treffsikkerhet klarer å klassifisere de mottatte data i riktig gruppe, men resultatene varierer mye mellom metodene. Riktignok viser resultatene totalt sett at det er visse kjennetegn som er forskjellig for de to gruppene, og disse kjennetegnene kan benyttes til en gruppering av skatteytere. Men som regel vil det være metoder som bruker kombinasjoner av mange kjennetegn som gir de beste resultatene, dvs. det er relativt komplekse modeller med mange kjennetegn (”sorte bokser”) som gir en mest riktig klassifisering. Basert på våre resultater vil vi fremheve metodene ”boosted” klassifikasjonstrær og ikke-parametrisk logistisk regresjon som de mest aktuelle metodene ved en kjennetegnsanalyse. Vi refererer videre til kapittel 9 for hovedanalysen.

Vi konkluderer med at basert på resultatene fremkommet i rapporten, kan bruk av denne typen analyser gjøre det mulig for Skattedirektoratet å foreta en bedre kontrollobjektutvelgelse. I praksis må dette skje i to steg; først må det tilpasses en modell til et utvalg av skatteytere som vi vet klassetilhørigheten til, en ærlig gruppe og en unndrager-gruppe. Deretter kan denne modellen benyttes på et nytt utvalg av skatteytere som vi ikke vet er ærlig eller unndrar. Modellen vil så kunne klassifisere dette nye utvalget av skatteytere, og de skatteytere som blir klassifisert som unndragere bør prioriteres for kontroll. Det må imidlertid tas hensyn til at det selvsagt kan være ærlige skatteytere som blir klassifisert som unndragere.

Vi avslutter med et mer spesifikt problem for analysene utført i denne rapporten, men som er helt essensielt i forhold til generalisering av resultatene fra rapporten. Problemet er at vi har med et skjevt utvalg å gjøre i frivillig retting gruppen, dvs. siden disse skatteytere har meldt seg frivillig gjennom skatteamnestiordningen, er de ikke nødvendigvis representative for den generelle gruppen av skatteunndragere. Vi kan derfor ikke trekke sikre slutninger om kjennetegn til de skatteyterne som unndrar, men som ikke har meldt seg, basert på de resultatene vi har funnet om de som har meldt seg frivillig. Det vil derfor være interessant å få analysert data om de skatteytere som Skatteetaten har avslørt gjennom ordinært kontrollarbeid i et fremtidig arbeid.

Vi mener også at kjennetegnsanalyser kan være aktuelle for å undersøke hvilke selskaper som driver med skatteunndragelse, noe som målt i kroner muligens er et større problem. Imidlertid kan det være enda større utfordringer å gjennomføre kjennetegnsanalyser i en slik setting, bl.a. fordi selskaper kan ha større muligheter for skatteunndragelse, for eksempel fordi det er større utfordringer med skatteregler på tvers av land, internprising mv.

Referanser

Allingham, M. & Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics,* 323-338.

Becker, G. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76 (2), 169-217.

Chipman, H.A., George, E.I. and McCulloch, R.E. (1998). Extracting Representative Tree Models From a Forest. *Working Paper 98-07, Department of Statistics and Actuarial Science, University of Waterloo*.

Collins, J.H., Milliron, V.C. & Toy, D.R. (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.

Engström, P. & Holmlund, B. (2009). Tax Evasion and Self-Employment in a High-Tax Country: Evidence from Sweden. *Applied Economics* 41, 2419-2430.

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

Feld, L.P. & Frey, B.S. (2002). Trust Breeds Trust: How Taxpayers are Treated. *Economics of Governance* 3, 87-99.

Kleven, H.J., Knudsen, M.B., Kreiner, C.T, Pedersen, S. & Saez, E. (2011). Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark. *Econometrica* 79 (3), 651-692.

Lee, J-S. & Carley, K.M. (2009). Predicting International Tax Error Using Open Source Literature and Data. CASOS Technical Report, Carnegie Mellon University.

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

Skatteunndragelsesutvalget (NOU 2009:4). Tiltak mot skatteunndragelser.

Torgler, B. (2005). Tax Morale and Direct Democracy. *European Journal of Political Economy* 21, 525-531.

I denne rapporten presenteres resultatet av et prosjekt utført for Skattedirektoratet (SKD). Hensikten har vært å finne kjennetegn ved personlige skattytere som har unndratt skatt gjennom å skjule formuer og/eller inntekt i utlandet/skatteparadiser. En viktig del av dette arbeidet har bestått i å evaluere tilgjengelige analysemetoder.

I rapporten presenteres flere analysemetoder relativt grundig, og vi gjennomgår styrker og svakheter ved de ulike metodikkene. Varianter av logistisk regresjon og klassifikasjonstrær (C&RT) utmerker seg som de mest lovende metodene.

Som regel er det relativt komplekse modeller med mange kjennetegn som gir en mest riktig klassifisering. Det er imidlertid noen variabler som ser ut til å være uunnværlige for å gjøre en god klassifikasjon. Eksempler på disse er inntekt, formue, kjønn, alder og hvorvidt en skatteyter bor i sentrale strøk.

SNF
Et selskap i NHH-miljøet

SAMFUNNS- OG NÆRINGSLIVSFORSKNING AS

Institute for Research in Economics and Business Administration

Breiviksveien 40
N-5045 Bergen
Norway
Phone: (+47) 55 95 95 00
Fax: (+47) 55 95 94 39
E-mail: publikasjon@snf.no
Internet: http://www.snf.no/

Trykk: Allkopi Bergen