Utvikling og implementering av nytt system for verdsettelse av fritidsboliger

Prosjektrapport
17. november 2020

Forord

Stortinget har bedt regjeringen om å gjennomgå og forbedre verdsettelsen av fritidsboliger i Norge. I mai 2019 inviterte Finansdepartementet tjenesteleverandører til å delta i utvikling av en ny verdsettelsesmodell for fritidsboliger. Formålet med modellen er å gi en mer treffsikker verdifastsettelse av fritidsboliger ved beregning av formuesskatt. Oppdraget med å utvikle modellen ble tildelt PwC 25. september 2019.

Denne rapporten gjennomgår arbeidet med utvikling av verdsettelsesmodellen og tilhørende resultater. Rapportens struktur følger prosjektets aktiviteter, herunder gjennomgang av tidligere arbeid, vurderinger av datakilder, variabler og modelleringsvalg. Verdsettelsesmodellen overføres til Statistisk Sentralbyrå (SSB) ved prosjektets avslutning.

Thomas Hansen og Daniel Kinn har ledet prosjektet og Nora Hansen, Hermann Andersen, og William Eilertsen har vært prosjektmedarbeidere. Henrik Gran har vært ansvarlig partner og kvalitetssikrer.

Oslo, 17.11.2020

Henrik Gran,
Partner

T: +47 952 60 046

Sammendrag

Dagens formuesverdier for fritdsboliger reflekterer i liten grad boligenes markedsverdi

Fritidsbolig inngår i grunnlaget for formuesskatt og beregnes med utgangspunkt i boligens markedsverdi. Formuesverdien fastsettes i forbindelse med nybygg og endres sjelden utover sporadiske og generelle prosentvise justeringer. Verdiene har derfor ikke holdt tritt med markedsutviklingen for fritidsboliger. Omsetningstall viser at det er et gjennomsnittlig avvik på 48 prosent[1] mellom grunnlaget for formuesskatt og faktisk salgspris. Verdsettelsen av fritidsboliger kan dermed oppfattes som lite presis og gi vilkårlig verdsettelse av like boliger.

Prosjektet har utviklet en modell som estimerer markedsverdien av norske fritidsboliger

Prosjektets mandat har vært å utvikle en verdsettelsesmodell for norske fritidsboliger som gir økt treffsikkerhet og likebehandling i formuesbeskatning. Den utviklede verdsettelsesmodellen gir en estimert markedsverdi basert på boligenes underliggende egenskaper som areal, alder og beliggenhet. Siden modellen skal benyttes for skatteformål er det lagt vekt på at verdsettelsene skal kunne forklares til den enkelte skatteyter. Videre har det blitt fokusert på å utvikle en modell som gir en stabil utvikling i markedsverdier, slik at modellen kan benyttes for årlige oppdateringer av formuesverdier. Den utviklede modellen verdsetter alle landets 400 tusen boliger på en presis, stabil og forklarbar måte basert på dagens tilgjengelige data.

Fritidsboliger har særegne karakteristikker som innebærer at de er lite egnet for tradisjonelle modeller

Verdsettelsesmodeller for bolig er i dag implementert i minst 15 land. Landene har typisk et mindre omfang av fritidsboliger, og de fleste modellene er derfor utviklet for primærboliger. I Norge har SSB utviklet en verdsettelsesmodell for formuesbeskatning av primær- og sekundærboliger. Fritidsboliger har imidlertid noen særegne karakteristikker som medfører at dagens system for primær- og sekundærboliger er lite egnet for å estimere boligenes markedsverdi. For det første har fritidsboliger en større variasjon i bygningsstandard, adkomst og fasiliteter enn primærboliger. Videre er boligenes bruksverdi tett knyttet til det umiddelbare nærområdet. Eksempelvis kan egenskaper som strandlinje og nærhet til skibakker gi store utslag i markedsverdi. I tillegg eksisterer det ikke et system for innrapportering av areal og alder av fritidsboliger til Skatteetaten, slik det gjør for primær- og sekundærboliger. Konsekvensen er at en slik verdsettelsesmodell må baseres på registrerte bygningsopplysninger fra det norske eiendomsregisteret som er av varierende kvalitet.

Verdsettelsesmodellen kombinerer maskinlæring og geodata for å gi treffsikre verdsettelser

Den utviklede verdsettelsesmodellen er en maskinlæringsmodell. Denne typen modeller egner seg godt for å fange opp de komplekse sammenhengene som påvirker markedsverdien av norske fritidsboliger. Begrepet maskinlæring innebærer at modellene lærer slike sammenhenger gjennom å trenes på et utvalg av tidligere solgte boliger. Boligegenskapene som benyttes for å trene modellen baserer seg på dagens registrerte eiendomsopplysninger i matrikkelen. I tillegg er det beregnet flere variabler knyttet til fritidsboligenes nærområde. Variablene inkluderer blant annet strandlinje, solforhold, utsikt, antall naboer, nærhet til skibakke og nærhet til vei. Variablene er basert på en kombinasjon av tidligere studier innenfor verdsettelse, meglerundersøkelser, og fremhevede egenskaper i annonser for fritidsboliger.

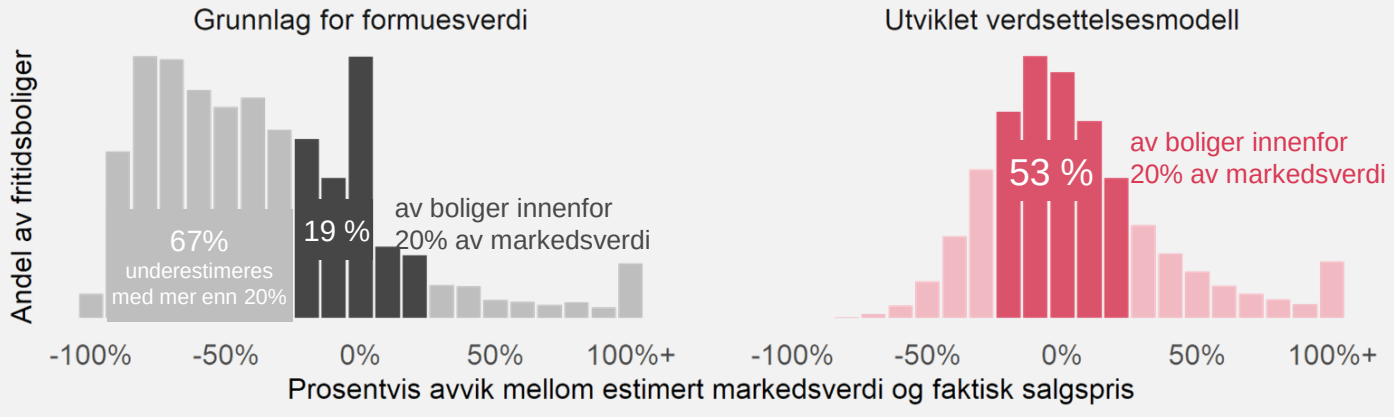
Tilnærmingen skiller seg fra tidligere arbeid med boligverdsettelser. For å belyse forskjellene mellom maskinlæring og tidligere arbeid er det også utviklet en tradisjonell verdsettelsesmodell. Modellen er en utvidelse av dagens modell for primær- og sekundærbolig og benyttes som sammenligningsgrunnlag i prosjektet.

1) Basert på grunnlaget for formuesverdiene for omsatte fritidsboliger på FINN.no i 2018.

Sammendrag

Verdsettelsesmodellen bidrar til økt treffsikkerhet og likebehandling i formuesbeskatning

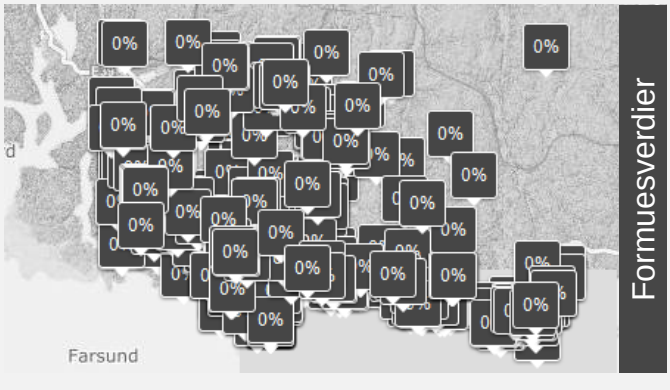
Den utviklede verdsettelsesmodellen oppnår et gjennomsnittlig absoluttavvik fra markedsverdi på 26,6 prosent. Dette er en forbedring fra dagens grunnlag for formuesverdier[2] på 22 prosentpoeng. Resultatene er basert på et utvalg av 38 tusen boliger der modellens estimerte markedsverdi er sammenlignet med faktisk salgspris. Antall boliger som underestimeres av modellen er like stort som omfanget av boliger som overestimeres. Dette skiller seg fra dagens grunnlag for formuesverdier hvor om lag 67 prosent av boligene underestimeres med mer enn 20 prosent. Modellen gir dermed betydelig større treffsikkerhet i fastsettelse av formuesverdier, og bidrar også til en mer ensartet verdsettelse av fritidsboliger.



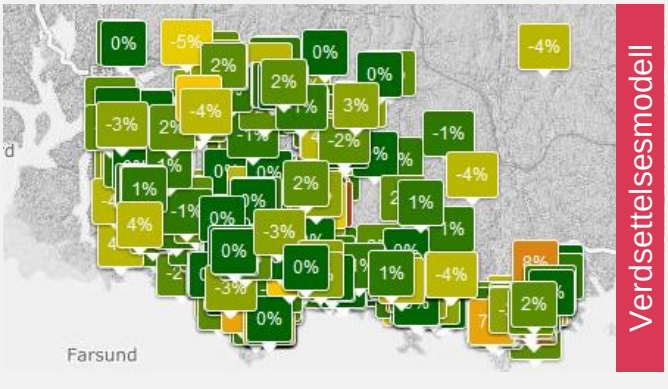
Verdsettelsesmodellen hensyntar markedsutvikling på en stabil måte uten store lokale variasjoner

For å belyse hvordan modellen vil oppleves for skattytere er det gjennomført simuleringer av skatteoppgjør for årene 2018 til 2020. Resultatene viser at modellen gir en stabil utvikling i boligenes markedsverdier uten store årlige variasjoner eller lokale forskjeller. Om lag 95 prosent av boligene har en årlig markedsutvikling som avviker mindre enn fem prosentpoeng fra kommunens gjennomsnittlige markedsutvikling.

Dagens formuesverdier ble sist oppjustert i 2014



Årlig endring i markedsverdi fra 2019 til 2020

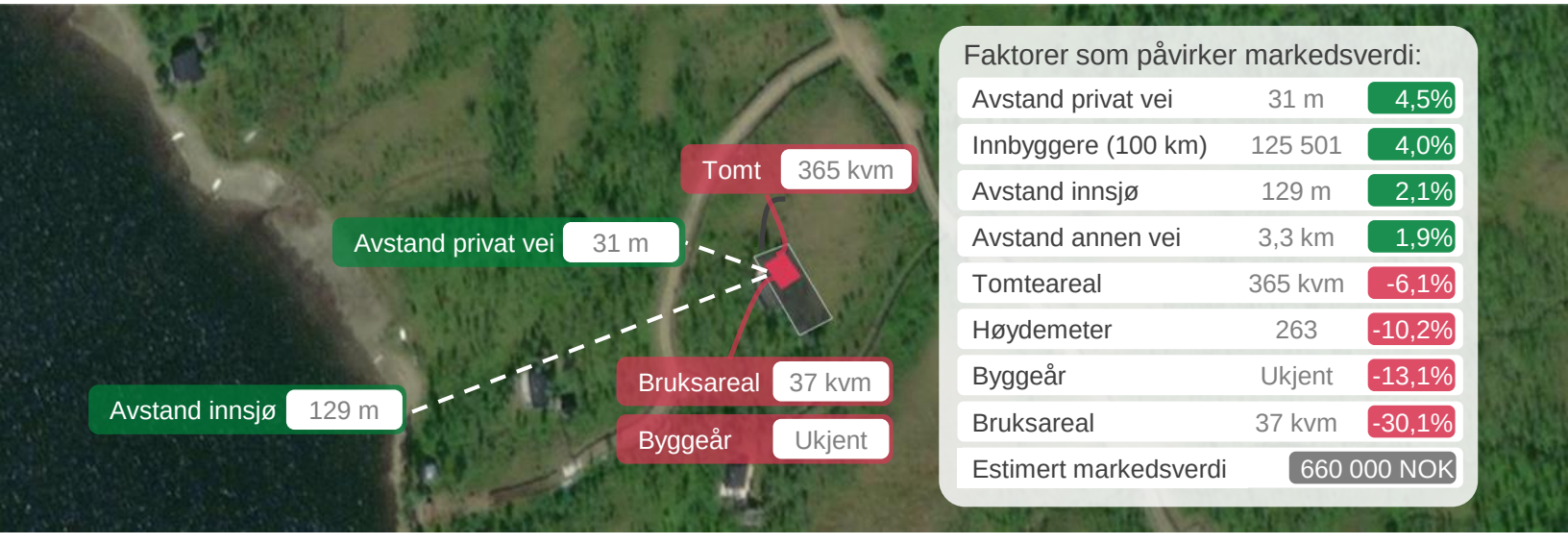


2) Formuesskatt beregnes for de fleste formål med utgangspunkt i eiendelenes markedsverdi. Fritidsboliger verdsettes ikke til full markedsverdi, men en noe lavere verdi kalt formuesverdi. Formuesverdien for nye bygg settes til maks 30 prosent av fritidsboligens markedsverdi eller 30 prosent av fritidsboligens kostpris inkludert grunn. Skatteetaten registrerer i dag ikke fritidsboligens estimerte markedsverdi, og formuesverdien dividert med 30 prosent er dermed beste estimat av markedsverdi i Skatteetatens grunnlag.

Sammendrag

Hver enkelt verdsettelse fra modellen kan forklares med bakgrunn i boligens egenskaper

Modellens verdsettelser kan dekomponeres i bidraget som hver enkelt egenskap har gitt til boligen. Dette innebærer at det er mulig å beregne og begrunne hvordan de ulike egenskapene har bidratt til å øke eller redusere verdsettelsen sammenlignet med en gjennomsnittlig fritidsbolig. I gjennomsnitt utgjør strukturelle egenskaper som bruksareal og byggeår over 30 prosent av differansen, mens antall nærliggende fritidsboliger og boligens nærhet til sjø utgjør om lag 19 prosent. Antall nærliggende fritidsboliger skiller mellom boliger med tett og avsidesliggende bebyggelse.



Figuren viser estimert markedsverdi for en fritidsbolig og de viktigste egenskapene som påvirker verdsettelsen.

Bruk av geodata og maskinlæring gir økt treffsikkerhet i estimering av boligformue

Den utviklede modellen gir åtte prosentpoeng lavere avvik fra markedsverdi enn den tradisjonelle lineære modellen som benyttes som sammenligningsgrunnlag. Begge modeller gir forklarbare verdsettelser med sammenlignbar stabilitet. Resultatene viser et stort potensiale for bruk av geodata og maskinlæring i verdsettelse av boligformue, og kan potensielt være overførbart til verdsettelse av primær- og sekundærbolig.

Verdsettelsesmodellen vil forbedre seg ytterligere over tid

Modellens stabilitet og treffsikkerhet påvirkes av kvaliteten på registrerte bygningsopplysninger i matrikkelen. Over halvparten av boligene mangler opplysninger om byggeår og 18 prosent mangler registrert bruksareal. Det pågående arbeidet med kvalitetsheving av bygningsopplysninger vil øke kvaliteten på modellens verdsettelser. I tillegg vil datagrunnlaget for modellen økes etter hvert som flere fritidsboliger omsettes. Beregninger prosjektet har gjennomført viser at økt datakvalitet i matrikkelen kan bidra til en forbedring i presisjon på 5,5 prosentpoeng. Hvert ekstra år med omsatte boliger gir 0,2 til 0,4 prosentpoeng forbedring.

Veien videre

Verdsettelsesmodellen er overlevert til SSB ved prosjektets avslutning med tilhørende teknisk dokumentasjon. Finansdepartementet eier modellen og står fritt til å benytte og videreutvikle denne for fremtidig verdsettelse av fritidsboliger.

Innhold

1 Bakgrunn og formål.....7

2 Tilnærming.....11

3 Kriterier for modell.....19

4 Variabler og datakilder.....28

5 Innledende analyse.....36

6 Modelleringsvalg.....48

7 Resultater.....57

8 Begrensninger og potensielle forbedringer.....75

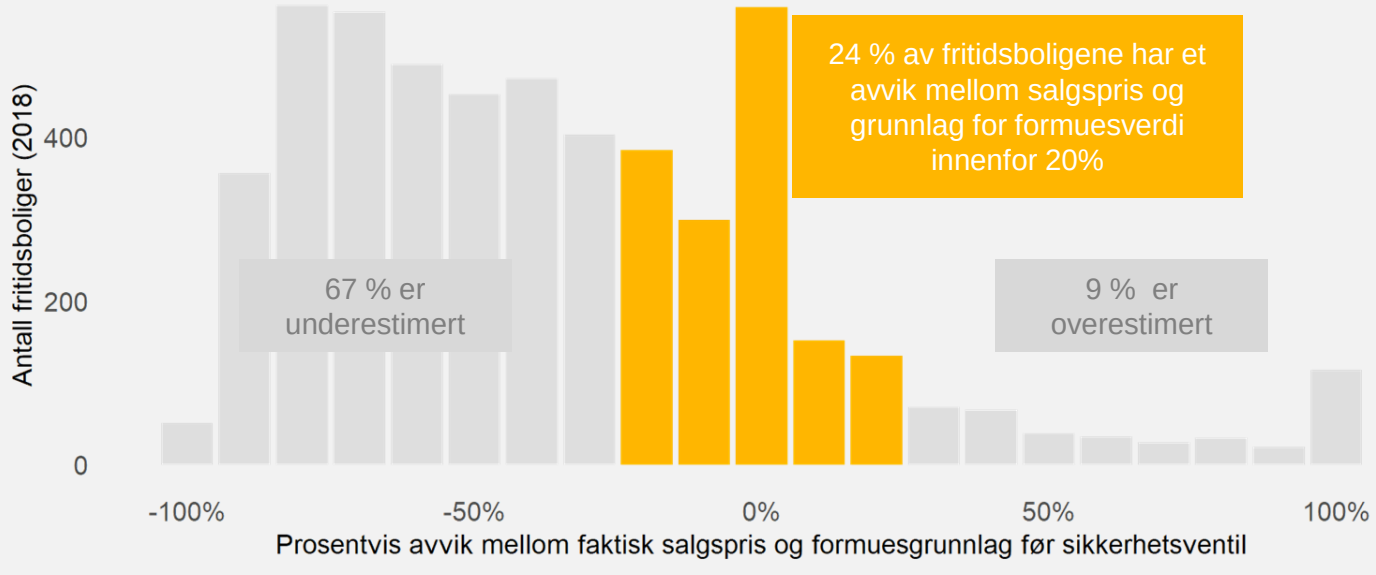
A Appendiks.....78

1. Bakgrunn og formål

Bakgrunn

I forbindelse med behandling av Meld. St. 4 (2015 - 2016) “Bedre Skatt” ba Stortinget regjeringen om å utrede hvordan et verdsettingssystem for fritidsboliger kan utformes. Formålet var å gjøre verdsettelsen av fritidsboliger mer ensartet og forbedre samsvaret mellom formuesverdier og faktiske verdier. Avviket mellom formuesverdier og faktiske verdier skyldes at formuesverdier for fritidsboliger hovedsakelig fastsettes i forbindelse med nybygg. Verdiene oppjusteres ikke i forbindelse med salg og endres kun gjennom sporadiske og generelle justeringer. Formuesverdiene har derfor ikke holdt tritt med prisutviklingen for fritidsboliger. Utviklingen av en mer presis verdsettelsesmodell vil bidra til økt treffsikkerhet og likebehandling ved formuesbeskatning av fritidsboliger.

Beregninger prosjektet har gjennomført viser at det var et gjennomsnittlig avvik på 48 prosent mellom grunnlaget for formuesverdier og salgsprisen for fritidsboliger som ble solgt i 2018. Resultatene viser at markedsverdien underestimeres for om lag 67 prosent av fritidsboligene med mer enn 20 prosent. Det er spesielt eldre fritidsboliger som verdsettes uforholdsmessig lavt. Om lag ni prosent av boligene overestimeres.



Figuren viser prosentvis avvik mellom grunnlag for formuesverdi per 01.01.2018 og faktisk salgspris for fritidsboliger omsatt på FINN.no i 2018. Det gjennomsnittlige absolutte avviket er på 48 prosent. (n = 4 551)

Det eksisterer i dag ingen modell for verdsettelse av fritidsboliger for formuesbeskatning. Det eksisterer imidlertid en modell for fastsettelse av formuesverdier for primær- og sekundærboliger (Takle og Medby, 2019). Modellen er basert på innrapporterte opplysninger om boligenes alder og areal til Skatteetaten, og tar høyde for boligenes beliggenhet gjennom ulike prissoner og tettstedsvariabler.

Fritidsboliger har imidlertid noen særegne karakteristikker som medfører at dagens system for primær- og sekundærboliger er lite egnet for å estimere boligenes markedsverdi. For det første har fritidsboligene en større variasjon i bygningsstandard, adkomst og fasiliteter enn primærboliger. Videre er boligenes bruksverdi gjerne tett knyttet til det umiddelbare nærområdet, og egenskaper som strandlinje og nærhet til skibakker kan gi store utslag i markedsverdi. I tillegg eksisterer det ikke et system for innrapportering av fritidsboligenes areal og alder til Skatteetaten, slik det gjør for primær- og sekundærboliger. Konsekvensen er at utviklingen av modellen må baseres på registrerte bygningsopplysninger fra det norske eiendomsregisteret som kan være av varierende kvalitet. Prosjektets formål er å utvikle en modell som forbedrer samsvaret mellom grunnlaget for formuesverdier og markedsverdier på en presis, stabil og forklarbar måte basert på dagens tilgjengelige data.

Prosjektets omfang

Formål

Prosjektets formål er å utvikle en verdsettelsesmodell for norske fritidsboliger som bidrar til økt treffsikkerhet og likebehandling i formuesbeskatning. Modellen skal estimere markedsverdier av alle norske fritidsboliger på en presis, stabil og forklarbar måte ved bruk av dagens tilgjengelige data.

Krav til verdsettelsesmodell

Presis

Modellen skal estimere markeds-verdien av norske fritidsboliger med størst mulig grad av treffsikkerhet. Dette innebærer å oppnå et lavest mulig prosentvis avvik mellom estimerte og faktiske markeds-verdier.

Stabil

Store årlige variasjoner i formues-verdier vil medføre uforutsigbarhet i privatøkonomien til den enkelte skattyter. Modellen må derfor gi stabile estimater på markedsverdi over tid og innenfor geografiske områder.

Forklarbar

Oppdraget stiller krav om at de estimerte markedsverdiene skal være "tilstrekkelig transparente" for den enkelte skattyter. Dette innebærer at bakgrunnen for hver enkelt verdsettelse må kunne forklares til skattyter.

Leveranser

Verdsettelsesmodell

Modellen med tilhørende dataflyt overleveres til SSB ved prosjektets slutt.

Prosjektrapport

Prosjektrapporten redegjør for arbeidet med utvikling av modell og effektene ved eventuell innføring.

Teknisk dokumentasjon

Den tekniske dokumentasjonen beskriver hvordan modellen kan benyttes i praksis.

Roller

Finansdepartementet

Finansdepartementet er oppdrags-giver for prosjektet.

PwC har ledet prosjektet og vært ansvarlig for utvikling av modellen.

Statistisk sentralbyrå
Statistics Norway

SSB har bidratt i diskusjoner og er ansvarlig for årlige oppdateringer.

Avgrensing av fritidsboliger

Det eksisterer ikke en enhetlig definisjon av begrepet fritidsbolig innenfor skatterett. Skatteetaten benytter begrepet "fritidseiendom" og beskriver at: "*Skillet mellom boligeiendom og fritidseiendom bygger på karakter og særtrekk ved eiendommen. Som fritidseiendom regnes typiske hytter, sommerhus og leilighets-kompleks bygget utelukkende for fritidsformål.*" SSB benytter begrepet "fritidsbygg" i bygningsstatistikk og legger til grunn den registrerte bygningstypen som benyttes i det norske eiendomsregisteret, matrikkelen. Dette omfattet i 2020 om lag 468 tusen "fritidsbygg". Prosjektet har tatt utgangspunkt i at modellen er utviklet for skatteformål og benytter derfor de 400 tusen boligene som Skatteetaten definerer som "fritidseiendommer".

Ulike typer fritidsboliger

Skatteetaten benytter de registrerte enhetene i matrikkelen som grunnlag for fastsettelse av formuesverdier. Matrikkelenhetene omtales for de fleste formål som eiendommer i dagligtale og kan for fritidsboliger deles i kategoriene grunneiendom, festegrunn og eierseksjoner.

Grunneiendom

En grunneiendom er en fast eiendom som er avgrenset av eiendomsgrenser. Den strekker seg vertikalt så langt oppover og nedover som den er praktisk nyttbar. Om lag **79%** av fritidsboligene er grunn-eiendommer.

Festegrunn

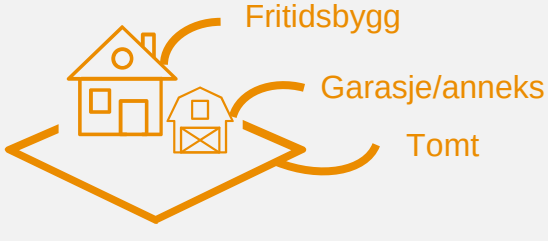
En festegrunn er en del av en grunneiendom som kan festes (leies) bort eller som noen har en tilsvarende eksklusiv og lang-varig bruksrett til. Det finnes ulike former for feste avhengig av om grunnen er oppmålt. Om lag **17%** av fritidsboligene er festegrunner.

Eierseksjon

Eierseksjon er en eierandel i en seksjonert eiendom, der eieren har enerett til å bruke en bestemt del av eiendommen. Den delen eieren har bruksrett til omtales gjerne som bruksenhet. Om lag **4%** av fritidsboligene er eier-seksjoner.

Hva kjennetegner en fritidsbolig?

De fleste fritidsboliger er grunneiendommer som innebærer at det er fast oppmålte eiendommer som hjemmelshavere kan pantsette, selge og overføre. Eiendommer som skattlegges som fritidseiendommer har typisk én eller flere bygninger eller bruksenheter som klassifiseres som fritidsbygninger eller boligseksjoner. Eiendommene kan i tillegg ha andre bygg som annekser, båthus eller garasjer som ikke regnes som fritidsboligareal, men som like fullt kan påvirke markedsverdien av boligen. Fritidsboliger kan ha betydelig variasjon, fra én enkel eierseksjon til en stor grunneiendom med flere bygninger.



Begrepet "fritidsbolig" benyttes om matrikkelenheter som Skatteetaten skattlegger som "fritidseiendommer". Fritidsbolig kan i denne sammenhengen omfatte både bygninger, grunneiendommer, festetomter og eierseksjoner.

2. Tilnærming

Verdsettelsesmodell

For å bidra til økt treffsikkerhet og likebehandling i formuesbeskatningen har prosjektet utviklet en verdsettelsesmodell for alle norske fritidsboliger. En slik modell omtales gjerne som en masseverdsettelses-modell. Ifølge Kolbe et al (2019) har minst 15 land implementert slike modeller for skatteformål. I Norge har SSB utviklet en tilsvarende modell for verdsettelse av norske primær- og sekundærboliger. Modellen benyttes for estimering av boligformue, men flere kommuner benytter også grunnlaget for beregning av eiendomsskatt.

En gjennomgang av internasjonal praksis viser at verdsettelsesmodellene ofte er begrenset til verdsettelse av primærboliger. Videre har landene typisk et mindre omfang av fritidsboliger enn Norge og Skandinavia. Dette innebærer at distinkte prisdrivere for fritidsboliger har fått lite oppmerksomhet i utvikling av modellene.

Utviklingen av en verdsettelsesmodell for fritidsboliger skiller seg derfor fra tidligere arbeid både nasjonalt og internasjonalt. De grunnleggende prinsippene for utvikling av modellen er likevel de samme. Det er vanlig å skille mellom tre ulike modeller for verdsettelse av boliger (International Association of Assessing Officers, 2018). De tre modellene omtales ofte som inntektsmodeller, kostnadsmodeller og markedsmodeller.

Inntektsmodeller
Baseres på boligens potensielle leieinntekter

Kostnadsmodeller
Baseres på kostnaden av boligens bygninger og eiendom

Markedsmodeller
Baseres på en statistisk modell for tidligere boligomsetninger

Inntektsmodeller tar utgangspunkt i boligens potensielle leieinntekter, der markedsverdien fastsettes som nåverdien av boligens fremtidige kontantstrøm. En slik modell brukes primært for å estimere den kommersielle verdien av forretningseiendom, og fordrer at det eksisterer et omfattende leiemarked for fritidsboliger.

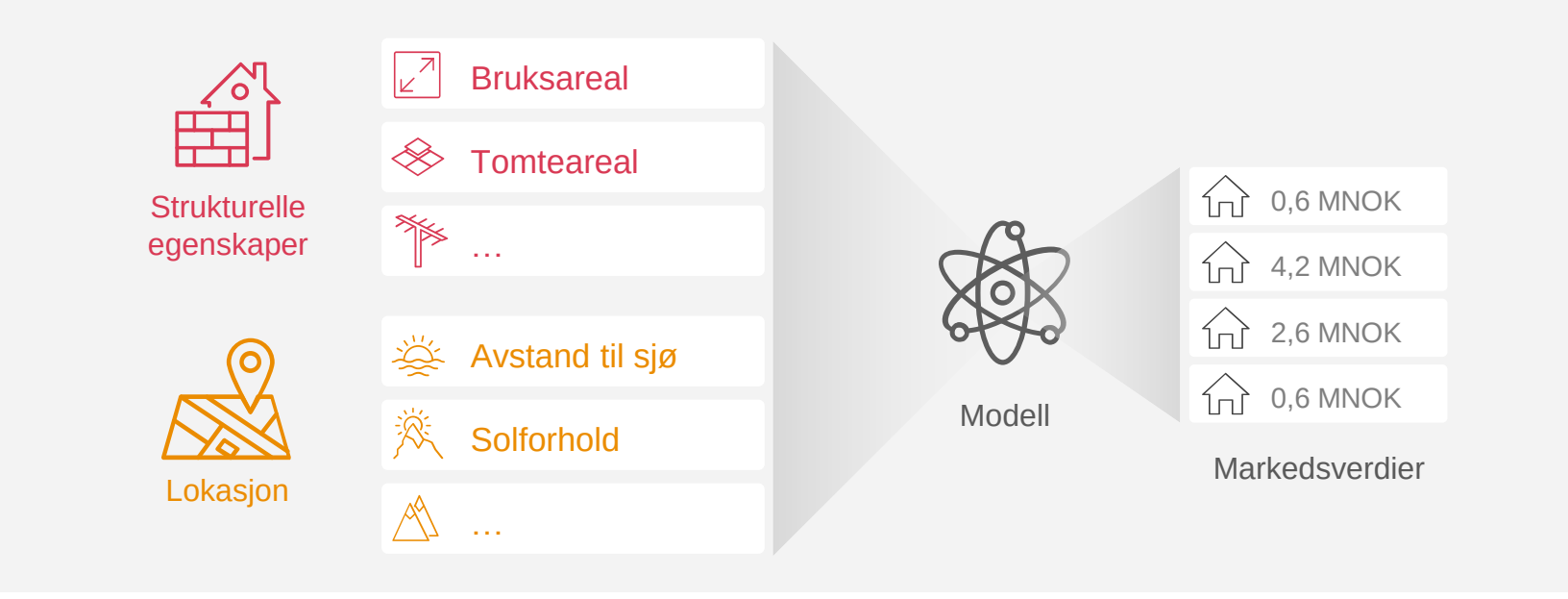
Kostnadsmodeller tar utgangspunkt i den estimerte kostnaden av boligens bygninger og eiendom. Modellene er basert på en økonomisk forutsetning om at informerte kjøpere ikke er villige til å betale mer enn kostnaden ved å oppføre en tilsvarende bolig med samme nytteverdi. Boligenes tekniske verdi fastsettes i dag ofte i forbindelse med taksering. Dette kan være en enkel øvelse for nyere boliger, men den tekniske verdien kan være vanskelig å estimere for eldre boliger med ulik grad av vedlikehold. Metoden fordrer også en manuell inspisering av landets boliger som vil være både tidkrevende og utfordrende å gjennomføre.

Markedsmodeller estimerer boligens markedsverdi gjennom en statistisk modellering av tidligere boligsalg. Modellene omtales i litteraturen som «direkte markedsmodeller» og er ofte den foretrukne verdsettelses-modellen for boliger og eiendommer med mange tilgjengelige markedsomsetninger (International Association of Assessing Officers, 2017). Premisset for modellene er at boligens markedsverdi er relatert til boligens individuelle egenskaper eller karakteristikker som bruksareal, alder og lokasjon. Slike underliggende egenskaper kan som regel ikke omsettes separat i markedet, og markedsprisen er gjerne sterkt avhengig av hvilken bruksverdi de ulike egenskapene gir på eiendommen.

Direkte markedsmodeller har vist seg å gi presise og stabile estimater for boligers markedsverdi gjennom flere studier og implementeringer. Dette inkluderer blant annet dagens verdsettelsesmodell for primær- og sekundærboliger. De er derfor det foretrukne valget for den utviklede verdsettelsesmodellen.

Markedsmodeller

Markedsmodeller skiller gjerne mellom boligens strukturelle egenskaper og egenskaper knyttet til boligenes lokasjon. Strukturelle egenskaper omfatter typisk bruksareal, tomtestørrelse og standard på bygninger, mens forhold som utsikt, solforhold, adkomst og strandlinje gjerne knyttes til boligenes lokasjon. Premisset for markedsmodellene er at sammenhengen mellom boligenes egenskaper og markedspris kan utledes fra et utvalg av tidligere solgte fritidsboliger. Det innebærer at modellen trenes til å lære hvordan egenskaper som areal, alder og beliggenhet påvirker fritidsboligenes salgspris. Modellen kan deretter brukes for å verdsette populasjonen av norske fritidsboliger.



Markedsmodellene verdsetter fritidsboliger basert på boligenes strukturelle egenskaper og lokasjon

Utvalget av fritidsboliger som benyttes for å trene modellen kalles ofte for treningsdata. For å utvikle en presis modell for verdsettelse er det avgjørende at treningsdataene er representative for populasjonen som skal verdsettes. Dette betyr at de underliggende egenskapene ved fritidsboligene som omsettes ikke bør avvike vesentlig fra populasjonen av alle norske fritidsboliger. Ettersom markedsmodellene baserer seg på tidligere omsetninger av fritidsboliger er det også avgjørende at salgsprisen reflekterer boligens faktiske markedsverdi. Dette innebærer at det er nødvendig å skille omsetninger til markedspris fra andre omsetninger som arv og gaver. Markedspris tilsvarer salgsprisen som selgeren vil kunne få for boligene i det åpne markedet, uten noen form for interessefellesskap med kjøper. Et slikt salg omtales ofte som et *fritt salg*.

Modeller for masseverdsettelse baseres typisk på offentlig registrerte boligopplysninger. Dette kan omfatte forhold som bruksareal, byggeår, tomteareal og bygningskvalitet. Omfanget av registrerte opplysninger er derfor ofte den begrensende faktoren for hvilke boligegenskaper modellen kan hensynta. Videre påvirkes modellenes treffsikkerhet i stor grad av kvaliteten og tilgjengeligheten av opplysningene som registreres.

Tidligere innføringer av nasjonale markedsmodeller for verdsettelse har avdekket vesentlige mangler i registrerte boligopplysninger. I Nord-Irland måtte om lag 20 prosent av boligene korrigeres som følge av mangler og feilregistreringer i datagrunnlaget (Gloudemans, 2008) . Ved innføring av verdsettelsessystemet for primær- og sekundærboliger i 2010 måtte boligeeiere selv innrapportere byggeår og primærrom til Skatteetaten. Det har vært utenfor prosjektets mandat å innhente eller korrigere registrerte boligopplysninger om fritidsboliger. Utviklingen av verdsettelsesmodellen er derfor basert på de offisielle registrene for eiendomsinformasjon og tinglysninger, matrikkelen og grunnboken.

Datautvalg

Verdsettelsesmodellen skal estimere markedsverdien av alle norske fritidsboliger. Ifølge Skatteetaten omfatter dette om lag 400 tusen fritidsboliger i 2018. Registrerte tinglysninger viser at om lag 25 prosent av disse boligene har blitt omsatt i løpet av de siste ti årene. Mange av boligene går i arv gjennom generasjoner, og over 60 prosent har ikke blitt omsatt siden årtusenskiftet.

Ettersom modellen skal verdsette alle fritidsboliger, uavhengig av hvorvidt de er tidligere omsatt, benyttes en direkte markedsmodell. Modellen baserer verdsettelsen på boligenes underliggende egenskaper. Det innebærer at modellen lærer hvordan egenskaper som areal, alder og beliggenhet påvirker boligens markedsverdi basert på tidligere omsatte boliger med kjente egenskaper og markedsverdi. Utviklingen av modellen forutsetter at det finnes et tilstrekkelig utvalg av tidligere omsatte fritidsboliger som kan benyttes som treningsdata.

De fleste omsetninger av boliger tinglyses. Tinglyste overdragelser utløser i de fleste tilfeller en dokumentavgift som beregnes basert på boligens markedsverdi. Det norske tinglysingsregisteret, grunnboken, er derfor et naturlig utgangspunkt for å identifisere markedsverdier for tidligere omsatte boliger. Undersøkelser fra SSB viser at en betydelig andel av de registrerte frie salg i grunnboken i realiteten er salg til familiemedlemmer med lavere pris enn boliger solgt uten tilsvarende interessefellesskap. Slike salg vil potensielt medføre at modellen underestimerer verdien av populasjonen av fritidsboliger. Prosjektet har derfor koblet tinglyste boligomsetninger til annonserte boliger på FINN.no for å sikre at boligene er solgt til markedspris.

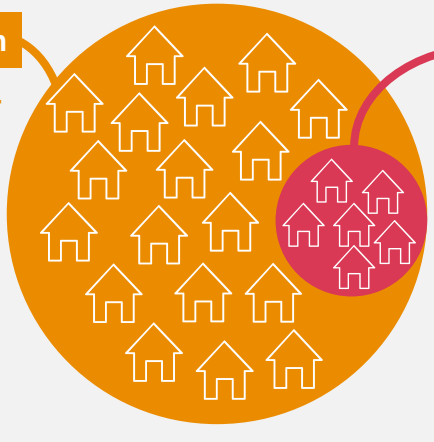
I perioden 2010 – 2020 ble om lag 110 tusen fritidsboliger annonsert gjennom FINN.no. Om lag 94 tusen av annonsene inneholder adresseopplysninger som gjør det mulig å knytte annonsen til en tinglyst omsetning. Nesten 20 tusen av annonsene omfatter boliger som Skatteetaten ikke skattlegger som fritidsboliger. Dette kan blant annet skyldes at boligene skattlegges som sekundærboliger. En del av boligene annonseres på FINN.no uten at annonseringen ender med et salg. Dette gjelder om lag 13 tusen av boligene, hvor enkelte av disse er annonsert flere ganger. Etter at disse ekskluderes utgjør utvalget 58 tusen annonser. Når boligannonsene kobles mot tinglysningsdata utgjør det endelige utvalget 38 tusen boliger. Det kan gjerne gå flere måneder fra fritidsboligen er solgt til kommunen tinglyser skjøtet. Følgende prioritering er derfor lagt til grunn for å knytte en annonsert bolig til en faktisk omsetning:

1. Tinglyst overdragelse med samme beløp som meglers salgspris innenfor 1 år fra rapportert salgsdato
2. Tinglyst overdragelse med samme beløp som meglers salgspris innenfor 1 år fra annonseringsdato
3. Tinglyst overdragelse med salgspris innenfor 25 % av prisantydning innenfor 1 år fra annonseringsdato

Populasjon: 400 tusen

Modellen skal verdsette alle boliger som skattlegges som fritidsbolig

Dette omfatter en stor mengde fritidsboliger som aldri har blitt solgt tidligere, samt boliger hvor det i dag finnes svært lite registrert informasjon om sentrale prisdrivere som areal og byggeår.



Utvalg: 38 tusen

Modellen trenes basert på tidligere boligomsetninger til markedspris

Modellen trenes basert på et utvalg av historiske markedsomsetninger i perioden 2010 – 2020. For å sikre at boligene er omsatt til reell markedspris er tinglysningene koblet mot annonserte fritidsboliger på FINN.no.

Tradisjonelle modeller for verdsettelse vs. maskinlæring

Bruken av markedsmodeller for verdsettelse har tradisjonelt vært ensbetydende med å benytte lineære regresjonsmodeller. Modellene kalles hedoniske verdsettelsesmodeller og deles gjerne inn i additive modeller, multiplikative modeller og hybridmodeller. Felles for modellene er at de forutsetter en antakelse om hvordan boligens markedsverdi avhenger av boligens egenskaper. Dette kalles ofte modellens *funksjonsform*. Deretter estimeres parametre som bestemmer hvordan de ulike egenskapene påvirker boligens markedsverdi.

Additive modeller antar at verdien av boligens underliggende egenskaper kan summeres til boligens markedsverdi. Et eksempel på et additivt forhold kan være boligens bygninger og tomt. Bygninger og tomt kan i teorien omsettes uavhengig av hverandre, og det er dermed naturlig å anta at verdiene kan summeres til boligens markedsverdi. Multiplikative modeller antar at verdien av boligens underliggende egenskaper kan multipliseres til boligens markedsverdi. Et eksempel på et multiplikativt forhold kan være boligens størrelse og alder. Dersom boligen er dårlig vedlikeholdt er det naturlig å anta at manglende vedlikehold kan redusere boligens kvadratmeterpris. Boligmarkedet er i virkeligheten basert på forhold som kan være både additive og multiplikative. Hybridmodeller er fleksible modeller som er utviklet for å kunne håndtere nettopp dette. Alle de tre modelltypene har vist seg å være godt egnet for verdsettelse av primær- og sekundærboliger. Den additive modellen er den mest brukte modellen som følge av enkelhet og tradisjon, mens multiplikative og hybride modeller ofte har vist seg å være mest treffsikre (O'Connor, 2002; Rossini og Kershaw, 2006). Dagens modell for verdsettelse av norske primær- og sekundærboliger er basert på en multiplikativ modell.

De tradisjonelle modellene har gjerne et begrenset antall variabler. Modellene er derfor enkle å forstå og transparente for skattytere. Det begrensede antallet variabler innebærer imidlertid at modellene i liten grad fanger opp mer granulerte egenskaper ved boligene som avstand til sjø og skibakker. Videre bygger markedsverdien ofte på komplekse avhengigheter mellom slike egenskaper som kan være krevende å fange opp med lineære modeller. Avhengighetene omtales som interaksjoner og kan omfatte flere variabler. Interaksjonseffektene representerer ofte sammenhenger som påvirker betalingsvilligheten i enkelte markedssegmenter. Et eksempel på en treveis interaksjon kan være at fritidsboligen (i) har en tinglyst rett til båtplass (ii) med sjødybde over 2,5 meter (iii) i Færder kommune. En slik kompleks interaksjon kan gjøre boligen mer attraktiv for eiere av store båter.

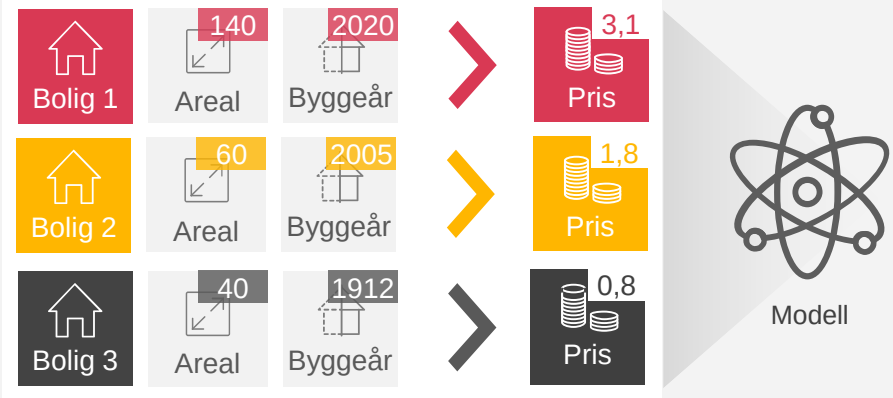
De siste årene har det vært en stor utvikling innenfor feltet maskinlæring, og det finnes i dag flere modeller som automatisk identifiserer og fanger opp denne typen interaksjonseffekter. I motsetning til de tradisjonelle metodene forutsetter ikke maskinlæringsmodellene en spesifisert funksjonsform, men overlater til modellen å identifisere et passende forhold mellom boligens egenskaper og markedsverdi. Siden modellene tillater komplekse sammenhenger mellom variablene oppnår de ofte en høyere treffsikkerhet enn tradisjonelle statistiske modeller. Den økte presisjonen går ofte på bekostning av modellens enkelhet og forklarbarhet.

Det har blitt argumentert for at maskinlæring har begrenset bruksverdi for boligverdsettelse så lenge skattytere ikke forstår hvordan de underliggende egenskapene ved boligen har bidratt til den estimerte markedsverdien. De siste årene har det imidlertid skjedd en stor utvikling innenfor ulike rammeverk for å forklare estimater fra flere typer maskinlæringsmodeller (Lundberg, 2017). Rammeverkene gjør det mulig å forklare komplekse modeller på måter som tidligere var forbeholdt lineære modeller. Shapley-verdier er et eksempel på et slikt rammeverk. Konseptet er basert på spillteori og gjør det mulig å forklare hvordan de ulike egenskapene ved en bolig har bidratt til modellens verdsettelse. Dagens maskinlæringsmodeller er dermed i stand til å fange opp komplekse sammenhenger innenfor boligenes nærområde, samtidig som hver enkelt verdsettelse fra modellen kan forklares.

Maskinlæring for verdsettelse

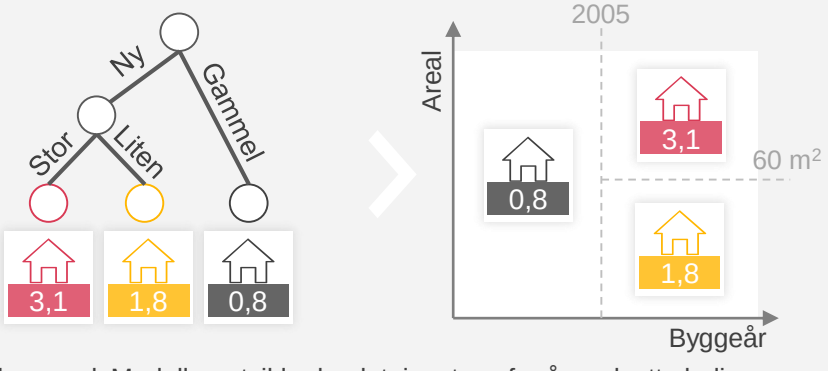
Maskinlæringsmodeller lærer automatisk fra erfaring, uten å bli instruert om hvordan et problem skal løses. Måten modellene lærer på kalles gjerne for algoritmer. Det finnes mange forskjellige maskinlæringsalgoritmer, men prinsippet for hvordan de lærer er det samme. Prosessen starter med at modellen lærer hvordan boligenes egenskaper påvirker markedsverdi basert på tidligere omsatte boliger med kjente markedsverdier. Modellen kan så bruke de lærte sammenhengene for å verdsette boliger den ikke har sett tidligere.

Steg 1: Modellen får tilgang til boliger med kjente markedsverdier



Modellen får tilgang til et utvalg av fritidsboliger med kjente egenskaper og markedsverdier. Boligens egenskaper kan eksempelvis være boligens alder og størrelse. Utvalget brukes for å trene modellen og kalles derfor for treningsdata. Verdsettelsesmodeller har, i likhet med eiendomsmeglere og takst-menn, behov for mye treningsdata for å gi gode estimater på boligenes markeds-verdi. I dette eksemplet får modellen tilgang til tre tidligere salg, men en presis modell for hele Norge vil kreve tusenvis av tidligere salg.

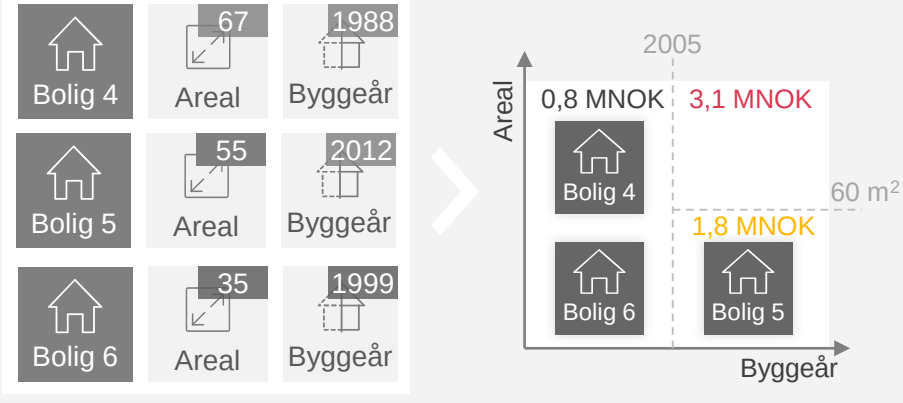
Steg 2: Modellen trenes til å lære sammenhengen mellom boligenes egenskaper og markedsverdi



Eksempel: Modellen utvikler beslutningstrær for å verdsette boligene

Eksempelet benytter beslutningstrær som er en av de enklere algoritmene å forklare. Et belsutningstre verdsetter boligene ved å dele dem inn i mindre og mindre grupper basert på boligenes egenskaper. Målet er å ende opp med grupper av boliger som er mest mulig like, og som dermed kan gis samme markedsverdi. Algoritmen velger derfor inndelingen basert på hvilken egenskaper som gir størst mulig reduksjon i varians. I eksemplet har modellen først delt boligene i nye og og gamle boliger, og deretter i små og store boliger for å ende med tre boliggrupper.

Steg 3: Modellen verdsetter boliger med ukjent markedsverdi basert på boligenes egenskaper



Når modellen har lært sammenhengene mellom boligenes egenskaper og markeds-verdi kan modellen benyttes for å verdsette boliger som modellen aldri har sett tidligere. I eksemplet har modellen lært tre regler: Boliger bygd før 2005 har en markedsverdi på 0,8 MNOK. Boliger bygd etter 2005 har en markedsverdi på 1,8 MNOK dersom arealet er under 60 kvm og 3,1 MNOK kr dersom arealet er over 60 kvm. Et beslutningstre med to delinger er imidlertid alene for enkelt til å verdsette alle typer boliger. I praksis benyttes derfor mange dype trær for verdsettelsen.

Kriterier for modell

En verdsettelsesmodell er en forenkling av virkeligheten, og det vil aldri være mulig å modellere alle faktorer ved etterspørsel og tilbud. Verdsettelse av boliger for skatteformål er masseoperasjoner, og varierende kvalitet i bygningsopplysninger innebærer at det ikke er mulig å stille samme krav til verdsettelsen som det stilles til takster ved enkeltstående salg. Formålet med verdsettelsesmodellen er derfor å oppnå et grunnlag for formuesbeskatning som er så riktig som mulig.

Siden modellen utvikles for skatteformål bør den også gi stabile og forklarbare verdsettelser over tid. Stabile verdsettelser innebærer at det ikke bør være store uforklarbare variasjoner når modellen reestimeres hvert år. At verdsettelsene er forklarbare innebærer at det er mulig å begrunne hver enkelt verdsettelse, og at det er mulig å forstå hvordan hvordan de ulike boligegenskaper påvirker markedsverdien i modellen. Prosjektet har formulert dette i tre kriterier som har fungert som en rettesnor for utviklingen av modellen.

Presisjon
Treffsikkerhet

Reduserer modellen avviket mellom grunnlaget for formues-verdi og reelle markedsverdier?

Stabilitet
Usikkerhet

Gir modellen en forutsigbar utvikling i formuesverdier for skattytere, uten store variasjoner?

Forklarbarhet
Intuitive resultater

Kan markedsverdiene forklares og er prisdriverne gjenkjennelige for skattyter?

Overordnede vurderingskriterier for utvikling av verdsettelsesmodellen.

De tre kriteriene kan stå i motsetning til hverandre. En mer kompleks modell vil gi økt presisjon, men redusert stabilitet og forklarbarhet. Utviklingen av verdsettelsesmodellen involverer dermed en avveining mellom kriteriene. En slik avveining kan kvantifiseres gjennom å knytte kriteriene til målbare størrelser. Det er mulig å knytte statistiske mål til presisjon og stabilitet, men det finnes i liten grad tilsvarende kvantifiserbare mål for forklarbarhet. Forklarbarhet er derfor formulert som fire kvalitative kriterier. En nærmere begrunnelse og beskrivelse av de ulike vurderingskriteriene er gjengitt i Kapittel 3.

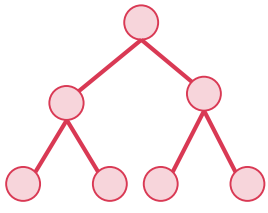
Presisjonsmål	Gjennomsnittlig absolutt avvik fra markedsverdi i prosent (MAPE)
Stabilitetsmål 1	Gjennomsnittlig variasjonskoeffisient for enkeltboliger ved årlig reestimering
Stabilitetsmål 2	Gjennomsnittlig standardavvik for boliger innenfor en kommune ved årlig reestimering
Forklarbarhetsmål 1	Verdsettelsen er basert på et håndterlig antall anerkjente prisdrivere for fritidsboliger
Forklarbarhetsmål 2	Hver enkelt verdsettelse kan forklares med bakgrunn i boligens egenskaper
Forklarbarhetsmål 3	Bidraget de ulike boligegenskapene gir til verdsettelsen er i tråd med forventninger
Forklarbarhetsmål 4	Verdsettelser fra modellen kan forklares til skattytere på en enkel måte

Konkrete vurderingskriterier for utvikling av verdsettelsesmodell.

Valg av modell

Kriteriene presisjon, stabilitet og forklarbarhet legger føringene for utvikling av verdsettelsesmodellen. For å identifisere en modell som ansees for å best tilfredsstille modellkriteriene har prosjektet benyttet en kombinasjon av studier om boligverdsettelse, rammeverk for forklarbarhet i maskinlæring og tidligere modelleringserfaring. Den valgte modellen er en modell basert på beslutningstrær. Trebaserte modeller er kjent for å oppnå god presisjon og nyere studier viser at modellene kan oppnå bedre resultater enn tradisjonelle metoder for verdsettelse. (McCluskey, 2014; Antipov og Pokryshevskaya, 2010; Hong og Kim, 2020). Utviklingen av nyere rammeverk for forklarbare maskinlæringsmodeller (Lundberg, 2017) innebærer at estimatene fra trebaserte modeller også kan forklares på en tilfredsstillende måte.

Andre maskinlæringsmodeller som blant annet nevrale nettverk, KNN (*K-Nearest Neighbor*) og SVM (*Support Vector Machines*) har blitt testet i prosjektet. Disse modellene viste seg å gi lavere presisjon enn trebaserte modeller, og kan i tillegg være vanskeligere å forklare. For å belyse i hvilken grad den trebaserte modellen tilfredsstiller modellkriteriene er det også utviklet en tradisjonell lineær verdsettelsesmodell. Modellen er en multiplikativ modell som bygger på SSBs modell for verdsettelse av primær- og sekundærboliger og tidligere studier. For enkelhet estimeres den lineære modellen for hele landet. Dette avviker fra SSBs modell som estimerer separate modeller for ulike regioner. Den lineære modellen fungerer som et sammenlignings-grunnlag for den utviklede verdsettelsesmodellen og resultatene for begge modeller presenteres gjennomgående i rapporten. En mer detaljert beskrivelse av modellene finnes i James et al (2013).

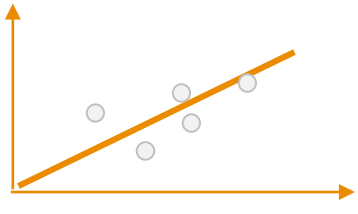


Trebasert modell

Maskinlæringsmodell

Trebaserte modeller er bygd opp av beslutningstrær. Gjennom treet deles boligene inn i mindre og mindre grupper basert på underliggende egenskaper. Boliger som ender sammen i treets nederste gren har tilsvarende egenskaper og tillegges den samme markedsverdien.

Prosjektet har benyttet en metode som kalles “boosted trees”, som benytter en sekvensiell samling av slike beslutningstrær. Trebaserte modeller har gjerne høy prediksjonsevne, men kan i noen tilfeller være ustabile på grunn av modellens kompleksitet. Modellene kan forklares ved bruk av SHAP-verdier (Lundberg, 2017) som for hvert estimat beskriver hver enkelt egenskaps bidrag til verdsettelsen.



Lineær modell

Sammenligningsgrunnlag

Lineære modeller antar en lineær sammenheng mellom boligens underliggende egenskaper og markedsverdi. Modellene er i stand til å modellere ikke-lineære sammenhenger gjennom numeriske transformasjoner og kurvetilpasninger, men dette forutsetter at modellens funksjonsform spesifiseres.

Videre er modellene i stand til å modellere avhengigheter mellom variabler, men det forutsettes at slike forhold identifiseres og representeres på en passende måte. Dette medfører at modellene gjerne krever en del forarbeid for å oppnå god presisjon på komplekse problemstillinger. Lineære modeller har typisk en enklere struktur enn maskinlæringsmodeller og er følgelig både svært stabile og forklarbare.

Beskrivelse av trebasert modell og lineær modell.

3. Kriterier for modell

Presisjon

En presis modell gir verdsettelser som avviker minst mulig fra boligens faktiske markedsverdi. Det er vanlig å benytte det gjennomsnittlige absolutte avviket i prosent for alle boliger som et mål på modellens presisjon. Det absolutte avviket benyttes for å unngå at overestimeringer og underestimeringer nuller hverandre ut i beregningen. For å likebehandle billige og dyre boliger benyttes et prosentvis avvik relativt til faktisk markedsverdi. Avviket omtales på engelsk som MAPE (Mean Absolute Percentage Error) og er det foretrukne presisjonsmålet i prosjektet. Når begrepet *avvik* benyttes i rapporten henviser dette til MAPE.

Presisjonsmål

Prosjektet benytter gjennomsnittlig absoluttavvik i prosent (MAPE) som et mål på modellens treffsikkerhet.

Upresis modell med høy MAPE

	Estimert markedsverdi	Faktisk salgspris	Absolutt avvik
Bolig 1	1,2 MNOK	2,3 MNOK	48%
Bolig 2	0,5 MNOK	0,3 MNOK	67%
Gjennomsnittlig absolutt avvik (MAPE)			59%

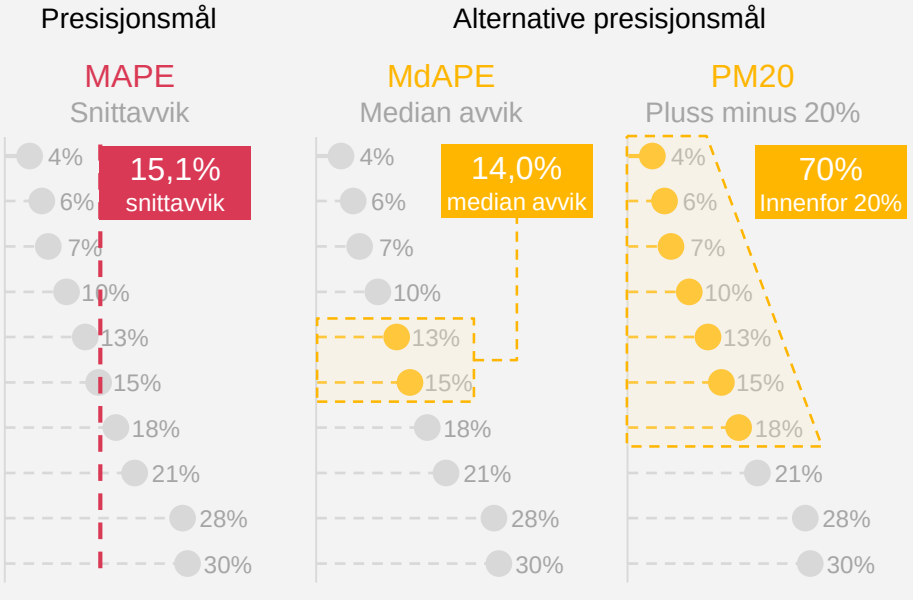
Presis modell med lav MAPE

	Estimert markedsverdi	Faktisk salgspris	Absolutt avvik
Bolig 1	2,2 MNOK	2,3 MNOK	4%
Bolig 2	0,35 MNOK	0,3 MNOK	17%
Gjennomsnittlig absolutt avvik (MAPE)			10,5%

Presisjonsmål: Gjennomsnittlig absoluttavvik i prosent (MAPE)

MAPE påvirkes av at enkelte boliger kan være vanskelig å verdsette. Slike boliger gir potensielt store utslag ved beregning av gjennomsnittlige avvik. Medianavviket er et alternativt mål som er mer robust mot store avvik for enkeltboliger. Målet omtales som MdAPE (Median Absolute Percentage Error). PM20 (Pluss minus 20 prosent) er et tilsvarende mål som beskriver hvor stor andel av boligene som har et avvik innenfor 20 prosent av markedsverdi. MdAPE og PM20 representerer alternative mål i prosjektet, og rapporteres i resultatene.

MNOK	Estimert markedsverdi	Faktisk salgspris	Absolutt avvik
	1,82	1,75	4%
	0,85	0,90	6%
	1,71	1,60	7%
	0,66	0,73	10%
	1,50	1,33	13%
	1,28	1,50	15%
	1,30	1,10	18%
	1,38	1,75	21%
	2,30	1,80	28%
	0,51	0,73	30%



Beregning av MAPE, MdAPE og PM20 for et utvalg av ti boliger.

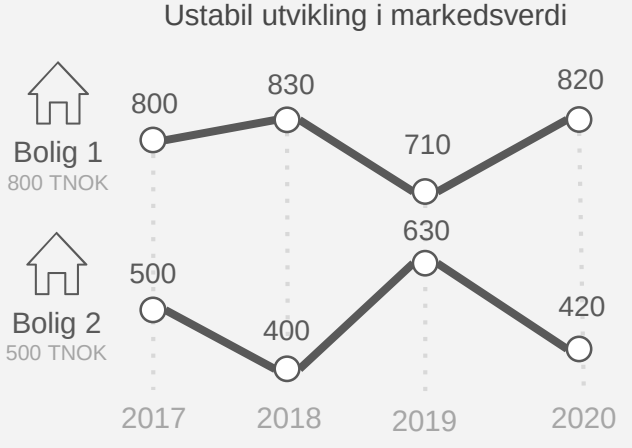
Stabilitet

Formuesbeskatning bør være forutsigbart og rettferdig for den enkelte skattyter. Dette innebærer at det ikke bør være store uforklarlige variasjoner i boligenes skattemessig verdi fra år til år. Enkelte modeller kan være svært sensitive for utvalget av boliger som benyttes for å trene modellen. For slike modeller kan små endringer i treningsgrunnlaget potensielt gi store utslag for boligenes estimerte markedsverdier. Ustabiliteten kan oppstå av flere årsaker. I lineære modeller kan det eksempelvis skyldes ekstremverdier i datagrunnlaget. I maskinlæringsmodeller kan det skyldes at modellene er så komplekse at de lærer seg sammenhenger i treningsdataene som ikke er generaliserbare til alle fritidsboliger.

Siden verdsettelsesmodellen skal ivareta boligenes fremtidige markedsutvikling må datagrunnlaget oppdateres årlig og modellen trenes på nytt. Modellen bør derfor gi stabile verdsettelser over tid som er i tråd med skattyteres forventede markedsutvikling. Gjennom å simulere årlige skatteoppgjør i perioden 2017 til 2020 er det mulig å knytte et statistisk mål til den årlige variasjonen i markedsverdi. Prosjektet har benyttet den gjennomsnittlige variasjonskoeffisienten for boligenes markedsutvikling som et mål på årlig variasjon for skattytere. En variasjonskoeffisient på 0 tilsvarer ingen markedsutvikling. Dette representerer dagens situasjon der formuesverdiene ikke har blitt justert siden 2014. Det er derfor ønskelig at variasjonskoeffisienten skal være lav, men større enn 0.

Stabilitetsmål 1

Som et mål på variasjonen i årlig markedsutvikling for enkeltboliger benyttes variasjonskoeffisienten fra simulerte skatteoppgjør. Variasjonskoeffisienten er standardavviket dividert med gjennomsittet for de årlige markedsverdiene. For å gi et aggregert mål på variasjonen benyttes gjennomsnittlig variasjonskoeffisient.



	Årlig markedsutvikling				Var.koeff
Bolig 1	800	830	710	820	7%
Bolig 2	500	400	630	420	21%
Gjennomsnittlig variasjonskoeffisient					14%

	Årlig markedsutvikling				Var.koeff
Bolig 1	800	800	820	840	2%
Bolig 2	500	510	510	540	4%
Gjennomsnittlig variasjonskoeffisient					3%

Stabilitetsmål 1: Gjennomsnittlig variasjonskoeffisient for boligenes markedsutvikling.

Stabilitet

Etterspørselen i boligmarkedet varierer over tid og sesongbaserte variasjoner kan påvirke etterspørselen av ulike boligtyper og områder ulikt. Fritidsboliger innenfor samme geografiske område bør likevel følge mer eller mindre samme markedsutvikling i fravær av lokale endringer som nye skiheiser eller oppgraderte naturområder. Likebehandling er et underliggende skatteprinsipp som innebærer at like boliger bør verdsettes likt. Innenfor samme geografiske område bør modellen derfor gi tilsvarende årlig utvikling i markedsverdi for fritidsboliger, forutsatt at boligene er i samme stand.

Standardavviket av enkeltboligers utvikling i markedsverdi innad i en kommune benyttes som et mål på modellens stabilitet innenfor tilsvarende geografiske områder. Et lavt standardavvik indikerer at boligene følger en tilsvarende markedsutvikling

Stabilitetsmål 2

Som et mål på variasjonen i årlig markedsutvikling innenfor et område benyttes standardavviket for boligenes årlige markedsutvikling innenfor hver enkelt kommune. For å gi et aggregert mål på variasjonen benyttes det gjennomsnittlige standardavviket for alle kommuner.

Ustabil utvikling i markedsverdi innenfor kommune

	Årlig markedsutvikling				Std.avvik
Kommune 1	8%	-10%	2%	▶	9%
Kommune 2	5%	-20%	-12%	▶	13%
Gjennomsnittlig standardavvik					11%

Stabil utvikling i markedsverdi innenfor kommune

	Årlig markedsutvikling				Std.avvik
Kommune 1	2%	-1%	0%	▶	2%
Kommune 2	2%	0%	2%	▶	1%
Gjennomsnittlig standardavvik					1,5%

Stabilitetsmål 2: Gjennomsnittlig standardavvik for årlig markedsutvikling innenfor kommuner.

De to stabilitetsmålene skal reflektere hvordan skattytere vil oppleve årlige endringer i markedsverdi. Målene påvirkes derfor både av modellens stabilitet og den faktiske etterspørselen i boligmarkedet. Markeds-utviklingen innebærer at en perfekt stabil modell ikke vil oppnå en variasjonskoeffisient på 0 eller et standardavvik på 0 innenfor en kommune. Det er mulig å benytte stabilitetsmål som isolerer modellens stabilitet gjennom å trekke et tilfeldig utvalg av boliger gjentatte ganger for samme skatteår. Prosessen omtales som bootstrapping innenfor statistikk. Prosjektet har benyttet bootstrapping i tillegg til de to stabilitetsmålene. Dette viser seg å gi tilsvarende resultater, og prosjektet har derfor fokusert på målene som best reflekterer skatteyteres oppfatning av årlige endringer i markedsverdi.

Forklarbarhet

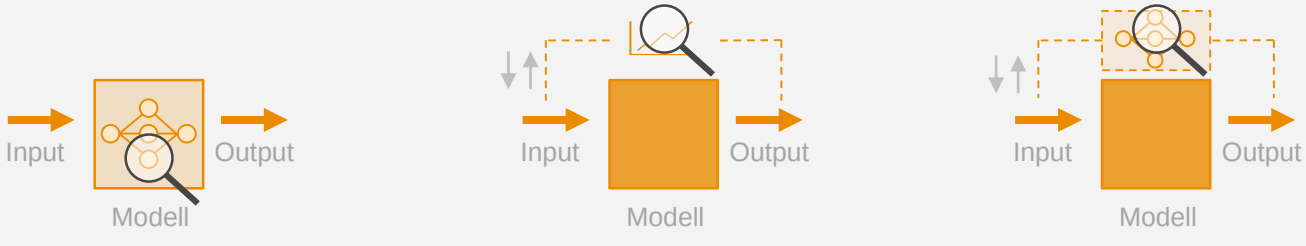
Presisjon og stabilitet kan knyttes til målbare størrelser som gjør det mulig å kvantifisere avveiningen mellom de to kriteriene. Modellens forklarbarhet er vanskeligere å kvantifisere siden den faktiske forklaringen bak en boligverdsettelse er ukjent. Videre er det i liten grad konsensus om metoder for å kvantifisere hvor forklarbar en modell er, og hvor riktig en forklaring er. Utviklingen i bruken av maskinlæringsmodeller har medført at temaet har fått stor oppmerksomhet det siste tiåret, men det mangler fremdeles en omforent definisjon av begrepet forklarbarhet (Molnar et al, 2020; Doshi-Velez og Kim 2017). For å definere forklarbarhet som et vurderingskriterie må prosjektet derfor ta stilling til hva en forklarbar modell for verdsettelse innebærer.

Diskusjonen om forklarbare modeller for boligverdsettelse har pågått siden det første kjente systemet for masseverdsettelse ble implementert i St Paul, Minnesota i 1896 (Silverherz, 1936). I senere tid har Gloudemans (1982) beskrevet syv kjennetegn ved forklarbare verdsettelsesmodeller:

- Modellene har en enkel funksjonsform og det er et rasjonale for bruken av variablene
- Modellenes variabler gir en rimelig effekt på boligens markedsverdi i tråd med forventninger
- Modellene gir stabile markedsverdier over tid uten uforklarbare variasjoner
- Modellene er enkle å forklare til skattytere, og beregningen kan demonstreres på en simpel måte
- Modellene gir bedre innsikt dersom de kan forklares gjennom tidligere sammenlignbare salg
- Det bør være overensstemmelse mellom variablenes effekt i eventuelle sub-modeller for ulike boligtyper
- Det bør være mulig å isolere verdien av tomt og bolig i områder der disse skattelegges ulikt

Gloudemans beskrev kjennetegnene ved forklarbare modeller på et tidspunkt der lineære modeller var i ferd med å få sitt gjennombrudd for boligverdsettelse. Dette var før de fleste av dagens maskinlæringsmodeller ble utviklet, og det eksisterte få alternative metoder. Wu og Gloudemans (2019) beskriver maskinlæring-modellenes rolle for verdsettelser som uunngåelig i dag , men få nyere studier diskuterer modellenes forklarbarhet i en kontekst av boligverdsettelser. McCluskey (2014) påpeker at modellenes resultater er oppløftende, men at algoritmene er lite gjennomsiktige og at modellene mangler metoder for å beskrive effekten av de ulike variablene.

Siden McCluskeys studie har det skjedd en stor utvikling av metoder for å forklare og forstå ulike typer maskinlæringsmodeller. Molnar et al (2020) skiller mellom tre ulike metoder for forklarbarhet avhengeig av om metodene analyserer (1) modellens komponenter, (2) modellens sensitivitet eller (3) om forklaringen baserer seg på surrogatmodeller.



Metoder som analyserer modellenes komponenter
Lineære modeller og beslutningstrær er eksempler på naturlig forklarbare modeller som kan forklares ved å analysere modellenes komponenter, dvs. modellens struktur og parametre

Metoder som analyserer modellenes sensitivitet
Metodene er typisk modell-agnostiske der modellen anses som et lukket system. Metodene analyserer hvordan modellens estimater påvirkes av endringer i variablenes verdier.

Metoder som baserer seg på surrogatmodeller
Surrogatmodeller er naturlig forklarbare modeller som er utviklet for å kopiere oppførselen til en ikke-forklar modell. Metodene baserer forklaringene på surrogatmodellen.

Tre ulike metoder for å forklare maskinlæringsmodeller som beskrevet i Molnar et al (2020).

Forklarbarhet

Metoder som analyserer modellens komponenter inkluderer blant annet tradisjonelle verdsettelsesmodeller som kan forklares gjennom regresjonskoeffisienter. Slike lineære modeller omtales ofte som naturlig forklarbare. Modellene ansees bare for å være forklarbare opp til en viss dimensjon. Eksempelvis vil en lineær regresjonsmodell med hundrevis av variabler og interaksjoner ikke være spesielt forklarbar.

Maskinlæringmodellenes struktur er ofte så kompleks at de er tilnærmet umulige å forstå gjennom å analysere modellens komponenter. Det er derfor utviklet metoder som kan forklare modellene gjennom å analysere modellens sensitivitet. Denne tilnærmingen innebærer at modellen behandles som et lukket system som tar inn data og gir tilbake estimater. Metodene forklarer den underliggende modellen gjennom å manipulere dataene som går inn i modellen og deretter analysere de respektive endringene i modellens estimater. Den mest populære anvendelsen av slike metoder er forklaringsrammeverket SHAP (Shapley Additive Explanations) av Lundberg og Lee (2017). Metoden har blitt utvidet for å bedre håndtere avhengigheter mellom variablene av Aas, Jullum og Løland (2019).

Surrogatmodeller er en blanding av de to tilnærmingene. Metoden behandler den underliggende modellen som et lukket system og benytter en naturlig forklarbar modell for å kopiere modellens oppførsel. Den mest populære anvendelsen av denne metoden er forklaringsrammeverket LIME (Local Interpretable Model-agnostic explanations) av Ribeiro, Singh, og Guestrin (2016).

Det er vanlig å skille mellom lokale og globale tilnærminger til forklarbarhet av maskinlæringsmodeller. En global forklaring er en forklaring av hele modellen. Slike forklaringer forutsetter at mottakeren er i stand til å forstå aspekter ved hele modellen på en gang. For boligverdsettelse kan dette være en beskrivelse av hvordan de ulike variablene i gjennomsnitt bidrar til boligens markedsverdi. En lokal forklaring innebærer å forklare en enkelt verdsettelse. En slik tilnærming ser bort fra store deler av modellen, og lokale forklaringer gir derfor bare en nøyaktig representasjon av en spesifikk del av modellen. For boligverdsettelse kan en lokal forklaring være en beskrivelse av hvordan de ulike egenskapene ved boligen har bidratt til verdsettelsen.

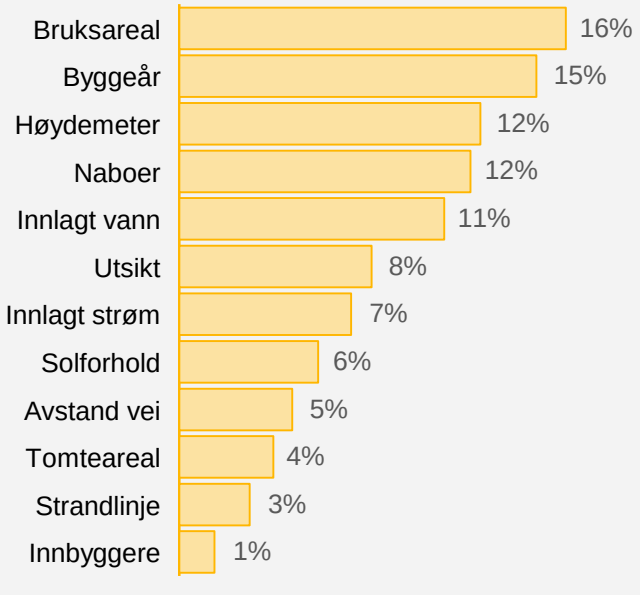
Lokal forklarbarhet

Hvordan bidrar egenskapene for en enkelt fritidsbolig til boligens estimerte markedsverdi?

Gjennomsnittlig markedsverdi		1 200 000
Egenskaper som øker verdi		240 000
Avstand til privat vei	31 m	120 000
Innbyggere (100 km)	125 501	60 000
Avstand innsjø	129 m	60 000
Egenskaper som senker verdi		360 000
Tomteareal	365 kvm	160 000
Høydemeter	263	120 000
Byggeår	Ukjent	80 000
Estimert markedsverdi		1 080 000

Global forklarbarhet

Hvordan påvirker de ulike boligegenskapene den estimerte markedsverdien for alle fritidsboliger?



Fiktive eksempler på lokale og globale forklaringer for en verdsettelsesmodell.

Forklarbarhet

SHAP-rammeverket representerer en enhetlig tilnærming for å gi lokale og globale forklaringer for alle typer maskinlæringsmodeller. Gjennom SHAP er det mulig å forklare hvordan egenskaper ved hver enkelt bolig har bidratt til boligens estimerte markedsverdi. Bidraget kan være i kroneverdi eller prosent og omtales som variabelens SHAP-verdi. Metoden er basert på et spillteoretisk fundament som sikrer at egenskapenes bidrag til markedsverdi fordeles på en rettferdig måte. Dette skiller metoden fra surrogatmodeller som LIME som ikke garanterer en tilsvarende optimal fordeling.

Ulempen med SHAP er at forklaringene fordrer en simulering av alle mulige kombinasjoner av boligegenskaper. Antall kombinasjoner øker eksponentielt med antall variabler i modellen, som innebærer at det ikke er praktisk gjennomførbart å bruke SHAP for modeller med mange variabler. Det er derfor utviklet algoritmer som estimerer SHAP-verdier for ulike typer maskinlæringsmodeller. Maskinlæringsmodeller basert på beslutningstrær har en spesiell struktur som gjør de spesielt godt egnet for bruk av SHAP. For disse modellene er det utviklet algoritmer som gir eksakte SHAP-verdier på en effektiv måte. Den raske beregningen innebærer at det er mulig å beregne alle de nødvendige SHAP-verdiene for å gi globale modellforklaringer. Slike forklaringer inkluderer blant annet analyser av variablenes gjennomsnittlige viktighet, samt variablenes avhengigheter og interaksjoner. SHAP-verdier for trebaserte modeller ansees derfor å være et tilfredsstillende rammeverk for forklaring av en verdsettelsesmodell.

Selv om det eksisterer rammeverk for lokale og globale forklaringer innebærer ikke dette at modellene er umiddelbart forståelige for skattytere. I praksis kan regresjonskoeffisientene fra en lineær modell offentliggjøres, men få skattytere vil oppleve dette som en tilfredsstillende forklaring på boligens estimerte markedsverdi. Det er derfor mange hensyn som må tas i utviklingen av en forklarbar modell. Disse hensynene omfatter blant annet hvilke variabler som inngår i modellen, hvordan disse variablene skal representeres og hvordan forklaringene presenteres til skattyter.

Bør forklaringen være kontrastiv?

En kontrastiv forklaring er en forklaring som besvarer hvorfor modellen ga en verdsettelse på X kroner og ikke Y kroner. Lipton (1990) konkluderer med at menneskelige forklaringer bør være kontrastive. En kontrastiv forklaring for boligverdsettelse kan være å ta utgangspunkt i hvilke egenskaper ved boligen som skiller seg fra en gjennomsnittlig fritidsbolig. Differansen fra gjennomsnittlig markedsverdi kan dermed forklares i at boligen var 5 kvadratmeter større og 20 år eldre enn en gjennomsnittsbolig. SHAP gir denne typen kontrastive forklaringer, og det er også mulig å basere forklaringen på gjennomsnittlig markedsverdi innenfor en kommune eller fylke.

+200 000 KRONER VS 4,2%

Bør forklaringen være relativ?

Bidraget som de ulike egenskapene ved boligen har gitt til verdsettelsen uttrykkes i en additiv modell som kroneverdier. En forklaring fra en slik modell kan eksempelvis vise at boligens byggeår reduserer verdsettelsen med 200 tusen kroner. I en multiplikativ modell uttrykkes bidraget som prosenter. I et slikt tilfelle kan boligens byggeår redusere verdsettelsen med 5 prosent. En positiv egenskap ved å forklare bidraget i prosent er at dette er relative størrelser som gjør det enkelt å sammenligne forklaringen på tvers av alle boliger. Ved å beregne gjennomsnittet av bidragene er det mulig å gi en global forklaring av modellen som dermed også kan uttrykkes i prosent.

Forklarbarhet

Er variablene gjenkjennelige for skattytere?

En forklarbar modell bør baseres på anerkjente prisdrivere for fritidsboliger som skattytere vil kjenne seg igjen i. Dette betyr at modellen bør bygge på de samme egenskapene som eiendomsmeglere og boligeiere vektlegger når boligene annonseres for salg. Videre bør de beregnede verdiene for de ulike egenskapene også være gjenkjennelige. Eksempelvis kan en skattyter oppleve det som urettferdig om modellen har vektlagt gode adkomstforhold som følge av kort luftlinjeavstand til vei. Dersom boligen i realiteten ligger på en bratt høyde som ikke gjør det praktisk mulig å benytte korteste avstand til vei vil en slik forklaring oppleves lite tillitsvekkende.

Hvor mange variabler skal inngå i modellen?

Dersom modellen skal gi treffsikre verdsettelser må den inkludere variabler som har signifikant betydning for boligens markedsverdi. Samtidig vil en modell med mange variabler bli vanskeligere å forstå for skattytere. Det er derfor nødvendig å finne en avveining mellom treffsikkerhet og forklarbarhet. Et eksempel kan være boligens størrelse. I det norske eiendomsregisteret kan dette registreres som både bruksareal, bruttoareal og bebygd areal. Modellen kan oppnå økt treffsikkerhet ved å inkludere alle arealene siden hver av dem kan mangle verdier eller inneholde feil. Samtidig vil tre ulike arealvariabler i en forklaring fremstå som vanskelig å forstå for skattyter.

Hvordan bør variablene representeres?

Boligenes egenskaper kan representeres med numeriske verdier på ulike måter. Solforhold bestemmes eksempelvis av hvor lenge solas vertikale vinkel er større enn den vertikale vinkelen på terrengets horisont rundt boligen. Dermed kan solforhold både representeres som antall timer med sol, eller som den gjennomsnittlige vertikale vinkelen på terrengets horisont. Sistnevnte kan være den matematisk representasjonen som gir best resultater for modellen, men er mindre forståelig og gjenkjennelig for skattyter. Dersom den oppnådde forbedringen i modellen er liten, bør antall soltimer per døgn eller år favoriseres for forklarbarhet i modellen.

Hvordan håndtere manglende verdier?

De fleste modeller forutsetter at data som benyttes for å trene modellen er komplett uten manglende verdier. Dette er sjelden tilfelle for boligopplysninger som ofte mangler vesentlig informasjon som alder og størrelse. En vanlig fremgangsmåte er å erstatte den manglende verdien med en sannsynlig verdi, eksempelvis benytte det vanligste bruksarealet for fritidsboliger. En annen strategi kan være å estimere boligens bruksareal basert på grunnflate og høyde fra kartdata. Når modellens verdsettelser skal forklares bør det synliggjøres i hvilke tilfeller den faktiske verdien er ukjent, og det bør begrunnes hvordan den brukte verdien har fremkommet.

I tillegg vil skattyters forståelse av modellen påvirkes av hvordan forklaringene visuelt presenteres. For en modell med mange variabler kan det virke overveldende dersom alle variablene presenteres samlet. Et alternativ er å gruppere sammen effekten av variabler som tilhører samme kategori. For adkomstforhold kan eksempelvis bidraget fra avstand til nærmeste vei, differanse i høydemeter og nærmeste veitype summeres til et samlet adkomstbidrag som skattyter selv kan ekspandere eller kollapse i et egnet brukergrensesnitt.

Forklarbarhet

Modellens forklarbarhet er et resultat av en lang rekke valg hvor det i liten grad finnes definisjoner og fasitsvar. Prosjektet har koblet Gloudemans (1982) kjennetegn ved forklarbare verdsettelsesmodeller og nyere studier for å definere fire vurderingskriterier for forklarbarhet.

Forklarbarhetskriterier for verdsettelsesmodell

1 Verdsettelsen baseres på et håndterlig antall boligegenskaper som påvirker markedsverdi
Verdsettelsen bør bygge på anerkjente prisdrivere for fritidsboliger. Prisdriverne bør ha signifikant påvirkning på boligens markedsverdien og representeres på en distinkt måte, dvs. uten overlapp.

2 Hver enkelt verdsettelse kan forklares med bakgrunn i boligens egenskaper
Bidragene fra hver enkelt egenskap ved boligen kan summeres eller multipliserer til boligens estimerte markedsverdi, avhengig om modellen er en en additiv eller multiplikativ modell.

3 Bidraget de ulike boligegenskapene gir til verdsettelsen er i tråd med forventninger
Det forventes at egenskaper som større bruksareal og strandlinje gir et positivt bidrag til boligens markedsverdi. Modellen bør bekrefte slike forventninger gjennom lokale og globale forklaringer.

4 Verdsettelser fra modellen kan forklares til skattytere på en enkel måte
Det bør være enkelt for skattyter å forstå hvordan ulike egenskaper ved boligen bidrar til en lavere eller høyere verdsettelse. Egenskapene bør representeres på en gjenkjennelig måte.

4. Variabler og datakilder

Valg av variabler

Det er potensielt mange variabler som påvirker markedsverdien av en fritidsbolig. Ettersom det ikke er praktisk gjennomførbart å innhente alle relevante variabler for verdsettelsen av fritidsboliger, har det vært nødvendig å prioritere hvilke variabler som skal benyttes i modellen. For å foreta denne prioriteringen har prosjektet først tatt utgangspunkt i variabler som er brukt i tidligere akademiske studier for verdsettelse av boliger. Videre er det gjort en gjennomgang av meglerundersøkelser og egenskaper som fremheves for fritidsboliger annonsert på FINN.no. I tillegg har det blitt gjennomført arbeidsmøter med deltakere fra Finansdepartementet, PwC og SSB for å identifisere ytterligere variabler som kan være relevante. Variablene er deretter vurdert ut fra kriteriene forventet tilgjengelighet, enkelhet og aksept for å prioritere hvilke variabler som skal innhentes.

Turrute | Strandlinje | Solforhold | Skibakker
Tomteareal | Avstand til vei | Båtplass
Bruksareal | Utsikt | Byggeår
Reguleringsplaner | … | … | …

Steg 1: Bruttoliste med variabler

Dette er en liste over variabler som ble vurdert ved oppstart for prosjektet. Listen er basert på:

- Akademiske studier for verdsettelse av boliger.
- Egenskaper ved annonserte fritidsboliger.
- Arbeidsmøter i utvidet prosjektgruppe.

Steg 2: Prioritering av variabler

Hver enkelt variabel vurderes ut fra forventet forklaringskraft i modellen og følgende tre kriterier:

- **Tilgjengelighet:** Forventet tilgjengelighet for variabelen for alle fritidsboliger.
- **Enkelhet:** Forventet ressursbruk for å innhente data for variabelen.
- **Aksept:** Forventet aksept for variabelen som prisdriver blant boligeiere.

Bruksareal
Byggeår
…
Strandlinje

Steg 3: Innhenting og testing av variabler

Data for de prioriterte variablene innhentes for alle fritidsboliger. Den reelle forklaringskraften for hver variabel testes deretter i modellen for å ende opp med et endelig utvalg av variabler.

Prosessen for identifisering og utvelgelse av variabler for modellen.

Kriterier for valg av variabler

Som nevnt i kapittel 2 skiller man gjerne mellom variabler som er knyttet til boligens strukturelle egenskaper og variabler som er knyttet til boligens lokasjon. Disse omtales henholdsvis som strukturelle variabler og lokasjonsvariabler i rapporten.

Strukturelle variabler	Lokasjonsvariabler
Bruksareal	Avstand til sjø
Tomteareal	Solforhold
Energikilde	Antall høydemeter

Eksempler på strukturelle variabler og lokasjonsvariabler.

Det er lagt vekt på at skattemessige avgjørelser skal fattes basert på korrekte data og at likebehandling og transparens setter føringer for innholdet i verdsettelsesmodellen. Det er derfor forutsatt at de strukturelle variablene baserer seg på offisiell registrert informasjon i matrikkelen, mens lokasjonsvariabler benytter den tilgjengelige datakilden med best dekning for alle fritidsboliger. Variablene er vurdert etter skalaen lav, medium og høy for hvert av vurderingskriteriene tilgjengelighet, enkelhet og aksept.

Tilgjengelighet

En variabel som forventes å være tilgjengelig for alle fritidsboliger vurderes å ha høy grad av tilgjengelighet. Et eksempel på dette er koordinatene for en fritidsbolig. Dette er et obligatorisk registreringsfelt i matrikkelen som er utfylt for alle boliger. Dersom variabelen er tilgjengelig for en mindre gruppe fritidsboliger vurderes tilgjengeligheten som medium. Et eksempel på dette kan være energikilde som sjelden er utfylt i matrikkelen. Dersom variabelen bare er tilgjengelig for noen få fritidsboliger vurderes tilgjengeligheten som Lav. Dette kan eksempelvis være megleropplysninger som kun er tilgjengelig for tidligere omsatte boliger.

Enkelhet

Innhenting av nye datakilder og variabler vil typisk medføre ekstra ressursbruk i form av teknisk arbeid eller juridiske avklaringer. Dersom en variabel er tilgjengelig gjennom åpne grensesnitt som geonorge.no vil det typisk kreve lite ekstra arbeid å innhente variabelen, og enkelhet vil dermed vurderes som Høy. Dersom variabelen kan beregnes basert på andre kilder, eller ved å kombinere datakilder, vurderes den som Medium. Hvis det forventes å være svært ressurskrevende å innhente eller beregne variabelen rangeres den som Lav.

Aksept

Dersom skattytere ikke kjenner seg igjen i de underliggende prisdriverne bak modellen vil verdifastsettelsen kunne oppfattes som tilfeldig. Dette er et viktig kriterie som både legger føringer for variablene som inkluderes, og i hvilken grad de er korrekt utfylt. Boligens bruksareal og alder er eksempler på anerkjente prisdrivere som vurderes å ha høy grad av aksept blant boligeiere. Nærhet til skibakker vil være en anerkjent prisdriver for fritidsboliger i populære områder som Hemsedal og Trysil, mens boligeiere i nærheten av en skibakke med lite aktivitet vil kunne oppleve dette som lite relevant for verdifastsettelsen. Disse vil derfor vurderes som Medium. Hvis variabelen oppfattes som irrelevant for verdsettelse av fritidsboliger vurderes den som Lav.

Valgte variabler

Data for variablene som tilfredsstiller kriteriene til tilgjengelighet, enkelhet og aksept er innhentet og inkludert i modellen, mens øvrige variabler er forkastet. Forkastede variabler er beskrevet i Appendiks C.

Strukturelle variabler

Bruksareal
Registrert bruksareal i matrikkel

Byggeår
Basert på byggesak-statuser i matrikkel

Eiendomstype
Registrert eiendoms-type i matrikkel

Andre bygg
Areal for andre bygg (garasje, annekser, etc)

Tomteareal
Registrert tomte-areal i matrikkel

Tomtehelning
Beregnet helning og orientering på tomt

Vann og avløp
Registrert avløps- og vannforsyningskode

Energi og varme
Registrert energi- og oppvarmingskode

Lokasjonsvariabler

Posisjon
- Breddegrad
- Lengdegrad

Adkomst
- Distanse og veitype
- Høydeforskjell

Høyde
- Antall høydemeter på fritidsboligens lokasjon

Naboer
- Antall fritidsboliger innenfor radius

Sjø
- Distanse til sjø
- Strandlinje på tomt

Innsjø
- Distanse til innsjø
- Innsjølinje på tomt

Skibakker
- Distanse til skibakke
- Areal innen radius

Turruter
- Distanse til turrute
- Turlengde innen radius

Innbyggere
- Antall innbyggere innenfor radius

Solforhold
- Antall soltimer
- Snittvinkel horisont

Utsikt
- Utsikt innenfor radius
- Utsikt sjø/innsjøareal

Admin. enhet
- Kommune
- Fylke

Valgte strukturelle variabler og lokasjonsvariabler for verdsettelsesmodell.

Strukturelle variabler

De strukturelle variablene tar utgangspunkt i registrerte og avledede opplysninger fra matrikkelen.

Bruksareal

Det benyttes flere definisjoner for arealberegninger av boliger innenfor offentlig og privat virksomhet. Boligens bruksareal er et mål på boligens innvendige størrelse. Prosjektet har benyttet matrikkelens definisjon av bruksareal som inkluderer innvendig areal innenfor omsluttende vegger. Definisjonen tar utgangspunkt i NS 3940.

Byggeår

Boligens byggeår gir en indikasjon på fritidsboligens standard og fasiliteter. Byggeår registreres ikke som et eget felt i matrikkelen og må derfor utledes fra datoene i byggesakhistorikken til boligene. For bygninger som har fått tilbygg benyttes byggesakhistorikken fra hovedbygget, altså det første bygget som ble oppført.

Eiendoms-type

Skatteetatens avgrensning av fritidsboliger omfatter i all hovedsak de tre eiendomstypene grunneiendom, festegrunn og eierseksjon. Majoriteten av fritidsboliger er grunneiendommer, som innebærer at boligens eier også eier grunnen bygget står på. Dette er ikke tilfelle for festetomter og eierseksjon og kan påvirke verdien av boligen.

Andre bygg

Over 45 prosent av fritidsboligene har andre registrerte bygg på eiendommen enn fritids- eller boligbygg. Dette inkluderer blant annet anneks, garasje, landbruksbygninger, koier, naust, og seterhus. Ettersom det kan være noe variasjon i bygningsklassifiseringen mellom kommuner, benyttes bygningenes summerte bruksareal.

Tomteareal

Tomtearealet til fritidsboligen defineres av de registrerte grensene i matrikkelen. Arealet beregnes basert på den registrerte geometrien av fritidsboligens eiendomsteiger. For festetomter finnes det ulike former for festeavtaler som regulerer festerens rettigheter på tomten. I eierseksjoner eies typisk tomten i sameie med andre seksjonseiere.

Tomte-helning

Bruksverdien av fritidsboligens tomt kan påvirkes av tomtens helning. Stor helning på tomten kan være positivt for boligens utsikt. Samtidig kan et stort tomteareal redusere bruksverdien av tomten dersom den preges av bratte skråninger som kan være vanskelig å utnytte.

Vann og avløp

Nyere fritidsboliger bygges ofte med innlagt vann og kloakk, og fasilitetene trekkes ofte frem som egenskaper som øker etterspørsel. Matrikkelen skiller mellom ulike former for vannforsyning basert på om boligen er tilknyttet offentlig vannverk, privat vannverk eller har privat innlagt vann. Det skiller også tilsvarende mellom offentlig og privat kloakk.

Energi og varme

Innlagt strøm og varme kan ha stor betydning for boligens bruksverdi. Matrikkelen skiller mellom ulike typer energikilder der elektrisitet, biobrensel og andre energikilder er de mest brukte registreringskodene. For varme skilles det mellom elektrisitet, sentralvarme og annen oppvarming.

Lokasjonsvariabler

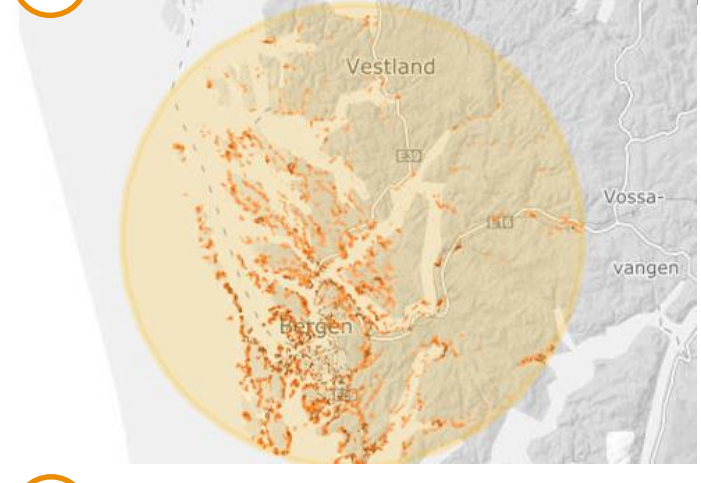
Lokasjonsvariablene beregnes med utgangspunkt i koordinatene til fritidsboligens hovedbygg eller tomtens avgrensning. En komplett liste med lokasjonsvariablene som inngår i modellen er gjengitt i appendiks.

Distanse til sjø

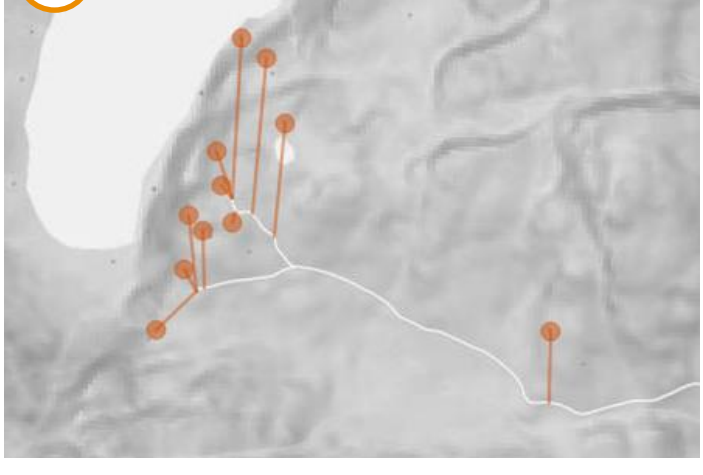
Strandlinje

Distanse til innsjø

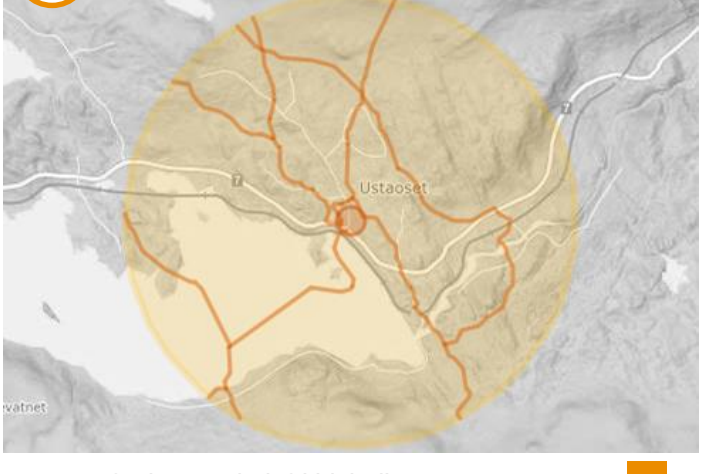
Innbyggere innenfor 50 kilometer



Avstand til privat vei



Turruter innenfor 3 kilometer

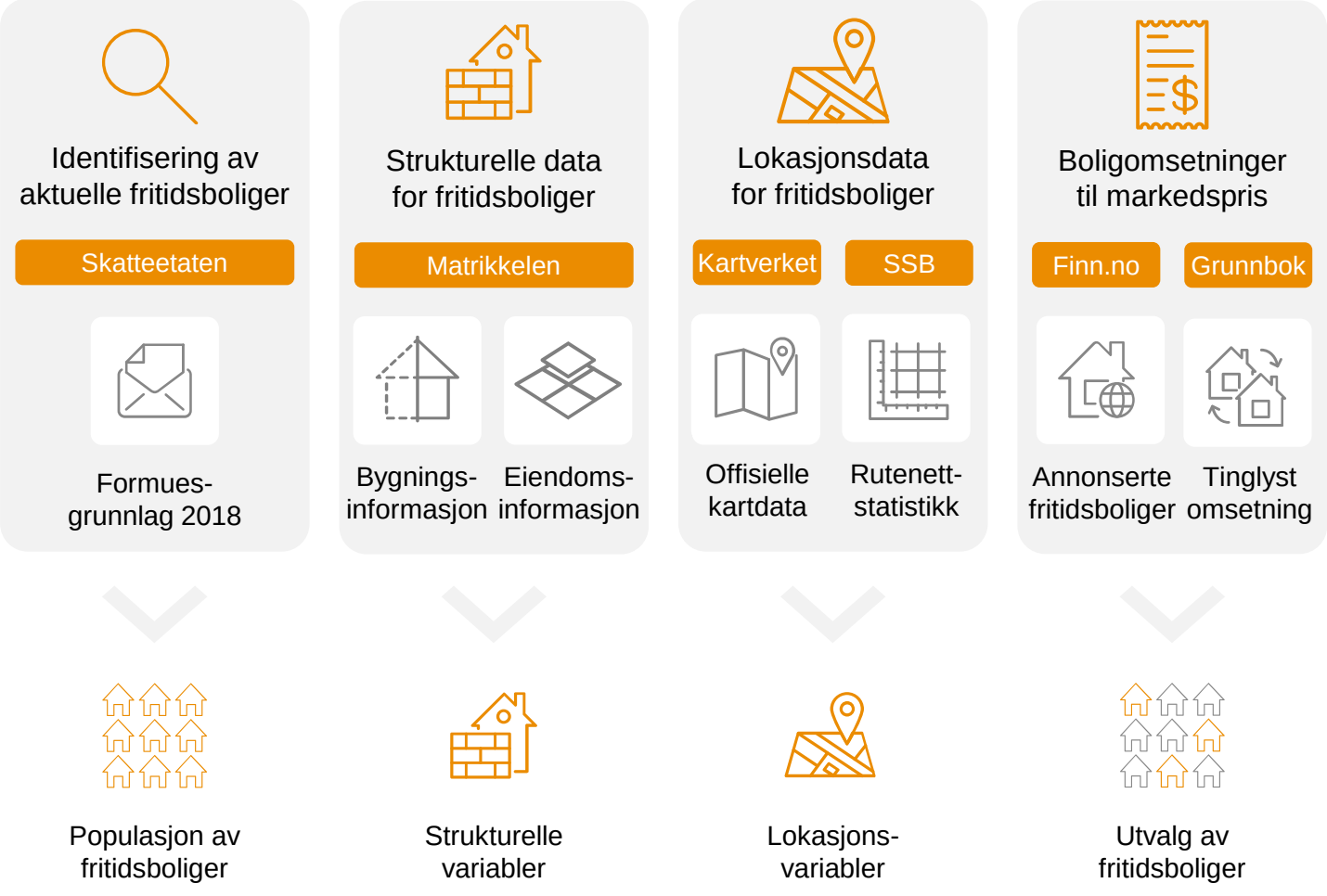


Eksempler på lokasjonsvariabler som beregnes for hver enkelt fritidsbolig.

Datakilder

Prosjektet har benyttet Skatteetatens formuesgrunnlag fra 2018 for å identifisere fritidsboliger som skal verdsettes. Matrikkelen er brukt som kilde for strukturelle data som bygning- og eiendomsopplysninger. Kartdata fra Kartverket er blant annet benyttet for å identifisere innsjøer, skibakker, strandsoner, veier og terrengdata. Felles Kartdatabase (FKB) er benyttet for å estimere bruksareal basert på geometrien av bygningenes fotavtrykk der bruksareal ikke er tilgjengelig i matrikkelen.

Annonserte boligannonser på markedsplassen FINN.no er benyttet for å identifisere fritidsboliger som er solgt til markedspris. Boligannonsene er koblet til tinglyste omsetninger for å verifisere riktige salgspriser.



Datakilder som er benyttet for ulike formål i prosjektet

I tillegg til datakildene over ble det vurdert å innhente data for blant annet jakt- og fiskemuligheter, reguleringsplaner og andre tinglyste rettigheter som båtplass, parkeringsplass og veirett. Enkelte av disse ble forkastet som følge av lav dekningsgrad for fritidsboliger, mens andre viste seg å være krevende å innhente og klassifisere for alle boliger.

Datakilder

De fleste datakildene som er brukt i modellen må oppdateres årlig. Unntaket er enkelte av de geografiske datasettene som gjennomgår små endringer fra år til år. Dette omfatter eksempelvis den norske kystlinjen og innsjøer. De fleste datakildene er tilgjengelige gjennom åpne grensesnitt (API-er) som geonorge.no. Prosjektet har automatisert all innhenting og bearbeiding av data slik at det skal være enkelt å oppdatere modellen.

Identifisering av aktuelle fritidsboliger	Kilde	Metode	Oppdatering
Fritidsboliger som inngår i formuesgrunnlag	Skatteetaten	Uttrekk	Årlig

Omsetninger av aktuelle fritidsboliger	Kilde	Metode	Oppdatering
Registrerte omsetninger	Grunnboken	API	Årlig
Boliger annonsert gjennom FINN.no	FINN.no	Uttrekk	Årlig

Strukturelle data	Kilde	Metode	Oppdatering
Registrerte eiendomsopplysninger	Matrikkelen	API	Årlig

Geografiske data	Kilde	Metode	Oppdatering
Sjø	Kartverket (N50)	API	Ved behov
Innsjøer	Kartverket (N50)	API	Ved behov
Skibakker	Kartverket (N50)	API	Årlig
Veier	Kartverket (N50)	API	Årlig
Tur- og friluftsruter	Kartverket	API	Årlig
Digital terrengmodell	Kartverket (Hoydedata.no)	API	Ved behov
Bebygd areal for bygning	Felles kartdatabase (FKB)	API	Årlig
Befolkningsstatistikk rutenett	SSB	API	Årlig
Bygningsstatistikk rutenett	SSB	API	Årlig
Tur- og friluftsruter	Kartverket	API	Årlig

Datakilder og benyttet metode for å innhente data.

5. Innledende analyse

Innledende analyse

Utviklingen av en statistisk modell starter gjerne med en innledende analyse. Den innledende analysen er en metodisk gjennomgang av datagrunnlaget der formålet er å avdekke eventuelle mangler og underliggende mønstre. Analysen skal blant annet sikre at det ikke er systematiske skjevheter mellom datautvalget som benyttes for trening av modellen og populasjonen.

Dersom modellen skal oppnå presise estimater av markedsverdier er det avgjørende at dataene reflekterer boligenes reelle egenskaper så godt som mulig. Manglende data og feilregistreringer forringer modellens kvalitet og kan ofte være tidkrevende og vanskelig å identifisere.

En analyse av de aktuelle variablenes distribusjon av verdier og sammenheng med markedsverdi gir gjerne en indikasjon på hvordan variablene bør modelleres. Dette er spesielt viktig for lineære modeller siden sammenhengen mellom boligenes strukturelle egenskaper og markedspris ofte er basert på ikke-lineære forhold. Eksempelvis kan prisøkningen for et tilbygg på 10 kvadratmeter være stor for små boliger, mens effekten av et tilsvarende tilbygg for en stor bolig kan gi et mindre utslag for markedsverdien. Resultatene fra den innledende analysen danner grunnlag for modelleringsvalgene som er beskrevet i senere kapitler og inneholder følgende tre steg:

1. Dekningsgrad

Er datautvalget som benyttes til å trene modellen representativt for populasjonen av alle fritidsboliger?

2. Datakvalitet

Er datagrunnlaget beskrivende for populasjonens faktiske egenskaper?

3. Analyse av variabler

Er det sammenheng mellom variablene og markedspris? Hvordan ser distribusjonen ut?

Den innledende analysen inneholder en vurdering av (1) dekningsgrad av omsetninger, (2) datakvalitet for variabler og (3) en analyse av de aktuelle variablenes fordeling og sammenheng med markedspris.

Oppsummering

Ettersom flere fritidsboliger omsettes utenfor fritt salg (bl.a. arv og gaver), men registreres som salg til markedsverdi, er det vanskelig å foreta en sammenligning av hvorvidt datautvalget av boliger er representativt for populasjonen av omsatte fritidsboliger til markedsverdi. Utvalget er basert på annonserte boliger på FINN.no og utgjør 44 prosent av sammenlignbare omsetninger i perioden 2010 til 2020. Dekningsgraden er lavest for boliger med salgspris under 500 tusen kroner hvor utvalget utgjør om lag 16 prosent av sammenlignbare omsetninger. Utvalget har en noe høyere andel av boliger med utfylt bruksareal og byggesaksstatus enn populasjonen. Dette kan indikere at eldre boliger omsettes mindre hyppig enn nyere boliger.

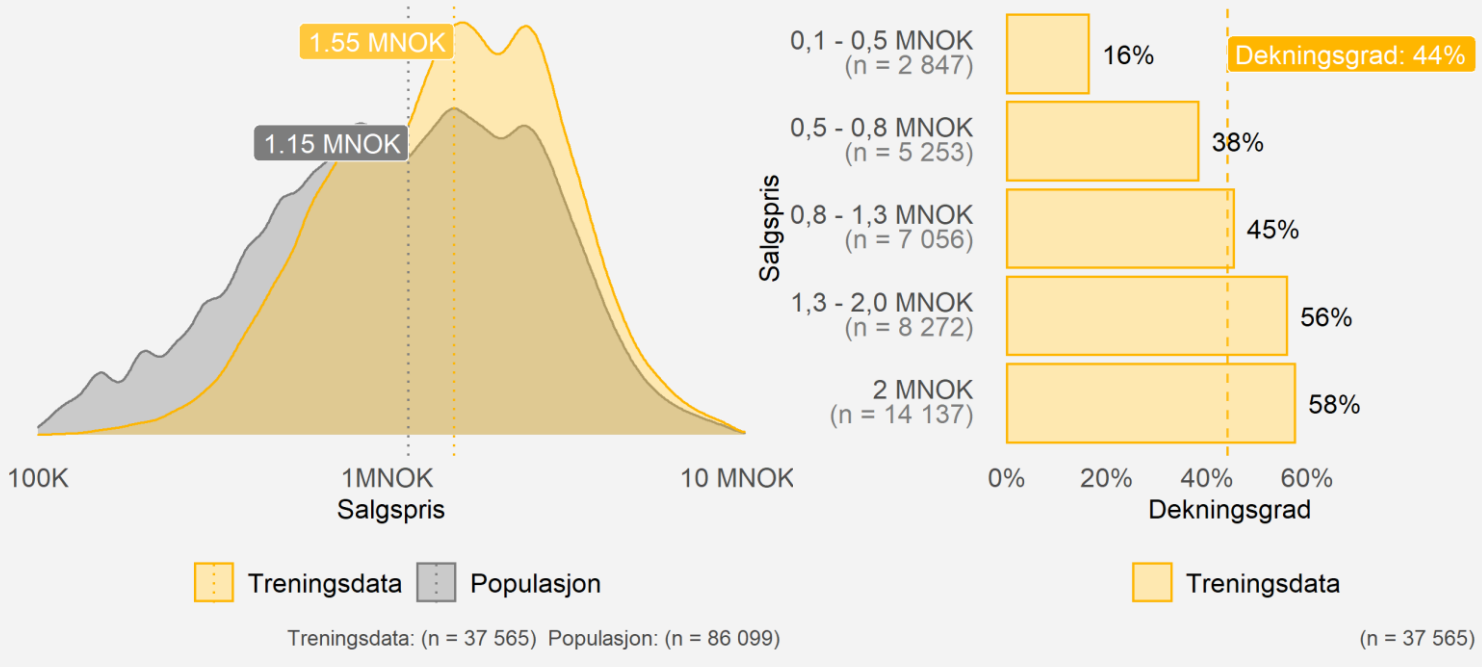
Gjennom den innledende analysen har det blitt avdekket utfordringer med datakvaliteten i matrikkelen og grunnboken. Manglene er spesielt omfattende i bygningsdelen, hvor det i dag mangler bruksareal for 18 prosent av fritidsboligene og byggesaksstatus for 54 prosent av fritidsboligene. For om lag 19 prosent av boligene finnes det ingen registrerte opplysninger om størrelse og alder.

Lokasjonsvariablene er beregnet med utgangspunkt i boligenes koordinater og offentlig tilgjengelige kartdata. Disse datakildene holder gjennomgående høy kvalitet, og den innledende analysen fokuserer derfor på de strukturelle variablene med størst mangler som bruksareal og byggeår.

Dekningsgrad for boligsalg

Verdsettelsesmodellen baseres på historiske omsetninger av fritidsboliger annonsert på FINN.no i perioden 2010 til 2020. Utvalget omfatter 37 565 av 399 787 fritidsboliger[1] og utgjør dermed 9,4 prosent av populasjonen. Dersom modellen skal oppnå presise verdsettelser av norske fritidsboliger bør det ikke være store systematiske forskjeller mellom utvalget og populasjonen av alle fritidsboliger. Systematiske forskjeller kan eksempelvis være at nyere fritidsboliger omsettes hyppigere enn eldre fritidsboliger.

Erfaringene fra SSBs modell for primærboliger er at svært billige og svært dyre boliger i mindre grad omsettes gjennom FINN.no. For billige fritidsboliger kan dette skyldes at honoraret for annonsering er uforholdsmessig høyt i forhold til forventet salgspris. Dyre fritidsboliger omsettes utenfor det åpne markedet av andre årsaker, som diskresjon og eksklusivitet. Ved å sammenligne annonserte boliger på FINN.no mot tinglyste omsetninger er det mulig å belyse hvor stor andel omsetningene på FINN.no utgjør av det totale fritidsboligmarkedet. Dette omtales gjerne som dekningsgrad. Resultatene viser at de annonserte boligene på FINN.no utgjør om lag 44 prosent av frie boligsalg i perioden 2010 til 2020. Dekningsgraden kan blant annet tilskrives at mange fritids-boliger mangler utfylt matrikkelenhet i boligannonsene. Fritidsboliger annonsert på FINN.no har en median salgspris på 1,55 MNOK sammenlignet med 1,15 MNOK for tinglyste omsetninger over 100 tusen kroner.



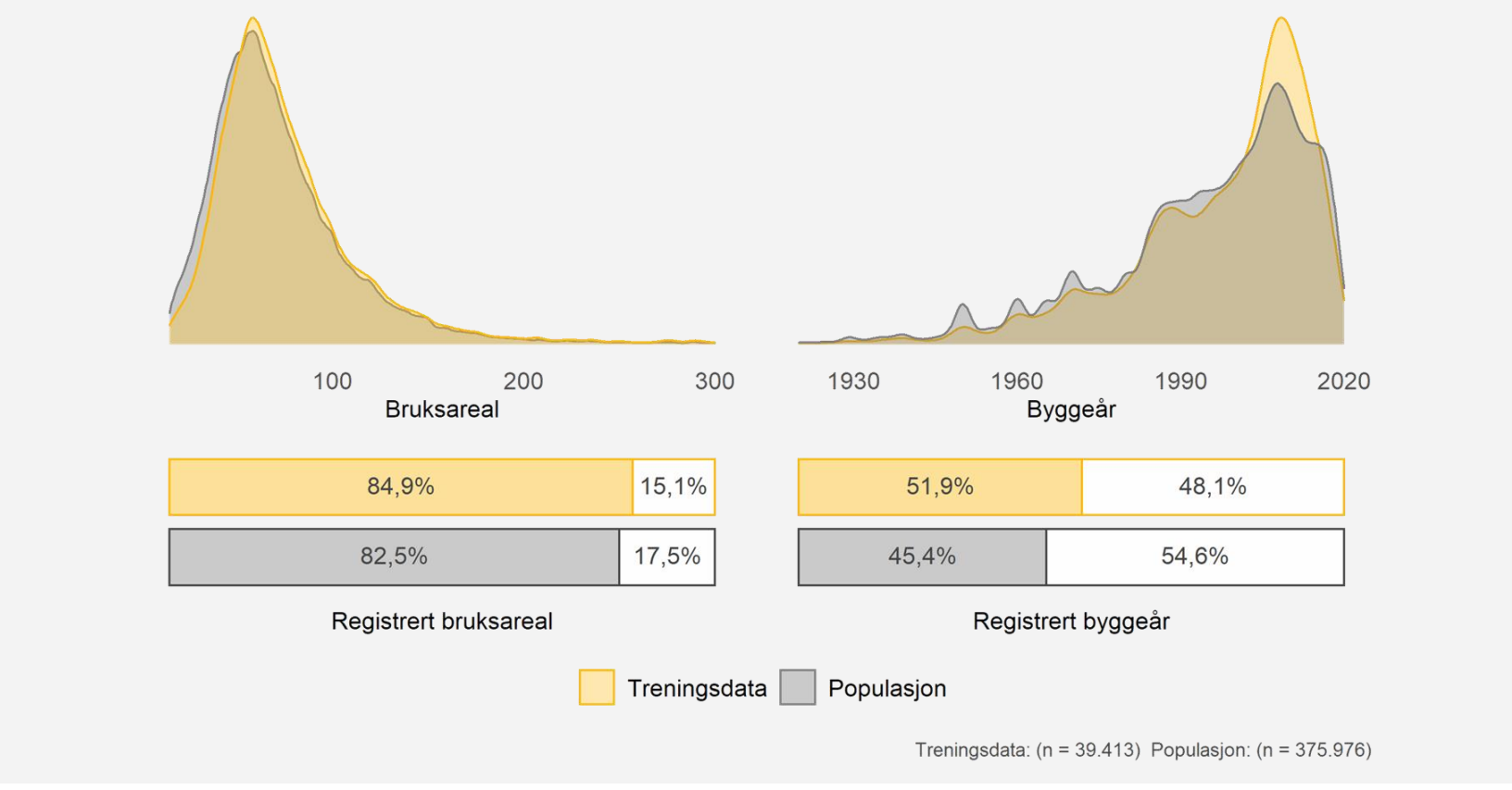
Fordeling av salgspris for fritidsboliger annonsert på FINN.no (treningdata) og omsetningstall for alle boliger som skattelegges som fritidsboliger (populasjon) i perioden 2010 - 2020. Populasjonen inkluderer fritidsboliger med omsetningstype "Fritt salg" og inkluderer omsetninger av fulle andeler for unike boliger med salgspris mellom 100 tusen kroner og 10 millioner kroner.

Dekningsgraden er størst for dyre boliger, og utvalget inkluderer over halvparten av boligene med salgspris over 1,2 millioner kroner. Dekningsgraden er lavest for boliger med salgspris under 500 tusen kroner, hvor utvalget utgjør 16 prosent av alle omsetninger. En årsak til dette kan være at boligsalg med interessefellesskap mellom selger og kjøper ofte omsettes til lave salgspriser utenfor det åpne markedet. Slike omsetninger vil være riktig å ekskludere fra utvalget av boliger solgt til markedspris. En annen årsak kan være at boligene omsettes gjennom andre kanaler enn FINN.no. At billige fritidsboliger er underrepresentert i treningsgrunnlaget for modellen kan potensielt medføre lavere presisjon i verdsettelsen av denne typen boliger. Omsetninger under 500 tusen kroner utgjorde 16 prosent av populasjonen av tinglyste omsetninger i 2019, og grunnlaget på om lag 2 800 boliger kan likevel være tilstrekkelig for å oppnå akseptable estimater for disse boligene.

1) 399 787 inkluderer alle boliger som ble skattlagt som fritidsboliger i 2018 og omtales som populasjonen

Dekningsgrad for boliger

Dekningsgraden for boligsalg viser at om lag 44 prosent av tinglyste omsetninger er med i utvalget. For å undersøke om populasjonen av boliger har andre karakteristikker enn boligene som omsettes kan boligenes strukturelle egenskaper sammenlignes. Gjennom å sammenligne bruksareal og byggeår for de registrerte fritidsbygningene som boligene disponerer er det mulig å belyse forskjeller i boligenes bygningsmasse.



Registreringsgrad og fordeling av bruksareal og bruksareal for fritidsbyggene i utvalget og populasjonen av fritidsboliger. En bolig kan ha 0 eller flere fritidsbygg og tallene er derfor ikke direkte sammenlignbare.med utvalg og populasjon av boliger.

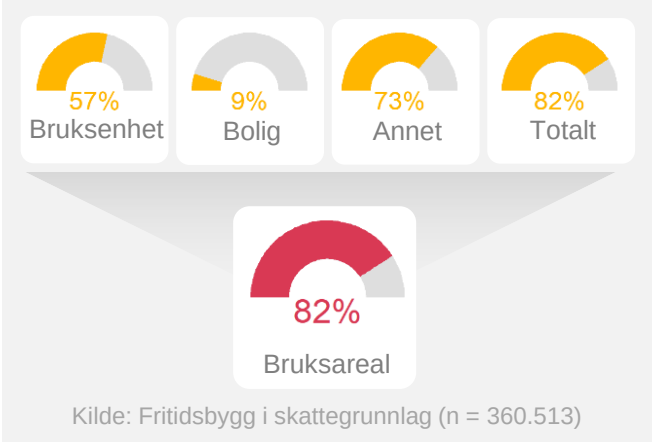
Det ser ut til å være god overensstemmelse mellom registrert bruksareal i utvalget og populasjonen. Andelen av fritidsbygg som mangler bruksareal er 2,4 prosentpoeng lavere for utvalget (15 prosent) enn populasjonen (18 prosent). Det er en liten overvekt av nye fritidsbygg bygd etter 1990 i utvalget. Andelen av bygninger som mangler registrert byggeår er 6,5 prosentpoeng lavere for utvalget (48 prosent) enn populasjonen (55 prosent) av fritidsbygninger.

Registrering av bygningsopplysninger som bruksareal ble obligatorisk for nybygg i 1983 og det forventes at registreringsgraden er høyere for nye fritidsbygg. Resultatene kan indikere at nyere fritidsboliger omsettes noe hyppigere gjennom FINN.no enn eldre fritidsboliger. Dette kan blant annet skyldes at eldre boliger går i arv gjennom generasjoner, eller at tilstandsgraden for enkelte eldre boliger medfører at annonseringshonoraret oppleves uforholdsmessig høyt sammenliknet med forventet salgspris.

Avvikene i strukturelle boligegenskaper mellom utvalget og populasjonen fremstår ikke som uforholdsmessig store. Videre forventes det at kvalitetsheving av bygningsopplysninger i matrikkelen vil øke registreringsgraden av byggeår over tid. Dette vil bidra til å utjevne forskjellene i registreringsgrad mellom eldre og nyere boliger.

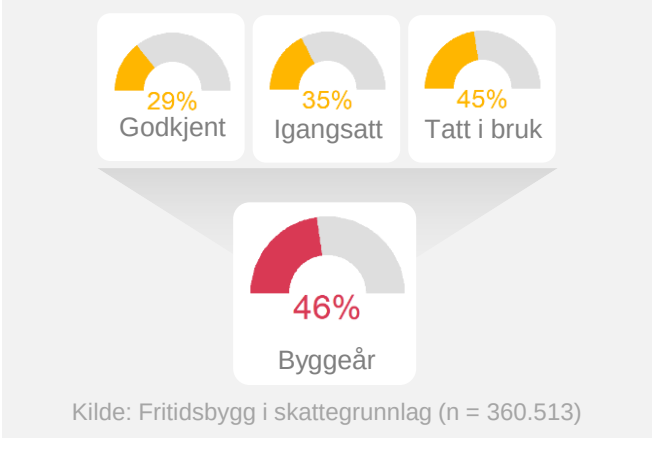
Registreringsgrad

Det er kommunene som er ansvarlig for å føre eiendomsopplysninger i matrikkelen. Det kan dermed være betydelige variasjoner i graden av registrering mellom kommuner. Føring av bygningsopplysninger på nasjonalt plan ble innført i 1983, og bygninger registrert før 1983 inneholder normalt ikke bygningsinformasjon med mindre kommunene har komplettert opplysningene i ettertid. Enkelte kommuner som har innført eiendomsskatt har etterregistrert informasjon for eldre bygg, men omfanget er større for primærboliger enn fritidsboliger. Det pågår for tiden et langsiktig arbeid med kvalitetsheving av matrikkelen, og det ventes at grunnlaget vil bedres over tid.



Bruksareal

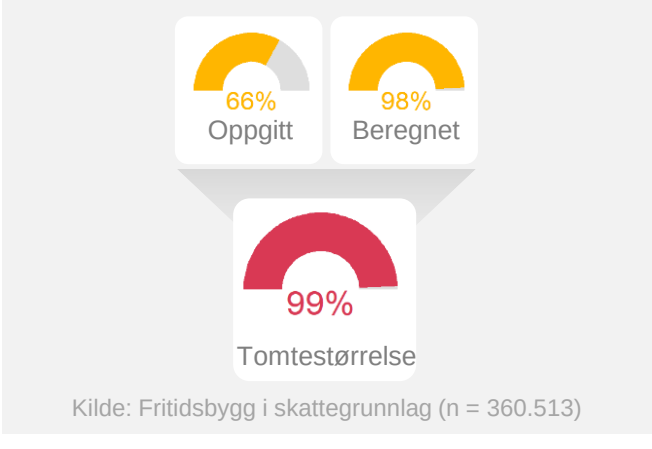
Om lag 82 prosent av fritidsbyggene i populasjonen har registrert bruksareal i matrikkelen. Majoriteten av disse er utfylt på bygningsnivå, mens det er noe lavere registreringsgrad for bruksenhetene. Om lag 16 tusen av fritidsboligene er eierseksjoner hvor bygget er tilknyttet flere seksjoner. For disse boligene må bruksenhetsareal benyttes, og eierseksjonene vil dermed ha flere manglende verdier enn grunneiendommer. Ifølge matrikkelførings-instruksen skal ikke fritidsbygg registreres med bruksareal til boligformål, men om lag ni prosent av byggene har likevel et utfylt bruksareal til bolig.



Byggeår

46 prosent av fritidsbyggene har utfylt én av de tre byggesakstatusene “Igangsatt”, “Godkjent” og “Tatt i bruk”. Dette innebærer at byggeåret for over halvparten av fritidsbyggene er ukjent.

Graden av utfylling for eldre bygg er betydelig lavere enn for nye bygninger. For annonserte fritidsboliger på FINN.no med byggeår før 1980 mangler registrert status på byggesak i matrikkelen for over 80 prosent av boligene. Dette indikerer at en stor andel av byggene som mangler byggesaksstatus er eldre bygg.

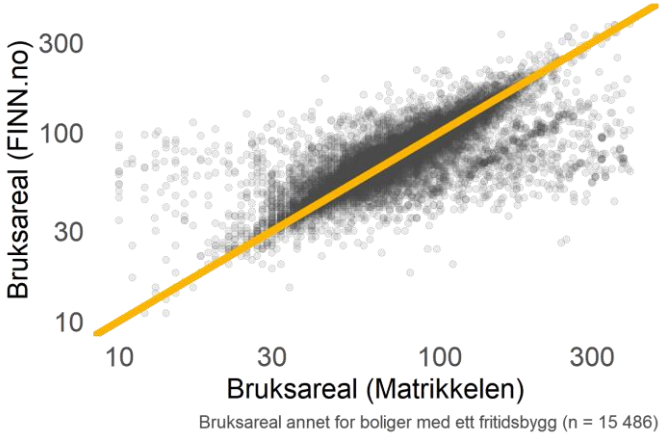


Tomtestørrelse

Matrikkelen skiller ikke mellom manglende tomtestørrelser og tomter som er oppført med 0 kvadratmeter. Fritidsboliger som er festegrunner eller eierseksjoner kan ofte ha et reelt registrert tomteareal på 0 kvadratmeter. Det registreres to ulike typer tomteareal, “Beregnet areal” og “Oppgitt areal”. Det beregnede arealet er basert på registrerte koordinater og har høyest grad av utfylling. Der beregnet arealet ikke er tilgjengelig kan oppgitt areal benyttes. Oppgitt areal kan være hentet fra forretningsdokumenter som f.eks. skylddelinger.

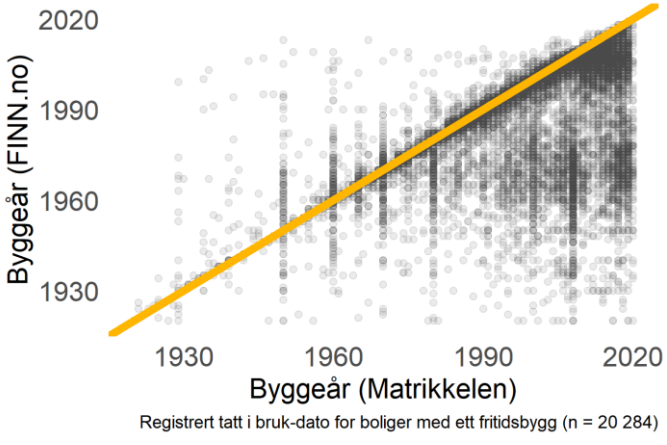
Kvalitet i registrering

Matrikkelen er den autoritative kilden til norsk eiendomsinformasjon, og det er dermed vanskelig å benytte alternative kilder for å ettergå kvaliteten på registrerte opplysninger for alle fritidsboliger. Prosjektet har hatt tilgang til annonserte fritidsboliger på FINN.no, der selger blant annet har fylt ut bruksareal, byggeår og tomtestørrelse. Siden opplysningene benyttes i forbindelse med salg av bolig, antas det å utgjøre det beste mulige grunnlag for å belyse hvorvidt det utfylte innholdet i matrikkelen reflekterer de faktiske egenskapene ved boligene.



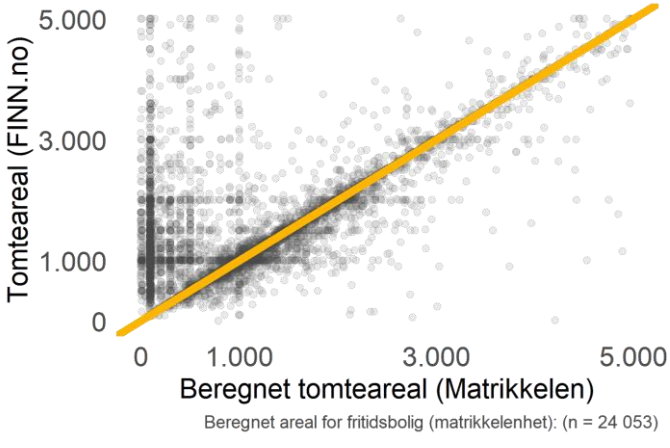
Bruksareal

For fritidsbygg i matrikkelen er det et gjennomsnittlig prosentvis avvik mot annonsert bruksareal på FINN.no på 22 prosent. Bygninger med lave registrerte bruksarealer i matrikkelen (under 20 kvadratmeter) ser ut til å potensielt inneholde feilregistreringer, da disse ofte har et høyere annonsert bruksareal på FINN.no. En del av bygningene med store registrerte bruksarealer (over 100 kvadratmeter) ser ut til å ha et lavere annonsert bruksareal på FINN.no. Dette kan blant annet skyldes dobbelttelling av bruksarealer for tilbygg dersom disse ikke er sammenslått med hovedbygget i matrikkelen etter ferdigstillelse.



Byggeår

For fritidsbygg i matrikkelen er det et median absoluttavvik mellom registrert byggeår på FINN.no og registrert dato tatt i bruk på ett år. Om lag 10 prosent av fritidsbyggene har et avvik på mer enn 34 år. Byggesaksstatusen “Tatt i bruk” ser ut til å bære preg av en del etterregistreringer siden 1980-tallet og historiske innlagte datoer settes gjerne til runde datoer som 1. januar 1950 eller 1. januar 1960. Når “Tatt i bruk” benyttes for å utlede byggeår forventes det at standarden for bygningene som er etterregistrert overvurderes i modellen da flere av disse bygningene ser ut til å være bygd mange år tidligere.

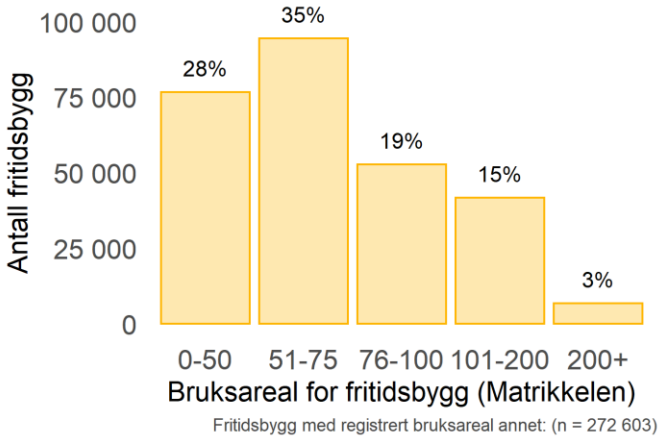


Tomtestørrelse

Det ser jevnt over ut til å være god overensstemmelse mellom beregnet tomteareal for grunneiendommer i matrikkelen og annonsert tomteareal på FINN.no. Unntaket er fritidsboliger som er registrert med lav tomtestørrelse i matrikkelen. Disse har typiske beregnet areal på 0 eller 10 kvadratmeter. Tomtene med 0 kvadratmeter er ofte eierseksjoner hvor megler har benyttets sameie totale tomt. De er derfor ikke indikasjoner på kvalitetsutfordringer ved opplysningene. Tomten er en fysisk avgrensning av grunnen som gir sterke insentiver til riktig rapportering av grenser, og kvaliteten ansees som tilfredsstillende.

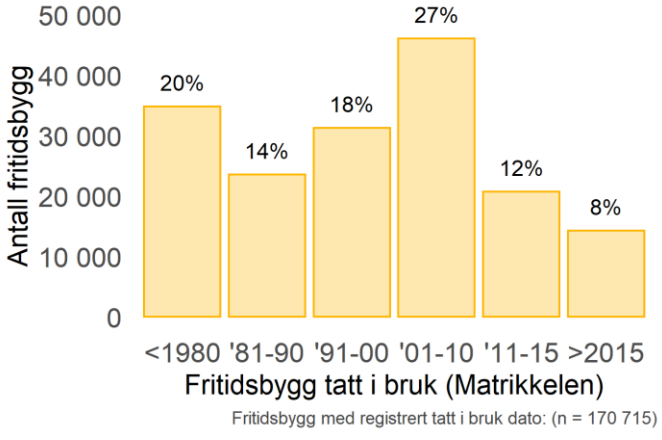
Distribusjon av verdier

Variablenes distribusjon av verdier gir gjerne en indikasjon på hvordan disse bør modelleres. Analysene kan også avdekke ekstremverdier som potensielt kan påvirke modellene. Variablenes distribusjon er spesielt viktig for lineære modeller som er sensitive for ekstreme verdier.



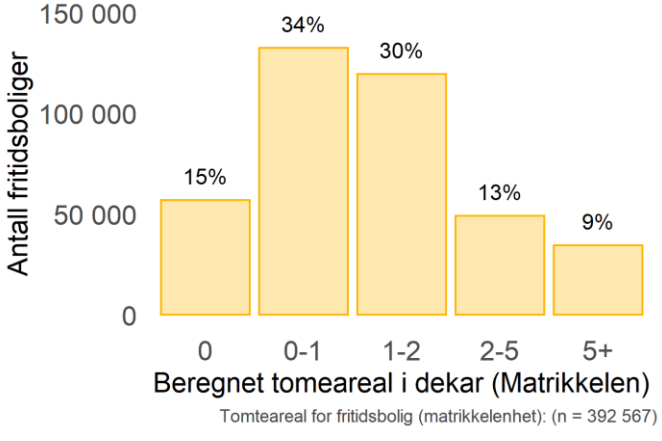
Bruksareal

Gjennomsnittlig registrert bruksareal for et norsk fritidsbygg er 88 kvadratmeter. Tallet påvirkes av enkelte store fritidsbygg og halvparten av boligene har et bruksareal under 65 kvadratmeter. Om lag en tredjedel av fritidsbyggene har registrert bruksareal mellom 51 og 75 kvadratmeter. Om lag 6 800 bygninger har et bruksareal over 200 kvadratmeter. Dette utgjør om lag 3 prosent av populasjonen av fritidsbygg. Om lag 1 000 bygninger har et registrert bruksareal over 1 200 kvadratmeter



Byggeår

Om lag en tredjedel av boligene er registrert tatt i bruk mellom 2001 og 2010 i matrikkelen. Det er spesielt årene 2007 til 2010 som inneholder påfallende mange registreringer. Dette sammenfaller med at begrepet “tatt i bruk” hadde en annen betydning i det gamle GAB-registeret som ble erstattet av matrikkelen i samme periode. Årsaken kan dermed skyldes feilaktige registreringer i forbindelse med overgang til ny kodepraksis. Om lag 500 fritidsbygg er registrert tatt i bruk i år 1 (31.12.0001). Det ser dermed ut til å være varierende kvalitet i registrering og validering av statusen på byggesaken. Det er likevel den beste indikasjonen på byggeår som eksisterer i matrikkelen.

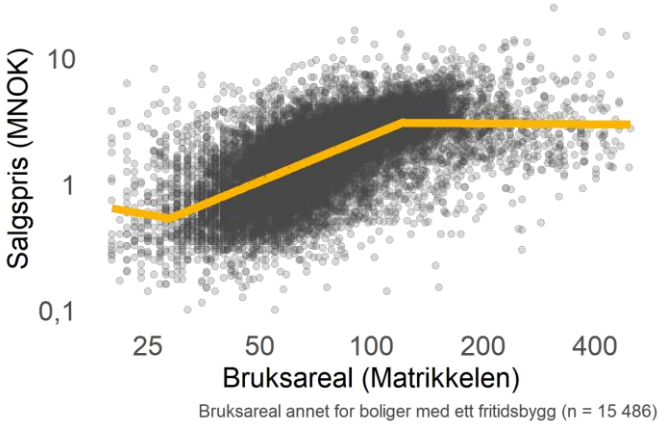


Tomtestørrelse

Om lag 15 prosent av fritidsboligene har ingen beregnet areal. En betydelig andel av boligene er festegrunner og eierseksjoner og forventes dermed heller ikke å ha beregnet areal. Om lag 91 prosent av tomtene er under 5 dekar. Den største tomten er på om lag 2 millioner dekar (2000 kvadratkilometer). Dette er fire ganger så stort som Oslo fylke er i utstrekning. Om lag 3 prosent av tomtene er større enn 5 dekar, og 50 tomter er større enn 100 tusen dekar (100 kvadratkilometer).

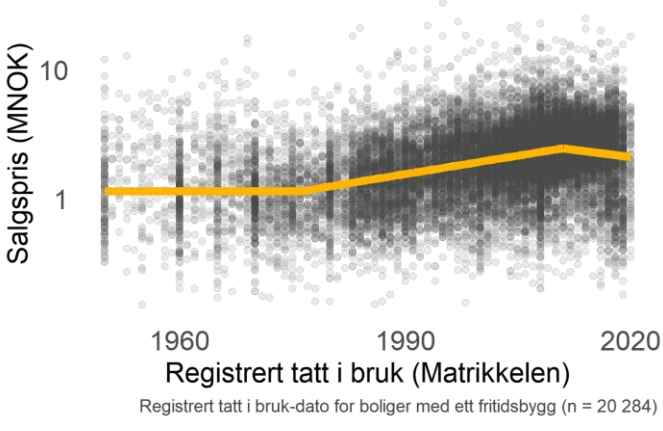
Sammenheng med pris

Fritidsboligenes strukturelle egenskaper er blant de viktigste faktorene som påvirker boligens markedsverdi. Analyser av sammenhengen mellom variablene og salgspris gir en indikasjon på hvordan disse forholdene kan modelleres. Modelleringsvalget er spesielt viktig for modeller som antar en lineære sammenheng mellom variablene og salgspris. Dette kan innebære at det er nødvendig å gjøre transformasjoner av variablene i den lineære modellen. Det er eksempelvis vanlig å benytte den naturlige logaritmen av salgspriser og arealer.



Bruksareal

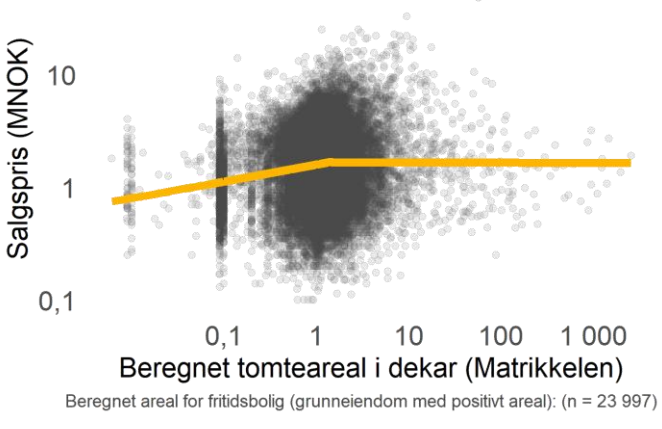
Det er en tydelig lineær sammenheng mellom logaritmen av salgspris og bruksareal mellom 30 kvadratmeter og 100 kvadratmeter. En økning i areal over 100 kvadratmeter ser i liten grad ut til å påvirke markedsverdien, men dette er en forenklet analyse som ikke kontrollerer for boligens beliggenhet. Markedsverdien ser ut til å være fallende mellom 15 kvadratmeter og 25 kvadratmeter som kan understøtte tidligere funn om at flere fritidsboliger har et feilaktig lavt registrert bruksareal.



Byggeår

Det er en lineær sammenheng mellom logaritmen av salgspris og boliger bygd mellom 1970 og 2010. Dette er som forventet med tanke på utviklingen i standarder for nybygg. Verdien ser ut til å avta for de aller nyeste boligene. Dette er overraskende, og det må igjen påpekes at dette er en forenklet analyse som ikke kontrollerer for markedsutvikling.

Det ser ut til å være liten forskjell om boligen er bygd i 1950 eller 1970. Dette kan indikere at utviklingen i bygnings-standard først og fremst har skjedd de siste 50 årene.



Tomtestørrelse

Det ser ut til å være en lineær sammenheng mellom logaritmen av tomteareal og logaritmen av salgspris for fritidsboliger med tomteareal under 1 dekar (1 000 kvadratmeter). En økning i tomteareal over 1 dekar ser i liten grad ut til å påvirke markedsverdien, men dette vil være sterkt avhengig av tomtens beliggenhet. Festetomter og eierseksjoner er utelatt fra figuren som bare inkluderer grunneiendommer med positivt registrert areal.

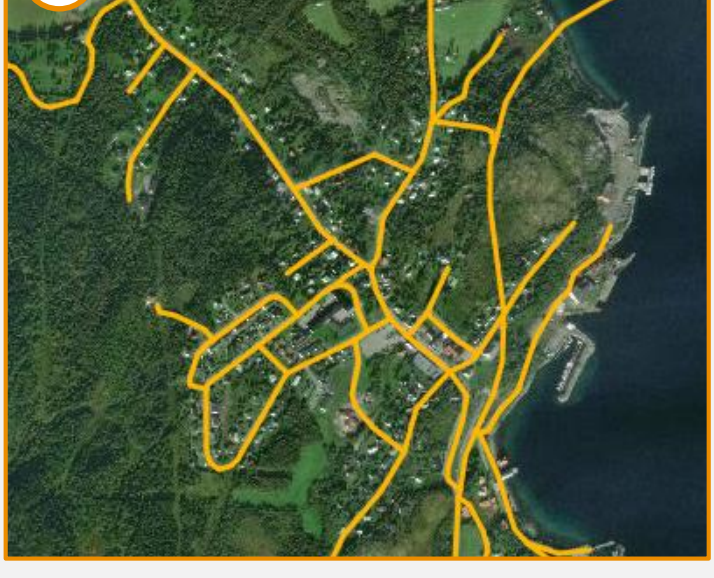
Kvalitet på lokasjonsdata

Lokasjonsvariablene er beregnet basert på offisielle kartdata og koordinatene til fritidsboligenes hovedbygg. Kartdataene som er benyttet for sjøgrenser, innsjøer, skibakker og veier stammer fra Kartverkets N50-serie som første gang ble utgitt i 1997. Siden kartdataene er basert på kvalitetssikrede data som har eksistert over flere år har det ikke vært nødvendig å ettergå kvaliteten på samme måte som det er gjort for de strukturelle variablene. Ved å overlegge objektene på satellittbilder er det enkelt å verifisere at konturene følger terrenget og at beregnede arealer og distanser er korrekt oppmålt.

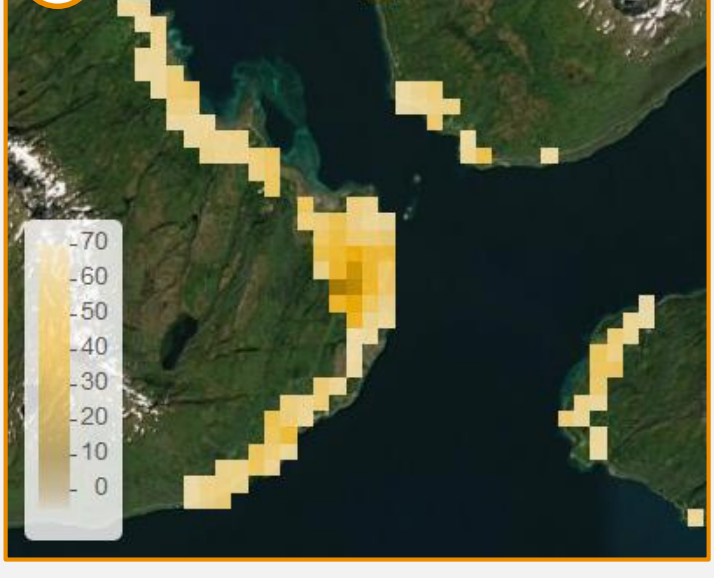
Innsjøer og regulerte innsjøer

Skibakker

Vei



Innbyggere (250 m x 250 m)



Beregningen av lokasjonsvariablene er basert på offisielle kartdata fra Kartverket og SSB. Korrekt lokasjon og utstrekning av kystlinje, innsjøer skibakker, veier og innbyggertall er verifisert ved bruk av satellittbilder.

Strukturelle variabler

Tabellen viser utfyllingsgrad og fordeling av alle numeriske verdier for de strukturelle variablene som inngår i modellen. Kolonnen “manglende” angir hvor mange fritidsboliger i utvalget som mangler verdi for variabelen. I tillegg kommer de kategoriske variablene eiendomstype (0% manglende), avløpskode (73% manglende), vannforsyningskode (60% manglende), oppvarmingskode (70% manglende) og energikildekode (30% manglende). For noen av variablene er laveste og høyeste verdi trolig feilutfyllinger i datakilden. Alle variablene er beskrevet i Appendiks B.

	Manglende	Laveste	Median	Høyeste	Histogram
Omsetninger					
Salgspris	0	54.000	1.600.000	48.000.000	
Salgsår	0	2010	2016	2020	
Salgsmåned*	0	1	8	12	
Bruksareal*					
Bruksareal (Annet)	9.279 (26%)	1	70	8.104	
Bruksareal (Bolig)	32.554 (90%)	1	69	2 357	
Bruksareal (Totalt)	5.822 (16%)	1	70	8.104	
Bruksareal (Bruksenhet)	14.483 (40%)	1	70	880	
Bruttoareal (Totalt)	34.844 (96%)	1	77	570	
Bebygd Areal (Matrikkel)	26.035 (72%)	0	83	2.673	
Bebygd Areal (FKB)	424 (1%)	3	97	6.040	
Byggeår*					
Status: Tatt i bruk	17.090 (47%)	1799	2003	2019	
Status: Igangsatt	20.263 (56%)	1899	2003	2020	
Status: Godkjent	23.365 (64%)	707	2001	2019	
Andre bygg					
Areal andre bygg	0	0	0	1.165	
Tomteareal*					
Oppgitt areal	0	0	285	299.000.000	
Beregnet areal	0	0	958	3.742.053	
Tomtehelning					
Helning	1.862 (5%)	0	7	37	
Orientering	1.862 (5%)	1	156	357	

*) For bruksareal, byggeår og tomteareal vises alle felter som benyttes i matrikkelen.
Senere kapitler belyser hvordan disse inngår i modellen som én variabel.

Lokasjonsvariabler

Tabellen viser utfyllingsgrad og fordeling av verdier for lokasjonsvariablene som inngår i modellen. Alle variablene er beskrevet i Appendiks B.

	Manglende	Laveste	Median	Høyeste	Histogram
Posisjon					
Breddegrad	0	58	60	78	
Lengdegrad	0	5	10	31	
Adkomst					
Avstand til vei (meter)*	0	0	376	20.320	
Avstand til privat vei (meter)*	0	0	35	11.851	
Høyde					
Høydemeter	0	0	408	1.251	
Naboer					
Antall fritidsboliger innenfor 1 km	0	0	104	953	
Antall fritidsboliger innenfor 3 km	0	0	202	3.070	
Antall primærboliger innenfor 1 km	0	0	78	2.159	
Sjø					
Avstand til sjø (meter)*	0	0	496	25.000	
Kystlinje (meter)	0	0	0	10.341	
Innsjø					
Avstand til innsjø (meter)*	0	0	892	25.000	
Innsjølinje (meter)	0	0	0	252.463	
Skibakker					
Avstand til skibakke (meter)*	0	0	10.313	25.000	
Areal av skibakke innenfor 10 km (kvm)	0	0	0	1.852.188	
Turruter					
Avstand til turrute (meter)*	0	0	2.824	25.000	
Turruter innenfor 1 km (meter)	0	0	0	23.402	

* For å forenkle beregningene er det satt en øvre grense ved beregning av avstander på 25 kilometer. Altså er alle avstander større enn 25 kilometer satt lik 25 kilometer i stedet for den reelle distansen.

Lokasjonsvariabler

Tabellen viser utfyllingsgrad og fordeling av verdier for lokasjonsvariablene som inngår i modellen. Alle variablene er beskrevet i Appendiks B.

	Manglende	Laveste	Median	Høyeste	Histogram
Antall fastboende					
Antall innbyggere innenfor 1 km	0	0	8	5.508	
Antall innbyggere innenfor 5 km	0	0	338	108.970	
Antall innbyggere innenfor 50 km	0	0	77.736	1.706.891	
Antall innbyggere innenfor 100 km	0	0	402.901	2.327.158	
Solforhold					
Gj. vinkel horisont	0	0	10	84	
Årlige soltimer	0	0	3.290	4.553	
Utsikt					
Synlig areal innenfor 1 km (100 kvm)	153	9	2.994	27.217	
Synlig areal innenfor 5 km (100 kvm)	153	9	30.130	504.718	
Synlig sjø/innsjø 1 km (100 kvm)	153	0	0	20.404	

* For å forenkle beregningene er det satt en øvre grense ved beregning av avstander på 25 kilometer. Altså er alle avstander større enn 25 kilometer satt lik 25 kilometer i stedet for den reelle distansen.

6. Modelleringsvalg

Modelleringsvalg

Kapittel 2 beskrev de tre vurderingskriteriene for modellen; presisjon, stabilitet og forklarbarhet. Kapittel 3 beskrev hvordan prosjektet har identifisert strukturelle variabler og lokasjonsvariabler som forventes å påvirke markedsverdien av norske fritidsboliger. Den innledende analysen i kapittel 4 viste variablenes fordeling av verdier og belyste også utfordringer med kvaliteten i registreringen av strukturelle variabler i matrikkelen.

Generelle modelleringsvalg

Modellens vurderingskriterier og den innledende analysen setter føringer for hvordan de valgte variablene bør modelleres. Analysen viser blant annet at flere av variablene kan representeres på ulike måter. Eksempelvis kan fritidsboligens bruksareal både ta utgangspunkt i bruksenhetens bruksareal og byggets bruksareal der bygget bare inkluderer én bruksenhet. Videre viser analysen at mange av boligene mangler registrering av bruksareal og byggeår. Dette er problemstillinger som må håndteres av begge modellene.

Modellering av strukturelle variabler

Opplysninger om bruksareal, byggeår og tomtestørrelse registreres på flere måter i matrikkelen. Hvordan bør disse inngå i modellen?

Håndtering av manglende data

Mange fritidsboliger mangler registrert bruksareal og byggeår. Hvordan skal manglende verdier håndteres i modellen?

Spesifikke modelleringsvalg

Ettersom lineære og trebaserte modeller har ulik teknisk oppbygging, må det i tillegg gjøres modellvalg som er spesifikke for de to modellene. Dette inkluderer blant annet håndtering av ikke-numeriske variabler som eiendomstype og kommune. Slike variabler omtales ofte som kategoriske variabler, og må gis en numerisk verdi i modellen. Den foretrukne fremgangsmåten vil typisk variere mellom ulike modelltyper.

Videre forutsetter lineære modeller en lineær sammenheng mellom boligens underliggende egenskaper og markedsverdi. I virkeligheten er få sammenhenger fullstendig lineære. Eksempelvis er den relative prisøkningen ved å gå fra 50 til 100 kvadratmeter bruksareal langt større enn en tilsvarende økning fra 200 til 250 kvadratmeter. Lineære modeller er i tillegg sensitive for ekstremverdier og skjevheter i variablenes distribusjon. For å gi gode resultater må disse problemstillingene håndteres gjennom ulike tilpasninger som transformasjoner og diskretisering av variabler. Lineære modeller krever derfor ofte en del forarbeid for å gi gode resultater for komplekse problemer.

Trebaserte modeller er basert på rangering av variablenes verdier og er derfor immune mot skjeve distribusjoner og ekstremverdier. Modellene kan derfor fange opp svært komplekse ikke-lineære sammenhenger uten at det er nødvendig å identifisere passende transformasjoner for variablene. I noen tilfeller kan modellene være så komplekse at modellene fanger opp sammenhenger som bare eksisterer for boligene i treningsdataene. Når de lærte sammenhengene ikke er generaliserbare til populasjonen av fritidsboliger omtales modellen gjerne som overtilpasset. En slik overtilpasning resulterer ofte i en mer ustabil modell. Den enkleste måten å unngå overtilpasning er å redusere modellens kompleksitet. Dette gjøres vanligvis gjennom et søk blant parametrene som regulerer modellens kompleksitet. Et slikt søk omtales som et parametersøk og benyttes ofte for å finne den optimale avveiningen mellom presisjon og stabilitet.

Videre i dette kapittelet beskrives både de generelle modelleringsvalgene, som hvordan strukturelle variabler modelleres og manglende data håndteres, samt spesifikke modellvalg for den lineære og trebaserte modellen.

Modellering av strukturelle variabler

Som vist i kapittel 4 finnes det flere felter i matrikkelen som benyttes for å beskrive boligens bruksareal, byggeår og tomteareal. For fritidsbygg finnes det fire ulike bruksareal med varierende grad av registrering og kvalitet i innhold som kan benyttes. Det er fullt mulig å utvikle en modell som benytter alle feltene for verdsettelse. Dette innebærer at det er opp til modellen å ta stilling til hvilke variabler som skal vektlegges for ulike boliger. Ettersom modellen har all tilgjengelig informasjon kan dette øke treffsikkerheten i verdsettelsene.

Ulempen ved denne tilnærmingen er at variablene ofte er korrelerte. Eksempelvis vil fritidsboliger med én bruksenhet typisk ha samme registrerte areal på bruksenhetsnivå og bygningsnivå. Dette introduserer et problem kjent som multikollinearitet. Når arealet for bygget og bruksenheten er det samme blir det i praksis vilkårlig hvilken variabel som benyttes for å utføre en splitt i en trebasert modell. Dette påvirker typisk ikke modellens presisjon, men gjør det i praksis umulig å tolke viktigheten av de ulike variablene. Problemstillingen omfatter både trebaserte og lineære modeller. I en lineær modell med vesentlig multikollinearitet vil koeffisientene for de ulike variablene kunne være statistisk ugyldige.

Et annet argument for å bare benytte ett av feltene for bruksareal er at modellens form blir enklere å forstå for skattytere. Dersom modellen baseres på fire ulike bruksarealer med varierende grad av registrering, blir det vanskelig å forstå hvordan bruksarealet faktisk påvirker verdsettelsen. Modellen vil derfor være mer forklarbar dersom en egenskap kun representeres av én variabel. Variabler som representerer tilsvarende strukturelle egenskaper slås derfor sammen til én verdi basert på funn fra innledende analyse og SSBs erfaringer. For bruksareal og byggeår benyttes reglene på bygningsnivå siden en fritidsbolig kan bestå av flere bygninger:

Regler for bruksareal:	
1. Eierseksjon med bruksenhetsareal over 15 kvm	Bruksenhet
2. Fritidsbygning med én matrikkelenhet og både bruksareal «bolig» og «annet» over 15 kvm	Totalt
3. Fritidsbygning med én matrikkelenhet og bruksareal «annet» over 15 kvm	Annet
4. Fritidsbygning med én matrikkelenhet og bruksareal «bolig» over 15 kvm	Bolig
5. Annen bygning med én matrikkelenhet og bruksareal «bolig» over 15 kvm	Bolig
6. Annen bygning med én matrikkelenhet og bruksareal «totalt» over 15 kvm	Totalt
Regler for byggeår:	
1. Har «tatt i bruk»-dato	Tatt i bruk
2. Har «igangsatt»-dato	Igangsatt
3. Har «godkjent»-dato	Godkjent
Regler for tomteareal:	
1. Har «beregnet areal» > 0 kvm	Beregnet areal
2. Har «oppgitt areal» > 0 kvm	Oppgitt areal

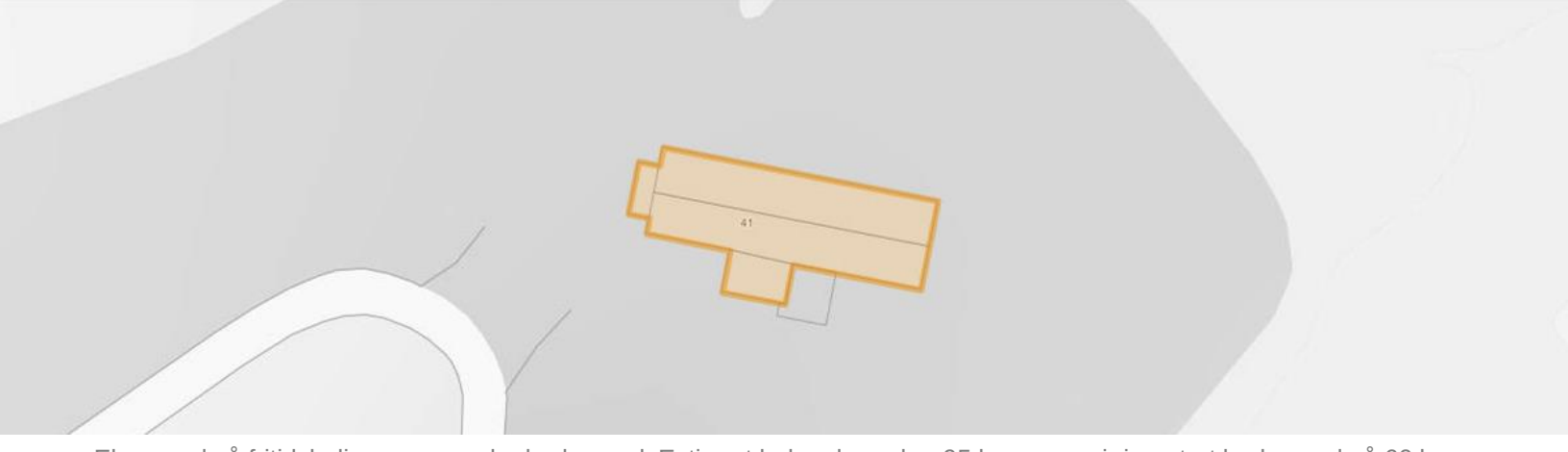
Regler for hvilke felter fra matrikkelen som benyttes for strukturelle variabler.

Håndtering av manglende strukturelle data

Selv om det identifiseres regler for å utnytte maksimalt av innholdet i matrikkelen er det likevel mange fritidsboliger som mangler verdier for strukturelle variabler. Den innledende analysen viste at 18 prosent av fritidsbyggene mangler bruksareal og 53 prosent mangler byggesaksstatus.

Lineære modeller kan ikke håndtere manglende verdier og forutsetter derfor at manglene erstattes med en annen numerisk verdi. En enkel fremgangsmåte er å erstatte manglende verdier med median eller gjennomsnittlig verdi for hver enkelt variabel. Denne typen utfylling av manglende verdier kalles gjerne imputering. Imputering ved bruk av snitt- og medianverdier er godt egnet dersom verdien mangler av tilfeldige årsaker. I mange tilfeller skyldes manglene systematiske årsaker, eksempelvis at kommunene ikke etterregistrerer byggesakstatus for eldre boliger. Manglende byggesakstatus er dermed ofte en systematisk indikasjon på at bygningen er gammel.

En effektiv måte å håndtere systematisk manglende verdier er å behandle manglene som et estimeringsproblem. Målet er dermed er å utvikle en modell som gir en mest mulig presis estimering av hvert manglende datapunkt. Ved å benytte en slik modellbasert imputering er det eksempelvis mulig å lære fra kommunene som har etterregistrert byggesakstatus for eldre boliger. Prosjektet har valgt å benytte en trebasert modell for imputering av manglende verdier. Imputeringsmodellen benytter en kombinasjon av registrerte variabler i matrikkelen og andre variabler. Eksempelvis gir boligens grunnflate fra kartdata en god indikasjon på boligens bruksareal.



Eksempel på fritidsbolig som mangler bruksareal. Estimert bebygd areal er 95 kvm som gir imputert bruksareal på 60 kvm.

Prosjektet har utviklet tre ulike imputeringsmodeller for å imputere bruksareal og byggeår.

Imputeringsmodell	Grunnlag for imputering				
Bruksareal for bruksenheter*	Antall enheter	Antall enheter per etasje	Bebygd areal	...	
Bruksareal for bygg med en bruksenhet	Bygningstype	Etasjer	Har tilbygg	Bebygd areal	...
Byggeår for bygninger	Bygningstype	Etasjer	Posisjon	Bebygd areal	...

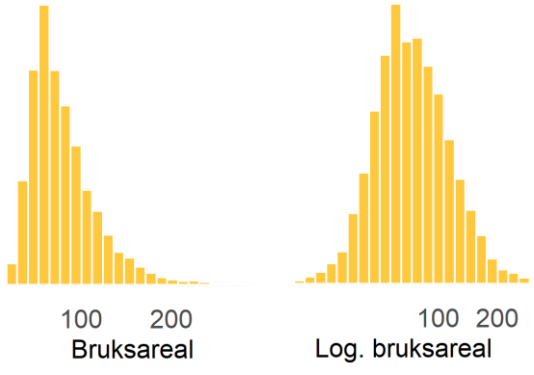
Strukturelle variabler som imputeres og enkelte av variablene som benyttes som grunnlag for imputeringen
*) Bruksenheter imputeres dersom bygget inneholder flere bruksenheter (typisk eierseksjoner).

Modellvalg i lineær modell

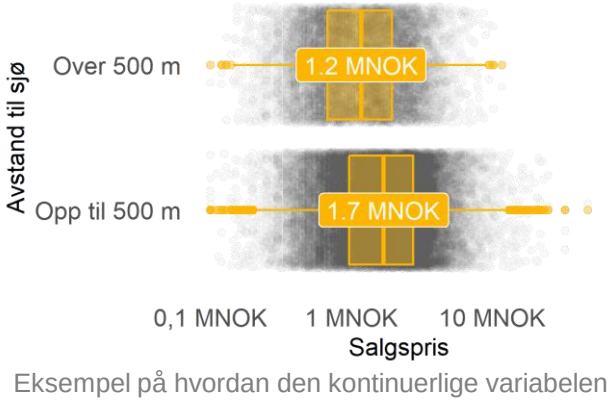
Lineære modeller antar en lineær sammenheng mellom boligens egenskaper og markedsverdi. Modellering av ikke-lineære forhold forutsetter derfor at variablene transformeres slik at sammenhengen kan representeres på en lineær måte. Vanlig teknikker inkluderer indikatorkoding av kategoriske variabler og logaritmiske transformasjoner og diskretisering av numeriske variabler.

Eiendomstype	Festegrunn	Eierseksjon
Grunneiendom	0	0
Grunneiendom	0	0
Festegrunn	1	0
Grunneiendom	0	0
Eierseksjon	0	1

Eksempel på hvordan den kategoriske variabelen *eiendomstype* kan representeres i en lineær modell.



Eksempel på hvordan en logaritmisk transformasjon gjør distribusjonen av *bruksareal* mer symmetrisk.



Eksempel på hvordan den kontinuerlige variabelen *avstand til sjø* kan diskretiseres til en indikatorvariabel.

Koding av kategoriske variabler

Lineære modeller antar en lineær sammenheng mellom variabler og utfall. Det er sjeldent at kategoriske variabler har verdier som er naturlig å representere med lineære sammenhenger. Slike forhold håndteres derfor typisk ved å opprette indikatorvariabler for variablenes verdier. Dette er svært enkelt når den kategoriske variabelen bare har to verdier, eksempelvis om boligen har peis. Da får variabelen verdien 1 dersom boligen har peis, og 0 dersom boligen ikke har peis. Kategoriske variabler har ofte flere verdier. I slike tilfeller benyttes gjerne en av kategoriene som referansekategori, mens det opprettes en indikatorvariabel for hver øvrige verdi.

Transformasjon av numeriske variabler

For å linearisere sammenhengen mellom variabler og utfall benyttes ofte numeriske transformasjoner av variablene. Slike transformasjoner kan gi større grad av symmetri for variabler med skjeve distribusjoner og kan gjøre modellen mindre sensitiv for ekstreme verdier. En vanlig transformasjon er å bruke den naturlige logaritmen av variabelens verdi fremfor den faktiske verdien. Dette er en effektiv måte å modellere sammenhenger hvor effekten av endringer avtar med størrelse. Arealer modelleres eksempelvis ofte ved bruk av logaritmer siden effekten av økning i bruksareal eller tomteareal er avtakende for store bygninger og tomter.

Diskretisering av numeriske variabler

Enkelte variabler kan ha stor betydning for utfallet dersom variabelens verdi ligger innenfor et gitt intervall, men være tilnærmet ubetydelig dersom verdien er utenfor dette intervallet. Et eksempel på dette er avstand til sjø. Fritidsboliger med kort avstand til sjø har gjerne en høyere markedsverdi enn fritidsboliger med stor avstand til sjøen, men markedsverdien synker ikke proporsjonalt ettersom avstanden blir større. Eksempelvis vil det for de fleste formål være likegyldig om en fritidsbolig ligger 40 eller 45 kilometer fra sjøen. Slike forhold modelleres ofte ved å dele variabelen inn i ulike intervaller. Dette kalles diskretisering av variabelen.

Modellvalg i trebasert modell

Trebaserte modeller krever svært lite forarbeid utover håndtering av kategoriske variabler. Dette skyldes at trebaserte modeller, i motsetning til lineære modeller, ikke antar noen bestemte forhold mellom boligens egenskaper og markedsverdi. Det overlates til algoritmen å identifisere slike underliggende sammenhenger. Algoritmen kan påvirkes gjennom å justere modellens parametre. Parametrene påvirker både presisjonen og stabiliteten i modellens verdsettelser. Selv om trebaserte modeller gjerne gir gode resultater med standardparametre, er det ofte ønskelig å utforske om andre parametre kan gi mer presise eller stabile verdsettelser. Et slikt parametersøk er viktig for at modellen skal gi stabile resultater.

Eiendomstype	Eiendomstype
Grunneiendom	1
Grunneiendom	1
Festegrunn	2
Grunneiendom	1
Eierseksjon	3

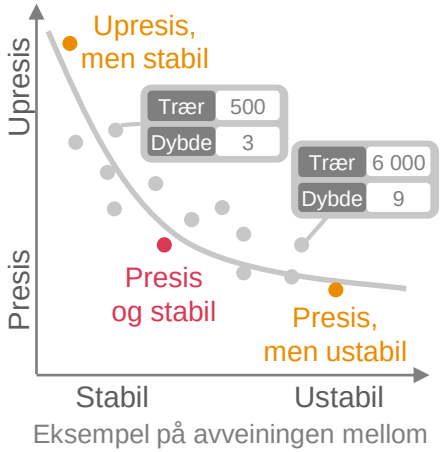
Eksempel på hvordan den kategoriske variabelen *eiendomstype* kan representeres i en trebasert modell.

Koding av kategoriske variabler

Trebaserte modeller er basert på rangering av variablenes verdier. Boligene deles i mindre og mindre grupper i treet basert på verdienes rangering. Ved å først skille mellom boliger med bruksareal under 80 kvm, og på neste nivå skille mellom boliger med bruksareal over 40 kvm er det mulig å behandle boliger mellom 40 og 80 kvm på samme måte. Modellen antar dermed ingen lineære sammenhenger for verdiene. Kategoriske variabler kan derfor modellers gjennom å tilegne hver kategoriske verdi en unik tallverdi. I eksempelet får grunneiendom verdien 1, festegrunn får verdien 2 og eierseksjon får verdien 3.

Parametersøk

De trebaserte modellene kan fange opp komplekse interaksjoner mellom boligens egenskaper gjennom dype trær med mange grener. Trærne kan være så dype og komplekse at modellene i noen tilfeller lærer sammenhenger som bare eksisterer for boligene i treningsdataene. Når de lærte sammenhengene ikke er generaliserbare til populasjonen av fritidsboliger er modellen overtilpasset. Modeller som er overtilpasset til treningsdataene gir ustabile verdsettelser og kan resultere i store årlige variasjoner i markedsverdi for enkeltboliger. For å gi stabile verdsettelser kan det derfor være ønskelig å begrense modellens kompleksitet. En enkel fremgangsmåte er å begrense antallet trær og trærnes dybde. Videre er det mulig å stille et minimumskrav til antall boliger som som skal inngå i en gren av treet. Dette hindrer modellen i å lære sammenhenger som bare er gyldig for et fåtall boliger. Antallet trær, trærnes dybde og antall boliger i en gren er eksempler på parametere som kan spesifiseres for trebaserte modeller.



Eksempel på avveiningen mellom presisjon og stabilitet for parametrene *antall trær* og *maksimal tredybde.*

Selv om modellene kan være overtilpasset til treningsdataene, kan modellene fremdeles fange opp andre gyldige sammenhenger. Slike sammenhenger kan potensielt forsvinne når trærnes kompleksitet begrenses. Konsekvensen er at komplekse modeller kan gi bedre presisjon i verdsettelsen enn enklere modeller selv om de er ustabile. Det må derfor ofte gjøres en avveining mellom modellens presisjon og stabilitet. For å identifisere hvordan avveiningen skal gjøres er det vanlig å gjennomføre et parametersøk for verdiene som regulerer modellens kompleksitet. Et slikt parametersøk innebærer at modellen testes med ulike kombinasjoner av parametre. Deretter er det mulig å velge parametrene som gir den mest presise modellen som også tilfredsstiller kriteriene for stabilitet. For den trebaserte modellen er det antatt at modellen er tilstrekkelig stabil dersom den er tilsvarende stabil som den lineære modellen. Parametersøket er nærmere beskrevet i Appendiks D.

Oppsummering av alle modellvariabler

Salgspris

Kategori	Utfallsvariabel	Imputering	Type	Trebasert	Lineær
Salgspris	Salgspris	Nei	Numerisk	Log	Log
Salgstidspunkt	Solgt år	Nei	Kontroll*	Direkte	Indikatorvariabler
Salgstidspunkt	Solgt måned	Nei	Kontroll*	Direkte	Utelatt

Strukturelle variabler

Kategori	Modellvariabel	Imputering	Type	Trebasert	Lineær
Bruksareal	Bruksareal	Ja	Numerisk	Direkte	Log
Byggeår	Byggeår	Ja	Numerisk	Direkte	Diskretisert, indikatorvariabler
Eiendomstype	Eiendomstype	Nei	Kategorisk	Numerisk koding	Indikatorvariabler
Andre bygg	Annet areal på tomten	Nei	Numerisk	Direkte	Log
Andre bygg	Antall boliger	Nei	Numerisk	Direkte	Utelatt
Andre bygg	Antall fritidsbygg	Nei	Numerisk	Direkte	Utelatt
Tomteareal	Tomteareal	Nei	Numerisk	Direkte	Log
Tomtehelning	Retning på tomtehelning	Nei	Numerisk	Direkte	Utelatt
Tomtehelning	Tomtehelning	Nei	Numerisk	Direkte	Utelatt
Vann og avløp	Avløpskode	Nei	Kategorisk	Numerisk koding	Indikatorvariabler
Vann og avløp	Vannforsyningskode	Nei	Kategorisk	Numerisk koding	Utelatt
Energi og varme	Energikildekode	Nei	Kategorisk	Numerisk koding	Utelatt
Energi og varme	Oppvarmingskode	Nei	Kategorisk	Numerisk koding	Utelatt

* Kontrollvariabler benyttes for å knytte salgsprisen til et gitt tidspunkt, slik at faktorer som markedsutvikling, inflasjon, o.l. kan hensyntas av modellen.

Oppsummering av alle modellvariabler

Lokasjonsvariabler

Kategori	Modellvariabel	Type	Trebasert	Lineær
Posisjon	Lengdegrad	Numerisk	Direkte	Direkte
Posisjon	Breddegrad	Numerisk	Direkte	Direkte
Adkomst	Avstand til offentlig vei	Numerisk	Direkte	Log
Adkomst	Høydeforskjell til offentlig vei	Numerisk	Direkte	Utelatt
Adkomst	Veitype for nærmeste offentlige vei	Kategorisk	Numerisk koding	Utelatt
Adkomst	Avstand til privat vei	Numerisk	Direkte	Diskretisert, indikatorvariabler
Adkomst	Høydeforskjell til privat vei	Numerisk	Direkte	Utelatt
Høyde	Høydemeter	Numerisk	Direkte	Andregradspolynom
Naboer	Naboer innenfor 3 km	Numerisk	Direkte	Log
Naboer	Fritidsboliger innenfor 1 km	Numerisk	Direkte	Log
Naboer	Primærboliger innenfor 1 km	Numerisk	Direkte	Log
Sjø	Avstand til sjø	Numerisk	Direkte	Diskretisert, indikatorvariabler
Sjø	Lengde på strandlinje	Numerisk	Direkte	Diskretisert, indikatorvariabler
Innsjø	Avstand til innsjø	Numerisk	Direkte	Diskretisert, indikatorvariabler
Innsjø	Lengde på innsjølinje	Numerisk	Direkte	Diskretisert, indikatorvariabler

Oppsummering av alle modellvariabler

Lokasjonsvariabler

Kategori	Modellvariabel	Type	Trebasert	Lineær
Skibakker	Avstand til skibakke	Numerisk	Direkte	Diskretisert, indikatorvariabler
Skibakker	Skibakke innenfor 10 km	Numerisk	Direkte	Diskretisert, indikatorvariabler
Turruter	Avstand til turrute	Numerisk	Direkte	Utelatt
Turruter	Lengde på turrute innenfor 1 km	Numerisk	Direkte	Diskretisert, indikatorvariabler
Innbyggere	Innbyggere innenfor 1 km	Numerisk	Direkte	Log
Innbyggere	Innbyggere innenfor 5 km	Numerisk	Direkte	Utelatt
Innbyggere	Innbyggere innenfor 50 km	Numerisk	Direkte	Utelatt
Innbyggere	Innbyggere innenfor 100 km	Numerisk	Direkte	Utelatt
Solforhold	Helning horisont	Numerisk	Direkte	Direkte
Solforhold	Årlige soltimer	Numerisk	Direkte	Direkte
Utsikt	Synlig areal	Numerisk	Direkte	Log
Utsikt	Synlig sjø/innsjøareal	Numerisk	Direkte	Diskretisert, indikatorvariabler
Admin.enhet	Fylke	Kategorisk	Numerisk koding	Utelatt
Admin.enhet	Kommune	Kategorisk	Numerisk koding	Indikatorvariabler

7. Resultater

Oppsummering

Verdsettelsesmodellen skal verdsette alle norske fritidsboliger og oppfylle de definerte vurderingskriteriene for presisjon, stabilitet og forklarbarhet. I dette kapittelet sammenlignes modellens oppfyllelse av kriteriene mot resultater for en tradisjonell lineær verdsettelsesmodell og dagens grunnlag for formuesverdier.

Verdsettelsesmodellen bidrar til økt treffsikkerhet og likebehandling i formuesbeskatning

Den utviklede modellen har et gjennomsnittlig avvik fra markedsverdi på 26,6 prosent. Dette er en reduksjon på 21,8 prosentpoeng fra dagens grunnlag for formuesverdier (48,4 prosent), og 8,7 prosentpoeng lavere enn den lineære modellen (35,3 prosent). Avviket mellom estimert og faktisk markedsverdi er symmetrisk. Dette innebærer at antall boliger som underestimeres er tilsvarende omfanget av boliger som overestimeres. Dette skiller seg fra dagens grunnlag for formuesverdier hvor to tredjedeler av boligene er underestimerte med mer enn 20 prosent. Resultatene viser at den trebaserte modellen vil bidra til økt likebehandling og riktigere formuesbeskatning av fritidsboliger.

Verdsettelsesmodellen vil forbedre seg ytterligere over tid

Modellens treffsikkerhet påvirkes av lav kvalitet på registrerte bygningsopplysninger i matrikkelen. Over halvparten av fritidsboligene mangler datoer for å identifisere bygningsalder og om lag 18 prosent mangler registrert bruksareal. For fritidsboliger med registrerte opplysninger om bruksareal og alder oppnår den trebaserte modellen et avvik mot markedsverdi på 21,1 prosent. Dette indikerer at modellens presisjon vil forbedres etter hvert som kvaliteten på bygningsopplysninger i matrikkelen økes.

Modellen vil også bli mer treffsikker over tid etter hvert som flere fritidsboliger omsettes. Resultater fra simuleringer indikerer at den trebaserte modellen oppnår en reduksjon i avvik fra markedsverdi på mellom 0,2 og 0,4 prosent per ekstra år med data.

Modellen har lavest avvik mot markedsverdi på sør- og innlandet, mens avviket er størst for Møre og Romsdal og de nordligste fylkene. Dette er fylkene med færrest omsatte boliger. Avviket varierer fra 23,2 prosent i Agder til 34,8 prosent i Troms og Finnmark. En mulig forklaringen er at kombinasjonen av færre boliger i treningsdataene, samt fritidsboliger på øyer uten fastlandsforbindelse og spredt bebyggelse gjør disse boligene utfordrende å estimere.

Modellen hensyntar markedsutvikling på en stabil måte uten store lokale variasjoner

Simuleringer av årlige skatteoppgjør for årene 2018, 2019 og 2020 viser at den trebaserte modellen gir en stabil utvikling i enkeltboligers markedsverdier uten store årlige variasjoner. Variasjonskoeffisienten for årlig markedsutvikling er på 3 prosent for både den trebaserte modellen og den lineære modellen.

Om lag 95 prosent av boligene har en årlig markedsutvikling som avviker mindre enn fem prosent fra gjennomsnittlig markedsutvikling i kommunen og 99,8 prosent avviker mindre enn 10 prosent. Resultatene indikerer at boliger innenfor en kommune i stor grad følger tilsvarende markedsutvikling. Den trebaserte og lineære modellen har et gjennomsnittlig standardavvik for årlig markedsutvikling i kommuner på henholdsvis 3,3 og 3,2 prosent.

Hver enkelt verdsettelse fra modellen kan forklares med bakgrunn i boligens egenskaper

Modellens verdsettelser kan dekomponeres i bidraget som hver enkelt egenskap har gitt til boligen. I gjennomsnitt utgjør strukturelle egenskaper som bruksareal og byggeår om lag 34 prosent av differansen fra markedsverdien av en gjennomsnittlig fritidsbolig. Antall nærliggende fritidsboliger og boligens avstand til sjø utgjør om lag 19 prosent av differansen.

Presisjon i verdsettelse

Verdsettelsesmodellen skal redusere dagens avvik mellom grunnlag for formuesverdier og reelle markedsverdier. Den trebaserte modellen gir et gjennomsnittlig avvik (MAPE) på 26,6 prosent fra boligenes faktiske markedsverdier. Til sammenligning er det et avvik mellom dagens grunnlag for formuesverdier og markedsverdi på 48,4 prosent. Modellen gir dermed en reduksjon i avvik på 21,8 prosentpoeng. Det gjennomsnittlige avviket påvirkes av at enkelte fritidsboliger med store avvik er utfordrende å verdsette. Modellen har et medianavvik (MdAPE) på 19 prosent. Dette innebærer at halvparten av de estimerte fritidsboligene avviker mindre enn 19 prosent fra markedsverdi.

Den lineære modellen har et gjennomsnittlig avvik fra markedsverdi på 35,3 prosent. Dette er 8 prosentpoeng høyere enn den trebaserte modellen. Modellen har et medianavvik på 25 prosent, og 41 prosent av de estimerte markedsverdiene har et absoluttavvik innenfor 20 prosent av boligens reelle markedsverdi (PM20).

Resultatene påvirkes av at mange fritidsboliger mangler registrerte bygningsopplysninger i matrikkelen. Om lag 18 prosent av fritidsboligene manger bruksareal, og over halvparten mangler informasjon om byggeår. For å belyse effekten av bedre datakvalitet har prosjektet foretatt en separat evaluering av treffsikkerheten for boliger med registrerte opplysninger om areal og byggeår. Dette omtales som boliger med “god datakvalitet”. For disse boligene oppnår modellen en verdifastsettelse med et gjennomsnittlig avvik på 21 prosent. Dette innebærer en reduksjon i avvik på 5 prosentpoeng. Den lineære modellen oppnår også en tilsvarende forbedring i prediksjonsevne. Resultatene gir indikasjoner på treffsikkerheten modellen kan oppnå med forbedret kvalitet på registrerte bygningsopplysninger i matrikkelen.

Rapporteringen av presisjonsmål er basert på kryssvalidering. Dette innebærer at målene reflekterer modellens presisjon for boliger som modellen ikke har blitt trent til å verdsette.

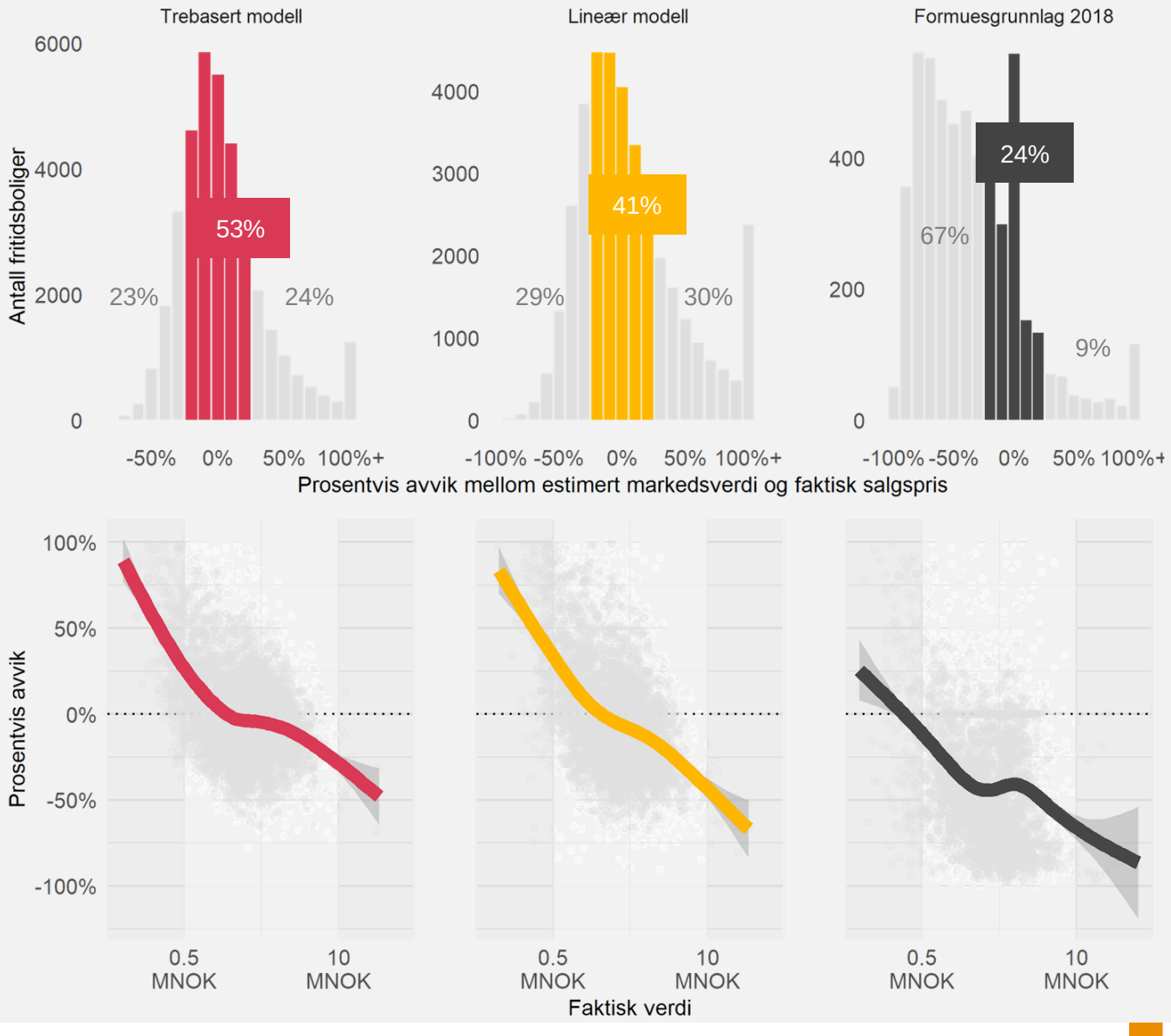
MAPE Trebasert modell		MAPE Lineær modell		MAPE Grunnlag for formuesverdi	
26,6 %		35,3 %		48,4 %	
God datakvalitet: 21%		God datakvalitet: 28%		God datakvalitet: 42%	
MdAPE	19%	MdAPE	25%	MdAPE	47%
PM20	53%	PM20	41%	PM20	24%
37 565		37 565		4 551	

Oppsummering av presisjonsmål for trebasert modell, lineær modell og grunnlag for formuesverdier for 2018.

Fordeling av avvik

Den utviklede verdsettelsesmodellen har en tilnærmet symmetrisk fordeling av avvik fra markedsverdi. Dette innebærer at omfanget av underestimerte boliger er like stort som omfanget av overestimerte boliger. Den trebaserte modellen underestimerer og overestimerer henholdsvis 23 og 24 prosent av boligene med mer enn 20 prosent. Til sammenligning underestimerer dagens grunnlag for formuesverdier 67 prosent av fritidsboligene med mer enn 20 prosent i forhold til reell markedsverdi.

For å forhindre at enkelte skattytere får uforholdsmessig høye formuesverdier benyttes det i dag en sikkerhetsventil i verdsettelsen av fritidsbolig. Sikkerhetsventilen innebærer at boligens formuesverdi ikke skal overstige 30 prosent av markedsverdien. Skattegrunnlaget viser at formuesverdien med sikkerhetsventil oversteg markedsverdien for 0,5 prosent av boligene i 2018. Når sikkerhetsventilen benyttes for verdsettelses-modellene inntreffer dette for henholdsvis 0,3 prosent og 0,7 prosent av boligene i den trebasert og lineære modellen. Modellen vil dermed føre til mer presise formuesverdier, uten å gi uforholdsmessig høy skatt.



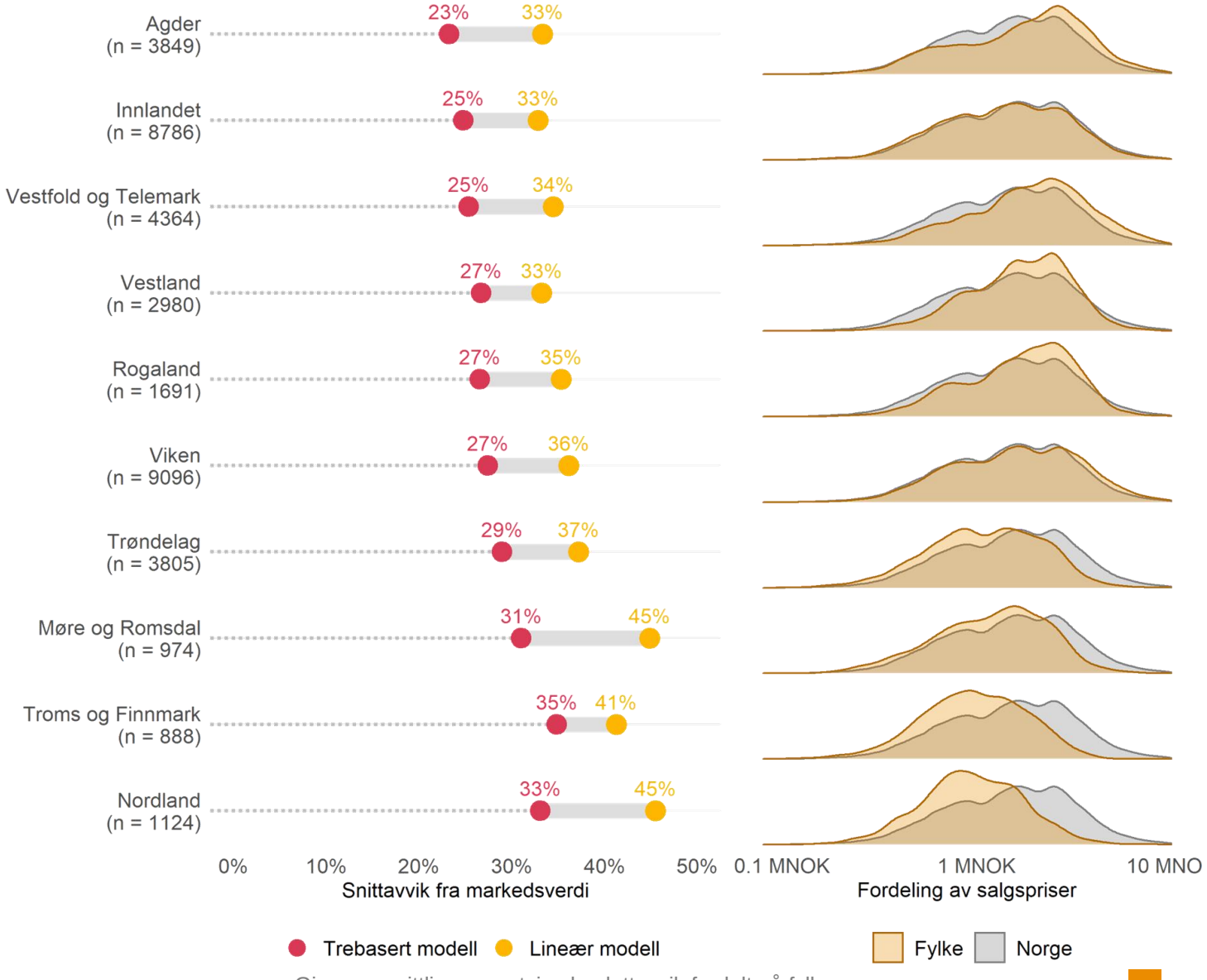
Sammenligning av (1) fordeling av avvik og (2) avvik og faktisk markedsverdi for den trebaserte modellen, den lineære modellen og Skatteetatens grunnlag for formuesverdi for 2018.

Regionale forskjeller

Verdsettelsesmodellen er mest presis i Agder, Innlandet og Vestfold og Telemark. I disse fylkene er snittavviket fra markedsverdi 25 prosent eller lavere. Modellen ser gjennomgående ut til å prestere best i områder med mange fritidsboliger og høyere prisnivå.

Snittavviket fra markedsverdi er størst for den nordlige landsdelen og Møre og Romsdal. Disse fylkene preges av en lavere omsetning av fritidsboliger og et lavere prisnivå enn resten av landet. Videre er alle de nevnte fylkene, med unntak av Møre og Romsdal, blant de største fylkene målt i areal. Dette medfører større geografisk spredning av fritidsboliger og færre nærliggende historiske omsetninger. Til sammenligning er Møre og Romsdal er et lite fylke, preget av mange fjorder og fjell. Lokasjonsvariablene er basert på luftlinjeavstander og kan dermed underestimere reiseavstander og overvurdere adkomstforhold og tilgjengelighet.

Den trebaserte modellen er mest treffsikker i alle fylker og oppnår typisk et avvik som er 6 til 10 prosentpoeng lavere enn den lineære modellen. Den lineære modellen gjør det spesielt dårlig i Møre og Romsdal, samt i Nordland hvor avviket er på 45 prosent. Dette er 12 til 14 prosentpoeng høyere enn den trebaserte modellen.



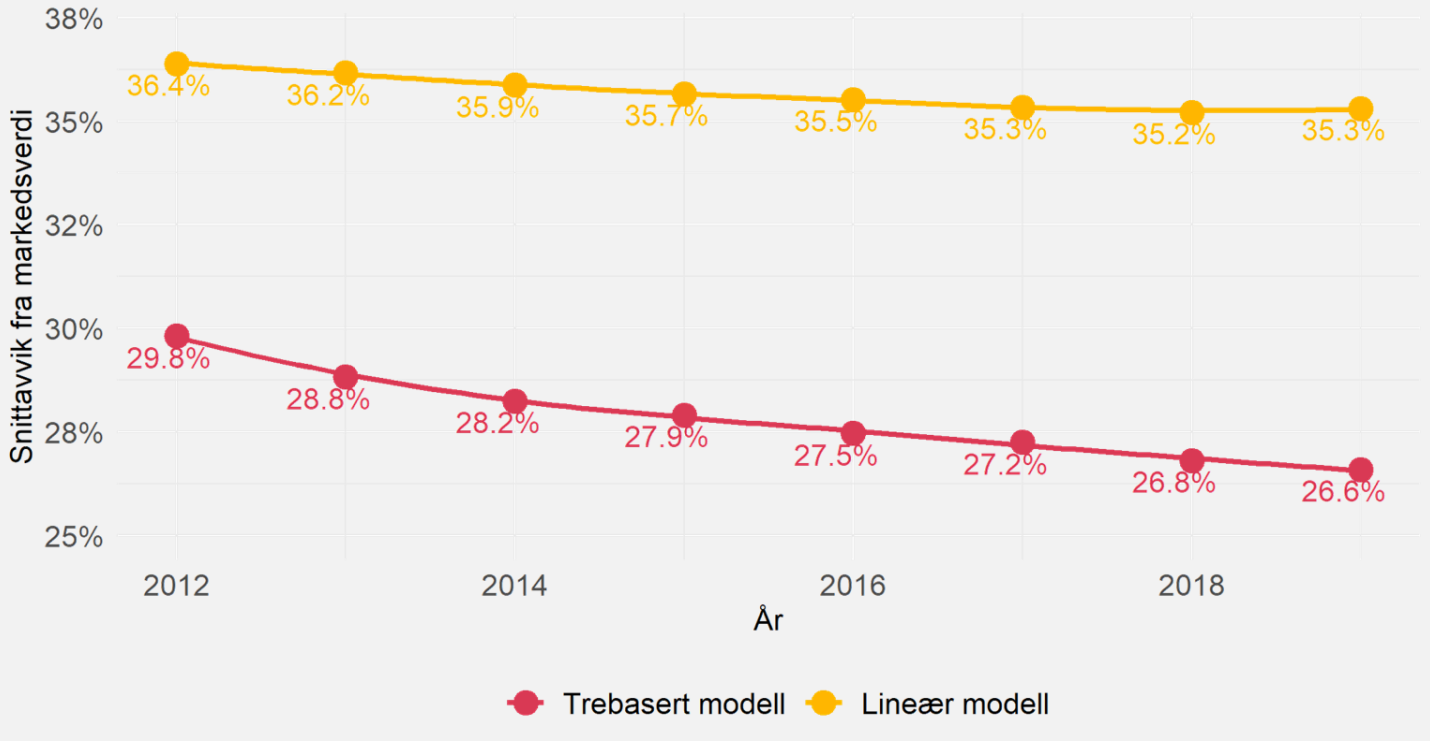
Gjennomsnittlig prosentvis absoluttavvik fordelt på fylker.

Effekten av mer data på modellens presisjon

Verdsettelsesmodellens treffsikkerhet avhenger av antallet tidligere boligomsetninger i utvalget. Dette skyldes at modellen er avhengig av flere eksempler for å identifisere sammenhenger som er generaliserbare for alle fritidsboliger. Etter hvert som modellen trenes på et større utvalg av tidligere boligsalg, vil den være i stand til å fange opp mer komplekse sammenhenger.

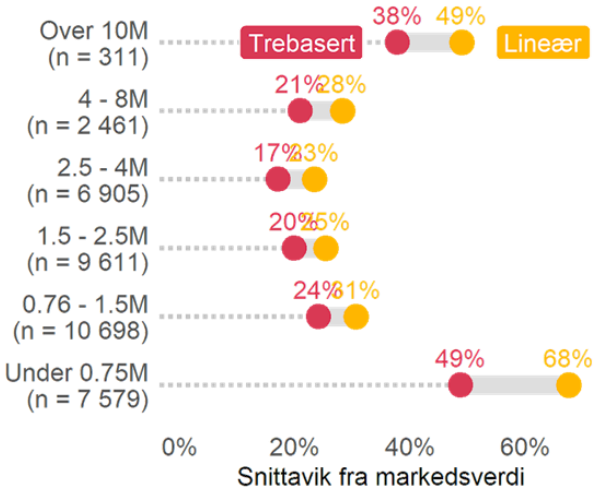
Den utviklede modellen er trent på 10 år med historiske omsetninger, og vil hvert år utvides med om lag 4 tusen boliger. For å belyse hvordan dette påvirker modellens presisjon har prosjektet evaluert modellenes årlige treffsikkerhet over tid. Dette innebærer at hvert årlige skatteoppgjør simuleres. Boligsalg som har skjedd etter skatteåret ekskluderes fra simuleringen. Resultatene viser at den trebaserte modellen hadde et gjennomsnittlig avvik på 29,8 prosent i 2012 mot 26,6 prosent i 2019. Dette innebærer at syv år med ekstra treningsdata forbedrer modellens presisjon med 3,2 prosentpoeng.

Den lineære modellen oppnår til sammenligning en mer konservativ forbedring i presisjon fra 2012 til 2019 på 1,4 prosentpoeng. Utviklingen i presisjon er flat fra 2017 til 2019. Dette reflekterer at den lineære modellen har en enkel struktur som i mindre grad nyttiggjør seg av de komplekse sammenhengene som kan identifiseres. Dersom den lineære modellen hadde vært utvidet med flere interaksjoner, eller om det hadde blitt utviklet en lineær modell for hver region, kunne den i større grad nyttiggjort seg av mer tilgjengelig data.



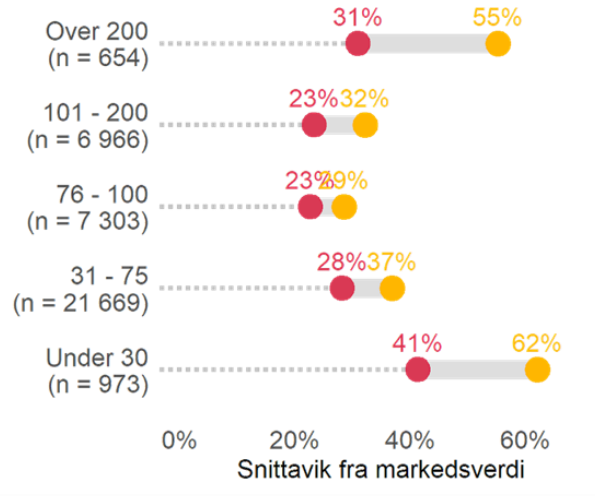
Gjennomsnittlig avvik for verdsettelsesmodellene når treningsdatasettet gradvis øker fra omsetninger registrert i perioden 2010 til 2012, til omsetninger registrert i perioden 2010 til 2019.

Presisjon i boligsegmenter



Salgspris

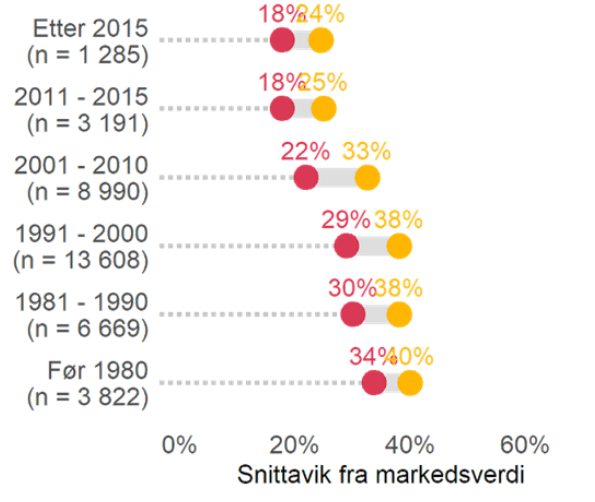
Modellene treffer best for boliger med markedsverdier mellom 1,5 og 4 millioner kroner. For boliger med salgspris under 750 tusen kroner er det gjennomsnittlige avviket i den trebaserte modellen på 49 prosent. De billigste boligene overvurderes gjennomgående i modellen og nærmere undersøkelser av de største avvikene viser at boligene holder gjennomgående lav bygningsstandard med behov for oppgradering eller totalrenoveringer. Variasjonen i bygningsstandard og fasiliteter er typisk større for fritidsboliger enn primærboliger, og byggeår er den eneste indikasjonen på byggets standard i modellen. Byggeåret er ikke et direkte mål på standard, og mange eldre fritidsboliger er i dag svært godt vedlikeholdt.



Bruksareal

Modellene treffer best for boliger med bruksareal mellom 30 og 200 kvadratmeter. Verdsettelsene fra den trebaserte modellen avviker i snitt 23 prosent fra reell markedsverdi for boliger med bruksareal mellom 76 og 200 kvadratmeter.

For små boliger med bruksareal under 30 kvadratmeter er avviket på 41 prosent for den trebaserte modellen. Dette utgjør om lag 2,5 prosent av boligene i utvalget. Om lag 35 prosent av boligene med bruksareal under 30 kvadratmeter har fått imputert bruksareal for minst ett bygg, sammenlignet med 25 prosent for hele utvalget. En av årsakene til lavere treffsikkerhet kan følgelig være at mindre boliger har en høyere grad av imputerte verdier.



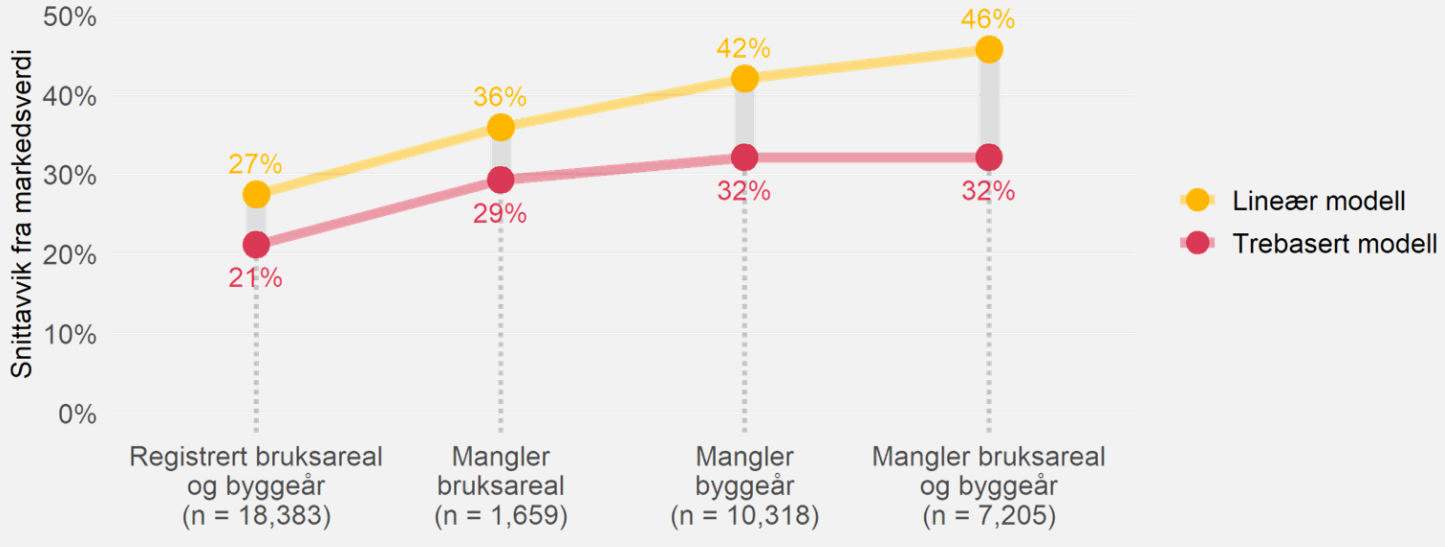
Byggeår

Modellene treffer best for nyere fritidsboliger. Boliger bygd etter 2010 har et gjennomsnittlig avvik på 18 prosent i den trebaserte modellen. Avviket blir gradvis større for eldre boliger. Dette kan skyldes at boligene har en større variasjon i bygningsstandard, fasiliteter og vedlikehold. Boliger bygd før 1980 har et snittavvik fra markedspris på 34 prosent.

Byggeår mangler for over halvparten av boligene i utvalget. Dette er primært eldre boliger som ikke har hatt søknadspliktige tiltak i nyere tid. Byggeåret er estimert for disse boligene. Ettersom utvalget i liten grad inkluderer eldre boliger med registrert byggeår kan en overvekt av disse imputert med en senere byggeår enn det som faktisk er tilfelle.

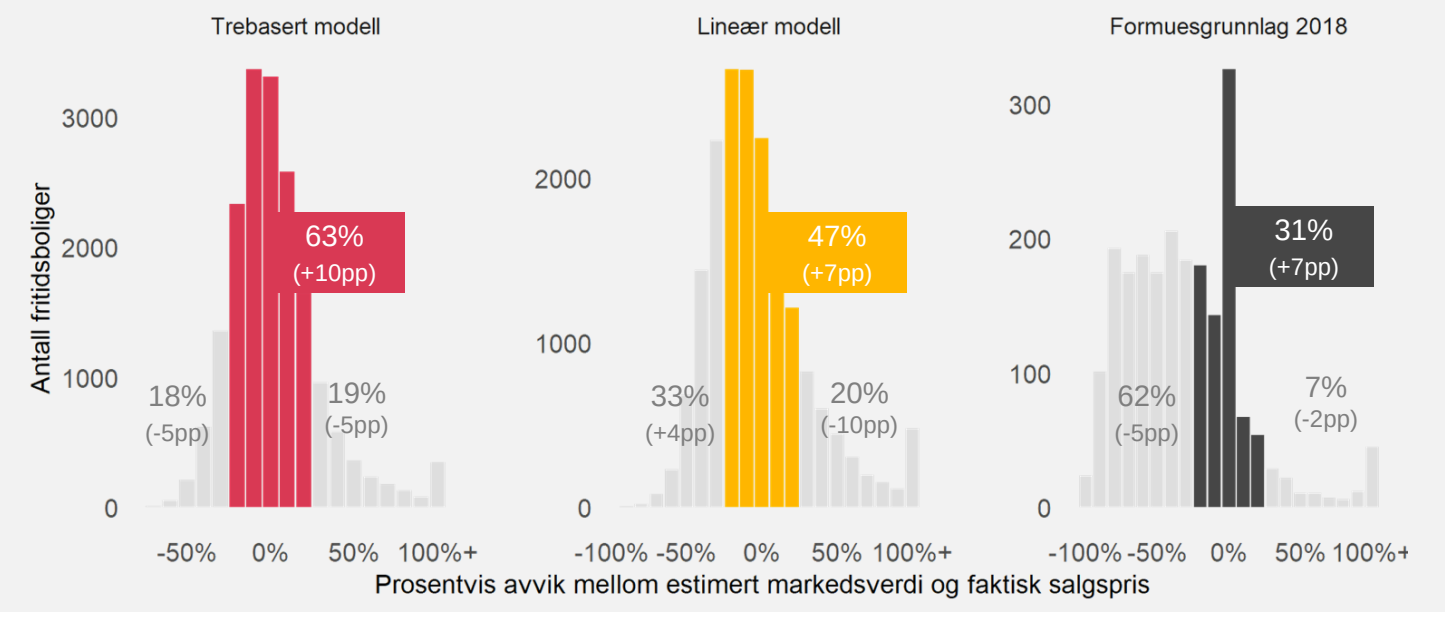
Presisjon og datakvalitet

Som vist i den innledende analysen er det varierende registreringsgrad og kvalitet på bygningsopplysninger i matrikkelen. Det pågår for tiden et langsiktig arbeid med kvalitetsheving av innholdet og det forventes at registreringsgraden vil bedres over tid. Økt datakvalitet vil resultere i økt presisjon for verdsettelsesmodellene. For boliger med fullstendig registrerte boligopplysninger har den trebaserte modellen et avvik fra gjennomsnittlig markedsverdi på 21 prosent. Dette innebærer en forbedring i presisjon på 6 prosentpoeng. Boliger som mangler bruksareal har et avvik på 29 prosent, mens boliger som mangler byggeår har et avvik på 32 prosent. Resultatene gir en indikasjon på modellens potensielle forbedring ved å øke kvaliteten i matrikkelen.



Gjennomsnittlig prosentvis absoluttavvik for fritidsboliger med manglende verdier.

Figuren under viser distribusjonen av avviket mellom estimerte markedsverdier og faktiske salgspriser for fritidsboligene med utfylt areal og byggeår. Fordelingen viser at en større andel av fritidsboligene har et avvik innenfor 20 prosent enn tidligere. Det er fremdeles noen avvik i halene av fordelingene.



Fordeling av avvik fra markedsverdi for med utfylt areal og byggeår i matrikkelen.

Variasjon i bygningsstandard

Gjennom å analysere boligene med størst negativt og positivt avvik fra markedsverdi er det mulig å belyse hvilke karakteristikker ved boligene modellen ikke fanger opp. Bygningsstandard er en av de gjennomgående årsakene til at markedsverdiene undervurderes eller overvurderes. En manuell gjennomgang av de største underestimeringene viser at boligene kjennetegnes av omfattende renoveringsbehov.

SALGSPRIS	TREBASERT	LINEÆR
0,16 MNOK	1,2 MNOK	1,5 MNOK

"Eiendommen har et godt potensiale, men bygningene bør anses som kondemnable."

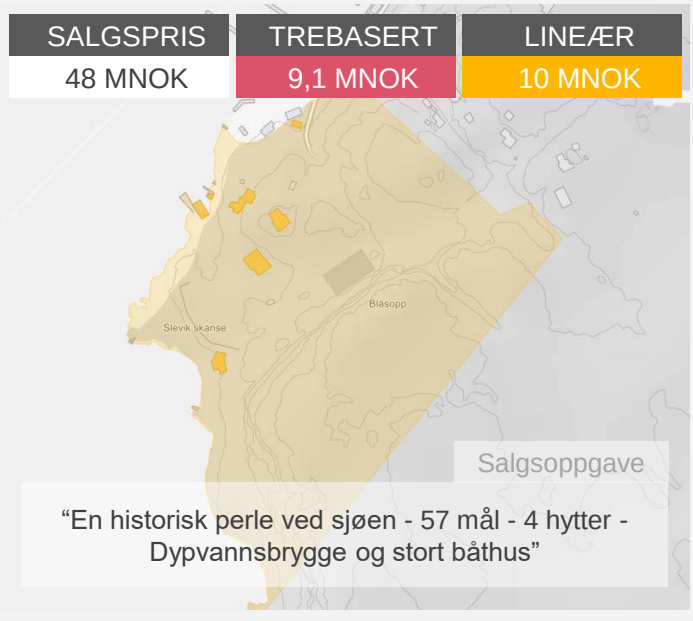
SALGSPRIS	TREBASERT	LINEÆR
0,15 MNOK	1,1 MNOK	1,2 MNOK

"Hytten fremstår i redusert stand, og det er påvist store mengder med museekskrementer i bygget. Dette ses i alle bygningskonstruksjonene"

Boligene som modellen i størst grad overvurderer bærer preg av dårlig standard og behov for totalrenovering.

I den andre enden av skalaen undervurderer modellen boliger med svært høy standard. En medvirkende årsak til at modellen undervurderer dyre boliger kan være at det er under 150 boliger i treningsdataene med salgspris over 10 millioner kroner

SALGSPRIS	TREBASERT	LINEÆR
48 MNOK	9,1 MNOK	10 MNOK



"En historisk perle ved sjøen - 57 mål - 4 hytter - Dypvannsbrygge og stort båthus"

SALGSPRIS	TREBASERT	LINEÆR
15 MNOK	4,5 MNOK	5,2 MNOK

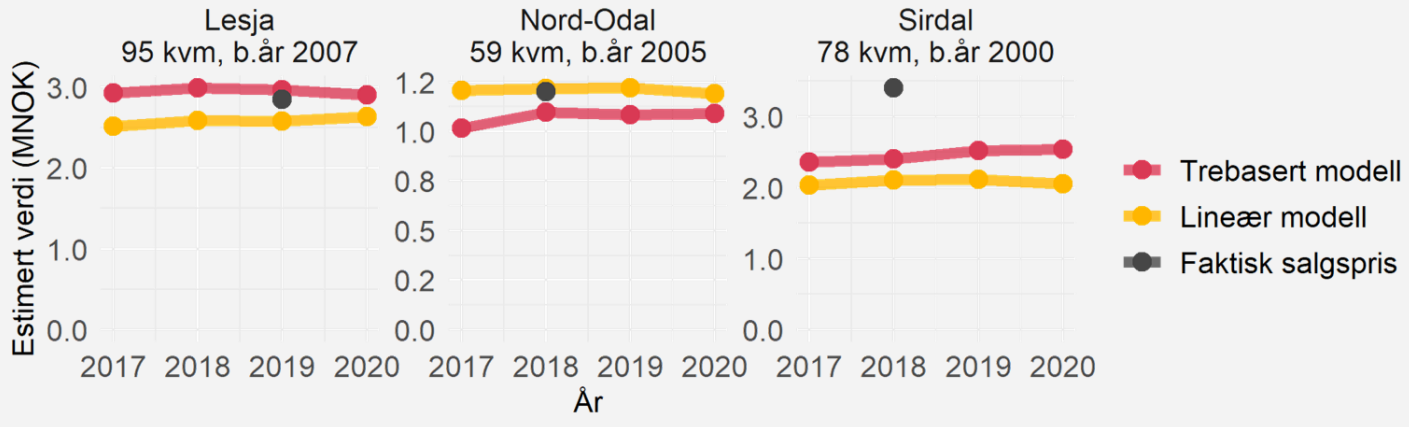
"Eksklusiv, moderne hytte med høy standard og høy håndverksmessig kvalitet. Oppført i klassisk sørlandsstil. Upåklagelig utsikt"

Boligene som modellen i størst grad undervurderer bærer preg av høy standard.

Stabilitet i verdsettelser

Det andre kravet til verdsettelsesmodellen er at modellen skal være stabil. Dette er et todelt krav. For det første skal modellen gi en forutsigbar utvikling i formuesverdier for skattyter uten store variasjoner utover generell markedsutvikling. I tillegg skal boliger innenfor et avgrenset område følge tilsvarende overordnede utvikling, med mindre utviklingen for enkeltboliger kan tilskrives lokale forhold eller strukturelle endringer.

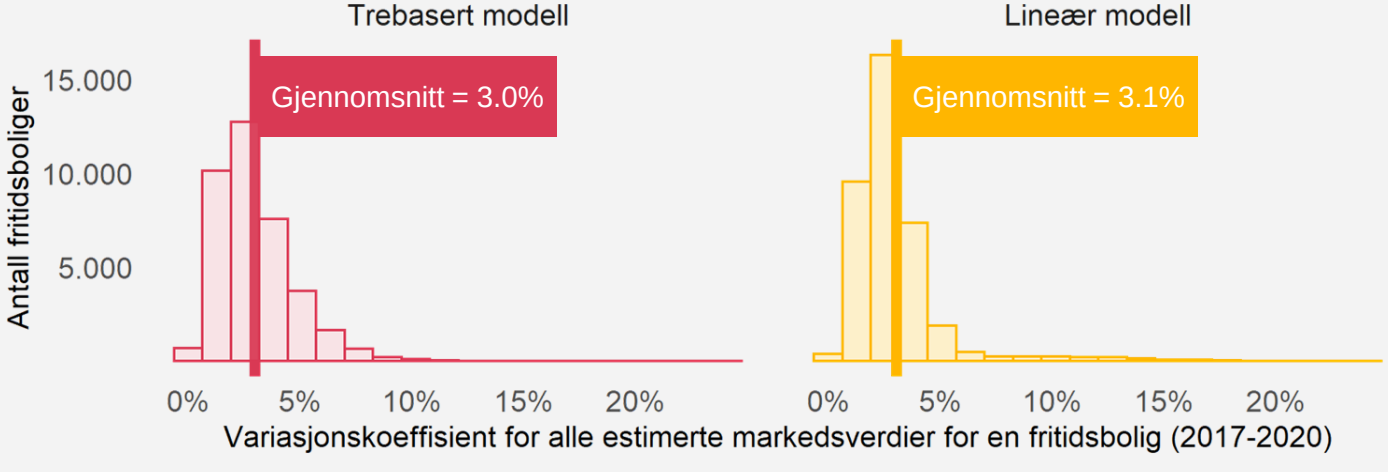
For å vurdere modellenes stabilitet er det gjennomført en simulering av årlige oppdateringer av modellene. Simuleringene gjøres på samme måte som et årlig skatteoppgjør ville blitt utført. For 2017 innebærer dette at modellene trenes på boligomsetninger opp til 2017. Modellen estimerer så markedsverdien for alle fritidsboliger i 2017. Simuleringen gjennomføres deretter for årene 2018, 2019 og 2020.[1]



Stabilitet i årlig estimert markedsverdi for tre tilfeldig valgte fritidsboliger basert på historiske simuleringer.

Stabilitetsmål 1: Forutsigbar utvikling i markedsverdier for skattyter

Simuleringene viser at enkeltboliger har en gjennomsnittlig variasjonskoeffisient på om lag 3 prosent i begge modellene. Dette betyr at den årlige justeringen for en bolig verdt 1 million kroner vil ha et standardavvik på 30 tusen kroner i perioden 2017 - 2020. En variasjonskoefffisient på 0 prosent innebærer at modellen antar ingen markedsutvikling. Det er følgelig ikke en målsetning at variasjonskoeffisienten skal være lavest mulig, men heller tilstrekkelig lav til å gi en stabil utvikling. Resultatene viser at modellene gir forutsigbare estimater over tid for de fleste fritidsboliger, samtidig som de fanger opp den generelle prisutviklingen i markedet fra år til år.



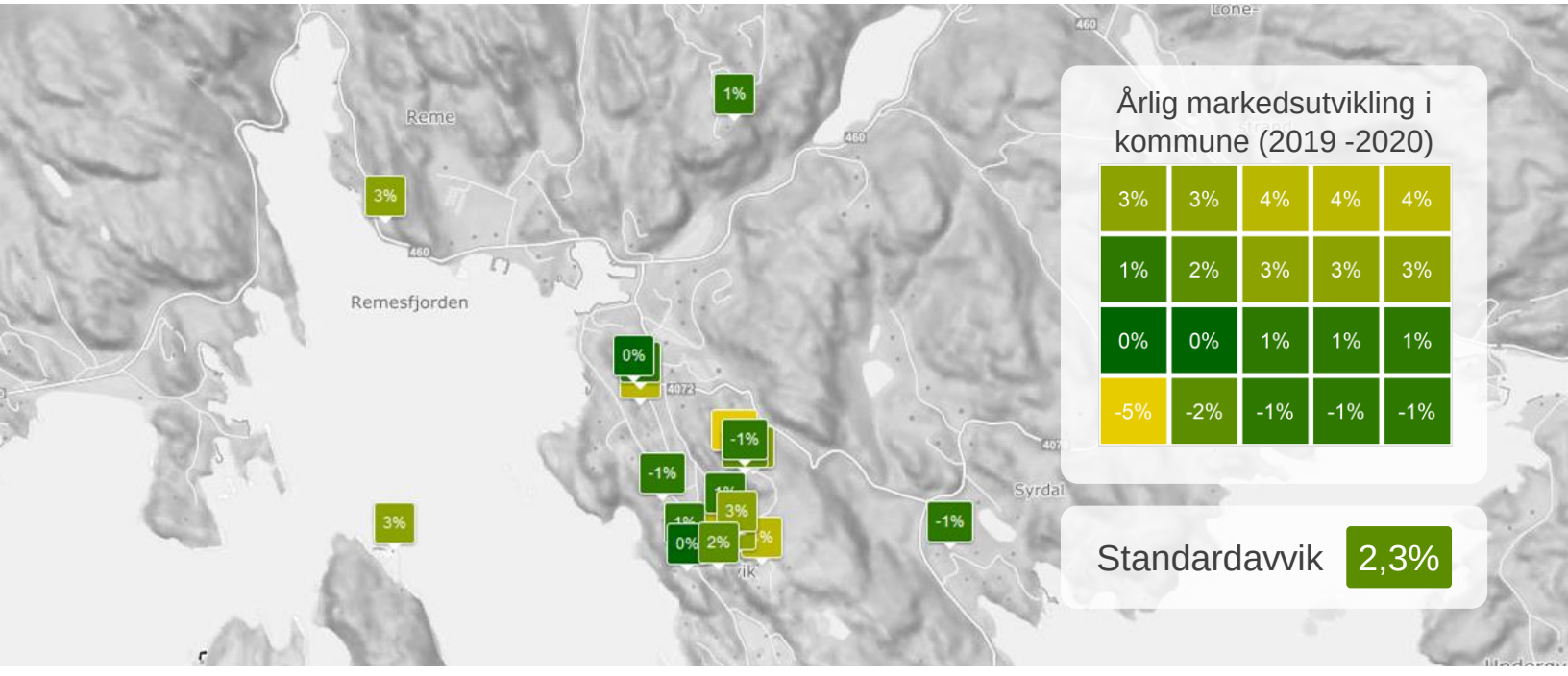
Variasjonskoeffisienter for alle fritidsboliger.

1) Prosjektet benytter kryssvalidering for å simulere årlige verdsettelser

Stabilitet i verdsettelser

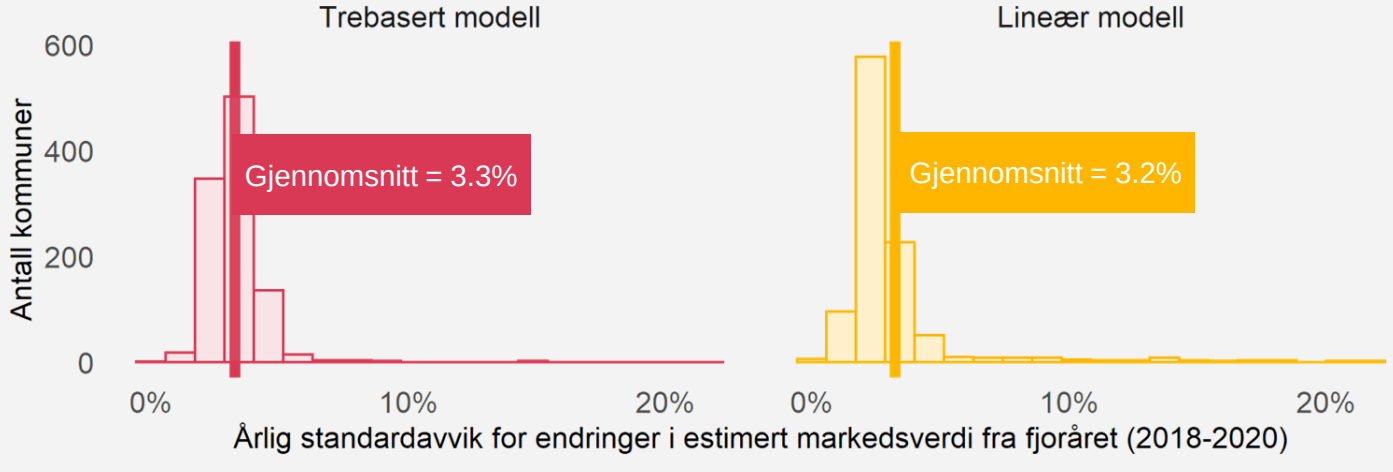
Stabilitetsmål 2: Tilsvarende utvikling for skattytere innenfor samme område

For å beregne variasjonene i utvikling for skattytere innenfor samme område beregnes standardavviket for årlige endringer innenfor kommunen. For å gi et nasjonalt mål på tvers av kommuner beregnes det gjennomsnittlige standardavviket for alle kommuner.



Årlig markedsutvikling fra 2019 til 2020 med den trebaserte modellen for et utvalg av boliger innenfor en kommune

Begge modellene har et gjennomsnittlig standardavvik i markedsutvikling for boligene innad i kommunen på mellom 3,2 og 3,3 prosent. Dette indikerer at modellene gir en stabil utvikling innenfor samme område uten store utslag. Resultatene for den trebaserte modellen for 2020 viser at 99,8 prosent av boligene var innenfor 10 prosent av kommunens gjennomsnittlige markedsutvikling, og over 95 prosent var innenfor 5 prosent.



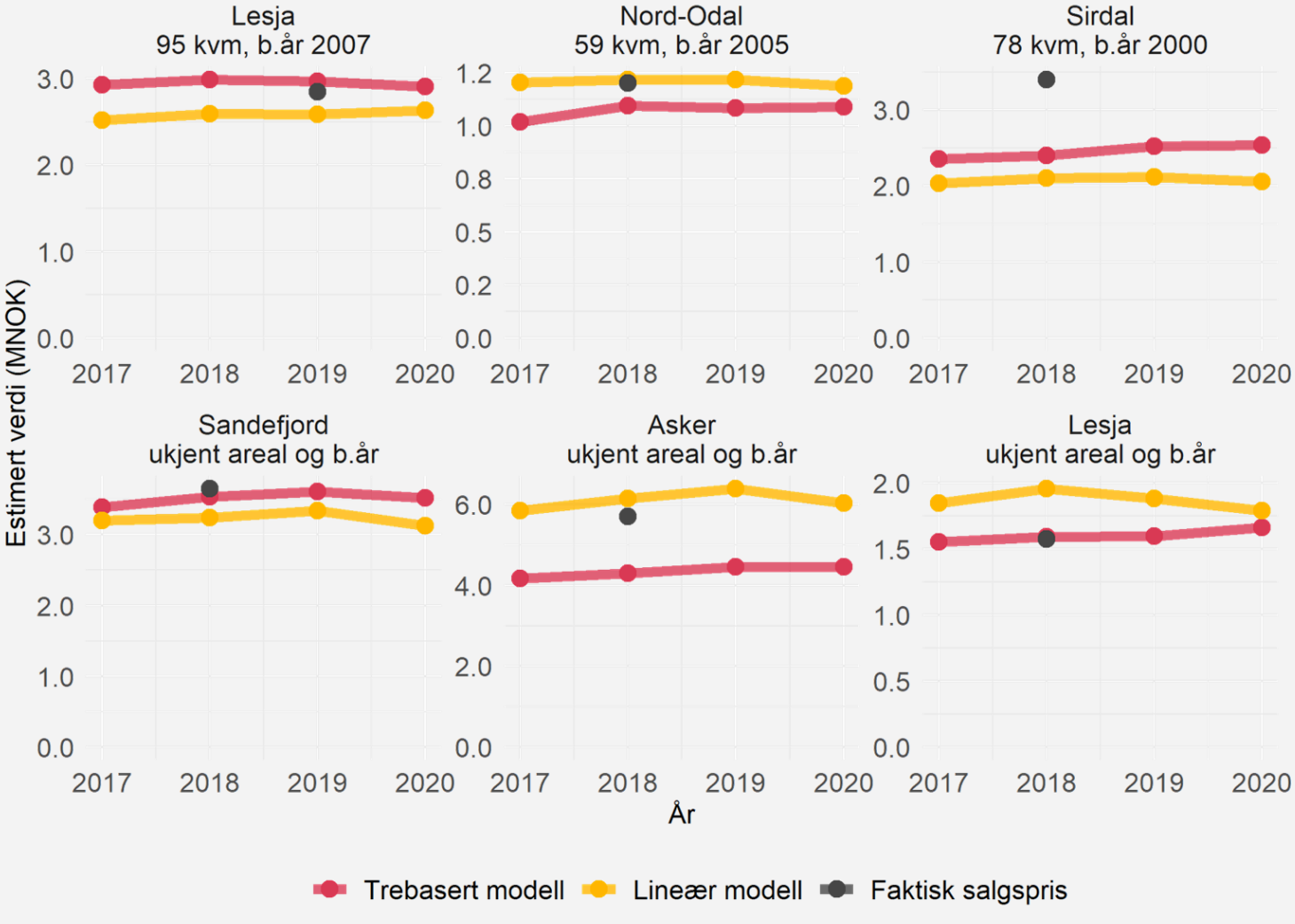
Det gjennomsnittlige standardavviket for kommunene viser at modellen gir stabile resultater innenfor avgrensede områder.

Stabilitet og datakvalitet

Modellens stabilitet påvirkes av kvaliteten på registrerte bygningsopplysninger. Den trebaserte modellen har en gjennomsnittlig variasjonskoeffisient (Stabilitetsmål 1) på 3 prosent for alle fritidsboliger. For boliger med fullstendig registrerte boligopplysninger oppnår modellen en variasjonskoeffisient på 2,8 prosent. Den lineære modellen oppnår en tilsvarende forbedring på 0,2 prosentpoeng. For markedsutvikling innenfor en kommune (Stabilitetsmål 2) reduseres det gjennomsnittlige standardavviket fra 3,3 prosent til 2,9 prosent for den trebaserte modellen, mens det forblir uendret i den lineære modellen. Begge modellene er i utgangspunktet stabile, men resultatene indikerer at det mulig å forbedre stabiliteten i modellene ytterligere ved å heve kvaliteten på bygningsopplysninger i matrikkelen.

Årlig estimert markedsverdi for en fritidsbolig

Figuren viser utvikling i markedsverdi for seks tilfeldige fritidsboliger i perioden 2017-2020. På øverste rad har alle boligene fullstendig registrerte boligopplysninger. Nederste rad viser boliger med manglende opplysninger. Figuren viser at det er mer variasjon for boligene med manglende opplysninger og at den trebaserte modellen i de fleste tilfeller treffer bedre enn den lineære modellen.



Utvikling av markedsverdi for seks tilfeldige fritidsboliger i perioden 2017-2020.

Lokal forklarbarhet

Den trebaserte modellen kan forklares gjennom modellens SHAP-verdier. SHAP-verdiene kvantifiserer bidraget som hver enkelt boligegenskap gir til den estimerte markedsverdien.[2] Beregningen tar utgangspunkt i den gjennomsnittlige markedsverdien av alle fritidsboliger. Forklaringen for en enkeltbolig beskriver dermed hvordan boligens egenskaper har bidratt til å øke eller redusere verdsettelsen sammenlignet med en gjennomsnittlig fritidsbolig.

For en verdsettelsesmodell med mange variabler vil det alltid være enkelte egenskaper som gir et lavere bidrag til verdsettelsen enn andre egenskaper. Dersom alle store og små bidrag vises samtidig kan en slik forklaring virke overveldende for skattytere. For å gi en en enkel og håndtering forklaring kan egenskapene som gir størst positivt og negativt bidrag vises enkeltvis, og bidraget fra resterende egenskapene summeres.

Eksempel 1: Grimstad

Gjennomsnittlig markedsverdi		1 960 000
Topp 5 egenskaper som øker verdi		
Bruksareal	128 kvm	23%
Breddegrad	58	19%
Høydemeter	17 m	18%
Strandlinje	76 m	16%
Tomteareal	3 200 kvm	9%
Topp 5 egenskaper som senker verdi		
Areal andre bygg	0	-3%
Fylke	Agder	-3%
Antall fritidsboliger (1 km)	71	-2%
Vannforsyning	Privat	-2%
Tomtehelning (retning)	152 (sør)	-2%
Andre egenskaper		29%
Estimert markedsverdi		4 900 000

Eksempel 2: Hol

Gjennomsnittlig markedsverdi		1 960 000
Topp 5 egenskaper som øker verdi		
Areal skibakke (1 km)	1 300 dekar	20%
Bruksareal	87 kvm	16%
Fylke	Viken	5%
Lengdegrad	8	4%
Innbyggere (5km)	2 800	4%
Topp 5 egenskaper som senker verdi		
Avstand vei	32 m	-5%
Byggeår	Ukjent (est. til 1998)	-3%
Antall fritidsboliger (1 km)	101	-2%
Areal andre bygg	37 kvm	-2%
Avstand innsjø	612 m	-1%
Andre egenskaper		-60%
Estimert markedsverdi		1 080 000

Forklaring av verdsettelser for to vilkårlig valgte fritidsboliger. De fem egenskapene ved boligen som gir størst positivt og negativt bidrag vises i figuren, mens bidraget fra de resterende egenskapene samles i kategorien “Andre egenskaper”

For å gi forklaringer som er gjenkjennelige for skattytere bør egenskapenes bidrag sees i lys av den faktiske verdien som legges til grunn for verdsettelsen. Modellen kan eksempelvis tilegne et ukjent bruksareal et signifikant negativt bidrag. Dersom eieren av en stor fritidsbolig ikke får opplyst at verdsettelsen baserer seg på et ukjent bruksareal, vil en slik verdsettelse fremstå som lite intuitiv. Synliggjøring av manglende eller feilregistrerte verdier vil også kunne bidra til å forbedre kvaliteten på bygningsopplysninger i matrikkelen.

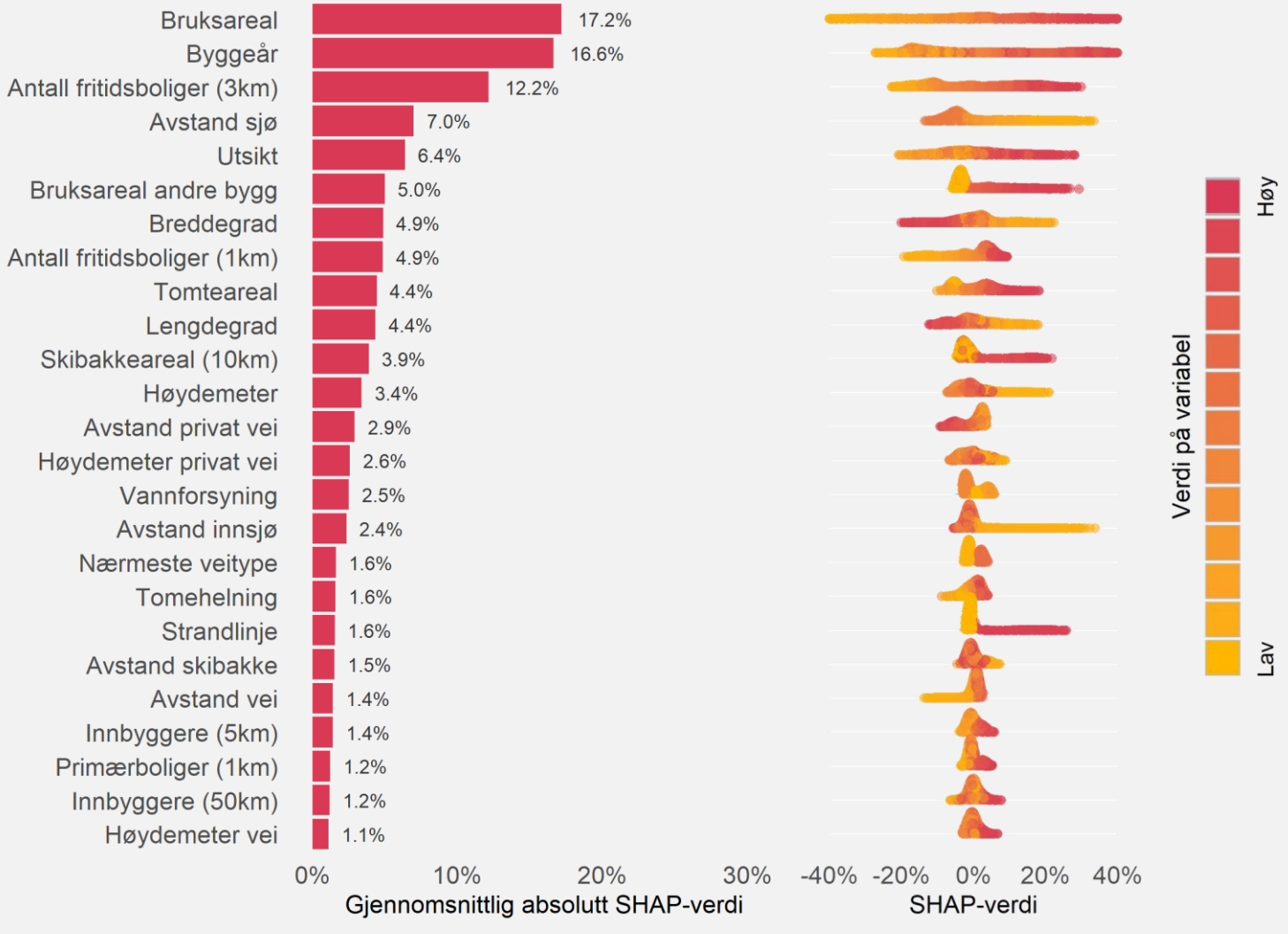
2) Siden prosjektet benytter en multiplikativ modell uttrykkes SHAP-verdiene i log-skala. Log-skalaen har attraktive numeriske egenskaper, men kan være vanskeligere å forstå for skattytere. En vanlig approksimasjon er derfor å uttrykke verdiene i prosent.

Global forklarbarhet

Ved å aggregere bidragene fra alle verdsettelsene er det mulig å forklare hvilken egenskaper som har størst påvirkning på boligens markedsverdi. For å unngå at positive og negative bidrag nuller hverandre ut i beregningen benyttes ofte den gjennomsnittlige absolutte SHAP-verdien for hver enkelt egenskap.

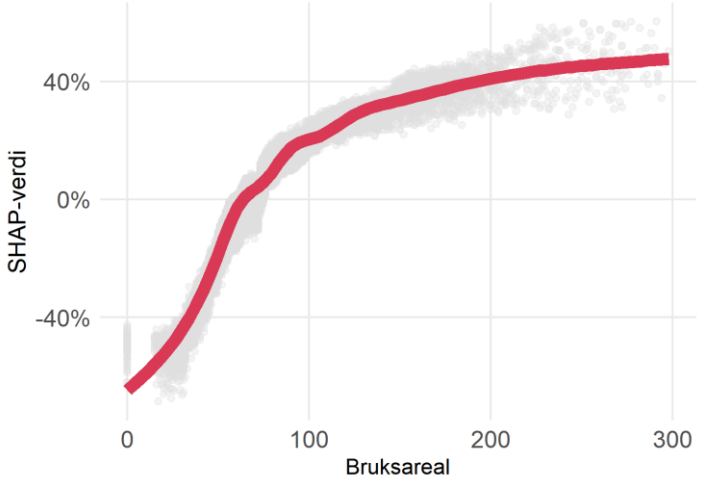
I gjennomsnitt utgjør strukturelle egenskaper som bruksareal og byggeår om lag 34 prosent av differansen fra markedsverdien av en gjennomsnittlig fritidsbolig. Den relative viktigheten av boligens alder og størrelse er i tråd med funn fra tidligere studier. Antall nærliggende fritidsboliger og boligens avstand til sjø utgjør om lag 19 prosent av differansen. Antall nærliggende fritidsboliger skiller mellom fritidsboliger med tett og avsidesliggende bebyggelse.

Selv om fire variabler i gjennomsnitt utgjør over halvparten av differansen, kan andre variabler gi store utslag for enkelte verdsettelser. Et eksempel på dette er lengden på boligens strandlinje. Fem prosent av fritidsboligene har strandlinje, og for disse boligene gir strandlinje et positivt bidrag på opp til 35 prosent.



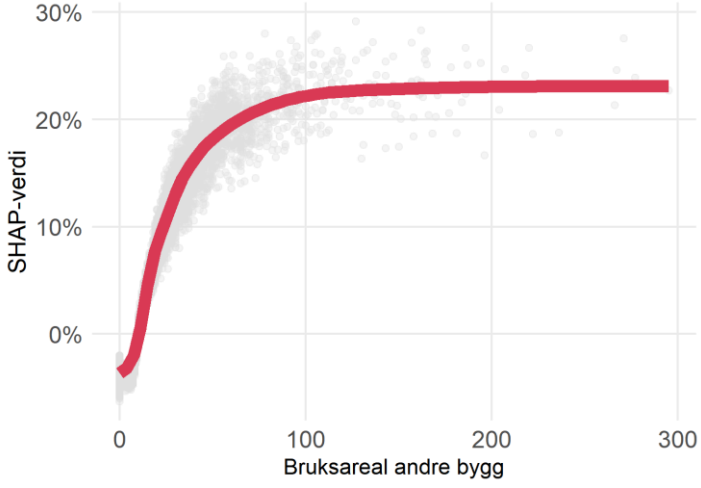
Gjennomsnittlig bidrag for de ulike boligegenskapene (venstre) og bidraget avhengig av egenskapenes verdi (høyre).

Boligens størrelse og alder



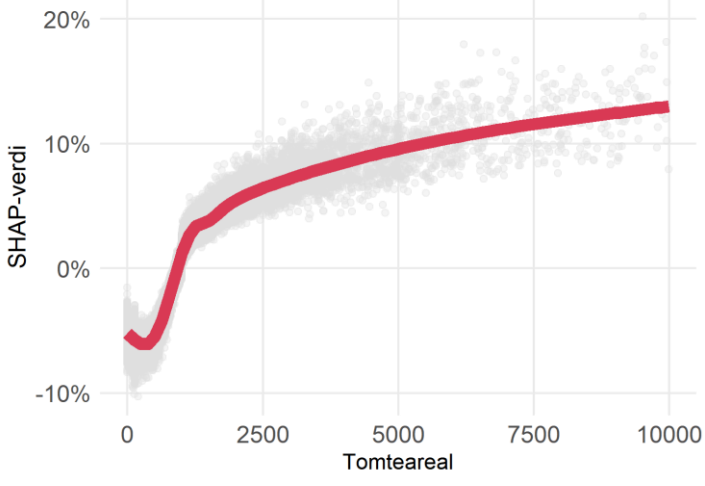
Bruksareal

Lavt bruksareal reduserer boligens markedsverdi med opp til 80 prosent. Markedsverdien ser ut til å ha en lineær prosentvis økning opp til 80 kvadratmeter, og flater deretter gradvis ut. Den relative effekten av en økning fra 300 til 400 kvadratmeter er langt mindre enn en økning fra 100 til 200 kvadratmeter.



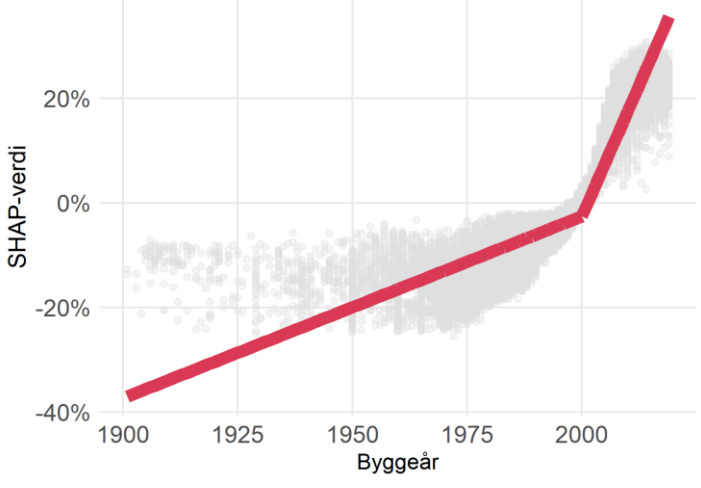
Bruksareal andre bygg

Flere fritidsboliger har bygninger som garasjer, naust og annekser på tomten. Arealet av disse byggene gir et positivt bidrag til markedsverdien som er tilnærmet lineært opp til 50 kvadratmeter før det gradvis flater ut. Som forventet følger forholdet samme form som bruksareal med lavere påvirkning av markedsverdien.



Tomteareal

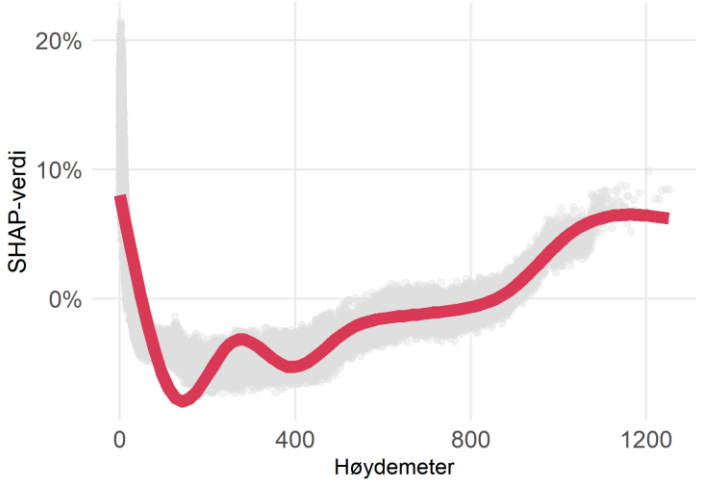
Større tomteareal gir et positivt bidrag til boligens markedsverdi. Effekten av større areal ser ut til å avta ved om lag 1 000 kvadratmeter, og flater ytterligere ut ved om lag 5 000 kvadratmeter. Festegrunner og eierseksjoner har ofte et tomteareal på 0 kvadratmeter som gir et negativt bidrag til boligens markedsverdi.



Byggeår

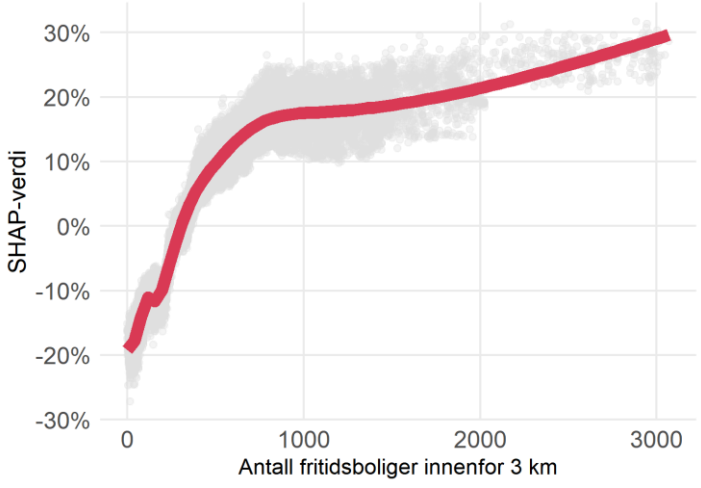
Hvert bruksår reduserer boligens markedsverdi. En nyere bolig kan bidra til å øke markedsverdien med opp opp til 25 prosent. Til sammenligning gir en 40 år gammel fritidsbolig et negativt bidrag på opp til 20 prosent. Verditapet er størst de første 40 årene etter boligen er bygd.

Boligens beliggenhet



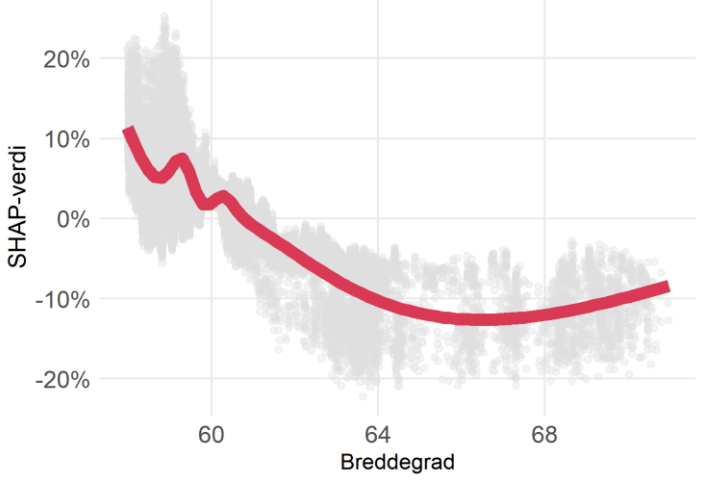
Høydemeter

Fritidsboliger som ligger under 100 meter eller over 900 meter får et positivt bidrag til markedsverdi. Dette er typisk fritidsboliger som enten ligger nært sjøen, eller høyt i fjellet. Boliger som ligger mellom 200 og 400 høydemeter får det største negative bidraget til markedsverdi.



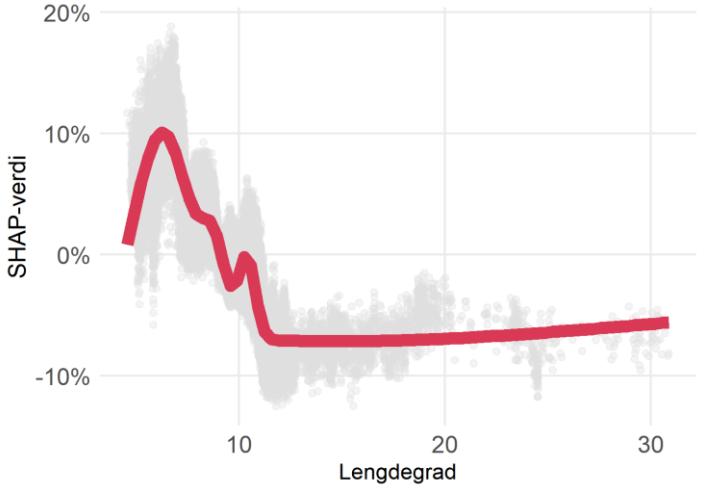
Antall fritidsboliger innenfor 3 km

Avsidesliggende fritidsboliger uten naboer kan gi et negativt bidrag til markedsverdi på opp til 25 prosent. Flere naboer gir et positivt bidrag til boligens markedsverdi. Effekten flater ut ved om lag 750 fritidsboliger, mens boliger med over 3000 fritids-boliger innenfor 3 km får et yttereligere positivt bidrag.



Breddegrad

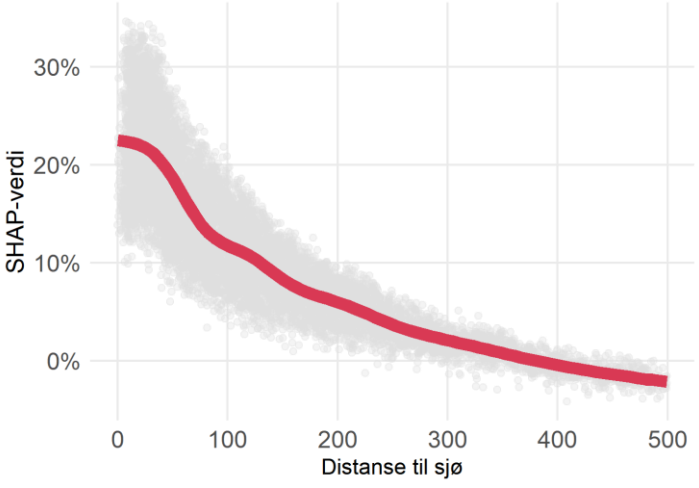
Breddegrad representerer sør-nord aksen av Norge. Fritidsboliger lengst sør i landet får et positivt bidrag til markedsverdi på opp til 25 prosent. Dette kan også være en konsekvens av at boligene typisk er lokalisert nært sjøen. Deretter faller bidraget med hver breddegrad. I nord gir breddegraden et negativt bidrag til markedsverdi på opp til 20 prosent.

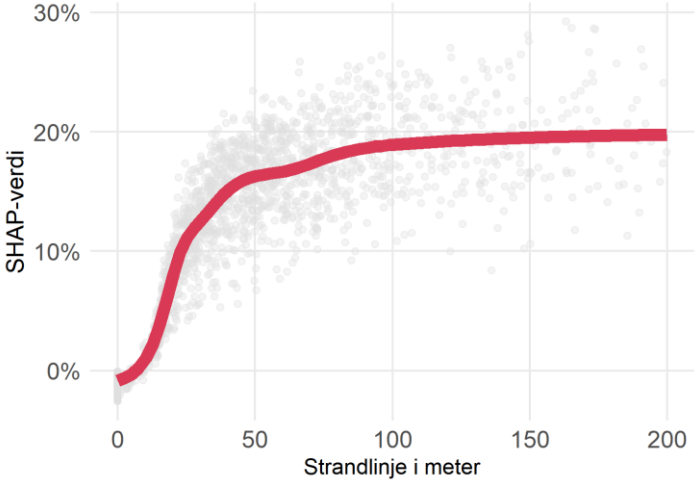


Lengdegrad

Lengdegrad representerer vest-øst aksen av Norge. Fritidsboliger vest i landet får et positivt bidrag til markedsverdi på opp til 20 prosent. Deretter faller bidraget for hver lengdegrad østover. Øst for lengdegrad 12 gir lengdegraden en konstant negativ effekt på mellom 5 og 15 prosent. Denne effekten slår i praksis kun ut for Nordland og Troms og Finnmark.

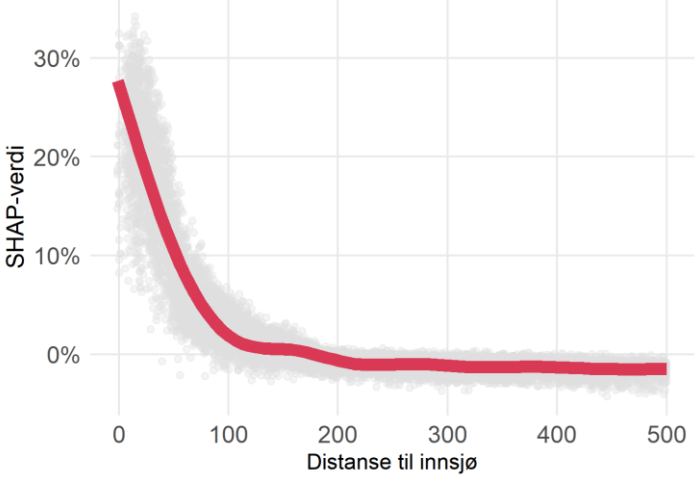
Beliggenhet ved sjø og innsjø

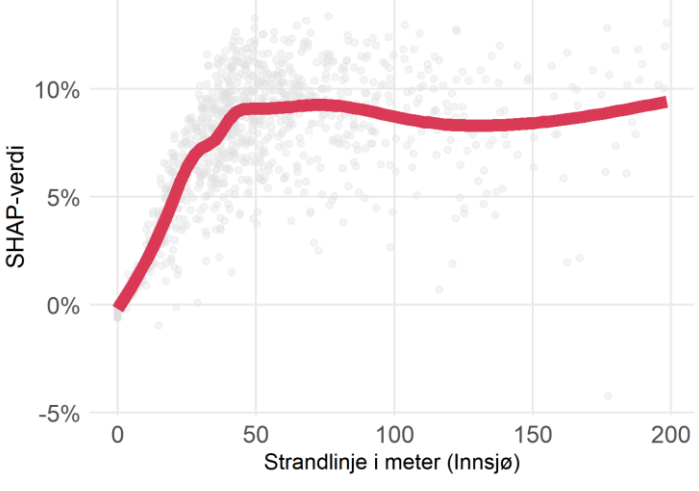




Beliggenhet ved sjø

Beliggenhet ved sjø kan både beskrives av boligens luftlinjeavstand til sjø, og boligens strandlinje. Kort avstand til sjø kan gi opp til 35 prosent økning i markedsverdi for enkelte boliger. Hver meter av strandlinje ser ut til å gi en tilnærmet lineær økning i prosentvis markedsverdi opp til 50 meter hvor utviklingen flater ut. Strandlinjen kan for enkelte boliger gi et bidrag på opp til 30 prosent. Det bør legges til at en del av effekten av beliggenhet ved sjø også kan fanges opp av boligens breddegrad og lengdegrad, samt variabelen for sjøutsikt.

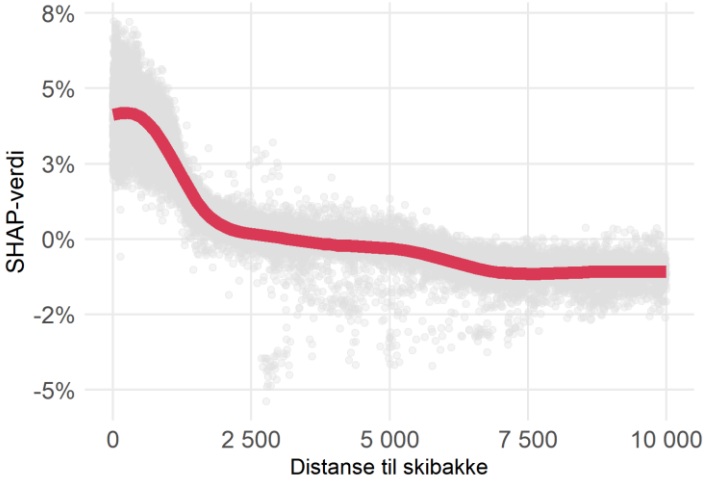


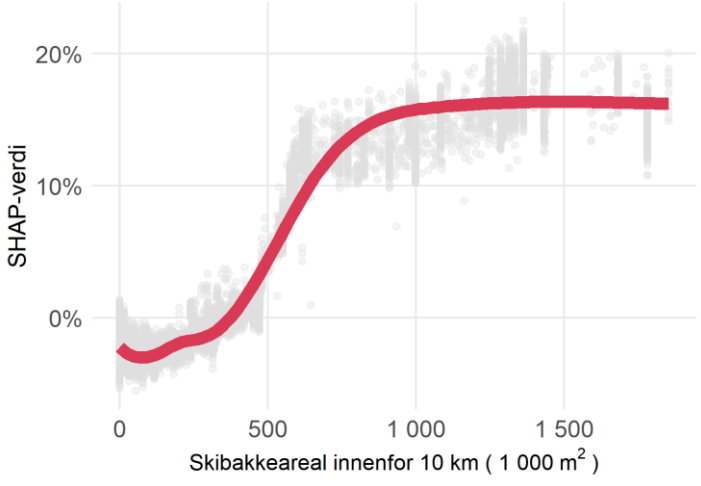


Beliggenhet ved innsjø

Distanse til innsjø gir et tilsvarende bidrag som distanse til sjø. Kort avstand til innsjø kan gi opp til 35 prosent økning i boligens markedsverdi. Bidraget er fallende med avstanden til innsjøen. Dersom boligen ligger 100 meter eller mer fra innsjøen har avstanden til sjø i praksis liten betydning. Hver meter tomtegrense mot innsjø gir en tilnærmet lineær økning i prosentvis markedsverdi opp til 50 meter, og flater deretter ut. Tomtegrense mot innsjø gir et relativt bidrag til markedsverdi tilsvarer om lag halvparten av tomtegrense mot sjø.

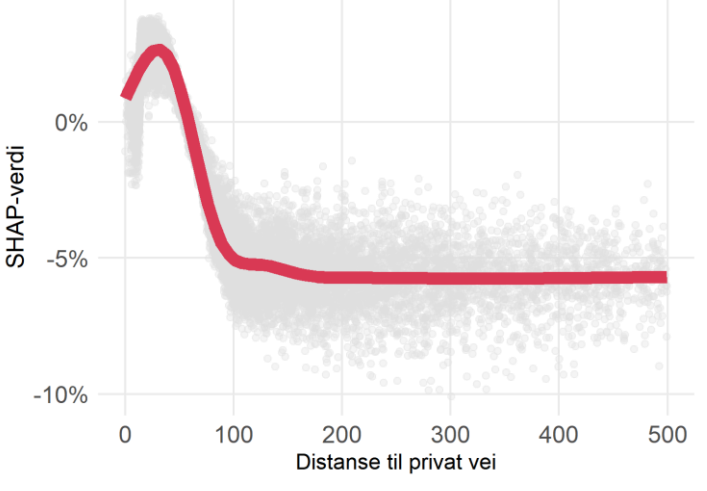
Beliggenhet ved skibakke og vei

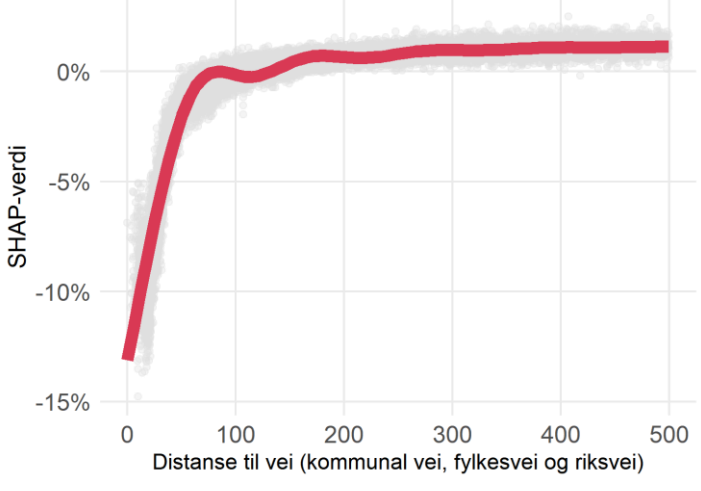




Beliggenhet ved skibakke

Beliggenhet ved skibakke gir et positivt bidrag til boligens markedsverdi. Bidraget avtar bratt med distansen fra bakken de første to kilometerene, og flater deretter ut. Nærhet til store skibakker (over 1 kvadratkilometer) kan gi et positivt bidrag til markedsverdien på opp til 20 prosent. Disse skibakkene inkluderer blant annet Voss og Hafjell Alpinsenter.





Beliggenhet ved vei

Kort avstand til privat vei gir et positivt bidrag til boligens markedsverdi. Dersom det er mer enn 100 meter til nærmeste private vei har det i praksis liten betydning hvor lang avstanden er. Kort avstand til riksvei, kommunal vei eller fylkesvei kan redusere markedsverdien med opp til 15 prosent. Dette indikerer at det er lite attraktivt å ligge svært nært en trafikkert vei.

8. Begrensninger og potensielle forbedringer

Begrensninger ved arbeidet

Datakvaliteten i matrikkelen setter begrensninger for modellens presisjon

I prosjektets innledende analyse ble det avdekket at en stor andel av fritidsboligene mangler utfylte verdier for flere relevante variabler i matrikkelen. Dette gjelder spesielt bruksareal og byggeår som er de viktigste variablene i verdsettelsesmodellen. Konsekvensen er at avviket mellom estimert og faktisk markedsverdi for boliger med ukjent bruksareal og byggeår er opp mot 11 prosent høyere enn for fritidsboliger med registrert bruksareal og byggeår.

Ved sammenligning mot FINN.no har det blitt avdekket store variasjoner i kvaliteten på bruksareal og byggesaksstatuser som er registrert i matrikkelen. Dersom skattytere får et feilaktig lavt eller høyt registrert bruksareal eller byggeår i modellen vil dette kunne medføre at modellen beregner feil markedsverdi. Det forventes at kvaliteten på bygningsopplysningene i matrikkelen vil forbedres over tid.

Verdsettelsesmodellen hensyntar ikke boligens bygningsstandard

Modellen er basert på offisielt registrerte bygningsopplysninger i matrikkelen. Dette innebærer at den eneste informasjonen modellen har om boligens bygningsstandard er det registrerte byggeåret i matrikkelen. Boligene med størst avvik fra markedsverdi i modellen holder gjennomgående veldig lav eller høy standard. Dette indikerer at det er potensiale for å forbedre modellen gjennom taksering eller egenrapportering av standard.

Ikke alle fritidsboliger omsettes gjennom FINN.no

Fritidsboliger utgjør en mindre andel av det norske boligmarkedet. Dette innebærer at datasettet som brukes for å trene modellen er vesentlig mindre enn for tilsvarende modell for primærboliger. Videre har prosjektet begrenset treningsgrunnlaget til å kun gjelde fritidsboliger som er publisert på FINN.no i perioden 2010-2020, hvor både matrikkelinformasjon og faktisk salgspris er registrert. Denne begrensningen er ansett som nødvendig for å sikre at fritidsboligene i utvalget er omsatt til markedspris, men begrenser modellens evne til å fange opp boliger som omsettes utenfor FINN.no. Dette gjelder spesielt boliger med salgspris under 500 tusen kroner, hvor utvalget utgjør 16 prosent av alle omsetninger.

Ytterligere arbeid kan forbedre treffsikkerheten for den lineære modellen

Maskinlæringsmodeller pekte seg tidlig ut som en god kandidat som tilfredsstilte verdsettelsesmodellens vurderingskriterier. For å belyse hvordan maskinlæringsmodellen presterer i forhold til en tradisjonell tilnærming har prosjektet utviklet en lineær modell som benyttes som sammenligningsgrunnlag. Den lineære modellen har en enkel funksjonsform inspirert av SSBs verdsettelsesmodell. Modellen skiller seg fra SSBs tilnærming ved at den estimerer fritidsboliger for hele landet, fremfor å bruke regionale modeller.

Det kan ikke utelukkes at ytterligere arbeid med den lineære modellen kunne hevet kvaliteten på modellens verdsettelser. Arbeidet kunne eksempelvis involvert å se nærmere på ulike regionale inndelinger, eller klassifisere boligene i ulike boligtyper basert på om de er lokalisert i fjellet eller ved sjøen. Erfaringer fra SSBs tilnærming har vist at modellen for primær- og sekundærboliger legger stor vekt på inndelingen i kommunale prissoner. Dette gir potensielt utfordringer med store variasjoner i verdsettelser når fylker og kommuner sammenslås. En annen utfordring ved regionale modeller er det begrensede omfanget av omsatte boliger.

Prosjektet har utforsket noen ulike tilnærminger for den lineære modellen uten at dette har gitt vesentlige forbedringer i modellens treffsikkerhet. Det har blant annet blitt utviklet lineære modeller basert på LASSO-regresjon (*Least absolute shrinkage and selection operator*) med flere interaksjonsledd. MARS (*multivariate adaptive regression splines*) er et eksempel på en lovende metode for å fange opp ikke-lineære sammenhenger som kunne vært utforsket i sammenheng med dypere interaksjoner. Samtidig er prosjektets erfaring at en lineær modell fremdeles vil gi en vesentlig lavere treffsikkerhet enn maskinlæringsmodellen.

Potensielle forbedringer

En nasjonal modell for verdsettelse av alle private boliger og eiendommer

Verdsettelsesmodellens treffsikkerhet begrenses av omfanget av omsatte fritidsboliger. Fritidsboliger deler mange av de samme prisdriverne som primær- og sekundærboliger som bruksareal og alder, men modellen utnytter ikke disse likhetene i dag. En nasjonal modell for alle private boliger og eiendommer vil gi modellen maksimal tilgang til slike sammenhenger, og vil også kunne dekomponere verdiene i boligverdi og tomteverdi.

Innhenting av boligopplysninger fra eiere av fritidsboliger

Rapporten har tidligere beskrevet utfordringene med kvaliteten på registrerte bygningsopplysninger i matrikkelen. Et tiltak for å heve kvaliteten på innholdet kan være å be boligeeiere rapportere og verifisere opplysninger om bruksareal, byggeår, tomtestørrelse, bygningsstandard og innlagt strøm og vann.

Kompletterende datakilder og videreutvikling av beregninger

Prosjektet har benyttet offisielt registrerte bygningsopplysninger i matrikkelen og grunnboken. Det er et potensiale for å utnytte andre kompletterende kilder som kommunenes egne eiendomsopplysninger og Skatteetatens registre. Videre er det mulig å videreutvikle beregningen av de ulike geovariablene for å blant annet ta høyde for reiseavstander. Tabellen viser en ikke-uttømmende liste med potensielle utvidelser.

Variabel	Potensiell utvidelse
Byggeår	Mange boliger mangler informasjon om byggeår. Ved bruk av historiske flyfoto kan det være mulig å gi indikasjoner på boligenes byggeår, men prosjektet har ikke undersøkt de tekniske mulighetene for dette i detalj.
Energikilde	For å komplettere data om energikilde for fritidsboliger kan Elhub-data benyttes. Elhub inneholder data for alle norske strømmålere og er en datakilde som SSB har pilotert for bruk i statistikk. Pilotrapporten indikerer at strømmålerene kan kobles til fritidsboliger ved bruk av adresser eller koordinater.
Adkomst	Adkomst kan utvides med informasjon om boligen har fastlandsforbindelse, og hvor mange innbyggere som er innenfor 3 timers reiseavstand.
Skibakker	Skibakke-variablene kan utvides med informasjon som sier noe om anleggets attraktivitet. Dette kan for eksempel være antall heiser, åpningstider, gjennomsnittlig snødybde, antall nedfarter, totalt antall høydemeter mellom topp og bunn, lengde på bakke, type løyper eller priser på heiskort.
Turruter	Turrutevariablene kan utvides med variabler som beskriver rutenes attraktivitet. Eksempelvis kunne det vært ønskelig å si noe om hvor ofte langrennsløyper prepareres, eller om turstiene har belysning.
Solforhold	Solforhold er beregnet basert på Kartverkets høydedata med en oppløsning på 10 x 10 meter. Det pågår utvikling av en nasjonal overflatemodell med oppløsning på 1 x 1 meter. Modellen har foreløpig ikke fullstendig nasjonal dekning, men vil kunne gi mer nøyaktige beregninger av solforhold
Utsikt	Beregnes basert på samme overflatemodell som solforhold, og kan tilsvarende utvides med å benytte den mer nøyaktige overflatemodellen med oppløsning på 1 x1 meter.

Ikke-uttømmende liste over mulige forbedringer ved dagens variabler

A. Appendiks

Appendiks A

Prosess

Prosess

Prosjektet er eid av Finansdepartementet og har blitt ledet av PwC og gjennomført som et samarbeid mellom Finansdepartementet, PwC og SSB. SSB er ansvarlig for å forvalte modellen og årlige oppdateringer av estimerte markedsverdier for populasjonen av fritidsboliger. I løpet av oppstartsmøtet 22. oktober 2019 ble det definert en prosjektplan med tre faser:

Fase 1: Valg av variabler

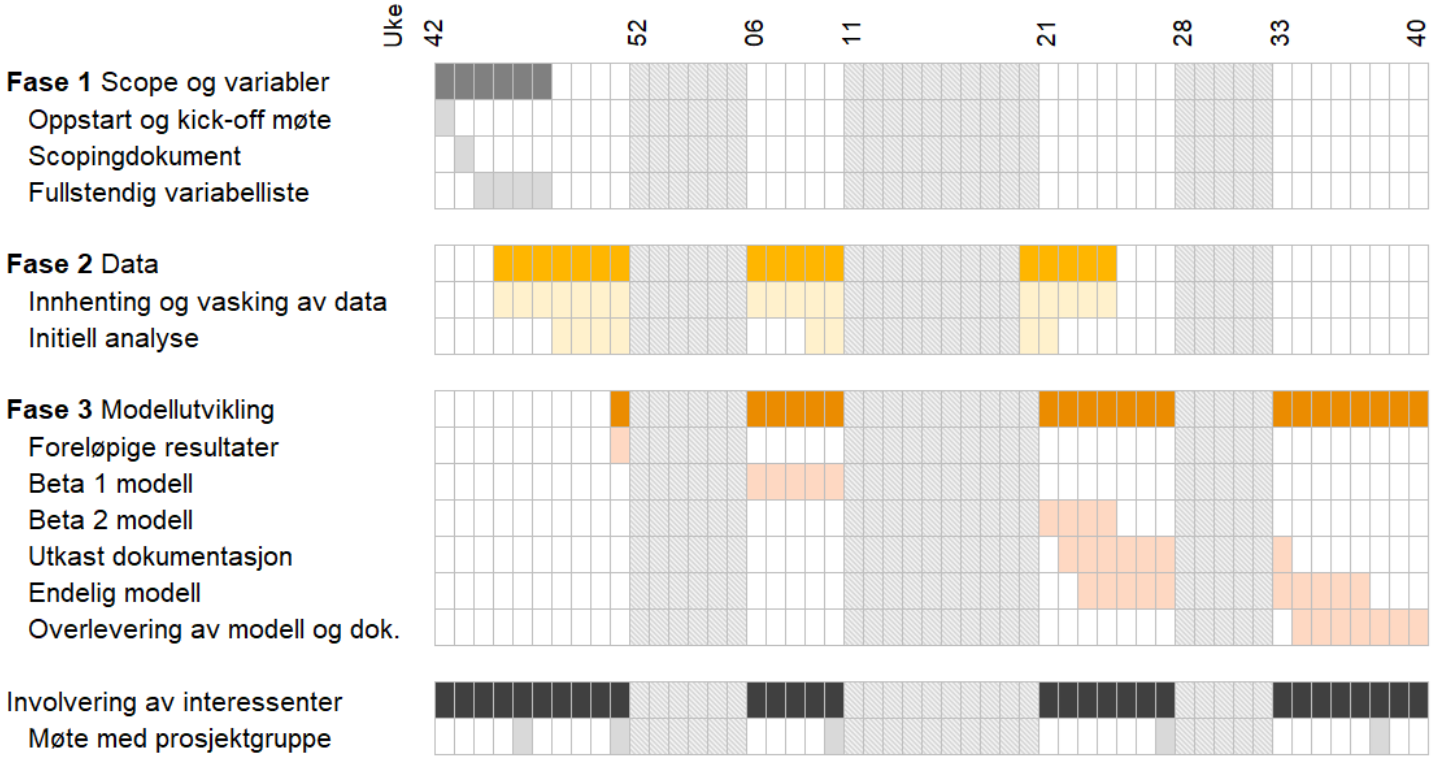
I den innledende fasen ble det avholdt en samling med deltakere fra Finansdepartementet, PwC og SSB. Formålet med samlingen var å definere en liste med boligegenskaper som forventes å påvirke markedsverdi. Metodikk for valg av variabler og liste over inkluderte variabler er beskrevet i seksjon 3.

Fase 2: Innhenting og analyse av data

Basert på samlingen i fase 1 ble det besluttet å innhente data for utvalgte variabler. Ved prosjektets oppstart var det forventet at SSBs datatilganger til matrikkel og grunnbok kunne benyttes av PwC, men prosjektet måtte til slutt inngå egne databehandleravtaler med Kartverket. Dette innebar en frysperiode i januar i påvente av datatilgang. Deretter ble det gjort innledende vurderinger av datakvaliteten i materialet.

Fase 3: Utvikling av endelig modell

I siste fase ble det utviklet to utkast til modeller med fokus på treffsikkerhet i verdsettelser. Gjennom arbeidet ble det avdekket utfordringer med å identifisere hvilke av fritidsboligene i grunnboken som var omsatt til markedsverdi. Det ble derfor besluttet å innhente annonserte fritidsiendommer fra FINN.no for tidsrommet 2010 til 2020, for å kunne identifisere faktiske markedssalg. Dette medførte en ny frysperiode fra uke 11 til uke 21 i påvente av data fra FINN.no. Når dataene ble mottatt kunne prosjektet utvikle endelig modell og rapport. Endelig versjon av rapporten ble levert 17. november 2020.



Prosjektets endelige prosjektplan med tilhørende faser

Appendiks B

Modellvariabler

Modellvariabler

Salgspris (Kilde: Grunnboken)

Kategori	Utfallsvariabel	Beskrivelse
Salgspris	Salgspris	Beløpet en fritidsbolig i treningsdataene ble omsatt for i det frie markedet. Dette er utfallsvariabelen modellen forsøker å estimere.
Salgstidspunkt	Solgt år	Året fritidsboligen sist ble omsatt. Brukes for å fange opp årlige variasjoner i markedsverdi når modellen trenes. Når modellen skal estimere markedsverdier i forbindelse med skatteoppgjør vil solgt år settes lik året for skatteoppgjøret.
Salgstidspunkt	Solgt måned	Måneden fritidsboligen sist ble omsatt, brukes for å fange opp sesongvariasjoner når modellen trenes. Når modellen skal estimere markedsverdier i forbindelse med skatteoppgjør vil solgt måned settes til januar.

Strukturelle variabler (Kilde: Matrikkelen)

Kategori	Modellvariabel	Beskrivelse
Bruksareal	Bruksareal	Totalt bruksareal for bygg på matrikkelenheten som enten er kategorisert som fritidsbolig eller bolig i henhold til SSB sin standard for bygningstype.
Byggeår	Byggeår	Året bygget ble ferdigstilt. Variabelen baserer seg på datoen bygget ble tatt i bruk. Fordi det er en del forsinkelser i registrering, brukes henholdsvis datoen bygging ble igangsatt eller byggetillatelse gitt dersom tatt i bruk dato mangler.
Eiendomstype	Eiendomstype	Matrikkelenhets eiendomstype. Dette kan enten være grunneiendom, festegrunn eller eierseksjon.
Andre bygg	Antall boliger	Totalt antall bruksenheter tilhørende matrikkelenheten.
Andre bygg	Antall fritidsbygg	Totalt antall fritidsbygninger for matrikkelenhet..
Andre bygg	Bruksareal andre bygg	Totalt bruksareal for bygg på matrikkelenheten som ikke er kategorisert som fritidsbolig eller bolig i henhold til SSB sin standard for bygningstype. Dette kan typisk være garasjer, anneks eller lignende.
Tomteareal	Tomteareal	Tomtens totale areal. Arealet er basert på beregnet areal registrert i matrikkelen. I de tilfellene beregnet areal ikke er registrert brukes historisk oppgitt areal registrert i matrikkelen.
Tomtehelning	Tomtehelning	Angir hvor bratt terrenget på tomten er, oppgitt i grader. Variabelen oppgir minimum og maksimum helning for 75 prosent av tomtens areal (25- og 75-persentil). Beregnet basert på tomtens geometri fra matrikkelen.
Tomtehelning	Retning på tomtehelning	Angir kompassretningen tomtens helning peker mot. Variabelen oppgir minimum og maksimum retning for 75 prosent av tomtens areal (25- og 75-persentil). Beregnet basert på tomtens geometri fra matrikkelen.
Vann og avløp	Avløpskode	Kode for byggets kloakksystem. Bygget kan enten være tilknyttet offentlig kloakk, privat kloakk eller ingen tilknytning.
Vann og avløp	Vannforsyningskode	Kode for byggets vannforsyning. Bygget kan være tilknyttet offentlig vannverk, privat vannverk, ha privat innlagt vann eller ikke ha innlagt vann.
Energi og varme	Energikildekode	Kode for byggets energikilde. Det kan være enten elektrisitet, olje/parafin, solenergi, varmepumpe, gass, biobrensel eller annen energikilde.
Energi og varme	Oppvarmingskode	Kode for hvordan bygget varmes opp. Bygget kan ha sentralvarme, elektrisk varme eller annen oppvarming.

Modellvariabler

Lokasjonsvariabler

Kategori	Modellvariabel	Kilde	Beskrivelse
Posisjon	Lengdegrad	Matrikkelen	Byggets lengdegrad.
Posisjon	Breddegrad	Matrikkelen	Byggets breddegrad.
Adkomst	Avstand til vei	Kartverket (N50)	Avstand fra boligen til nærmeste offentlige vei, målt i luftlinje. Inkluderer kommunal vei, fylkesvei, riksvei og europavei.
Adkomst	Høydeforskjell til offentlig vei	Kartverket (N50/Hoydedata.no)	Differanse i høydemeter mellom boligen og nærmeste offentlige vei
Adkomst	Veitype for nærmeste offentlige vei	Kartverket (N50)	Inkluderer kommunal vei, fylkesvei, riksvei og europavei.
Adkomst	Avstand til privat vei	Kartverket (N50)	Avstand fra bygget til nærmeste private vei, målt i luftlinje.
Adkomst	Høydeforskjell til privat vei	Kartverket (N50/Hoydedata.no)	Differanse i høydemeter mellom boligen og nærmeste private vei
Høyde	Høydemeter	Kartverket (Hoydedata.no)	Høydemeter for byggets plassering i terrenget.
Naboer	Fritidsboliger innenfor 1 km	SSB	Antall andre fritidsboliger innenfor en radius på én kilometer målt i luftlinje fra boligen.
Naboer	Fritidsboliger innenfor 3 km	SSB	Antall andre fritidsboliger innenfor en radius på tre kilometer målt i luftlinje fra boligen.
Naboer	Primærboliger innenfor 1 km	SSB	Antall primærboliger innenfor en radius på én kilometer målt i luftlinje fra boligen.
Sjø	Avstand til sjø	Kartverket (N50)	Korteste avstand i luftlinje fra boligen til sjøen.
Sjø	Lengde på strandlinje	Kartverket (N50)	Antall meter av tomten som grenser mot sjø.
Innsjø	Avstand til innsjø	Kartverket (N50)	Korteste avstand i luftlinje fra fritidsboligen til nærmeste innsjø.
Innsjø	Lengde på innsjølinje	Kartverket (N50)	Antall meter av tomten som grenser mot innsjø.
Skibakker	Avstand til skibakke	Kartverket (N50)	Korteste avstand målt i luftlinje fra boligen til nærmeste skibakke.
Skibakker	Skibakkeareal innenfor 10 km	Kartverket (N50)	Antall kvadratmeter med skibakke innenfor en radius på ti kilometer målt i luftlinje fra boligen.

Modellvariabler

Lokasjonsvariabler

Kategori	Modellvariabel	Kilde:	Beskrivelse
Turruter	Avstand til turrute	Kartverket	Korteste avstand målt i luftlinje fra boligen til nærmeste turrute.
Turruter	Lengde på turruter innenfor 1 km	Kartverket	Antall meter med turruter innenfor en radius på én kilometer i luftlinje fra boligen.
Innbyggere	Innbyggere innenfor 1 km	SSB	Antall fastboende innenfor radius av én km fra bygget.
Innbyggere	Innbyggere innenfor 5 km	SSB	Antall fastboende innenfor radius av fem km fra bygget.
Innbyggere	Innbyggere innenfor 50 km	SSB	Antall fastboende innenfor radius av 50 km fra bygget.
Innbyggere	Innbyggere innenfor 100 km	SSB	Antall fastboende innenfor radius av 100 km fra bygget.
Solforhold	Helning horisont	Kartverket (Hoydedata.no)	Gjennomsnittlig vinkel på horisont basert på omkringliggende terreng innenfor 10 km
Solforhold	Årlige soltimer	Kartverket (Hoydedata.no)	Totalt antall soltimer i året.
Utsikt	Synlig areal	Kartverket (Hoydedata.no)	Totalt synlig areal og synlig areal nedenfor bygg innenfor en radius av henholdsvis én, fem og ti kilometer fra bygget.
Utsikt	Synlig sjø/innsjøareal	Kartverket (Hoydedata.no)	Totalt synlig sjø- og innsjøareal innenfor en radius av henholdsvis én, fem og ti kilometer i luftlinje fra bygget.
Admin.enhet	Fylke	Matrikkelen	Fylket boligen tilhører.
Admin.enhet	Kommune	Matrikkelen	Kommunen boligen tilhører.

Appendiks C

Andre vurderte variabler

Andre vurderte variabler

I første fase av prosjektet ble det gjennomført et omfattende arbeid for å identifisere en bruttoliste over potensielle variabler. Hver variabel ble prioritert basert på forventet forklaringskraft og kriteriene tilgjengelighet, enkelhet og aksept. Data for de prioriterte variablene innhentet for alle fritidsboliger. Den reelle forklaringskraften for hver variabel ble så testet, for å ende opp med et endelig utvalg av variabler. Prosessen for valg av variabler er nærmere beskrevet i *Kapittel 3 - Variabler og datakilder*. I dette appendikset oppsummeres variablene fra bruttolisten som ikke ble inkludert i modellen. Variablene grupperes etter begrunnelsen for ekskludering av variabelen. Det kan være flere årsaker til at en variabel er ekskludert, men den er bare listet én gang i denne oversikten. Felles for alle variablene er antagelsen om lav prediksjonskraft.

Kategori 1: Variabler med lav reell prediksjonskraft

Denne kategorien inkluderer alle variabler som ble prioritert høyt nok til at data for variabelen ble innhentet og testet, men som ikke ble inkludert i den endelige modellen på grunn av lav prediksjonskraft. Dette kan for eksempel skyldes at effekten av variabelen plukkes opp av andre variabler eller at variabelen ikke har like stor forklaringskraft som først antatt.

Variabel	Beskrivelse
Skibakkeareal innenfor x km	Tilsvarende variabel som skibakkeareal innenfor 10 kilometer som er inkludert i modellen ble beregnet for 1, 5 og 20 kilometer.
Lengde på turrute innenfor x km	Tilsvarende variabel som turrute innenfor 1 kilometer som er inkludert i modellen ble beregnet for 3, 5 og 10 kilometer.
Bebyggelse innenfor 1 km	Antall industribygg, kulturbygg, helsebygg og fengsels- og beredskapsbygg i henhold til SSBs definisjoner, innenfor 1 kilometer fra fritidsboligen.
Avstand til by	Korteste avstand til nærmeste sentrumssone i henhold til SSB's definisjon, målt i luftlinje.
Avstand til vei for ulike veityper	Tilsvarende variabel som avstand til offentlig og privat vei som er inkludert i modellen, for hver av veitypene europavei, fylkesvei, kommunal vei og riksvei.
Avstand til jakt og fiske	Korteste avstand målt i luftlinje fra fritidsboligen til nærmeste område for henholdsvis innlandsfiske, laksefiske, småviltjakt og storviltjakt basert på inatur.no
Areal av øy	Areal av øyen fritidsboligen ligger på, dersom den ligger på en øy.
Høyde bygning	Høyden på fritidsboligen oppgitt i meter basert på Kartverkets overflatemodell
Punktfeste	Angir om tomten er et punktfeste som gir disposisjonsrett til punktet bygget står på eller tomtefeste med definerte grenser, dersom tomten er en festegrunn.
Antall andre bygg	Antall garasjer, uthus og anneks og totalt antall andre bygg på tomten fritidsboligen står på.
Bygningstyper på tomt	Variabler som angir om det er henholdsvis garasje, rorbu eller naust, enebolig, våningshus eller landbrukshus etter SSB's definisjoner på tomten fritidsboligen står på
Delområde og grunnkrets	Delområde og grunnkrets fritidsboligen tilhører.

Andre vurderte variabler

Kategori 2: Variabler utelatt på grunn av forklarbarhet

I *Kapittel 5 - Modelleringsvalg* ble det forklart at variabelen bruksareal ble slått sammen til én variabel, blant annet for at modellen skulle være mer forklarbar for skattyter. Av samme grunn er enkelte lignende variabler ekskludert fra modellen. Eksempelvis vil ikke skattyter nødvendigvis ha et forhold til forskjellen mellom bruksareal og bruttoareal. I tillegg er disse variablene gjerne korrelert og brukes til å imputere verdien der hvor bruksareal mangler.

Variabel	Beskrivelse
Bebygd areal	Areal av grunnflaten til en fritidsbolig.
Bruttoareal	Totalt areal omsluttet av fritidsboligens yttervegger.

Kategori 3: Variabler hvor det antas at effekten plukkes opp av andre variabler

Effekten av enkelte variabler kan plukkes opp av andre mer generelle variabler. Eksempelvis kan enkelte fritidsboliger ligge i områder med stor rasfare. En slik effekt kan føre til lavere markedsverdier. Effekten kan indirekte fanges opp av lengde- og breddegrad, ved at modellen lærer seg at verdien er lavere i enkelte områder uten at det eksplisitt kommer frem at årsaken er stor rasfare. Variabler hvor det i stor grad antas at effekten plukkes opp av andre variabler er ekskludert fra modellen.

Variabel	Beskrivelse
Avstand til nærmeste nabo	Korteste avstand mellom fritidsboligen og nærmeste bygg (fritidsbolig eller primærbolig), målt i luftavstand.
Gjennomsnittlig snødybde	Gjennomsnittlig snødybde i vinterhalvåret for lokasjonen.
Antall regndager	Gjennomsnittlig antall regndager per år for lokasjonen.
Avstand til nasjonalpark	Korteste avstand til nærmeste nasjonalpark, målt i luftavstand.
Faresone	Indikasjon på om området er ras- eller flomutsatt.
Forurensning og støy	Variabler som indikerer forurensing (eksempel: svevestøv eller utslipp fra vedfyring) og støy (eksempel: støy fra vei, lufthavn eller skytebane). Flere eksempler finnes på https://miljoatlas.miljodirektoratet.no/.

Andre vurderte variabler

Kategori 4: Variabler som er vanskelig å innhente for alle fritidsboliger

Dersom dataene som ligger til grunn for en variabel ikke er tilgjengelig gjennom åpne grensesnitt eller har lav dekningsgrad vil det både være ressurskrevende å innhente dataene og beregne variabelen for alle fritidsboliger. Eksempel på dette kan være boligens fasiliteter, som for eksempel peis, som ikke registreres i et offentlig register, men som kan fremkomme i boligens salgsoppgave.

Variabel	Beskrivelse
Offentlig transport	Korteste avstand til offentlig transport og hyppighet av forbindelsen.
Antall rom*	Antall rom boligen består av, fordelt på type rom som for eksempel antall soverom og antall bad.
Balkong/terrasse	Om fritidsboligen har tilhørende balkong eller terrasse.
Avstand til båtplass	Korteste avstand til offentlige båtplass, målt i luftavstand.
Tilhørende fasiliteter	Tilgjengelige fasiliteter tilhørende fritidsboligen, som for eksempel peis, basseng eller privat båtplass.
Mobildekning og internettilgang	Kvalitet på mobildekning (eksempel: edge, 3G, 4G eller 5G) og tilgjengelig internett (eksempel: bredbånd eller fiber).
Tilstandsgrad	Tilstandsgrad på bad og kjøkken

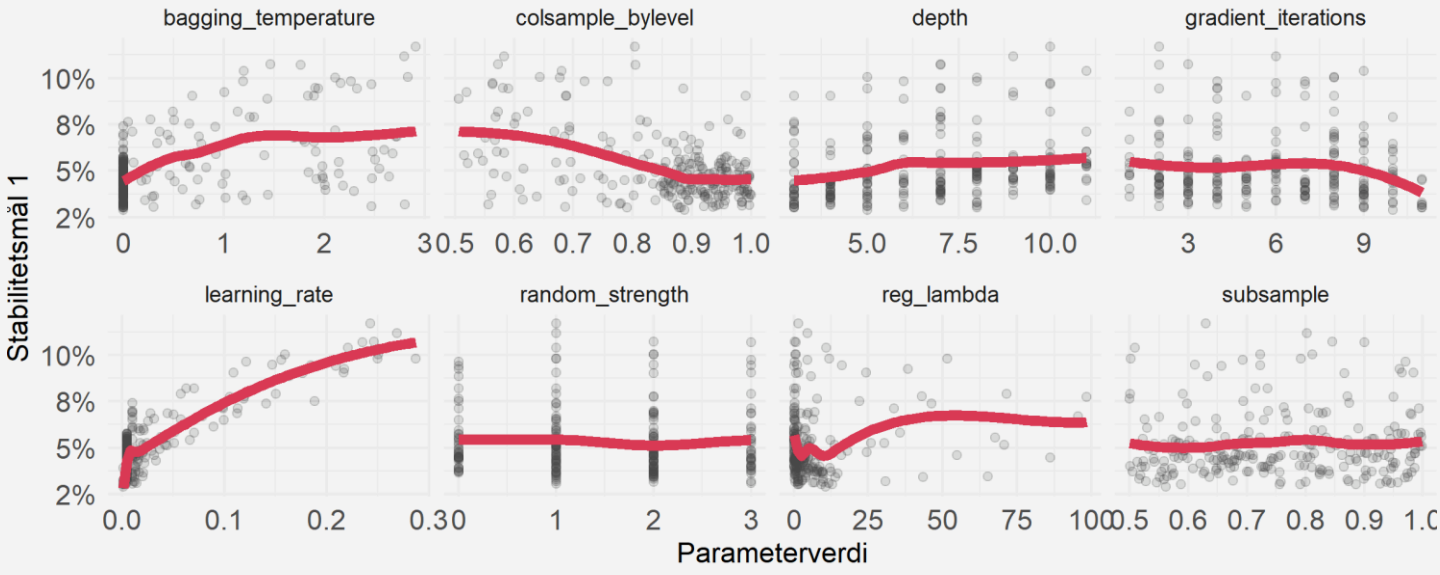
*Det er egne felt i matrikkelen som registrerer antall rom, bad og wc for bruksenheter, men variablene har per dags dato så lav utfyllingsgrad og kvalitet at de ikke kan brukes.

Appendiks D

Avveining mellom presisjon og stabilitet for trebasert modell

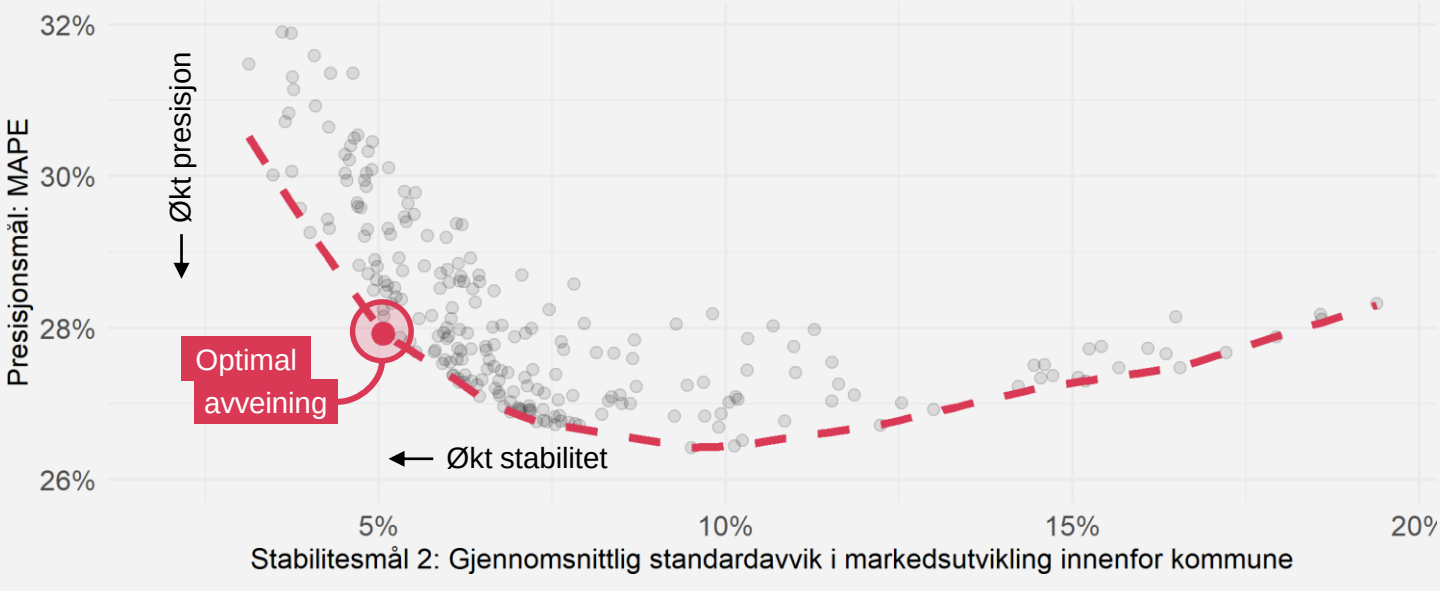
Optimal avveining mellom presisjon og stabilitet

For å identifisere den optimale avveiningen mellom presisjon og stabilitet er det gjennomført et parametersøk for den trebaserte modellen. Parametersøket gjennomføres ved å simulere ulike verdier for parametrene som regulerer modellens kompleksitet. Deretter evalueres modellens presisjon og stabilitet for de ulike kombinasjonene av parametre.



Ulike kombinasjoner av parameterverdier og tilhørende påvirkning på Stabilitetsmål 1.

Avveiningen mellom presisjon og stabilitet blir tydelig når presisjonsmålet sees i sammenheng med stabillitets-målene. Prosjektet har valgt å benytte stabiliteten for den lineære modellen som det førende kriteriet for en tilstrekkelig stabil modell. [1] Dette valget innebærer at den trebaserte og lineær modellene oppnår samme stabilitet.



Optimal avveining av presisjon og stabilitet for den trebaserte modellen. Siden prosjektet har benyttet en mindre kryssvalidering (3 folder) i parametersøket er presisjonen i figuren noe lavere enn reell presisjon

1) De to stabilitetsmålene er svært tett korrelerte og det er derfor vilkårlig hvilket av disse som benyttes for å finne riktig avveining.

Referanser

Aas, K., Jullum, M., Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values

Antipov, E., Pokryshevskaya, E. (2010). Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics

Doshi-Velez, F., Finale Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning

Gloudemans, R.J., Eckert, J., Dotson, M. (1983). Computer Assisted Appraisal Systems. Legends of CAMA, Lincoln Institute of Land Policy, Massachusetts

Gloudemans, R. J., Montgomery, E. J. (2008). Domestic revaluation in Northern Ireland. Journal of Property Tax Assessment & Administration, 5(4), 17-31.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R.

Hong, J., Choi, H., Kim, W. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24(3), 140-152.

International Association of Assessing Officers (2017). Standard on Mass Appraisal of Real Property

International Association of Assessing Officers (2018). Standard on Automated Valuation Models (AVMs)

Kolbe, J., Schulz, R., Wersing, M., Werwatz, A. (2019). Land value appraisal using statistical methods, FORLand-Working Paper, No. 07 (2019)

Lipton, P. (1990). Contrastive Explanation

Lundberg, S. Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4768–4777.

McCluskey, W.J., Zulkarnain, D, Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia"

Molnar, C., Casalicchio, G, Bernd, B. (2020). Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges

O'Connor, P.M. (2002). Comparison of Three Residential Regression Models: Additive, Multiplicative and Nonlinear. Assessment Journal July/August 2002: 37–43.

Rossini, P., Kershaw, P. (2006). Can a Hybrid Automated Valuation Model Outperform Individually Assessed Capital and Site Values.

Silverherz, J. (1936). The assessment of real property in the United States

Takle, M., Medby, P. (2019). Modell for beregning av boligformue: Oppdatert med tall for 2018.

Wu, A., Gloudemans R.J. (2019). Machine Learning Algorithms: The Future of Property Valuation

Programvare og biblioteker

All programvare og biblioteker som er benyttet i prosjektet er basert på åpen kildekode:

Bivand et al., (2020). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package
https://CRAN.R-project.org/package=rgdal

Dorogush et al., (2018). CatBoost: gradient boosting with categorical features support.
https://arxiv.org/abs/1810.11363

Friedman et al., (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
http://www.jstatsoft.org/v33/i01/

Hijmans, R.J., (2020). raster: Geographic Data Analysis and Modeling. R package version 3.3-13.
https://CRAN.R-project.org/package=raster

Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.
https://www.tidymodels.org

Lundberg S., (2020). shap: A unified approach to explain the output of any machine learning model.
https://pypi.org/project/shap/

Pebesma, E., (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1)
https://doi.org/10.32614/RJ-2018-009

Pebesma, E., (2020). stars: Spatiotemporal Arrays, Raster and Vector Data Cubes. R package version 0.4-3.
https://cran.r-project.org/web/packages/stars/index.html

Python Software Foundation (2020). The Python Language Reference — Python 3.9.0
https://docs.python.org/3/reference/

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA
http://www.rstudio.com/

Therneau et al., (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15.
https://CRAN.R-project.org/package=rpart

Vaughan et al., (2020). furrr: Apply Mapping Functions in Parallel using Futures. R package version 0.2.1.
https://CRAN.R-project.org/package=furrr

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686,
https://doi.org/10.21105/joss.01686

pwc.com

Denne rapporten er utarbeidet for Finansdepartementet i samsvar med signert kontrakt, datert 25 .september 2019. PwC har ikke utført noen form for revisjon eller kontrollhandlinger av Finansdepartementets virksomhet i dette oppdraget. Finansdepartementet har rett til å benytte informasjonen i denne rapporten i sin virksomhet, i samsvar med forretningsvilkårene som er definert i kontrakten. Rapporten og/eller informasjon fra rapporten skal ikke benyttes for andre formål uten skriftlig samtykke fra PwC. Enhver handling som gjennomføres på bakgrunn av vår rapport foretas på eget ansvar.

© 2020 PwC. Med enerett. I denne sammenheng refererer "PwC" seg til PricewaterhouseCoopers AS, Advokatfirmaet PricewaterhouseCoopers AS, PricewaterhouseCoopers Accounting AS, PricewaterhouseCoopers Skatterådgivere AS og PricewaterhouseCoopers Services AS som alle er separate juridiske enheter og uavhengige medlemsfirmaer i PricewaterhouseCoopers International Limited.