



# Nye beregningsområder i boligformuesmodellen

TALL

SOM FORTELLER

NOTATER / DOCUMENTS

2023/3

Erlend Eide Bø og Odd Erik Nygård

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

© Statistisk sentralbyrå

Publisert: 3. februar 2023

ISBN 978-82-587-1671-3 (elektronisk)

ISSN 2535-7271 (elektronisk)

<b>Standardtegn i tabeller</b>	<b>Symbol</b>
<b>Ikke mulig å oppgi tall</b> Tall finnes ikke på dette tidspunktet fordi kategorien ikke var i bruk da tallene ble samlet inn.	.
<b>Tallgrunnlag mangler</b> Tall er ikke kommet inn i våre databaser eller er for usikre til å publiseres.	..
<b>Vises ikke av konfidensialitetshensyn</b> Tall publiseres ikke for å unngå å identifisere personer eller virksomheter.	:
<b>Desimaltegn</b>	,

## Forord

Den eksisterende boligverdsettingsmodellen utarbeidet av SSB har vist seg å være lite robust med hensyn til omfattende endringer i kommuneinndelingen. På oppdrag fra Finansdepartementet har SSB gjennomført et prosjekt for å undersøke hvordan verdsettelse av boliger i større grad kan frigjøres fra kommunegrensene. Dette prosjektet er en videreutvikling av det begrensede forprosjektet beskrevet i SSB Notater 2022/11. Deler av metodebeskrivelsen og resultatene i dette notatet er derfor hentet fra SSB Notater 2022/11.

Det oppdaterte notatet er skrevet av Erlend Eide Bø og Odd Erik Nygård. Per Medby og Mona Takle var medforfattere ved forrige utgave, SSB Notater 2022/11. Erik Engelién og Sturla Løkken har gitt nyttige bidrag underveis.

Statistisk sentralbyrå, 25. januar 2023

Linda Nøstbakken

## Sammendrag

SSBs eksisterende modell for beregning av boligformue har vist seg å være lite robust med hensyn til endringer i kommueinndelingen. Dette notatet beskriver mulige endringer i modell for verdsettelsesmodellen for å frigjøre seg fra kommunestrukturen, og hvor mye treffsikkerheten kan forbedres ved disse endringene.

Statistisk sentralbyrå (SSB) har tidligere utviklet en modell for å beregne markedsverdier for boliger som ble tatt bruk i 2009, og siden har blitt brukt til å beregne formuesgrunnlaget for boliger. Flere kommuner bruker også dette formuesgrunnlaget ved utskrivning av eiendomsskatt.

I den eksisterende modellen beregnes boligverdi separat for forskjellige regioner og boligtyper (beregningssområde). De enkelte beregningsområdene er inndelt i soner basert på prisnivå. Hvert beregningsområde deles opp i prissoner, der hovedregelen er at soner består av hele kommuner. Kommunereformen i 2020 førte til at modellen treffer dårligere enn før, fordi kommuner i forskjellige soner ble slått sammen.

Vi utarbeider her en modell basert på grunnkretser, som er betraktelig mindre enn kommuner. Vi bruker en maskinlæringsalgoritme for å gruppere grunnkretsene til beregningsområder med tilstrekkelig antall observasjoner. De 14 000 grunnkretsene i Norge deles opp i ca. 1 200 beregningsområder, separat for hver boligtype. For hver kombinasjon av geografisk område og boligtype kjøres så en separat hedonisk regresjon for å predikere boligverdi.

Vi diskuterer også kort andre metoder som kan brukes for å bedre utnytte informasjonen som er tilgjengelig i dagens data.

Den grunnleggende avveiningen ved aggregering av grunnkretser er at vi ønsker å skape små områder som er så homogene som mulig, samtidig som vi er avhengig av et visst antall omsetninger i den hedoniske regresjonen. Det resulterer i mange, geografisk små beregningsområder i sentrale strøk, og geografisk større beregningsområder i mindre sentrale strøk. Slik utnyttes den store mengden informasjon i byer til bedre å predikere lokale prisnivåer.

Vi viser resultater fra grunnkretsmodellen, og sammenligner dem med resultater fra den eksisterende modellen for både 2019 og 2020. Grunnkretsmodellen gir tydelige forbedringer i prediksjonskraften for beregnet boligverdi. Dette gjelder for alle fylker, og for fire forskjellige mål på prediksjonskraft. I tillegg predikerer grunnkretsmodellen boligverdien for dyre boliger i Oslo mye bedre. Dette har vært en kjent svakhet ved den eksisterende modellen. Resultatene for den nye modellen er svært like for 2019 og 2020, noe som tyder på at sammenslåing av kommuner har liten effekt på modellen, og at den er relativt stabil over tid.

Notatet avsluttes med en diskusjon av resultatene, modellens robusthet, samt en gjennomgang av forskjellige problemstillinger man må ta hensyn til hvis man skal bruke grunnkretsmodellen for beregning av boligverdi for hele boligmassen.

# Innhold

<b>Forord</b> .....	<b>3</b>
<b>Sammendrag</b> .....	<b>4</b>
<b>1. Bakgrunn</b> .....	<b>6</b>
1.1. Effekter av regionreformen.....	7
1.2. Andre svakheter ved modellen.....	8
<b>2. Ny modell</b> .....	<b>10</b>
2.1. Maskinlæring på grunnkretser .....	10
2.2. Alternative metoder .....	11
<b>3. Resultater</b> .....	<b>14</b>
3.1. 2019.....	14
3.2. 2020.....	16
3.3. Dyre boliger .....	17
<b>4. Diskusjon</b> .....	<b>19</b>
4.1. Validering med trenings- og testdatasett.....	19
4.2. Minimumsantall observasjoner .....	20
4.3. Grunnkretser uten transaksjoner.....	20
4.4. Oppdatering av modellen.....	21
<b>5. Oppsummering</b> .....	<b>24</b>
<b>Referanser</b> .....	<b>25</b>
<b>Vedlegg A: Ekstra figurer og tabeller</b> .....	<b>26</b>

# 1. Bakgrunn

Statistisk sentralbyrå (SSB) har utviklet en metode for å beregne markedsverdier for boliger som ble tatt i bruk i 2009. Siden modellen ble etablert, har SSB på oppdrag fra Skatteetaten brukt modellen til å beregne formuesgrunnlaget for boliger. Fra og med 2014 har flere kommuner tatt i bruk dette formuesgrunnlaget ved utskrivning av eiendomsskatt. For verdsetting av bolig i formuesskatten erstattet modellen en eksisterende, svært unøyaktig verdsetting basert på historisk byggekostnad (Finansdepartementet, 1996).

Grunnlagsdata i modellen er boligtransaksjoner fra Finn.no over en 10-årsperiode. Boligverdi beregnes separat for hver kombinasjon av boligtype og region (beregningssområde). De enkelte beregningsområdene er inndelt i soner basert på prisnivå. Det kjøres en regresjon per beregningsområde med soner som forklaringsvariabler sammen med tettstedsinndeling, byggeår og areal. I modellen er hovedregelen at hele kommunen tilhører én sone, og ingen soner går på tvers av kommunegrenser. Unntaket er de fire største byene, der prissonene er basert på bydeler. Modellen er nærmere beskrevet i Medby og Takle (2021).

Kommuneinndelingen ble vesentlig endret per 1.1.2020. Det førte til at modellen traff dårligere enn før i noen regioner, fordi kommuner i forskjellige soner ble slått sammen. På denne bakgrunn fikk SSB i oppdrag av Finansdepartementet å vurdere et alternativ til dagens verdsettelsesmodell som i mindre grad er basert på administrative grenser slik at den blir mer robust overfor endringer i kommunestruktur.

Hovedformålene med dette notatet er å kartlegge mulighetene for å frigjøre seg fra bindingen til kommunestruktur som finnes i den eksisterende modellen, og å undersøke hvor mye treffsikkerheten kan forbedres ved slike endringer. Resultatene vil også kunne fungere som et sammenligningsgrunnlag i forbindelse med en eventuell større revisjon av dagens verdsettelsesmodell, ved f.eks. bruk av maskinlæring. Det vises i denne sammenhengen til prosjektet med utvikling av en ny verdsettelsesmodell for fritidsbolig i regi av PWC, som i et høringsnotat er skissert implementert fra og med inntektsåret 2024 (PWC, 2020; Finansdepartementet, 2021).

I et tidligere forprosjekt har modellen blitt beskrevet og utprøvd for regionene Oslo, Hedmark og Møre og Romsdal (Bø, Medby, Nygård og Takle, 2022a). Her undersøker vi egenskapene ved modellen også for de resterende regionene i Norge. I tillegg går vi grundigere gjennom en del punkter som er nødvendige for å få modellen til å fungere for Skatteetatens verdsettingsformål. En omarbeidet versjon av Bø et al. (2022a) er fagfellevurdert og publisert som Bø et al. (2022b).

Vi utforsker her muligheten for å bruke grunnkretser som geografisk enhet for verdsettelsesmodellen, som et alternativ til dagens kommunebaserte modell. Grunnkretser er en mer stabil enhet enn kommuner. Formålet med å dele kommuner inn i grunnkretser er å lage små, stabile geografiske enheter som kan gi et fleksibelt grunnlag for å arbeide med og presentere regionalstatistikk. Norge er delt inn i ca. 14 000 grunnkretser, mens det per 2020 er 356 kommuner. For å gruppere grunnkretser til enheter med tilstrekkelig antall observasjoner bruker vi en maskinlæringsalgoritme. Vi diskuterer også kort andre metoder som kan brukes for bedre å utnytte informasjonen som er tilgjengelig i dagens data, men som krever noe mer ressurser til analyse og utvikling.

Den eksisterende modellen benytter få forklaringsvariabler. I dette prosjektet viser vi hvordan informasjonen fra disse forklaringsvariablene kan utnyttes mer effektivt. Datagrunnlaget fra Finn.no gir også mulighet for å inkludere flere forklaringsvariabler enn de som faktisk brukes, og det er mulig å hente ut ytterligere informasjon fra andre kilder. Et viktig hensyn i så fall er at de samme opplysningene må foreligge for hele boligmassen når verdiene skal beregnes. Videre var det ved

innføring av modellen et uttrykt ønske om at den skulle være enkel og transparent (Finansdepartementet, 2009).

## 1.1. Effekter av regionreformen

Inndelingen i prissoner som brukes i dagens modell er lite robust ved endringer i kommunestruktur. I modellen er landet delt inn i:

- 21 regioner for eneboliger (2019-fylker og storbyer)
- 11 regioner for småhus (storbyer, Akershus og fylkesgrupper)
- 11 regioner for leiligheter (storbyer, Akershus og fylkesgrupper)

Beregningene gjennomføres separat for hver kombinasjon av boligtype og region (beregningssområde). Innenfor de enkelte regionene er kommunene inndelt i soner basert på prisnivået. Soneinndelingen er ulik for hvert beregningsområde. Sone 4 for eneboliger i en region er ikke det samme som sone 4 for leiligheter i samme region osv. Det kjøres en regresjon per boligtype per region med soner som forklaringsvariabler sammen med tettstedsinndeling, byggeår og areal. I modellen er hovedregelen at hele kommunen tilhører én sone, og ingen soner går på tvers av kommunegrensene. Unntaket er de fire største byene, der prissonene er basert på bydeler.

Fra og med 1. januar 2020 ble det gjort vesentlige endringer i kommunestrukturen. 112 tidligere kommuner ble berørt av kommunereformen fra dette tidspunktet. Mange av sammenslåingene har bestått i at en sentral kommune med høyere boligprisnivå ble slått sammen med en eller flere omegnskommuner med lavere prisnivå. Som ved tidligere sammenslåinger har dette stort sett blitt håndtert ved at sonen for den nye kommunen har blitt satt lik sonen for den mest folkerike «gamle» kommunen. Dette fordi de små kommunene i liten grad påvirker gjennomsnittlig pris i den store kommunen. Dette innebærer at boliger i noen randkommuner vil få til dels stor økning i de beregnede boligverdiene selv om boligverdiene reelt sett ikke har økt.

For noen mer jevnstore kommuner ble den nye kommunen plassert midt mellom sonene til de gamle kommunene. Kommunesammenslåinger kan også påvirke kommuner som ikke har blitt gjenstand for sammenslåing, ved at det kan bli færre kommuner i de gjenstående sonene. I noen tilfeller har soner blitt slått sammen som følge av dette.

Når en folkerik kommune slås sammen med omegnskommuner kan dette medføre at mer av prisvariasjonen er innad i soner, og mindre mellom soner. Dette vil isolert sett føre til redusert forklaringskraft i modellen.<sup>1</sup> Detaljerte tall for fallet i forklaringskraft fordelt på beregningsområder finnes i Medby og Takle (2021).

Forklaringskraften for enebolig falt kraftig i regionene der det var mange sammenslåinger (for eneboliger er regionene de gamle fylkene). Dette gjelder f.eks. i regioner som Sogn og Fjordane, Vest-Agder, Trøndelag og Møre og Romsdal. Alle disse har hatt mange sammenslåinger mellom folkerike kommuner og omegnskommuner. Sterkest nedgang i forklaringskraften hadde Sogn og Fjordane, hvor den falt med nesten 20 prosent. I Sogn og Fjordane har de to «gamle» kommunene med de høyeste boligprisene, Sogndal og Førde, blitt slått sammen med flere kommuner med lavere prisnivå.

Samtidig var forklaringskraften tilnærmet uendret i regioner som Hedmark og Oppland som var upåvirket/lite påvirket av kommunereformen. Det er imidlertid ikke fall i forklaringskraften i alle

---

<sup>1</sup> Forklaringsgraden eller forklaringskraften er her målt ved  $R^2$ .  $R^2$  ligger mellom 0 og 1. Den forteller oss hvor mye variasjonen i avhengig variabel den estimerte modellen fanger opp.  $R^2 = 1$  betyr at all variasjon i avhengig variabel fanges opp av forklaringsvariablene.  $R^2 = 0$  betyr at ingen variasjon i avhengig variabel fanges opp av forklaringsvariablene.

regionene hvor det har vært kommunesammenslåinger. Østfold og Telemark er eksempler på dette. Det kan skyldes at sammenslåingene der skjedde mellom kommuner hvor prisnivået i utgangspunktet ikke varierte veldig sterkt.

Modellens forklaringskraft for småhus falt ikke like kraftig som den gjorde i flere av regionene for eneboliger. Det er heller ikke en like iøynefallende sammenheng mellom fall i forklaringskraft og kommuneendringer som for eneboligenes del. Regionen med det kraftigste fallet, Trondheim, ble riktignok påvirket av kommunesammenslåing. I Trondheim falt forklaringskraften med om lag 7 prosent. En forklaring på at kommunesammenslåinger har hatt mindre effekt for denne boligtypen, er at sonene i utgangspunktet var større enn sonene for eneboliger.

Modellens forklaringskraft for leiligheter falt stort sett enda mindre fra 2019 til 2020 enn den gjorde for småhus. Heller ikke for leiligheter var det noen iøynefallende sammenheng mellom fall i forklaringskraft og kommuneendringer. Regionen med det kraftigste fallet, Agder og Rogaland, var riktignok påvirket av kommunesammenslåing, her faller forklaringskraften med 4,65 prosent. Som for småhus er større soner en mulig forklaring på at kommunesammenslåinger har hatt mindre effekt. En annen mulig forklaring er at leiligheter, som ofte er konsentrert i byer og tettsteder, har mindre prisvariasjon mellom sentrum og periferi enn eneboliger.

## 1.2. Andre svakheter ved modellen

Håndteringen av kommunereformen er ikke den eneste svakheten med dagens boligformuesmodell selv om dette er hovedfokus her. Modellen underestimerer også boligverdiene for dyre boliger i sentrale strøk. Disse svakhetene er velkjente, blant annet inneholder Hurdalsplattformen følgende punkt: *Metoden for verdsetting av dyre eiendommer gjennom SSBs sjablongmodell må oppdateres for å reflektere eiendommenes reelle verdi. Øvrige eventuelle endringer i eiendomsskatten må utredes i samarbeid med kommunene* (Regjeringen, 2021).

Modellen benytter få forklaringsvariabler. Innen hvert beregningsområde tas det hensyn til boligens størrelse målt som P-ROM, og dummyvariabler for prissone (bydel/grupper av bydeler og kommune/grupper av kommuner), boligens alder inndelt i fire intervaller, en syvdelt gruppering av tettsteder etter innbyggertall, samt omsetningsår.

$$\ln(kvmpris) = konst + \beta_1 \ln(p\_rom) + \beta_2 prissone_p + \beta_3 ald_g + \beta_4 tskode_s + \beta_5 \text{år}_t + \varepsilon$$

Inkludering av flere forklaringsvariabler har blitt lansert som en mulig kilde til forbedring. Datagrunnlaget fra Finn.no inneholder for eksempel flere variabler som ikke utnyttes (som for eksempel peis/ildsted og balkong/terrasse), men disse har vi bare for omsatte boliger. Det finnes mye mindre data om boliger som ikke omsettes enn om dem som omsettes. **For å bruke modellen for å fastsette boligverdier for alle boliger kreves det at slike data må være tilgjengelige for alle boliger.** Modellen ble da også i sin tid bevisst utformet med få forklaringsvariabler for å minimere innhenting av data fra skatteyterne.

Når det gjelder mange av forklaringsvariablene fra finn.no er det vanskelig å se for seg muligheter for å avlede tilsvarende variabler for den totale boligmassen. Bruk av geodata er imidlertid lettere fordi innhenting kan baseres på koordinater som er tilgjengelige for (tilnærmet) alle boliger.

Problemet med underestimering av boligverdier i Oslo ble nylig analysert i en masteroppgave (Vinje, 2020). Her drøftes det hva som kan oppnås ved en annen modell for boligpriser. I sammendraget konkluderes det med at *«Resultatene viser at det er mulig å forbedre SSB sin boligprismodell betraktelig uten å kreve at brukerne av modellen må innhente mer informasjon om boligene enn i dag».*

Maskinlæringsmodeller skisseres i oppgaven som mulig løsning. Oppgaven viser at bruk av slike modeller kan gi økt forklaringskraft, men noe underestimering av dyre boliger var det fortsatt.



Oppgaven fokuserte bare på Oslo, og det er uvisst om resultatene kan generaliseres til resten av landet. Oppgaven berørte naturlig nok heller ikke virkninger av kommunereformen som er hovedfokus i forprosjektet. I kapittel 3.3 viser vi at samme type forbedring av verdsetting for dyre boliger kan oppnås ved hjelp av tradisjonelle hedoniske regresjonsmodeller i kombinasjon med en maskinlæringsmetode for å effektivt utnytte lokal geografisk informasjon.

Et annet problem når det gjelder underestimering av verdien på dyre boliger er at undersøkelser tyder på at de aller dyreste boligene er sterkt underrepresenterte i datagrunnlaget fra finn.no, noe som ikke vil kunne avhjelpe ved en bedre modell (Medby og Takle, 2020). Underrepresentasjonen er for øvrig enda større for de aller billigste boligene.

## 2. Ny modell

Som vist i forrige kapittel, er dagens modell lite robust ved kommuneendringer. Prosjektet har som mål å undersøke hvordan verdsettelsen av boliger i større grad kan frigjøre seg fra kommunegrensene.

En metode er å undersøke hvordan prissonene kan fastsettes på en mer fleksibel måte, ved å basere seg på koordinatene til de omsatte boligene. I teorien kan boligverdier beregnes ved hjelp av boligadressen til hver enkelt bolig, hvor hver adresse har et sett koordinater. Dette er en metode som omtales i Ahlfeldt, Heblich og Seidel (2021). Denne typen metode utnytter tilgjengelig informasjon på en svært fleksibel måte, men utvikling krever mer ressurser enn tilgjengelig i dette prosjektet.

I dette notatet utforsker vi derfor muligheten for å knytte sammen små områder (grunnkretser) ved hjelp av maskinlæringsalgoritmer for å optimalisere avveiningen mellom å bruke geografisk homogene eiendommer, og å få nok observasjoner. Dette omtales nærmere i kapittel 2.1. I kapittel 2.2 diskuteres flere mulige metoder.

### 2.1. Maskinlæring på grunnkretser

Ved å frigjøre seg fra kommunegrensene får man muligheten til å skape geografiske enheter som er mer fleksible. I tillegg vil endringer av kommunegrensene slå mindre direkte ut i boligverdsettning. Her har vi valgt å bruke grunnkretser som geografiske grunnenheter i stedet for kommuner. Formålet med grunnkretser beskrives som:<sup>2</sup>

Formålet med å dele kommuner inn i grunnkretser er å ha små stabile geografiske enheter som kan gi et fleksibelt grunnlag for å arbeide med og presentere regionalstatistikk. Grunnkretser består av et geografisk sammenhengende område som er mest mulig ensartet når det gjelder natur og næringsgrunnlag, kommunikasjonsforhold og bygningsmessig struktur.

Fordelen ved å bruke grunnkretser fremfor for eksempel et rutenett er at grunnkretser er laget for å dekke homogene områder. De følger i stor grad naturlige geografiske grenser som elver, motorveier, endringer i boligstruktur o.l. Ulempen ved grunnkretser er at de ikke er konstante (se grundigere omtale i kapittel 4).

Grunnkretser er små enheter (det er ca. 14 000 grunnkretser i Norge). En enkelt grunnkrets mangler derfor ofte antallet boligtransaksjoner som er nødvendig for å estimere en regresjonsmodell, selv en med få forklaringsvariabler. I vår situasjon vil det særlig være mange landlige grunnkretser som har få observasjoner av leiligheter og rekkehus. Derfor er man avhengig av å slå sammen flere grunnkretser til større enheter.<sup>3</sup>

Vi bruker en enkel maskinlæringsalgoritme kalt beslutningstre (*Decision tree*) for å sette sammen grunnkretser som ligner på hverandre.<sup>4</sup> Algoritmen setter sammen grunnkretser i beregningsområder slik at det kvadrerte avviket mellom kvadratmeterpriser for enkeltobservasjoner og gjennomsnittsprisen innad i beregningsområdene blir minimert. Det vil si at likhet mellom grunnkretser måles ved gjennomsnittlig kvadratmeterpris for gitt boligtype. Forklaringsvariablene er geografisk plassering (gjennomsnitt av lengdegrad og breddegrad for omsetninger innad i

<sup>2</sup> <https://www.ssb.no/a/metadatas/conceptvariable/vardok/135/nb>.

<sup>3</sup> Selv kommuner kan ha for få transaksjoner for å estimere en modell; i den eksisterende verdsettelsesmodellen slås mange kommuner sammen i grupper.

<sup>4</sup> Mullainathan og Spiess (2017) er en god introduksjon til maskinlæring. En annen vanlig maskinlæringsalgoritme, random forest, kombinerer resultater fra mange beslutningstrær.

grunnkretsen), kommune og fylke. En algoritme deler rekursivt opp observasjonene i datasettet basert på forklaringsvariablene for å skape beregningsområder som best mulig predikerer kvadratmeterprisene. Oppdelingen fortsetter til videre oppdeling vil føre til beregningsområder med færre observasjoner enn vårt ønskede minimum (500).<sup>5</sup> Grupperingen gjøres separat for forskjellige boligtyper (enebolig, småhus og leiligheter).

For hver enkelt beregningsområde gjøres så en hedonisk regresjon, med hovedsakelig de samme variablene som i den eksisterende modellen. Det vil si at man har en separat regresjon for hvert enkelt geografisk område og boligtype. Sonevariablene og tettstedsinndelingen er overflødige i dette rammeverket, men vi legger til dummyvariabler for hver enkelt grunnkrets. Totalt sett blir det rundt 1 200 regresjoner for hele landet med formen:

$$\ln(kvmpri) = konst + \beta_1 \ln(p_{rom}) + \beta_2 ald_g + \beta_3 \text{år}_t + \beta_4 grkrets_i + \varepsilon$$

Å ha mange separate regresjoner gir større fleksibilitet fordi alle parameterverdiene i modellen kan være forskjellige for hver kombinasjon av boligtype og gruppering. I den eksisterende modellen begrenses sammenhengen mellom f.eks. kvadratmeterpris og areal til å være lik for alle boliger av en boligtype i en hel region.

Den grunnleggende avveiningen ved aggregering av grunnkretser er at man ønsker å skape områder som er så homogene som mulig (noe som tilsier små enheter) samtidig som man er avhengig av et visst antall omsetninger i den hedoniske regresjonen (noe som tilsier større enheter). Vi krever at det er minimum 500 omsetninger i en gruppe av grunnkretser, for å få et tilfredsstillende antall observasjoner for den hedoniske regresjonen. Det resulterer i mange, geografisk små beregningsområder i sentrale strøk, og geografisk større beregningsområder i mindre sentrale strøk. Slik utnytter man den store mengden informasjon (antall transaksjoner) i byer til å bedre predikere lokale prisnivåer.

Modellen som presenteres her, er så langt mulig lik den eksisterende boligverdsettingsmodellen. Dette for best mulig sammenligning. Det betyr at vi bruker logaritmen av kvadratmeterprisen som avhengig variabel, mens det mest vanlige i hedoniske regresjoner er å bruke logaritmen av prisen (Hill, 2013). Vi har undersøkt hvorvidt det betyr noe, ved å lage en versjon av den nye modellen med logaritmen av pris som avhengig variabel. Forskjellen mellom resultatene er marginal, og ikke i entydig retning (resultatene fra den alternative modellen presenteres ikke her av plasshensyn).

## 2.2. Alternative metoder

Det finnes flere alternative metoder man kan tenke seg for å estimere boligverdi for den samlede norske boligmassen. En enkel metode for å løsrive seg fra kommunegrenser er å erstatte sonene knyttet til kommunegrenser i dagens modell med SSBs sentralitetsindeks på grunnkretsnivå.<sup>6</sup> Vi har testet en modell der vi beholder beregningsområdene fra dagens modell, og alle de andre variablene i den hedoniske regresjonene, men erstatter prissoner med sentraliteten til grunnkretsen en eiendom ligger i. Denne metoden gir gjennomgående dårligere treffsikkerhet enn den eksisterende modellen i 2019, og forverrer resultatene for storbyer, der grunnkretsene er homogene når det gjelder sentralitetsnivå. Resultatene finnes i Bø et al. (2022a).

Det er også mulig å løsrive seg helt fra de administrative inndelingene. En enkel måte vil være å dele landet i kvadrater av lik størrelse. SSB publiserer for eksempel årlig noen typer statistikk for ruteinndelingene 250m\*250m, 1km\*1km og 5km\*5km. Man kunne tenkt seg at rutene utgjorde den

<sup>5</sup> En introduksjon til rekursive grupperingsmetoder, inkludert beslutningstrær, finnes i Strobl et al. (2009).

<sup>6</sup> Sentralitetsindeksen beskrives i Høydahl (2017).

minste byggesteinen i modellen. Allikevel må man fortsatt utvikle en metode for å slå sammen ruter, da noen ruter vil ha for få observasjoner til at man kan estimere boligverdier. Denne metoden unngår problemer som kan oppstå ved at kommuner eller grunnkretser endres, men en ulempe er at rutene kan dele opp områder som naturlig hører sammen. Dersom et nabolag tilfeldigvis blir splittet av rutegrensene, kan det oppstå et uønsket skille mellom boligverdiene på hver side av grensen.

Det finnes også en større litteratur basert på mer avanserte metoder, som romlig autokorrelasjon (spatial autocorrelation) og geografisk vektet regresjon (se for eksempel Anselin, 2010 for en oversikt). Disse metodene gir mulighet til å løsrive modellen fra administrative eller arbitrære geografiske oppdelinger. Her brukes informasjon om priser fra nærliggende transaksjoner som ekstra forklaringsvariabler, henholdsvis gjennom en vektingsmatrise (av parametere eller feilledd) og ved lokal estimering av parameterverdier, der nærliggende observasjoner gis større vekt. Modeller med romlig autokorrelasjon er avhengige av sterke antakelser om strukturen på romlig sammenheng (McMillen, 2012), noe som er særlig problematisk ved et så stort og heterogent område som hele Norge. Modeller med geografisk vektet regresjon lar i større grad geografisk variasjon bestemmes av data.

En interessant variant av geografisk vektet regresjon er Ahlfeldt et al. (2021) med en kombinasjon av parametrisk, hedonisk regresjon og ikke-parametrisk romlig regresjon (spatial regression). Her kan man teoretisk beregne verdien separat for hver enkelt eiendom. Algoritmen de har utviklet, bruker i stor grad salg i nærområdet der det er mange tilgjengelige observasjoner, og i større grad en parametrisk regresjon der det er få nærliggende observasjoner. Dermed utnyttes tilgjengelig informasjon på best mulig måte: Der det er mange lignende observasjoner i datasettet, settes verdi basert på verdien av boliger helt i nærheten, noe som fanger opp lokal variasjon i priser, men modellen har også en struktur som prissetter boliger som ikke har noe direkte sammenligningsgrunnlag i data.

Hill og Scholz (2018) bruker en lignende, men enklere tilnærming, der en «spline surface» kombineres med en parametrisk hedonisk modell.<sup>7</sup> Den parametriske delen tar hensyn til fysiske attributter ved boligen slik som i vår modell (størrelse, alder osv.), mens den ikke-parametriske delen tar hensyn til betydningen av beliggenhet. Hill og Scholz kan dermed konstruere en topografisk flate som beskriver hvordan prisene varierer med de geografiske koordinatene (lengdegrad og breddegrad), gitt de fysiske attributtene. Modellen sammenlignes med en enklere modell hvor spline-funksjonen erstattes med dummyvariabler for postnummer og region med data fra Sydney. Dummyvariabler for postnummer gir om lag samme treffsikkerhet som bruk av spline-funksjonen. Årsaken er at postnummer er så finmasket at det er lite å vinne på andre metoder for å ta hensyn til boligens beliggenhet. Å kontrollere for hvert postnummer fordrer at det er nok observasjoner innenfor hvert postnummer, noe som ikke er et problem i en storby som Sydney.

Metodene vi har omtalt til nå, baserer seg alle på å benytte informasjon om salg av nærliggende boliger på en bedre måte enn i den eksisterende modellen. Alle de nevnte modellene vil kunne brukes med data som er tilgjengelige i dag. En felles problematikk med de beskrevne metodene er at empiriske undersøkelser av metodenes treffsikkerhet er basert på markeder (i stor grad byer) med større tetthet av transaksjoner enn det er i store deler av Norge. En verdsettingsmodell for alle boliger i Norge må kunne verdsette boliger der det eksisterer få transaksjoner i geografisk nærhet, samtidig som modellen også bør være i stand til å utnytte en større mengde informasjon der tettheten av transaksjoner er større. Vår metodikk er en enkel måte å få til dette på, ved å sette

---

<sup>7</sup> Man kan tenke seg en «spline surface» som høydekurver som ligger over et kart, og beskriver prisnivået for en «normal» bolig på et gitt punkt.

sammen lignende grunnkretser til beregningsområder med et minimumsantall observasjoner. Ved å ha mange beregningsområder tillater den også stor geografisk variasjon i parameterverdiene, om enn ikke like stor fleksibilitet som i geografisk vektet regresjon.

Modellen som benyttes i dag, inkluderer få forklaringsvariabler. En annen tenkbar retning er å bruke et større antall forklaringsvariabler. En moderne tilnærming her er å inkludere et stort antall mulige variabler og interaksjonsledd, og la en maskinlæringsmodell bestemme formen på modellen (se Mullainathan og Spiess, 2017 for en grundigere gjennomgang). Dette vil være en tilnærming som ligner den nye modellen for verdsetting av fritidsboliger, som i et høringsnotat er skissert implementert fra og med inntektsåret 2024 (PWC, 2020; Finansdepartementet, 2021).

Ulempen med en slik modell er at datatilgangen kan være utfordrende. Den eksisterende modellen ble i sin tid utformet med få forklaringsvariabler delvis for å minimere innhenting av data fra skattemyterne. For å bruke modellen til å fastsette boligverdier for alle boliger kreves det kjennskap til forklaringsvariabler for alle boliger, ikke bare omsatte boliger.<sup>8</sup> Datagrunnlaget fra Finn.no inneholder flere variabler som ikke utnyttes (som for eksempel ildsted og balkong/terrasse), men disse har vi bare for omsatte boliger. Det finnes mye mindre data om boliger som ikke omsettes, enn om dem som omsettes. Bruk av geodata vil være lettere fordi innhenting kan baseres på koordinater som er tilgjengelige for (tilnærmet) alle boliger.

Det er videre uklart hvor stor gevinst som kan oppnås ved å inkludere en mengde lokaliseringsvariabler utover en optimalisering av bruken av geografisk plassering dersom man har godt med omsetningsdata. PWC (2020) finner tydelig bedre forklaringskraft ved en maskinlæringsmodell (av typen random forest) enn ved en lineær modell for fritidsboliger, men utforsket dessverre ikke en lineær modell med mer enn én region. Fagereng, Blom og Torstensen (2020) sammenligner prediksjoner fra en hedonisk modell med prediksjoner fra en ensemble maskinlæringsmodell. Maskinlæringsmodellen predikerer klart bedre på landsbasis, men verken datasettet eller den hedoniske modellen kan sammenlignes direkte med den eksisterende skatteverdsettingsmodellen.<sup>9</sup>

Vinje (2020) tester ut en maskinlæringsmodell (også random forest) for verdsetting på boliger omsatt i Oslo. Han finner en markert økning i forklaringskraft ved å gå fra en hedonisk regresjon til en maskinlæringsmodell. Datasettet er ikke helt sammenlignbart med datasettet brukt for Oslo i denne artikkelen, men resultatene kan indikere en noe større økning av forklaringskraft ved en maskinlæringsmodell enn ved grunnkretsmodellen. Det er uvisst om resultatene kan generaliseres til resten av landet. Vinje (2020) konkluderer også med at ekstra forklaringsvariabler utover dem som allerede brukes i den eksisterende modellen, samt avstander beregnet ved hjelp av koordinater (avstand til kollektivtransport, skole, butikk m.m.), ikke bidrar i vesentlig grad til økt forklaringskraft.

---

<sup>8</sup> Om informasjon mangler for enkelte observasjoner kan verdien imputeres, men dette bør ikke gjelde en stor andel av observasjonene. Å imputere verdier for alle ikke omsatte boliger basert på omsatte boliger er trolig en dårlig ide.

<sup>9</sup> Det private selskapet Eiendomsverdi har laget en boligverdsettingsmodell, e-takst, til bruk for banker og eiendomsmeglere. Det finnes ingen offentlig tilgjengelig informasjon om hvordan denne modellen er utformet, eller hvor nøyaktig den er.

### 3. Resultater

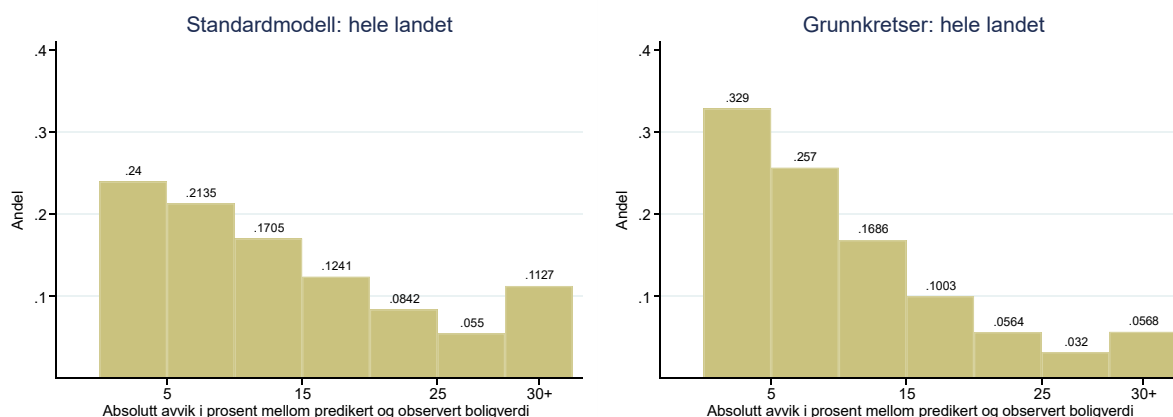
Her viser vi resultater for den nyutviklede modellen beskrevet i kapittel 2.1 sammenlignet med den eksisterende modellen. Vi bruker det samme datagrunnlag som i den eksisterende modellen, ca 800 000 boliger omsatt gjennom Finn.no over en periode på 10 år. Som i den eksisterende modellen ekskluderes ca. 0,2 prosent av observasjonene på grunn av ekstreme arealer eller kvadratmeterpriser (se Medby & Takle, 2021 for detaljer).

Først viser vi resultater for år 2019 (dvs. at vi bruker data for perioden 2010 – 2019). Å vise modellene for 2019 er best egnet for sammenligning, da det gir en bedre illustrasjon av forbedringspotensialet ved den nye metodikken. Da sammenligner man den nye modellen med en “velfungerende” versjon av den eksisterende modellen. Vi vil også vise resultater fra den enklere oppdateringen av modellen hvor soner erstattes med en sentralitetsindeks. Senere viser vi også resultater for 2020, og vi ser nærmere på hvordan modellene fungerer for dyre boliger i Oslo.

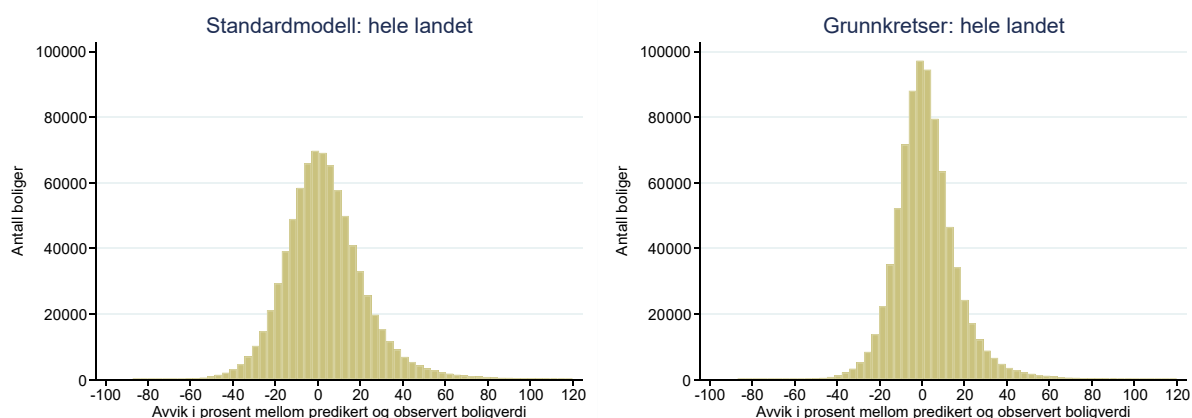
#### 3.1. 2019

Ved grupperingen av grunnkretser som beskrevet i kapittel 2.1 dannes totalt 1 184 beregningsområder i Norge (eksempelvis 245, 49 og 48 i Oslo, Hedmark og Møre og Romsdal). Grupperingen av grunnkretser for de tre eksempefylkene illustreres i vedlegg A, figur A.1 – A.3.

**Figur 3.1. Fordeling av absolutte prediksjonsfeil, eksisterende og ny modell, hele landet, 2019**



**Figur 3.2. Fordeling av prediksjonsfeil, eksisterende og ny modell, hele landet, 2019**



Deretter kjøres 1 184 separate hedoniske regresjoner for å predikere priser for boliger i hvert beregningsområde. Predikerte priser fra modellen sammenlignes så med transaksjonsprisene fra

Finn-datasettet. I figur 3.1 viser vi fordelingen av absolutte prediksjonsfeil for hele landet for den eksisterende modellen, og for modellen basert på gruppering av grunnkretser i 2019. Det er tydelig fra figurene at standardavviket for prediksjonsfeil er lavere ved grunnkretsmodellen. Fordelingen av prediksjonsfeil vises i figur 3.2. Begge modellene har en normalfordelt kurve rundt 0, men spredningen er mye større for den eksisterende modellen.

I Tabell 3.1 viser vi oppsummerte tall for treffsikkerheten til den eksisterende og den nye modellen. For alle fylker, og alle fire mål viser modellen basert på grunnkretser mindre feil. Standardavviket for feilen blir redusert med cirka 20 prosent i alle regionene.  $R^2$  stiger med 7,7 prosent totalt for hele landet.

**Tabell 3.1. Gjennomsnitt og median for absolutt prosentvis feil, standardavvik for prosentvis feil, og  $R^2$ , 2019**

Fylke	Gjennomsnitt, absolutt		Median, absolutt		Standardavvik		R2	
	Eks.	Ny	Eks.	Ny	Eks.	Ny	Eks.	Ny
Alle	15,14	11,26	11,25	8,13	21,64	16,61	0,841	0,905
Østfold	16,32	12,53	12,52	9,30	22,32	17,71	0,731	0,835
Akershus	14,04	9,56	10,75	7,16	19,90	14,11	0,814	0,910
Oslo	11,91	9,13	9,44	7,07	17,26	13,79	0,880	0,925
Hedmark	18,31	14,34	13,39	10,48	25,55	20,26	0,723	0,826
Oppland	18,19	14,23	13,27	10,11	25,65	20,58	0,727	0,817
Buskerud	17,69	11,68	13,31	8,60	24,48	16,91	0,743	0,863
Vestfold	17,77	13,69	13,35	9,97	24,34	19,29	0,644	0,761
Telemark	20,03	14,20	14,82	10,34	27,16	19,82	0,663	0,783
Aust-Agder	21,05	16,82	15,33	11,69	28,86	24,01	0,457	0,619
Vest-Agder	19,55	12,81	15,28	9,37	25,74	18,08	0,588	0,796
Rogaland	13,67	11,03	10,08	7,97	19,21	15,84	0,781	0,846
Hordaland	13,66	10,47	10,35	7,70	19,21	15,08	0,786	0,863
Sogn og fjordane	20,91	15,22	15,83	10,59	30,19	22,09	0,680	0,818
Møre og Romsdal	17,56	13,95	12,66	9,84	25,20	20,32	0,749	0,826
Trøndelag	15,14	10,78	11,34	7,69	21,37	15,92	0,790	0,890
Nordland	17,61	13,50	12,48	9,40	25,76	19,92	0,768	0,851
Troms	15,51	11,35	11,21	8,05	22,57	17,10	0,826	0,893
Finnmark	19,10	13,58	13,43	9,69	28,98	19,63	0,712	0,833

I vedlegg A viser vi tabell A.1, med transaksjonspriser og predikerte priser for eksisterende og ny modell fordelt på fylker (18), boligtyper (3), prisgrupper (7) og sentralitet (6). Gjennomsnittspriser for fylker, boligtyper og sentralitet blir stort sett godt predikert av både den eksisterende modellen og den nye, men den eksisterende modellen bommer litt for noen av fylkene med få observasjoner. Den nye modellen treffer også litt bedre for de mest og minst sentrale boligene. Det er verdt å merke seg, i tabell A.1, panel c), at den nye modellen predikerer prisene betraktelig bedre for boligene i toppen og bunnen av prisfordelingen. Tabell A.2 i vedlegg A viser transaksjonspriser og predikerte priser for prisgrupper fordelt på landsdeler. Mønstret er det samme i alle landsdelene, med kraftig forbedring av treffsikkerheten for de dyreste og billigste boligene.

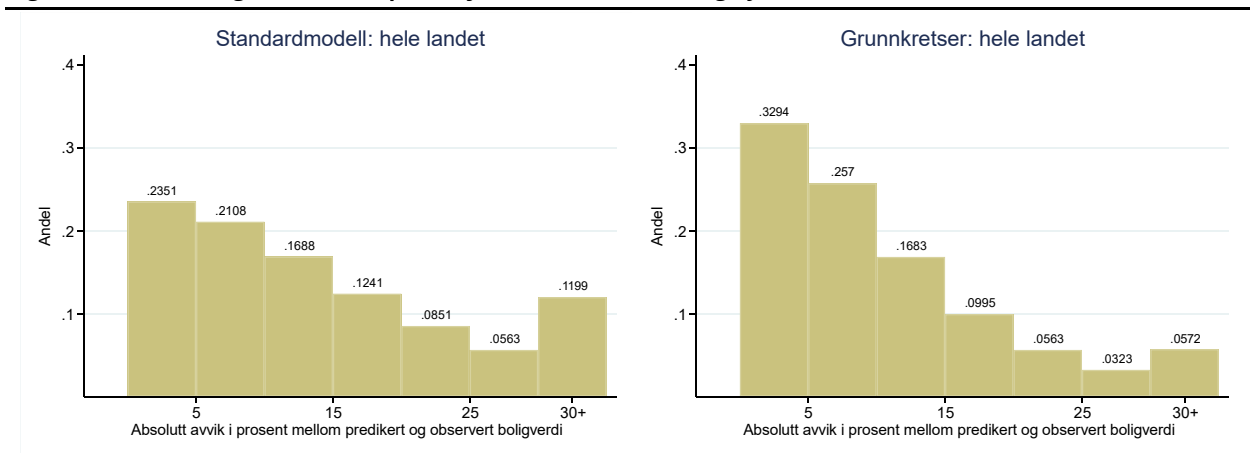
Tilslutt viser figur A.4 – A.33 i vedlegg A fordelingen av prediksjonsfeil innad i prisgrupper for hver sentralitet.<sup>10</sup> Ikke uventet er prediksjonene dårligere for begge modeller der det er få observasjoner, dvs. i de minst sentrale områdene, samt på toppen og bunnen av prisfordelingen. Gjennomgående har den nye modellen mindre prediksjonsfeil, synlig som en større andel av observasjonene nær 0-punktet.

<sup>10</sup> Det er seks sentraliteter, beskrevet i Høydahl (2017). Prisgruppene er her redusert til 5, da de minst sentrale områdene har svært få observasjoner for de høyeste prisene.

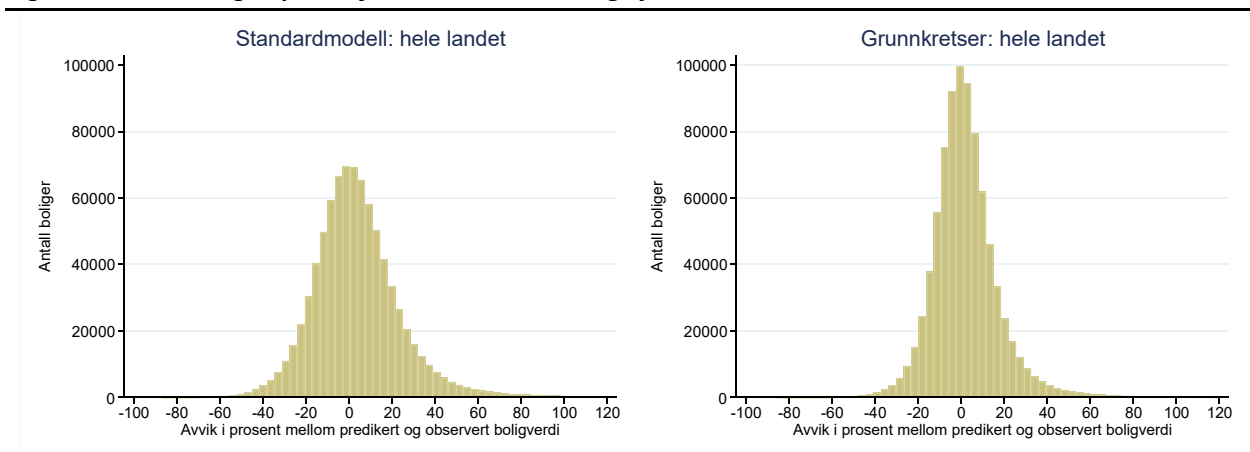
### 3.2. 2020

Her viser vi tilsvarende resultater som i kapittel 3.1 for år 2020 (dvs. data fra 2011 – 2020). For den nye modellen dannes totalt 1 208 beregningsområder i Norge (eksempelvis henholdsvis 242, 55 og 45 i Oslo, Hedmark og Møre og Romsdal). Figur 3.3 og 3.4 viser fordelingen av absolutte og vanlige prediksjonsfeil for den eksisterende og den nye modellen. Hvis man sammenligner resultatene for den eksisterende modellen i 2019 og 2020 er det tydelig at den treffer dårligere i 2020. Den nye modellen har svært like resultater som for 2019.

**Figur 3.3. Fordeling av absolutte prediksjonsfeil, eksisterende og ny modell, hele landet, 2020**



**Figur 3.4. Fordeling av prediksjonsfeil, eksisterende og ny modell, hele landet, 2020**



I Tabell 3.2 viser vi oppsummerte tall for treffsikkerheten til den eksisterende og den nye modellen. Også i 2020 gir modellen basert på grunnkretser mindre feil for alle regioner og alle fire mål. Standardavviket for feilen blir redusert med cirka 25 prosent nasjonalt, 20 prosent i Oslo og Hedmark, og med nesten 25 prosent i Møre og Romsdal.  $R^2$  forbedres med 8,6 prosent nasjonalt.

I vedlegg A viser vi tabell A.3, med transaksjonspriser og predikerte priser for eksisterende og ny modell fordelt på fylker (11), boligtyper (3), prisgrupper (7) og sentralitet (6). Som for 2019 er gjennomsnittspriser for fylker, boligtyper og sentralitet godt predikert av både den eksisterende modellen og den nye, men den nye modellen predikerer prisene betraktelig bedre for boligene i toppen og bunnen av prisfordelingen. Tabell A.4 i vedlegg A viser transaksjonspriser og predikerte priser for prisgrupper fordelt på landsdeler.

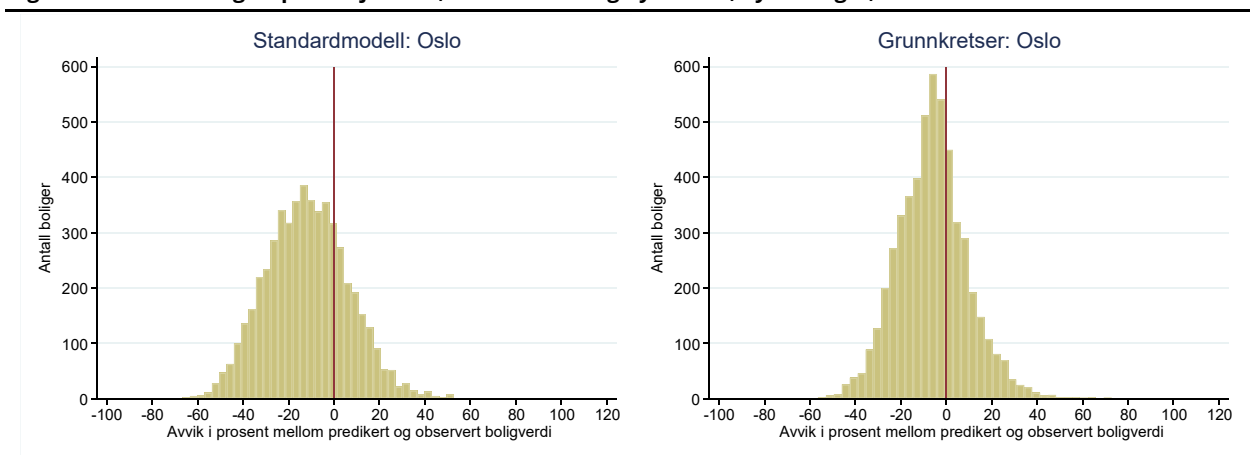
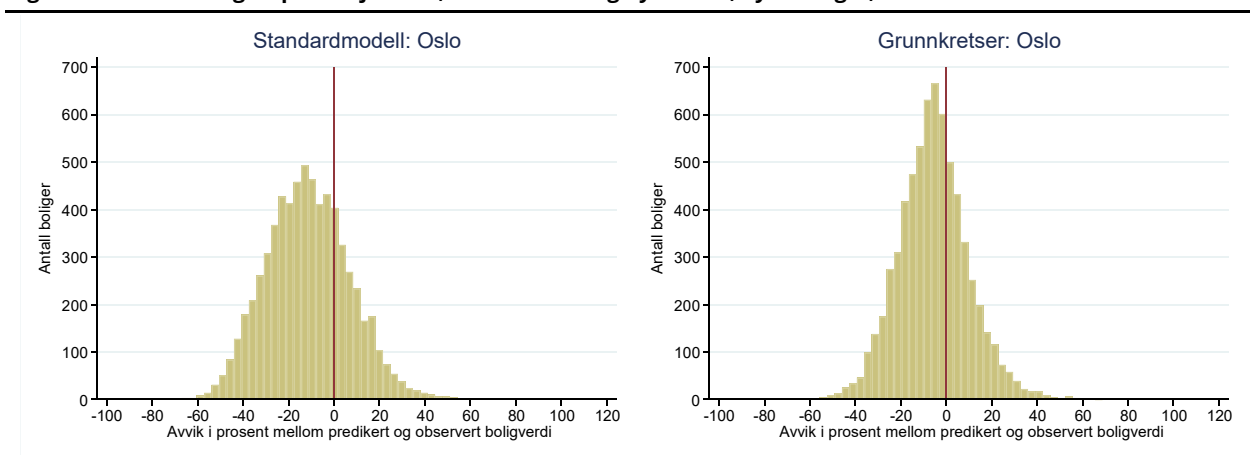


**Tabell 3.2. Gjennomsnitt og median for absolutt prosentvis feil, standardavvik for prosentvis feil, og R<sup>2</sup>, 2020**

Fylke	Gjennomsnitt, absolutt		Median, absolutt		Standardavvik		R <sup>2</sup>	
	Eks.	Ny	Eks.	Ny	Eks.	Ny	Eks.	Ny
Hele landet	15,58	11,29	11,48	8,11	22,44	16,72	0,833	0,905
Viken	16,16	10,76	12,24	7,88	22,70	15,79	0,803	0,908
Oslo	11,84	9,04	9,40	7,02	17,16	13,60	0,877	0,923
Innlandet	19,03	14,47	13,65	10,38	26,91	20,64	0,719	0,825
Vestfold og Telemark	18,87	13,95	14,06	10,17	25,85	19,59	0,659	0,774
Agder	20,96	14,28	15,98	10,15	27,92	20,57	0,533	0,745
Rogaland	14,05	11,19	10,29	8,01	19,92	16,21	0,774	0,842
Vestlandet	14,59	10,85	10,80	7,85	21,09	15,83	0,774	0,862
Møre og Romsdal	18,67	14,16	13,27	9,97	26,93	20,75	0,724	0,823
Trøndelag	15,51	10,84	11,48	7,66	22,28	16,18	0,781	0,889
Nordland	18,92	13,68	13,14	9,56	28,41	20,17	0,742	0,849
Troms og Finnmark	16,14	11,88	11,36	8,34	24,08	17,93	0,824	0,890

### 3.3. Dyre boliger

Det har vært en del oppmerksomhet i den senere tid om at den eksisterende modellen for boligverdsetting underestimerer verdien på dyre boliger. Derfor undersøker vi her om modellen med grunnkretsgrupper forbedrer prediksjon av priser for dyre boliger i Oslo. Dyre boliger definerer vi som boliger som selges for over 10 millioner (inkludert fellesgjeld).

**Figur 3.5. Fordeling av prediksjonsfeil, eksisterende og ny modell, dyre boliger, 2019****Figur 3.6. Fordeling av prediksjonsfeil, eksisterende og ny modell, dyre boliger, 2020**

Figur 3.5 og 3.6 og tabell 3.3 viser at den nye modellen gir en kraftig forbedring i prediksjon for dyre boliger. Gjennomsnittlig og median absolutt prosentvis feil blir redusert med omtrent en

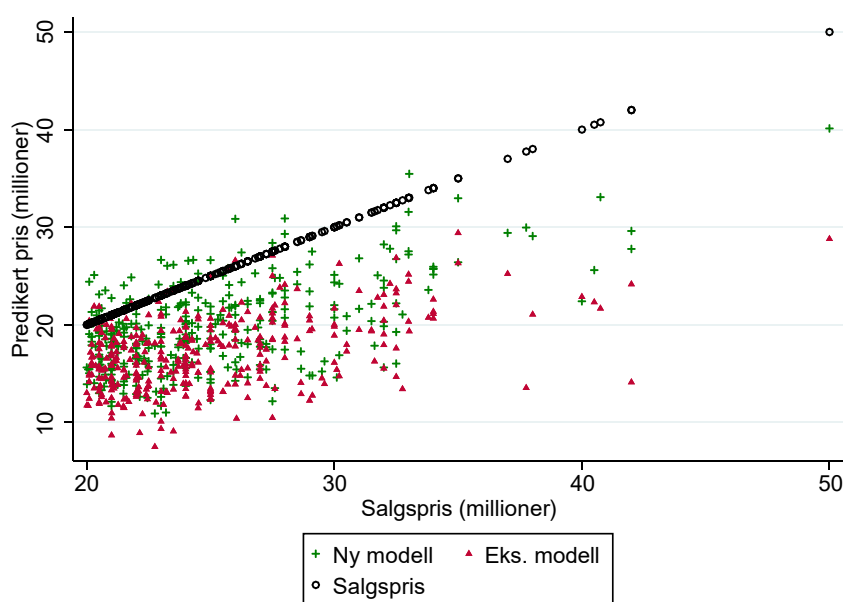
tredjedel, mens standardavviket til den prosentvise feilen reduseres med rundt 15 prosent.  $R^2$  øker med rundt 40 prosent, fra et veldig lavt nivå. Det er verdt å merke seg, i figur 3.5 og 3.6 at prediksjonsfeilene, både for den eksisterende og den nye modellen, er konsentrert under 0. Det betyr at begge modellene underestimerer verdien på dyre boliger, men problemet er synlig mindre for den nye modellen. Figurer som viser fordelingen av absolutte avvik i prediksjonene finnes i Vedlegg A, figur A.34 – A.35.

**Tabell 3.3. Gjennomsnitt og median for absolutt prosentvis feil, standardavvik for prosentvis feil, og  $R^2$ , dyre boliger i Oslo**

Modell	2019		2020	
	Eks.	Ny	Eks.	Ny
Gjennomsnitt, absolutt	17,54	13,04	17,49	12,87
Median, absolutt	15,34	10,44	15,44	10,50
Standardavvik	18,11	15,44	18,22	15,38
$R^2$	0,396	0,565	0,415	0,590

Som en illustrasjon på økt treffsikkerhet sammenligner vi i figur 3.7 transaksjonsprisene for boliger solgt for over 20 millioner i 2020-datasettet (369 observasjoner) med predikert pris fra eksisterende og ny modell.<sup>11</sup> De svarte sirklene er faktisk pris, og danner en 45°-linje. Punkter under denne linjen har for lav predikert pris, punkter over for høy. Det er tydelig hvordan de grønne kryssene med prediksjoner fra den nye modellen ligger nærmere de reelle markedsverdiene enn de røde trekantene som markerer prediksjonene fra den eksisterende modellen.

**Figur 3.7 Predikerte priser, boliger over 20 millioner i Oslo, 2020**



Som omtalt i kapittel 1.2 blir noen av de dyreste boligene som selges, ikke annonsert via Finn. Det er vanskelig å vurdere hvordan den nye modellen vil treffe for disse, men siden den tillater mer variasjon i prisene på forskjellige nabolag vil den trolig også her gi noe mer treffsikker verdsetting.

<sup>11</sup> Vi bruker her 2020-datasettet, da det inneholder flere observasjoner av svært dyre boliger. For Oslo, i motsetning til fylker der det var kommunesammenslåinger, er treffsikkerheten for den eksisterende modellen svært lik i 2019 og 2020 (sammenlign tabell 3.1 og 3.2.).

## 4. Diskusjon

Til nå har vi vist lovende resultater fra grunnkretsmodellen for 2019 og 2020. Men hvis modellen skal brukes for skatteverdsetting må den på en robust måte kunne beregne boligverdier for alle boliger i Norge år etter år. Her omtaler vi noen problemstillinger som er viktige i den sammenheng.

Selv om grunnkretser er ment å være stabile over tid hender det at grunnkretser splittes eller endres. De fleste endringene i grunnkretsstrukturen er oppdeling i mindre kretser, noe som er mest vanlig i utbyggingsstrøk. I Ullensaker kommune ble for eksempel syv grunnkretser delt i to eller tre nye grunnkretser mellom 2017 og 2018, på grunn av befolkningsøkning. Endringer i kommunegrensene vil føre til at grunnkretsene får nytt nummer (de fire første sifrene i grunnkretsnummeret er kommunenummer), men selve grunnkretsen består som regel.

I dette notatet har vi brukt grunnkretsdata fra 2019 og 2020 respektivt, for alle observasjoner i 2019- og 2020-datasettet. Noen boligtransaksjoner i datasettene er ikke koblet til en av disse grunnkretsene, enten på grunn av endringer i grunnkrets siden tidspunktet for boligtransaksjon eller pga. manglende utfylling. Disse boligtransaksjonene er plassert i en oppdatert grunnkrets ved hjelp av geokoordinater. For å unngå unødvendig manuelt arbeid ved grunnkretsendringer bør en rutine for geografisk plassering av transaksjoner i oppdaterte grunnkretser opprettes i driftsversjonen av modellen. Det vil si at alle observasjoner i datasettet plasseres i oppdaterte grunnkretser ved et program som kobler geografiske koordinater med grunnkrets.

### 4.1. Validering med trenings- og testdatasett

En vanlig problemstilling ved bruk av maskinlæringsmetoder er overtilpasning (*overfitting*) (Mullainathan og Spiess, 2017). Det er risikoen for at modellen tilpasses for godt til tilfeldige trekk ved datasettet som brukes, slik at modellen blir mindre nyttig ved prediksjoner på et datasett med andre observasjoner. Et eksempel i vårt rammeverk ville være hvis man predikerer pris for hver enkelt grunnkrets. I en grunnkrets med svært få salg kan en enkelt omsatt enebolig med uvanlig lav standard trekke ned predikerte priser for alle ikke omsatte eneboliger i grunnkretsen. Dette er en av grunnene til at vi har satt krav til et minimum antall observasjoner i beregningsområdene våre.

I maskinlæringslitteraturen validerer man ofte modeller ved å se på prediksjonskraften til modellen utenfor datautvalget (*out-of-sample*). Vi følger denne metodikken ved å dele datasettet vårt (for 2019) inn i to tilfeldige utvalg. Treningsdatasettet, som består av 90 prosent av observasjonene, brukes for å estimere modellen. Den ferdige modellen brukes så for å predikere boligverdier på testdatasettet (10 prosent av totalt antall observasjoner). Dette har vi gjort både for den eksisterende modellen og for den nyutviklede. For den nye modellen har vi valgt å skalere ned minimumsantallet observasjoner med 10 prosent (til 450), for å få omtrent like mange beregningsområder i denne analysen som i den fulle modellen.

**Tabell 4.1. Resultater for full modell, og for data delt i trenings- og testdata, 2019**

Modell	Eksisterende			Ny		
	Full modell	Trening	Test	Full modell	Trening	Test
Gjennomsnitt, absolutt	15,14	15,14	15,16	11,26	11,22	12,07
Median, absolutt	11,25	11,25	11,22	8,13	8,10	8,57
Standardavvik	21,64	21,65	21,52	16,61	16,54	18,11
R <sup>2</sup>	0,841	0,841	0,837	0,905	0,906	0,892

I tabell 4.1 ser vi at prediksjonskraften fra den eksisterende modellen er tilnærmet uendret når man validerer med testdatasettet (kolonnen merket Test). Det er ikke uventet, da det er svært mange observasjoner i hvert beregningsområde, og en endring i utvalg verken påvirker prissoner eller

regioner. For den nye modellen blir det litt større forskjeller. Den fulle modellen og resultatene for treningsdatasettet er ganske like, dvs. at å redusere antallet observasjoner med 10 prosent ikke påvirker utfallet særlig mye. Men modellen har en noe svakere prediksjonskraft for testdatasettet. - Treffsikkerheten til den nye modellen er fortsatt bedre enn den eksisterende for alle fire mål, men eksempelvis gir den nye modellen en økning i  $R^2$  på 6,6 prosent mot 7,7 prosent for det fulle datasettet.

## 4.2. Minimumsantall observasjoner

Vi har valgt å bruke minimum 500 observasjoner per beregningsområde. Minimumstallet er satt for å unngå at uvanlige enkeltobservasjoner påvirker resultatene for mye. For å gjøre en mer formell undersøkelse av betydningen av valgt minimum utfører vi også en analyse av 2019-data med minimum 1000, 250 og 100 observasjoner per beregningsområde.

Tabell 4.2 viser resultater for et treningsdatasett og et testdatasett (ikke brukt i ved estimering av modellen), som forklart i kapittel 4.1. I treningsdatasettet gir en reduksjon av minimumsantallet bedre resultater, om enn ikke så mye. 100 observasjoner gir bedre resultater på alle mål enn 250, som er bedre enn 500 og 1000.<sup>12</sup> Men det er en risiko for at dette kan skyldes overtilpasning i de små datasettene. Det er en avveining mellom økt fleksibilitet (ved flere grupperinger) og større risiko for påvirkning fra ekstreme enkeltobservasjoner ved bruk av mindre beregningsområder. Resultater fra testdatasettet er mye likere. Treffsikkerhetsmålene er marginalt dårligere for 1000 observasjoner per beregningsområde, men for 500 og 250 observasjoner ser modellen ut til å treffe tilnærmet likt. For 100 observasjoner er resultatene dårligere igjen. Dermed ser det ut som om overtilpasning kan bli et problem ved for små beregningsområder. Disse resultatene indikerer at 500 observasjoner som minimum er et godt utgangspunkt, men det er også trolig mulig å gå litt ned i antall.

**Tabell 4.2. Modeller med forskjellig minimumsantall observasjoner, trenings- og testdata, 2019**

Min. Antall obs.	1000		500		250		100	
	Trening	Test	Trening	Test	Trening	Test	Trening	Test
Gjennomsnitt, absolutt	11,43	12,14	11,22	12,07	10,97	12,05	10,53	12,20
Median, absolutt	8,28	8,64	8,10	8,57	7,91	8,50	7,68	8,49
Standardavvik	16,87	18,21	16,54	18,11	16,15	18,19	15,77	18,53
$R^2$	0,903	0,891	0,906	0,892	0,910	0,892	0,914	0,889

## 4.3. Grunnkretser uten transaksjoner

Datasettene som brukes til beregningene dekker boligtransaksjoner over ti år. En del grunnkretser har ikke noen transaksjoner over disse periodene. Det kan skyldes at det ikke finnes boliger i grunnkretsen (grunnkretser uten transaksjoner i Oslo er f.eks. Bjerkebanen, Sjursøya og diverse grunnkretser i sentrum og Marka), men det kan også være grunnkretser der det finnes et fåtall boliger som omsettes sjelden. Særlig for boligtypene småhus og leiligheter finnes det et større antall landlige grunnkretser uten observerte transaksjoner. Alle grunnkretser bør tildeles et beregningsområde, fordi det kan være boliger der som skal verdsettes for skatteformål.

I tabell 4.3 viser vi antallet grunnkrets/boligtypekombinasjoner uten transaksjoner i hele perioden vi har tilgjengelig data, dvs. de 11 årene 2010 – 2020 (se også figur A1-A3 i vedlegg A for en indikasjon på det geografiske aspektet).<sup>13</sup> Cirka 40 prosent av grunnkrets/boligtypekombinasjonene har ikke noen transaksjoner, og ytterligere 14 prosent har transaksjoner i tre eller færre av de 11 årene. I har vi også fordelt antallet grunnkretser uten transaksjoner på boligtype. Det er tydelig flere grunnkretser som mangler transaksjoner av boligtypene småhus og leiligheter, for begge typer er

<sup>12</sup> Resultatene for 500 observasjoner er de samme som vist i tabell 4.1.

<sup>13</sup> Her ser vi på 2020-grunnkretser. Tallene for 2019 – grunnkretser er svært like, da det kun var små endringer i grunnkretser mellom årene.

dekningsgraden godt under 50 prosent. Av 14 096 grunnkretser mangler 1 786 transaksjoner for alle tre boligtyper. Dette er i stor grad grunnkretser med få eller ingen beboere. Nær en tredjedel (568) av grunnkretsene har ikke registrerte beboere, og de 1 786 grunnkretsene har totalt sett ca 64 000 innbyggere, dvs. rundt 1,2 prosent av landets befolkning. Så selv om antallet grunnkrets/boligtypekombinasjoner uten observasjoner er stort, så vil antallet boliger som ligger i de relevante områdene, og som dermed omfattes av en unntaksregel, være lite.

**Tabell 4.3. Antallet grunnkrets/boligtypekombinasjoner fordelt på antall år med transaksjoner**

Antall år med transaksjoner	Alle boligtyper	Enebolig	Småhus	Leilighet
0	17 863	2 444	7 566	7 853
1	2 552	1 053	1 001	498
2	1 851	872	641	338
3	1 663	792	525	346
4	1 431	690	482	259
5	1 424	717	420	287
6	1 419	736	423	260
7	1 386	726	411	249
8	1 593	846	449	298
9	1 753	960	439	354
10	2 506	1 341	623	542
11	6 847	2 919	1 116	2 812
Total	42 288	14 096	14 096	14 096

At det er såpass mange grunnkretser uten småhus eller leiligheter kan være et argument for å oppdatere modellen ofte, for å kunne inkludere nybygg av boliger av disse boligtypene.

Fordi unntaksreglene trolig vil omfatte få boliger, og fordi det er vanskelig å teste forskjellige løsninger (dette er grunnkretser uten observasjoner, dermed er transaksjonsprisene uobserverbare), foreslår vi en enkel løsning, der grunnkretser uten gruppe tildeles det beregningsområdet med den nærmeste grunnkretsen. Mer spesifikt beregner vi det geografiske midtpunktet for alle grunnkretser (hav ekskludert), og måler avstanden mellom disse midtpunktene. Grunnkretser uten transaksjoner settes i samme beregningsområde som den grunnkretsen som er nærmest i avstand og har beregningsområde. Som et eksempel på denne metodikken vises kart for Bærum, uten og med tildeling av grunnkretser uten transaksjoner til beregningsområder, i figur A.36 – A.38 i vedlegg A.

For grunnkretser uten observasjon av en eller to boligtyper ville det vært mulig å plassere den/de manglende grunnkrets/boligtypekombinasjonene i samme beregningsområde som for kombinasjonene som har transaksjoner. Men basert på visuell inspeksjon av grunnkretser der alle tre boligtyper eksisterer, kan det virke som dette ikke er en god strategi, da beregningsområder ofte er satt sammen av ulike grunnkretser for de forskjellige boligtypene.

#### 4.4. Oppdatering av modellen

Det må også vurderes hvorvidt beregningsområdene skal oppdateres fra år til år, eller med for eksempel 5- eller 10-årsintervaller, og hvilke konsekvenser det får med hensyn til stabilitet i ligningsverdier over tid. Spørsmålet om hvor ofte man skal oppdatere er relatert til stabilitet. Generelt sett ønsker en at verdsettingen av enkeltboliger skal variere minst mulig fra år til år. Hvis årlige oppdateringer tilfører ekstra støy, som skaper unødvendige svingninger i verdiene kan det være bedre å oppdatere sjeldnere. Samtidig vil man ønske å oppdatere ofte hvis det gir mer nøyaktige resultater, og hvis sjeldne oppdateringer gjør at endringene i beregnede verdier blir veldig store ved oppdateringene.

Først har vi gjort en enkel beregning på korrelasjonen mellom predikerte verdier for eiendommer i datasettet som predikeres i både 2019- og 2020-modellen (transaksjoner 2011 – 2019). Resultatene vises i tabell 4.4. Korrelasjonene mellom predikerte priser er noe høyere for den nye modellen enn for den eksisterende totalt sett. Men det kan se ut som om den eksisterende modellen har høyere korrelasjon i fylker uten endring i kommunegrenser, men tydelig lavere for fylker som Møre og Romsdal med mange kommunesammenslåinger.

**Tabell 4.4. Korrelasjon mellom predikerte verdier 2019 og 2020**

Modell	Hele landet		Oslo		Hedmark		Møre og Romsdal	
	Eks.	Ny	Eks.	Ny	Eks.	Ny	Eks.	Ny
Korrelasjon	0,9956	0,9977	0,9999	0,9985	0,9968	0,9952	0,9836	0,9943

Videre har vi testet ut effekten av å ikke oppdatere beregningsområdene mellom 2019 og 2020. Det vil si at vi gjør hedoniske regresjoner på 2020-datasettet basert på beregningsområdene laget på 2019-datasettet. I tabell 4.5 vises data for trenings- og testdatasett for 2019, 2020, og for 2020 basert på beregningsområdene fra 2019. Resultatene er svært like, og viser kun en marginal forverring for 2020-resultater ved bruk av beregningsområder fra 2019. Men i underkant 500 av 2020-observasjonene blir ikke fordelt til beregningsområde, da de kommer fra grunnkretser uten observasjoner i 2019-datasettet.

**Tabell 4.5. Modeller for 2019, 2020 og 2020 med 2019-grupper, trenings- og testdata**

Min. Antall obs.	2019		2020		2020 m. 2019-grupper	
	Trening	Test	Trening	Test	Trening	Test
Gjennomsnitt, absolutt	11,22	12,07	11,25	12,00	11,27	12,09
Median, absolutt	8,10	8,57	8,08	8,53	8,11	8,59
Standardavvik	16,54	18,11	16,64	18,19	16,66	18,53
R <sup>2</sup>	0,906	0,892	0,906	0,890	0,906	0,890

Til slutt har vi gjort en analyse av hvordan predikerte boligverdier endres fra år 2019 til år 2020. Vi sammenligner resultatene fra den eksisterende modellen, den nye modellen der beregningsområdene oppdateres i 2020, den nye modellen der 2019-beregningsområdene brukes også i 2020, og den nye modellen hvor grunnkretsene oppdateres fra 2016-beregningsområder<sup>14</sup> til 2020-beregningsområder. Her ser vi på den faktiske endringen som vil oppstå i skatteverdsettingen, det vil si at vi predikerer alle verdier for henholdsvis 2019 og 2020.<sup>15</sup>

**Tabell 4.6. Endring i beregnet verdi fra 2019 til 2020**

	Gj.snitt	St.avvik	p1	p10	p25	p50	p75	p90	p99
Eks.	0,0453	0,0564	-0,088	0,006	0,026	0,044	0,059	0,070	0,311
Ny	0,0427	0,0739	-0,196	-0,035	0,010	0,047	0,079	0,117	0,236
Ny, 2019-grupper	0,0430	0,0384	-0,054	-0,001	0,020	0,043	0,065	0,087	0,143
Ny, fra 2016-grupper	0,0427	0,0751	-0,198	-0,036	0,010	0,046	0,080	0,119	0,239

Tabell 4.6 viser resultatene. Snittveksten i verdi er relativt lik for alle fire beregningene, og tilsvarer den gjennomsnittlige prisveksten i landet. Spredningen i vekst er derimot merkbart forskjellig. Den nye modellen har noe større spredning i vekst, spennet mellom veksten i 10. og 90. prosentil er for eksempel -3,5 til 11,7 prosent, mot 0,6 til 7 prosent for den eksisterende modellen, noe som gjenspeiles i høyere standardavvik. Derimot gir den nye modellen uten endring av

<sup>14</sup> Vi har brukt år 2016 fordi grunnkretsdata fra 2016 var lett tilgjengelig for oss. Ved beregning av 2016-beregningsområder brukes kun 7 år for gruppering (da vi har data fra 2010), og minimumsantallet observasjoner for et beregningsområde er derfor skalert fra 500 til 350.

<sup>15</sup> I alle tidligere resultater ser vi på hvordan predikert verdi er i en boligs omsetningsår. Dette for å kunne sammenligne med den observerte transaksjonsprisen.

beregningsområder, et lavere standardavvik enn den eksisterende modellen. Det er særlig synlig i mindre endringer i ytterkanten av fordelingen. Den eksisterende modellen har svært høy prosentvis verdivekst for eiendommer i toppen av fordelingen, noe som kan være et resultat av de tidligere nevnte kommunesammenslåingene. Når man går fra bruk av 2016-beregningsområder i 2019 til nye 2020-beregningsområder blir endringene kun marginalt større enn endringene fra 2019-beregningsområder i 2019 til 2020-beregningområder (sammenligning av rad 2 og 4 i tabell 4.6).

Å beholde beregningområder fra 2019 til 2020 gir altså mindre endringer enn å oppdatere områdene. Samtidig blir det ikke noe særlig større endringer ved å gå fra 4 år gamle 2016-beregningområder til 2020-beregningområder enn ved å gå fra 2019-beregningsområder til 2020-beregningsområder. Dette kan være en indikasjon på at noen av endringene som kommer ved å oppdatere beregningsområder er tilfeldig støy, ikke reelle endringer i markedsverdi. Det taler for å ikke oppdatere områdene årlig, men heller med 3 – 5-års mellomrom. Ulempen ved å ikke oppdatere årlig er at noen observasjoner som får tildelt 2019-beregningsområder ikke er tildelt 2016-beregningsområder (se omtale i kapittel 4.3). De vil bli verdsatte etter en dårligere unntaksregel. Det gjelder ca 0,3 prosent av transaksjonene, men muligens en noe større andel av boligmassen, og er et argument for årlige oppdateringer.

## 5. Oppsummering

Et stort antall kommuner ble slått sammen i 2020. Det har medført dårligere prediksjoner fra den eksisterende boligverdsettingsmodellen. Hovedmålet ved dette notatet er å undersøke hvordan modellen kan baseres på mindre og mer robuste geografiske enheter enn kommuner.

Dagens modell benytter få forklaringsvariabler. Datagrunnlaget gir mulighet for å inkludere flere forklaringsvariabler enn de som faktisk brukes, men de samme opplysningene må foreligge også for hele boligmassen når verdiene skal beregnes. Videre var det ved innføring av modellen, et uttrykt ønske om at den skulle være enkel og transparent.

Dette notatet gir en beskrivelse av en maskinlæringsmetode som bruker grunnkretser som geografisk grunnenhet ved vurdering av boligverdier. Nærliggende grunnkretser aggregeres til beregningsområder med lignende prisnivå. Størrelsen på beregningsområdene bestemmes av at det kreves et minimum antall transaksjoner i hvert område. Grupperingen gjøres separat for forskjellige boligtyper. Det resulterer i mange, geografisk små beregningsområder i sentrale strøk, og geografisk større beregningsområder i mindre sentrale strøk. En hedonisk regresjon, lignende den eksisterende modellen, kjøres så separat for hvert enkelt beregningsområde. Slik utnytter man den store mengden informasjon (antall transaksjoner) i byer til bedre å predikere lokale prisnivåer.

Resultatene, beskrevet i kapittel 3, viser tydelig bedre treffsikkerhet enn den eksisterende modellen, for flere forskjellige mål og for alle fylker. Treffsikkerheten er kraftig forbedret også sammenlignet med den eksisterende boligverdsettingsmodellen i 2019, før kommunesammenslåingsproblematikken. Treffsikkerheten forbedres langs flere dimensjoner og forbedringen er tydeligst for boliger med lav og høy pris. Resultatene er svært like for 2019 og 2020, noe som tyder på at sammenslåing av kommuner har liten effekt på modellen, og at den er relativt stabil over tid. Modellen gir også en kraftig forbedring av prediksjonene for dyre boliger i Oslo, selv om de fortsatt verdsettes noe lavt.

Videre problemstillinger utforskes i kapittel 4. Der ser vi på validering av modellen ved hjelp av data utenfor datautvalget. Datasettet deles inn i to tilfeldige utvalg, og modellen estimeres på et treningsdatasett, som består av 90 prosent av observasjonene. Den ferdige modellen brukes så for å predikere boligverdier på de resterende testdatasettet. Også her gir den nye modellen betydelig bedre forklaringskraft enn den eksisterende. Vi utforsker det optimale antallet minimumsobservasjoner i et beregningsområde, og finner at de beste resultatene ligger ved 500 – 250 observasjoner. Vi diskuterer også måter å tilknytte grunnkretser uten observasjoner til en av de eksisterende beregningsområdene, samt hvor ofte modellen bør oppdateres. Vi anbefaler en enkel løsning for grunnkretser uten observasjoner, nemlig tilknytning til det beregningsområdet som inneholder den nærmeste grunnkretsen (målt som avstand mellom de geografiske midtpunktene av grunnkretsene). Det er forhold som taler for både årlig og noe sjeldnere oppdatering av beregningsområder, totalt sett tror vi oppdatering hvert 3. – 5. år kan være å foretrekke, da det vil minske støy som kan oppstå i verdsetting av boliger som bytter beregningsområde.

Det vil være mulig å endre antall minimumsobservasjoner, tilknytning av grunnkretser uten observasjoner og oppdateringsfrekvens også etter at modellen eventuelt er satt i drift, uten at modellapparatet må endres vesentlig.



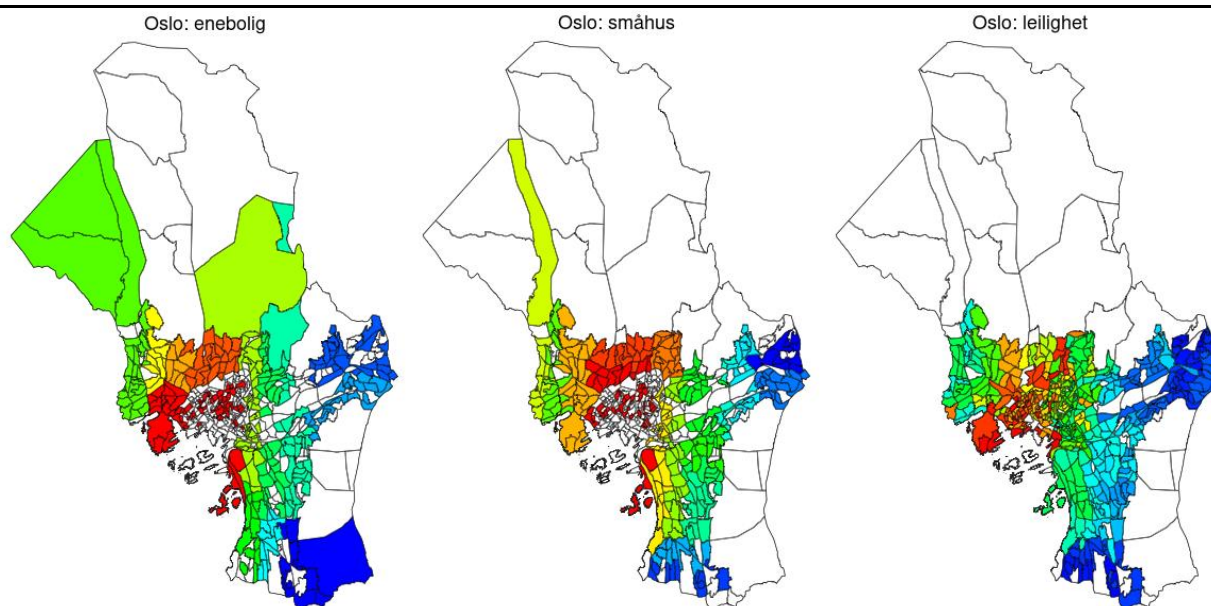
## Referanser

- Ahlfeldt, Gabriel M., Stephan Hebllich og Tobias Seidel (2021): «Micro-geographic property price and rent indices», kommende i *Regional Science and Urban Economics*.
- Anselin, Luc (2010) «Thirty years of spatial econometrics», *Papers in regional science*, 89 (1): 3-25.
- Bø, Erlend E., Per Medby, Odd Erik Nygård og Mona Takle (2022a): «Forprosjekt om boligformuesmodellen», Statistisk sentralbyrå, Notat 2022/11.
- Bø, Erlend E., Per Medby, Odd Erik Nygård og Mona Takle (2022b): «Hedonisk verdsettelse av bolig med fleksibel geografisk inndeling», *Tidsskrift for boligforskning*, 5 (2): 70-87.
- Fagereng, Andreas, Martin Blomhoff Holm og Kjersti Næss Torstensen (2020): «Housing wealth in Norway, 1993–2015», *Journal of Economic and Social Measurement*, 45 (1): 65-81.
- Finansdepartementet (1996): «Boligtaksering og prinsipper for boligbeskatning», Stortingsmelding nr. 45 (1995-96).
- Finansdepartementet (2009): «Prop. 1 S (2009-2010): For budsjettåret 2010: Skatte-, avgifts- og tollvedtak», Proposisjon til Stortinget.
- Finansdepartementet (2021): «Høringsnotat – nytt verdsettelsessystem for fritidsboliger i formuesskatten», Saksnr. 21/3144, datert 07.10.2021.
- Hill, Robert J. (2013) «Hedonic price indexes for residential housing: A survey, evaluation and taxonomy», *Journal of economic surveys*, 27 (5): 879-914.
- Hill, Robert J. og Michael Scholz (2018): «Can geospatial data improve house price indexes? A hedonic imputation approach with splines», *Review of Income and Wealth* 64 (4): 737-756.
- Høydahl, Even (2017): «Ny sentralitetsindeks for kommunene», Statistisk sentralbyrå, Notat 2017/40.
- McMillen, Daniel P. (2012): «Perspectives on spatial econometrics: linear smoothing with structured models», *Journal of Regional Science*, 52 (2): 192-209.
- Medby, Per og Mona Takle (2020): «Modell for beregning av boligformue: Oppdatert med tall for 2019», Statistisk sentralbyrå, Notat 2020/9.
- Medby, Per og Mona Takle (2021): «Modell for beregning av boligformue: Oppdatert med tall for 2020», Statistisk sentralbyrå, Notat 2021/10.
- Mullainathan, Sendhil og Jann Spiess (2017): «Machine learning: an applied econometric approach», *Journal of Economic Perspectives*, 31 (2): 87-106.
- PWC (2020): «Utvikling og implementering av nytt system for verdsettelse av fritidsboliger», <https://www.regjeringen.no/contentassets/e9f862f850474dca89826e25e4b01314/pwc-rapport-utvikling-og-implementering-av-nytt-system-for-verdsettelse-av-fritidsboliger.pdf>.
- Regjeringen (2021): «Hurdalsplattformen», <https://www.regjeringen.no/no/dokumenter/hurdalsplattformen/id2877252/?ch=1>.
- Strobl, Carolin, James Malley & Gerhard Tutz (2009): «An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests», *Psychological methods*, 14 (4): 323-348.
- Vinje, Thomas (2020): «Hva oppnår en med en ny modell for boligpriser», Masteroppgave Economics, UiO, juni 2020.

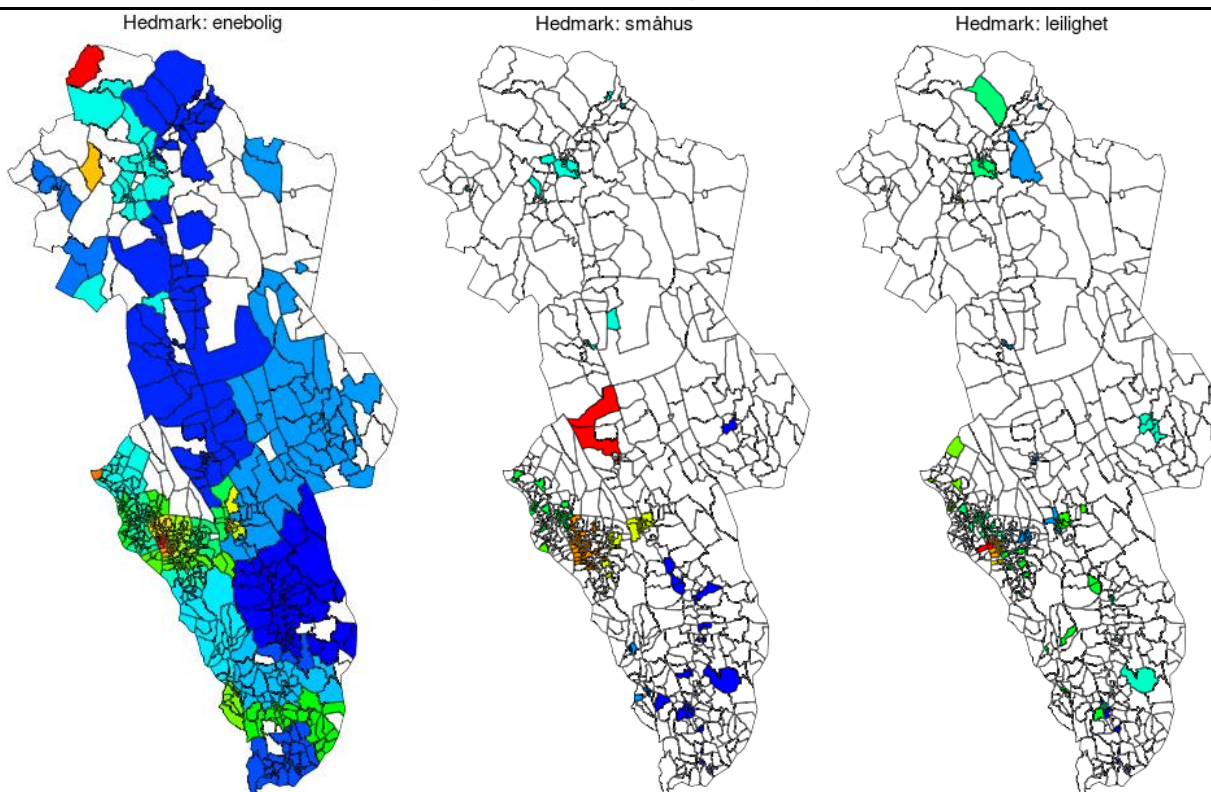
## Vedlegg A: Ekstra figurer og tabeller

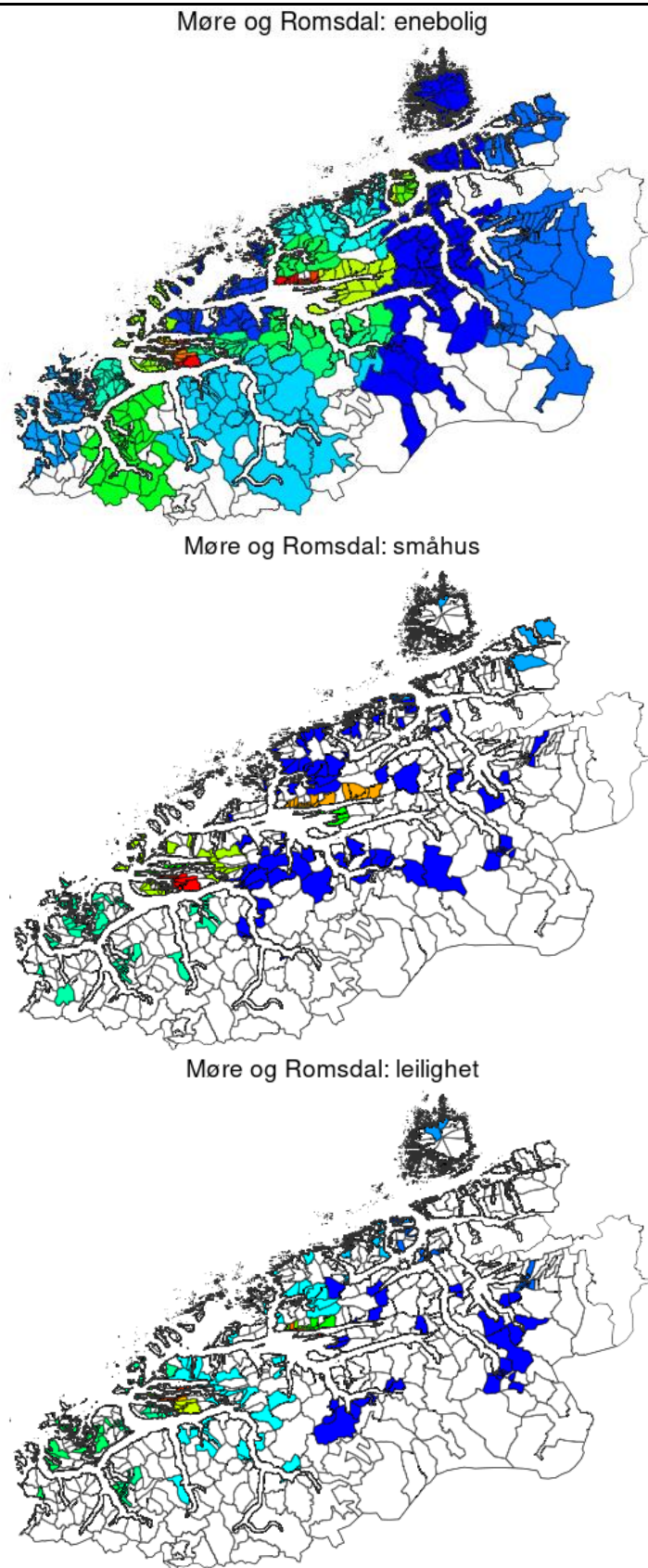
### Kart over inndeling av beregningsområder

**Figur A.1** Kart over grunnkretser med predikert kvadratmeterpris, boligtype, Oslo, 2019



**Figur A.2** Kart over grunnkretser med predikert kvadratmeterpris, boligtype, Hedmark, 2019



**Figur A.3 Kart over grunnkretser med predikert kvadratmeterpris, boligtype, Møre og Romsdal, 2019**

Figur A1 – A3 viser predikerte kvadratmeterpriser for forskjellige beregningsområder. Varme farger er høye priser, kalde farger lave. Kartene viser resultater for respektivt eneboliger, småhus og leiligheter i hvert av de tre fylkene.

**Ekstra resultater, 2019****Tabell A.1 Sammenligning av transaksjonspriser og predikerte priser, 2019**

a) Fylke	Transaksjonspris	Eksisterende modell	Ny modell	Antall
Østfold	2 371 912	2 357 204	2 373 452	46 864
Akershus	3 888 793	3 876 773	3 889 452	94 944
Oslo	3 928 779	3 911 957	3 926 209	182 410
Hedmark	2 089 853	2 088 122	2 090 859	24 011
Oppland	2 218 279	2 226 806	2 224 937	18 794
Buskerud	2 695 567	2 696 315	2 698 067	41 307
Vestfold	2 608 346	2 588 277	2 603 212	41 099
Telemark	1 971 799	2 009 333	1 973 490	23 878
Aust-Agder	2 323 681	2 318 096	2 318 656	12 845
Vest-Agder	2 545 063	2 513 883	2 543 088	26 581
Rogaland	3 177 712	3 187 657	3 177 828	66 855
Hordaland	2 951 706	2 951 740	2 950 411	76 986
Sogn og fjordane	2 394 294	2 326 464	2 402 672	5 890
Møre og Romsdal	2 342 094	2 354 991	2 346 520	29 682
Trøndelag	2 778 637	2 778 548	2 779 982	72 732
Nordland	2 349 813	2 356 705	2 356 918	25 017
Troms	3 004 605	3 026 653	3 009 007	19 404
Finnmark	2 355 873	2 356 672	2 363 690	5 557

b) Boligtype	Transaksjonspris	Eksisterende modell	Ny modell	Antall
Enebolig	3 506 644	3 509 147	3 510 765	245 341
Småhus	3 411 892	3 410 745	3 413 461	106 323
Leilighet	2 773 179	2 762 769	2 770 904	463 192

c) Prisgruppe	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	811 470	1 189 327	1 046 255	18 499
< 2 500 000	1 882 893	2 019 864	1 964 045	344 564
< 5 000 000	3 377 868	3 383 235	3 380 364	365 573
< 7 500 000	5 928 533	5 517 674	5 698 105	62 091
< 10 000 000	8 435 520	7 476 720	7 937 118	15 853
< 20 000 000	12 423 723	10 334 588	11 313 825	7 915
> 20 000 000	24 332 383	15 383 151	18 513 727	361

d) Sentralitet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
1	3 920 739	3 897 429	3 918 773	252 884
2	3 081 380	3 075 910	3 080 121	240 805
3	2 592 437	2 604 018	2 593 304	202 504
4	2 190 458	2 199 666	2 196 311	90 702
5	1 840 092	1 834 370	1 848 317	23 672
6	1 604 147	1 566 190	1 609 565	4 289

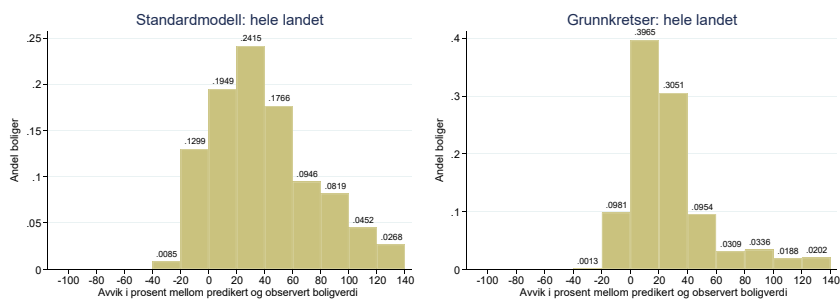
Tabell A.1 viser gjennomsnittlig transaksjonspris (inkludert fellesgjeld), og gjennomsnittlig predikert verdi fra den eksisterende og den nye modellen, for boliger fordelt på fylke, på boligtype, på prisgruppe og på sentralitet. Tabellen bruker 2019 – datasettet.

**Tabell A.2 Sammenligning av transaksjonspriser og predikerte priser, prisgrupper per landsdel, 2019**

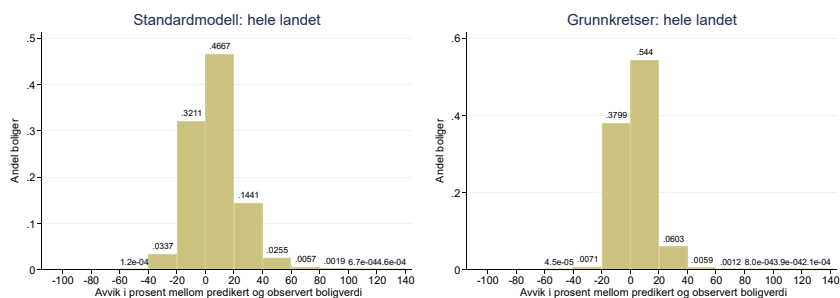
a) Østlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	819 149	1 156 917	1 033 502	10 279
< 2 500 000	1 867 148	2 000 706	1 943 283	185 592
< 5 000 000	3 421 137	3 457 925	3 437 898	213 009
< 7 500 000	5 964 438	5 644 761	5 806 200	43 678
< 10 000 000	8 462 867	7 693 492	8 114 235	13 139
> 10 000 000	13 026 651	10 831 959	11 888 226	7 610
b) Sørlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	823 181	1 270 340	1 114 742	935
< 2 500 000	1 842 711	2 048 343	1 958 100	23 885
< 5 000 000	3 259 852	3 064 489	3 179 766	13 036
< 7 500 000	5 860 665	4 097 562	4 965 612	1 273
< 10 000 000	8 284 194	4 368 965	5 843 260	228
> 10 000 000	12 613 296	4 696 047	6 732 468	69
c) Vestlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	818 179	1 265 640	1 103 581	3 161
< 2 500 000	1 932 869	2 074 337	2 024 121	76 422
< 5 000 000	3 335 941	3 326 269	3 335 880	86 465
< 7 500 000	5 842 809	5 291 084	5 450 497	11 312
< 10 000 000	8 306 696	6 596 866	7 061 127	1 673
> 10 000 000	12 035 195	7 727 381	8 718 903	380
d) Trøndelag	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	788 626	1 181 645	1 007 898	2 039
< 2 500 000	1 916 024	2 031 311	1 983 850	33 489
< 5 000 000	3 293 972	3 273 552	3 276 490	32 786
< 7 500 000	5 854 414	5 256 329	5 543 118	3 662
< 10 000 000	8 333 599	6 514 242	7 503 830	590
> 10 000 000	11 868 735	7 437 723	9 350 594	166
e) Nord-Norge	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	780 528	1 204 599	1 029 016	2 085
< 2 500 000	1 841 309	1 953 494	1 914 043	25 176
< 5 000 000	3 313 632	3 223 793	3 262 574	20 277
< 7 500 000	5 817 397	5 414 784	5 504 011	2 166
< 10 000 000	8 215 121	7 029 448	7 360 574	223
> 10 000 000	11 196 863	8 338 813	8 491 356	51

Tabell A.2 viser gjennomsnittlig transaksjonspris (inkludert fellesgjeld), og gjennomsnittlig predikert verdi fra den eksisterende og den nye modellen, for boliger fordelt på prisgruppe i hver landsdel. Tabellen bruker 2019 – datasettet. Tabellen inkluderer ikke en egen gruppe for boliger solgt for over 20 000 000 kroner, da det er svært få slike boliger utenfor Østlandet.

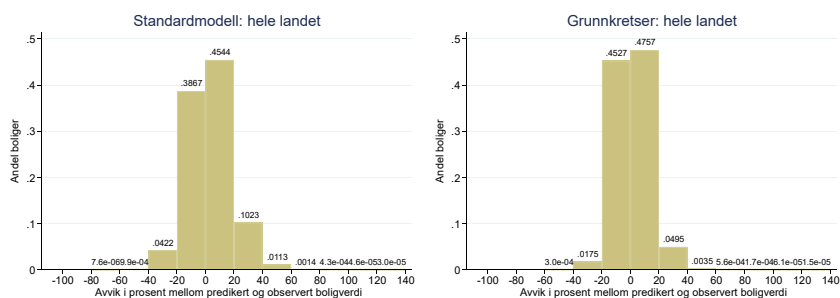
**Figur A.4 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 1, prisgruppe 1, 2019**



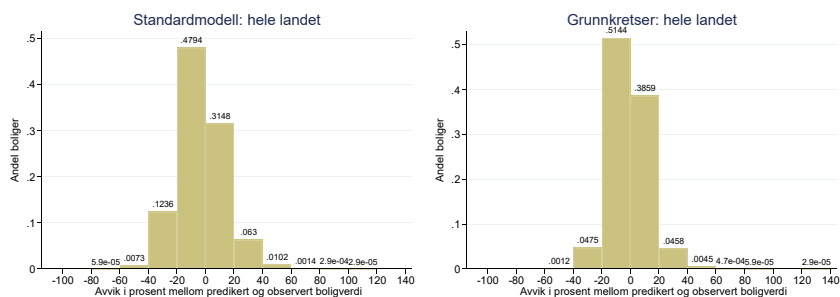
**Figur A.5 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 1, prisgruppe 2, 2019**



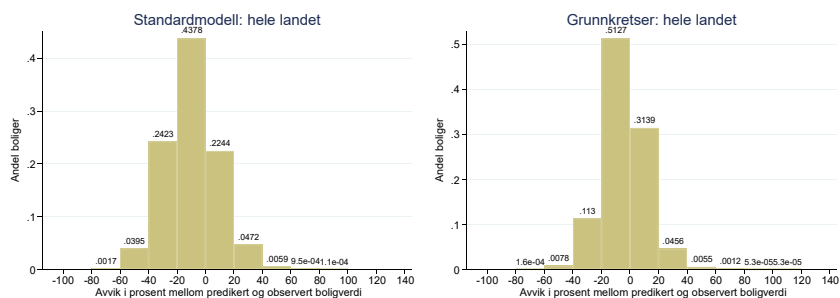
**Figur A.6 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 1, prisgruppe 3, 2019**



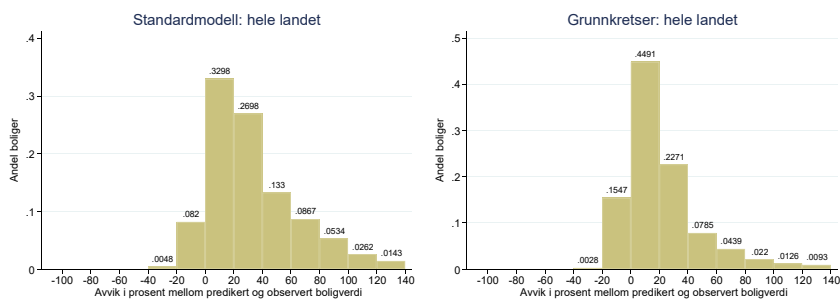
**Figur A.7 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 1, prisgruppe 4, 2019**



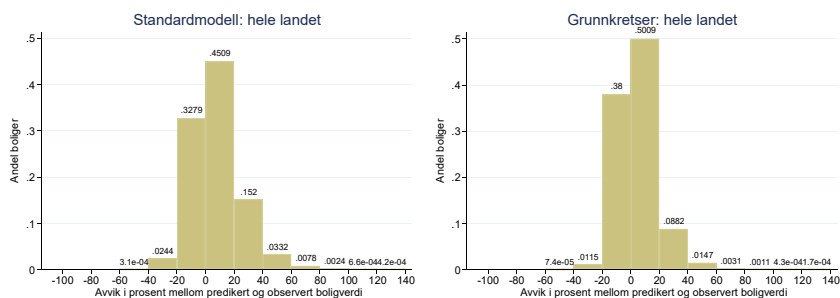
**Figur A.8 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 1, prisgruppe 5, 2019**



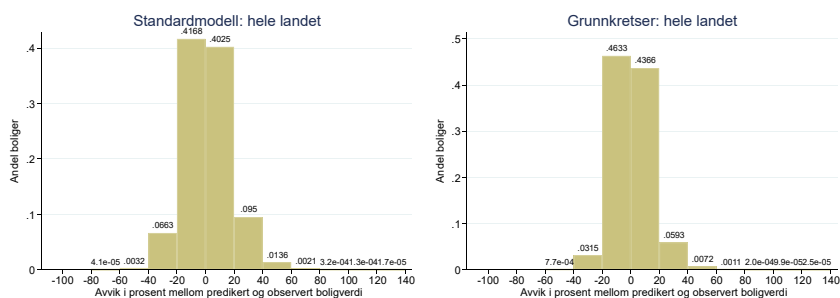
**Figur A.9 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 2, prisgruppe 1, 2019**



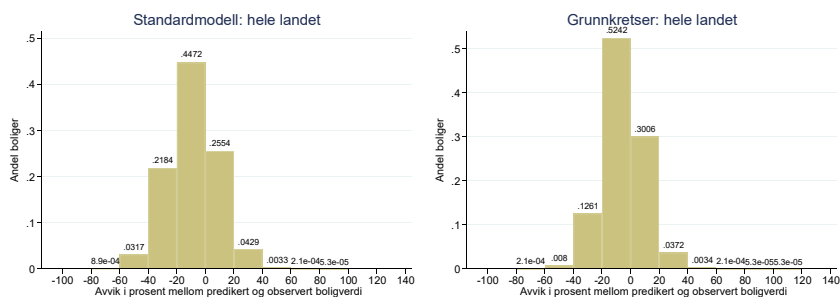
**Figur A.10 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 2, prisgruppe 2, 2019**



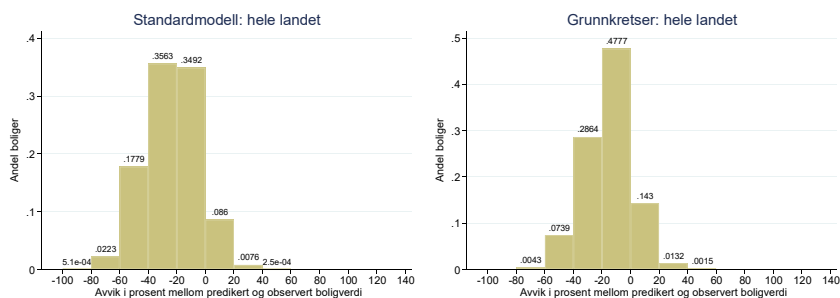
**Figur A.11 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 2, prisgruppe 3, 2019**



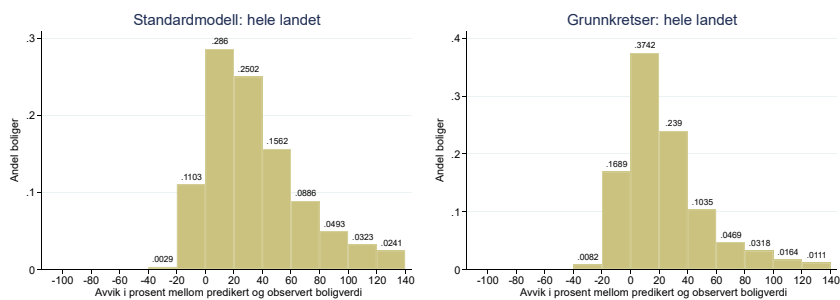
**Figur A.12 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 2, prisgruppe 4, 2019**



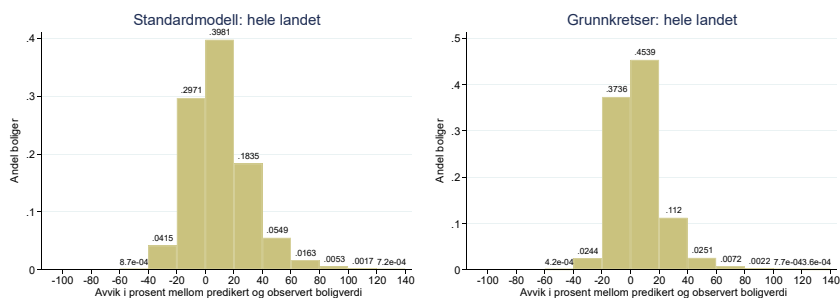
**Figur A.13 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 2, prisgruppe 5, 2019**



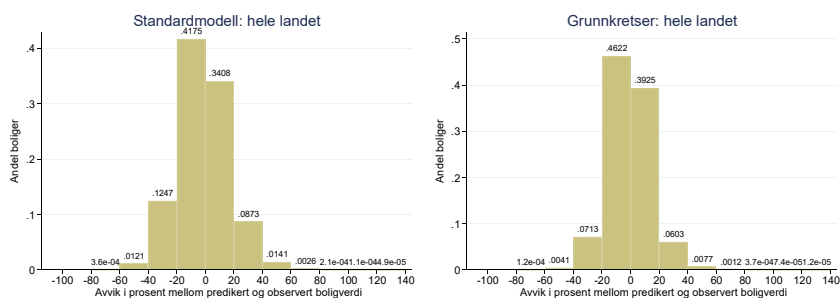
**Figur A.14 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 3, prisgruppe 1, 2019**



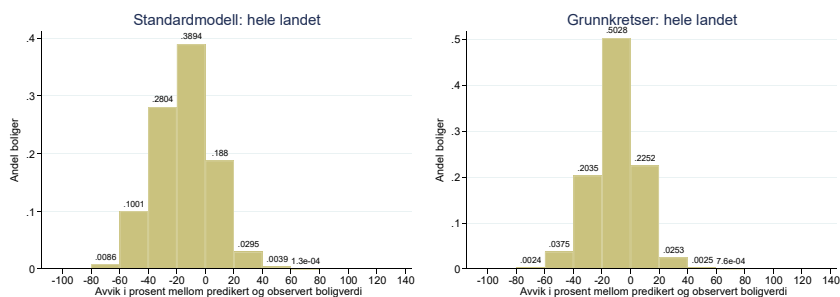
**Figur A.15 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 3, prisgruppe 2, 2019**



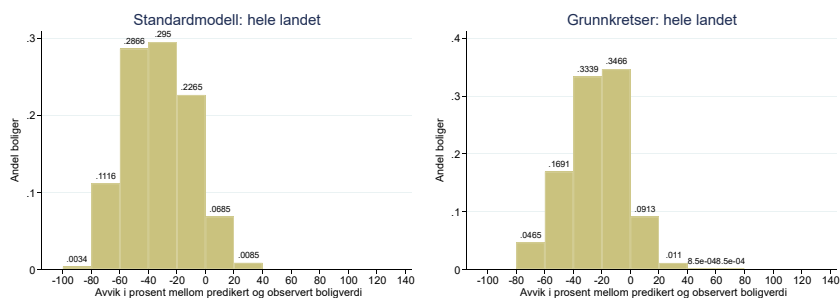
**Figur A.16 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 3, prisgruppe 3, 2019**



**Figur A.17 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 3, prisgruppe 4, 2019**

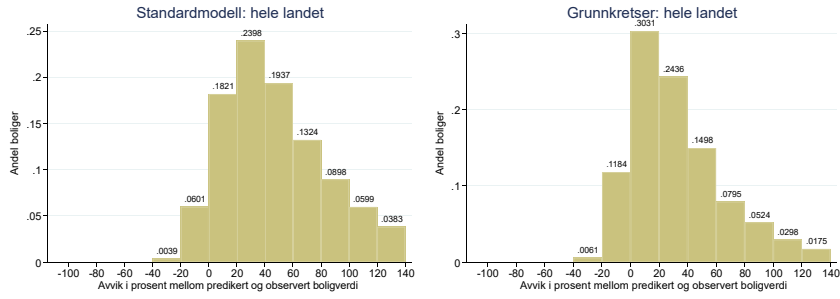


**Figur A.18 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 3, prisgruppe 5, 2019**

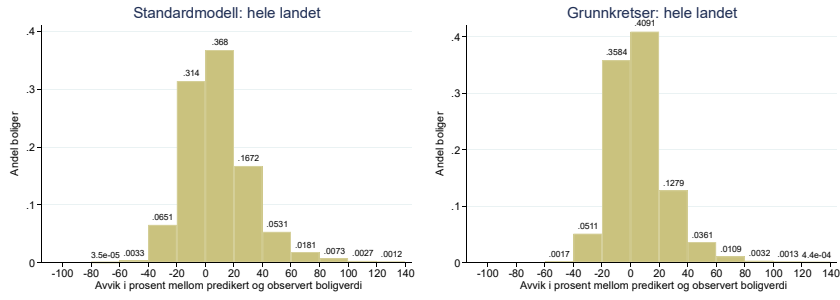




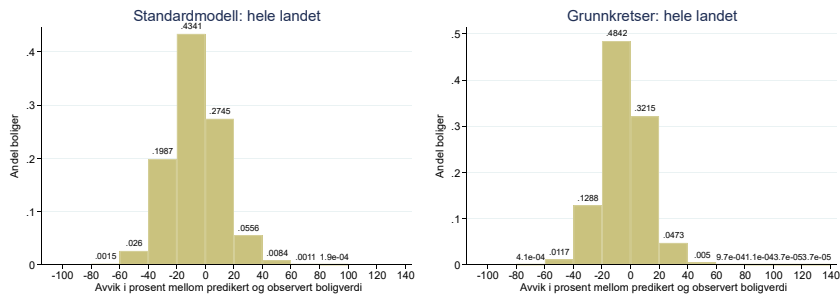
**Figur A.19 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 4, prisgruppe 1, 2019**



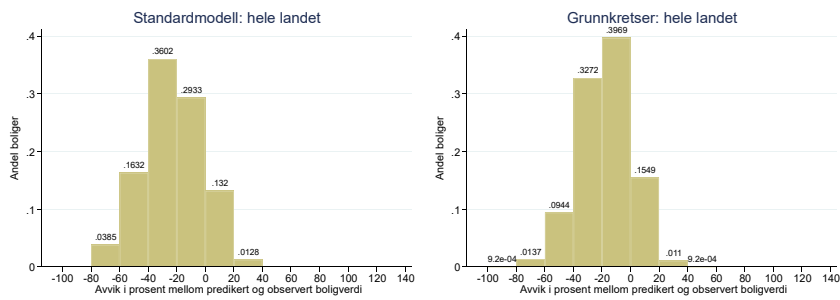
**Figur A.20 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 4, prisgruppe 2, 2019**



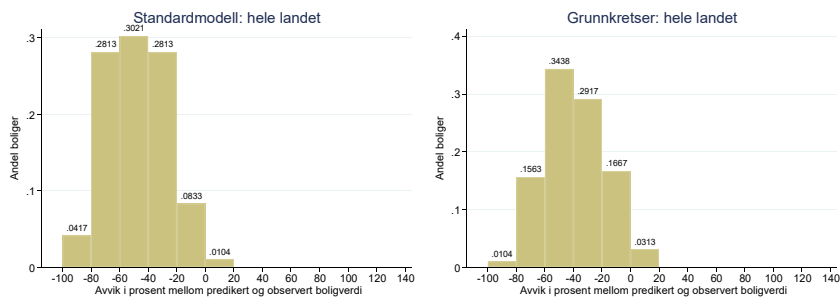
**Figur A.21 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 4, prisgruppe 3, 2019**



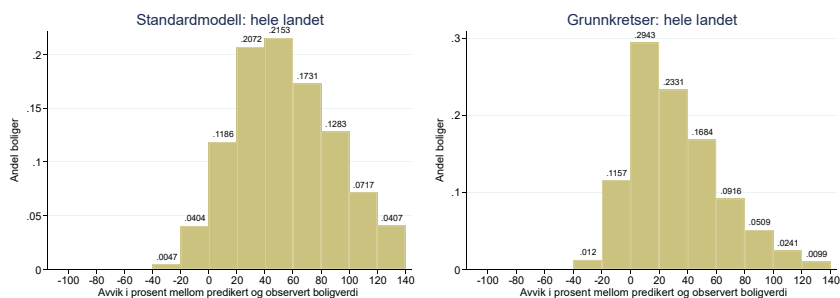
**Figur A.22 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 4, prisgruppe 4, 2019**



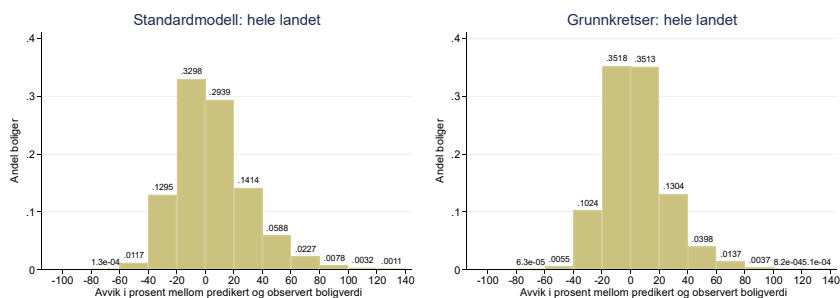
**Figur A.23 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 4, prisgruppe 5, 2019**



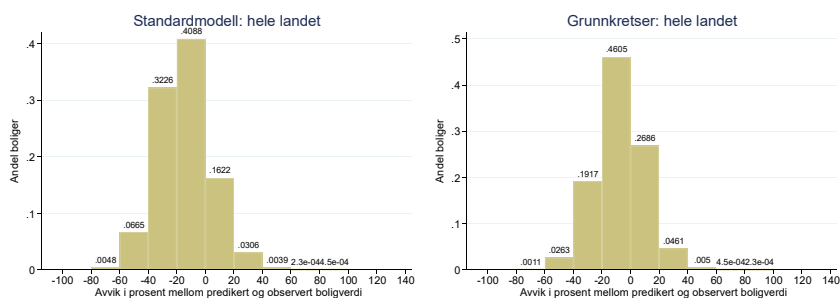
**Figur A.24 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 5, prisgruppe 1, 2019**



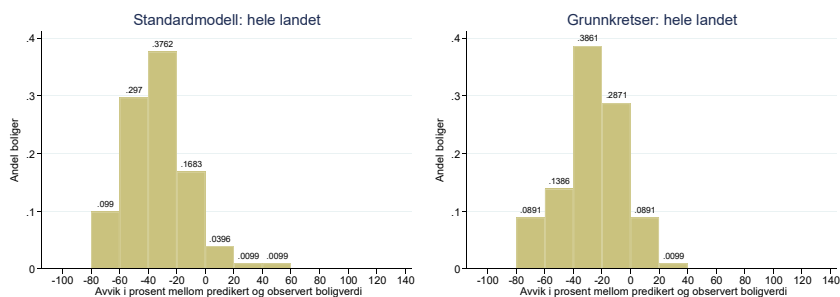
**Figur A.25 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 5, prisgruppe 2, 2019**



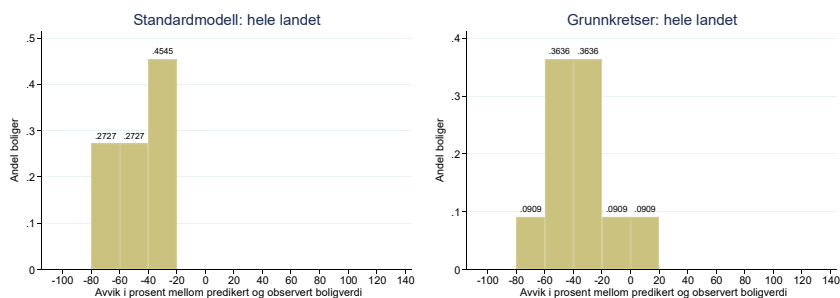
**Figur A.26 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 5, prisgruppe 3, 2019**



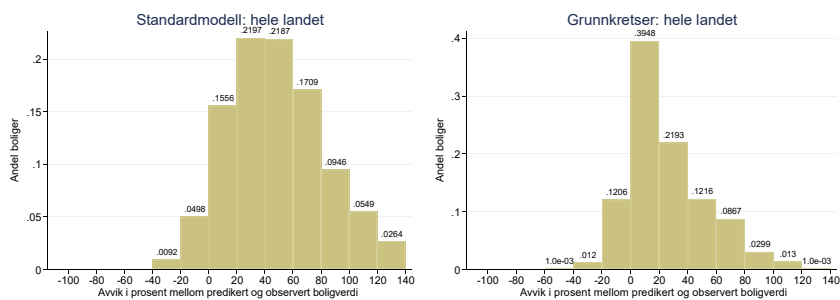
**Figur A.27 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 5, prisgruppe 4, 2019**



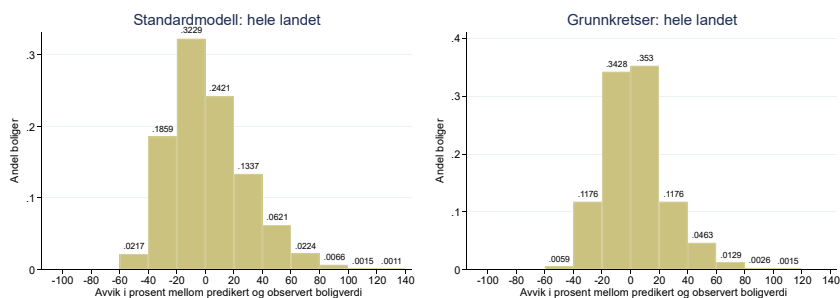
**Figur A.28 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 5, prisgruppe 5, 2019**



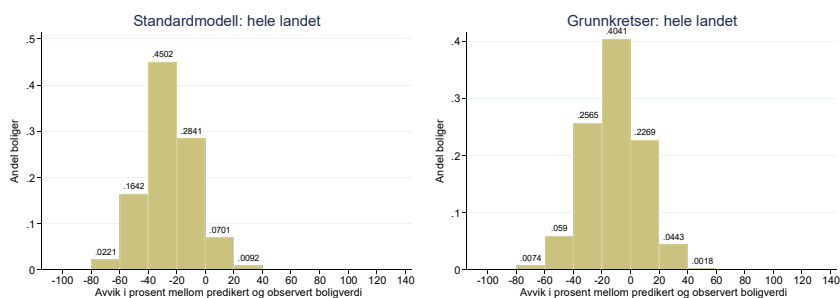
**Figur A.29 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 6, prisgruppe 1, 2019**



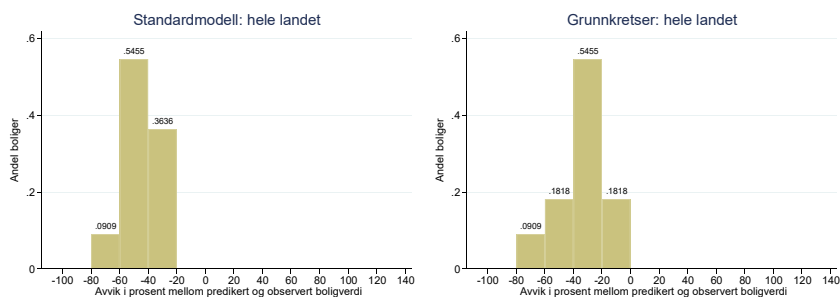
**Figur A.30 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 6, prisgruppe 2, 2019**



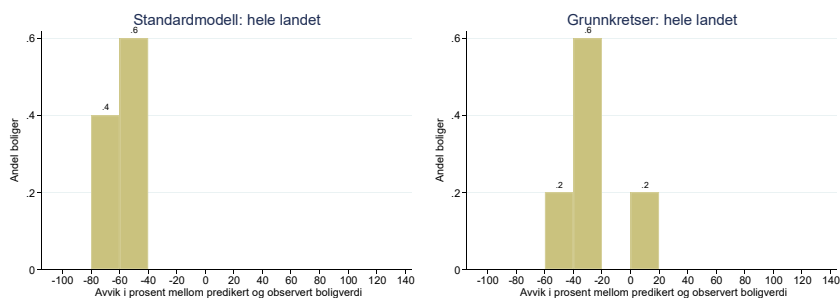
**Figur A.31 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 6, prisgruppe 3, 2019**



**Figur A.32 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 6, prisgruppe 4, 2019**



**Figur A.33 Fordeling av prediksjonsfeil, eksisterende og ny modell, sentralitet 6, prisgruppe 5, 2019**



Figur A.4 – A.33 viser fordelingen av prosentvis prediksjonsfeil, for den eksisterende og den nye modellen, fordelt på fem prisgrupper i seks sentraliteter, i 2019. Sentralitetskategoriene er beskrevet i Høydahl (2017). Prisgruppene er som i tabell A.1 og A.3, foruten at de 3 toppgruppene er slått sammen, slik at alle boliger over 7 500 000 er i en gruppe. Dette da noen av sentralitetene har svært få transaksjoner med høye priser. Prisgruppene er dermed: 1 < 1 000 000; 2 < 2 500 000; 3 < 5 000 000; 4 < 7 500 000; og 5 > 7 500 000.

**Ekstra resultater, 2020****Tabell A.3 Sammenligning av transaksjonspriser og predikerte priser, 2020**

a) Fylke	Transaksjonspris	Eksisterende modell	Ny modell	Antall
Viken	3 412 595	3 400 893	3 413 536	189 995
Oslo	4 202 330	4 182 335	4 198 318	185 234
Innlandet	2 237 515	2 255 931	2 240 398	42 220
Vestfold og Telemark	2 501 796	2 494 433	2 497 880	65 636
Agder	2 533 053	2 509 664	2 530 218	40 248
Rogaland	3 224 255	3 234 553	3 223 406	67 662
Vestlandet	3 018 976	3 010 570	3 018 985	85 329
Møre og Romsdal	2 420 886	2 444 426	2 426 015	30 150
Trøndelag	2 887 923	2 888 204	2 889 214	74 894
Nordland	2 464 392	2 468 124	2 471 684	26 047
Troms og Finnmark	2 982 914	3 007 278	2 988 041	26 275

b) Boligtype	Transaksjonspris	Eksisterende modell	Ny modell	Antall
Enebolig	3 662 012	3 664 823	3 666 058	249 604
Småhus	3 560 363	3 559 015	3 561 917	110 379
Leilighet	2 930 048	2 918 058	2 926 932	473 707

c) Prisgruppe	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	810 655	1 253 588	1 071 579	15 858
< 2 500 000	1 906 816	2 062 893	1 996 390	320 249
< 5 000 000	3 400 672	3 426 424	3 413 114	396 188
< 7 500 000	5 938 024	5 541 669	5 723 600	71 967
< 10 000 000	8 439 501	7 450 659	7 949 451	18 929
< 20 000 000	12 471 504	10 415 482	11 433 788	9 988
> 20 000 000	24 560 992	15 639 729	18 739 684	511

d) Sentralitet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
1	4 184 852	4 146 369	4 181 987	255 230
2	3 219 835	3 218 595	3 218 173	246 491
3	2 705 752	2 715 024	2 705 708	207 364
4	2 283 114	2 302 710	2 289 892	93 832
5	1 921 210	1 959 314	1 929 915	25 681
6	1 693 345	1 744 618	1 692 585	5 092

Tabell A.3 viser gjennomsnittlig transaksjonspris (inkludert fellesgjeld), og gjennomsnittlig predikert verdi fra den eksisterende og den nye modellen, for boliger fordelt på fylke, på boligtype, på prisgruppe og på sentralitet. Tabellen bruker 2020 – datasettet.

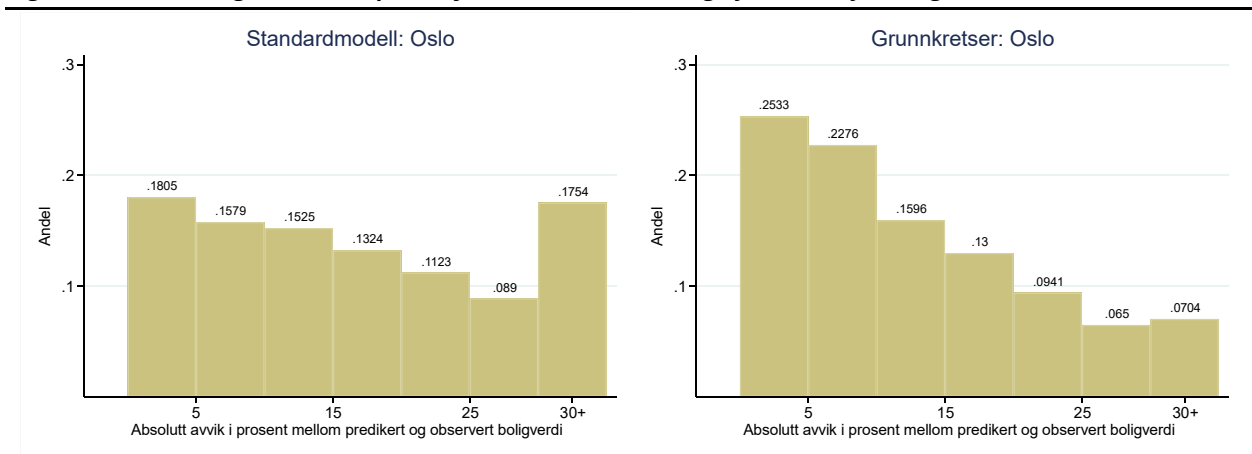
**Tabell A.4 Sammenligning av transaksjonspriser og predikerte priser, prisgrupper per landsdel, 2020**

a) Østlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	819 541	1 212 735	1 056 716	8 353
< 2 500 000	1 892 942	2 043 941	1 977 462	167 133
< 5 000 000	3 451 588	3 519 143	3 480 658	231 346
< 7 500 000	5 972 777	5 664 017	5 827 385	50 948
< 10 000 000	8 463 514	7 651 702	8 117 767	15 648
> 10 000 000	13 150 892	10 940 200	12 041 409	9 657
b) Sørlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	821 919	1 307 451	1 138 801	876
< 2 500 000	1 854 672	2 088 961	1 976 729	23 578
< 5 000 000	3 269 231	3 075 504	3 198 398	13 990
< 7 500 000	5 865 701	4 136 305	5 032 375	1 460
< 10 000 000	8 319 293	4 429 611	6 038 945	271
> 10 000 000	12 957 294	4 717 149	6 875 020	73
c) Vestlandet	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	814 813	1 325 249	1 127 923	2 884
< 2 500 000	1 952 251	2 114 674	2 050 598	73 288
< 5 000 000	3 346 033	3 340 725	3 352 165	91 971
< 7 500 000	5 854 040	5 303 310	5 475 925	12 546
< 10 000 000	8 319 255	6 646 171	7 137 515	1 983
> 10 000 000	12 044 853	7 842 107	8 869 160	469
d) Trøndelag	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	790 270	1 260 708	1 035 054	1 812
< 2 500 000	1 943 064	2 072 684	2 019 859	31 840
< 5 000 000	3 308 778	3 298 423	3 297 326	35 990
< 7 500 000	5 855 619	5 311 486	5 563 455	4 314
< 10 000 000	8 371 267	6 586 762	7 509 605	709
> 10 000 000	11 832 789	7 582 991	9 611 999	229
e) Nord-Norge	Transaksjonspris	Eksisterende modell	Ny modell	Antall
< 1 000 000	780 058	1 292 126	1 055 515	1 933
< 2 500 000	1 868 482	1 999 245	1 951 612	24 410
< 5 000 000	3 330 436	3 249 401	3 288 628	22 891
< 7 500 000	5 843 235	5 468 262	5 545 648	2 699
< 10 000 000	8 262 281	7 075 117	7 338 934	318
> 10 000 000	11 454 437	8 639 736	8 874 793	71

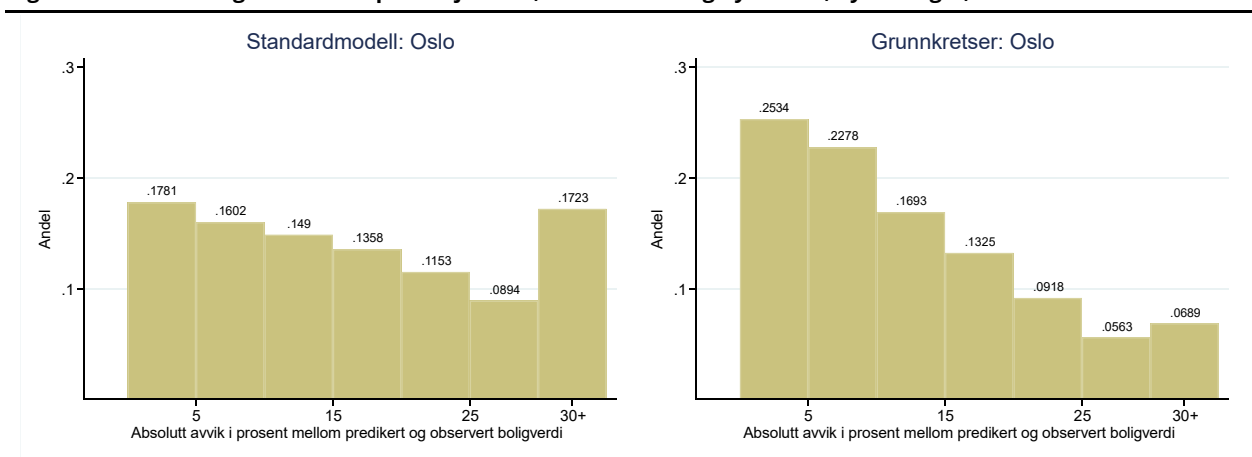
Tabell A.4 viser gjennomsnittlig transaksjonspris (inkludert fellesgjeld), og gjennomsnittlig predikert verdi fra den eksisterende og den nye modellen, for boliger fordelt på prisgruppe i hver landsdel. Tabellen bruker 2020 - datasettet. Tabellen inkluderer ikke en egen gruppe for boliger solgt for over 20 000 000 kroner, da det er svært få slike boliger utenfor Østlandet.

## Dyre boliger

**Figur A.34 Fordeling av absolutt prediksjonsfeil, eksisterende og ny modell, dyre boliger, 2019**



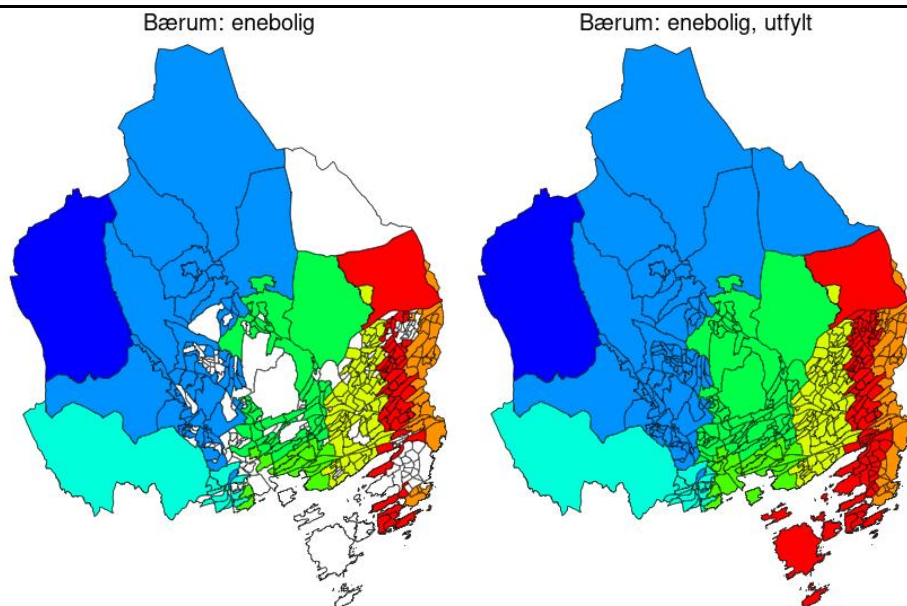
**Figur A.35 Fordeling av absolutt prediksjonsfeil, eksisterende og ny modell, dyre boliger, 2020**



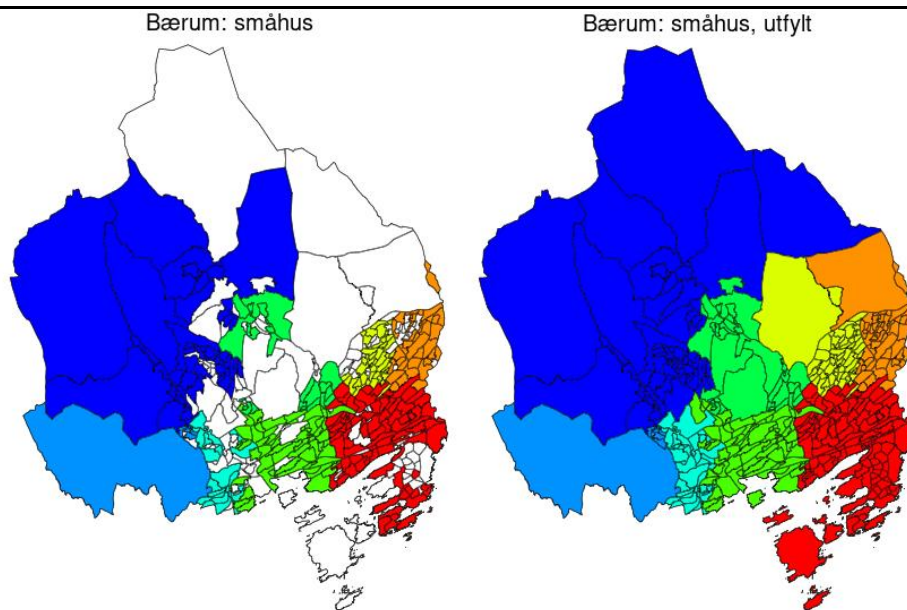
Figur A.34 – A.35 viser fordelingen av absolutte prosentvis prediksjonsfeil, for dyre boliger i Oslo i den eksisterende og nye modellen, i 2019 og 2020. Den siste stolpen i hver figur, merket 30+, inneholder alle observasjoner med absolutt prediksjonsfeil større enn 30 prosent.

## Fordeling av grunnkretser uten observasjoner

**Figur A.36** Kart over grunnkretser med predikert kvadratmeterpris, uten og med utfylling av grunnkretser uten observasjoner, enebolig, Bærum, 2020



**Figur A.37** Kart over grunnkretser med predikert kvadratmeterpris, uten og med utfylling av grunnkretser uten observasjoner, småhus, Bærum, 2020





**Figur A.38** Kart over grunnkretser med predikert kvadratmeterpris, uten og med utfylling av grunnkretser uten observasjoner, leilighet, Bærum, 2020

