De Økonomiske Råd

d. 06.10.2023
JLC

Baggrundsnotat til analyse om huslejeregulering og fordeling

Notatet beskriver data, metode og resultater for analysen om huslejeregulering og fordeling, som indgår i afsnit IV.5 af De Økonomiske Råds formandskab (2023).

1 Indledning

Huslejereguleringen i Danmark sætter en række regler for, hvordan huslejen skal fastsættes, og formålet er at sænke huslejen i forhold til en husleje, der er sat på markedsvilkår. Huslejeregulering begrundes ofte med fordelingspolitiske hensyn. Reguleringen skal sænke huslejen blandt andet for at give bedre mulighed for, at alle typer af husstande kan bo i billigere boliger, særligt i de største byer.

Dette notat uddyber de analyser, der er foretaget i forbindelse med analyserne af fordelingen af besparelsen i huslejen som følge af huslejereguleringen, de såkaldte reguleringsgevinster, i afsnit IV.5 i De Økonomiske Råds formandskab (2023). I analysen blev det undersøgt, om gevinsten ved lavere husleje i ældre (regulerede) private lejeboliger (opført frem til 1991) tilfalder husstande med lav indkomst. Hovedkonklusionen er, at huslejereguleringen ikke synes at være målrettet husstande med lav indkomst. For det første er den årlige reguleringsgevinst i ældre lejeboliger højere for husstande med høje indkomster. For det andet er der store forskelle i gevinsten for husstande med omtrent samme indkomst.

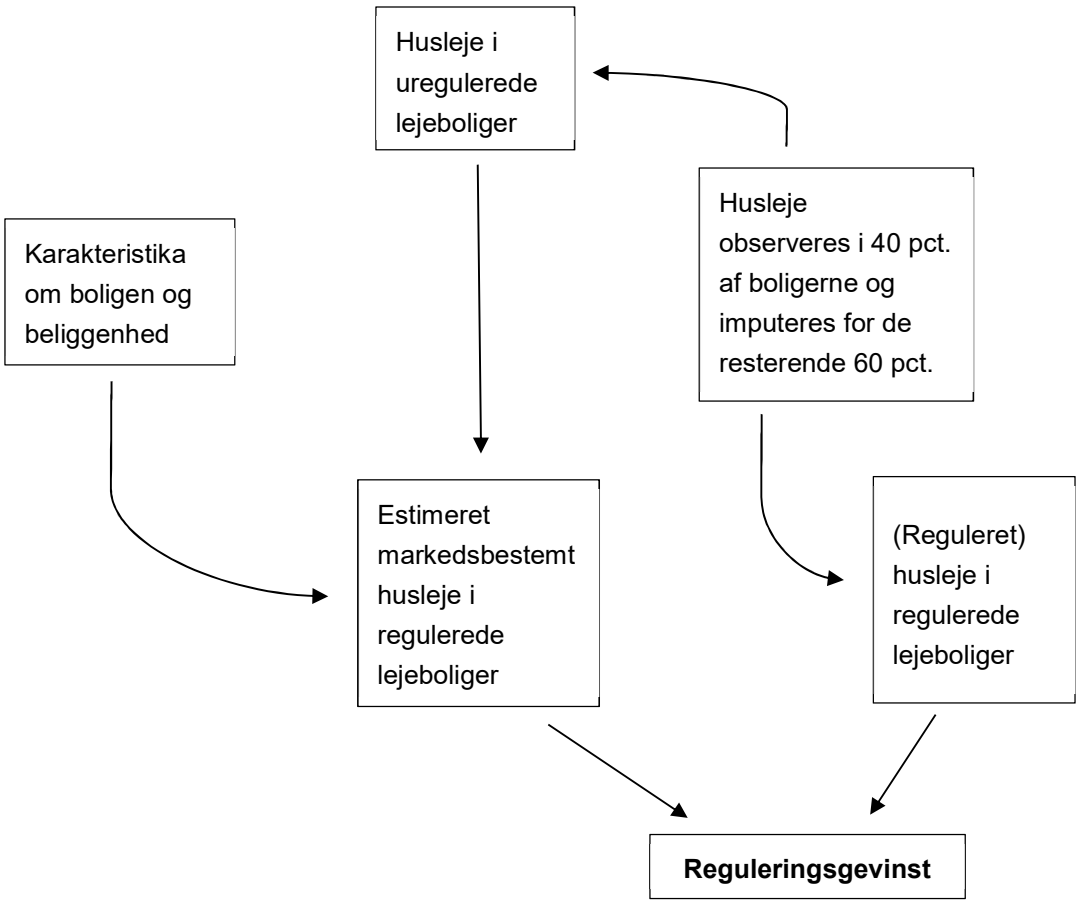
Reguleringsgevinsten er forskellen på den faktiske (regulerede) husleje og den markedsbestemte husleje i fravær af huslejereguleringen. Reguleringsgevinsten skal beregnes, da den markedsbestemte husleje ikke observeres for de regulerede lejeboliger, jf. figur 1. Den markedsbestemte husleje estimeres på baggrund af huslejen samt bolig- og områdekarakteristika i nyudlejede boliger opført efter 1991 (uregulerede lejeboliger), da huslejen i disse boliger aftales frit. Huslejen i både ældre (regulerede) lejeboliger og nyere (uregulerede) boliger observeres for ca. 40 pct. af boligerne, mens den imputeres for de resterende 60 pct.

Afsnit 2 beskriver data, som er brugt i analysen.

Afsnit 3 redegør for metoden til at beregne reguleringsgevinsterne. Fokus er i høj grad på at beskrive estimationen af den markedsbestemte husleje for ældre lejeboliger. Dertil vises en række mål for kvaliteten af estimationen af den markedsbestemte husleje.

Afsnit 4 viser først udviklingen i den estimerede markedsbestemte husleje for ældre lejeboliger. Herefter præsenteres analysens hovedresultater, inden afsnittet afslutningsvis viser en række følsomhedsberegninger.

Figur 1: Oversigt over beregning af reguleringsgevinster



2 Data

Analyserne bygger på registerdata fra Danmarks Statistik i perioden 2010-21. Der fokuseres på husstande i private lejeboliger, der er defineret som boliger beboet af en *lejer* og ejet af privatpersoner eller selskaber. Almene boliger, der er lejeboliger ejet af almene boligselskaber, indgår som sammenligningsgrundlag for fordelingsvirkningerne for private lejeboliger. Kollegier og offentlig ejede lejeboliger indgår ikke i analysen, da disse boliger er atypiske i forhold til det generelle lejemarked. Det er ikke muligt ud fra data fra Danmarks Statistik at identificere, hvilke boliger som er udlejede i perioden 2019-21. Det antages derfor, at en bolig er en lejebolig i 2019-21, hvis den var en lejebolig i 2016-18. Det gør det muligt at identificere op til 90 pct. af alle lejeboligerne i 2019-21.[1] Analysen fokuserer på husstande, der er personer med samme bopælsadresse i slutningen af året.

Data om lejeboligernes fysiske og geografiske karakteristika er centralt i at estimere den markedsbestemte husleje, jf. figur 1 og afsnit 3. Oplysninger om boligens karakteristika

[1] Denne beregning bygger på, hvor stor en del af lejeboliger i 2018 (hvor lejestatus kan observeres), der også var lejet ud i perioden 2015-17. Bemærk at boliger bygget fra 2019-21 ikke kan indgå i analysen, da det ikke på noget tidspunkt kan observeres, om disse boliger er lejeboliger.

og beliggenhed stammer fra BBR-registret. Der er også brugt data om miljøfaktorerne, som støj, luftforurening og nærhed til natur. Disse data stammer fra en analyse i De Økonomiske Råds formandskab (2019). Se boks 1 for en deltaljeret beskrivelse af analysens data.

BOKS 1 DATA I ANALYSEN

Analysens primære datakilde er registerdata fra Danmarks Statistik i perioden 2010-21. Denne boks giver detaljeret overblik over datakilderne og databehandlingen.

Huslejeoplysninger
Huslejen observeres på månedsniveau, og som udgangspunkt bruges huslejen fra december måned. Huslejeoplysningerne stammer fra tre kilder. For det første bygger huslejen for private lejeboliger i perioden 2010-21 primært på boligstøtteregisteret. Boligstøtteregistret bygger på data fra Udbetaling Danmark, og datakvaliteten for huslejeoplysningerne vurderes derfor til at være høj. I 2018 er huslejedata mangelfulde, da udbetalingerne overgik fra KMD til ATP, og 2018 indgår af denne årsag ikke i analysen. For det andet suppleres huslejeoplysningerne for ældre private lejeboliger i årerne 2014-21 med data fra EjendomDanmark, som har indsamlet huslejen for et større udsnit af deres medlemsorganisationer. Disse huslejeoplysninger betragtes som værende af høj kvalitet og er i god overensstemmelse med huslejen fra boligstøtteregistret for de boliger, som optræder i begge kilder, jf. bilag A. For det tredje stammer oplysninger om huslejen for almene boliger fra 2012 og frem fra registret om almene boliger.[a)]

Boligkarakteristika
Karakteristika om boligen og dens beliggenhed stammer fra BBR-registret. Disse oplysninger bruges til at estimere den markedsbestemte husleje, jf. afsnit 3. Udlejers udgifter til ejendomsskat påvirker huslejen, og ejendomsskatten indgår derfor også i estimationen af den markedsbestemte husleje. Oplysninger om lejernes andel af ejendomsskatten stammer fra IND-registret. Oplysningerne er på personniveau, så den samlede ejendomskat i en bolig fås ved at lægge skatten for alle beboere i en lejebolig sammen.

Miljøkarakteristika
Miljøfaktorerne; støj, luftforurening og nærhed til natur, indgår også i estimationen af den markedsbestemte husleje. Data stammer fra en analyse i De Økonomiske Råds formandskab (2019). Det anvendte datasæt har kun oplysninger om miljøfaktorerne for boliger bygget senest i 2016. Det betyder, at miljøfaktorerne ikke observeres for boliger opført herefter. For disse boliger imputeres miljøfaktorerne ud fra den gennemsnitlige værdi for de øvrige boliger i samme *grundværdisområde*. Grundværdisområderne er geografisk opdeling af hele Danmark foretaget af SKAT.[b)] Hvert områder er karakteriseret ved, at der i området er nogenlunde samme karakteristika og afstande til forskellige aktiviteter. Danmark er opdelt i ca. 65.000 grundværdiområder.

[a)] Huslejen i almene boliger før 2013 stammer fra boligstøtteregistret.

[b)] SKAT har senest grundværdisområderne i 2011.

BOKS 1 **DATA I ANALYSEN, FORTSAT**

Husstandskarakteristika
Husstande defineres som personer med samme bopælsadresse i slutningen af året. Husstandens årlige disponible indkomst stammer fra IND-registret. Husstandene inddeles i ti grupper ud fra deres placering i fordelingen af den ækvivalerede disponible indkomst for alle husstande i Danmark. Gennemsnitsindkomsten og antal husstande i ældre lejeboliger i hver indkomstgruppe fremgår af tabel A.

Antallet af år siden indflytning bruges til at dele nyere lejeboliger op i husstande, som har boet højest to år i boligen, og husstande der har boet mere end to år. Huslejen i nyere lejeboliger aftales fri ved indflytning og kan efterfølgende stige med nettoprisindekset. Lejen indenfor de første to år kan derfor opfattes som en markedsbestemt husleje. For hvert år opgøres tiden siden indflytning ud fra BEF-BOP-registret. Variablen viser antallet af år, måneder og dage fra indflytning til 31/12 i det respektive år. Hvis flere personer bor i samme bolig og er flyttet ind på forskellige tidspunkter, bruges indflytningen for den person, som har boet længst tid i boligen.

Øvrige husstandskarakteristika om socioøkonomisk status stammer fra registrene BEF, UDDA og RAS.

Tabel A: Gennemsnitlig indkomst i indkomstgrupper, ældre lejeboliger

Indkomst-gruppe	Gns. disp. indkomst (1.000 kr.)	Antal tusinde husstande
1	103	60,7
2	162	56,5
3	186	42,4
4	211	41,3
5	239	38,2
6	271	32,2
7	307	25,3
8	352	19,3
9	416	14,3
10	770	10,5

Kilde: Egne beregninger på baggrund af registerdata

Salgspriser for ejerboliger
Salgspriserne på ejerboliger bruges ifm. estimationen af den markedsbestemte husleje, jf. afsnit 3.2. Der bruges kun salg mellem privatpersoner. Derudover fjernes salg som ikke er 'repræsentative' for det generelle boligmarked. Det gøres ved at udelade salg, hvor købesummen er mindre 300.000 og større end 30 mio. kr., jf. Nationalbanken (2023).

Private lejeboliger inddeles i to grupper: ældre lejeboliger, der er opført frem til 1991, og nyere lejeboliger, der er opført efter 1991, jf. faktaboksen. I ældre lejeboliger reguleres både den initiale husleje og stigningen i huslejen. I analyserne fokuseres på fordelingen af reguleringsgevinsterne for ældre lejeboliger. Huslejen i nyere lejeboliger aftales frit ved nyindflytning og kan herefter højest stige med nettoprisindekset. Huslejen ved indflytning er derfor det tætteste, man kan komme på den markedsbestemte husleje. Nyere lejeboliger underopdeles i to grupper: nyudlejede og ikke-nyudlejede boliger opført efter 1991. En bolig er nyudlejet, hvis lejerne har boet i boligen i højest to år. I 2021 var der ca. 372.000 ældre lejeboliger og 71.000 nyere lejeboliger, heraf var 28.000 nyudlejet.

NOMENKLATUR FOR PRIVATE LEJEBOLIGER

Der skelnes mellem to overordnede kategorier af private lejeboliger:

- Ældre boliger
- Nyere boliger.

Ældre boliger omfatter boliger opført frem til 1991. Nyere boliger omfatter boliger opført efter 1991. Nyere boliger underopdeles i to grupper:

- Nyudlejede boliger opført efter 1991
- Ikke-nyudlejede boliger opført efter 1991.

Nyudlejede boliger opført efter 1991 omfatter boliger, hvor lejeren har boet i op til to år. Lejen i disse bolig kan opfattes som en markedsbestemt husleje.

2.1 Imputation af husleje

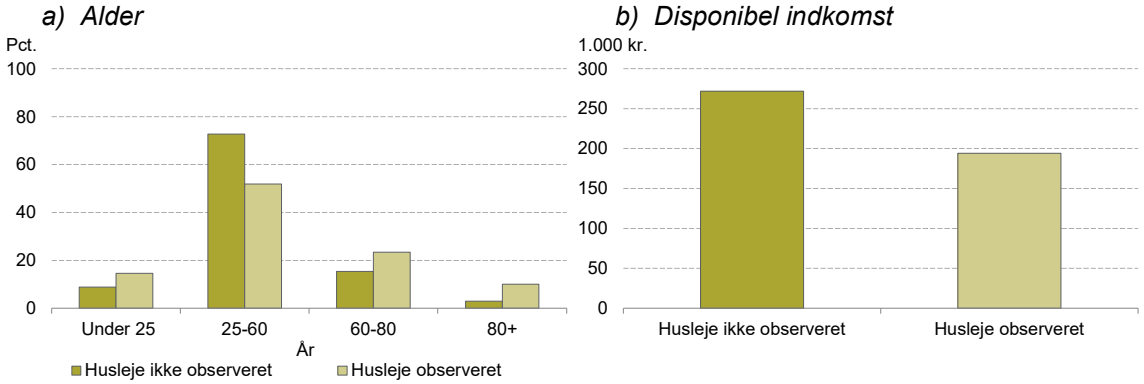
Huslejen er en central variabel i at beregne besparelsen i huslejen som følge af huslejereguleringen. For en del private lejeboliger observeres huslejen dog ikke. I det følgende beskrives, dels hvilke datakilder der er huslejen, og dels metoden til at beregne huslejen for lejeboliger, hvor lejen ikke observeres.

Huslejen observeres på månedsniveau, og som udgangspunkt bruges huslejen fra december måned. Huslejeoplysningerne i private lejeboliger stammer fra to kilder. For det første bygger huslejen for private lejeboliger i perioden 2010-21 primært på boligstøtteregisteret. Boligstøtteregistret bygger på data fra Udbetaling Danmark. I 2018 er huslejedata mangelfulde, da udbetalingerne overgik fra KMD til ATP, og 2018 indgår af denne årsag ikke i analysen. For det andet suppleres huslejeoplysningerne i årerne 2014-21 med data fra EjendomDanmark, som har indsamlet huslejen for et større udsnit af deres medlemsorganisationer.

Samlet set observeres huslejen fra de tre kilder for ca. 40 pct. af alle private lejeboliger og for ca. 65 pct. af alle lejeboliger inkl. almene boliger i perioden 2010-21. Huslejen observeres for en lidt større andel af lejeboligerne i de seneste år.

Boligerne, hvor huslejen er observeret, er dog ikke repræsentativ for alle private lejeboliger, jf. figur 2. Husstandene i disse boliger har i gennemsnit en lavere indkomst. Dertil er de enten ældre eller yngre end husstandene, der bor i boliger uden observeret husleje. Det skyldes, at huslejen i private lejeboliger primært observeres, hvis husstanden modtager boligstøtte, der i højere grad gives til husstande med lavere indkomst og folkepensionister, jf. De Økonomiske Råds formandskab (2023).

Figur 2: Husstandskarakteristika opdelt på om husleje observeres i data



Anm.: Figuren i panel a) viser fordelingen af husstandene ud fra alderen på den ældste person, mens figuren i panel b) viser den gennemsnitlige husstandens samlede disponible indkomst.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

For at sikre at analysen kan gennemføres på et repræsentativt udsnit af husstandene boende i lejeboliger imputeres huslejen i de boligerne, hvor den ikke observeres. Det gøres i overensstemmelse med metoden i Kristensen (2012), jf. boks 2. Imputationen har to dele. For husstande i lejeboliger, som bor i en bygning, hvor huslejen er observeret i mindst én anden bolig i bygningen, bruges den gennemsnitlige husleje pr. kvm for de husstande, hvor huslejen er observeret. Herved beregnes huslejen for ca. 30 pct. af alle husstande i private lejeboliger. Metoden har den fordel, at boliger i samme bygning har samme beliggenhed, opførelsesår og ca. samme stand, der alle er centrale faktorer i lejefastsættelsen.

For de resterende 30 pct. af husstandene, som bor i bygninger uden andre boligstøttemodtagere, beregnes huslejen ved hjælp af Heckmans to trins selektionsmodel (Heckman (1979)), jf. boks 2. Denne metode tager højde for den skævhed, der er i boligerne, hvor huslejen ikke observeres, jf. figur 2. I første trin estimerer modellen sandsynligheden for, at en husstand modtager boligstøtte. I andet trin bruges en OLS-model til at estimere huslejen ud fra en række karakteristika om boligen og dens beliggenhed. Den beregnede sandsynlighed for at modtage boligstøtte fra første trin inkluderes som en forklaringsfaktor i en OLS-modellen. På den måde tages der højde for den selektion, der er i boliger, hvor huslejen observeres.

BOKS 2 **IMPUTATION AF HUSLEJE**

Huslejen imputeres for boliger, hvor huslejen ikke observeres. Det gøres i to trin, jf. Kristensen (2012).

Gennemsnitlig husleje i bygningen
For boliger, hvor huslejen kendes for mindst én anden bolig i bygningen, bruges den gennemsnitlige husleje pr. kvadratmeter for alle andre boliger i bygning. På den måde imputeres huslejen for halvdelen af boligerne, hvor huslejen ikke observeres. Den grundlæggende antagelse ved denne metode er, at boliger i samme bygning har samme beliggenhed, opførelsesår og ca. samme stand, der alle er centrale faktorer i lejefastsættelsen. En mulig problemstilling ved denne metode er, at den ikke tager højde for, at boliger i samme bygning kan være reguleret forskelligt. Nogle kan have omkostningsbestemt husleje, mens andre kan være gennemgribende renoveret og derved have det lejedes værdi, der er en lempeligere form for regulering.

Heckmans to-trins model
Det er ikke alle boliger, der ligger i en bygning, hvor huslejen observeres i mindst én bolig. Huslejen i disse boliger imputeres ved brug af Heckmans to-trins selektionsmodel (Heckman (1979)). Denne model kan på baggrund af en række husstands- og boligkarakteristika estimere den faktiske husleje, hvor der tages højde for selektionsproblemerne i de observerede faktiske huslejer. Modellen estimeres separat for hvert år (t) og for hhv. ældre og nyere private lejeboliger (r), idet lejefastsættelsen er forskellig i disse to lejeboligtyper.

I trin 1 estimeres en probit model for sandsynligheden for, at en husstand modtager boligstøtte, og huslejen derved observeres:

$$BS_{it}^{r} = \delta^{r} z_{it}^{r} + u_{it}^{r}$$

hvor BS_{it}^{r} er en dummy for, om husstanden modtager boligstøtte, og z_{it}^{r} er en række husstandskarakteristika, der bestemmer, om man modtager boligstøtte. Konkret inkluderer z_{it}^{r}: husstandens indkomst (før skatter og overførelser), antal børn, om der er en folkepensionist i husstanden samt lejlighedstypen, kommunen og om boligen ejes af et selskab eller privat person.

I trin 2 estimeres en OLS-model på baggrund af den observerede husleje i de boliger, der modtager boligstøtte, hvor der tages højde for selektion vha. af et korrektionsled, λ:

$$husleje_{it}^{r} = \beta^{r} x_{it}^{r} + \pi^{r} \lambda(\delta^{r} z_{it}^{r}) + \varepsilon_{it}^{r}$$

hvor huslejen måles pr. kvadratmeter, og x_{it}^{r} er en rækkehusstandskarakteristika: boligens areal, tagtypen, opvarmningskilden, opførelsesåret, ejerforhold, boligens anvendelse (lejlighed, rækkehus eller parcelhus) og postnummeret, som boligen ligger i. For nyere lejeboliger inkluderes også, hvor længe husstanden har boet i boligens. $\lambda(\delta^{r} z_{it}^{r}) = \frac{\phi(\delta^{r} z_{it}^{r})}{\Phi(\delta^{r} z_{it}^{r})}$ er *Mills ratio* (ϕ og Φ er hhv. PDF og CDF for en standardnormalfordeling).

Huslejen imputeres som $\widehat{husleje}_{it}^{r} = \hat{\beta}^{r} x_{it}^{r} + \hat{\pi}^{r} \lambda(\hat{\delta}^{r} z_{it}^{r})$.

Som vist i tabel 1 er der en vis usikkerhed i imputationen af huslejen. En robusthedsanalyse viser dog, at de overordnede konklusioner ikke ændres ved kun at bruge boliger, hvor huslejen observeres, jf. afsnit 4.4.

Prædiktionskvalitet af imputation af huslejen

Kvaliteten af metoden til imputation af huslejen kan undersøges. Det gøres ved at sammenligne den imputerede husleje med den observerede husleje for boliger, hvor huslejen observeres. Sammenligningen foretages for 30 pct. af lejeboligerne, der ikke indgik i imputationsberegningen (en såkaldt out-of-sample prædiktion), men hvor huslejen observeres. På den måde undersøges imputationskvaliteten på data, som ikke er indgået

i imputationsberegningerne, hvilket giver et mere retvisende billede af præcisionen for boliger, hvor huslejen ikke observeres.[2]

De imputerede huslejer er forholdsvis tætte på den faktiske husleje. R^2 ligger mellem 55-75 pct. og mellem 70-90 pct. er indenfor 20 pct., jf. tabel 1. Metoden med at bruge den gennemsnitlige husleje i bygningen synes at være bedre end Heckman-metoden, idet R^2 samt andelen indenfor 10 og 20 pct. er størst ved førstnævnte metode.[3] Det vurderes derfor, at begge metoder til at imputere huslejen har en tilfredsstillinde præcision. En robusthedsanalyse viser derudover, at hovedkonklusionen om fordelingen af reguleringsgevinsterne i De Økonomiske Råds formandskab (2023) er den samme, hvis analysen udføres på det udsnit af lejeboliger, hvor huslejen kan observeres, jf. afsnit 4.4.

Tabel 1: Prædiktionskvalitet af huslejeimputation, 2021

Metode	R^2	Andel indenfor 10 pct.	Andel indenfor 20 pct.
Gns. husleje i bygning	0,76	0,82	0,91
Heckman to-trins model	0,55	0,41	0,69

Anm.: Tabellen viser prædiktionskvaliteten af huslejeimputationen. Tabellen bygger på 30 pct. af private lejeboliger, som ikke indgik i imputationsberegningen, altså en out-of-sample prædiktion. $R^2 = \frac{MSE}{var(husleje)}$, hvor MSE er summen af kvadrerede fejlled. Andel indenfor 10 (20) pct. viser, hvor stor en andel af observationerne som er indenfor 10 (20) pct. af den observerede husleje.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

3 Beregning af reguleringsgevinster

Huslejereguleringen indebærer en økonomisk gevinst for lejerne, da den sænker huslejen i forhold til en husleje, der er sat på markedsvilkår. Denne gevinst betegnes som *reguleringsgevinsten*. I dette afsnit beskrives, hvordan reguleringsgevinsten beregnes. Der er primært fokus på, hvordan den markedsbestemte husleje estimeres.

Reguleringsgevinsten (RG) beregnes som forskellen mellem den estimerede markedsbestemte husleje (hl^{mb}) og den faktiske husleje (hl):

$$RG = hl^{mb} - hl$$

Den faktiske husleje observeres eller imputeres, jf. afsnit 2.1. Den markedsbestemte husleje estimeres på baggrund af huslejen i nyudlejede boliger opført efter 1991. Lejen indenfor de første to år kan nemlig opfattes som den markedsbestemte husleje, jf. afsnit 2. Denne metode til at estimere den markedsbestemte husleje er tidligere blevet brugt af Bloze og Skak (2013).

[2] Prædiktionskvaliteten er også opgjort ud fra boliger, hvor huslejen kun fremgår af data fra EjendomDanmark, jf. bilag A. Overordnet set er prædiktionskvaliteten den samme.

[3] Hvis prædiktionskvaliteten kun vurderes ud fra boliger, hvor huslejen fremgår af EjendomDanmark, er det modsatte gældende, jf. bilag A.

3.1 Estimation ved random forest

Den markedsbestemte husleje for regulerede lejeboliger estimeres ud fra en lang række forklaringsfaktorer. Der bruges en såkaldt *random forest*-estimator, og denne metode er valgt, da formålet er at prædiktere den markedsbestemte husleje, og ikke at finde årsagssammenhængen mellem huslejen og forklaringsfaktorerne. Her er random forest-estimatoren (og lignende Machine Learning-estimatorer) velegnet, da den tillader en smule bias, hvilket øger prædiktionskvaliteten.

En random forest-estimator opstiller ikke en *a priori* funktionsform for den model, der prædikterer den markedsbestemte husleje. I stedet opstilles et meget stort antal modeller, som hver især prædikterer den markedsbestemte husleje, med forskellige udpluk af forklaringsfaktorer i hver model. I hver model udvælges de variable, som er vigtigst for at prædiktere den markedsbestemte husleje. Den endelige estimation opgøres herefter som et simpelt gennemsnit over alle de opstillede modellers prædiktioner. Random forest-estimatoren er nærmere gennemgået i bilag B.

Konkret prædikteres den markedsbestemte husleje pr. kvadratmeter ud fra boligernes fysiske og geografiske karakteristika, jf. tabel 2. Disse karakteristika er blandt andet boligens størrelse, antal værelser, geografisk placering defineret ud fra postnumre og miljøfaktorer som støj, luftforurening samt nærhed til natur.[4] Disse karakteristika er valgt ud fra de faktorer, som anses for vigtigst i at bestemme boligers værdi, jf. Freeman (1993) og Sopranzetti (2015).

Tabel 2: Variable i random forest-modellen

Outcome variabel	Husleje pr. kvm
Forklarende variable:	
Kontinuerte variable	Boligens størrelse, antal etager i bygningen, antal badeværelser, antal værelser, antal toiletter, afstand til kyst, større søer og skove, luftforurening, støj, ejendomsskatten og kalenderår.
Kategoriske variable (dummy for hver kategori)	Elevator i opgang, varmeinstallation, opvarmningsmiddel, køkkenforhold og postnummer[a].

[a)] Der er ikke en dummy for hvert postnummer, da der i stedet bruges *sufficient representation*-tilgangen, jf. Johannemann mfl. (2019) og bilag B.

3.2 Boligens opførelsesår

En ulempe ved at estimere den markedsbestemte husleje på baggrund af huslejen på nyudlejede boliger er, at disse lejeboliger er opført efter 1991, mens de regulerede lejeboliger primært er opført frem til 1991. Herved kan boligens opførelsesår ikke medtages i de boligkarakteristika, som bruges til at estimere den markedsbestemte husleje. Ældre boliger kan være af dårligere kvalitet, og derfor kan reguleringsgevinsten blive overvurderet, hvis der ikke tages højde for, hvornår boligen er opført.

I beregningen af den markedsbestemte husleje tages der højde for opførelsesåret for de regulerede boliger ved at beregne effekten af opførelsesår på salgsprisen for ejerboliger.

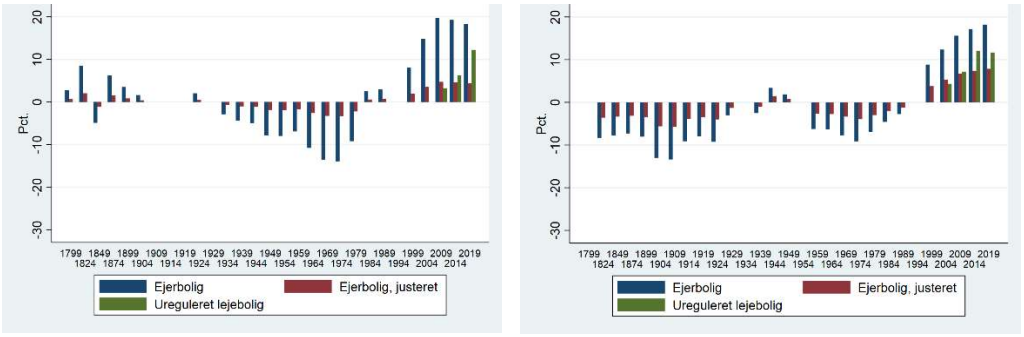
[4] Det kan være problematisk at inkludere områdevariable (her postnumre) som fixed effects i en random forrest-estimator, jf. Johannemann mfl. (2019) og Hastie mfl. (2008). Der tages derfor højde for boligens beliggenhed vha. *sufficient representation*-tilgangen, jf. Johannemann mfl. (2019) og Nationalbanken (2023). Metoden er nærmere beskrevet i bilag B.

Metoden udnytter, at salgsprisen for ejerboliger er markedsbestemt for ejendomme opført i alle år, altså både før og efter 1991. Effekten af opførelsesåret for salgsprisen nedjusteres, da effekten af opførelsesåret på den markedsbestemte husleje siden 1991 har været mindre end effekten på ejerboligpriser. Metoden er beskrevet i dybden i bilag C.

Først opstilles en OLS-model, der finder sammenhængen mellem salgsprisen på ejerboliger og boligens opførelsesår, når der kontrolleres for en række af boligernes fysiske og geografiske karakteristika. Modellen estimeres separat for lejligheder, rækkehus og parcelhuse, jf. figur 3. Nybyggede ejerlejligheder har som forventet de højeste salgspriser, men lejligheder bygget omkring år 1900 er dyrere sammenlignet med lejligheder bygget fra 1950-80 (blå søjler).[5] Det afspejler nok, at ældre lejligheder har en herlighedsværdi, der gør dem mere attraktive. Figuren viser også, at effekten af opførelsesår på huslejen i uregulerede lejeboliger (nyudlejede boliger opført efter 1991) er mindre i årerne efter 1991 (grønne søjler). Effekten på ejerboliger justeres derfor for ikke at overvurdere effekten på huslejen i årerne frem til 1991. Den justerede effekt fremgår af de røde søjler, og den estimerede markedsbestemte husleje for ældre lejeboliger justeres med denne faktor.

Figur 3: Ejerboligpriser og bygningens opførelsesår

Lejligheder *Rækkehuse*



Anm.: Figuren viser opførelsesårs effekt på salgspriserne for ejerboliger (blå søjler) og huslejen for uregulerede lejeboliger opført efter 1991 (grønne søjler). De røde søjler viser den justerede effekt, hvor der er taget højde for, at effekt på huslejen er mindre end effekten på ejerboligpriserne. X-aksen viser opførelsesåret i femårsintervaller (25-årsintervaller før år 1900). Effekten måles ift. boliger opført i de fem år af 1990'erne.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Niveauet af reguleringsgevinsterne afhænger af, hvordan der kontrolleres for boligens opførelsesår, men som det fremgår i afsnit 4.4, er analysens hovedkonklusioner robust for at kontrollere for boligens opførelsesår på forskellige måder.

3.3 Prædiktionskvalitet af estimation

De estimerede sammenhænge mellem huslejen og de forklarende faktorer ved random forest-metoden er svære at gennemskue. Eksempelvis viser metoden ikke marginale effekt af de forskellige forklarende variable ligesom en OLS-model. Det kan gøre det

[5] Bemærk, at der er taget højde for boligernes beliggenhed og andre karakteristika ved boligen.

sværere at vurdere, om modellen er velspecificeret, og hvor god modellen er til at prædiktere den markedsbestemte husleje.

Man kan dog teste, hvor præcis modellen er til at prædiktere den markedsbestemte husleje for uregulerede lejeboliger. Det gøres ved at sammenligne den observerede husleje pr. kvm for nyudlejede boliger opført efter 1991 med den markedsbestemte husleje pr. kvm, der er prædikteret med random forest-metoden.

Prædiktionskvaliteten måles ved at sammenligne den estimerede husleje med den faktiske husleje i et sample af 30 pct. af nyudlejede boliger opført efter 1991, som ikke indgik i samplet til at estimere modellerne.[6] Dvs. at der laves en out-of-sample prædiktion. Den foretrukne specifikation af random forest-modellen er bl.a. valgt ud fra, hvornår den bedste prædiktionskvalitet opnås vurderet ud fra R^2 samt andelen indenfor 10 og 20 pct. Alle mål ville være 100 pct., hvis modellen helt nøjagtig kunne prædiktere den markedsbestemte husleje.

Prædiktionskvaliteten er høj, jf. tabel 3. I den foretrukne specifikation (baseline-modellen) er R^2 ca. 89 pct. og i 83 (94) pct. af tilfældene er den prædikterede markedsbestemte husleje indenfor 10 (20) pct. af den faktiske husleje. Det er også tydeligt at se, at de geografiske karakteristika er vigtige. Således falder R^2 til 63 pct., hvis der kun bruges boligkarakteristika som forklarende variable. Prædiktionskvaliteten ændres kun lidt ved at kontrollere for boligens beliggenhed defineret ud fra kommuner eller skoledistrikter, men random forest-metoden er markant bedre end en OLS-model, jf. det nederste panel i tabel 3.

Tabel 3: Prædiktionskvalitet af random forest-modellen

	R^2	Andel indenfor 10 pct.	Andel indenfor 20 pct.
	Baseline model		
Baseline	**0,89**	**0,83**	**0,94**
Kun boligkarakteristika	0,63	0,48	0,76
	Alternative metoder		
Kommuner	0,88	0,83	0,94
Skoledistrikter	0,88	0,83	0,94
OLS	0,72	0,59	0,85

Anm.: Første panel viser prædiktionskvaliteten for baseline modellen, jf. afsnit 3.1 og 3.2, og baseline modellen, hvor kun variable om boligens karakteristika og ikke områdets karakteristika. Andet panel viser kvaliteten ved ændringen i baseline metoden. De to første rækker tager højde for boligernes beliggenhed ud fra kommuner hhv. skoledistrikter i stedet for postnumre, mens den sidste række estimerer modellen vha. en OLS-model i stedet for en random forest-model. Tabellen bygger på 30 pct. af nyudlejede boliger opført efter 1991, som ikke indgik i estimationen, altså en out-of-sample prædiktion. $R^2 = \frac{MSE}{var(husleje)}$, hvor MSE er summen af kvadrerede fejlled. Andel indenfor 10 (20) pct. viser, hvor stor en andel af observationerne som er indenfor 10 (20) pct. af den faktiske husleje. OLS-modellen medtager de samme inputvariable som random forest modelle.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

[6] Prædiktionskvaliteten kan kun testes på nyudlejede boliger opført efter 1991 og ikke de regulerede lejeboliger, da den markedsbestemte husleje ikke kendes for disse boliger.

Til forskel fra en OLS-model estimerer random forest-modellen ikke marginal effekter af de forskellige forklarende variable. Det kan gøre det svært at vurdere, om modellen er velspecificeret, og dermed om den estimerede sammenhæng mellem huslejen og de forklarende variable er plausibel. Random forest-metoden tillader dog forskellige teknikker, der gør det muligt at vurdere, om modellen er velspecificeret, jf. figur 4.

Panel a) viser det såkaldte *variable importance plot*. Vigtigheden af en variabel beregnes ud fra den procentvise reduktion i summen af de kvadrerede fejlled (MSE) grundet denne givne variabel, jf. Hastie mfl. (2008).[7] På den måde får man et indtryk af, hvilke variable, der er vigtigst i at prædiktere den markedsbestemte husleje pr. kvm. De to vigtigste variable er beboelsesarealet og kalenderåret, men diverse miljøfaktorer som støj- og luftforurening og antallet af (bade)værelser er også vigtige i at forklare den markedsbestemte husleje pr. kvm.[8]

Man kan ikke bruge variable importance beregningen til at sige, hvordan de forskellige variable er korreleret med den markedsbestemte husleje pr. kvm. Ud fra det såkaldte *partial dependence plot* kan man dog vise sammenhængen mellem huslejen pr. kvm og de forskellige variable, jf. Hastie mfl. (2008). Denne beregning prædikterer en sammenhæng mellem en given variabel, X_S, og den gennemsnitlige husleje under to forudsætninger. For det første tildeles alle boliger samme den værdi af X_S, f.eks. tildeles alle boliger en størrelse på 100 kvm. For det andet forbliver boligernes øvrige karakteristika uændret. På den måde kan den gennemsnitlige husleje beregnes ved forskellige værdier af X_S. Det giver en ide, om hvordan huslejen ændres ved forskellige værdier af X_S, men der er dog ikke tale om en kausal sammenhæng.

Figur 4.b viser partial dependence plottet for beboelsesåret og kalenderåret, der er de to faktorer, som er vigtigst i at prædiktere den markedsbestemte husleje pr. kvm. Den prædikterede markedsbestemte husleje pr. kvm er højest i små boliger og lavest i store boliger. Det er i overensstemmelse med forventningerne, da man typisk vil betale mere for én ekstra kvadratmeter i små boliger ift. store boliger. Dertil er den prædikterede husleje højest i de seneste år, hvilket harmonerer med den faktiske udvikling i den markedsbestemte husleje pr. kvm, jf. De Økonomiske Råds formandskab (2023).

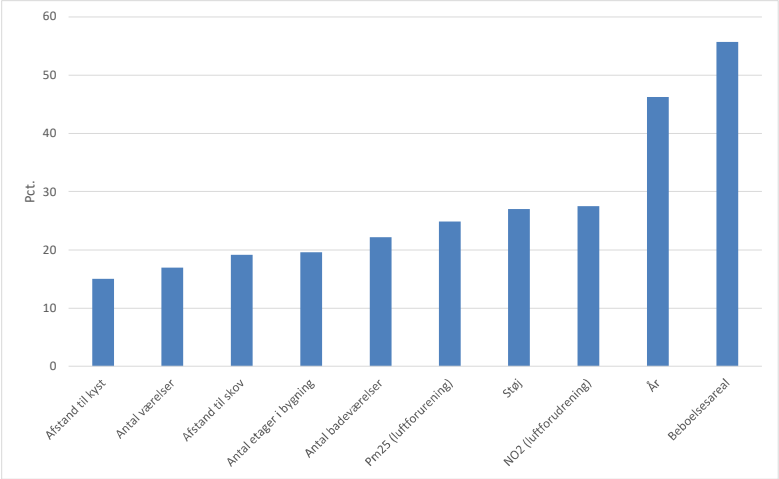
Alt i alt viser resultaterne i dette delafsnit, at random forest-modellen er god til at prædiktere den markedsbestemte husleje og synes at være velspecificeret.

[7] Hastie mfl. (2008) giver en grundigere beskrivelse af, hvordan variable importance beregnes.

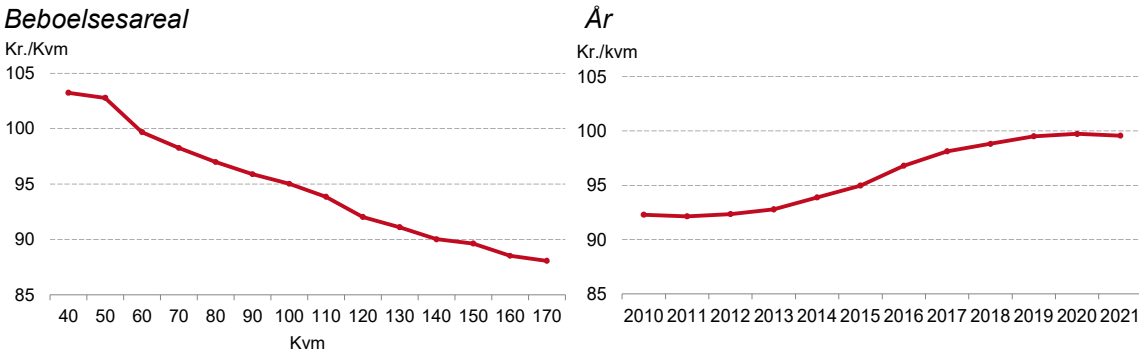
[8] Måden, hvorpå der er kontrolleret for boligens beliggenhed, gør, at det er svært vise vigtigheden af beliggenheden på denne måde.

Figur 4: **Specifikationstest af random forest-estimationen**

a) Variable importance plot



b) Partial dependence plot



Anm.: Panel a) viser et variable importance plot over de ti vigtigste variable i estimationen af den markedsbestemte husleje pr. kvm. Figuren viser den procentvise reduktion i summen af de kvadrerede fejlled (MSE) grundet hver variabel. Måden, hvorpå der er kontrolleret for boligens beliggenhed, gør, at det er svært vise vigtigheden af beliggenheden på denne måde. I panel b) præsenteres to partial dependence plot. Figuren viser den gennemsnitlige estimerede markedsbestemte husleje pr. kvm ved forskellige værdier af hhv. beboelsesstørrelsen og kalenderåret.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

3.4 Tolkning af reguleringsgevinster

Den beregnede reguleringsgevinst er ikke den samme som forskellen mellem den faktiske leje og den markedsbestemte leje i det tilfælde, hvor huslejen for alle lejeboliger blev fastsat på markedsvilkår. I stedet skal gevinsten tolkes som den forventede huslejestigning, hvis én enkelt lejebolig overgik til fri lejefastsættelse. Huslejen i nyudlejede boliger opført efter 1991 kan nemlig adskille sig fra den markedsbestemte husleje i fravær af nogen form for huslejeregulering af to årsager, jf. afsnit IV.4 i De Økonomiske Råds formandskab (2023). For det første kan misallokering medføre huslejen i nyere lejeboliger. Misallokering indebærer, at husstande, der efterspørger nyere lejeboliger, har en højere betalingsvilje. Det får huslejen til at blive højere end den markedsbestemte husleje. For det andet kan lejer være villig til at acceptere en forøget startleje, hvis den forventede stigningstakst for huslejen reduceres på grund af huslejereguleringen.

3.5 Estimation af markedsbestemte husleje ud fra salgspriser på ejerboliger

I stedet for at estimere den markedsbestemte husleje ud fra huslejen i nyudlejede boliger opført efter 1991 er en alternativ metode at bruge salgspriserne på ejerboliger. Ved denne fremgangsmåde estimeres først værdien af en lejebolig på baggrund af værdien af sammenlignelige ejerboliger. Værdien af lejeboligen konverteres herefter til en 'markedsbestemt' husleje ved at antage, at værdien af lejeboligen skal forrentes med afkastet på den bedste alternative investering. Afkastet størrelse herefter kalibreres ud fra en række faktorer, f.eks. den risikofrie rente, risikopræmien, omkostningerne ved at vedligeholde en lejebolig, osv. Denne metode er tidligere blevet brugt til at estimere den markedsbestemte husleje i Det Økonomiske Råds formandskab (2001) og Kristensen (2012).

Nærværende analyse bruger ikke denne metode, da det indebærer større usikkerhed i estimationen af den markedsbestemte husleje, end når den estimeres på baggrund af huslejen i nyudlejede boliger opført efter 1991. Der synes nemlig både at være en vis usikkerhed ved at beregne det korrekte afkast på den bedste alternative investering. Derudover er ejerboliger af meget forskellig karakter, hvilket gør det svært at estimere boligværdien med stor præcision.

4 Resultater

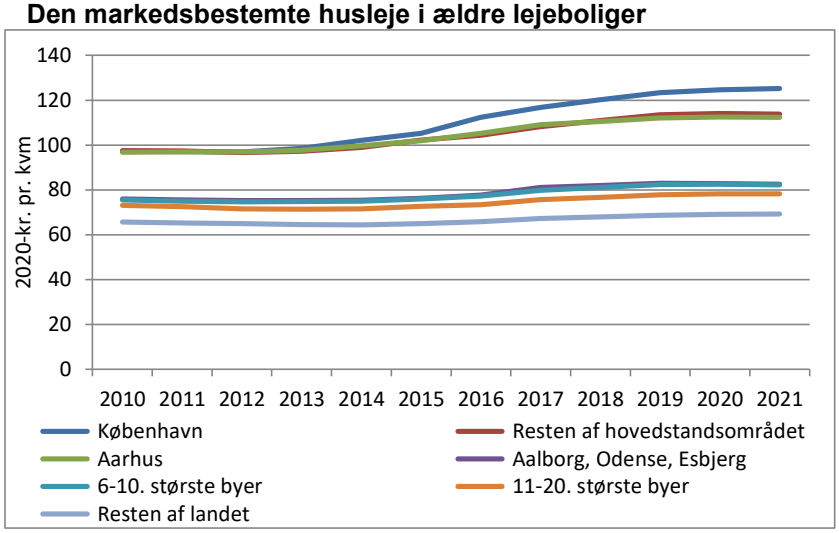
I dette afsnit vises først udviklingen i den estimerede markedsbestemte husleje i ældre lejeboliger, da det er et grundelement i beregningen af reguleringsgevinsten, jf. afsnit 3. Herefter vises forskelle i reguleringsgevinsterne på tværs af indkomstgrupperne og Theil-indekset beskrives grundigt, da dekomponeringsanalysen, der beregnes ud fra Theil-indekset, er centralt i for hovedkonklusionen. Slutteligt understøtter en række følsomhedsanalyser, at hovedkonklusionen holder, når den foretrukne estimationsmetode (baseline-metoden) ændres på forskellige måder.

4.1 Den markedsbestemte husleje

Den estimerede markedsbestemte husleje for ældre lejeboliger er højest i København, efterfulgt af det øvrige hovedstadsområde og Aarhus, jf. figur 5. I 2021 er huslejen pr. kvadratmeter 10 pct. højere i København end i Aarhus og 40-50 pct. højere end de større provinsbyer. Siden 2010 er den markedsbestemte husleje steget med knap 30 pct. i København, 16 pct. i Aarhus samt resten af hovedstadsområdet og 5-10 pct. i resten af landet. Den estimerede stigning i den markedsbestemte husleje for ældre lejeboliger svarer derfor i høj grad til den faktiske stigning i huslejen for nyudlejede lejeboliger opført efter 1991, jf. De Økonomiske Råds formandskab (2023).[9]

[9] Det samme gør sig gældende for den estimerede markedsbestemte husleje for almene boliger, jf. bilag D.

Figur 5: Den markedsbestemte husleje i ældre lejeboliger



Anm.: Figuren viser den gennemsnitlige markedsbestemte husleje pr. kvm. for ældre lejeboliger for hvert år og i hver byområde. Tallene for 2018 er baseret på en lineær interpolation mellem 2017 og 2019, jf. afsnit 2. København omfatter både Københavns Kommune og Frederiksberg Kommune. De resterende kommuner grupperes efter bystørrelse, således at "6.-10. største byer" eksempelvis betegner de kommuner, der rummer de 6.-10. største byer. De 6.-10. største kommuner er: Randers, Horsens, Kolding, Vejle og Roskilde. De 11.-20. største kommuner er: Herning, Silkeborg, Hørsholm, Helsingør, Næstved, Viborg, Fredericia, Køge, Holstebro, Taastrup.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

4.2 Forskelle på tværs af indkomstgrupper

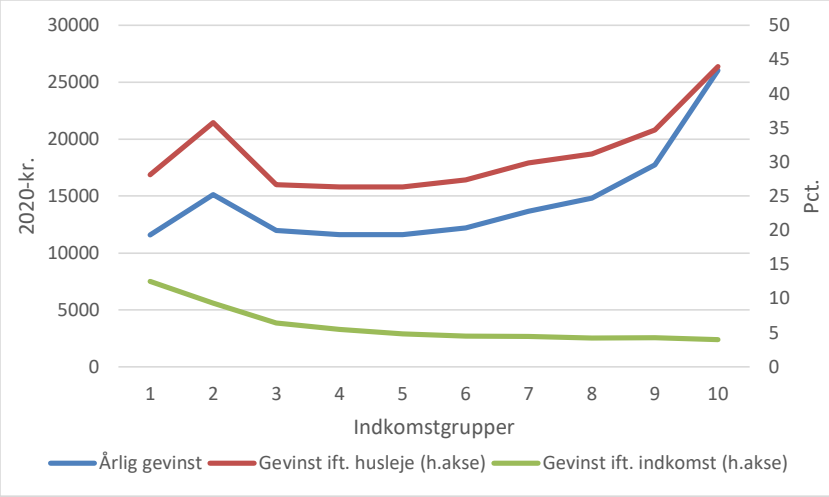
Husstandene i ældre lejeboliger inddeles i ti grupper ud fra deres placering i fordelingen af den ækvivalerede disponible indkomst for alle husstande i Danmark. Den gennemsnitlige ækvivalerede reguleringsgevinst beregnes indenfor hver indkomstgruppe. På den måde undersøges det, om gevinsten ved lavere husleje i højere grad tilfalder husstande med lav indkomst. Gevinsten er ækvivaleret, for at den bliver den sammenlignelig på tværs af husstande med forskellige størrelser, og der tages samtidig højde for stordriftsfordele ved at bo flere personer i samme husstand.

Den gennemsnitlige absolutte reguleringsgevinst er størst i den højeste indkomstgruppe for husstande i ældre lejeboliger, jf. figur 6. Reguleringsgevinsten er ca. 12.000 kr. i den laveste indkomstgruppe, hvilket er 35 pct. lavere end gevinsten i næsthøjeste indkomstgruppe og ca. halvt så meget som reguleringsgevinsten i den højeste indkomstgruppe. Det afspejler, at de attraktive ældre lejeboliger med lavere husleje ofte tilfalder husstande med høj indkomst, blandt andet fordi disse husstande i højere grad bor i Københavnsområdet. Samme billede gælder også, når gevinsterne måles ift. til huslejen, hvor gevinsterne ift. huslejen er 10-15 pct.point højere for højindkomsthusstande sammenlignet med lavindkomsthusstande. Det Økonomiske Råds formandskab (2001) og Kristensen (2012) fandt ligeledes, at de højeste indkomstgrupper har de største reguleringsgevinster.

Reguleringsgevinsten relativ til husstandenes disponible indkomst er dog størst for husstande med lave indkomster. I gennemsnit er gevinsten i ældre lejeboliger omkring 12 pct. af indkomsten i den laveste indkomstgruppe og ca. 4 pct. af indkomsten i den højeste

gruppe. Det betyder, at reguleringsgevinsten udgør en mindre del af husstandenes indkomst, jo højere indkomsten er.

Figur 6: Reguleringsgevinst fordelt på indkomstgrupper, ældre lejeboliger i 2021



Anm.: Figurerne viser den gennemsnitlige reguleringsgevinst opgjort på lejeboligtype og indkomstgruppe. Reguleringsgevinsten, huslejen og den disponible indkomst er ækvivaleret. Indkomstgrupperne er dannet ud fra indkomstdecilerne beregnet ud fra alle husstande i Danmark uanset boligform, men kun husstande, som bor i lejeboliger, er inkluderet i figuren. Observationer med årlig husleje under 1.000 kr. og disponibel ækvivaleret indkomst under 10.000 kr. er fjernet.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

4.3 Forskel i gevinster på tværs og indenfor indkomstgrupper ved Theil-indeks

Huslejereguleringen gælder for alle private ældre lejeboliger, men der kan være store forskelle i reguleringsgevinsterne for husstande i disse boliger. Reguleringen af disse boliger indebærer, at der er for mange, som ønsker at bo i de billigere regulerede lejeboliger i forhold til udbuddet af boliger. Det betyder, at de ældre lejeboliger fordeles ud fra eksempelvis ventelister, netværk eller udlejernes præferencer. Det tilsiger, at det til en vis grad kan være tilfældigheder, som afgør, om en husstand får tilbudt en billigere lejebolig. Der kan derfor være store forskelle i gevinsterne for husstande med samme indkomst. Det undersøges i dette afsnit ved brug af det såkaldte Theil-indeks.

De samlede forskellene i reguleringsgevinsterne kan opgøres ved det såkaldte Theil-indeks, jf. boks 3. Theil-indekset er, ligesom Gini-koefficienten, et summarisk mål for uligheden i en fordeling. Indekset måler, hvor stor afvigelsen er mellem den faktiske fordeling af reguleringsgevinsten og den fordeling af reguleringsgevinsten, der ville være, hvis reguleringsgevinsterne var ligeligt fordelt. Hvis alle har samme reguleringsgevinst, vil Theil-indekset være 0, og en større værdi betyder større forskelle i reguleringsgevinsterne. Idet formålet med huslejeregulering er at omfordele til gavn for lavindkomsthusstandene, skal indekset altså gerne være større end 0.

I modsætning til Gini-koefficienten kan Theil-indekset også dekomponere disse forskelle for at se, hvor stor en del af forskellene, som skyldes forskelle på tværs henholdsvis

indenfor indkomstgrupperne.[10] På den måde kan det belyses, hvor store forskelle i gevinsterne der er for husstande med samme indkomst.

BOKS 3 THEIL-INDEKSET

Theil-indekset (T) er et summarisk mål for ulighed eller forskelle i en fordeling og beregnes på følgende måde:

$$T = \sum_{i=1}^{N} s_i ln\left(\frac{s_i}{p_i}\right)$$

hvor s_i er andelen af husstand i's andel af de samlede reguleringsgevinster[a)], p_i er husstandens andel af alle husstande (svarende til $1/N$), og N er antallet af alle husstande.

Theil-indekset måler, hvor stor afvigelsen er mellem den faktiske fordeling og den helt lige fordeling. Det kan ses i formlen ved, at s_i/p_i udtrykker, om husstand i har en større andel af reguleringsgevinsten (s_i) end husstandens andel af alle husstande (p_i). Uligheden stiger jo større Theil-indeks. Ved en helt lige fordeling er Theil-indekset lig 0, hvilket kan ses i formlen ved at sætte $s_i = p_i$ for alle husstande. I et andet ekstreme tilfælde, hvor kun én husstand har en reguleringsgevinst, er Theil-indekset lig $\ln(N)$. I praksis betyder, at Theil-indekset øvre grænse i dette tilfælde er ca. 13.

Theil-indekset kan, i modsætning til Gini-koefficienten, dekomponeres, så det er muligt at belyse, hvor stor en del af den samlede ulighed, der skyldes forskelle henholdsvis på tværs og indenfor delgrupper, jf. Theil (1967):

$$T = \underbrace{\sum_{k=1}^{K} s_k ln\left(\frac{s_k}{p_k}\right)}_{Ulighed\ på\ tværs} + \underbrace{\sum_{k=1}^{K} s_k T_k}_{Ulighed\ indenfor}$$

hvor s_k er indkomstgruppe k's andel af reguleringsgevinsterne, p_k er andelen af husstande i indkomstgruppe k's ift. alle husstande, og T_k er Theil-indekset for indkomstgruppe k. Der anvendes ti indkomstgrupper. Det første led viser, hvor stor en del af uligheden i reguleringsgevinsterne, der skyldes forskelle på tværs af indkomstgrupperne. Ligesom ved den overordnede formel for Theil-indekset opfanger dette led, om husstande i hver indkomstgruppe får en større andel af de samlede reguleringsgevinster (s_k), end deres andel 'berettiger' til i en lige fordeling (p_k). Det andet led viser den del af uligheden, som skyldes forskelle indenfor hver indkomstgruppe. Det gøres ved at beregne Theil-indekset indenfor hver gruppe og vægte det med, hvor stor en del af reguleringsgevinsterne gruppen får (s_k).

Theil-indekset kan kun medtage husstande med positive reguleringsgevinster. Reguleringsgevinsten er negativ for ca. 20 pct. af alle husstande i ældre lejeboliger, og det betyder, at Thiel-indekset vil undervurdere den faktiske ulighed. Dette problem mindskes ved at tildele husstande med negative reguleringsgevinster den laveste værdi blandt husstande med positive reguleringsgevinster.

a) Hvis Theil-indekset beregnes på baggrund af reguleringsgevinsten ift. indkomsten eller lignende vil det være andelen af den samlede sum af dette mål, som udtrykkes i s_i.

Der er betydelig forskelle i reguleringsgevinster blandt husstande i ældre lejeboliger, jf. tabel 4. Det samlede Theil-indeks er ca. dobbelt så stort for ældre private lejeboliger ift. almene boliger.[11] Det gælder uanset, om reguleringsgevinsterne opgøres i absolutte værdier eller relativt til indkomsten.

[10] Tidligere studier (f.eks. De Økonomiske Råds formandskab (2016) og De Økonomiske Råds formandskab (2019)) har også brugt Theil-indekset til at dekomponere ulighed i hhv. indkomst og miljøgoder på tværs og indenfor grupper.

[11] Theil-indekset for almene boliger er ca. 0,22, når det opgøres ud fra den årlige reguleringsgevinst, jf. bilag D.

I ældre lejeboliger skyldes 90-97 pct. af forskellene i reguleringsgevinsterne forskelle indenfor indkomstgrupperne. Forskellen i reguleringsgevinsterne kan altså næsten udelukkende tilskrives forskelle blandt husstande med samme indkomst, og det gælder også selvom husstandene opdeles i 100 indkomstgrupper i stedet for ti, jf. bilag D. I almene boliger afspejler hovedparten af forskellene i reguleringsgevinsterne også forskelle indenfor indkomstgrupperne, jf. bilag D.

Forskelle på tværs af indkomstgrupperne bidrager mere til de samlede forskelle, når gevinsten måles ift. indkomsten, hvor 11-18 pct. skyldes forskelle på tværs af indkomstgrupperne. Det skyldes, at husstande i de laveste indkomstgrupper har en reguleringsgevinst ift. indkomsten, der er større end for højindkomsthusstande, jf. figur 6. Størstedelen af de samlede forskelle i gevinsterne afspejler dog stadig forskelle indenfor indkomstgrupperne, også når gevinsten måles ift. indkomsten.

Ca. 73 pct. af de samlede forskelle i boligstøtten skyldes forskelle indenfor indkomstgrupperne. Det er knap 25 pct.point mindre end ved reguleringsgevinsterne. Størstedelen af forskellene i boligstøtten skyldes dog forskelle indenfor indkomstgrupperne, og det afspejler, at boligstøtten afhænger af flere andre faktorer end blot husstandsindkomsten. Boligstøtten bestemmes nemlig også af eksempelvis niveauet af huslejen, boligens størrelse, antal børn, og om der er folkepensionister i husstanden.

Tabel 4: Theil-indeks for reguleringsgevinsterne og boligstøtte, ældre lejeboliger

	Årlig regule-ringsgevinst	Gevinst ift. husleje	Gevinst ift. indkomst	Årlig boligstøtte
Samlet Theil-indeks	0,65	0,67	0,77	0,76
	---------- Pct. ----------			
Forskel på tværs af indkomstgrupper	2,6	1,2	10,9	28,4
Forskel indenfor indkomstgrupper	97,4	98,2	89,1	71,6

Anm.: Tabellen viser, hvor stor en del af de samlede forskelle i reguleringsgevinsterne (og boligstøtten), som skyldes forskelle på tværs af og indenfor indkomstgrupperne. Gevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset. Reguleringsgevinsten er ækvivaleret. Indkomstdecilerne er beregnet ud fra alle husstande i Danmark uanset boligform.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

4.4 Følsomhedsberegninger

Dette afsnit præsenterer resultaterne af en række følsomhedsanalyser. Overordnet set er konklusionerne for alle følsomhedsanalyser i tråd med hovedkonklusionen i De Økonomiske Råds formandskab (2023). Fokus i følsomhedsanalyserne er at undersøge forskelle i reguleringsgevinster på tværs af og indenfor indkomstgrupperne. Resultaterne vises kun for ældre lejeboliger.

Først vises, at der er forskelle i reguleringsgevinsterne indenfor alle indkomstgrupper. Dernæst undersøges fordelingen af gevinsterne, hvis den foretrukne estimationsmetode

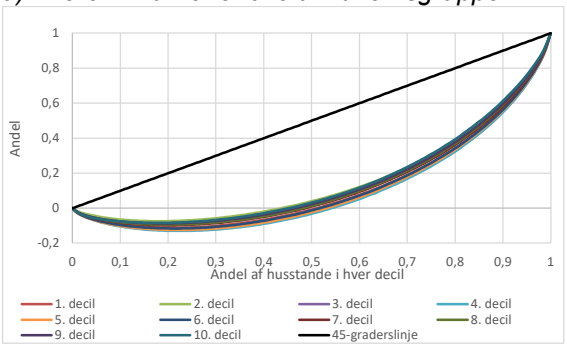
og udsnit af lejeboliger modificeres i forskellige aspekter. Til sidst vises effekten på fordelingen af gevinsterne ved at kontrollere for aldersforskelle i husstandene på tværs og indenfor indkomstgrupperne.

Forskelle i reguleringsgevinster indenfor indkomstgrupper

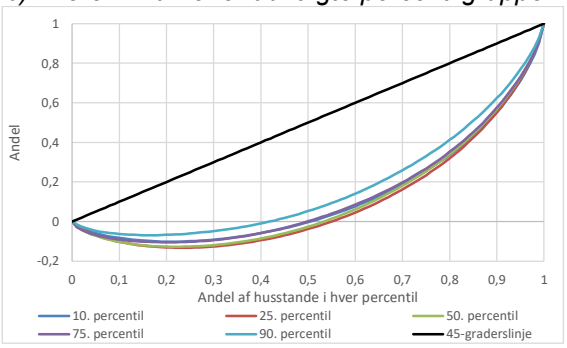
Analysen i De Økonomiske Råds formandskab (2023) viste på baggrund af Lorenz-kurver store forskelle i reguleringsgevinsterne indenfor anden og niende indkomstgruppe. Denne store forskel findes også i de andre otte grupper, jf. figur 7a. Lorenz-kurverne hænger lavt for alle indkomstgrupper, og Gini-koefficienten er mellem 0,7-0,8 i alle grupper. Det samme gælder, hvis indkomstgrupperne inddeles ud fra percentiler i stedet for deciler (100 grupper i stedet for ti), jf. figur 7b. I dette tilfælde er Gini-koefficienten mellem 0,6 og 0,8. Hermed er der en forventet forskel i gevinsterne for to husstande med samme indkomst på 120-160 pct., jf. Atkinson (1975).

Figur 7: Forskelle i gevinster indenfor indkomstgrupper, ældre lejeboliger

a) Lorenz-kurve for alle ti indkomsgrupper



b) Lorenz-kurve for udvalgte percentilgrupper



Anm.: Figurerne viser Lorenz-kurver for fordelingen af ækvivalerede reguleringsgevinster indenfor hver indkomstgruppe (decil eller percentil). Lorenz-kurven opgøres ved at opstille alle husstande i lejeboliger efter størrelsen på reguleringsgevinsten. Derefter beregnes den andel af den samlede sum af reguleringsgevinsterne, som en husstand og alle husstande med lavere reguleringsgevinster tilsammen har. Et punkt (x,y) på Lorenz-kurven viser, hvor stor en andel (y pct.) af de samlede reguleringsgevinster, der haves af husstandene med de x pct. laveste reguleringsgevinster. Visse husstande har en negativ reguleringsgevinst, hvilket ses ved, at Lorenz-kurven i starten er under nul. De negative gevinster skyldes usikkerhed i estimaterne og tolkes som ingen gevinst, da beboerne i givet fald vil foretrække en bolig til markedsleje. Indkomstgrupperne er dannet ud fra indkomstdecilerne eller perceltiler beregnet ud fra alle husstande i Danmark uanset boligform.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Modifikationer i estimationsmetode og udsnit af lejeboliger

Dette afsnit viser, at konklusionerne i De Økonomiske Råds formandskab (2023) er robuste over for at modificere den foretrukne estimationsmetode og udsnit af lejeboliger, *baseline modellen*, i forskellige aspekter. Det gøres ved at vise forskelle i reguleringsgevinsterne på tværs af de ti indkomstgrupper samt dekomponere de samlede forskelle i gevinsterne vha. Theil-indekset. På den måde vil det fremgå, om ændringer i baseline modellen ændrer på fordelingen af reguleringsgevinster på tværs og indenfor indkomstgrupperne.

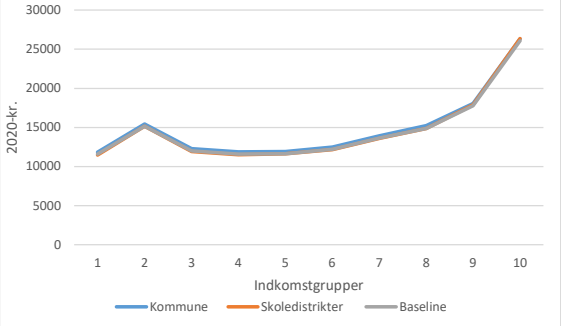
I baseline modellen tages der højde for boligernes beliggenhed i form af postnumre. Som det fremgår af tabel 3, er prædiktionskvaliteten dog ca. lige så god, hvis beliggenheden

defineres ud fra kommunen eller skoledistriktet. Det ændrer dog ikke på fordelingen af reguleringsgevinster, hvis kommuner eller skoledistrikter bruges i stedet for postnumre (baseline), jf. figur 8a. Både den gennemsnitlige gevinst i hver indkomstgruppe og Theil-dekomponeringen er uændret, uanset om beliggenheden defineres ud fra postnumre, kommuner eller skoledistrikter.

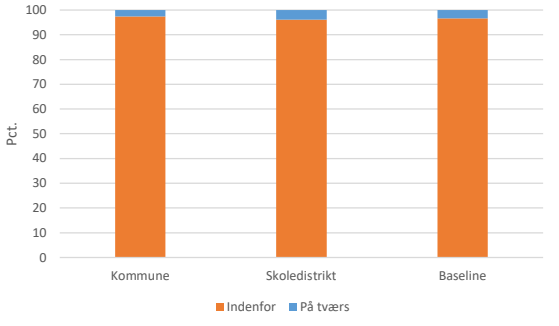
Figur 8: Forskelle i gevinster på tværs og indenfor indkomstgrupper ved modifikation af estimationsmetode, ældre lejeboliger

a) Kommune og skoledistrikt som områdedefinition:

Forskelle på tværs af indkomstgrupper

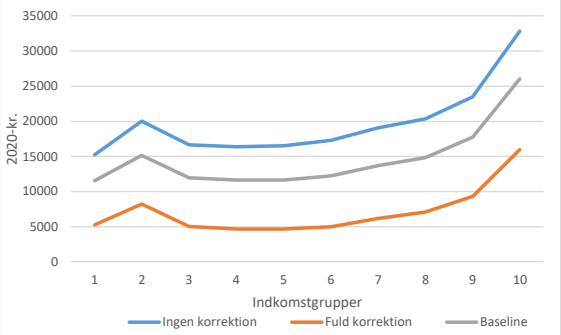


Dekomponering af forskelle i gevinsterne

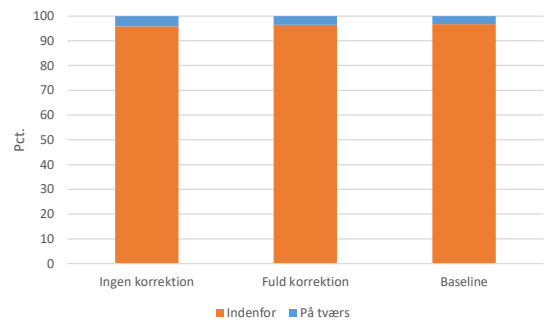


b) Korrektion af boligens opførelsesår

Forskelle på tværs af indkomstgrupper



Dekomponering af forskelle i gevinsterne



Anm.: De venstre figurer viser den gennemsnitlige reguleringsgevinst i hver indkomstgruppe. De højre figurer viser dekomponeringen af de samlede forskelle i gevinsterne vha. Theil-indekset. Søjlerne summer pr. definition til 100 pct. og viser, hvor stor en andel af forskellene, som skyldes hhv. forskelle på tværs af og indenfor indkomstgrupperne. I panel a) defineres beliggenheden ud fra kommunen eller skoledistriktet i stedet for postnumre (baseline). I panel b) tages der højde for opførelsesårets effekt på huslejen på forskellige måder. Reguleringsgevinsten og den disponible indkomst er ækvivaleret. Indkomstgrupperne er dannet ud fra indkomstdecilerne beregnet ud fra alle husstande i Danmark uanset boligform, men kun husstande, som bor i lejeboliger, er inkluderet i figuren. I de højre figurer er reguleringsgevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Boligernes opførelsesår kan ikke direkte indgå i estimationsmodellen af den markedsbestemte husleje, hvorfor der tages højde for dette ud fra effekten på salgsprisen på ejerboliger, jf. afsnit 3.2. Fordelingen af reguleringsgevinsterne ændres dog ikke det store, hvis man tager højde for opførelsesårs betydning på huslejen på en anden måde, jf. figur 8b. Gevinsten er højest for husstande i de højeste indkomstgrupper, og over 95 pct. af alle forskelle i gevinster skyldes forskelle indenfor indkomstgrupperne. Det gælder både,

hvis der ikke tages højde for opførelsesåret, eller hvis der er 'fuld' korrektion svarende til de blå søjler i figur 3. Niveauet for reguleringsgevinsterne varierer dog en del alt efter, hvordan der tages højde for effekten af opførelsesår. Det mest sandsynlige scenarie er, at den blå kurve i figur 8b er et overkantsskøn, da den ikke tager højde for boligens opførelsesår. Omvendt formodes det, at den orange kurve er et underkantsskøn, da effekten på ejerboligpriserne synes større end på den markedsbestemte husleje, jf. figur 3.

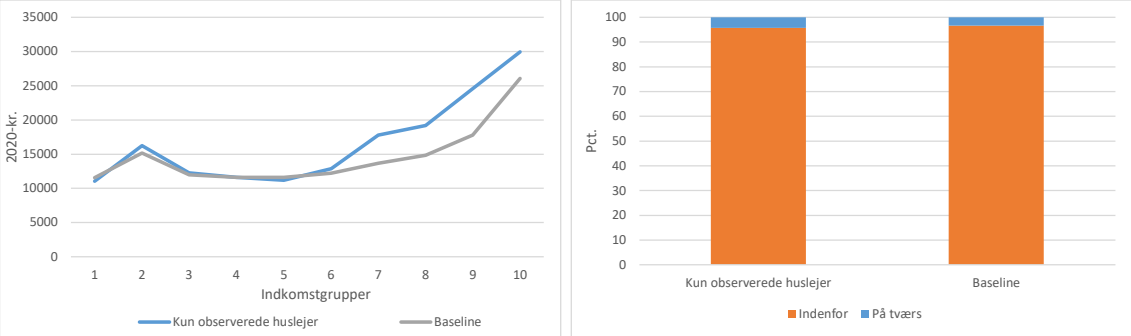
Som nævnt i afsnit 2, imputeres huslejen for ca. 60 pct. af alle private lejeboliger. Det giver en større usikkerhed om den faktiske husleje for disse boliger. Fordelingen af reguleringsgevinster ændres dog ikke det store ved kun at medtage boliger, hvor huslejen direkte observeres gennem boligstøtteregistret eller data fra EjendomDanmark, jf. figur 9a. Gevinsterne bliver endnu større for højindkomsthusstandene, og forskellene indenfor indkomstgrupperne forklarer 96 pct. i stedet for 97 pct. Hovedkonklusionerne afhænger derfor ikke af, at huslejen imputeres for en del lejeboliger.

I baseline modellen medtages alle lejeboliger, uanset om det er lejligheder, rækkehuse eller parcelhuse. Fordelingen af gevinster er ca. uændret, hvis der fokuseres på lejligheder, jf. figur 9b.

Figur 9: Forskelle i gevinster på tværs og indenfor indkomstgrupper ved modifikation i udsnit af lejeboliger, ældre lejeboliger

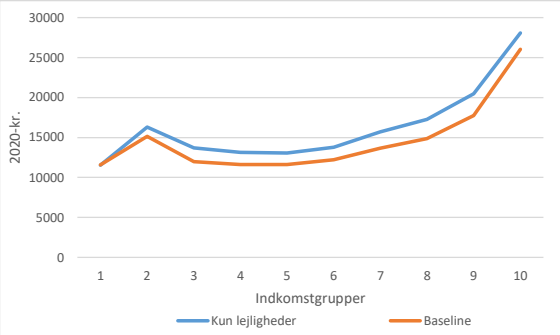
a) Kun boliger med observerede huslejer

Forskelle på tværs af indkomstgrupper *Dekomponering af forskelle i gevinsterne*

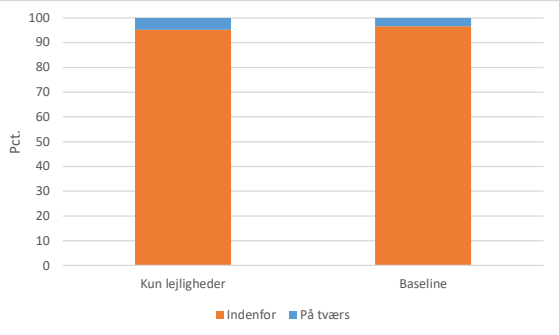


b) Kun lejligheder

Forskelle på tværs af indkomstgrupper



Dekomponering af forskelle i gevinsterne



Anm.: De venstre figurer viser den gennemsnitlige reguleringsgevinst i hver indkomstgruppe. De højre figurer viser dekomponeringen af de samlede forskelle i gevinsterne vha. Theil-indekset. Søjlerne summer pr. definition til 100 pct. og viser, hvor stor en andel af forskellene, som skyldes hhv. forskelle på tværs af og indenfor indkomstgrupperne. I panel a) medtages kun lejeboliger, hvor huslejen observeres, mens panel b) kun medtager lejligheder. I begge tilfælde gælder det både i estimationen af den markedsbestemte husleje og i fordelingen af gevinsterne for ældre lejeboliger. Reguleringsgevinsten og den disponible indkomst er ækvivaleret. Indkomstgrupperne er dannet ud fra indkomstdecilerne beregnet ud fra alle husstande i Danmark uanset boligform, men kun husstande, som bor i lejeboliger, er inkluderet i figuren. I de højre figurer er reguleringsgevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

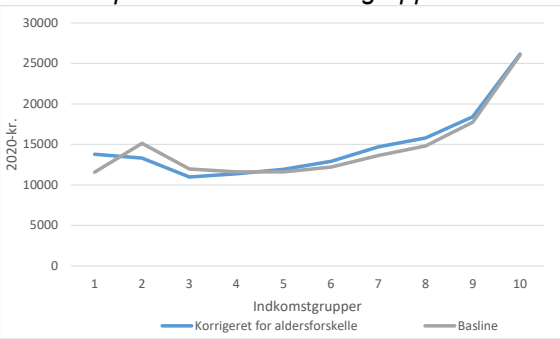
Korrektion for aldersforskelle

En mulig årsag til forskellene i reguleringsgevinsterne på tværs af og indenfor indkomstgrupperne er aldersforskelle mellem husstande. Således blev det vist i De Økonomiske Råds formandskab (2023), at gevinsterne i gennemsnit er større for husstande med ældre personer. Ligeledes er højindkomsthusstandene oftere husstande med ældre personer.

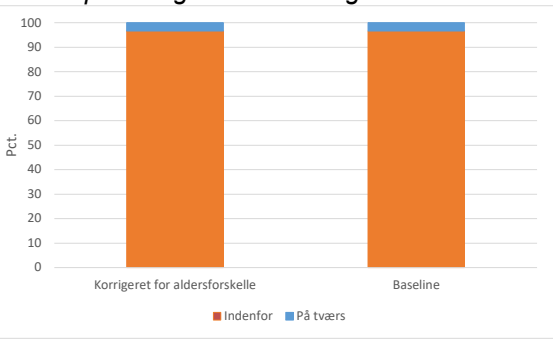
Fordelingen af reguleringsgevinsterne på tværs af og indenfor indkomstgrupperne ændres dog ikke ved at tage højde for aldersforskelle, jf. figur 10. I figuren er der foretaget en standardberegning, der sikrer, at alderssammensætningen er den samme i alle ti indkomstgrupper. Der er dog stadig husstande med høj indkomst, som får de højeste gevinster, ligesom over 96 pct. af indkomstforskellene skyldes forskelle indenfor indkomstgrupperne.

Figur 10: Forskelle i gevinster på tværs og indenfor indkomstgrupper ved korrektion af aldersforskelle, ældre lejeboliger

Forskelle på tværs af indkomstgrupper



Dekomponering af forskelle i gevinsterne



Anm.: De venstre figurer viser den gennemsnitlige reguleringsgevinst i hver indkomstgruppe. De højre figurer viser dekomponeringen af de samlede forskelle i gevinsterne vha. Theil-indekset. Søjlerne summer pr. definition til 100 pct. og viser, hvor stor en andel af forskellene, som skyldes hhv. forskelle på tværs af og indenfor indkomstgrupperne. Korrektionen for alder foretages ved en standardberegning, der sikrer, at den aldersfordelingen er den samme i alle indkomstgrupperne. Reguleringsgevinsten og den disponible indkomst er ækvivaleret. Indkomstgrupperne er dannet ud fra indkomstdecilerne beregnet ud fra alle husstande i Danmark uanset boligform, men kun husstande, som bor i lejeboliger, er inkluderet i figuren. I de højre figurer er reguleringsgevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Litteratur

Athey, S., og G. Imbens (2019): Machine Learning Methods Economists Should Know About. March 2019.

Atkinson, A.B. (1975): The Economics of Inequality. *Oxford University Press.*

Bloze, G, og M. Skak (2013): Rent Control and Misallocation. *Urban Studies* 50(10), s. 1988-2005.

Breiman, L. (2001): Random forests. *Machine Learning* 45(1), s. 5-32.

De Økonomiske Råds formandskab (2019): *Økonomi og Miljø, 2023.*

De Økonomiske Råds formandskab (2023): *Dansk Økonomi, efterår 2023, kapitel IV.*

Det Økonomiske Råds formandskab (2001): *Dansk Økonomi, forår 2001.*

Freeman, A.M. (1993) Property value models, i The measurement of environmental and resource values, Washington, Resources for the future, 367-420.

Hastie, T., Tibshirani, R., og J. Friedman (2008): *The Elements of Statistical Learning – Data Mining, Inference and Prediction*. 2nd edition.

Heckman, J. J. (1979): Sample Selection as a Specification Error. *Econometrica*, 47, s. 153-161.

James, G., Witten, D., Trevor, H., og R. Tibshirani (2021): An introduction to Statistical Learning. *With Applications in R*. Springer.

Johannemann, J., Hadad, V., Athey, S., og St. Wager (2019): Sufficient Representation for Categorical Variables. arXiv:1908.09874v3. Submitted on 26 Aug 2019, last revised 28 Oct 2021.

Kristensen, J.B. Konsekvenser af huslejeregulering på det private udlejningsmarkedet – En mikroøkonomisk analyse af 2000'erne, DREAM arbejdspapir.

Nationalbanken (2023): Housing wealth and consumption during Covid-19. Economic memo

Pyatt, G. (1976): On the Interpretation and Disaggregation of Gini Coefficients. *Economic Journal*, 86 (342), s. 243-255.

Sopranzetti, B.J. (2015): Hedonic Regression Models I Lee, E. C.-F og J.C. Lee (red.): *Handbook of Financial Econometrics and Statistics*. Springer

Theil, H. (1967): Economics and Information Theory. North-Holland Publishing Company

Bilag

A Datakvalitet

Tabel A1: Sammenligning af husleje fra boligstøtteregistret og EjendomDanmark

Gennemsnitlig afvigelse	-0,7
Indenfor 1 pct.	48,1
Indenfor 5 pct.	76,2
Indenfor 10 pct.	86,4
Indenfor 20 pct.	94,1

Anm.: Tabellen viser forskellen mellem den observerede huslejen i boligstøtteregistret og EjendomDanmark for private lejeboliger, hvor huslejen observeres i begge kilder. Huslejen i boligstøtteregistet er brugt som referencegruppe, så den gennemsnitlige afvigelse viser eksempelvis, at huslejen i EjendomDanmark er lidt mindre end i boligstøtteregistret.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Tabel A2: Prædiktionskvalitet af huslejeimputation, 2021 (ED)

Metode	R^2	Andel inden for 10 pct.	Andel inden for 20 pct.
Gns. husleje i opgang	0,63	0,53	0,74
Heckman to-trins model	0,62	0,63	0,85

Anm.: Tabellen viser prædiktionskvaliteten af huslejeimputationen, hvor der kun ses på huslejen oplyst ud fra data fra EjendomDanmark. Dvs. tabellen bygger på 30 pct. af private lejeboliger, hvor huslejen kun kendes gennem data fra EjendomDanmark, og som ikke indgik i imputationsberegningen, altså en out-of-sample prædiktion. $R^2 = \frac{MSE}{var(husleje)}$, hvor MSE er summen af kvadrerede fejlled. Andel indenfor 10 (20) pct. viser, hvor stor en andel af observationerne som er indenfor 10 (20) pct. af den observerede husleje.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

B Random Forest-estimatoren

Helt grundlæggende kan sammenhængen mellem boligkarakteristika X og huslejen Y skrives som:

$$Y = f(X) + u \tag{B1}$$

hvor $f(\cdot)$ er en ikke-specificeret funktion af en række faktorer X, som kan være korrelerede med huslejen, og u er et fejlled.

Til en prædiktion er en *random forest*-estimator (og lignende Machine Learning-estimatorer) velegnet, da den har fokus at estimere den bedste approksimation af $f(\cdot)$, $\hat{f}$, og tillader en smule bias i prædiktionen for at mindske prædiktionsvariansen. Til sammenligning fokuserer en OLS-estimator alene på at finde unbiased estimater for parametrene i $f(X)$, $\hat{\beta}$.

Anvendt på den aktuelle problemstilling, hvor målet er at prædiktere markedsbestemte huslejer for boliger, består fremgangsmåde i *Random Forest* overordnet set af følgende elementer, jf. Hastie mfl. (2008):

Gentag for $i = 1$ til $i = B$:

1. Udtræk en tilfældig stikprøve Z_i fra populationen af boliger.
2. Opdel Z_i i dataområder $R_{ij}, j = 1, \dots, J$ ved hjælp af et beslutningstræ:
    i. Udvælg tilfældigt m ud af de p kovariater til huslejen
    ii. Opdel data i to dele ved at lave det optimale split (kriteriet for "optimal" forkalres nærmere nedenfor) ved at udvælge variablen X_k og splitværdi t_k blandt de m kovariater. Denne opdeling fortsætter (for hver gren af træet) indtil et forud valgt mindste antal observationer pr. dataområde er nået.
3. Pkt. 2 resulterer i B træer, der hver især består af J dataområder R_{ij}. For hvert dataområde dannes en prædikteret husleje for en bolig og den endelige random forest model tager gennemsnittet af alle dataområder for alle B træer.

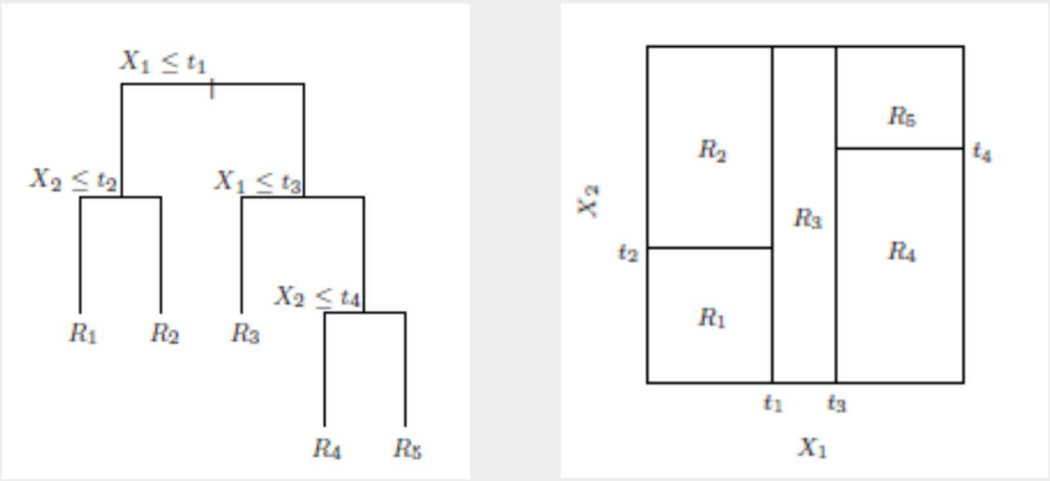
Random Forest-metoden kombinerer dermed to metodiske tilgange: a) kalibrering af prædiktionsmodellen ud fra et antal tilfældigt udtrukne stikprøver, og b) opdeling af hver stikprøve i mindre dataområder via et beslutningstræ. Detaljerne ved denne kombination forklares i de følgende underafsnit.

Som den centrale reference for random forest nævnes oftest Breiman (2001), men metoden er velbeskrevet i nyere publikationer og lærebøger, herunder Hastie mfl. (2008), James mfl. (2021), Mullainathan og Spiess (2017) og Athey og Imbens (2019).

Figur B1 illustrerer beslutningstræet og opdeling af en stikprøve Z i fem dataområder. I første trin vælges X_1 og stikprøven opdeles i observationer med $X_1 \leq t_1$ og $X_1 > t_1$, jf. figur B1.a. Dette svarer i figur B1.b til den vertikale opdeling af stikprøve ved t_1. Denne proces gentages for hver af de to grene af træet (figur B1.a). Delstikprøven med $X_1 \leq t_1$ (venstre gren) underopdeles ved t_2-værdien for X_2, mens delstikprøven med $X_1 > t_1$ (højre gren) underopdeles ved t_3–værdien for X_1. Og så videre, indtil opdelingen i dataområder R_1 til R_5 er opnået.

FIGUR B1 **ILLUSTRATION AF BESLUTNINGSTRÆER**

a) Illustration af beslutningstræ **b) Opdeling af 'udfaldsrummet'**



Anm.: [Tekst]

Kilde: Hastie mfl. (2008), figur 9.2.

Prædiktioner ud fra flere tilfældige stikprøver

Som det ene element i random forest-estimatoren udtrækkes B tilfældige stikprøver fra data. Litteraturen betegner dette som *bagging* (bootstrap aggregation), jf. Hastie mfl. (2008).

Den generelle begrundelse for *bagging* er, at det danner et gennemsnit over mange estimater, der har høj varians, men er omtrent unbiased, jf. Hastie mfl. (2008). Dette skyldes, at de tilfældige stikprøver alle er udtrukket fra samme population (af boliger). De følger derfor samme fordeling som det originale data og derfor er den forventede prædiktion af gennemsnittet over alle stikprøver det samme som den forventede prædiktionen ud fra populationen, $E\hat{f}(\cdot) = E\frac{1}{B}\sum \hat{f}_b(\cdot)$. Prædiktionen har derfor en uændret bias. Samtidig reducerer det større antal observationer/prædiktioner variansen af den opnåede prædiktion.

Dannelse af beslutningstræer og korrelation mellem træerne

Som det andet element af Random Forest-metoden opdeles hver bootstrappede stikprøve i et antal dataområder vha. et beslutningstræ, hvor variablene trinvist opdeler data i subområder/blade, jf. figur B1a. I figur B1 er der eksempelvis fem områder, R_1-R_5, så fx $\bar{y}_1 = \frac{1}{N_1}\sum_{i \in R_1} y_i$ med N_1 værende antal observationer i R_1.

På øverste niveau af beslutningstræet opdeles data i to dele ud fra den variablen X_k og den værdi c af variablen, der giver den største reduktion i tabsfunktionen. Det giver følgende minimeringsproblem:

$$\underset{k,c}{\operatorname{argmin}} L(k,c) = \sum_{i:x_{ik}\leq c}^{N} \left(y_i - \bar{y}_{k,c,l}\right)^2 + \sum_{i:x_{ik}>c}^{N} \left(y_i - \bar{y}_{k,c,r}\right)^2$$

hvor de kvadrerede afvigelser nu dannes separat for hvert at de to delstikprøver hhv. til venstre (left, l) og til højre (right, r) for cut-off c, jf. Athey og Imbens (2019) På de følgende niveauer fortsættes på samme måde, hvor hvert af de nye subsamples igen opdeles ud fra den variabel, X_k, og den cutoff-værdi, c, der giver størst reduktion i tabsfunktionen.

Ved hvert split af data er det dog kun en tilfældig delmængde af x-variablene, som random forest-modellen kan vælge imellem. Hvis der er p x-variable i alt, vil der kun blive valgt fra $m < p$ mulige x-variable, der er tilfældigt valgt ved hvert split. Normalt er $m = \sqrt{p}$ eller så lavt som 1, jf. Hastie mfl. (2008).

Begrundelsen for kun at vælge en delmængde af variable er at det mindsker korrelationen mellem hvert beslutningstræ og dermed forbedre variansreduktionen for prædiktionen fra *bagging* (jf. foregående afsnit). For B prædiktioner, der følger samme fordeling (men har en parvis korrelation ρ) er variansen $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$. *Bagging* for sig selv kan reducere variansen via andet led, da $\frac{1-\rho}{B}$ går mod nul med stigende B. Men variansen går ikke samlet set mod nul, hvis $\rho > 0$, da det første led ikke forsvinder. Ved at tilfældigt udvælge m x-variable ved hvert split mindskes træernes korrelation, ρ. Optimalt set fjernes korrelationen helt, så variansen vil være $\frac{1}{B}\sigma^2$ og derfor gå mod nul med stigende B, jf. Hastie mfl. (2008).

Implementering af område-fixed effects i random forest

Det kan være problematisk at inkludere områdevariable som fixed effects i machine learning metoder som random forrest, jf. Johannemann mfl. (2019) og Hastie mfl. (2008). Implementeringen af fixed effects vil ske med dummies for hvert område, hvilket med M områder giver $2^{M-1} - 1$ mulige splits i en RF-model, jf. Johannemann mfl. (2019). Effekten for en del områder vil blive sat til 0, hvis der er for få observationer, eller huslejen i området ikke adskiller sig i 'tilstrækkelig grad' fra det generelle huslejeniveau. Det kan mindske modellens prædiktionsevne.

Område-specifikke effekter vurderes dog at være vigtige forklaringsfaktorer for markedsbestemte huslejer. Til opgørelsen af reguleringsgevinster anvender derfor en metode, der muliggør brugen af fixed effects ved mindske dimensionerne af de kategoriske område-variable, jf. Johannemann mfl. (2019).

Den centrale antagelse er en *sufficient latente variable* antagelse, jf. figur B2Figur B2: den kategoriske variabel, G, for et område har ingen kausal indflydelse på markedsbestemte huslejer, Y, men er korreleret med en *kategorisk* latent (uobserverbar) variable, L, som har en direkte effekt på Y. Den kategoriske variabel G korrelerer kun med huslejer gennem L, og har ikke selvstændig effekt på huslejer. *Sufficient latente variable*-antagelsen vil derfor være, at det ikke er f.eks. selve skoledistriktet, der direkte påvirker huslejen, men derimod en række faktorer som er korreleret med distrikterne, fx skolens kvalitet, gode transportforbindelser eller grønne områder.

Figur B2: Sufficient latent variable-antagelse

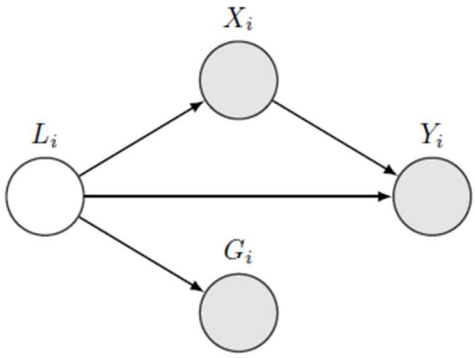


Figure 1: Causal graph depicting the key assumption that Y_i and X_i are independent of group membership G_i conditionally on latent state L_i. The grayed-out nodes are observed.

Kilde: Johannemann mfl. (2019).

Til opgørelsen af markedsbestemte huslejer, hvor antallet af kategorier i den kategoriske variabel er stort ift. antallet af variable i X, dvs $p \ll M$, anvender vi derfor såkaldt *means encoding*. *Means encoding* betegner gruppespecifikke gennemsnit for alle X-variable for hvert område i G. Metoden danner herved p nye variable (p er antal X-variable), som erstatter de M dummies for den kategoriske variabel, G. Johannemann mfl. (2019) viser viser, at metoden forbedrer prædiktionsevnen ift. bare at medtage område-fixed effects som dummies.

C Korrektion af boligens opførelsesår

Effekten af boligens opførelsesår indarbejdes vha. en metode, der består af seks trin. Metoden udnytter, at ejerboliger er opført i alle år, både før og efter 1991. Herved kan den relative effekt af opførelsesår på ejerboligpriserne estimeres. Denne effekt konverteres til en effekt på huslejen.

Før random forest-modellen estimeres, opjusteres huslejen for alle uregulerede lejeboliger (nyudlejde boliger opført efter 1991), så den svarer til, at boligen er opført i 2010'erne. Justeringen sker på baggrund af effekten af opførelsesåret på salgsprisen for ejerboliger, jf. trin 1-4. Den estimerede markedsbestemte husleje estimeres på baggrund af disse

justerede huslejer (trin 5) og vil på den måde afspejle huslejen, hvis boligen var opført i 2010'erne. De estimerede markedsbestemte huslejer for ældre lejeboliger skal nedjusteres, så niveauet for huslejen afspejler, at disse boliger er opført i årerne frem til 1991 og ikke i 2010'erne. Det gøres i trin 6. Metoden er estimeret særskilt for lejligheder, rækkehuse og parcelhuse (anv), og effekten for opførelsesår er opdelt på, om boligen er blevet ombygget eller væsentligt istandsat hhv. de seneste 30 eller 15 år (omb). De seks trin er beskrevet i detaljer nedenfor:

I **trin 1** estimeres to hedoniske OLS-regressioner. Den første estimerer sammenhængen mellem log-transformeret salgspris på ejerboliger pr. kvm. og en lang række forklarende variable, herunder dummies for opførelsesåret i 5-årsintervaller for ejerboliger. Den anden specifikation regresserer de den log-transformeret husleje for uregulerede lejeboliger (UR) på de samme forklarende variable:

$$\log(ejendomspris\ pr.\ kvm_{anv}) = \beta_{anv}^{EB} X + \pi_{t,omb,anv}^{EB} + \epsilon \quad \text{(A1)}$$

$$\log(husleje\ pr. kvm_{anv}) = \beta_{anv}^{UR} X + \pi_{t,omb,anv}^{UR} + \epsilon \quad \text{(A2)}$$

X er andre relevante faktorer til at forklare huslejen for salgsprisen. Konkret er de det samme faktorer som i random forrest modellen i fleksibel form samt fixed effekts for postnumre. Ligning (A2) estimeres kun for $t \geq 1992$, da de uregulerede lejeboliger ikke er opført før dette år.

I **trin 2** transformerer vi $\pi_{t,omb,anv}^{EB}$-koefficienterne, så de viser den eksakte procentvise ændring:

$$\tilde{\hat{\pi}}_{t,omb,anv}^{EB} = \exp\left(\hat{\pi}_{t,omb,anv}^{EB}\right) - 1 \quad \text{(A3)}$$

Det gøres, fordi log-kvotienterne kun viser den approksimative procentvise ændring. Denne approksimation er god for $\hat{\pi}_{t,omb,anv}^{EB}$ 'tæt' på nul, men vil undervurdere den sande procentvise ændring ved høje (numeriske) værdier af $\hat{\pi}_{t,omb,anv}^{EB}$.

Trin 3 tager højde for at effekten af opførelsesår kan være forskellig for ejendomspriser og husleje, jf. den blå og grønne søjle i figur 3. Derfor beregnes en skalleringsfaktor for hver af de tre anvendelsestyper ved at tage det gennemsnitlige forhold mellem de estimerede effekter for hvert interval fra 1991-2020:

$$scale_{anv} = \frac{1}{T}\sum_{t \in T} \frac{\pi_{t,anv}^{UR}}{\pi_{t,anv}^{EB}} \quad \text{(A4)}$$

Hvert element i summen svarer til forholdet mellem de blå og grønne søjler i figur 3. Denne skalleringsfaktor ganges på estimaterne fra ligning (A3) for at få den egentlige effekt svarende til de røde søjler i figur 3:

$$\tilde{\hat{\pi}}_{t,omb,anv}^{EB,scal} = \tilde{\hat{\pi}}_{t,omb,anv}^{EB} \cdot scale_{anv} \quad \text{(A5)}$$

I **trin 4** bruger vi estimaterne fra ligning (A5) til at korregere huslejen for uregulerede lejeboliger, så huslejen svarer til at alle, at boliger er opført i samme periode (i slutningen af 2010'erne):

$$hl^{RF} = \frac{hl^{AL}}{(1+\tilde{\pi}^{EB,scal}_{t,omb,anv})}, \qquad \text{for } t > 1991 \qquad \text{(A6)}$$

I **trin 5** bruger vi den korrigerede husleje fra ligning (A6), hl^{RF}, til at kalibrere random forest modellen. Den prædikterede markedsbestemte husleje ($\widehat{RF}$) vil herved være baseret på, at boligen er opført i slutningen af 2010'erne.

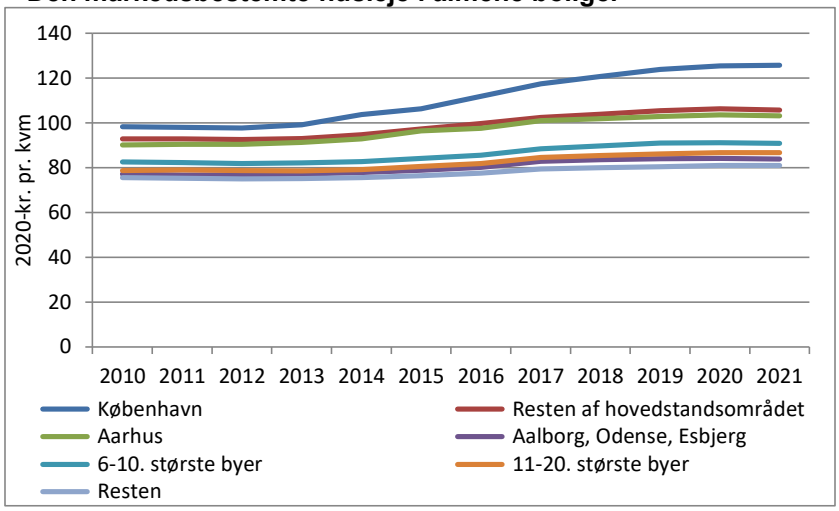
Til sidst i **trin 6** tilføjes effekten af opførelsesåret til den random forest-prædiktion af markedsbestemte husleje, $\widehat{RF}$, hvor vi bruger de skallerede estimater for opførelsesår fra OLS-modellen, jf. ligning (A5):

$$\widehat{hl^{mb}} = \widehat{RF}(1 + \tilde{\pi}^{EB,scal}_{t,omb,anv}) \qquad \text{(A6)}$$

Dermed er $\widehat{hl^{mb}}$ den estimerede markedsbestemte husleje, som tager højde for opførelsesåret.

D Supplerende resultater

Figur D1: Den markedsbestemte husleje i almene boliger



Anm.: Figuren viser den gennemsnitlige markedsbestemte husleje pr. kvm. for almene lejeboliger for hvert år og i hver byområde. Tallene for 2018 er baseret på en lineær interpolation mellem 2017 og 2019, jf. afsnit 2. København omfatter både Københavns Kommune og Frederiksberg Kommune. De resterende kommuner grupperes efter bystørrelse, således at "6.-10. største byer" eksempelvis betegner de kommuner, der rummer de 6.-10. største byer. De 6.-10. største kommuner er: Randers, Horsens, Kolding, Vejle og Roskilde. De 11.-20. største kommuner er: Herning, Silkeborg, Hørsholm, Helsingør, Næstved, Viborg, Fredericia, Køge, Holstebro, Taastrup.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Tabel D1: Theil-indeks for reguleringsgevinsterne og boligstøtte, almene boliger

	Årlig regule-ringsgevinst	Gevinst ift. husleje	Gevinst ift. indkomst	Årlig boligstøtte
Samlet Theil-indeks	0,32	0,33	0,40	0,64
	---------- Pct. ----------			
Forskel på tværs af indkomstgrupper	0,8	2,2	14,5	27,3
Forskel indenfor indkomstgrupper	99,2	97,8	85,5	72,7

Anm.: Tabellen viser, hvor stor en del af de samlede forskelle i reguleringsgevinsterne (og boligstøtten), som skyldes forskelle på tværs af og indenfor indkomstgrupperne. Gevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset. Reguleringsgevinsten er ækvivaleret. Indkomstdecilerne er beregnet ud fra alle husstande i Danmark uanset boligform.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.

Tabel D2: Theil-indeks for reguleringsgevinsterne og boligstøtte, 100 indkomstgrupper

	Årlig reguleringsgevinst		Reguleringsgevinst ift. indkomst		Årlig Boligstøtte	
	Ældre lejebolig	Almen bolig	Ældre lejebolig	Almen bolig	Ældre lejebolig	Almen bolig
Samlet Theil-indeks	0,65	0,32	0,77	0,40	0,76	0,64
	---------- pct. ----------		---------- pct. ----------		---------- pct. ----------	
Forskel på tværs af indkomstgrupper	3,0	1,0	17,5	20,1	30,6	29,5
Forskel indenfor indkomstgrupper	97,0	99,0	82,5	79,9	69,4	70,5

Anm.: Tabellen viser, hvor stor en del af de samlede forskelle i reguleringsgevinsterne (og boligstøtten), som skyldes forskelle på tværs af og indenfor indkomstgrupperne. Gevinster under 0 er sat til laveste værdi over nul for, at de kan medtages i Theil-indekset. Reguleringsgevinsten er ækvivaleret. Indkomstpercentilerne er beregnet ud fra alle husstande i Danmark uanset boligform. Ældre lejeboliger er private lejeboliger opført frem til 1991.
Kilde: Egne beregninger på baggrund af registerdata og data fra EjendomDanmark.