Fremskrivning af ejendomsværdien af parcelhuse

Peter P. Stephensen, Jonas Zangenberg Hansen, Esben Hvid Jørgensen og Marianne Frank Hansen

DREAM

Amaliegade 44, 1256 København K, Denmark

psp@dreammodel.dk

Juni 2015

1. Indledning

Dette notat er afrapportering af projektet "Fremskrivning af ejendomsværdien af parcelhuse". Der blev indgået kontrakt om dette projekt mellem Skatteministeriet og DREAM den 29. april 2015. Arbejdet blev udført af DREAM i perioden 1. maj - 12. juni 2015. Projektets resultat, er udviklingen af en robust metode til fremskrivning af ejendomsværdier og et datasæt bestående af fremskrevne ejendomsværdier (herefter kaldet priser) for samtlige parcelhuse solgt i perioden 2006-2010 til de 4 kvartaler i 2011.

Den benyttede metode bygger på geografisk vægtet regression (ofte kaldet GWR). Det er en moderne beregningstung metode der via sin geografiske vægtning giver en fin balance mellem lokale og globale egenskaber i beskrivelsen af et system. Metoden sammenlignes med *statiske forventninger* (svarende til at man fremskriver med den oprindelige pris) og en metode der bygger på *medianen af nabohandler (*NH-metoden*)*. Ekspertudvalget benyttede NH-metoden med 50 nabo-handler.

Ved at se på genhandler i perioden 2006-2011 undersøges det hvor mange naboer man skal inddrage under NH-metoden for at få en optimal fremskrivning. Det vises at der er et lokalt maksimum i træfsikkerhed ved 30-50 handler, man at det globale maksimum ligger tæt på 2.000 naboer. Det vises desuden at man kan risikere at træfsikkerheden ved anvendelse af 50 naboer, er ringere end ved anvendelse af den oprindelige pris (statiske forventninger).

Konklusionen er altså at man skal bruge store naboområder. Dette fører naturligt tanken hen på geografisk vægtet regression der typisk anvender meget store naboområder (5-10.000 naboer eller mere). Det vises at GWR er den af de 3 metoder der giver den mest præcise fremskrivning med en RMSE på godt 0,23 svarende til en pm20 på ca. 0,78.

2. *Hvad er en god fremskrivningsmetode?*

Det første og mest central krav til en fremskrivningsmetode er *træfsikkerhed*. Den fremskrevne kvadratmeterpris bør ligge så tæt som muligt på den "sande" handelsprisen for den betragtede bolig på fremskrivningstidspunktet (vi vender tilbage til hvorledes dette kan måles). Kvadratmeterpriser i et område, målt over tid, vil indeholde geografiske/spatiale og tidsmæssige informationer. Ved fremskrivning lægges der ifølge sagens natur vægt på de tidsmæssige informationer. Herved er der en fare for at de spatiale informationer ødelægges. Sker dette, vil anvendeligheden af de fremskrevne priser i den nye vurderingsmodel forringes. Formålet med at inddrage fremskrevne historiske handelspriser i vurderingsmodellen, er netop at opnå en mere præcis vurdering den geografiske variation. Måden at sikre dette på, er at vælge metoden med den bedste træfsikkerhed.

Et andet væsentligt krav er *spatial kontinuitet*. To boliger der ligger tæt på hinanden og som har beslægtede karakteristika, bør have relativt ens fremskrevne kvadratmeterpriser, uanset hvornår de oprindelige salg foregik. Dette er vigtigt både af principielle og praktiske årsager. Rent principielt ønsker vi som nævnt ovenfor at bibeholde den geografiske/spatiale variation. Ud fra en mere pragmatisk synsvinkel er det væsentligt at den endelige model ikke giver anledning til "vilde" udsving i priserne på tæt beliggende boliger. En nødvendig forudsætning for dette er, at de fremskrevne priser ikke udviser for store udsving.

Endelig vil der blive lagt vægt på *tidsmæssig kontinuitet*. Skat har efterspurgt muligheden for at kunne frem- og tilbageskrive handelspriser i hele den historiske periode. Sådanne bolig-specifikke prisindeks vil fx kunne bruges til dannelsen af aggregerede kvartalsvise/månedlige prisindeks. De vil ligeledes potentielt kunne bruges i sagsbehandling, hvis der skal gives et hypotetisk bud på en vurdering tilbage i tiden. Kombinationen af spatial og tidsmæssig kontinuitet indebærer at disse bolig-specifikke prisindeks dels ikke springer for

meget over tid, dels har en samlet tidsmæssig udvikling der er relativt ens for to boliger der ligger tæt på hinanden og har relativt ens karakteristika.

3. *Fremskrivningsmetoder*

Vi vil analysere 3 forskellige fremskrivningsmetoder: median af nabohandler, statiske forventninger og 'Geographically Weighed Regression'.

Median af nabohandler (NH) er den metode ekspertudvalget benyttede. Ekspertudvalget benyttede 50 nabohandler, men som vi skal komme tilbage til, kan et hvilken som helst antal naboer benyttes. Hvis kvadratmeterprisen for en given bolig P_0 skal fremskrives fra år t_0 til t, sker det ved først at beregne medianen M_0 af de 50 nærmeste handler i år t_0. Man benytter medianen (i stedet for gennemsnittet) for at gøre metoden *robust*, - dvs. nogenlunde uafhængig af 'outliers'. Herefter beregner man medianen M af de 50 nærmeste handler i år t. Man fremskriver herefter kvadratmeterprisen:

$$P = \frac{M}{M_0} P_0$$

Intuitionen er oplagt: hvis en bolig i år t_0 er 10 procent dyrere end medianprisen i nærområdet, så er den også 10 procent dyrere i et senere år t. Eller sagt på en anden måde: hvis priserne i nærområdet er steget 2 procent, da stiger prisen på den betragtede bolig også 2 procent. Grundideen er at man kan tale om en *lokal kvadratmeterpris*, og at den enkelte boligs værdi relaterer sig til denne lokale pris.

NH-metoden har en række ulemper. Metoden ser kun på kvadratmeterprisen, og tager derfor ikke hensyn til hvilke karakteristika boligerne i naboområdet har. Især vil de karakteristika som medianboligen har, være helt tilfældige, og bør derfor ikke bruges til noget i analysen. Et andet problem er, at naboudvælgelsen i det historiske år og fremskrivningsåret ikke er tæt relateret. Hvis der fx var mange handler i det historiske år og få handler i fremskrivningsåret, vil naboområdet være geografisk størst i fremskrivningsåret. En sammenligning af medianerne M_0 og M bliver derfor ikke nødvendigvis retvisende. Hvis vi fx

forestiller os at den betragtede bolig ligger på landet, da kunne naboområdet i det historiske år bestå af boliger der alle ligger på landet, mens naboområdet i fremskrivningsåret kunne udvide sig til udkanten af en provinsby på grund af det lave handelsniveau.

Vi vil nedenfor især beskæftige os med logaritmen til kvadratmeterprisen (log-kvadratmeterprisen). Hvis vi laver en stor/lille-bogstav-konvention der hedder: log('stort bogstav') = 'lille bogstav', kan vi omskrive ovenstående ligning til:

$$p = (m - m_0) + p_0$$

Den anden fremskrivningsmetode er *statiske forventninger* (SF). Her antages det ganske simpelt at

$$p = p_0$$

Den fremtidige pris antages at være den sammen som den oprindelige. Årsagen til at vi medtager denne oversimplificerede fremskrivningsregel er at den klarer sig overraskende godt. Vi berørte allerede i sidste afsnit det trade-off der kan være mellem måling af spatiale og temporære (tidsmæssige) informationer. Den enkelte boligs kvadratmeterpris må antages at være bestemt af individuelle, spatiale og temporære forhold. De individuelle forhold dækker over diverse karakteristika, herunder generel stand. De spatiale forhold dækker over beliggenhed i videste forstand: udsigt, afstand til gode ting (arbejdspladser, skole, butikker, natur), dårlige ting (støj, lugt) osv. De temporære forhold dækker over ændringer i de to andre faktorer, samt vigtigst, udviklingen i boligmarkedet. I NH-metoden antages det at, den temporære udvikling er uafhængig af den individuelle bolig, således at den kan måles ved at se på udviklingen i naboområdet. De individuelle og spatiale forhold antages derimod entydigt at være bestemt at den initiale pris p_0. Der er derfor fare for at metoden fokuserer for meget på den temporære udvikling, og derved ødelægger individuelle og spatiale informationer. I modsætning til dette, ignorerer statiske forventninger fuldstændig det temporære aspekt. Som følge af dette er metoden ekstremt skånsom overfor individuelle og spatiale informationer. Det er

sandsynligvis dette der forklarer at metoden er overraskende god sammenlignet med NH-metoden.

Den 3. metode er geografisk vægtet regression (GWR for Geographically Weighed Regression). Dette er en spatial metode der har vundet udbredelse i de seneste 15 år (Fotheringham, Brunsdon & Charlton 2002 ; Bivand, Nakaya & Garcia-Lopez 2013; Gollini, Lu, Charlton, Brunsdon & Harris 2015). Metoden forstås bedst hvis vi tager udgangspunkt i en almindelig lineær regression. Antag at log-kvadratmeterprisen er givet ved

$$p = \beta x + \varepsilon$$

hvor x er en vektor med boligens karakteristika, β er koefficienter og ϵ er et normalfordelt restled. Dette kunne fx være ekspertudvalgets lineære model eksl. nabo-priser og kommune-dummies og inkl. kvartalsdummies for hele perioden 2006-2011. Hvis vi estimerede denne for hele perioden 2006-2011 ville vi have en såkaldt *global model*. Her antages det implicit at den marginale priseffekt af fx tegltag eller en given afstand til motorvej, er den samme i hele landet. Det vi ønsker er imidlertid en *lokal model*. Vi antager at den enkelte bolig er defineret ved (p, x, s) hvor $s = (s_1, s_2)$ er boligens spatiale position, dvs. ved log-kvadratmeterpris, karakteristika og spatial position. Herefter ønsker vi at estimere modellen:

$$p = \beta(s)x + \varepsilon$$

Vi tillader at koefficienterne β er en ikke-lineær funktion af den spatiale position. Denne sammenhæng ønsker vi at estimere ikke-parametrisk, dvs. at vi ikke på forhånd har gjort antagelser om den funktionelle form. I de seneste 15-20 år er det i stigende grad blevet populært at benytte lokal regression i stedet for global regression (Cleveland, Grosse & Shyu 1992; Li & Racine 2007) . Dette kendes fx fra 'Local Polynomial Regression Fitting' (Loess kaldes metoden i R) og 'Kernel Density Estimation' (Li & Racine 2007). Ideen er at man i stedet for at lave en global regression, laver en regression for hvert eneste datapunkt. I princippet er alle datapunkter med i hver estimation, men bliver tilordnet en vægt der afhænger af afstanden til det centrale datapunkt i estimationen. Som regel er det

et problem at definere dette afstandsmål, men netop i spatiale analyser er valget af afstandsmål ret indlysende. Vægtningen 'tunes' således at modellens træfsikkerhed optimeres (mere om det senere).

Lad os tage et eksempel. Vi betragter en bestemt villa udenfor Roskilde der blev solgt 1. kvartal 2007 til log-kvadratmeterprisen p_0. Vi vil gerne fremskrive dens kvadratmeterpris til 1. kvartal 2011. I princippet laver vi nu en lineær regressionsanalyse med alle ekspertudvalgets variable (eksl. nabo-priser og kommune-dummies) og med alle handler i Danmark i perioden 2006-2011 som data; - kun til ære for denne ene bolig. Den enkelte handels vægt i regressionen afhænger af afstanden til den aktuelle bolig udenfor Roskilde. I praksis vil mange handler derfor ikke indgå i analysen, fordi de er så langt væk, at vægten er sat til 0. Som vi vil se senere vil antallet af handler der i praksis indgår i estimationen ligge i intervallet 3.000-10.000. I estimationen indgår kvartals-dummies for perioden 2006-2011. Modellen giver derfor et bud på den lokale kvadratmeterpris for alle kvartaler i denne periode. Lad os definere gwr_0 og gwr som modellens bud på log-kvadratmeterprisen 1. kvartal 2007 og 1. kvartal 2011. Vi sætter nu den fremskrevne pris i 1. kvartal 2011 til:

$$p = (gwr - gwr_0) + p_0$$

Intuitionen er den samme som den vi havde i relation til nabohandler (NH-metoden). Hvis en ejendom i 1. kvartal 2007 blev solgt 10 procent over den lokale kvadratmeterpris (her beregnet ved GWR-metoden), så forventes den også at blive solgt 10 procent over 1. kvartal 2011. GWR kan siges at være en (voldsom) generalisering af NH-metoden. Hvor NH-metoden ikke udnyttede informationer om naboboligernes karakteristika, gør GWR-metoden i høj grad dette. Og hvor NH-metoden havde problemer med forskellige naboområder i det historiske og det fremskrevne år, er der i GWR-metoden kun defineret et naboområde for alle kvartaler. Naboområderne vil i GWR-metoden typisk være meget store (ofte med en radius over 10 km), men de lokale egenskaber sikres via nedvægtningen af data-punkter der ligger langt væk. De store områder sikrer at der er data nok

til at estimere modellen, mens den geografiske vægtning sikrer målingen af de lokale egenskaber.

Tabel 1 Variable anvendt i GWR-regressioner

Variabel	Type	Tekst
AFST_HJSP_f	Faktor	Afstand til højspændingsledning
AFST_HOVEDVEJ_f	Faktor	Afstand til hovevej
AFST_JERNB_f	Faktor	Afstand til jernbane
AFST_KYST_f	Faktor	Afstand til kyst
AFST_LANDEVEJ_f	Faktor	Afstand til landevej
AFST_MVEJ_f	Faktor	Afstand til motorvej
BYG_VANDFORSY_KODE_f	Faktor	Vandforsyningskode
KYEST_FOERSTE_f	Faktor	Første række til kyst
OMBYG_ALDER_f	Faktor	Ombygningsalder
OPFOERELSE_AAR_f	Faktor	Opførelsesår
OPVARMNING_KODE_f	Faktor	Opvarmningskode
SOE_FOERSTE_f	Faktor	Første række til sø
TAG_KODE_f	Faktor	Tagkode
VARME_SUPPL_KODE_f	Faktor	Kode for supplerende varme
YDERVAEG_KODE_f	Faktor	Kode for ydervæg
ZONESTATUS_f	Faktor	Zonestatus (landzone, byzone)
ANTBADEVAERELSER	Tal	Antal badeværelser
ANTVANDSKYLTOIL	Tal	Antal toiletter
CARPORT_INDB_ARL	Tal	Carport, areal
GARAGE_INDB_ARL	Tal	Garage, areal
Grundareal	Tal	Grundareal
UDHUS_INDB_ARL	Tal	Udhus, areal
KAELDER_ARL_U_125M	Tal	Kælderareal under 125 m2
BYG_BEBYG_ARL	Tal	Bebygget areal
UDESTUE_ARL	Tal	Udestue areal
TAGETAGE_ARL_UDNYT	Tal	Areal af udnyttet tagetage
Bolig_Areal	Tal	Boligareal
Bolig_Areal^2	Tal	Boligareal, kvadreret
havuds_antal	Tal	Antal havudsigter
KVARTERTYPE_f	Faktor	Kvartertype (tæthedsmål)
AAR_KVARTAL_f	Faktor	Kvartal

4. Data og test

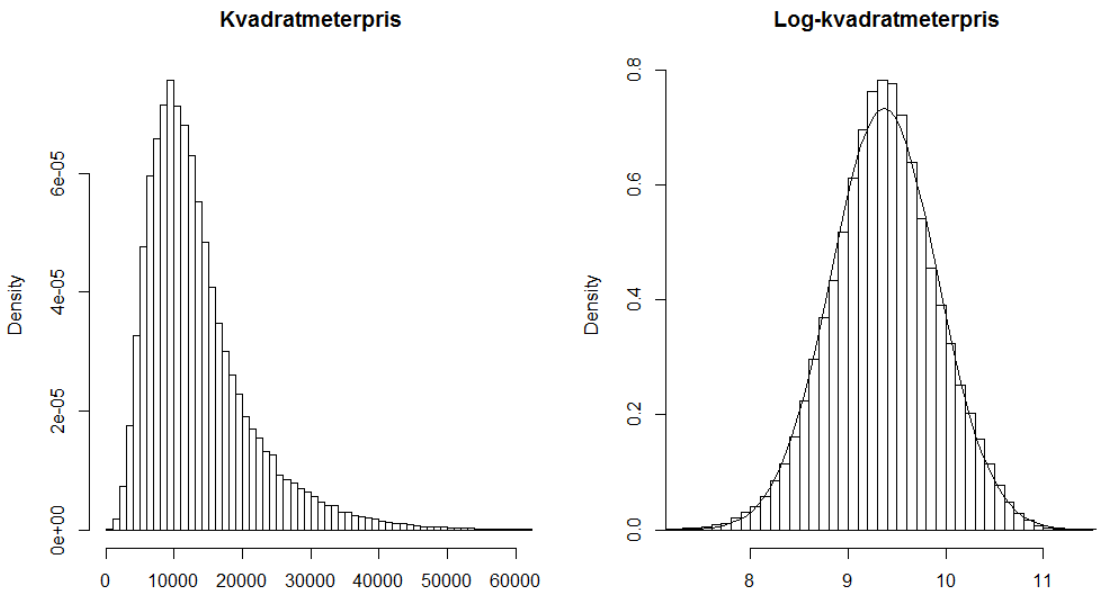
Analysen er baseret på grundlæggende de samme tal for parcelhuse som ekspertudvalget anvendte i deres rapport (Skatteministeriet 2014)[1]. Der var i perioden 2006-2011 139.049 handler med parcelhuse. Af disse blev 122.595 handler foretaget før 2011. Formålet med denne analyse er at fremskrive disse handler til 2011.

I GWR-analyserne regresseres på de 30 variable vist i tabel 1. En del af variablene er faktorer (også kaldet dummy-variable). I den udstrækning disse variable er opdelt på intervaller (fx for afstandsvariablene og opførselssår), er det sket på samme måde som i ekspertudvalgets rapport (Skatteministeriet 2014). For en nærmere beskrivelse af data henvises til ekspertudvalgets rapport.

Den centrale variabel der ønskes fremskrevet er *kvadratmeterprisen*. Som det fremgår af den venstre figur i Figur 1 er fordelingen af kvadratmeterpriser meget vestre-skæv. Fordelingens højre hale er "tyk". Betragtes i stedet *log-kvadratmeterprisen* (højre figur i Figur 1) fås en fordeling der er relativt tæt på en normalfordeling (den nærmeste normalfordeling er markeret med den solide linje). Boligpriserne er tydeligvis log-normalfordelte. Dette giver ret god mening, idet den centrale determinant for boligefterspørgselen – indkomsten – er approksimativt log-normalfordelt. Vi vil i det følgende anvende log-kvadratmeterprisen som forklaret variabel. Den klassiske antagelse om at restledene er normalfordelte giver mest mening ved modellering af log-kvadratmeterprisen.

[1] En enkel variable er tilføjet siden da, nemlig 'Kvartertype'. Denne variabel er et tæthedsmål udviklet af Skat.

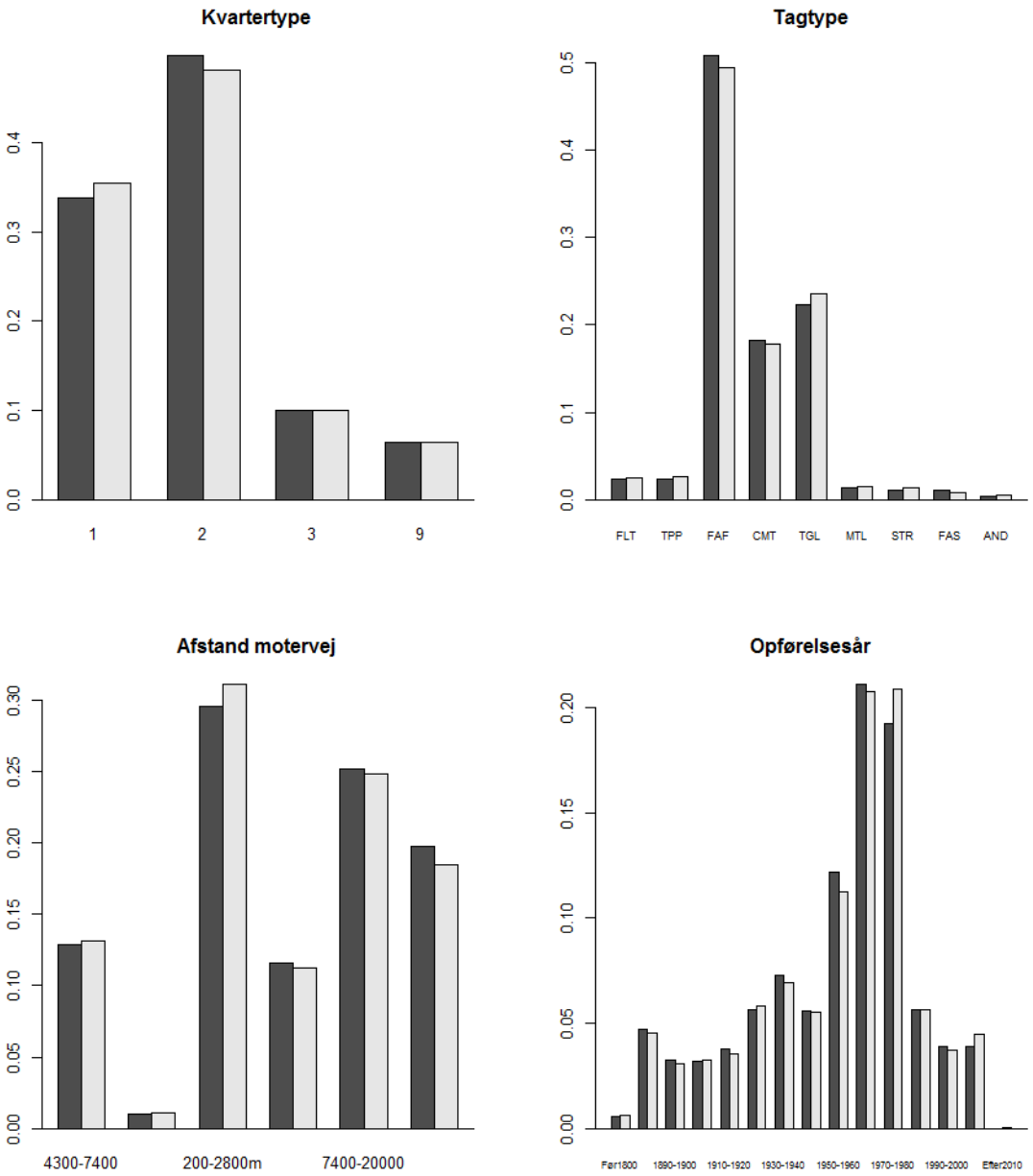
Figur 1 Fordeling af kvadratmeterpriser 2006-2011



Kilde: Egne beregninger

Det er umiddelbart vanskeligt at *teste* fremskrivninger af historiske boligpriser. Som regel er vi ude i en hypotetisk situation: hvad *ville* en ejendom koste *hvis* den blev solgt i 2011 og ikke i 2007? Vi har arbejdet med to potentielle løsninger på dette problem: test af den samlede vurderingsmodel og anvendelse af gensalg. Ved test af den samlede vurderingsmodel blev der konstrueret en "test bed" der efterlignede den endelige anvendelse af vurderingsmodellen. For given fremskrivningsmetode (der skal testes) blev der dannet lokale kvadratmeterpriser for alle salg i 2011. Herefter blev ekspertudvalgets oprindelige model estimeret med disse lokale kvadratmeterpriser som input. Dette blev udført i et 10-foldigt krydsvaliderings-setup (Hastie, Tibshirani, and Friedman 2009). Det viste sig vanskeligt at få præcise resultater ud af dette system. Årsagen er sandsynligvis den sammenblanding der sker af test af fremskrivningsmetode og test af den samlede metode.

Figur 2 Sammenligning af gensalg og alle salg



Kilde: Egne beregninger

I stedet har vi anvendt gensalg til at vurdere fremskrivningsmetodernes træfsikkerhed. Der er 8.864 gensalg i perioden 2006-2011 svarende til 6,3 procent af samtlige salg. Af disse forekommer 2.200 af gensalgene i 2011.

Gensalg gør det muligt at teste træfsikkerheden på fremskrivningsmetoden direkte. Problemet er hvis gensalg på en eller anden måde er atypiske i forhold til andre salg. Hvad angår boligernes karakteristika er der ikke betydelig forskel på gensalg og andre salg. Eksempler på dette er vist i Figur 2. Det ses her fx at fordelingen på tætheds-variablen kvartertype er meget ens. Dette betyder at gensalg er fordelt geografisk nogenlunde ligesom andre salg.

Tabel 2 Effekt af boligændringer

	Solgt en gang	Solgt flere gange
Tagets type ændret	0.023	0.025
Ekstra badeværelse	0.030	0.032
Garageareal forøget	0.006	0.007
Varmekilde ændret	0.059	0.096
Boligstørrelse forøget	0.072	0.080
Boligstørrelse formindsket	0.013	0.013
Stigning i boligareal, m^2	2.260	2.418

Anm: Cellerne viser andelen af boliger med den angivne ændring, undtagen nederste række, som viser stigningen i m^2 fra salgsåret til 2011.
Kilde: Egne beregninger

En antagelse kunne være, at huse der sælges igen efter en kort årrække, er blevet købt som 'håndværkertilbud', sat i stand, og derefter solgt med en fortjeneste. Dette er svært at se i data, da BBR giver et meget ufuldstændigt billede af en boligs stand. Der er dog nogle variable der kan opfange når en bolig bliver renoveret, og disse er sammenfattet i Tabel 2 for hhv. boliger solgt en og flere gange. Det fremgår af tabellen at en lidt større andel af gensalg har oplevet ændringer i karakteristika end enkeltsalg. Effekten ser dog ud til at være ret beskeden.

5. *Geografisk vægtet regression*

Der er udviklet en speciel version af GWR-metoden til dette projekt. Vi valgte selv at udvikle teknikken, i stedet for at anvende en af de eksisterende varianter i R, for at opnå fleksibilitet og sikre at metoden hastighedsmæssigt er optimeret

præcist til det den skal bruges til. Metoden er udformet som en R-funktion der har følgende format:

```
gwr_forecast(fml, forecastDate, dataDF, queryDF,
      kMin, kMax, r, kernel="bisquare", bw,
      verbose=TRUE, randomize=FALSE)
```

Argumenter:

Fml	Regressionsmodel formel som `formula` R-objekt
forecastDate	1) Vektor med samme længde som `nrow(quaryDF)` med kvartaler der skal fremskrives til. Formatet `"YYYYK"`, fx er `"20112"` 2. kvartal 2011. Alle boliger i `quaryDF` fremskrives til disse datoer. 2) Skalar med årstal. Alle boliger i `quaryDF` fremskrives til de 4 kvartaler i dette år.
dataDF	Boliger som `quaryDF` –boliger finder sine naboer blandt. `data.frame`.
queryDF	Boliger der skal fremskrives som `data.frame`.
kMin	Det mindste antal naboer
kMax	Det maksimale antal naboer
r	Radius
kernel	Funktion vælges således: gaussian: wgt = exp(-.5*(vdist/bw)^2) exponential: wgt = exp(-vdist/bw) bisquare: wgt = (1-(vdist/r)^2)^2 hvis vdist < r, wgt=0 ellers tricube: wgt = (1-(vdist/r)^3)^3 hvis vdist < bw, wgt=0

	ellers boxcar: wgt=1 hvis dist < bw, wgt=0 ellers (Kun `gaussian` og `bisquare` er implementeret)
Bw	Båndbredte i `gaussian` og `exponetial`.
Verbose	Hvis `TRUE`, udskrives status
randomize	Hvis `TRUE`, udskrives status i tilfældig orden

Funktionen returnerer et GWR-objekt. Dette objekt har elementerne:

Forecast	Hvis `forecastDate` har format 1: en vektor med fremskrevne log-kvadratmeterpriser. Hvis `forecastDate` har format 2: En matrix med 4 kollonner, indeholdende fremskrivning af log-kvadratmeterpriser i 4 kvartaler.
Itt	Antal iterationer i forbindelse med beregning af outliers (se nedenfor)
N	Effektivt antal naboer
Bw	Effektiv radius
RMSE	Root-mean-square-error beregnet på residualerne fra estimationen
pm20	pm20 beregnet på residualerne fra estimationen
coefficients	Et data.frame med alle estimerede parametre
SE	Et data.frame med beregnede 'standard errors' på alle estimerede parametre
t	Et data.frame med beregnede t-værdier på alle estimerede parametre

nSig	Antallet af variable der er signifikante, - dvs |t-værdi| < 2
x, y	x,y-korrodinater på boliger
nOutliers	Antallet af outliers (se nedenfor)
kMin, kMax, r, fml	Værdier af parametre der var input til proceduren

Lad os se på et eksempel. Vi vil fremskrive alle handler på Fyn for perioden 2006-2010 til alle 4 kvartaler i 2011. Dette kan gøres ved proceduren:

```
gwr_fyn = gwr_forecast(fmlEkspUdv, forecastDate = 2011,
                       dataDF = d_all,
                       queryDF = dd_fyn,
                       kMin = 3000, kMax = 5000, r = 10000)
```

Vi bruger ekspertudvalgets model beskrevet ved objektet `fmlEkspUdv`. `forecastDate = 2011` indebærer at vi fremskriver til alle 4 kvartaler i 2011 (format 2, jf. ovenfor). Alle fynske boliger ligger i data-frame'et `dd_fyn`, således at `queryDF = dd_fyn`. Vi søger naboer blandt alle boliger i hele landet og på alle tidspunkter (`dataDF = d_all`). Derfor kan en fynsk bolig i fx Middelfart i princippet inddrage boliger fra den jyske by Fredericia i sit naboområde. Vi fastsætter at et naboområde, som udgangspunkt skal bestå af alle boliger indenfor en radius på 10 km (`r = 10000`). Der skal dog mindst være 3.000 boliger i et naboområde og maksimalt 5.000. Hvis der for en given bolig i `queryDF` fx viste sig at være 2.500 boliger indenfor 10 km, da ville radius blive forøget automatisk således at der er 3.000 boliger i naboområdet. På samme måde; hvis det viser sig at fx er 7.600 boliger indenfor en radius på 10 km, da vil radius blive formindsket således at der er 5.000 boliger i naboområdet. Den effektive radius kan aflæses af elementet `gwr_fyn$bw`. Det ses at valg af kernefunktion ikke fremgår. Det skyldes at vi vælger default-værdien `"bisquare"` (beskrives nærmere i afsnit 6).

Funktionen `gwr_forecast` udfører grundlæggende 2 beregninger. Først finder den for hver bolig i `queryDF` samtlige naboer i `dataDF` som ligger mindre end `kMax` fra hinanden. Det gør den ved hjælp af proceduren `nn2` fra R-pakken `RANN`. Dette er en såkaldt ' Nearest Neighbour Search' der er kendt for sin hurtighed (Arya, Mount, Netanyahu, Silverman and Wu 1998; Jefferis 2015).

Herefter udfører funktionen en fuld vægtet estimation for hver bolig i `queryDF`. Vægtene beregnes ifølge den valgte kerne-funktion.

5.1. Robust estimation

Som nævnt i afsnit 3 gøres NH-metoden robust ved at betragte medianen i stedet for gennemsnittet i naboområdet. Derved får outliers en begrænset betydning. I GWR-metoden giver det ikke mening at betragte medianen, idet der udføres en estimation på hele naboområdets data. Vi bliver derfor nødt til at søge alternative veje for at sikre robusthed. I Fotheringham, Brunsdon & Charlton (2002) foreslås flere metoder. Vi har valgt den der kører hurtigst. Metoden finder outliers ved en iterativ proces. Efter en estimation beregnes de standardiserede residualer (residualer delt med standardafvigelse). Hvis den i'te observation i naboområdet har en standardiseret residual e_i^S da gives observationen en vægt:

$$w_i = 1 \; hvis \; e_i^S < 2$$

$$w_i = \left(1 - \left(e_i^S - 2\right)^2\right)^2 \; hvis \; e_i^S \in [2,3]$$

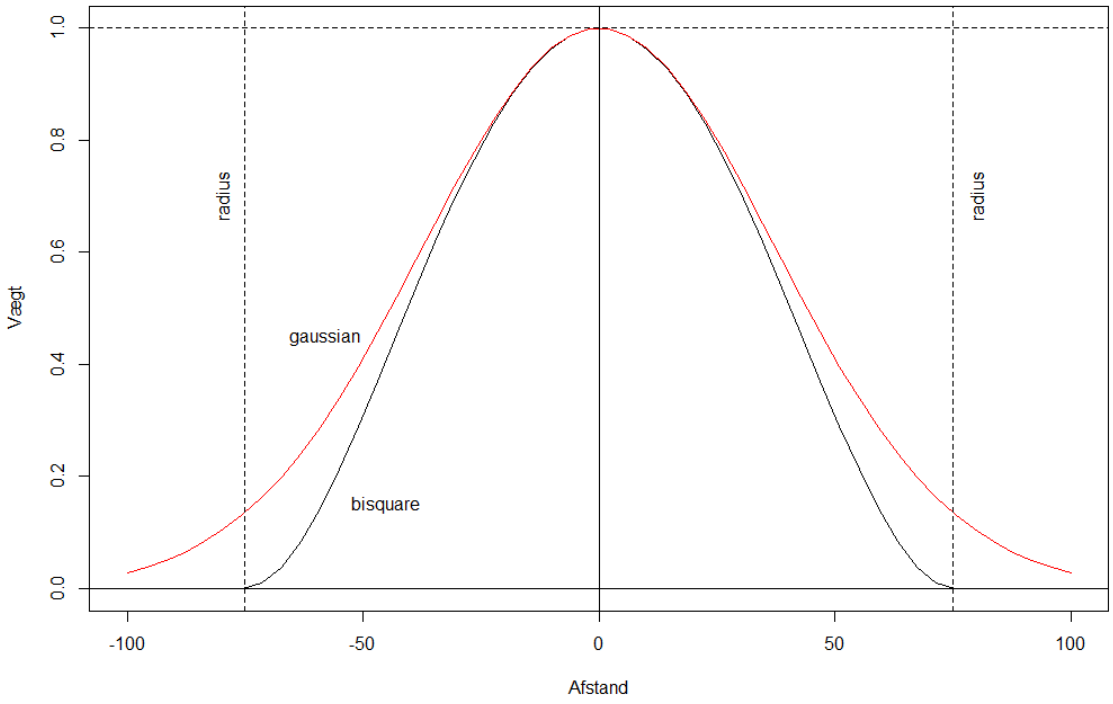
$$w_i = 0 \; hvis \; e_i^S > 3$$

Herefter udføres en regression hvor disse vægte indgår i tabs-funktionen. Denne procedure gentages indtil vægtene konvergerer. Det ses at hvis $w_i = 0$ svarer det til at observation i er en outlier. Hvis $w_i < 1$ er observationen stadig med i estimationen, men der er 'skuet ned' for dens indflydelse på resultatet.

Der vil typisk skulle foretages 10-15 iterationer før vægtene konvergerer. Hvis vi skal fremskrives 100.000 boliger skal der med andre ord udføres 1-1,5 mio.

estimationer. Dette skal sammenlignes med situationen hvor man har en global model. I dette tilfælde er antallet af estimationer: 1.

Figur 3 Kernefunktionerne 'gaussian' og 'bisquare'



6. *Tuning af modellen*

GWR-modellen har 4 parametre der kan skrues på: `kMin`, `kMax`, radius `r` og kerne-funktionen. Vi har i denne analyse lagt os fast på kerne-funktionen 'bisquare'. Den vægter tæt på centrum stort set ligesom en gaussisk kerne, men har i modsætning til den gaussiske kerne den egenskab at vægten går mod 0 ved fuld radius-afstand (se Figur 3).

De tre parametre `kMin`, `kMax` og `r` optimeres for hvert af områderne Sjælland, Jylland, Fyn og Bornholm. I det enkelte område vælges den kombination af de 3 parametre der giver størst træfsikkerhed. Som beskrevet i afsnit 4, benytter vi gensalg til at måle kvaliteten af den enkelte fremskrivning. Træfsikkerheden

måles via RMSE (Root-mean-squared-error). Den opnåede træfsikkerhed ses til højre i Tabel 3. At Sjællands RMSE er lig 0,2131 kan (approksimativt) fortolkes som at fremskrivningerne på Sjælland rammer gennemsnitligt 21,3 procent forkert. Det ses at fremskrivningerne på Sjælland er bedst, at de er nogenlunde lige gode på Fyn og i Jylland, og at Bornholm er dårligst. Til sammenligning er pm20 vist. Det ses at pm20 ligger i intervallet 0.76-0.79. Det er overraskende at Bornhold har den højeste pm20 på 0.789. Forklaringen er sandsynligvis at variationen i kvadratmeterpriser er mindre på Bornholm end i resten af landet.

Tabel 3 Optimalt valg af GWR-parametre

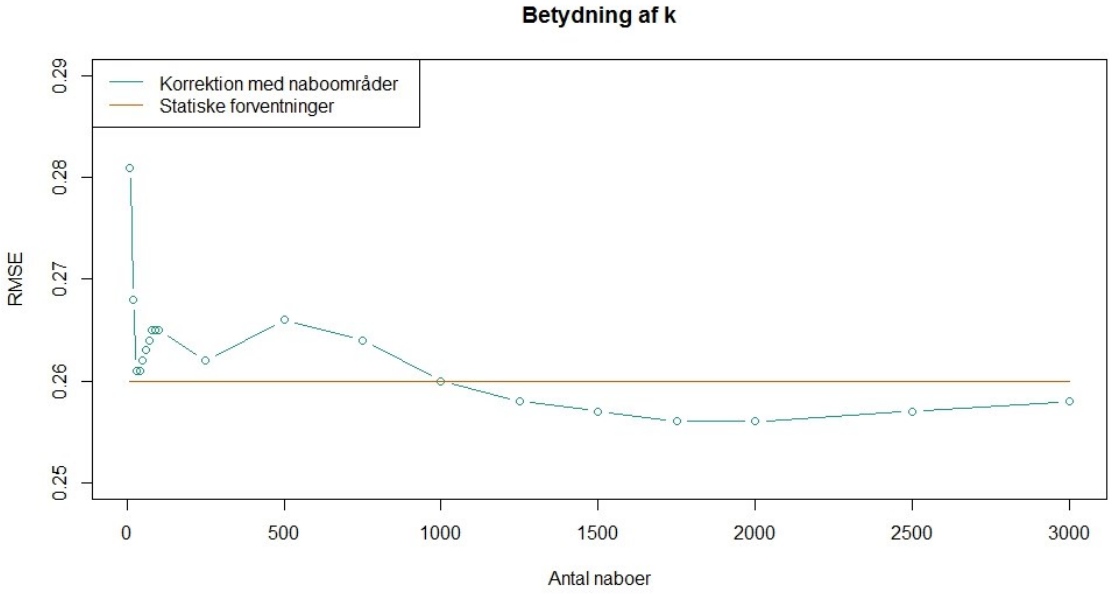
Område	kMin	kMax	Radius	Antal salg 2006-2011	Antal double-salg 2006-2011	RMSE	pm20
Sjælland	7500	7500	-	43786	2289	0.2131	0.780
Bornholm	1562	1562	-	1562	133	0.2821	0.789
Fyn	3000	5000	10000	13662	793	0.2318	0.760
Jyllan	10000	10000	-	80039	5649	0.2372	0.769

Kilde: Egne beregninger

Det optimale valg af parameterværdierne for de fire områder er sammenfattet i Tabel 3. Identisk valg af `kMin` og `kMax` svarer til, at estimationen altid baseres på det samme antal boliger, hvorved værdien for den maksimale søgeradius er irrelevant. For Bornholm, Sjælland og Jylland vurderes det mest hensigtsmæssigt at basere GWR estimationen på et konstant antal boliger, mens det for Fyn er optimalt at vælge mindst 3000 naboejendomme og lade enten en maksimal søgeradius på 10 km eller 5000 naboejendomme udgøre den øvre grænse for antallet af ejendomme i estimationen.

Der estimeres for hvert gensalg på det antal nabosalg som søgeparametrene giver anledning til. Naboerne vælges i udgangspunktet blandt samtlige handler i perioden 2006-2011. For Bornholm vurderes det dog mest hensigtsmæssigt udelukkende at se på nabohandler på Bornholm, hvilket sikres ved at anvende samme værdi for `kMin` og `kMax`, der er givet ved det samlede antal handler på Bornholm i perioden 2006-2011.

Figure 4 Sammenligning af statiske forventninger med NH-metoden



Kilde: Egne beregninger

En bolig, der er solgt mere end én gang, vil blive vurderet flere gange. Er en bolig eksempelvis solgt i 2006, 2008 og 2010 er der tale om gensalg. Salget i 2006 fremskrives til 2008 og 2010 og sammenholdes med de faktiske salgspriser i henholdsvis 2008 og 2010, mens salget i 2008 fremskrives til 2010 og sammenholdes med den faktiske salgspris i 2010. Ud fra forskellen mellem de faktiske og de fremskrevne salgspriser beregnes RMSE.

7. Sammenligning af fremskrivningsmetoder

Med udgangspunkt i datasættet med gensalg er de tre metoder blevet sammenlignet. Det viser sig at GWR giver de mest præcise fremskrivninger.

Som nævnt tidligere er det ikke indlysende hvor mange naboer man skal tage med i NH-metoden (median af nabohandler). For at undersøge dette, fremskrev vi de 2.200 handler der genhandles i 2011, med mange varierende antal naboer. Handelsprisen blev fremskrevet til 2011 og sammenlignet med den faktisk handlede pris. Resultatet af dette eksperiment ses i Figure 4. Til sammenligning

er der indlagt en orange kurve der viser RMSE ved statiske forventninger. Det ses at NH-metoden (den blå kurve) har et lokalt minimum ved 30 naboer; dvs. tæt på de 50 ekspertudvalget brugte. Dette lokale minimum er imidlertid dårligere end statiske forventninger (har højere RMSE). Vælger man imidlertid et meget større antal naboer (flere tusinde) bliver NH-metoden bedre end statiske forventninger. Som nævnt tidligere kan dette ses som tegn på at NH-metoden er for 'brutal' overfor de spatiale informationer. Eller sagt på en anden måde: det tyder på at de informationer der kan uddrages af data ligger på en meget højere/grovere 'spatial bølgelængde' end den 30-50 naboer giver anledning til. Dette fører naturligt frem til at inddrage GWR-metoden der typisk anvender et betydeligt antal naboer.

Table 4 Sammenligning af fremskrivningsmetoder. RMSE

	Alle år	2011
Statiske forventninger	0.2518	0.2571
50 naboer	0.2431	0.2609
2000 naboer	0.2388	0.2540
GWR	0.2315	0.2448

Kilde: Egne beregninger

I Table 4 ses RMSE for forskellige fremskrivningsmetoder og for to forskellige datasæt. Vi har dels målt træfsikkerheden på datasættet med 2.200 handler der genhandles i 2011, og dels for det fulde datasæt af alle godt 9 tusinde gensalg. Det ses at GWR-metoden har den laveste RMSE for begge datasæt. I det lille datasæt genfinder vi resultatet, at statiske forventninger er bedre en NH-metoden med 50 naboer. Dette resultat gælder imidlertid ikke i det fulde datasæt. NH-metoden med 2.000 naboer er den fremskrivningsmetode der ligger tættest på GWR-metoden.

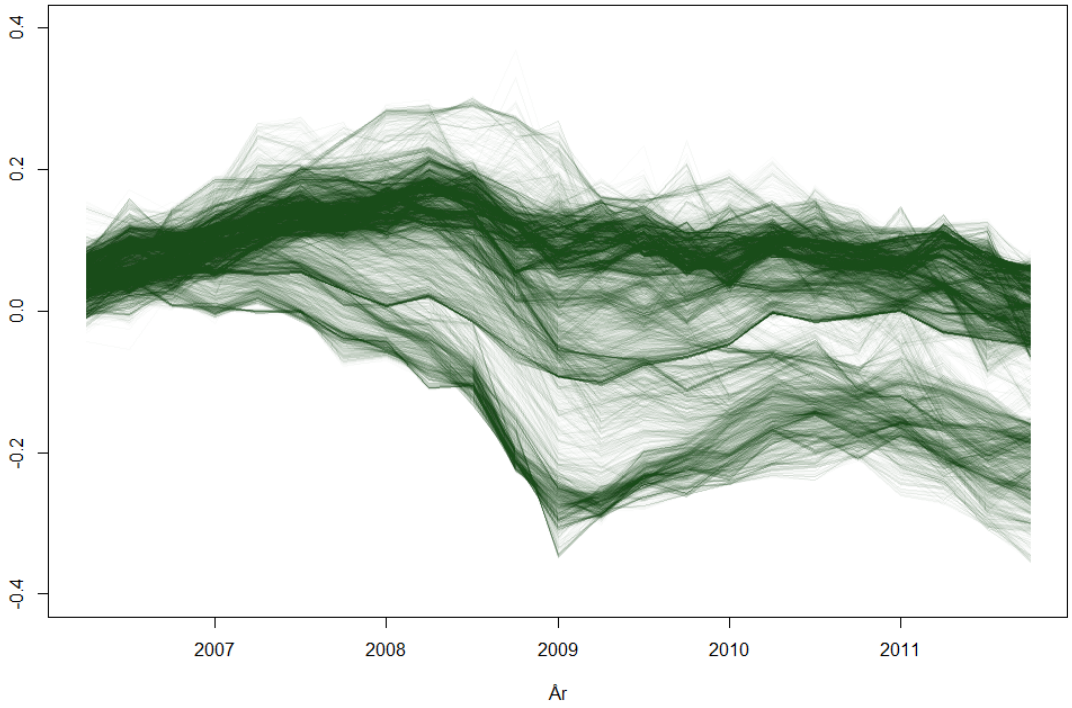
8. Resultater

For samtlige parcelhushandler i perioden 2006-2010 fremskrives boligprisen til hvert af de fire kvartaler i 2011 med udgangspunkt i GWR estimation, der for hvert geografisk område er baseret på det antal boliger, som de optimerede søgeområdeparametre giver anledning til. I perioden var der 38.459 parcelhussalg på Sjælland, 1.421 på Bornholm, 12.198 på Fyn og 70.517 i Jylland. Som i tuningsproceduren søges med undtagelse af Bornholm naboer blandt alle parcelhushandler i perioden 2006-2011. Da der alene laves fremskrivninger for boliger handler i 2006-2010, vil der ikke forekomme fremskrivninger baseret på salg i selve fremskrivningsåret. Der findes dermed ikke et skøn for prisudviklingen i 2.-4. kvartal i 2011 for en bolig handlet i 1. kvartal.

GWR-estimation giver anledning til meget output. For hver af de over 115 tusinde boliger der fremskrives, have estimerede koefficienter, standardafvigelser, t-værdier, RMSE, pm20, antal naboer osv. (se afsnit 5). Der gives nogle eksempler i dette afsnit.

Som nævnt i afsnit 2, er der interesse for at kunne danne et prisindeks for den enkelte bolig, således boligens pris kan frem- eller tilbageskrives til et hvilken som helst tidspunkt indenfor dataperioden. Med den nuværende GWR-metoden er dette mulig på kvartaltsniveau. Der indgår kvartalsdummies i estimationen. Koefficienterne til disse, er for den enkelte bolig et skøn over, hvad den lokale udvikling i kvadratmeterprisen har været, lige præcis på den adresse. I Figur 5 ses udviklingen i disse dummies for 10.000 boliger der er tilfældigt udvalgt af de over 115 tusinde boliger. Koefficienten til 1. kvartal 2006 er sat til 0 for alle boliger. Figuren starter 2. kvartal 2006 og slutter 4 kvartal 2011. Der ses en uhørt nuanceret beskrivelse af udviklingen i boligpriserne. En hovedstrøm udviser stigende boligpriser frem til starten af 2008, hvorefter der sker en lille korrektion nedad, fulgt at et svagt faldende forløb fra starten af 2009 til og med 2011. En større bi-strøm oplever faldende boligpriser allerede fra midten af 2007, og oplever ved årsskiftet 2008/2009 et fald på over 30 procent relativt til 1. kvartal 2006. Flere mindre bi-strømme kan identificeres.

Figur 5 Prisindeks. 10.000 tilfældigt valgte boliger



Kilde: Egne beregninger

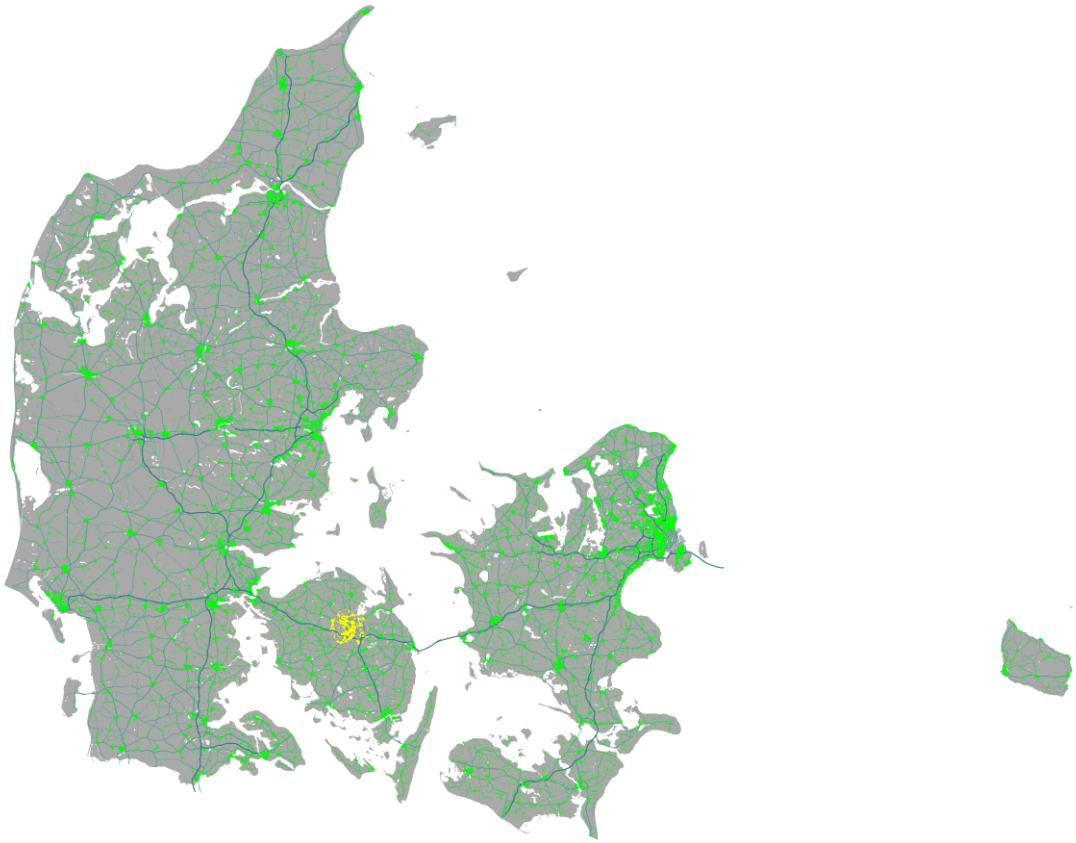
Der haves for alle boliger informationer om eventuel endogent bestemt naboområde. Hvis radius er givet, er antallet af naboer endogen; og omvendt. Hvis man vil have overblik over dette, er en grafisk analyse nødvendig.

Antallet af naboer, n, i GWR estimationen er for Sjælland, Bornholm og Jylland konstant for samtlige salg, hvorfor kMin = kMax = n. For salg på Fyn er antallet af naboer i estimationen nedadtil begrænset til 3.000 og vil grundet en begrænsning af den maksimale søgeradius på 10 km sjældent være højere.

I Figur 6 er antallet af naboer anvendt i estimationen for samtlige salg illustreret. Per definition er der kun mulighed for variation på Fyn, hvor det alene er i Odense og omegn, at antallet af naboer overstiger mindstegrænsen på 3.000. Ud af de 11.545 forskellige parcelhuse, der er solgt på Fyn i estimationsperioden er omkring 1/3, præcist 3.007 estimeret ud fra mere end 3.000 naboer. Således er

det altså kun for en i en relativt lille andel af de samlede parcelhussalg, at der ikke er anvendt et konstant antal naboer.

Figur 6 Valg af antal naboer ved estimation baseret på handler i perioden 2006-2010



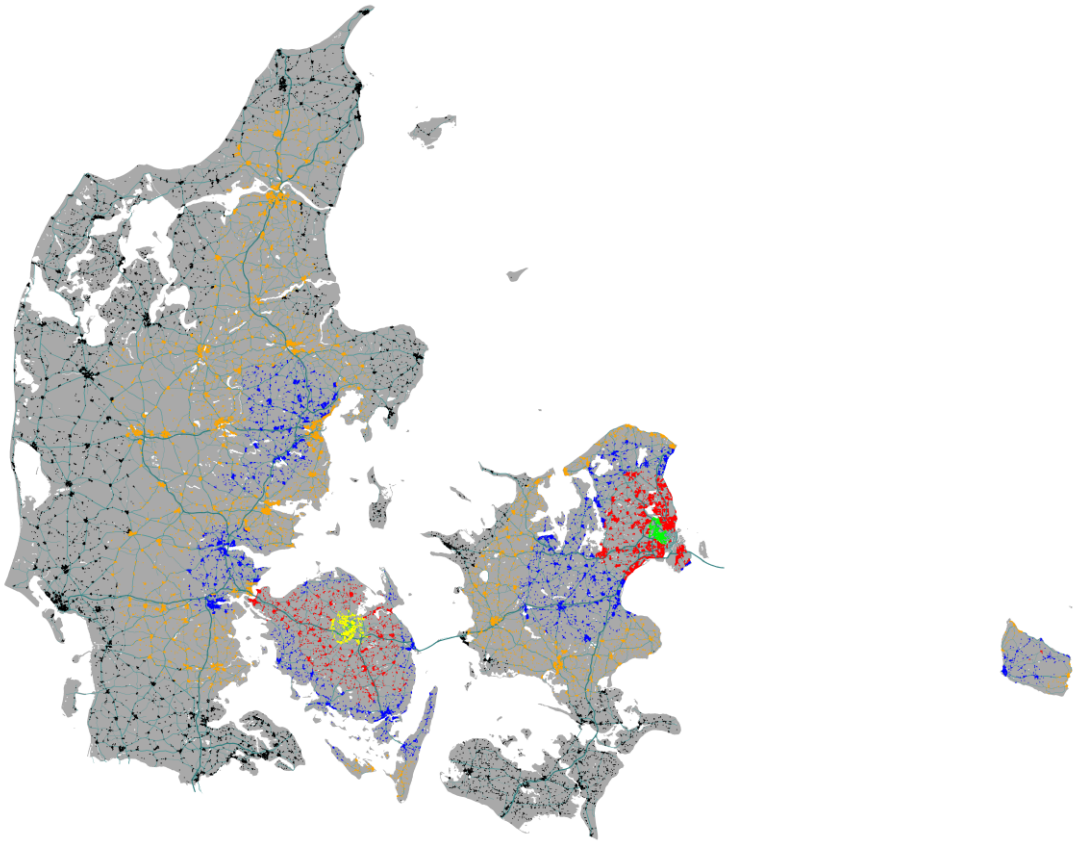
Anm: Grøn: antallet af naboer, n = kMin, Gul: kMin < n < kMax, Rød: n = kMax. For ejendomme handlet mere end én gang illustreres antallet af naboejendomme kun fra det nyeste salg.

Kilde: Egne beregninger

Anvendelsen af et fast antal naboer betyder, at der ikke er nogen geografisk begrænsning på søgeområdet. Søgeradius kan dermed variere betydeligt, hvilket kan ses af Figur 7. I Hovedstadsområdet og på Midtfyn findes nabosalg typisk inden for en radius af 20 km, mens søgeområdet gradvist øges ved stigende afstand til storbyerne. Ved valg af et fast antal naboer vil søgeområdet per definition øges i områder med relativt få salg. I den nordligste del af Jylland, Vestjylland, Sydjylland og Lolland-Falster søges der således typisk naboer inden for en afstand på mere end 40 km. En udvidelse af søgeradius motiveres også i

områder, hvor parcelhuse ikke er den hyppigste boligform. Omkring Århus og Ålborg ses derfor en søgeradius på over 20 km, hvilket selvfølgelig også er foranlediget af, at antallet af naboer ikke nuanceres inden for de fire geografiske områder.

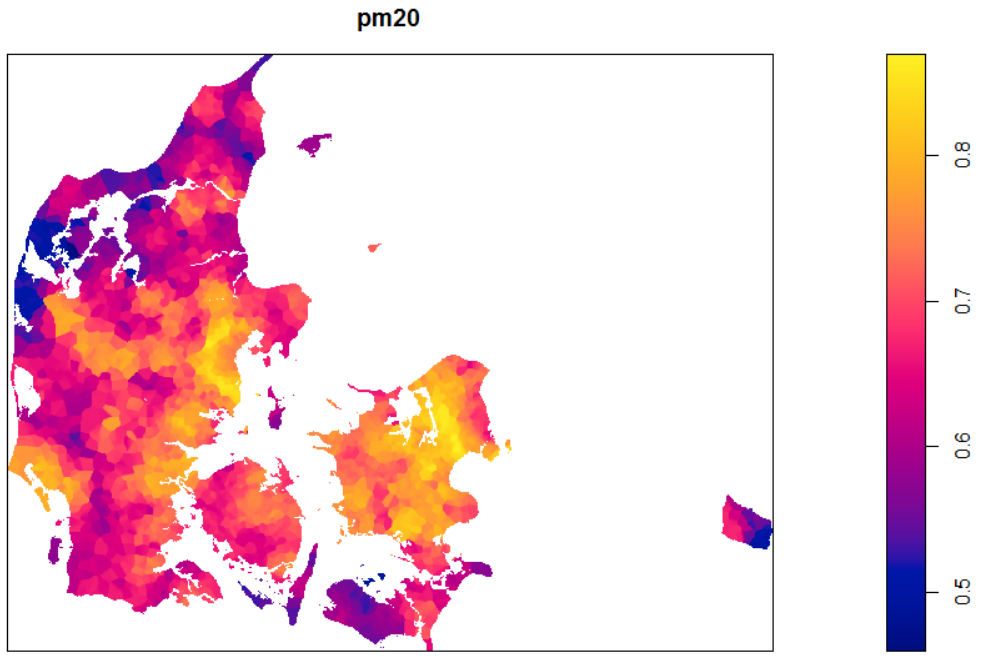
Figur 7 Søgeradius ved valg af antal naboer ved estimation baseret på handler i perioden 2006-2010



Anm.: Grøn: Søgeradius, bw < 10 km, Gul: bw = 10 km, Rød: 10 km < bw <= 20 km, Blå: 20 km < bw <= 30 km, Orange: 30 km < bw <= 40 km, Sort: bw > 40 km. For ejendomme handlet mere end én gang illustreres antallet af naboejendomme kun fra det nyeste salg.

Kilde: Egne beregninger

Figure 8 Lokale pm20-værdier fra GWR-estimation

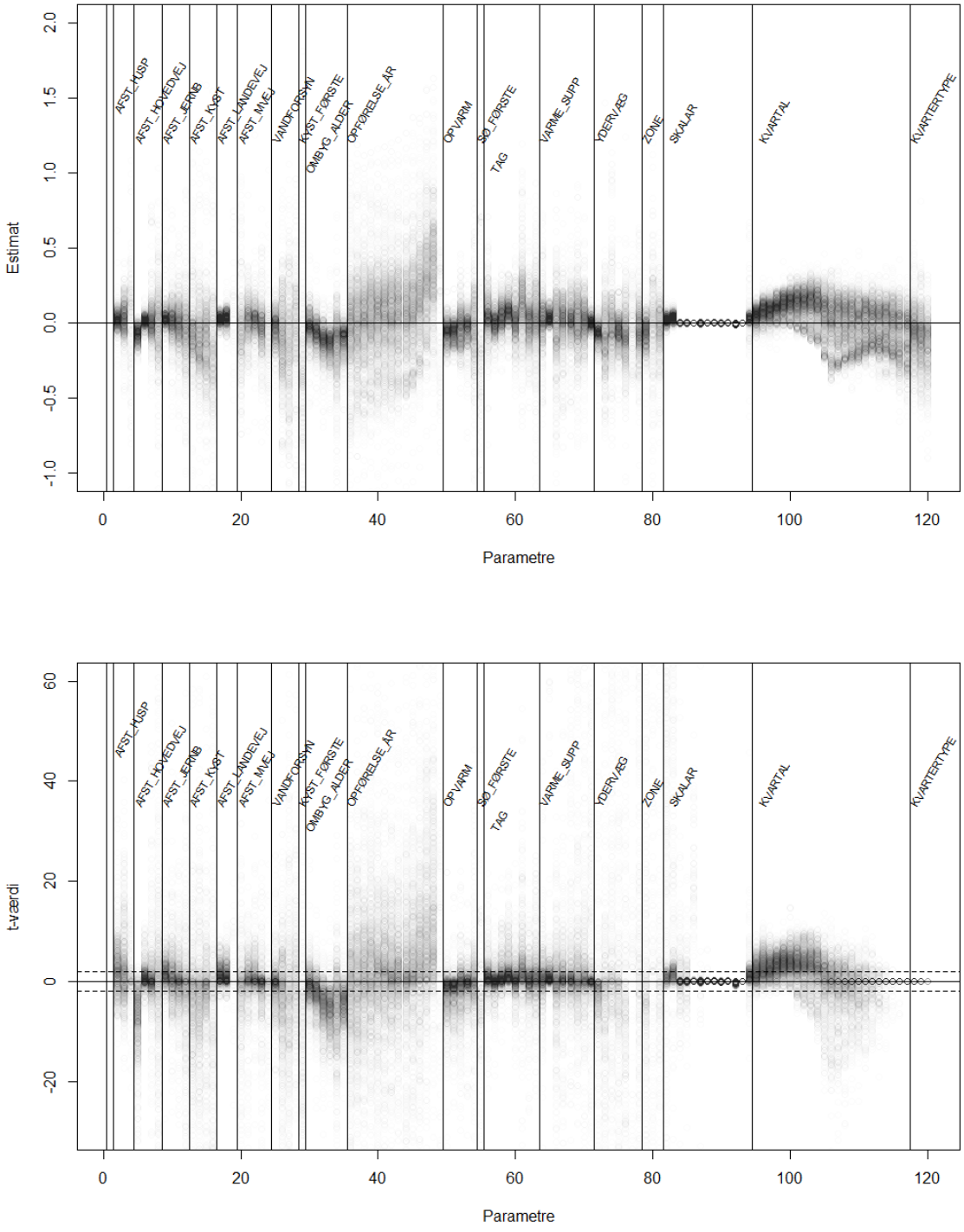


Kilde: Egne beregninger

Kvaliteten af estimationen, der for hver bolig udføres med data for hele perioden 2006-2011, måles med RMSE og pm20. I Figure 8 ses en grafisk beskrivelse af variationen i pm20. Man genfinder til dels billedet fra Figur 7. De vanskeligste områder er Lolland og nord-vest-Jylland (Figuren er lavet ved at smooth'e pm20-værdierne med en bandwidth på 500 m).

Vi har 30 variabel i den lineære estimation der indgår i GWR-metoden. Omregnet til binære dummy-variable svarer det til 120 variable. Vi har med andre ord over 115 tusinde estimater af 120 parametre. I Figure 9 ses et forsøg på at visualisere disse (baseret på en stikprøve på 1.000 boliger). I den øverste figur ses koefficienterne. Hvis vi fx betragter 'OPFØRELSE_ÅR' er dette en dummy-variabel der et opdelt i års-intervaller. To ting kan man se ud af figuren: der er en tendens til at nyere huse er dyrere, og der er betydelig spredning på estimaterne. Længere ude til højre ses 'KVARTAL'. Her genfindes billedet fra Figur 5.

Figure 9 Estimerede koefficienter (øverste) og t-værdier



Anm: t-værdi er stadig eksperimentel.

Kilde: Egne beregninger

I den nederste figur er koefficienterne delt igennem med deres standardafvigelser, således at vi får beregnet t-værdier. De stiplede linjer markerer -2 og 2. Koefficienter der ligger udenfor dette bånd er signifikant forskellige fra 0. Det ses at ombygningsalder og kvartal er signifikante, men at kvartertype ikke ser ud til at være det. Det skal understreges at beregningen af t-værdier stadig er på det eksperimentale stadie.

9. Konklusion og udviklingspotentialer

GWR-metoden er en moderne ikke-parametrisk metode. Sådanne metoder har typisk et betydeligt potentiale for 'tuning'; dvs. optimering af de parametre og variable der indgår i metoden.

Et godt sted at starte er sandsynligvis at kikke på de variable der indgår i estimationen. Her gør flere forhold sig gældende. Der er sandsynligvis en sammenhæng mellem det relativt store antal variable (30 variable svarende til 120 binære variabel) og de relativt store naboområder som viste sig optimale. Valgte men en mindre, men mere 'skarp' model, ville man sandsynligvis opnå mindre naboområder, og derfor en bedre beskrivelse af lokale informationer. Resultatet ville være bedre fremskrivninger. Den traditionelle måde at gøre dette på er at forsøge sig med at udelade 1 variabel af gangen, og måle effekten af dette. Problemet med denne fremgangsmåde er, dels at den er tidskrævende og dels (mere alvorligt), at det sikkert er forskellige variabel-sæt der er relevante forskellige steder i landet. En løsning på dette er at anvende en metode der automatisk vælger relevante variable. *Lasso* (Tibshirani 1996; Hastie, and Friedman 2009) er et eksempel på en sådan metode. Dette er en moderne udgave af ridge-regression der i den lineære regression automatisk sætter mindre vigtige variables koefficienter til 0. Metoden er meget udbredt og ekstrem effektiv. Indenfor biostatistik bruges den fx til modeller hvor antallet af variable er betydeligt højere end antallet af observationer. Metoden har været anvendt til GWR i fx Wheeler (2009).

I de her beskrevne fremskrivninger blev Danmark opdelt i Sjælland, Fyn, Jylland og Bornholm. Baggrunden for dette valg af områder var at undgå *syninger*. Hvis vi fx havde opdelt Jylland i nord og syd ville der opstå en linje henover Jylland. Boliger nord for linjen ville have andre parametre i deres GWR end boliger lige syd for linjen. Dette er en syning. Vi havde ikke indenfor projektets tidshorisont mulighed for at test om syninger er et problem. Det potentielle problem kunne være at boliger tæt på hinanden, med nogenlunde ens karakteristika, fremskrives forskelligt og har forskellige prisindeks. På den anden side er naboområderne så store at problemet muligvis er ret begrænset. Dette bør testes, idet muligheden for syninger (opdeling af fx Jylland og Sjælland) vil give nogle frihedsgrader til metoden.

Betragter man de 10.000 prisindeks i Figur 5 vil man observere at de fleste indeks ser relativt kontinuerte ud, men at et mindretal hopper op og ned. Det ville sandsynligvis være en god ide at 'smoothe' disse serier. En anden mulighed er at udskifte kvartalsdummis med en egentlig panel-modellering af prisudviklingen. Dette ville åbne mulighed for at modellere persistens i prisdynamikken, hvilket sandsynligvis er meget relevant. Gensalg ville være en naturlig del af dette approach. Det ville være relativt lige ud af landevejen at inkorporere R's panel-metoder i den eksisterende model. Der findes eksempler på en 'Geographically Weighted Panel Regression' i fx Yu (2010).

Referencer

S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman and A. Wu (1998) An optimal algorithm for approximate nearest neighbor searching, Journal of the ACM, 45(6):891-923, 1998.

R. Bivand, D. Yu, T. Nakaya, M.A. Garcia-Lopez (2013). spgwr: Functions for computing Geographically Weighted Regressions. R package version 0.6-22, URL http://CRAN.R-project.org/package=spgwr.

W. S. Cleveland, E. Grosse and W. M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

A.S. Fotheringham, C. Brunsdon, M. Charlton (2002). Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. John Wiley & Sons, Chichester.

I. Gollini, B. Lu, M Charlton, C. Brunsdon, P. Harris (2015). GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. Journal of Statistical Software, 63(17), 1-50.

URL http://www.jstatsoft.org/v63/i17/.

T. Hastie, R. Tibshirani, and J. Friedman (2009) The Elements of Statistical Learning. Springer, New York.

G. Jefferis (2015), RANN: Fast Nearest Neighbour Search (Wraps Arya and Mount's ANN Library). http://cran.r-project.org/web/packages/RANN/RANN.pdf

Q. Li and J. S. Racine (2007), Nonparametric Econometrics: Theory and Practice, Princeton University Press

Skatteministeriet (2014), Forbedring af ejendomsvurderingen. Resultater og anbefalinger fra regeringens eksterne ekspertudvalg

(http://www.skm.dk/media/1106957/Forbedring-af-ejendomsvurderingen_web.pdf)

R. Tibshirani (1996), Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B, 58, 267–288.

D. C. Wheeler (2009), Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environment and Planning A 2009, volume 41, pages 722 - 742

D. Yu (2010), Exploring spatiotemporally varying regressed relationships: the geographically weighted panel regression analysis. E. Guilbert, B. Lees, & Y. Leung (eds.). The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 38, pp. 134-139.