Interpretability vs. out-of-sample prediction performance in spatial hedonic models

Bjarke Christensen, Sydbank
Tony Vittrup Sørensen, Jyske Bank

Outline

1. Spatial autocorrelation in hedonic house price models
2. The data
3. Related literature
4. Methods
5. Results
6. Concluding remarks

Motivation

- Two uses for hedonic house price models:
  - Forecasting future sales prices (AVM, mass appraisal).
  - Estimating implicit value of externalities/amenities

- Spatial autocorrelation: a curse, and a blessing?
  - Causes biased and inconsistent parameter estimates
  - Can be used to improve forecast precision

- Methods for 'dealing with' spatial autocorrelation include
  - Fixed effects / GIS variables
  - Spatial econometric models
  - Semiparametric / spline based models an alternative?

(von Grewnitz and Panduro (LE, 2014)).

Motivation

- Does the objective dictate the approach?
  - For forecasting, interpretability is desirable but less important than accuracy
  - For externalities, interpretability is paramount

- The spatial variable of interest in this presentation is: proximity to wind turbines.

Data

- Western Jutland, 8 municipalities, 7.000 km^2.
- Period: 2002-2014
- Number of single family property sales:
  - In-sample: 21.066
  - Out-of-sample: 2.340
- Control variables: living area, number of rooms, age, number of bathrooms, detached vs. semi-detached, roof material, exterior wall material, proximity to large road.

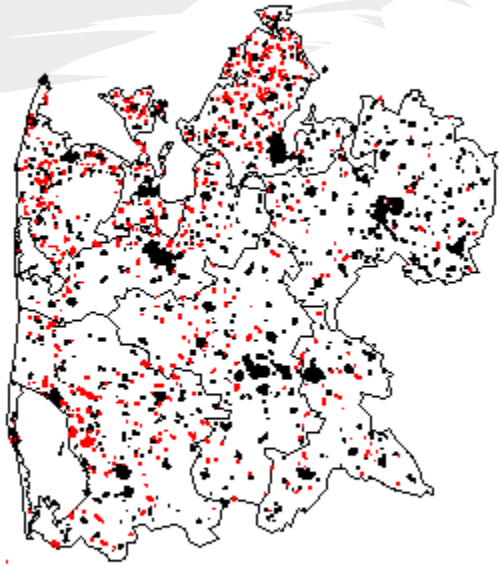
Spatial variable of interest: wind turbines

- Provided 30% of electric power consumption in Denmark in 2012.
- Expected to increase to 50% in 2020.
- Neighbours legally entitled to compensation for ‘loss of property value’.

	Number of turbines
In operation throughout the sample period	683
Commissioned during the sample period	279
Decommissioned during the sample period	661
Total	1615

Spatial distribution of sales and turbines

Distance to turbine (m)	Count
(0,500]	243
(500,1000]	1454
(1000,1500]	2440
(1500,2500]	7400
(2500,Inf]	11869



Measuring the impact of wind turbines on property values

Study	Temporal variation in spatial externatlity	Models	Impact of turbines
Sims and Dent (2008)	no	OLS	no
Jensen et al. (2008)	yes	SEM, SARAR	negative
Lang and Opaluch (2013)	yes	OLS	no
Hoen et al. (2011)	yes	Fixed effects OLS	no
Hoen et al. (2013)	yes	SEM	no
Heinzelman and Tuttle (2011)	yes	Fixed effects OLS	negative
Sunak and Madlener (2012)	yes	Geographically weighted regression	negative

Methods

- Ordinary least squares

$$y = X\beta + Z\gamma + \varepsilon$$

- Spatial Error Model

Anselin. (1988).

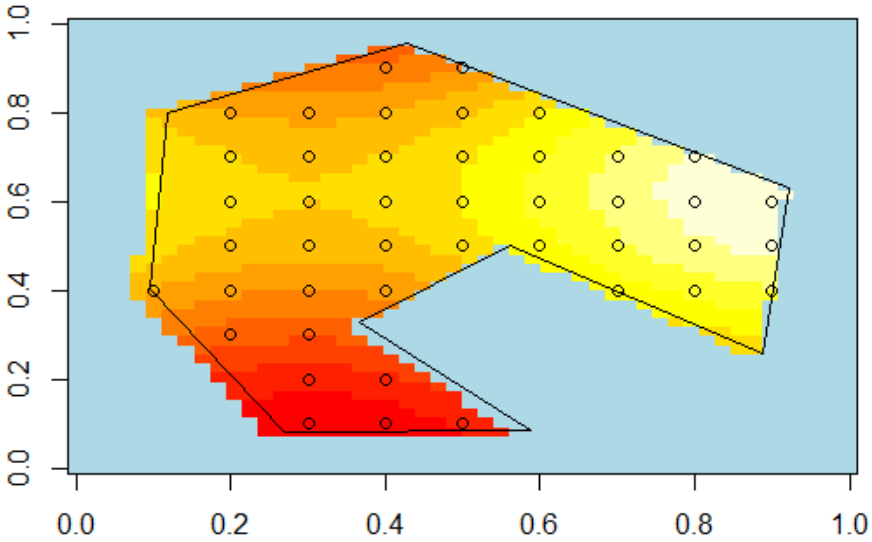
$$y = X\beta + \lambda Wy + \varepsilon$$

- Generalized additive model

Wood et al. (JRSS, 2008).

$$y = X\beta + f(x, y) + \varepsilon$$

- Interpretability



Estimation strategies

- Hedonic models with a small set of controls, varying only the spatial approach
- The proximity of a windmill is estimated as 1.5 km less the distance to the nearest active windmill
- The OLS model serves as a baseline
- SEM models
  - Row-standardized, k-nearest neighbours spatial weight matrix
  - Number of neighbours ranges from 5 to 50
  - Fitting by Maximum Likelihood
- GAM models
  - Maximum number of knots ranges from 50 to 2000
  - Penalized least squares, penalty determined by generalized cross validation
- Compare estimates to
  - Out-of-sample prediction errors
    - Median absolute percentage error
  - Tests for spatial autocorrelation
    - Moran's I, Perturbed Moran's I, Geary's C, Perturbed Geary's C

Results

Model	Turbine.prox	P-value	Moran's I	MAE out of sample
OLS	-0,296	<1%	0,49-0,57	21,19%
Fixed effects OLS	-0,086	<1%	0,06-0,14	16,39%
SEM.5	-0,103	<1%	0,02-0,08	16,10%
SEM.50	-0,039	<1%	0,003-0,10	15,74%
GAM.50	-0,154	<1%	0,21-0,31	15,77%
GAM.750	0,016	26%	0,01-0,07	13,83%
GAM.2000	0,013	47%	0,00-0,03	13,85%

Additional results - SEM models

	SEM.5	SEM.50
Shortest neighbour distance	4m	4m
Longest neighbour distance	6.500m	12.100m
Lambda	0,73656***	0,93578***

Additional results - GAM models

	GAM.50	GAM.750	GAM.2000
Area per knot	140 km2	9,3 km2	3,6 km2
Adj. R2	69,4%	77,5%	79%

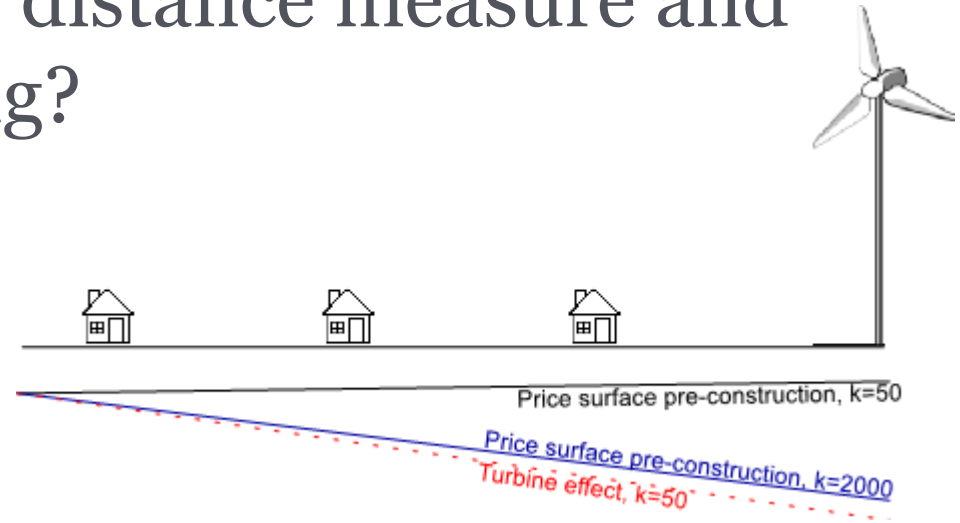
```
Knots Resid. Df Resid. Dev     Df Deviance       F    Pr(>F)
50        20948     1566.9
250       20768     1227.5 179.99   339.43 36.7015 < 2.2e-16 ***
500       20594     1165.2 174.41    62.24  6.9450 < 2.2e-16 ***
750       20387     1118.1 206.43    47.11  4.4412 < 2.2e-16 ***
1000      20215     1085.3 172.22    32.85  3.7118 < 2.2e-16 ***
1500      20089     1051.8 126.27    33.42  5.1510 < 2.2e-16 ***
2000      19956     1025.4 132.50    26.45  3.8850 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Additional results - Spatial autocorrelation tests

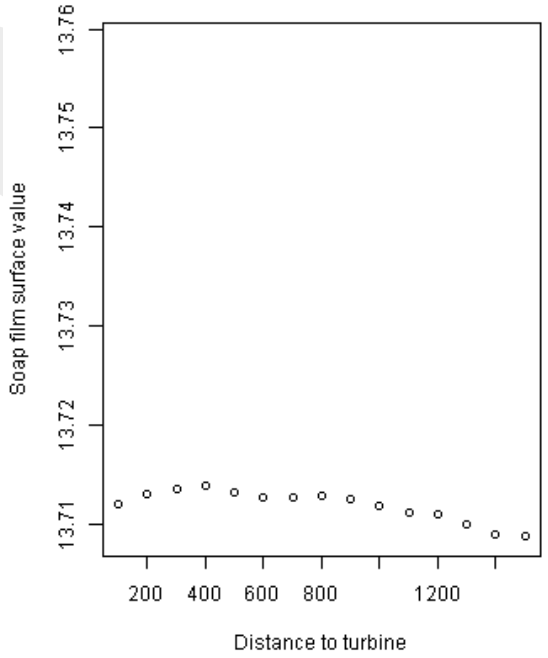
Model	Moran's I.mc	Geary's C	Geary's C.mc
SEM.5	0,02-0,08	0,91-1,01	0,90-1,01
SEM.50	0,003-0,09	0,89-0,97	0,88-0,97
GAM.50	0,21-0,31	0,68-0,76	0,67-0,76
GAM.750	0,01-0,07	0,92-0,98	0,91-0,97
GAM.2000	0,00-0,03	0,96-1,00	0,96-1,00

Where did the externality go?

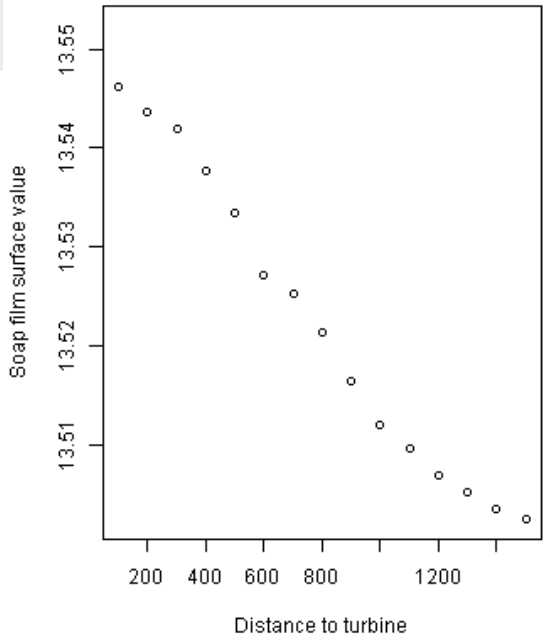
- Omitted variable bias / inconsistent estimates?
- Concurvity between distance measure and the spatial modelling?



The soap-film surface: model comparison



k=50



k=2000