Regional Science and Urban Economics 77 (2019) 239–252

Contents lists available at ScienceDirect

Regional Science and Urban Economics

journal homepage: www.elsevier.com/locate/regec

Learning from man or machine: Spatial fixed effects in urban econometrics

Åvald Sommervoll [a], Dag Einar Sommervoll [b,*]

[a] Department of Informatics, University of Oslo, Norway
[b] School of Economics and Business, Norwegian University of Life Sciences, Norway

ARTICLE INFO

JEL classification:
R31
R21
C45

Keywords:
Spatial fixed effects
House price prediction
Machine learning
Genetic algorithm
Spatial aggregation

ABSTRACT

Econometric models with spatial fixed effects (FE) require some kind of spatial aggregation. This aggregation may be based on postcode, school district, county or some other spatial subdivision. Common sense would suggest that the less aggregated, the better inasmuch as aggregation over larger areas tends to gloss over systematic spatial variation. On the other hand, low spatial aggregation results in thin data sets and potentially noisy spatial fixed effects. We show, however, how this trade-off can be substantially lessened if we allow for more flexible aggregations. The key insight is that if we aggregate over areas with similar location premiums, we obtain robust location premiums without glossing over too much of the spatial variation. We use machine learning in the form of a genetic algorithm to identify areas with similar location premiums. The best aggregations found by the genetic algorithm outperform a conventional FE by postcode, even with an order of magnitude fewer spatial controls. This opens the door for spatially sparse FEs, if economy in the number of variables is important. The major takeaway, however, is that the genetic algorithm can find spatial aggregations that are both refined and robust, and thus significantly, lessen the trade-off between robust and refined location premium estimates.

1. Introduction

Many economic variables have a considerable and systematic spatial variation. Even if the spatial variation is not of interest in itself, failure to control for such variation may bias the econometric analysis. Any model with spatial fixed effects (FE) relies on some spatial aggregation, say by block, block group, census tract, postcode, school district, county, city, or state. While all such aggregation may give rise to econometric models with spatial fixed effects, they will vary with respect to their ability to control for location. A challenge is that low spatial aggregation tends to give few observations for estimation of the location premium and therefore potentially noisy estimates. On the other hand, a high spatial aggregation may give robust estimates of the average location premium, but at the same time gloss over systematic spatial variation. This trade-off between potentially noisy spatially refined estimates and spatially crude but robust estimates may be lessened if we manage to aggregate over areas with similar location premiums. This is our point of departure.

The main ingredient in our approach is a genetic algorithm which relies on random variation and non-random selection in the search for larger areas with similar location premiums. Our illustrative example is house price prediction by a hedonic regression model with spatial FE. The data set we consider comprises all arm's-length transactions of apartments in the metropolitan area of Oslo in the years 2014–2015.

The key point is that the flexible aggregation allows us to find areas potentially that are spatially far apart, but have the same location premium. We may think of such areas as belonging to the same submarket.[1]

An identification of submarkets defined by a similar location premium comes with a double dividend for FE effects models. Most important from an econometric point of view, are the more robust estimates of the average location premium effect, as we get a higher signal to

* Corresponding author.
E-mail addresses: aavalds@ifi.uio.no (Å. Sommervoll), dag.einar.sommervoll@nmbu.no (D.E. Sommervoll).

[1] For our analysis it is not crucial whether or not these areas are part of the same submarket in the sense that agents consider these location as close substitutes. It is interesting to note that the high end of the housing market consists of pockets in the center, north, west, east and south of Oslo. These may or may not share key qualities like proximity to the fjord or a fjord view. Moreover, the distance to low-end neighborhoods may be short, and tends not to be determined by administrative boundaries. Although it is not our main point, our analysis contributes to the growing literature on non-spatially connected housing submarkets ((Pryce, 2013; Randolph and Tice, 2013)).

https://doi.org/10.1016/j.regsciurbeco.2019.04.005
Received 19 August 2018; Received in revised form 20 April 2019; Accepted 25 April 2019
Available online 15 May 2019

Table 1
Summary statistics of transactions of apartments in Oslo 2014–2015.

Statistic	N	Mean	St. Dev.	Min	Max
Value	14,036	3.86	1.63	1.33	11.50
LivingArea	14,036	69.31	28.69	24	271
Floor	14,036	3.03	1.70	1	13
BuildYear	14,036	1957	42.61	1800	2015

noise ratio. Moreover, as we are able to aggregate into larger submarkets with similar location premiums, we need fewer spatial controls. This opens for models that are more refined along other dimensions if data constraints call for economy in variable selection.

We consider two different types of aggregations. One is aggregations over postcodes, giving us aggregations in respect of administrative boundaries. The other is a more flexible aggregation over rectangles defined by introducing a grid of a given dimension, for instance 10 by 10, dividing the metropolitan area of Oslo into 100 rectangles or cells.

We find that an aggregation of the 53 two digit Oslo postcodes into 12 submarkets can be done without significant loss of out-of-sample performance measured by R^2. The aggregations of rectangular cells with augmented flexibility outperform the 53 postcode models not only with respect to aggregation into 12 submarkets, but also as few as 4 submarkets.

The major takeaway from our analysis is that controlling for location is not synonymous with introducing a large number of spatial dummies. On the contrary, less may be more, in the sense that a careful aggregation (by, for instance, a genetic algorithm) may give a few spatial controls that correlate with good out-of-sample properties.

We are not the first to consider machine learning for the purpose of house price prediction (Caplin et al., 2008). Use spatio-temporal closeness of observations in a gradient machine learning algorithm. Earlier contributions regarding genetic algorithms and housing markets are (NgMartin and Wong, 2008) and (Shekarian and Fallahpour, 2013). Our machine learning approach resembles (Plakandaras et al., 2015), as the population of regression models evolves according to a fitness measure given by the models explanatory power. However, our approach is transversal in two ways. First, whereas (Plakandaras et al., 2015), take the spatial division as given, our models differ only according to spatial aggregation. Second (Plakandaras et al., 2015), are less concerned with the different models' out-of-sample properties. For us, however, this is the main point.

The remainder of the paper is organized as follows. Section 2 describes the data set and gives the baseline hedonic regressions, which have spatial dummies for Oslo's 12 urban areas, commonly known as Oslo 1 to Oslo 12. We also present two aggregations by postcode into 12 submarkets based on square meter prices. In section 3 we describe a genetic algorithm for aggregating postcodes into 12 submarkets. Section 4 considers spatial aggregation in Oslo by rectangular cells. First, we vary cell size, but keep the aggregation into 12 submarkets. Second, we allow the number of submarkets to vary, and explore how out-of-sample performance varies with the number of submarkets. Section 5 concludes. Details regarding data set preparation and supplementary tables and figures are found in the appendix.

2. Data description and baseline hedonic regression models

We consider all arm's length market transactions of apartments in Oslo 2014 and 2015. The data set consists of 14,036 observations. Table 1 gives summary statistics of variables used in the subsequent analysis. Details regarding data preparation are given in Table 7 in appendix.

This data set is divided into three: a training set (8,400 observations), a validation set (2,836) and a test set (2,800).

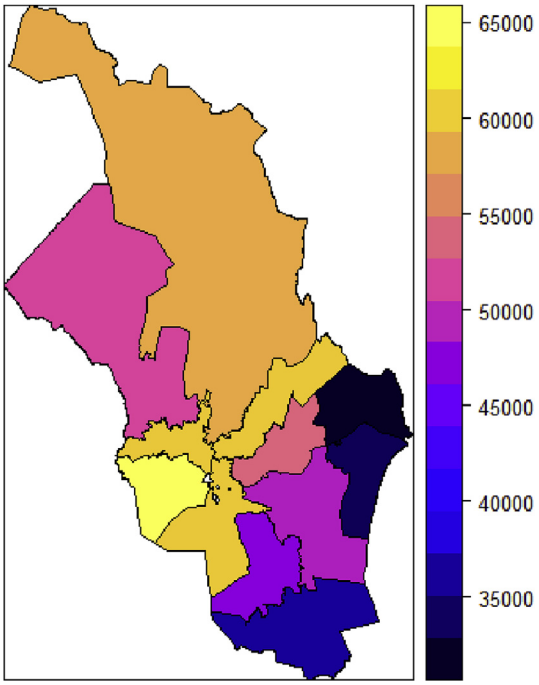


Fig. 1. Heat map of median square meter prices for Oslo 1 to Oslo 12.

A standard hedonic regression model for house prices is given by:

$$p_i = \alpha + \beta_{\text{logArea}} LogArea + \beta_{\text{logAge}} LogAge + \sum_{i=2}^{14} \theta_i Floor_i + \sum_{i=2}^{24} \nu_i Mth_i + \sum_{i=2}^{12} \gamma_i Sub_i + \epsilon, \quad (1)$$

where *LogArea* is log (Area in sqm.), *logAge* is the log (2017-construction year), $Floor_i$, Mth_i, and Sub_i are dummy variables for time and spatial FE (month and spatial submarket).

The submarkets in this baseline model are defined according to a commonly used division of the metropolitan area of Oslo, where the last two digits of the four digit postcodes are omitted.[2] This is our default division of Oslo into 12 submarkets as displayed in Fig. 1. In this paper we keep the number of spatial dummies fixed and equal to 12, when we compare aggregation by postcodes.[3]

Table 2 gives regression results for the baseline model used on the training and validation sets. Of particular interest is the explained price variation measured by R^2, 77.42 and 75.82. The former number is the in-sample R^2 (in percent) of the model estimated on the training set, and the latter is the out-of-sample R^2 on the validation set. The third column is the regression model without spatial controls. Note that the difference 77.42–65.70 = 11.72 is directly attributable as the added benefit of spatial FE (12 submarkets). A roughly 12 percent gain in explanatory power in a model that already has substantial explanatory power, tells us that these spatial submarkets indeed have different location premiums. Still there is reason to believe that these average location premiums mask systematic within submarket price variation. For this base line model there is a 1.6 percent (77.42–75.82) difference between the in-sample and out-of-sample R^2.[4]

[2] Postcode 1236, is in Oslo 12.

[3] This number is in some sense arbitrary (although it originates from Oslo's 12 areas defined by postcodes). In Section 4 where we consider aggregation by rectangular cells, we also vary the number of submarkets.

[4] Machine learning on the training set is overt data mining, and overfitting is a prime concern. A sign of overfitting is when the better in-sample fit comes at the expense of the out-of-sample fit.

Table 2
Baseline regression.[a]

	Dependent variable:	
	Transaction price	
	Baseline	Baseline without Spatial FE
logLiving	3.19***	3.31***
	(0.02)	(0.03)
logAge	−0.16***	−0.04***
	(0.01)	(0.01)
R^2 in − sample (Training set)	77.42	65.70
R^2 out − of − sample (Validation set)	75.82	63.07

*p < 0.1; **p < 0.05; ***p < 0.01.
[a] Note: On the validation is the R^2 for the predicted values of the baseline model estimated on the training set. Number of observations in the training set (Validation set) 8400 (2386).

We proceed to construct partitions of Oslo into 12 submarkets, which are likely to correlate with high explanatory power when used to define spatial dummies in a hedonic regression model.

2.1. A human approach to alternative spatial aggregations

We intuitively understand that there are many ways to partition of the metropolitan area of Oslo. At the same time, our intuition fails us when it comes to comprehending the immensity of all possible aggregations. In our data set we have 385 distinct four digit postcodes.[5] The number of possible partitions is given by the Bell number $B_{385} \approx 1.0 \cdot 10^{226}$.[6] In comparison the number of elementary particles in the universe is believed to be around 10^{80}. Most of these aggregations will naturally be poor candidates for submarkets for house price prediction purposes.

In the following we limit the number of aggregations to allow us in principle to pit models against each other. There are essentially two ways to do this. One is to impose some a priori constraints. The other is to have some data-driven aggregation rules. We will do both. A natural a priori constraint is to let the spatial building blocks be larger, reducing the ways to combine them. The Oslo postcodes are constructed in such a way that numerically close postcodes are also geographically close. This means that by skipping the last digit we create larger (spatially connected) building blocks for aggregation. In our data set there are 53 such three-digit postcodes. If we want to pit a given model against the benchmark model above, it makes sense to consider only aggregations into 12 submarkets. In this way, we compare models with the same number of explanatory variables, but where the spatial dummies differ in the underlying aggregation.

All of these aggregations will improve the explanatory power of the hedonic model without spatial controls ($R^2 = 65.70$). Table 3 gives summary statistics of 1000 random aggregations into 12 submarkets, and shows that on average a random aggregation gives an R^2 of 69.69. The span, however, is considerable ranging from 66.55 to 75.36. It is interesting to note that all models have significantly lower explanatory power than the baseline model, the reason being the systematic price variation across Oslo 1 to 12, and the probability that a random aggregation should be better is discouragingly low.

A data-driven approach is to aggregate by some observable summary statistics in a way that is likely to correlate with location premium. One such variable is average square meter price. Fig. 2 gives a heat chart of square meter prices for the three-digit postcodes for the training set. Note that this is (mild) data mining on the training set.

Table 3
Summary Statistics of 1000 random spatial aggregations into 12 submarkets.

Statistic	N	Mean	St. Dev.	Min	Max
R^2 in percent	1000	69.69	1.3	66.55	75.36

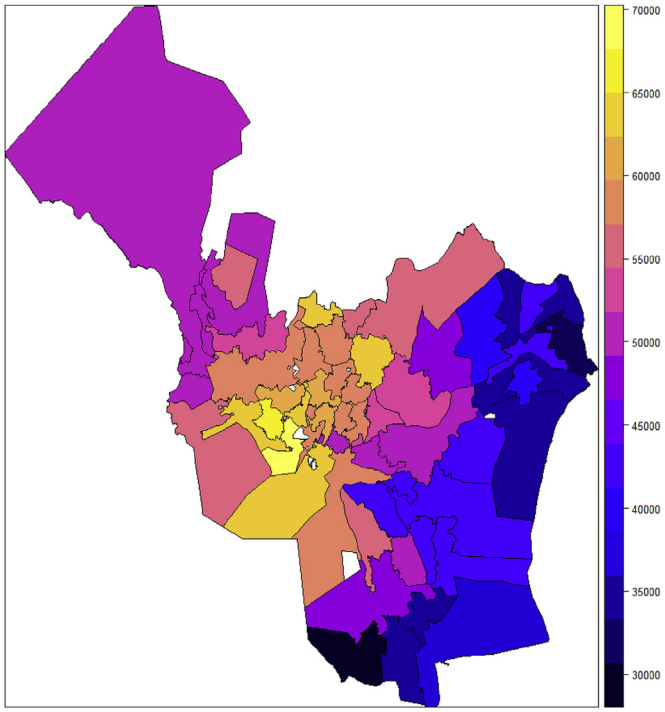


Fig. 2. Heat map of median square meter prices submarkets defined by three-digit postcodes. Postcodes with no transaction excluded (white).

It makes sense to assume that good candidates for spatial aggregation tend to aggregate postcodes with comparable square meter prices. If true, partitions that respect the ordering with respect to square meter prices is a good option.

Further natural limitations are to group the 53 postcodes into 12 groups (submarkets) of 4–5 postcodes. A more refined grouping is to choose the groups such that within group variance is minimized. In Table 4 we compare these two variants against the baseline (with and without spatial controls) and the full model controlling for all 53 postcodes.

The in-sample R^2 of these two models is virtually identical (78.73 and 78.74) and 1.3 percent higher than the baseline model with spatial controls. The aggregation with minimum variance within submarkets is slightly worse out-of-sample (77.24 versus 77.63). Since this aggregation not only uses the in-sample ranking, but also square meter

Table 4
Comparison of the R^2 in percent for selected spatial aggregations.

Model	in-sample	out-of-sample	Difference
Baseline without spatial controls	65.70	63.07	2.63
Baseline with spatial controls	77.42	75.98	1.44
Aggregation by m^2 (4–5 postcodes)	78.73	77.63	1.10
Aggregation by m^2 (Min. variance)	78.74	77.24	1.53
Model with 52 postcode dummies	80.27	79.24	0.91

[5] Note that subsamples like the training, validation and test set will in general have fewer distinct postcodes.
[6] The generating function for Bell numbers is given by $\sum_{i=0}^{\infty} B_n x^n = e^{e^x - 1}$ (Gian-Carlo Rota, 1964).

prices, this grouping involves slightly more data mining. This seems to translate into a somewhat lower out-of-sample performance.

To determine whether this is a substantial improvement the “full model” with 53 spatial controls is tabulated. This serves as an in-sample upper bound for aggregation improvement as any aggregation of the 53 postcodes will necessarily have a lower R^2. The in-sample R^2 of the model with 53 spatial controls is 80.27 percent. More interesting is the out-of-sample performance and the difference between in- and out-of-sample performance. The latter is a key statistic with respect to overfitting.

As we see, the difference between in-sample and out-of-sample fit is the lowest for the full model with 53 postcodes (0.91) and the highest for the base line with out spatial controls (2.63). The aggregation by square meter prices with 4–5 postcodes in each submarket is a good second (1.10).

In the next section we use a genetic algorithm (GA) to find spatial aggregations that give high in-sample R^2s. In light of the discussion above, the key question is to what extent high in-sample R^2s translate into high out-of-sample R^2s.

3. Three-digit postcode spatial aggregation and genetic algorithm

In this section we use a genetic algorithm (GA) to find spatial aggregations of the 53 three-digit postcodes into 12 submarkets. We aim to find aggregations that give high R^2's when used as spatial controls in regression model 1. Before we go into specifics regarding the genetic algorithm we use here, let us briefly discuss the mathematical intuition behind genetic algorithms. A search for maxima for a function (here R^2) tends to rely on some kind of gradient ascent.[7] That is, we evaluate the function to be maximized at two points, work out a proxy for the derivative and ”head uphill”. A genetic algorithm is a variant of gradient ascent. We can picture it as a herd of points corresponding to regression models, where the points highest up the hill, are used to create new models, by random variation. These replace the points/models with the lowest R^2s. The result is a new ”generation” of points/models which is further up the hill, and as generations pass, the herd of models/points moves to higher elevations.

This picture of a herd slowly moving up a hill, is a little misleading. The strength of a genetic algorithm is that the herd does not concentrate in one area. If it did, a more classical version of gradient ascent would be equivalent or better. The strong suit of the genetic algorithm is genetic diversity, which in our herd picture corresponds to a widely dispersed herd, which in principle glides up multiple hill sides, and by recombination can suddenly stumble on even higher hills and start climbing.

3.1. Genetic algorithm for spatial aggregation

A genetic algorithm mimics natural selection. The key is random variation and non-random selection. We consider a population of hedonic models that differ only in their spatial aggregation.

An aggregation of 53 postcodes to 12 submarkets is naturally represented by a 53-dimensional vector $(1, 12, 3, 3, \ldots)$, where a submarket is identified by an integer. We refer to this vector of integers as the genome or genotype of a given model.

Every generation consists of 50 models, and the first generation is 50 random draws of 12 submarkets. The fitness of a model is defined to be R^2 of the hedonic regression model (1) with the spatial controls defined by the models genotype (the 53-dimensional vector coding for the submarket aggregation). This means that the first generation average fitness is likely to be close to the average random fitness (69.69) given in 3. Any model is uniquely defined by it is spatial aggregation and easily visualized as a map. Fig. 3 is an example of such a model.

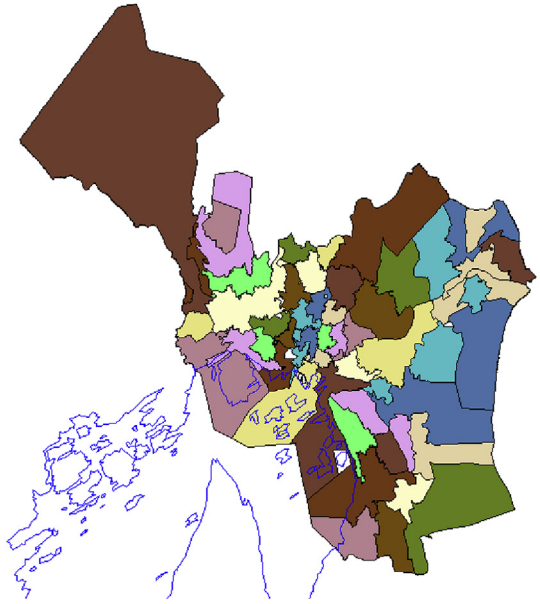


Fig. 3. Example of a model represented as a map with borders. Areas with the same colors aggregated to one submarket. Blue line defines the Oslo Fjord. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Specification of the GA.

population size (N)	Crossover	Mutations	Number of generations
50	Yes	3	3000

The next generation is created in the following way. The population is ranked according to R^2. The 24 highest ranked models are divided into two according to rank. Parent pairs are formed by pairing according to rank. That is, the highest ranked model is paired with the 13th rank (since it is the highest ranked in the second group), the second with the 14th, et cetera. Each parent pair gives rise to a one offspring. These 12 offspring replace the highest ranked models without offspring in this generation.[8]

The offspring is formed by genetic crossover. Let us illustrate genetic crossover by a genome only 6 integers long and only four groups:

Parent one: (1,2,1,3,3,4)
Parent two: (1,3,3,3,4,4)
Offspring: (1,2,1,3,4,4)

The offspring inherit the first three integers from Parent one and the last three from Parent two. It is customary to allow for mutations in order to preserve genetic diversity. A mutation tends to be just a random draw of a place in the genome, and a random replacement of the integer by another integer. In this example, say a random draw gave position 5, and group 1, then the resulting offspring would be:

Offspring: (1,2,1,3,1,4)

We have 53 different three-digit postcodes, so the genome does not allow for an even split of genetic inheritance between parents. We

[7] In the literature it is more common to use the notion of gradient decent, as the objective is usually to minimize a loss function (Marsland, 2009).

[8] As the 24 first models beget offspring, the offspring will replace ranks 25 to 36.

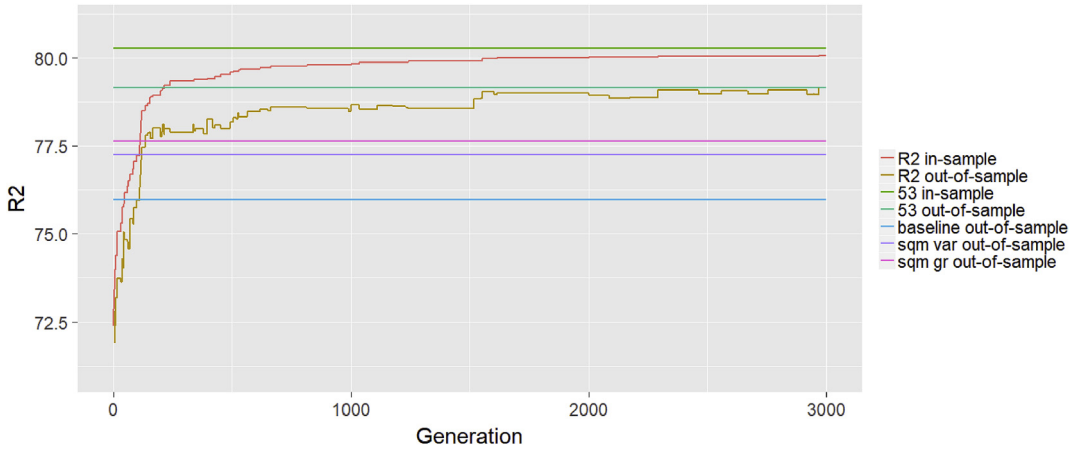


Fig. 4. The explanatory power (R^2) by generation number. The red curve is the highest R^2 in the population of models in the indicated generation. The yellow curve is the R^2 of the out-of-sample prediction on the validation set. The green (dark green) line is the R^2 of the full model with 53 submarkets in-sample (out-of-sample). The pink, violet and blue are the out-of-sample R^2 of the sq. meter 4–5 postcodes, m^2 min. variance, and the baseline with 12 submarkets, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

choose the first 26 elements of the DNA-strain from the most fit parent and 27 from the least fit parent. The offspring are also mutated on three randomly drawn places of the genome.[9]

Table 5 summarize the genetic algorithm.

The global in-sample R^2 maximum for aggregation into 12 submarkets is unknown, but it is lower by construction than the in-sample R^2 for the full model with 53 submarkets. More interesting, is it to compare out-of-sample R^2's of the best models from the GA with the baseline, the square meter models and the full model.

Fig. 4 displays a typical run of the GA presented in Table 5. We see that both the in-sample and the out-of-sample R^2, surpass the benchmark and models based on square meter prices after a few hundred generations. The fitness continues to improve for another 1,000 generations.

We also note that evolutionary pressure does indeed give models with high in-sample R^2s, and that these get close to the theoretical upper bound in-sample. More interestingly, however, the out-of-sample R^2s are even closer to full model with 53 submarkets out-of-sample (79.16 versus 79.24)[10]. In other words, going from 11 spatial dummies to 52 improves the R^2 by a mere 0.2 percent. Moreover, the GA out-of-sample performance outclasses the baseline model and the two models based on square meter prices.

The GA is overt data mining in-sample, and thus may suffer from overfitting. Fig. 4 shows no sign of overfitting. On the contrary, the difference between in- and out-of-sample performance is roughly constant at about 1 percent and thus in line with the in- and out-of-sample differences of the models presented in the previous section.[11]

4. Spatial aggregation based on grids and geocoordinates

A potential strength of the approach described in the preceding section is that the aggregation respects administrative boundaries. For example, it is well known that school districts ((Black, 1999)) can create sharp price boundaries, so that choosing spatial aggregations in hedonic regression models that respect administrative boundaries is likely to correlate with higher explanatory power ceteris paribus. At the same time, it is also a straightjacket that limits the possible spatial aggregations. We will now consider the more flexible approach and use a grid to partition Oslo into rectangular cells. As these cells can vary in size, we can address the question of optimal cell size for aggregation. This optimal cell size will likely vary with the number of submarkets. We will first keep the number of submarkets fixed at 12 in order to compare with the postcode aggregation of the preceding section. Subsequently, we will also explore the added benefit from a higher number of submarkets, asking to what extent we can choose a few submarkets without losing too much explanatory power.

Fig. 5 displays a 33×33 grid. The coloring indicates an aggregation of the 367 cells with housing market transactions into 12 submarkets.

4.1. Genetic algorithm

The basic construction of the GA resembles closely the approach described in section 3. The regression model considered is 1. The only thing that is different is the definition of the 12 submarkets. Consider an $n \times m$ grid. This grid partitions Oslo into $n \cdot m$ cells. An aggregation into 12 submarkets may be represented by assigning a number in $\{-1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ for each cell. Here -1 is included for cells with no housing market transactions.[12] Any such aggregation may be represented as a matrix, M, of size $n \times m$, where m_{ij} is the corresponding submarket number. We view this matrix as a model's genome in the same way that we represented the genome in the postcode GA as a 53 dimensional vector.

The population size, N, remains fixed from generation to generation. The next generation is created in the following way. The population is sorted according to its fitness measured by R^2. The upper third of the

[9] A GA tends not to be very sensitive to the details of recombination or mutation rates. In other words, we have some leeway in the choice of these parameters. The important thing is to strike a balance between R^2 reward and genetic diversity. The probability of getting stuck on some potentially low local maximum decreases with genetic diversity. One technical detail: We have assigned 12 submarkets centers, these are left untouched by mutations.

[10] The R^2_{adj} is equal to 78.48 and 78.79, respectively. In other words, by this measure the 12 submarkets found by the GA, outperform the full model of 53 postcodes out-of-sample.

[11] The best GA model arises from a herd of models that evolve by random variation and nonrandom selection. This means that two different runs will result in different models. These models tend to be both statistically and economically different in general, but the major differences are linked to postcodes with few transactions. A detailed discussion is found in the appendix.

[12] The actual grid is constructed by defining the longitude and latitude step size individually. In other words $lng\_step\_size = (max(longitude) - min(longitude))/n$ and $lat\_step\_size = (max(latitude) - min(latitude))/m$.

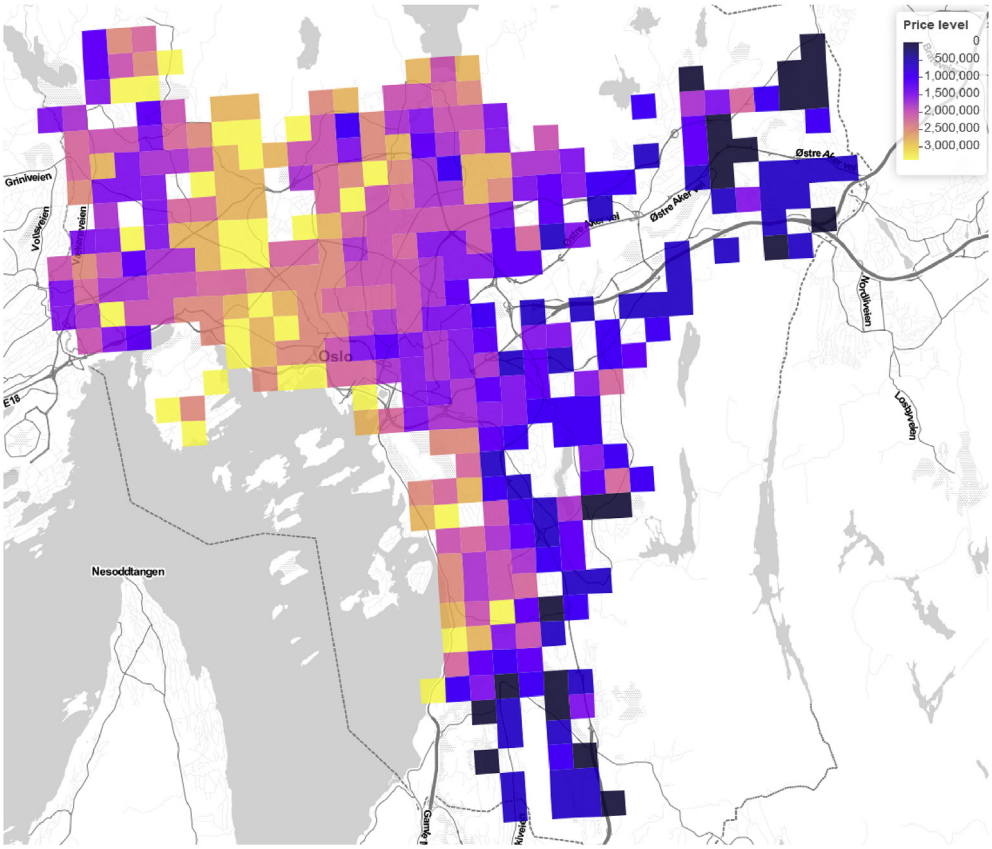


Fig. 5. Heat map of 12 submarket partition with in-sample $R^2 = 83.50$ arising from 33×33 grid of the metropolitan area of Oslo. Color defined by estimated submarket coefficient dummy. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

population is paired with the middle third according to fitness. Each pair gives rise to one offspring. The offspring and their parents constitute the next generation. In other words, the lower third ranked by fitness is replaced in every generation.[13]

The offspring are constructed by letting the six submarkets (aggregated submarkets) be inherited from one of the parents by random draw. These 6 submarkets may or may not be the lion's share of entries in the offspring's genome. In any case the entries of the offspring's genome matrix M corresponding to these six submarkets are inherited from this parent. The other matrix entries are inherited from the other parent.[14]

Mutations may occur in any element, m_{ij}, in the genome the $n \times m$ matrix. A mutation is a replacement of the integer m_{ij} by a random draw of a number in $\{0, 1, \ldots, 11\}$. For any element in the matrix, the mutation probability is 0.005. In other words, for a 10×10 genome the expected value of the probability of a mutation is 50 percent. The GA is summarized in Table 6.

Fig. 6 shows a typical run. We see that both the in-sample and out-of-sample R^2s climb above the 80 percent line after around 1000 generations. In this case we do not have a theoretical bound for the R^2, but if we view the grid approach as an alternative to aggregation over postcodes, we see that this GA surpasses even the theoretical bound for postcodes, not only in-sample, but also out-of-sample. In other words, the increased flexibility of the grid GA comes with the added potential of finding models with much higher out-of-sample explanatory power.

Table 6
Specification of the GA.

population size (N)	Crossover	Mutation probability	N. of generations
20	Yes	0.005	15,000

4.2. Grid size and in- and out-of-sample properties

The grid size defines the building blocks or cells that are used in the spatial aggregation. Too crude a grid may result in cells with a considerable spatial premium variation while too fine a grid may give noisy cell premiums.[15] The latter will manifest as a good in-sample fit at the expense of out-of-sample fit. The role of the validation set is to locate the point at which the in-sample fit comes at the expense of out-of-sample fit.

Since different runs tend to vary with respect to their in- and out-of-sample performance, we compare the distribution and averages of 10 runs. We consider only $n \times n$ grids. This will result in n^2 close to quadratic grid cells. Due to the clustered nature of housing market transactions the number of cells with housing market transactions does not grow quadratically with n.[16]

Fig. 7a is a notch plot showing that the in-sample fit increases up to grid size 70 to 80, whereas the out-of-sample fit levels off and falls

[13] Note that this implies that the evolutionary pressure is a bit higher in this GA-model compared to GA-postcode model ($\frac{N}{3}$ in contrast to $\frac{N}{4}$).

[14] The crossover algorithm allows for the contingency that one or more aggregation submarkets are lost, as it is no pre-assigned submarkets centers as in the postcode GA. That such models are going to dominate in the gene pool, is highly unlikely, as these models are likely to have lower fitness (R^2) compared to models with more submarkets.

[15] In statistical terms, this is the bias variance trade-off.

[16] The number of cells with housing transactions grows close to linearly with n. See appendix for details.

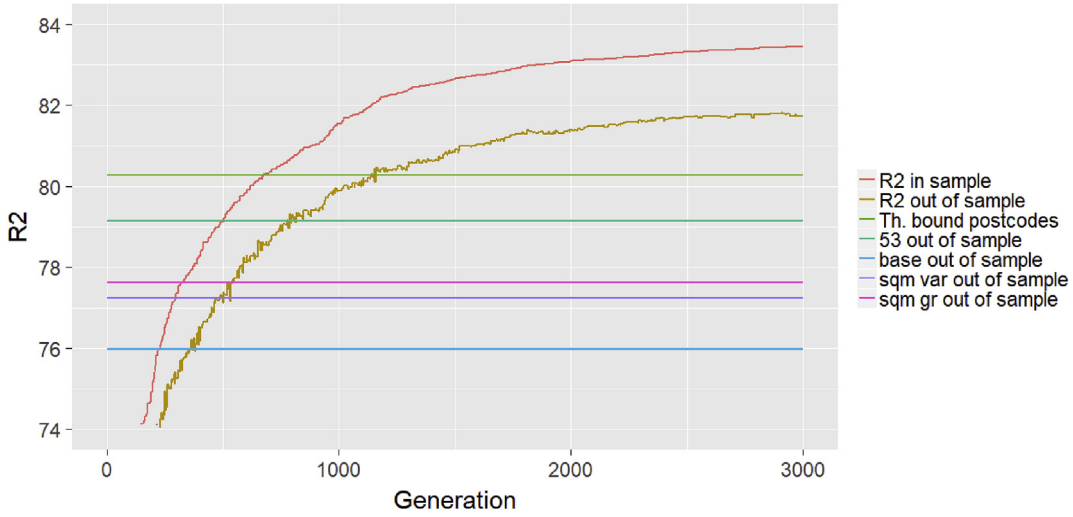
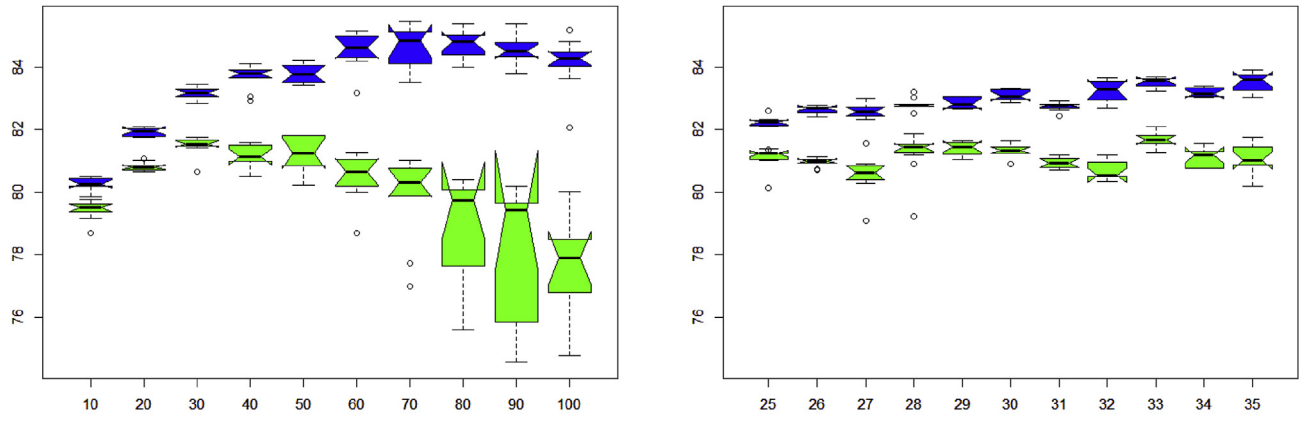


Fig. 6. The explanatory power (R^2 in percent) by generation number. A run of the 33 by 33 grid.



(a) Grid size 10 to 100 (b) Grid size 25 to 35

Fig. 7. Notch plot of R^2 for the best model training set (blue) and the best model used on the validation set (green). Note that the notch plot notches give the 95 confidence interval for the median and the colored box corresponds to the observations between the 25th and the 75th percentile. Outliers are represented by small circles. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

around 30. In other words, somewhere around $n = 30$, the in-sample improvement starts to come at the expense of out-of-sample performance. In order to pinpoint the grid size with the best out-of-sample fit, we run all grid sizes from 25 to 35. Fig. 7b displays the results. The figure shows a steady increase of in-sample fit with grid size. The out-of-sample R^2 (green notches), on the other hand has no obvious trend. Some grid sizes perform significantly worse than their nearest neighbors. The best is the 33×33 grid (with a mean of $R^2 = 81.86$) with the 29×29 a close second.[17]

The significant differences between close neighbors with respect to grid size is most likely driven by the clustered nature of housing market transactions. Some grids turn out to be unfortunate insofar as clusters are divided or joined in such a way that the in-sample fit is more likely to come at the expense of out-of-sample fit.[18]

At a higher level, fewer observations per cell make the actual location premium in the cell hard to assess both for man and machine. The 33×33 grid has 367 cells with housing market transactions giving on average 23 observations per cell. This number may be taken as a crude proxy for the number of observations per cell, and may give the best out-of-sample properties of the aggregation. This number, however, may be driven by the idiosyncratic spatial clustering of this particular housing market, and thus, at best serve as a natural starting point when choosing grid size.

[17] The comparison of out-of-sample performance, by estimating the models on the validation set, is data mining on the validation set. Thus we need a third data set to get a true out-of-sample estimate. We divided the original data set into three, and may use the test set for this purpose. The out-of-sample R^2 of the best 33×33 model is 81.57.

[18] Although the large out-of-sample differences between neighboring grid sizes are not driven by the particulars of a given run, it still makes sense to ask whether the within genetic variation and genetic variation across runs resemble the results found for postcode aggregation. It does. There is high within population genetic similarity and considerably lower across populations. See Tables 13 and 14 in Appendix.

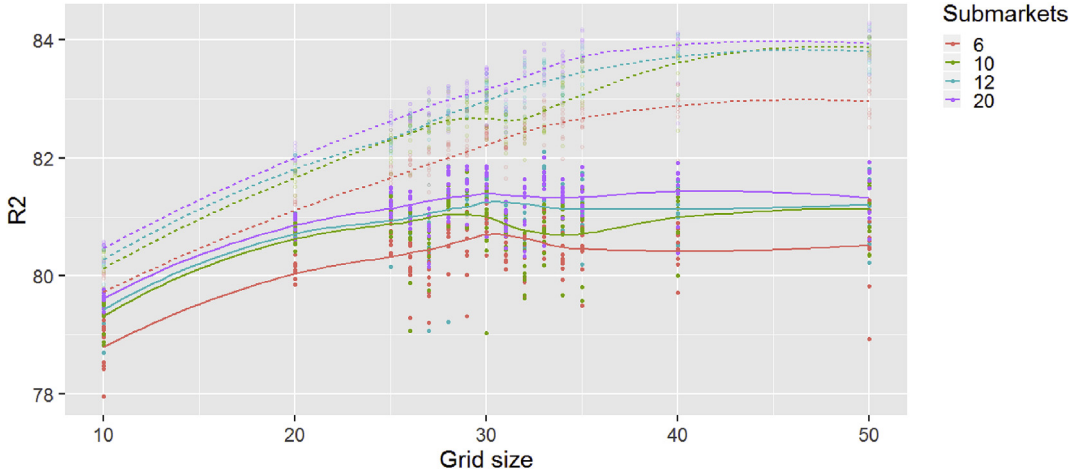


Fig. 8. A plot of in- and out-of-sample R^2s for best in-sample models in final generation for as function of grid size for the number of submarkets equal to 6, 10, 12 and 20. In-sample R^2 transparent dots and dashed curves. The corresponding out-of-sample R^2 non transparent dots and solid curves. 10 runs for each combination of grid size and number of submarkets.

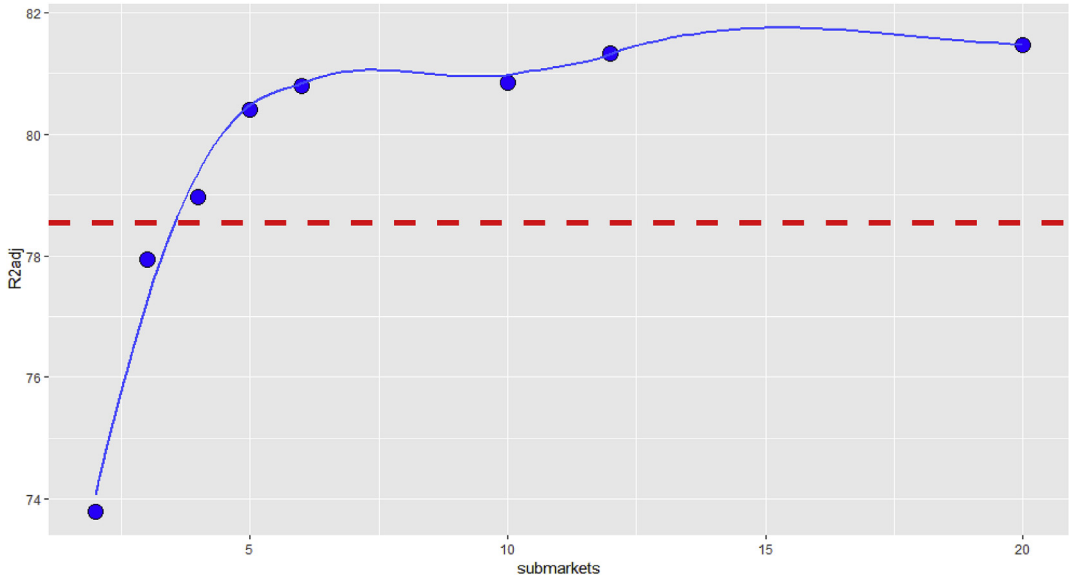


Fig. 9. The adjusted R^2 in percent for the best in-sample model used for out-of-sample prediction on the validation set as a function of number of submarkets. Red dashed line the adjusted R^2 in percent for the full model with 53 submarkets defined by postcode (out of sample on the validation set). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4.3. The number of submarkets

The number of submarkets required to adequately control for location is not obvious. In the previous analysis we aggregated to 12 submarkets. We now consider 6, 10 and 20 in addition to 12. Fig. 8 displays the results. We see that more submarkets correlate with higher R^2s as expected. More interesting to note is that the benefit of adding submarkets is substantial from 6 to 10 submarkets. The additional benefit of going from 12 to 20 submarkets is also significant, but on average gives an R^2 improvement of less 0.5 percent for all grid sizes.

This moderate increase in out-of-sample performance by adding variables is somewhat surprising, and tells us most of the spatial premium variation is accounted for by a few carefully chosen spatial controls. Fig. 8 shows that even the crudest 10 × 10 grid with only 6 submarkets may have a high R^2. It interesting to note that a 10 × 10 model with only 6 submarkets but with the highest in-sample R^2 (80.09), marginally outperforms the full postcode model with 53 submarkets out-of-sample (79.36 versus 79.24).[19]

More interesting here is the comparison with the best 6 submarket model in-sample in the grid region 25 to 35, as there is no reason to start with a crude grid. The best 6 submarket in-sample is for grid size 33 × 33 with an out-of-sample R^2 of 81.1. In other words, a 6 submarket 33 × 33 grid outperforms the full 53 postcode model out-of-sample by about 2 percent.

If economy in the number of variables is essential, we may ask how few submarkets are needed to be on a par with out-of-sample with the full model with 53 submarkets defined by postcode. Fig. 9 shows that just 4 submarkets (3 spatial dummies) are required to outperform the full postcode model (52 spatial dummies) out-of-sample.

That a conventional postcode FE with 52 spatial dummies is outperformed by a model with just 3 spatial dummies is surprising. The most likely explanation is twofold. First, there is considerable location premium heterogeneity within a given postcode. Fig. 10 shows

[19] The adjusted R^2 is 79.04 versus 78.55.

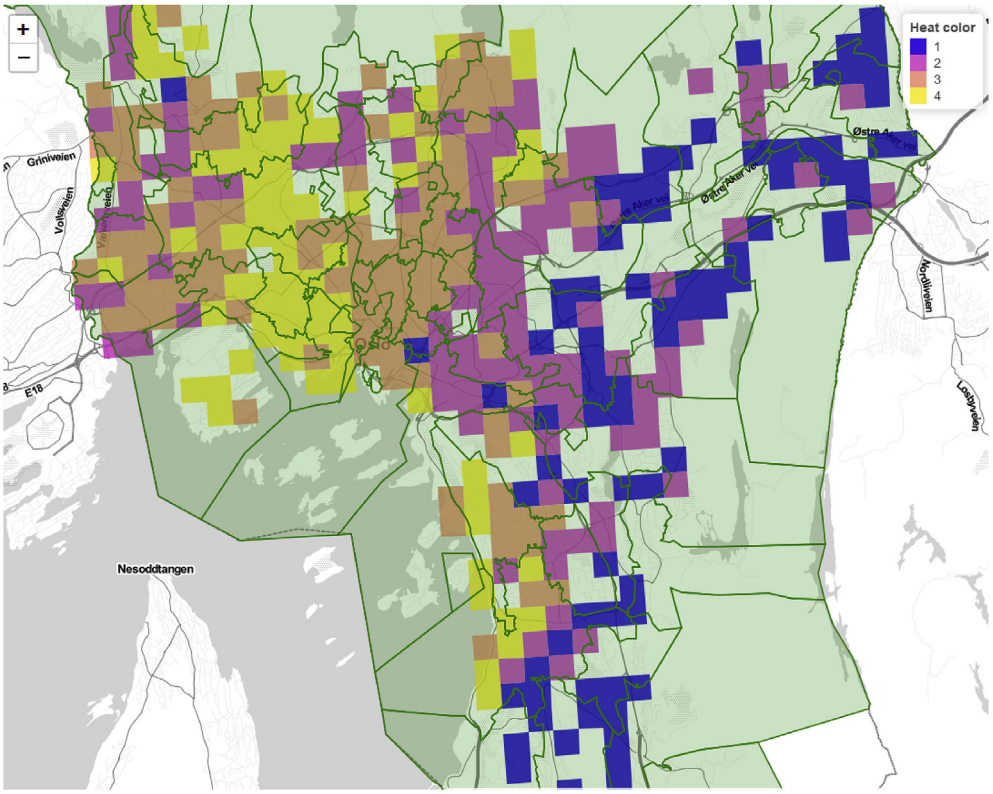


Fig. 10. A comparison of the best in-sample 4 submarkets (R^2 = 81.81) with the full postcode submarket subdivision (green lines). The out-of-sample adjusted R^2 is 79.04 versus 78.55. The aggregation into 4 submarkets is done by aggregation of a 34 × 34 grid with gives 386 squares with housing market transactions. The dimension of these squares are approximately 350 by 350 m. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the four submarket division that outperform the 53 postcode model out-of-sample. According to the four submarket model, there is considerable within postcode price variation for the lion share of postcodes. In other words, even an unbiased estimate of the postcode premium will gloss over systematic within postcode price variation. Second, 53 postcodes involve fewer observations within each postcode and more noisy average effects. Noisy average effects translate into poorer out-of-sample explanatory power. Our analysis suggests that these two adverse effects to a large extent outweigh the potential benefit of a full postcode model. Thus, we come to the counter-intuitive conclusion, that fewer spatial controls gives better spatial controls provided they are chosen with care. A genetic algorithm is one way to implement the "with care" proviso, as it is a highly non-trivial exercise to discover the actual submarkets by more conventional techniques.

At a higher level, this insight opens for two potential improvements when it comes to modeling spatial FEs. First, if economy in number of variables is essential, then we may choose a small set of spatial controls. Second, if economy in the number of variables is not essential, we may choose a somewhat sparser model (say 20 submarkets as in the analysis above) and obtain robust location premium estimates.[20]

5. Conclusion

Econometric models with spatial fixed effects require some kind of spatial aggregation. Conventionally, this is done by block, block group, census tract, postcode, school district, county, city, or state. We show that the trade-off between a low level spatial aggregation with noisy spatial effects and a high level of spatial aggregation with robust but crude spatial effects, can be lessened by using a more flexible spatial aggregation.

The key challenge is to find clusters with similar spatial premiums. We used a genetic algorithm and compared out-of-sample performance with much used approaches like aggregation by postcode.

We find that the out-of-sample performance of the genetic algorithm which aggregated 53 postcodes to 12 submarkets, was similar to that of the full postcode model with 52 postcode dummies. A more flexible approach using a grid to divide the metropolitan area into cells, which are then aggregated into a fixed number of submarkets, tells a striking story. The best combination of grid size and number of submarkets – 35 × 35 and 20 submarkets – had an out of sample R^2 that is about 2 percentage points higher than the model with 52 spatial dummies in-sample. An important takeaway from the analysis, is how few submarkets are actually needed to compete against the traditional FE postcode model. We found that as few as 4 submarkets (3 spatial dummies) could outperform the full postcode model with 53 submarkets (52 spatial dummies) out-of-sample.

The most likely reason is that our flexible spatial aggregation lessens the traditional trade-off between low spatial aggregation and number of observations in each aggregation group by identifying non trivial clusters of observations with similar premiums. It must be stressed that the number of possible aggregations is incomprehensibly large, so the success of the genetic algorithm rests on an efficient search through random variation and non-random selection. That the full postcode fails to harvest the potential benefit of 26 times the number of spatial dummies is interesting in itself. It tells us that, contrary to common wisdom, low spatial aggregation is not synonymous with controlling for more spatial variation. In this sense, what we end up with is a "less-is-more" result. For spatial fixed effects, it is not a question of introducing many spatial dummies, but of finding a few good ones.

Acknowledgement

The authors would like to thank two anonymous referees and participants at WEAI 2017 for valuable comments that helped improve the paper. Moreover, the authors would like to thank Eiendomsverdi AS in general and Erling Røed Larsen in particular for providing the data and giving valuable insights and suggestions.

[20] The best in-sample 35 × 35 grid with 20 submarkets (84.18) has an out-of-sample R^2 of 81.84.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.regsciurbeco.2019.04.005.

6. Appendix

6.1. Preparation of the data set

The data set of all realtor mediated housing market transactions in the metropolitan area of Oslo were acquired from Eiendomsverdi, a Norwegian firm that collects housing market data, and produces house price indices. Preparation of the data set is summarized in Table 7.[21]

Table 7
Dataset preparation.

Data operation	Number of sales
All transactions	98,599
Postcode between 100 and 1299	98,580
Observations with sale date	92,933
Living area 1–99th percentile	91,095
Observations with build year	90,889
Transaction price 1–99th percentile	89,115
Apartments	72,464
Observations in year 2014 or 2015	14,036
Observations in training set	8,400
Observations in validation set	2,836
Observations in test set	2,800

6.2. Genetic diversity and economically different models

There are essentially two ways to address how different two spatial aggregations are. One is to compute the extent to which two three-digit postcodes that are aggregated to the same submarket in aggregation A also are aggregated to the same submarket in aggregation B. The second is see whether the estimated submarket price effect is statistically and economically different. It might be beneficial to draw on notions from biology. The first measure will essentially be a question of genotype. That is, to what extent the two genomes differ. The latter is how the gene is expressed, or the extent to which differences in genotype translate into different phenotypes. For us, the different location premiums can be viewed as different phenotypes.

Let us look at the question of genotypes first. We are not concerned with the actual numbering of the submarkets, only with which postcodes are grouped together. Consider parent one given in Section 3.1:

Parent one: (1,2,1,3,3,4).

An equivalent representation of parent one is (3,2,3,1,1,4), as we just have switched the label of submarket 1 to submarket 3 and vice versa. To give a measure of genetic similarity, it is better to have a unique representation of a given aggregation.

Table 8 gives the unique the matrix representation of parent one (with rows and column labels in italics). This matrix contains the aggregation information. It is symmetric, since if area i is in the same submarket as area j, then j is in the same as i. Moreover, it has 1 on the diagonal as every area is necessarily in the same submarket as itself.

Table 8
Matrix representation of Parent one.

	1	2	3	4	5	6
1	1	0	1	0	0	0
2	0	1	0	0	0	0
3	1	0	1	0	0	0
4	0	0	0	1	1	0
5	0	0	0	1	1	0
6	0	0	0	0	0	1

We denote this matrix by G. The rows and columns correspond to a pregiven ordering of geographical areas, and two areas are aggregated to the same submarket if and only if $G_{ij} = 1$. All other elements in matrix G are equal to zero. Note that two matrices of this kind G^a and G^b, are equal if and only if they define the same partition into submarkets.

We define the similarity of G^a and G^b to be given by:

$$S(G^a, G^b) = \sum_{i<j} I(G^a_{ij} = G^b_{ij}),$$

[21] Zero floor recoded to first floor (corresponding to English ground floor (11 cases). Missing floor recoded to median floor (3) (2359 cases).

where $I()$ is an indicator function which is 1 if the equality is true, and 0 otherwise.

Note that we only sum over matrix the upper triangle and exclude the main diagonal. This similarity measure satisfies the criterion $\max L = L(G^a, G^a)$ for all x. The following measure normalizes the max to be one hundred:

$$NS(G^a, G^b) = 100 * \frac{S(G^a, G^b)}{(S(G^a, G^a)S(G^b, G^b))^{\frac{1}{2}}}$$

Furthermore, by construction it is a nonnegative number less than or equal to one hundred. We view this as a percentage similarity measure.[22]

It is evident that the lower bound of normalized similarity varies with the number of submarkets and the number of groups, and that it is decreasing in the number submarkets. With our genome of 53 postcodes that are aggregated into 12 submarkets the average similarity of 100 random models was 8.3 percent.[23]

We expect genetic variation to be lost in the fittest half of the population as the improvement in R^2 levels off. Table 9 confirms this.

Table 9
Genetic similarity for upper half according to fitness ($R2$) in last generation for 4 different populations.

Population	Min	Mean	Max
1	79.75	79.87	79.90
2	79.56	79.69	79.73
3	79.97	80.06	80.07
4	79.90	80.03	80.04

More interesting is whether different runs converge toward (essentially) the same aggregation. Table 10 shows four different runs and the genetic similarity varies considerably.

Table 10
Genetic similarity of the fittest models of different populations.

	1	2	3	4
1	100	55.2	39.6	50.7
2		100	33.7	48.6
3			100	36.5
4				100

We see that the genetic similarity between the best models in different populations is roughly midway between the similarity between two randomly drawn models and the within population similarity of the upper half of the last generation.[24]

We now turn to the question of whether the differences in genotype translate into differences in location premiums for the areas defined by 53 three-digit postcodes. Fig. 11 displays the heatmaps for the coefficient dummies where all coefficients are relative to the lowest priced submarket set to zero. We see unsurprisingly that high and low price areas are consistently identified as high and low priced. More interestingly, there seem to be statistically and economically significant differences between these models, as the location premium of the same areas defined by the dummy regression coefficient may vary by several hundred thousand NOK.[25]

[22] This measure is a close analogue of the (Pearson's) correlation coefficient $r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$.

[23] The chance of two random draws of numbers 1 to 12 to be equal, is $\frac{1}{12}$, or 8.3 percent. In other words the normalized similarity measure gives this degree of similarity (on average) for two random draws, though it is not obvious from the definition.

[24] This result is in line with the famous experiment by Lenski and Travisano, 1994. They considered a controlled lab experiment of twelve different populations of *Escherichia coli* bacteria. They followed thousands of generations and observed that all twelve populations evolved towards larger cell size. However, the twelve respective gene pools achieved this by different genetic mutations.

[25] 1 NOK = 0.13 USD.

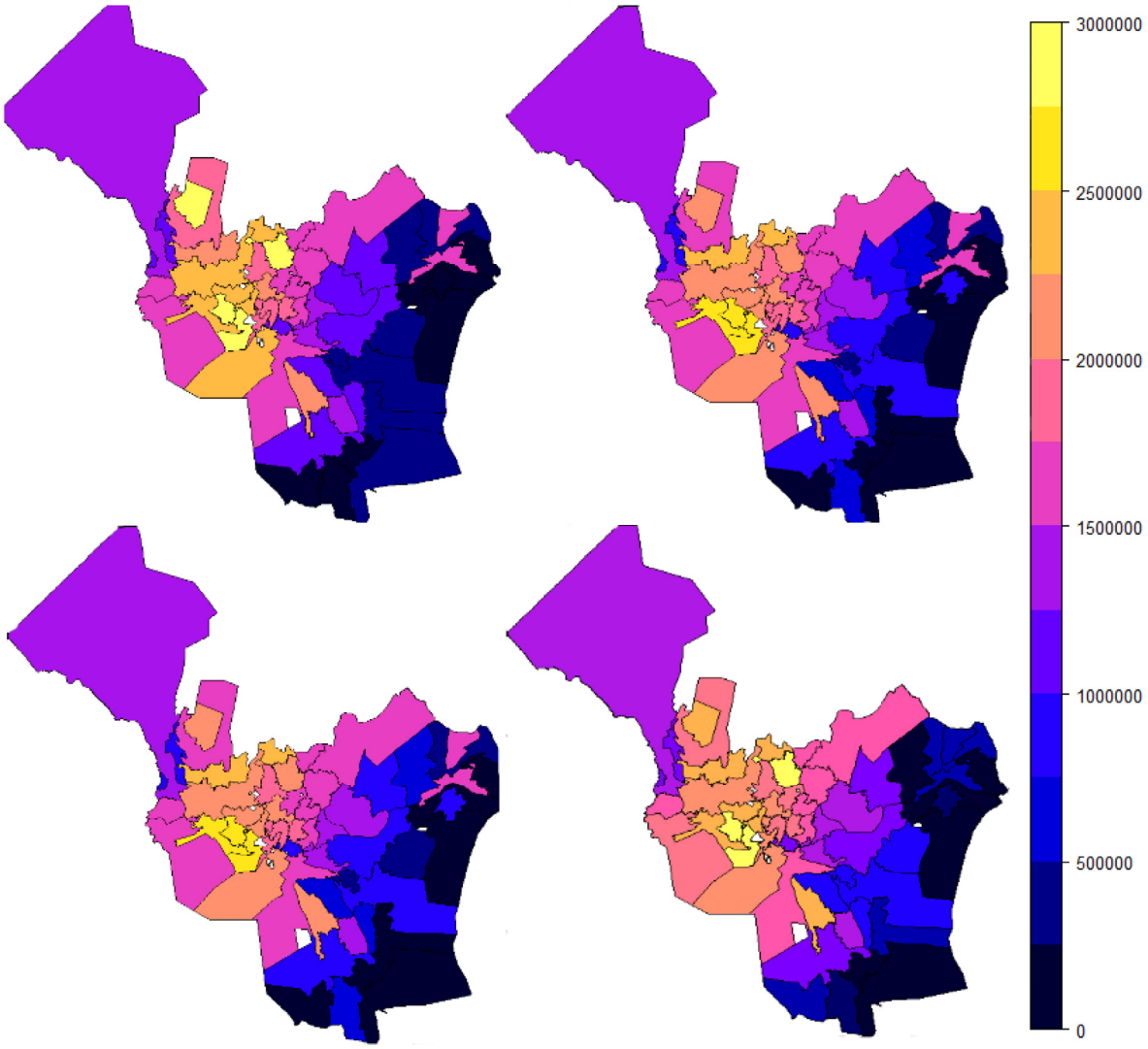


Fig. 11 Spatial aggregations into 12 submarkets given by fittest model in last generation of four different runs. Colors defined by estimated submarket location premium dummy in NOK.

Table 11
Percentage of statistically insignificant differences dummy coefficient estimates by three-digit postcode. Significance measure: zero not within two standard deviations for the coefficient difference.

	2	3	4
1	71.2	69.2	71.1
2		63.5	73.1
3			63.5

Table 11 shows that for most postcodes the difference in location premium across models is not statistically different from zero. At the same time, about a third are statistically different from zero. In other words, the models differ statistically and economically for a substantial number of three-digit postcodes.

It is puzzling that models that are less than 0.57 percent away from the theoretical R^2 bound differ so significantly economically in a third of the postcodes. At first glance, this is arguably even counter-intuitive as these models only differ on spatial aggregation of postcodes.

The solution to this puzzle lies with the differing number of transactions in each postcode. They range from 15 to 128 transactions, where the first quartile is 27. Postcodes with few transactions have a noisy location premium, and a modest impact on R^2. The first may be viewed as a general uncertainty. The latter implies that the evolutionary pressure is weaker for postcodes that have few transactions.

Table 12 illustrates this point. Few observations tend to be consistently misplaced in the sense that a new run is likely to give a statistically different location premium.

Table 12
Number of observations in the training set sorted by the number statistically significant differences of three-digit postcode levels of the 4 different runs.

Significant	No. of obs.
0	3847
1	1575
2	1313
3	1252
4	156
5	257
6	0
Sum	8400

Table 13
Genetic similarity for upper half according to fitness ($R2$) in last generation for 4 different populations. The 33 × 33 grid.

Population	Min	Mean	Max
1	93.4	97.1	100
2	95.5	97.1	100
3	95.1	97.7	100
4	93.4	99.3	100

Table 14
Genetic similarity of the fittest models of four different populations runs. The 33 × 33 grid.

	1	2	3	4
1	100	34.2	35.7	28.6
2		100	34.6	29.2
3			100	32.4
4				100

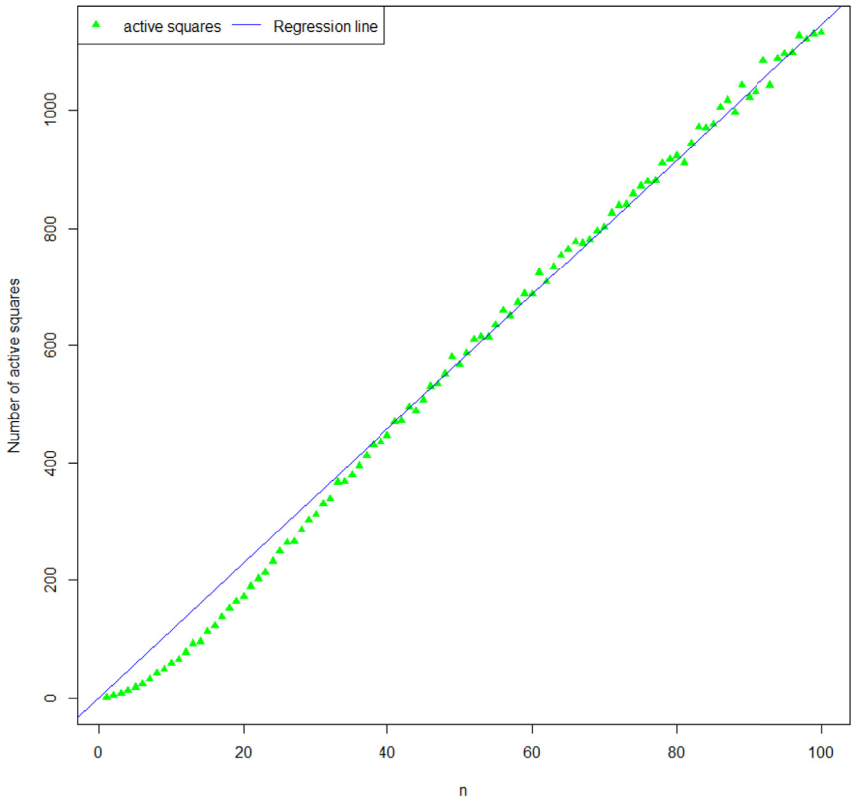


Fig. 12 : The number of cells with housing market transactions. Regression line without constant term, coefficient 11.4 and R^2 = 99.8 percent.

References

Black, Sandra E., 1999. Do better schools matter? parental valuation of elementary education. Q. J. Econ. 114 (2), 577–599.
Caplin, Andrew, Chopra, Sumit, Leahy, John V., LeCun, Yann, Thampy, Trivikraman, 2008. Machine Learning and the Spatial Structure of House Prices and Housing Returns.
Gian-Carlo Rota, 1964. The number of partitions of a set. Am. Math. Mon. 71 (5), 498–504.
Lenski, Richard E., Travisano, Michael, 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. Unit. States Am. 91 (15), 6808–6814.
Marsland, Stephen, 2009. Machine Learning: an Algorithmic Perspective, first ed. Chapman & Hall/CRC.
Ng, S Thomas, Martin, Skitmore, Wong, Keung Fai, 2008. Using genetic algorithms and linear regression analysis for private housing demand forecast. Build. Environ. 43 (6), 1171–1184.
Plakandaras, Vasilios, Gupta, Rangan, Gogas, Periklis, Papadimitriou, Theophilos, 2015. Forecasting the US real house price index. Econ. Modell. 45, 259–267.
Pryce, Gwilym, 2013. Housing submarkets and the lattice of substitution. Urban Stud. 50 (13), 2682–2699.
Randolph, Bill, Tice, Andrew, 2013. Who lives in higher density housing? A study of spatially discontinuous housing sub-markets in Sydney and Melbourne. Urban Stud. 50 (13), 2661–2681.
Shekarian, Ehsan, Fallahpour, Alireza, 2013. Predicting house price via gene expression programming. Int. J. Hous. Mark. Anal. 6 (3), 250–268.