Norwegian University of Life Sciences (NMBU)

Learning from man or machine: Spatial aggregation and house price prediction

Dag Einar Sommervoll, Åvald Sommervoll

Norwegian University of Life Sciences
Centre for Land Tenure Studies

Centre for Land Tenure Studies Working Paper 4/18

Learning from man or machine: Spatial aggregation and house price prediction

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

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

Abstract

House prices vary with location. At the same time the border between two neighboring housing markets tends to be fuzzy. When we seek to explain or predict house prices we need to correct for spatial price variation. A much used way is to include neighborhood dummy variables. In general, it is not clear how to choose a spatial subdivision in the vast space of all possible spatial aggregations. We take a biologically inspired approach, where different spatial aggregations mutate and recombine according to their explanatory power in a standard hedonic housing market model. We find that the genetic algorithm consistently finds aggregations that outperform conventional aggregation both in and out of sample. A comparison of best aggregations of different runs of the genetic algorithm shows that even though they converge to a similar high explanatory power, they tend to be genetically and economically different. Differences tend to be largely confined to areas with few housing market transactions.

Keywords: House price prediction, Machine learning, Genetic algorithm, Spatial aggregation
JEL: R31, R21, C45

1. Introduction

House price prediction is hard. Every house is essentially unique and a given house may be sold just few times during the course of a century. In addition, houses tend to vary in characteristics and amenities. The housing market affects the wider economy through multiple channels, and house price movements are therefore closely watched by banks, policy makers and the general public.

Email addresses: dag.einar.sommervoll@nmbu.no (Dag Einar Sommervoll), aavalds@student.matnat.uio.no (Åvald Sommervoll)

The authors would like to thank Eiendomsverdi for providing the housing market transactions data set, and Eiendomsverdis director of research Erling Røed Larsen for his insights and encouragement. Comments from participants of WEAI 2017, and RBNZ bank's research seminar provided valuable input at the earlier stages of paper.

A wide array of house price prediction models is in commercial and academic use. The models can be divided into two groups, hedonic models ([Ros74],[GT03]) and repeat sales models ([BMN63],[CS89]). There are also hybrid models that seek to utilize the combined strength of hedonic and repeat sales models ([EQR98]). Whether or not the aim of the house price model is to satisfy academic or commercial needs, model selection is a delicate task. Schooled professionals combine housing market insight with sophisticated statistical tools and typically pit different models against each other to find the most attractive.

The rapid advancement of machine learning has created intriguing scenarios for model selection. In fact, the future may already be here in the sense that machine learning is already on it's way into commercial house price prediction.[2] Machine learning is also making an entry in academic research [AI17]. In some ways, the shift has actually been pretty slow given that machine learning technology oftentimes outperforms traditional econometric approaches by a wide margin when it comes to predictive power. Part of this econometric inertia may be due to the "black box" nature of some machine learning methods which may fail the academic transparency requirements.[3] Moreover, some machine learning algorithms are known to have both poor asymptotics and biases ([CCD$^+$17]).

A third potential reason is that machine learning represents a paradigm shift, one that comes with a new vocabulary. In particular, taking the model to the data is learning, and the data set is routinely divided into three parts the training set, the validation set and test set.[4] This learning tends to be an overt data mining, and the role of the validation set is to pinpoint when in-sample prediction comes at the expense of out-of-sample prediction (measured on the validation set).

Apart from relabeling data sets and potentially dividing the data set into three (usually unequal) parts, one big difference between traditional econometric approaches and machine learning is the change in viewpoint. Often the goal is not one good model, but an ensemble of models. Each of these models may be far from impressive (weak learners), but in aggregate outperforming a single carefully selected model. The quiz show analog to this is instead of searching for the ultimate know-it-all quiz expert, we select two hundred good quizzers and select answers by way of majority vote.

The aim of this paper is twofold. First, to contribute to the house price modeling literature regarding spatial aggregation and in particular how to identify submarkets

[2] As of spring 2017 Zillow the online real estate database company, has launched a machine learning competition for house price prediction with a prize fund of 1.2 million USD.

[3] Some of the machine learning approaches that are close to much used econometric tools, like LASSO and Ridge- regression, already staples of most econometricians' tool boxes, and methodological advances warrant publication in top journals ([Koz17]).

[4] This is only true given enough data to warrant a division into three disjoint parts.

that are similar though potentially spatially separated. Second, to attempt to bridge the gap between standard econometric methods and machine learning by considering a hybrid approach. We rely on standard hedonic house price prediction models, but use a machine learning algorithm, a genetic algorithm, to help us find a good spatial aggregation. We show that the genetic algorithm (GA) consistently finds models with surprisingly good in- and out-of-sample properties (measured by R^2). We find the genetic variation between best models in the same GA run to be low. This is to be expected as all these models are close to (local) maximum in terms of fit measured by R^2. More intriguing, the genetic variation across different runs of the GA is high even though the best models across runs have strikingly similar fits measured by R^2. In biological terms, different genotypes lead to similar phenotypes. This is virgin land in econometrics, and we explore to what extent these different genotypes lead to *economically* different models. They do, and some clusters of observations which tend to get statistically and economically different estimates for different population runs. This insight may be of importance not only for machine learning in urban economics but for machine learning in econometrics in general.

We are not the first to consider machine learning algorithms for house price prediction. The lion's share of academic contributions within computer science regarding house price prediction is artificial neural network approaches

([CS13],[LWJ01][CCMO14][QTNH][KTTT],[Lim04],[LGL04]).

It is puzzling that artificial network approaches have been a favorite in machine learning and real estate pricing. As location is a key house price factor, other machine learning techniques which are good at finding spatial clusters would in some sense be more likely first attempts.[[OS16] and [AP12] are papers that consider explicitly geometric machine learning constructions like k-NN (the k nearest neighbors).

The papers cited thus far are computer science papers, where in some sense the application, house price prediction, is not interesting in itself. It is the performance of the algorithm that's the focus point. In contrast to this, and more ambitious from an economist's point of view, is [CCL$^+$08], where a spatio-temporal closeness of observations lies at the heart of the machine learning algorithm. Self-organizing maps (SOM), another promising machine learning technique, have been applied to housing markets ([KHH02],[Kau03]).

Other approaches seek to address the sparse nature of housing market data, by using machine learning algorithms that tend to perform well under such optimation constraints. [WWZW14],[PGGP15] and [WH15] are examples that utilize support vector machines (SVM) and particle swarm optimation to forecast real estate prices. [AP12] uses a random forest approach and ensembles of weak learners. In a similar vein is [PB15] where a wide array of different machine learning algorithms is put to the test of predicting house prices.

We use a genetic algorithm (GA) on housing market data. Earlier contributions regarding GAs and housing markets are [NSW08] and [SF13]. Our machine learning approach most resembles [PGGP15], as population of regression models evolve according to a fitness measure given by the models explanatory power. However, our approach is transversal in two ways. First, whereas [PGGP15], take the spatial division as given, our models differ only according to spatial aggregation. Second, [PGGP15] are less concerned with the different models' in- and out-of-sample properties. For us, however, this is a key point.

We also seek to contribute to the growing literature on non-spatially connected submarkets. Recently, both theoretic and empirically based research has challenged the legitimacy of spatially connected submarkets [Pry13]. Our genetic algorithm has no priors regarding spatial connectedness. The only objective is house price prediction. Spatially disconnected submarkets arise and they resemble submarket structures found in [RT13] for the Sydney market.

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. Moreover, we give aggregate neighborhoods defined by postcode based on square meter prices into 12 "prize zones". In section 3 we describe a genetic algorithm for aggregating these three-digit postcodes into 12 areas. This algorithm uses only the fit (measured by R^2) of the hedonic model with a given spatial aggregation. Section 4 considers spatial aggregation in Oslo by rectangular cells. This approach is more flexible than postcode aggregation as we may vary cell size. 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 arms 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 12 in the appendix.

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

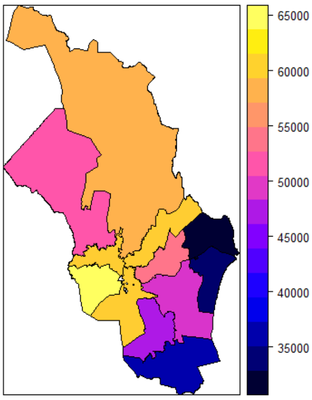

Figure 1: Heat map of median square meter prices for Oslo 1 to Oslo 12.

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

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

$$p_i = \alpha + \beta_{\text{logArea}} LogArea + \beta_{\text{logAge}} LogAge + \sum_{i=2}^{14} \theta_i Floor_i + \sum_{i=2}^{24} \nu_i Month_i + \sum_{i=1}^{11} \gamma_i Nbhoods_i + \epsilon, \tag{1}$$

where $LogArea$ is log(Area in sqm.), $logAge$ is the log(2017-construction year), $Floor_i$, $Month_i$, and $Nbhoods_i$ are dummy variables.

The neighborhoods 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 postcode removed.[5]. This is our default division of Oslo into 12 neighborhoods as displayed in figure 1. In this paper we keep the number of spatial dummies fixed and equal to 12.[6]

Table 2 gives regression results for the baseline model used on the training and validation sets. Of particular interest is the explained price variation,77.42 and 77.01 percent respectively. In the next section we discuss the extent to which the explanatory power of a model derived from a machine learning algorithm translates to the validation set. There are two different interpretations of out-of-sample performance. The classical one is to use the coefficient estimates on the validation sample to compute the explanatory power (R^2). This is the second column (TrainValidation). However, the litmus test is not whether the given regression coefficients have

[5] Postcode 1236, is in Oslo 12.

[6] This number is in some sense arbitrary (though it originates from Oslos 12 areas defined by postcodes). To find the best or a good number of spatial controls is depends on the analysis at hand. We do not address this problem, as we take spatial aggregation into 12 neighborhood as given.

good out-of-sample properties, but whether the hedonic model with a given spatial aggregation has good out-of-sample properties. This is given in the third column under the label "ValidationValidation". The fourth column is the regression model without spatial controls. This implies that the difference $77.42 - 65.70 = 11.72$ is directly interpretable as the added benefit of introducing 12 neighborhood dummies. A roughly 12 percent gain in explanatory power in a model that already has substantial explanatory power, may be taken as a sign that these neighborhoods defined by postcodes capture a substantial amount of the price variation driven by location. Still there is little reason to believe house prices closely mirror this broad brush postcodes selections.

Table 2: Baseline regression [a]

	Dependent variable:			
	Transaction price			
	TrainTrain	TrainValidation	ValidationValidation	No Spatial Controls
logLiving	3.19^{***}	3.19^{***}	3.05^{***}	3.31^{***}
	(0.02)	(0.02)	(0.04)	(0.03)
logAge	-0.16^{***}	-0.16^{***}	-0.14^{***}	-0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)
Observations	8,400	8,400	2,836	8,400
R^2in percent	77.42	75.82	77.03	65.70
Note:				*p<0.1; **p<0.05; ***p<0.01

[a] TrainTrain and ValidationValidation are OLS estimates of the regression model.TrainValidation uses the TrainTrain coefficients to predict on the validation set.

We proceed to construct partitions of Oslo into 12 neighborhoods, 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 choose partitions of the metropolitan area of Oslo. At the same time our intuition fails us when it comes to comprehend the immensity of all possible aggregations. In the data set at hand there are 385 distinct postcodes. The number of possible partitions is given by the Bell number $B_{385} \approx 1.0 \cdot 10^{226}$. [7] In comparison the number of elementary particles

[7] The generating function for Bell numbers is given by $\sum_{i=0}^{\infty} B_n x^n = e^{e^x - 1}$ [Rot64].

in the universe in believed to be around 10^{80}. [8] Most of these aggregations will naturally be poor candidates for neighborhoods for house price prediction purposes.

In the following we will limit the number of aggregations we to consider so that we in principle can pit models against each other. There are essentially two ways to do this. One is to impose some apriori 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 skipping the last digit, creates larger (spatially connected) building blocks for aggregations. In our data set there are 53 such three-digit postcodes. If we take them as the basis for further aggregation we're left with $B_{53} \approx 1.0 \cdot 10^{55}$ possible aggregations.[9]. This number is small compared to our original $1.0 \cdot 10^{226}$, but still far from a list of models we can run through to find the best one. If we want to pit a given model against the benchmark model above, it makes sense to consider only aggregations into 12 neighborhoods. In this way we compare models with the same number of explanatory variables, but where the spatial dummies differ in the underlying aggregation.

The number of possible aggregations into 12 neighborhoods is of the order of 10^{48}.[10] 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 1,000 random aggregations into 12 groups, 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 and less random approach is to aggregate by some observable summary statistics that is likely to correlate with price level. One such variable is average square meter price. Figure 2 gives a heat chart of square meter prices for the three-digit postcodes.

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. The

[8] Not only is the magnitude of these numbers hard to comprehend. It is hard fathom how much larger 10^{226} is compared to 10^{80}. If we glue a copy of our universe onto every elementary particle in our universe, and count elementary particles in this "augmented" universe, we get 10^{160}. This is still dwarfed by the staggering order of magnitude of 10^{226}.

[9] This is astronomically comprehensible in the sense that is higher than the number of stars in our universe (10^{22}), but lower than the number of elementary particles in the universe (10^{80}).

[10] It is of the same order as $(12)^{53}/12!$.

Table 3: Summary Statistics of 1000 random spatial aggregations into 12 neighborhoods

Statistic	N	Mean	St. Dev.	Min	Max
R^2 in percent	1,000	69.69	1.3	66.55	75.36

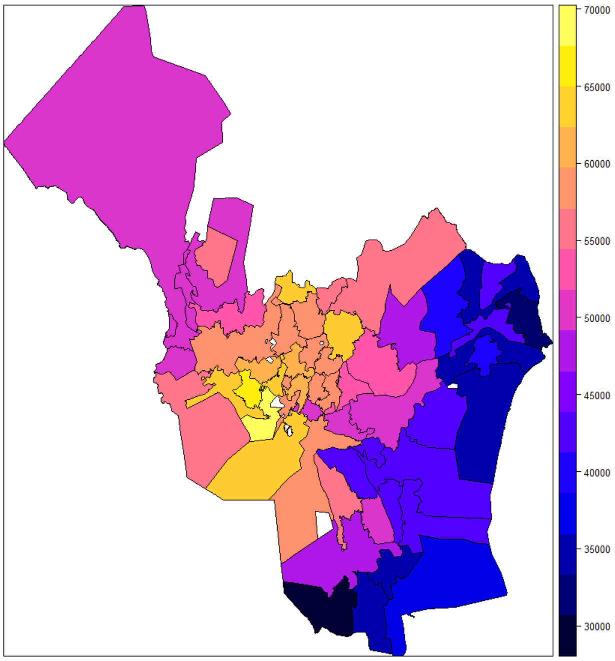

Figure 2: Heat map of median square meter prices neighborhoods defined by three-digit postcodes. Postcodes with no transaction excluded (white).

number of such partitions is $\binom{52}{11} = 60,403,728,840$. This is still a daunting number but maybe within the realm of what is computable by a supercomputer.[11]

Further natural limitations are to divide the 53 postcodes into 12 roughly equally sized groups. Here 4 or 5 postcodes in each group or more refined, choose 12 groups in such a way that the within group variance is minimized. In table 4 we pit these two variants against the in sample[12].

Table 4: Comparison of selected spatial aggregations.

Model	R^2 in percent
Baseline without spatial controls	65.70
Baseline with spatial controls, Oslo 1- Oslo 12	77.42
Aggregation by sqm. (4-5 three-digit post code in each neighborhood)	78.73
Aggregation by sqm. (Minimal within neighorhood sqm. variance)	78.74
Model with 53 postcode dummies	80.27

The explanatory power of these two models is virtually identical and 1.3 percent higher than the baseline model with spatial controls. To determine whether this is a substantial improvement the "full model" with 53 spatial controls is tabulated. This serves as an upper bound for aggregation improvement as any aggregation of the 53 postcodes necessarily would have a lower R^2. The R^2 of the model with 53 spatial controls is 80.27 percent. So not only do these square meter models yield 1.3 percent higher explanatory power than the baseline model, they are also just 1.5 percent away from the full model with 41 degrees of freedom more. In other words the improvement is close to half way to the theoretical upper bound.

In the next section we use a genetic algorithm (GA) to this problem of spatial aggregation. In light of the discussion above, the question is whether it can surpass the square meter models and get (even) closer to the theoretical upper bound defined by the model with 53 time dummies.

3. Three-digit postcode spatial aggregation and GA-algorithm

In this section we use a genetic algorithm (GA) to find spatial aggregations of the 53 three-digit postcodes into 12 neighborhoods which give high R^2's when used in regression model 1. Before we go into specifics regarding the genetic algorithm we use here, we will 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

[11] With a standard laptop and standard run times checking all alternatives would take in the ballpark of $6.6 \cdot 10^8$ seconds or a little more than 21 years.

[12] As there is no training or data mining apart from working out square meter prices, there is no significant difference between in- and out-of-sample properties.

ascent.[13] 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, where we rely on random variation but non-random selection. 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 (GA) 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 neighborhoods to 12, is naturally represented by a 53-dimensional vector $(1, 12, 3, 3, \dots)$ indicating which neighborhoods belong to the same group, where the group is naturally identified by its number. We will 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 neighborhoods. The fitness of each model is defined to be R^2 of the hedonic regression model corresponding to the spatial aggregation defined by genotype (the 53-dimensional vector coding for the neighborhood 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. Figure 3 is an example of such a model.

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

[13] In the literature [Mar09] it is more common to use the notion of gradient decent, as the objective is usually to minimize some kind of loss function.

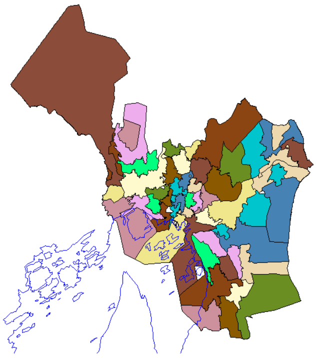

Figure 3: Example of a model represented as a map with borders. Areas with the same colors aggregated to one neighborhood. Blue line defines the Oslo Fjord.

one offspring. These 12 offspring replace the highest ranked models without offspring in this generation.[14]

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)

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 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.[15]

Table 5 summarize the genetic algorithm.

The global R^2 maximum for aggregation into 12 groups is unknown. We have an

[14] As the 24 first models gets offspring, the offspring replace ranks 25 to 36.

[15] Both choices are semi random in the sense that the GA tends not to be supersensitive to 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 groups centers, they are left untouched by mutations.

Table 5: Specification of the GA

population size (N)	Crossover	Mutations	Number of generations
50	Yes	3	3,000

upper bound (80.29) and our prime concern is comparison with the baseline model (77.42) and the more refined model where we aggregated into groups by square meter prices (78.7).

Figure 7 displays a typical run of the GA-algorithm presented in table 5. It is important to note that randomly drawn spatial aggregations are expected to perform significantly worse than the baseline model, as the baseline spatial aggregation captures some known price difference patterns in the metropolitan area of Oslo. We highlight this by the coloring of the plot. The pink region is where random draws are likely to fall. The green region is where the baseline model and square meter models fall. We can view this region as associated to models that tend to capture spatial price variation. The blue region is of special interest since it can be seen as associated with nontrivial (and inhumanly good) spatial aggregations. The question is whether or to what extent natural selection can find models with higher R^2s. 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.

Thus far the best model of 12 groups had an R^2 of 80.06[16] percent compared to the full model (80.27). In other words going from 11 spatial dummies to 52 improve explanatory power by 0.21 percent. This impressive in-sample performance is likely driven in part by overfitting, so the real test is out-of-sample performance (measured by R^2). The R^2 of the fittest model on the validation set is 79.40 (blue line in 7).[17]

The models we presented in the previous section, i.e. the baseline model and square meter models, are based on "natural criteria", political boundaries and square meter prices. They have essentially equal in- and out-of-sample properties. The GA algorithm, in contrast, gives models with better in-sample properties compared to out-of-sample properties. More importantly, the best GA model arises from a herd of models that evolve by random variation and nonrandom selection. This randomness raises new questions and concerns. One of them concerns the extent to which two different runs of the GA algorithm generate different models. This is a question of

[16] This is certainly not the global max. Longer runs with combinations of best models from different runs indicate that there is still room for improvement.The highest in these combined runs is 80.09.

[17] Note that this is the ValidationValidation out of sample. In other words, the aggregation of the fittest model is used, and the model parameters are estimated on the validation set.

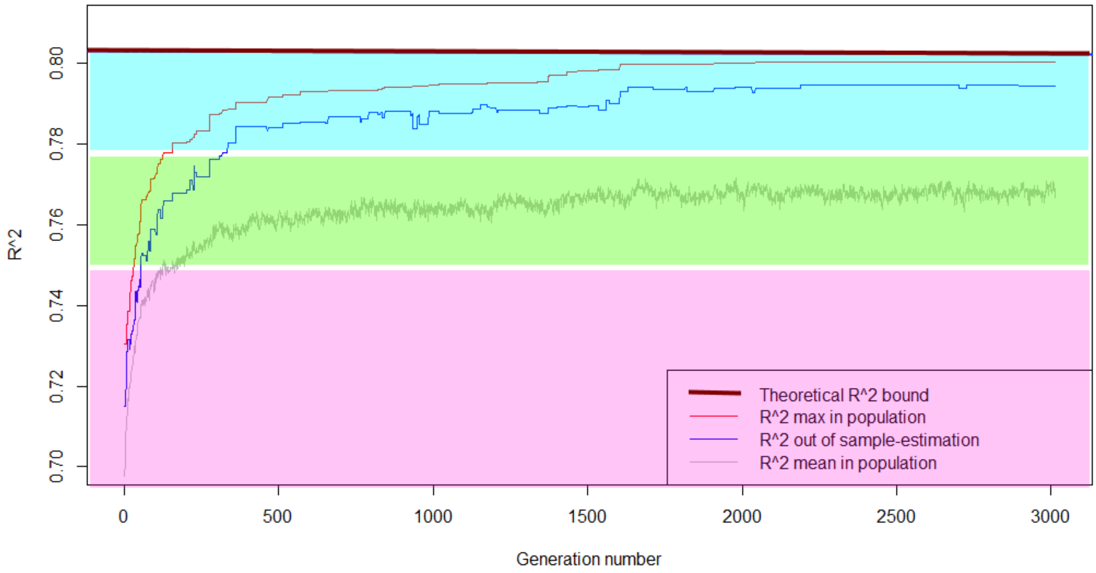

Figure 4: The explanatory power (R^2) by generation number. Pink region is the R^2 region mapped out by upper half of the random aggregations given in table 3. The green is the region contains the baseline model and the models based on square meter prices of the preceding section. The blue region has very high R^2's. Out-of-sample is the ValidationValidation model, i.e., the model corresponding to the fittest model estimated on the validation set.

potential genetic diversity between GA runs, and how far such diversity translates into economically different models.

3.2. 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 group in aggregation A also are aggregated to the same group in aggregation B. The second is see whether the estimated neighborhood 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, that is, to what extent differences in genotype translate into different phenotypes. For us, the different price levels 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 groups, only with which postcodes are grouped together. Consider parent one given above:

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 group 1 to group 3 and vice versa. To give a measure of genetic similarity, it is better to have a unique representation of a given aggregation.

Table 6 gives a 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 neighborhood 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 neighborhood as itself.

Table 6: 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 neighborhood if and only if $G_{ij} = 1$. All other elements in the 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 neighborhoods.

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

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

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 that $\max L = L(G^a, G^a)$ for all x. The following measure normalizes the max to be one:

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

Furthermore, by construction it is a nonnegative number less than or equal to one. We view this as a percentage similarity measure.[18]

It is evident that the lower bound of normalized similarity varies with the number of neighborhoods and the number of groups, and that it is decreasing in the number groups. With our genome of 53 postcodes that are aggregated into 12 groups the

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

average similarity of 100 random models was 8.3 percent.[19]

We expect that genetic variation will be lost in the fittest half of the population as the improvement in R^2 level off. Table 7 confirms this.

Table 7: 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 to what extent different runs converge toward (essentially) the same aggregation or not. Table 8 shows four different runs and the genetic similarity varies considerably.

Table 8: 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 model and the within population similarity of the upper half of the last generation.[20]

We now turn to the question of whether the differences in genotype translate to differences into price levels for the areas defined by 53 three-digit postcodes. Figure 3.2 displays the heatmaps for the coefficient dummies where all coefficients are relative to the lowest priced neighborhood 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 price level of the same areas defined by the dummy

[19] 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.

[20] This result is in line with the famous experiment by Lensky et al. [LT94]. 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.

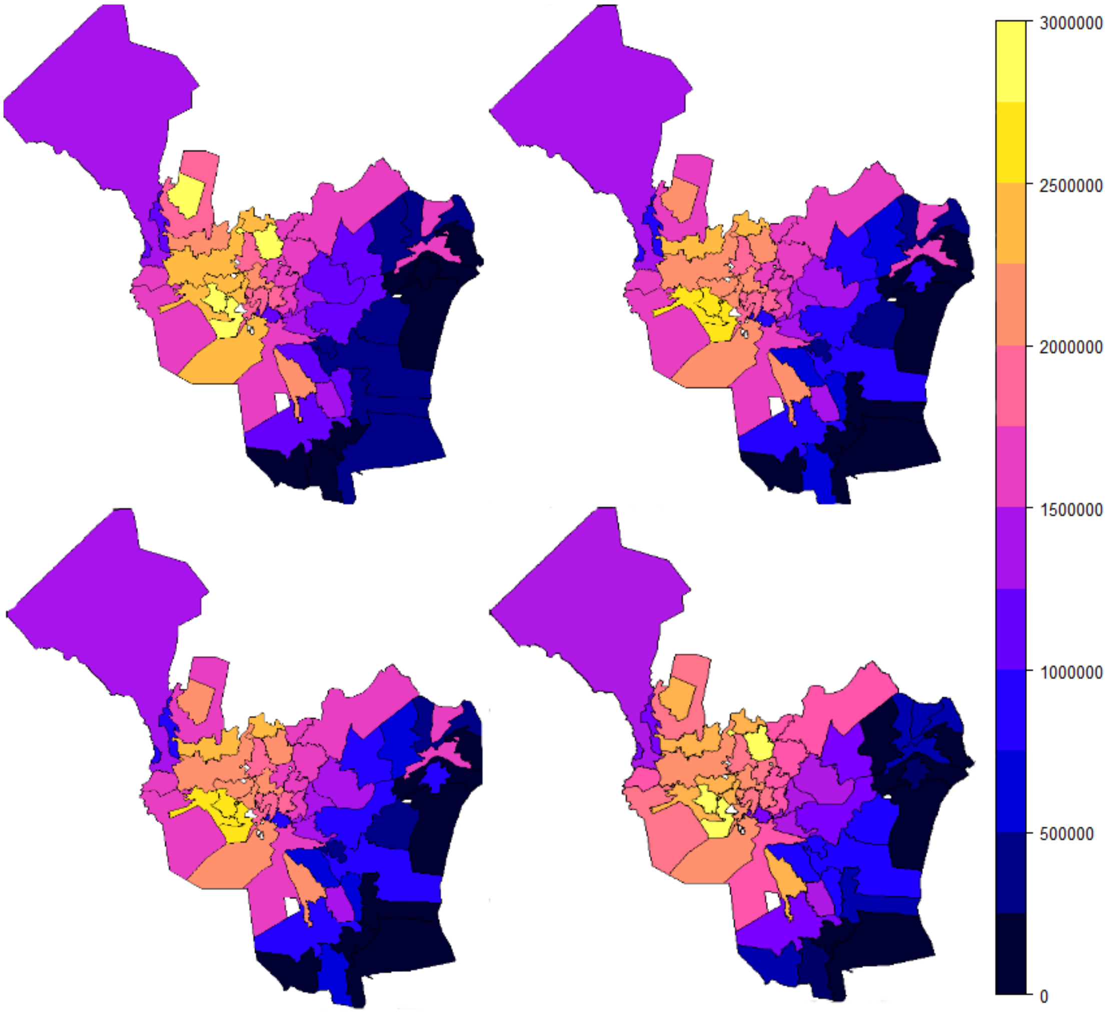

Figure 5: Spatial aggregations into 12 neighborhoods given by fittest model in last generation of four different runs. Colors defined by estimated neighborhood price level dummy in NOK.

regression coefficient may vary by several hundred thousand NOK. [21].

Table 9: 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 9 shows that for most postcodes the difference in price level across models is not statistically different from zero. At the same, 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 significantly economically in a third of the postcodes. At first glance, this is arguably even counterintuitive 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 price level, 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 10 illustrates this point. Few observations tend to be consistently misplaced in the sense that a new run is likely to give a statistically different price level.

Table 10: 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	3,847
1	1,575
2	1,313
3	1,252
4	156
5	257
6	0
Sum	8,400

[21] 1 NOK= 0.13 USD.

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 ([Bla99]) can create sharp price boundaries. In other words, it is likely that choosing spatial aggregations in hedonic regression models that respect administrative boundaries correlates with higher explanatory power ceteris paribus. At the same time, it is also a straight jacket 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.

Figure 4 displays 33 by 33 grid. The coloring indicates an aggregation of the 58 cells with housing market transactions into 12 neighborhoods.

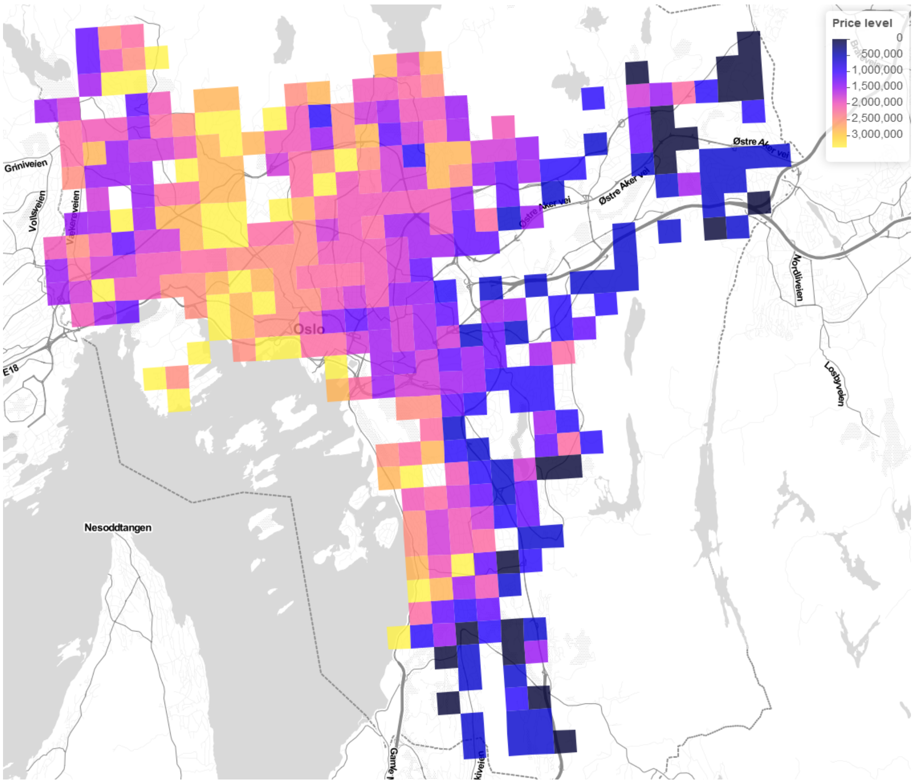

Figure 6: Heat map of 12 neighborhood partition with in sample $R^2 = 83.50$ arising from 33 by 33 grid of the metropolitan area of Oslo. Color defined by estimated neighborhood coefficent dummy.

4.1. Genetic Algorithm

The basic construction of the GA algorithm 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 neighborhoods. Consider an $n \times m$ grid. This grid partitions Oslo into nm cells. An aggregation into 12 groups 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.[22] Any such aggregation may be represented as a matrix, M, of size $n \times m$, where m_{ij} is the corresponding neighborhood number. We view this matrix as a model's genome in the same way that we represented the genome in the post code 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 their fitness measured by R^2. The upper third of the 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 dies in every generation. [23]

The offspring are constructed by letting the six neighborhoods (aggregated groups) be inherited from one of the parents by random draw. These 6 groups 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 groups are inherited from this parent. The other matrix entries are inherited from the other parent.[24]

Mutations may occur at 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 every element in the matrix is the mutation probability 0.005. In other words for a 10×10 genome the expected value of the probability of a mutation is 50 percent.

The GA algorithm is summarized in table 11.

Table 11: Specification of the GA

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

Figure 7 shows a typical run. We see that both the in-sample and out-of-sample R^2 climbs above the 80 percent line after around 1,000 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

[22] 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 $lat\_step\_size = (max(latitude) - min(latitude))/m$.

[23] 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 $\frac{N}{4}$).

[24] Note that this crossover algorithm allows for the contingency that one or more aggregation groups are lost, as it is no preassigned groups 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 groups.

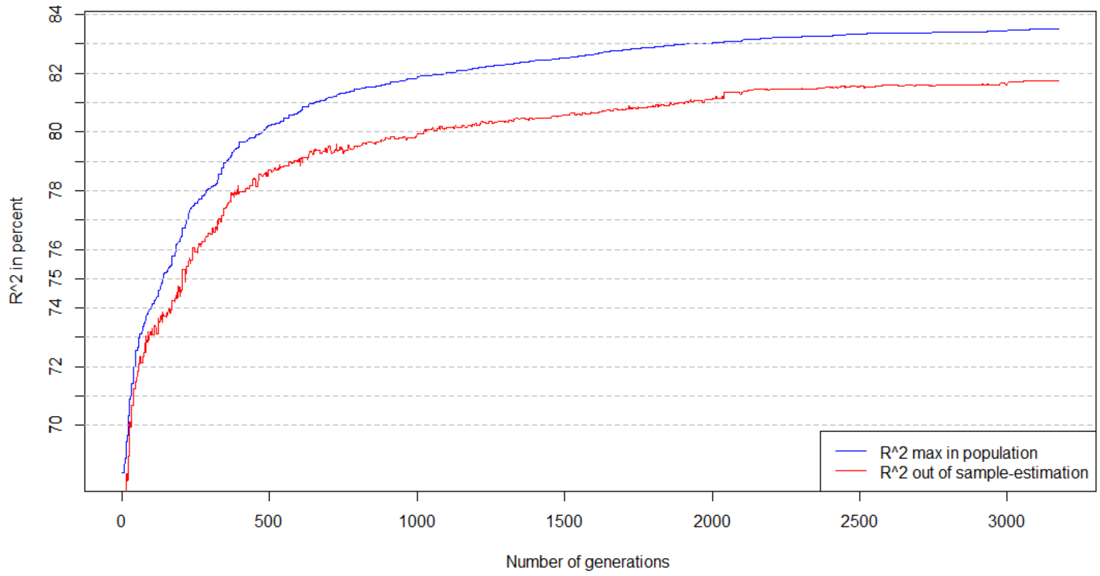

Figure 7: The explanatory power (R^2 in percent) by generation number. A run of the 33 by 33 grid.

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 potential to find models with much higher explanatory power.

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

The grid size defines the building blocks of the spatial aggregation.They are likely to have a bearing on the R^2 of the fittest models in the final generation. Too crude a grid will gloss over systematic spatial price variation, and too fine a grid will pick up transaction noise.[25] The latter will manifest as a good in-sample fit at the expense of out-of-sample fit. In machine learning this is called overfitting, and the most common way to deal with it, is to use a validation set to pinpoint the point at which the in-sample fit comes at the expense of out-of-sample fit. We will use this technique to pinpoint the best grid size with respect to out-of-sample properties.

To allow for the contingency of misleading runs, 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. [26]

Figure 8 is a notch plot showing that the in-sample fit increases up to grid size

[25] This is the common bias variance trade off in statistics.

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

70 to 80, whereas the two out-of-sample measures level off and fall 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 properties, we run all grid sizes from 25 to 35. Figure 9 displays the results. The figure shows a steady increase of in-sample fit with grid size. The out-of-sample R^2 (red notches) has no particular trend. Some grid sizes perform significantly worse than their nearest neighbors. The best is the 33 by 33 grid with the 29 by 29 a close second. The significant differences of 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 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.[27]

At a higher level, fewer observations per cell make the actual price level in the cell hard to assess both for man and machine. The 33 by 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. But as this number is likely to be driven by the spatial clustering of housing market transactions, it may at best serve as a natural starting point when choosing grid size.

[27] Though 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 post code aggregation. It does. There is high within population genetic similarity and considerably lower across populations. See table 13 and table 14 in the Appendix.

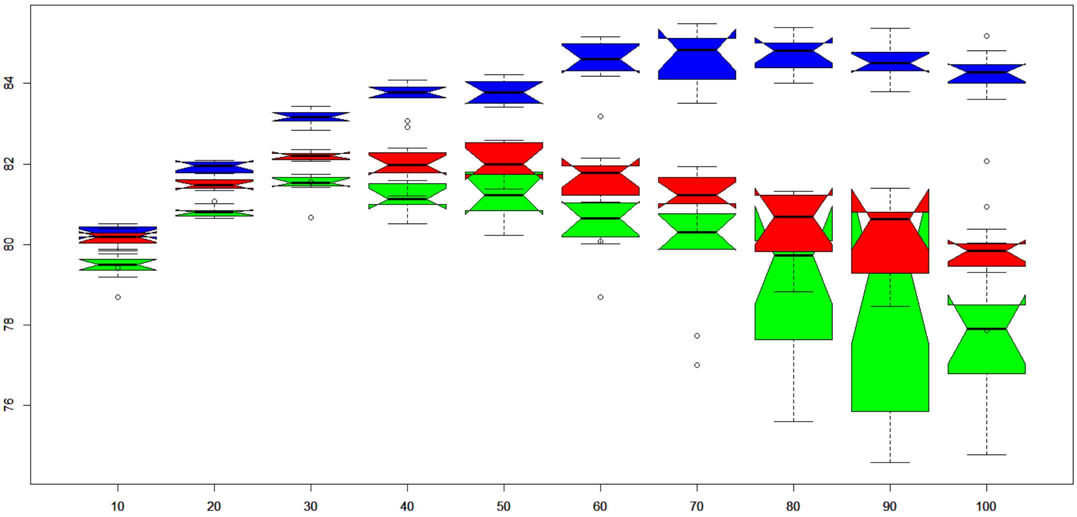

Figure 8: Notch plot of R^2 for the best model training set (blue), the best model used on the validation set (green) and the spatial aggregation of the best model on the training set used for estimating the hedonic model (red). Note that the notch plot notches give the 95 confidence interval for the median and the colored box corresponds to the observations between the 25th and the 75th percentile. Outliers are represented by small circles.

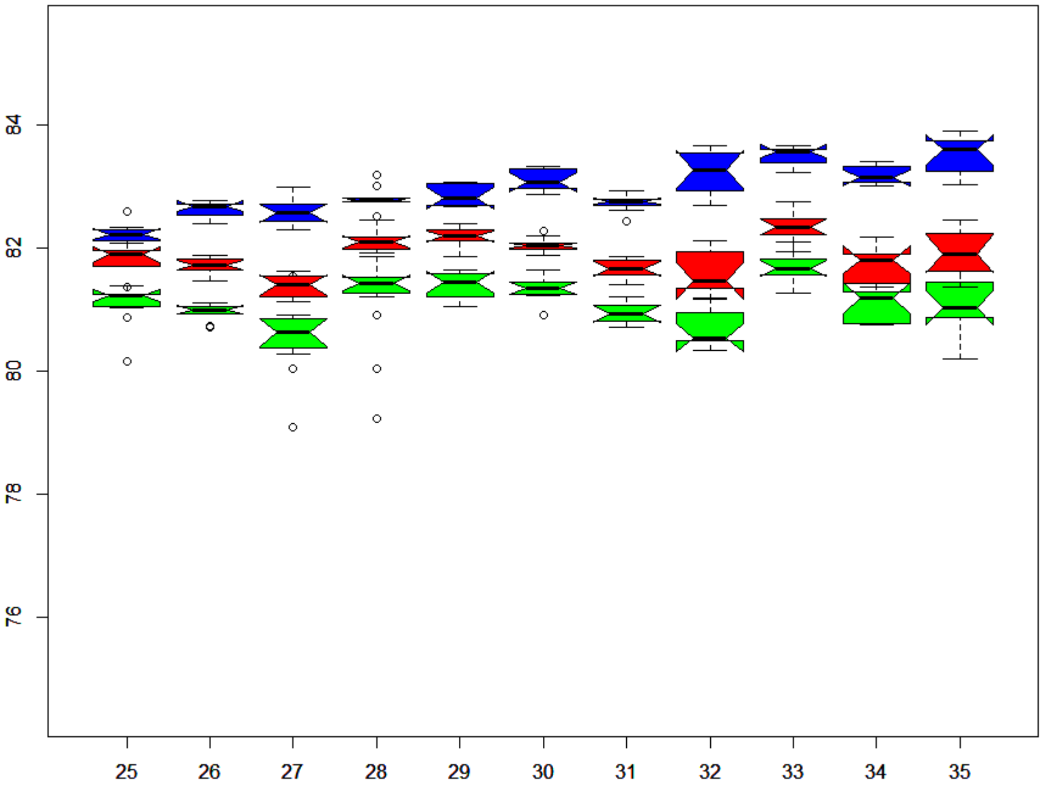

Figure 9: Notch plot of R^2 for the best model training set (blue), the best model used on the validation set (green) and the spatial aggregation of the best model on the training set used for estimating the hedonic model (red). Note that the notch plot notches give the 95 confidence interval for the median and the colored box corresponds to the observations between the 25th and the 75th percentile. Outliers are represented by small circles.

5. Conclusion

Model selection tends to be hard. In the case of hedonic housing market models the analyst wants to follow prices closely in time and across neighborhoods. As houses are seldom transacted, compromises need to be made in order to achieve a robust housing market model. One key ingredient may be to identify spatial submarkets. There are infinitely many ways to aggregate and even when one aggregate using predefined building blocks like postcodes or a grid partition, the number of possible aggregations remains incomprehensibly large. We used a genetic algorithm to search for spatial aggregations with good in- and out-of-sample properties with regard to explained price variation measured by R^2. The GA runs found consistently models with high explanatory power both in and out of sample.

The genetic algorithm works through random variation and non-random selection, and thus it is not necessarily so that the best models across GA runs give economically similar models. We found that genetic variation across runs remained high, though the runs converged to models of similar explanatory power. The genetic differences also lead to statistically and economically different models. However, the economic differences were largely confined to areas with few housing market transactions.

The grid approach allowed us to address the question of a good choice of aggregation building blocks. We found that the 33 by 33 grid gave the best out-of-sample performance. This amounts to ,on average, 23 observations per cell with housing market transactions. It is not obvious that this insight translates over to other housing markets, as the degree of transaction clustering is likely to play a major role. Future research may shed light on this interplay between genetic algorithms and spatial clustering. Thus far, the main take away from the grid analysis, is that it can be easily implemented and tailored to any housing market provided the individual transactions are geocoded. Moreover, the genetic algorithm consistently finds models with high in- and out-of-sample explanatory power.

[AI17] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2017.

[AP12] Evgeny A Antipov and Elena B Pokryshevskaya. Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. *Expert Systems with Applications*, 39(2):1772–1778, 2012.

[Bla99] Sandra E Black. Do better schools matter? parental valuation of elementary education. *The Quarterly Journal of Economics*, 114(2):577–599, 1999.

[BMN63] Martin J Bailey, Richard F Muth, and Hugh O Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942, 1963.

[CCD$^+$17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

[CCL$^+$08] Andrew Caplin, Sumit Chopra, John V Leahy, Yann LeCun, and Trivikraman Thampy. Machine learning and the spatial structure of house prices and housing returns. 2008.

[CCMO14] Vincenza Chiarazzo, Leonardo Caggiani, Mario Marinelli, and Michele Ottomanelli. A neural network based model for real estate price estimation considering environmental quality of property location. *Transportation Research Procedia*, 3:810–817, 2014.

[CS89] Karl E Case and Robert J Shiller. The efficiency of the market for single-family homes. *The American Economic Review*, 79(1):125–137, 1989.

[CS13] NB Chapalkar and Sayali Sandbhor. Use of artificial intelligence in real property valuation. *International Journal of Engineering Technology*, 5(3):2334–37, 2013.

[EQR98] Peter Englund, John M Quigley, and Christian L Redfearn. Improved price indexes for real estate: measuring the course of swedish housing prices. *Journal of Urban Economics*, 44(2):171–196, 1998.

[GT03] Allen C Goodman and Thomas G Thibodeau. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3):181–201, 2003.

[Kau03] Tom Kauko. On current neural network applications involving spatial modelling of property prices. *Journal of housing and the built environment*, 18(2):159–181, 2003.

[KHH02] Tom Kauko, Pieter Hooimeijer, and Jacco Hakfoort. Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies*, 17(6):875–894, 2002.

[Koz17] Damian Kozbur. Testing-based forward model selection. *American Economic Review*, 107(5):266–69, 2017.

[KTTT] Olgun Kitapci, Ömür Tosun, Murat Fatih Tuna, and Tarik Turk. The use of artificial neural networks (ann) in forecasting housing prices in ankara, turkey.

[LGL04] Visit Limsombunchao, Christopher Gan, and Minsoo Lee. House price prediction: hedonic price model vs. artificial neural network. *American Journal of Applied Sciences*, 1(3):193–201, 2004.

[Lim04] Visit Limsombunchao. House price prediction: hedonic price model vs. artificial neural network. 2004.

[LT94] Richard E Lenski and Michael Travisano. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences*, 91(15):6808–6814, 1994.

[LWJ01] Owen M Lewis, J Andrew Ware, and David Harrison Jenkins. Identification of residential property sub-markets using evolutionary and neural computing techniques. *Neural Computing & Applications*, 10(2):108–119, 2001.

[Mar09] Stephen Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, 1st edition, 2009.

[NSW08] S Thomas Ng, Martin Skitmore, and Keung Fai Wong. Using genetic algorithms and linear regression analysis for private housing demand forecast. *Building and Environment*, 43(6):1171–1184, 2008.

[OS16] Timothy Oladunni and Sharad Sharma. Hedonic housing theory—a machine learning investigation. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 522–527. IEEE, 2016.

[PB15] Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934, 2015.

[PGGP15] Vasilios Plakandaras, Rangan Gupta, Periklis Gogas, and Theophilos Papadimitriou. Forecasting the us real house price index. *Economic Modelling*, 45:259–267, 2015.

[Pry13] Gwilym Pryce. Housing submarkets and the lattice of substitution. *Urban Studies*, 50(13):2682–2699, 2013.

[QTNH] BUI Quang-Thanh and DO Nhu-Hiep. House price estimation in hanoi using artificial neural network and support vector machine: in considering effects of status and house quality.

[Ros74] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974.

[Rot64] Gian-Carlo Rota. The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498–504, 1964.

[RT13] Bill Randolph and Andrew Tice. Who lives in higher density housing? a study of spatially discontinuous housing sub-markets in sydney and melbourne. *Urban Studies*, 50(13):2661–2681, 2013.

[SF13] Ehsan Shekarian and Alireza Fallahpour. Predicting house price via gene expression programming. *International Journal of Housing Markets and Analysis*, 6(3):250–268, 2013.

[WH15] Heyong Wang and Ming Hong. Study on residential hedonic price classification model based on mdlp binning and support vector machine. *information Technology Journal*, 2015.

[WWZW14] Xibin Wang, Junhao Wen, Yihao Zhang, and Yubiao Wang. Real estate price forecasting based on svm optimized by pso. *Optik-International Journal for Light and Electron Optics*, 125(3):1439–1443, 2014.

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 collect housing market data, and produce house price indices. Preparation of the data set is summarized in table 12.[28]

6.2. Grid size and number of cells with housing market transactions

6.3. Genetic variation tables grid approach

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

Table 12: Dataset Preparation

Data operation	Number of sales
All transactions	$98,599$
Postcode between 100-1299	$98,580$
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 train set	$8,400$
Observations in validation set	$2,836$
Observations in test set	$2,800$

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

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 times 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

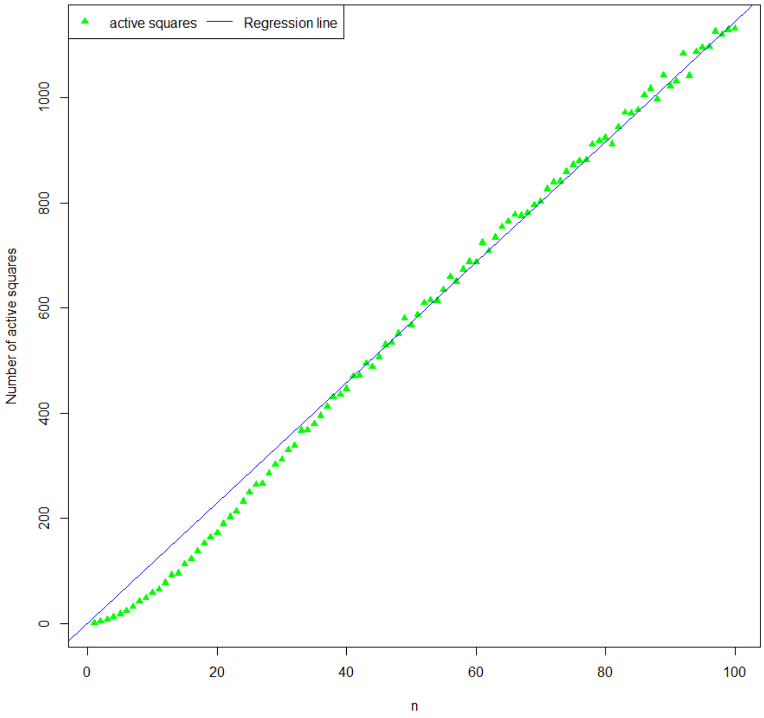

Figure 10: The number of squares with housing market transactions. Regression line without constant term, coefficient 11.4 and $R^2 = 99.8$ percent.