

PAPER2

ROY E. LOWRANCE, MUSTAFA ANIL KOCAK, ANDREAS MUELLER, YANN LECUN,
AND DENNIS SHASHA

ABSTRACT. TODO: write me

CONTENTS

1. Prior Work and Contribution of This Study	5
1.1. Features Often Used in Predicting Prices	5
1.1.1. [GT07]	6
1.1.2. [CM11]	7
1.1.3. [CIW17]	7
1.2. Impact On Value of Specific Features	8
1.2.1. Financial Features of the Property or Its Neighbors	8
1.2.2. School-Related Features	9
1.2.3. Climate and Environment-Related Features	9
1.2.4. Nearness To Water	10
1.2.5. Commuting Features	11
1.2.6. Other Features of Neighborhoods	12
1.2.7. Features of The Sales Transaction and Selling Process	13
1.2.8. Other Features Considered	14
1.3. Insights on Model Design	14
1.4. Insights on Downturns and Upturns in Prices	18
1.5. Contributions of the present study	19
1.5.1. Linear Models	20
1.5.2. Non-linear Models	23
1.5.3. Ensemble Model	23
2. Data Preparation	24
3. Linear Models	24
4. Nonlinear Models	24
5. Feature Importance	24
6. Model Hyperparameter Selection	24
7. Comparing the Best Linear and Nonlinear Models	24
8. Technical Appendix	24

TODO: write an introduction


1. PRIOR WORK AND CONTRIBUTION OF THIS STUDY

This literature review covers primarily articles about residential real estate values that were published from 2007 until the first quarter of 2018 in these journals:

- The Journal of Real Estate Finance and Economics
- Journal of Real Estate Research.
- Real Estate Economics.

A few articles references in the above but not in the three primary sources are also included in this review.

In the remainder of this section, we report on the features often used in predicting residential real estate prices, the impact on value of specific features, what submarkets were considered, the overall design of the models, and insights on handling of downturns and upturns in prices. We conclude with the contributions of this work.

1.1. Features Often Used in Predicting Prices.  Many studies used linear models of the form $\log(\text{price}) = \beta_0 + \sum_i \beta_i * \text{feature}_i$. The features used were often attributes of the property itself, attributes of the neighborhood containing the property, and indicator variables for time periods. The indicators were used to allow the model to adjust for trends in prices over time.

We summarize below the features that were use in three papers that we found to be typical of recent practice. We do not report on the β values, as these are sometimes surprising: a negative β can arise that appears to be non-economical in part because of omitted variables and omitted interactions among variables. For example, in some studies the coefficient for the number of bedrooms would be negative, possibly because given a fixed amount of interior space, more bedrooms mean smaller bedrooms.


Many of the house features come from the tax assessor's files. The files contains descriptions of the properties and their improvements. The files are used

to generate property tax bills. The property description may contain the census block identifier. When it does, additional features from the relevant census can be generated. Likewise, the property description may contain the school district, allowing information about school quality to be generated.

The other relevant data set is from the recorder of deeds. The files contains descriptions of the financial transaction, and, in many jurisdictions, the price, whether the transaction was at arms length, and whether the price was the entire consideration.



1.1.1.1. [GT07]. In [GT07], $\log(\textit{price})$ was estimated and a list of house features was given. These were:

- square feet of living area 
- log of square feet of living area
- square feet of servant's quarters
- log of square feet of servant's quarters (zero if there were no servant's quarters)
- dwelling age in years
- dwelling age in decades
- dwelling age in years squared
- dwelling age in years cubed
- number of bathrooms
- whether the house had central heating
- whether the house had non-central gas heating
- whether the house had another heating system
- whether the house had no air conditioning
- whether the house had window air conditioning
- whether the house had a wet bar
- whether the house had at least one fireplace
- whether the house had at least one pool
- whether the house had an attached garage

- whether the house had an attached carport
- whether the house had no covered parking facility
- three dummy variables: whether the house sold in quarter 1, 2, or quarter 3 (quarter 4 was the base category).

1.1.2. [CM11]. In [CM11], $\log(\textit{price})$ was estimated and a list of house features was given. These were:

- age of the house
- age of the house squared
- number of bedrooms
- number of bathrooms
- square footage of the house
- square footage of the lot.

[CM11] used these neighborhood features:

- distance from house to downtown
- natural log of distance from house to downtown,
- distance from house to nearest freeway
- natural log of distance from house to nearest freeway
- elementary school's based Academic Performance Index
- distance from house to coast
- natural log of distance from house to coast
- dummy variables derived from distance to coast.

1.1.3. [CIW17]. In [CIW17], a list of house features was given. These were:

- age of the house
- log of square footage of the house
- number of bathrooms
- number of bedrooms
- acreage
- whether the house has at least one garage

- whether the house has a walk-in closet
- whether the house has air conditioning
- whether the house has central air conditioning
- whether the house has a lake view
- whether the house has a pool
- whether the house was in fair condition
- whether the house was in good condition.

1.2. Impact On Value of Specific Features. Many authors studied the impact on value of specific features. All used transactions for particular geography and a particular time period. We report on the directional impact of these features and generally elide the details on place and time, as in many cases, the directional impact would seem to be generalizable. However, those who wish to incorporate these features into their models should of course test whether they hold for the specific spaces and times relevant to their work. This review blends together features for houses and condos as our focus in this section is on residences.

In general, these features are harder to construct than features that can be directly pulled from tax assessors, records of deeds, and the census.

Our review is organized by kind of feature and within kind, in increasing order of date of publication since 2007. One can perhaps spot trends in features that were of concern at various time periods.

1.2.1. Financial Features of the Property or Its Neighbors. Features related to financial status of the subject property or its neighbors:

- [LRY09] claimed that a foreclosure depressed prices on other houses that were near in space and time to the foreclosed house.
- [IM16] claimed that having REO in the neighborhood reduces prices of other residences.
- [CIW17]) claimed that prices for REO (foreclosed) properties were at a discount to market prices.



- [Li17] claimed that “foreclosed properties depress neighboring property prices.”
- [RRSW17] claimed that “a property initially sold as a real estate owned (REO) property . . . [later sells] at a market price.” The return of the property to normal price levels was attributed to improvements made to it subsequent to the REO sale.

1.2.2. *School-Related Features.* Features related to schools:

- [Car08] claimed that the introduction of court-ordered school busing did “not have a significant effect on house prices.”
- [ZHT08] claimed that changes in school quality affected both house prices and liquidity.
- [SS09] claimed that school quality as measured by school district ratings and performance indices are “readily capitalized into housing prices.”
- [SCN16] claimed that residences near a school face a price penalty.
- [SZ16] claimed that higher school quality is associated with higher prices.
- [BI18] claimed a price premium for school quality.

1.2.3. *Climate and Environment-Related Features.* Feature related to climate and other environmental concerns:

- [VB08] claimed that community gardens had a significant positive effect on housing prices, “especially in the poorest neighborhoods” and when the garden had higher quality.
- [AH09] claimed that home values were increased when the homes were proximate to trails, greenbelts, trails with greenbelts, neighborhood playgrounds, tennis courts, neighborhood pools views, and cul-de-sacs.
- [MLGC09] claimed that repeated forest fires reduced prices for houses near the fires.

- [CLW⁺10] claimed that “neighborhood greenspace at the immediate vicinity of houses had a significant impact on house prices even after controlling for spatial autocorrelation.”
- [HWC⁺11] claimed that “neither the view of [a] wind facility nor the distance of the home to [that] facility [has] a statistically significant effect on sales prices.”
- [Asa14] claimed that having permanent open spaces in clustered residential developments increases value.
- [FAMM02] claimed that research had not confirmed an “unambiguous positive relationship between housing prices and air quality.”
- [ZCKS14] claimed that air pollution reduced value.
- [BBS15] claimed that “the rate of land erosion negatively affect coastal residential property values” when “the ratio of the property’s distance from the shore to the rate of erosion is sufficiently low.”
- [FSY15] claimed that Leadership in Environmental Design (LEED) certification for buildings “adds a premium to condo sales prices” and that LEED certification for a neighborhood “fails to add value for condo buyers.”
- [SSR15] claimed that refinery “air pollution has a significant negative . . . on house prices” and that the effect generally diminishes with distance from the refinery.
- [HAP16] claimed that wind turbines do not impact house prices though electric transmission lines decrease prices and open spaces increase prices.
- [VP16]) claimed that prices adjust quickly to redrawn flood risk maps.

1.2.4. *Nearness To Water*. Many studies published prior to 2007 found that properties near water carried a price premium. Below are studies from 2007 on:

- [Ude10] claimed that views of lagoons in Nigeria had “a statistically significant impact on home values even after adjusting for other significant home value determinants.”

- [CM11] claimed that houses within six miles of the coast in San Diego Country had higher values than others.
- [RLvM17] claimed that residences in planned developments in the Netherlands near water carried a price premium and estimated that the value of this premium was lower than other studies had found.
- [SM18] claimed that properties on waterfront lots in 2018 in the United States still had substantial premiums that varied whether the water was an ocean, a lake, or a river in spite of the possible negative influences of “global climate warming and greater flood risks.” This work notes that “few studies have dealt with elevation risks from sea level rise” and claims evidence (to be published later) “the possibility of a decreasing premium for waterfront when the risks of floods are perceived as high or increasing.” Prior research, some before 2007, claimed that:
 - “Rivers and lakefronts are not as valuable as ocean fronts”
 - “Oceanfront sites with waves are highly valued”
 - “Larger lakes are better than smaller lakes”
 - “The greater the radius of unobstructed view the better”
 - “Waterfront premiums decline rapidly after 60 to 100 meters.”

1.2.5. *Commuting Features.* Features related to commuting to and from work:

- [DPR07] claimed that residence prices are higher where there is close enough proximity to a commuter railway station, especially when highways are not easily available.
- [PGB09] claimed that being within 100 meters of a train reduced apartment prices, being within 100 to 150 meters increased prices, and being further away decreased prices. The work was in Haifa, an urban environment in which commuting by train was important.
- [CRC11] claimed that while new highways are being constructed, values are lower the closer to the construction, that prices are not reduced appropriately in the period before construction starts, and that post

construction, values are increased at moderate distances from the new highway.

- [DCS16] claimed that houses sufficiently close to railway lines increased in value after the line agreed to cease operations.

1.2.6. *Other Features of Neighborhoods.* Other neighborhood features used in studies in our universe:

- [PBCR98, Liu13] claimed that the extent of spatial and temporal price dependencies were important.
- [MT07] claimed that “in auto-oriented developments, a more gridiron-like street pattern reduced house values.” In pedestrian-oriented developments, the impact of street layout on house values was ambiguous.
- [WW08] claimed that planting trees in an inner city environment increased property values for residences sufficiently close to the trees.
- [Leg10] claimed “that an increase in average house size of the eight nearest neighbors and [in] the largest houses in the district [had] a negative effective on predicted house price.” This effect was inverted for the ninth to sixteenth neighbors and for the smallest houses in the district.
- [Rea10] claimed that landfills that accept more than 500 tons of waste per day decreased adjacent property values and that the decrease diminishes as the distance from the landfill increased. Some lower volume land fills were claimed to “not impact nearby property values.”
- [NK11] claimed that the market values positively historic quality, and that the market values negatively the effects of regulations that control historic districts.
- [PF11] claimed that apartments and other commercial properties with a high Walk Score (properties in neighborhoods that are walkable) had a higher price.
- [SSZ11] claimed that higher residence “values in a subdivision may results from smaller blocks, interconnected greenways, and a single entrance.”

- [YCK12] claimed that property values rose for residences near but not too near to newly-constructed shopping centers.
- [HMM13] claimed that larger Gulf-of-Mexico views were associated with higher prices.
- [HMM13] claimed that residences purchased earlier in the life of a subdivision were bought for lower prices than those purchased later.
- [WWB14] claimed that proximity to a registered sex offender decreased property values. The decrease was larger for houses with more bedrooms and when the offender was designated as violent.
- [ZHG14] claimed that historic designation increased values.
- [RSS15] claimed that single-family “homes in gated communities carry significant price premiums relative to similar homes in non-gated communities.”
- [FhY16] claimed that proximity to certain types of open spaces increased residence values during boom periods and decreased residence values during bust periods.
- [Zab16] claimed that more accurate pricing estimates were obtained by using vacancy rates as well as residence features.

1.2.7. *Features of The Sales Transaction and Selling Process.* Features of the sales transaction and selling process:

- [BRA12] claimed that estate sales were associated with lower prices.
- [IM12] claimed that prices were higher when the purchaser was someone who lived distant to the property and when the purchaser came from a market with higher prices.
- [CLAH13] claimed that restricting purchases in condominiums to seniors reduced values.
- [AH14] claimed that sales at a foreclosure carried a 20 percent discount and that short sales carried a 13 percent discount.

- [ACRR15] claimed that prices were increased when brokers hold “public open houses, broker open houses, MLS virtual tours, and [post] MLS photographs.”
- [SGZH16] claimed that once a property is taken off the market by delisting it, the price was maximized by relisting it with the same agent with 30 days.
- [CIW17] investigated value differences associated with whether the house that was sold was acquired through a foreclosure auction by the lender, whether it was acquired through a foreclosure auction by a third party, and whether it was acquired through a foreclosure and sold to a third party.
- [GW17] claimed that reducing the listing price was associated with even larger selling price reductions.

1.2.8. *Other Features Considered.*

- [TZH11] claimed that vacant houses sell for less.
- [LAC13] claimed that allowing pets in condominiums increased values.
- [BS17] claimed “a significant price premium for housing with neo-traditional architecture” in the Netherlands for recently-developed homes.

1.3. Insights on Model Design. Researchers have found that training models on submarkets in a city rather than an all transactions in a city has led to improved predictions. Using submarkets has been found to be at least as accurate as using the more complicated-to-use spatial approaches. Approaches to defining submarkets algorithmically have been proposed. These approaches seek to avoid requiring a human expert to define the relevant submarkets. Another systematic finding has been that non-linear models are more accurate than linear models. This section explores the major findings regarding these aspects of model design.

[BHP03] claimed that accuracy of hedonic price predictions was improved by using dummy variables to indicate submarkets. Defining submarkets as neighborhoods which government appraisers considered to be “relatively homogeneous” was found to lead to more accurate models than “a statistically-generated aspatial classification” derived using principle component analysis based on characteristics of the properties.

[TSY07] claimed that defining and using submarkets “can improve the precision of price predictions by 17.5 percent.” The study defined submarkets using the correlations among the residuals of a global hedonic model, which is a spatial autocorrelation approach.

[BCH07] claimed that including 33 submarket dummy variables in an ordinary least squares model resulted in more accurate models than using geostatistical or lattice methods. That “conclusion is of practical importance as submarket dummy variables [are] substantially easier to implement than spatial statistical methods.” They claimed that omitting the submarket dummy variables from the OLS models reduced their accuracy to below the accuracy of the geostatistical models. The submarkets were defined by appraisers (not an algorithm) as “geographical neighborhoods within which house values are considered to be interdependent.”

[GT07] compared the accuracy of predictions from hedonic models using two definitions of submarkets. The first definition consolidated census block groups based on median per-square-foot prices and living space. The submarkets were built in two steps. In the first step, the percentile values for median house prices were determined in each census block group. Each census block was assigned to its median-percentile group, yielding 100 groups. In the second step, these 100 groups were subdivided by range of square feet of living space to yield 324 submarkets. The census block groups in these submarkets were not necessarily near to each other.

The second definition consolidated census block groups that were in the same school district and the same municipality. A submarket was grown until it had at least five sale transactions per parameter in the linear hedonic model. The census block groups in these submarkets were near to each other. A total of 372 submarkets were defined.

Hedonic models were trained on the subsets of the data defined by the submarkets. Thus about 700 models were trained.

The study claimed that the accuracy of hedonic models that used the two definitions was about equal as measured by mean error, mean absolute error, mean proportional error, and fraction of estimates within 20 percent. The study claimed that implementing the first definition was both faster and less costly.

[GZL08] explored an “ANFIS” model which combined a neural network and a fuzzy logic regressor in predicting prices for residential properties. According to the paper, neural networks were first proposed for this application in 1992 ([DG92], fuzzy logic was first proposed in 1995 ([Bym95]), and the combined system was first proposed in 1998 ([SG98]). The proposed model is was a classification model on the properties followed by a neural network that used a property’s class as one of the features. The ANFIS model was implemented by the fuzzy logic toolbox from Mathworks. The paper claimed that the accuracy of the ANFIS model was about equal to that of a standard multiple regression model.

[LVWW08] proposed a “replication method” to predict property values. The method predicts the price of a query property as a weighted average of the price of k comparable properties. The weight vector w was required to be such that when used to linearly combine the features of the comparables, the features of the query property are obtained. That requirement allows for multiple weight vectors, so the weight vector that is chosen is the one for which the mean price prediction was zero (so that the prediction is unbiased) and the variance of the price predictions was minimized. These constraints are claimed to uniquely determine w . The replication method was claimed to be advantaged where “the

analyst finds a statistically significant correlation in the prediction errors.” No empirical work was reported. (We note that a correlation in the prediction errors was found by others to occur if prices were correlated spatially, which is regarded as most often true.)

[ZZS08] claimed “that the purchasers of higher-prices homes [valued] certain housing characteristics such as square footage and the number of bathrooms differently from buyers of lower-priced homes.” (We note that a submarket approach may capture such differences.)

[KB09] claimed that in all U.S. regions except for the Midwest, a Smooth Transition Autoregressive model based on nonlinear properties of housing prices performed better than a linear model. The tested global linear models were less accurate than the tested spatial models.

[PF09] claimed that that artificial neural networks were more accurate the linear hedonic pricing models.

[BCH10] compared approaches for incorporating spatial dependence into hedonic models. The work claimed that a geostatistical model with disaggregated submarket variables performed better than an OLS model using typical features and an OLS model that incorporated as variables the residuals of the ten nearest neighbors. The geostatistical approach assumed that the covariance in price at two locations depends only on the distance between the two locations. The submarkets were defined by combining census block groups with similar median house values to build a submarket that had at least 200 transactions. For the OLS model, the most accurate predictions were obtained by using a single equation and hence dummy variables for the submarkets. The OLS model that used the nearest neighbor residuals performed slightly better than the OLS model not using the residual as features.

In [ZFR11], accuracy of predictions were improved by modeling spatial autocorrelations that could vary with direction.

In [ZSG11], linear regression was compared to neural networks and to “nontraditional regression methods”: M5P trees (decision trees with linear regression models at the leaves), additive regression (in which new models are added sequentially to an ensemble), SM-SMO regression (a support vector machine fit using sequential minimal optimization), RBFNN (a neural network in which each hidden unit implements a radial activation function), and MBR 10 (a 10-nearest neighbor regressor). The study claimed that the nontraditional methods were more accurate than both the linear regression and neural network methods.

[KHP15] claimed that for condominium prices in Hong Kong features of the property had a non-linear effect on the price quantiles of the property. For example, “an increase in the size of the gross floor area [was] more valuable at higher [price] quantiles.”

1.4. Insights on Downturns and Upturns in Prices. This section reviews the research we found that addressed the decline in real estate prices starting in mid 2007 and research that investigated lower bounds on house prices.

In [ET07], a model for residential housing market cyclical dynamics was developed. That model directly estimated supply and demand. The study claimed that “fundamentals, such as employment growth and interest rates are key determinants of the residential real estate cycle.”

In [Mil08], forecasting price levels during boom and bust cycles was studied. Generalized autoregressive (GAR) models were claimed to outperformed autoregressive-moving average (ARMA) and generalized autoregressive conditional heteroscedastic (GARCH) models “in many cases, especially in those markets traditionally associated with high home-price volatility.”

In [GD10], a Bayesian Vector Autoregressive model was used to forecast quarterly price levels in the next quarter for 20 states. This model was found to do a “fair” job in predicting the 2007 price downturn in 18 of the states.

In [BBD17], a lower bound based on investor incentives was developed for housing prices. This lower bound was claimed to be tight in that it infrequently overstated price declines in the 2007 downturn.


1.5. Contributions of the present study. Predicting housing prices has moved through three generations into an emerging fourth generation. In the first generation, linear models were used to predict *price* or $\log(\textit{price})$. The very early work described how to fit a linear model and how to judge its accuracy. Then, leveraging the correlation of prices spatially, a wave of innovation introduced second-generation spatial models that generally were more accurate than the first-generation linear models. More accuracy came at the expense of harder-to-understand and harder-to-fit models. The third generation was built around defining models for submarkets, with the idea that since prices were correlated spatially and that the extent of correlation was defined in part by fixed boundaries like highways and other permanent features of the landscape, define the spaces with high correlations before the models were fit. This third generation of models was easier to understand and fit than the second generation spatial models, and they were at least as accurate. The fourth generation is just starting: it is defined by using non-linear models. Research on non-linear models is sparse, but to date, researchers have found that non-linear models outperform linear models at the cost of being more complex than linear models. With increased complexity comes a decrease in explainability, which may limit the effectiveness of non-linear models in some applications such as real estate taxation, where the ability to explain to a tax payer how her tax bill was calculated may be more important than an accurate assessment.

In parallel to the movement through generations of model forms, the feature sets have been enriched. Many modelers started with tax roll data to obtain physical descriptions of properties. Some have augmented these descriptions with information from multiple listing services, as these services may have more accurate features for residences that they list. Prices have often come

from deeds records. Features of neighborhoods have been induced from census records and other sources.

NOTE TO READERS OF EARLY DRAFTS: THESE ARE HYPOTHETICAL CONTRIBUTIONS. WE HOPE TO DEMONSTRATE THEM.

We have made three sets of contributions using a large dataset of transactions from Los Angeles County starting in 2003 and ending in the first quarter of 2009. (TODO: Update to actual time period, possible ending 2011, before the 2010 census results became available.) This period includes the period before the real estate crash in mid 2007 and the period during the crash up to the start of the recovery in early 2009.

- We revisited linear models to systematically define and select the most accurate models. 
- We updated the literature on non-linear models to include easy-to-use non-linear models by testing a set of models provided by Scikit-learn, a popular open source machine learning library for Python.
- We developed a way to use an ensemble model that made the model-selection process self tuning. The ensemble model considered a large number of both linear and non-linear models and historic training periods, and then blended their predictions based on their recent accuracy. We assessed the extent to which more volatility in prices (as at the start and end of the 2007 - 2009 pricing crisis) led to certain types of models and to certain length training periods. In particular, we found that the ensemble model automatically shortened training periods when volatility was high and lengthened them when volatility was low.

In the remainder of this section, we explore these contributions in more depth.

1.5.1. *Linear Models.* Our first set of contributions was around designing linear models. We carried out a set of experiments designed to reveal the most accurate

design for linear models. We used three readily available feature sets, which were the the tax accessor’s data, the deeds, and the U.S. census.

These experiments considered a wide range of design choices for linear models.

- Whether to predict the *price* or the $\log(\textit{price})$. Many linear model builders have predicted the $\log(\textit{price})$ and we sought to justify that choice.
- Whether to transform the hedonic features that measure size into the log space. If a hedonic feature measures size and possibly doubling it could double the price, then a linear model that estimates $\log(\textit{price})$ using $\log(\textit{feature})$ would be appropriate for that feature. For example, doubling the square footage of a house could double its value. Most of the literature does not consider this type of transformation on the features, but doing so is a typical practice in machine learning. We did not evaluate other systematic transformations, such as including as a feature the square of the age of the property, even though it is common to square that particular feature. We left that systematic investigation to future work, where one approach could be to consider a range of transformation of the raw features, perhaps raising each to a power, perhaps multiplying each by another. (As an example of what could be explored, Vowpal Wabbit [Lan] contains invocation parameters that will systematically transform input features into the cross-products of subsets of those features.)
- To what extent to regularize the linear models. We evaluated including both an L1 and an L2 regularizer in the optimization objective. Most of the work in real estate around linear models has not regularized the models and the machine learning literature often regularizes linear models.
- Using readily available features of the house, neighborhood, and census tract, rather than construction of the many other feature claimed in the literature to improve prediction accuracy. We decided to focus on model design, not feature engineering.

- We considered defining submarkets by using census-based definitions, property city definitions, and city name definitions. For each, we explored the impact on prediction accuracy of further subdividing the market based on residence size, following the practice reported on in [BHP03]. (DENNIS: If we use Collateral Analytics data, we may be able to use their definitions of neighborhoods and compare accuracy using their definitions to other definitions in the literature.)
- We compared two approach to including submarkets in the models. One was by using dummy variables to capture the fixed effects of each submarket. The other was restricting the training set to samples just in the submarket.
- We compared two approaches to handling changing price levels. One was by using dummy variables for time periods. The other was a perhaps new-to-real estate method to test and train models on a rolling basis. In this method, we, for example, train models using data up to the end of January 2003 and use these trained models to predict prices in February 2003. In the security pricing literature, this type of process is called “walk forward.”
- In order to support replicability of our findings, we used the Python library Scikit-learn [PVG⁺11] to implement all of the models we tested and have provided GPL-licensed source code all for all of our code in one of the first author’s github account, under the account name *rlowrance*..
- For each set of design choices, we retain all predictions for all properties. We can then report on the distribution of errors and multiple summary statistics of those distributions. Other authors have noted that the most accurate model depends in part how errors are measured, and we sought to make our work extensible to new error measurements that might be driven by specific loss functions in applications.

1.5.2. *Non-linear Models.* Our first set of contributions arrived at a “best” linear model and related design practices that used open software and the most readily available data sets. We then updated the work of [ZSG11] to compare this model to more recently available non-linear models including random forests, gradient boosting, non-linear SVMs, and neural networks using Scikit-learn implementations. We compared the non-linear models to the best linear model and determined the best non-linear model.

For that model, we then reviewed its fitting data to understand what features it found to be important. We report on the best settings for the hyperparameters of this model.

1.5.3. *Ensemble Model.* In the first two parts of our work, we developed a best linear model and a best non-linear model. Subsequently, we developed an ensemble model that used all the models. We determined that training only on more recent data led to more accuracy when prices were more volatile, and that training on data that included transactions further back in time led to more accuracy when prices were less volatile. We determined that in less volatile pricing environments, some linear models trained on a long history of prices performed better, and in more volatile markets, some non-linear models trained only on recent prices performed better. (NOTE: That’s a working hypothesis.)

This work arrived an a self-tuning ensemble model that selected the linear and non-linear used by the ensemble model, their hyperparameters, and defined the extent of historic transactions used to train the model. All the work up to this point had been done without examining randomly-selected hold-out data. We then tested the ensemble model fitting and prediction procedure using the hold-out data.



2. DATA PREPARATION

3. LINEAR MODELS

4. NONLINEAR MODELS

5. FEATURE IMPORTANCE

6. MODEL HYPERPARAMETER SELECTION

7. COMPARING THE BEST LINEAR AND NONLINEAR MODELS

8. TECHNICAL APPENDIX

REFERENCES

- [ACRR15] Marcus T. Allen, Anjelite Cadena, Jessica Rutherford, and Ronald C. Rutherford, *Effects of real estate brokers' marketing strategies: Public open houses, broker open houses, mls virtual tours, and mls photographs*, *Journal of Real Estate Research* **37** (2015), 343–369.
- [AH09] Paul K. Asabere and Forrest E. Huffman, *The relative impacts of trails and greenbelts on home price*, *The Journal of Real Estate Finance and Economics* **38** (2009), 408–419.
- [AH14] Ramya Rajajagadeesan Aroul and J. Andrew Hansz, *The valuation impact on distressed residential transactions: Anatomy of a housing price bubble*, *The Journal of Real Estate Finance and Economics* **49** (2014), 277–302.
- [Asa14] Paul K. Asabere, *The value of homes in cluster development residential districts: The relative significance of the permanent open spaces associated with clusters*, *The Journal of Real Estate Finance and Economics* **48** (2014), 244–225.
- [BBD17] Alexander N. Bogin, Stephen D. Bruestle, and William M. Doerner, *How low can house prices go? estimating a conservative lower bound*, *The Journal of Real Estate Finance and Economics* **54** (2017), 97–116.
- [BBS15] Scott Below, Ali Beracha, and Hilla Skiba, *Land erosion and coastal home values*, *Journal of Real Estate Research* **37** (2015), 499–534.
- [BCH07] Steven C. Bourassa, Evan Cantoni, and Martin Hoesli, *Spatial dependence, housing submarkets and house price prediction*, *Journal of Real Estate Finance and Economics* **35** (2007), no. 2, 143–160.

- [BCH10] Steven C. Bourassa, Eva Cantoni, and Martin Hoesli, *Predicting house prices with spatial dependence: A comparison of alternative approaches*, *Journal of Real Estate Research* **32** (2010), no. 2, 139–159.
- [BHP03] Steven C. Bourassa, Martin Hoesli, and Vincent S. Peng, *Do housing submarkets really matter?*, Tech. report, Universite De Geneve, Geneva, Switzerland, 2003.
- [BI18] Eli Beracha and William G. Harding III, *The capitalization of school quality into renter and owner housing*, *Real Estate Economics* **46** (2018), 85–119.
- [BRA12] Justin D. Benefield, Ronald C. Rutherford, and Marcus T. Allen, *The effects of estate sales of residential real estate on price and market timing*, *The Journal of Real Estate Finance and Economics* **45** (2012), 965–981.
- [BS17] Edwin Buitelaar and Frans Schilderf, *The economics of style: Measuring the price effect of neo-traditional architecture in housing*, *Real Estate Economics* **45** (2017), 7–27.
- [Bym95] P. Byrne, *Fuzzy analysis: A vague way of dealing with uncertainty in real estate analysis*, *Journal of Property Valuation & Investment* **13** (1995), 22–41.
- [Car08] Steven Carter, *Court-ordered busing and housing prices: The case of pasadena and san mario*, *Journal of Real Estate Research* **30** (2008), 377–393.
- [CIW17] Peter Chinloy, William Harding III, and Zonghua Wu, *Foreclosure, reo, and market sales in residential real estate*, *The Journal of Real Estate Finance and Economics* **54** (2017), 188–215.
- [CLAH13] Charles C. Carter, Zhenguo Line, Marcus T. Allen, and William J. Haloupek, *Another look at effects of "adults-only" age restrictions on housing prices*, *The Journal of Real Estate Finance and Economics* **46** (2013), 115–130.
- [CLW⁺10] Delores Conway, Christina Q. Li, Jennifer Wolch, Christopher Kahle, and Michael Jerrett, *A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values*, *The Journal of Real Estate Finance and Economics* **41** (2010), 150–169.
- [CM11] Stephen J. Conray and Jennifer L. Milosch, *An estimation of the coastal premium for residential housing prices in san diego country*, *The Journal of Real Estate Finance and Economics* **42** (2011), 211–228.
- [CRC11] Ekaterina Chernobai, Michael Reibel, and Michael Carney, *Nonlinear spatial and temporal effects of highway construction on house prices*, *The Journal of Real Estate Finance and Economics* **42** (2011), 348–370.

- [DCS16] Mi Diao, Yu Chin, and Tien Foo Sing, *Negative externalities of rail noise and housing values: Evidence from the cessation of railway operations in singapore*, *Real Estate Economics* **44** (2016), 878–917.
- [DG92] Q. Do and G. Grudnitski, *A neural network analysis of residential property appraisal*, *Real Estate Appraiser* **58** (1992), 38–45.
- [DPR07] Ghebreegziabihir Debrezion, Eric Pels, and Piet Rietveld, *The impact of railway stations on residential and commercial property values: A meta-analysis*, *The Journal of Real Estate Finance and Economics* **35** (2007), 161–180.
- [ET07] Robert H. Edelstien and Desmond Tsang, *Dynamic residential housing cycle analysis*, *The Journal of Real Estate Finance and Economics* **35** (2007), 295–313.
- [FAMM02] Gema Fernandez-Aviles, Roman Minguez, and Jose-Maria Montero, *Geostastistical air pollution indexes in spatial hedonic models: The case of madrid, spain*, *Journal of Real Estate Research* **34** (2002), 243–274.
- [FhY16] Qin Fan, J. Andrew hanzs, and Ziaoming Yang, *The pricing effects of open space amenities*, *The Journal of Real Estate Finance and Economics* **52** (2016), 244–271.
- [FSY15] Julia Freybote, Hua Sun, and Xi Yang, *The impact of leed neighborhood certification on condo prices*, *Real Estate Economics* **43** (2015), 586–608.
- [GD10] Rangan Gupta and Sonali Das, *Predicting downturns in the us housing market: A bayesian approach*, *The Journal of Real Estate Finance and Economics* **41** (2010), 294–319.
- [GT07] Allen C. Goodman and Thomas G. Thibodeau, *The spatial proximity of metropolitan area housing submarkets*, *Real Estate Economics* **35** (2007), 209–232.
- [GW17] Bruce L. Gordon and Daniel T. Winkler, *The effect of listing price changes on the selling price of single-family residential homes*, *The Journal of Real Estate Finance and Economics* **55** (2017), 185–215.
- [GZL08] Jian Guan, Jozef Zurada, and Alan S. Levitan, *An adaptive neuro-fuzzy inference system based approach to real estate property assessment*, *Journal of Real Estate Research* **30** (2008), 395–421.
- [HAP16] Ben Hoen and Carol Atkinson-Palombo, *Wind turbines, amenities and disamenities: A study of home value impacts in densely populated massachusetts*, *Journal of Real Estate Research* **38** (2016), 473–504.
- [HHM13] Paul Hindsley, Stuart E. Hamilton, and Ashton Morgan, *Gulf views: Toward a better understanding of viewshed scope in hedonic property models*, *The Journal of Real Estate Finance and Economics* **47** (2013), 489–505.

- [HMM13] Harris Hollans, Richard W. Martin, and Henry J. Munneke, *Measuring price behavior in new residential subdivisions*, The Journal of Real Estate Finance and Economics **47** (2013), 227–242.
- [HWC⁺11] Ben Hoen, Ryan Wiser, Peter Cappers, Mark Thayer, and Gautam Sethi, *Wind energy facilities and residential properties: The effect of proximity and view on sales prices*, Journal of Real Estate Research **33** (2011), 279–316.
- [IM12] Keith Ihlanfeldt and Tom Mayock, *Information, search, and housing prices: Revisited*, The Journal of Real Estate Finance and Economics **44** (2012), 90–115.
- [IM16] ———, *The impact of reo sales on neighborhoods and their residents*, The Journal of Real Estate Finance and Economics **53** (2016), no. 3, 282–324.
- [KB09] Sei-Wan Kim and Radha Bhattacharya, *Regional housing prices in the usa: An empirical investigation of nonlinearity*, The Journal of Real Estate Economics **38** (2009), 443–460.
- [KHP15] Hyung-Gun Kim, Kwong-Chin Hung, and Sung Y. Park, *Determinants of housing prices in hong kong: A box-cox quantile regression approach*, The Journal of Real Estate Economics **50** (2015), 270–287.
- [LAC13] Zhenguo Lin, Marcus T. Allen, and Charles C. Carter, *Pet policy and housing prices: Evidence from the condominium market*, The Journal of Real Estate Finance and Economics **47** (2013), 109–122.
- [Lan] John Langford, *Vowpal wabbit*, Accessed: 2018-04-06.
- [Leg10] Susane Legulzamon, *The influence of reference group house size on house price*, Real Estate Economics **38** (2010), 507–527.
- [Li17] Lingxiao Li, *Why are foreclosures contagious?*, Real Estate Economics **45** (2017), 979–997.
- [Liu13] Xialong Liu, *Spatial and temporal dependence in house price prediction*, The Journal of Real Estate Finance and Economics **47** (2013), 341–369.
- [LRY09] Zhenguo Lin, Eric Rosenblatt, and Vincent W. Yao, *Spillover effects of foreclosures on neighborhood property values*, The Journal of Real Estate Finance and Economics **38** (2009), 387–407.
- [LVWW08] Tsong-Yue Lai, Kerry Vandell, Ko Wang, and Gerd Welke, *Estimating property values by replication: An alternative to the traditional grid and regression methods*, Journal of Real Estate Research **30** (2008), 441–460.
- [Mil08] W. Miles, *Boom-bust cycles and the forecasting performance of linear and non-linear models of house prices*, The Journal of Real Estate Finance and Economics **36** (2008), 249–264.

- [MLGC09] Julie Mueller, John Loomis, and Armando Gonzalez-Caban, *Do repeated wildfires change homebuyers' demand for homes in high-risk areas? a hedonic analysis of the short and long-term effects of repeated wildfires on house prices in southern california*, *The Journal of Real Estate Finance and Economics* **38** (2009), 155–172.
- [MT07] John W. Matthews and Geoffrey K. Turnbull, *Neighborhood street layout and property value: The interaction of accessibility and land use mix*, *The Journal of Real Estate Finance and Economics* **35** (2007), 111–141.
- [NK11] Douglas S. Noonan and Douglas J. Krupka, *Making-or picking-winners: Evidence of internal and external price effects in historic preservation policies*, *Real Estate Economics* **39** (2011), 379–407.
- [PBCR98] R. Kelley Pace, Ronald Barry, John M. Clapp, and Mauricio Rodriguez, *Spatiotemporal autoregressive models of neighborhood effects*, *Journal of Real Estate Finance and Economics* **17** (1998), no. 1, 5–13.
- [PF09] Steven Peterson and Albert B. Flanagan, *Neural network hedonic pricing models in mass real estate appraisal*, *Journal of Real Estate Research* **31** (2009), 147–164.
- [PF11] Gary Pivo and Jeffrey D. Fisher, *The walkability premium in commercial real estate investments*, *Real Estate Economics* **39** (2011), 185–219.
- [PGB09] Boris A. Portnov, Bella Genkin, and Boaz Barzilay, *Investigating the effect of train proximity on apartment prices: Haifa, israel as a case study*, *Journal of Real Estate Research* **31** (2009), 371–395.
- [PVG⁺11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011), 2825–2830.
- [Rea10] Richard C. Ready, *Do landfills always depress nearby property values?*, *Journal of Real Estate Research* **32** (2010), 321–339.
- [RLvM17] Jan Rouwendal, Or Levkovich, and Ramona van Marwijk, *Estimating the value of proximity to water, when ceteris really is paribus*, *Real Estate Economics* **45** (2017), 829–860.
- [RRSW17] Jessica Rutherford, Ronald C. Rutherford, Elizabeth Strom, and Lei Wedge, *The subsequent market value of former reo properties*, *Real Estate Economics* **45** (2017), 713–760.
- [RSS15] Evgeny L. Radetskiy, Ronald W Spahr, and Mark A. Sunderman, *Gated community premiums and amenity differentials in residential subdivision*, *Journal of Real Estate Research* **37** (2015), 405–438.

- [SCN16] Vivek Sah, Stephen J. Conroy, and Andrew Narwold, *Estimating school proximity effects on housing prices: the importances of robust spatial controls in hedonic estimations*, *The Journal of Real Estate Finance and Economics* **53** (2016), 50–76.
- [SG98] M. Sugan and G.T.Kang, *Structure identification of fuzzy model*, *Fuzzy Sets and Systems* **28** (1998), 15–33.
- [SGZH16] Patrick S. Smith, Karen M. Gibler, and Velma Zahiroic-Herbert, *The effect of relisting on house selling price*, *The Journal of Real Estate Finance and Economics* **52** (2016), 176–195.
- [SM18] Michael Sklarz and Norman Miller, *The impact of waterfront location on residential home values*, Tech. report, Collateral Analytics, March 2018.
- [SS09] Youngme Seo and Robert A. Simons, *The effect of school quality on residential sales price*, *Journal of Real Estate Research* **31** (2009), 307–327.
- [SSR15] Robert A. Simons, Younme Seo, and Paul Rosenfeld, *Modeling the effects of refinery emissions on residential property values*, *Journal of Real Estate Research* **37** (2015), 321–342.
- [SSZ11] Woo-Jin Shin, Jesse Saginor, and Shannon Van Zandt, *Evaluating subdivision characteristics on single family housing value using hierarchical linear modeling*, *Journal of Real Estate Research* **33** (2011), 317–348.
- [SZ16] Rui Wang Siqi Zheng, Wanyang Hu, *How much is a good school worth in beijing? identifying price premium with paired resale and rental data*, *The Journal of Real Estate Finance and Economics* **52** (2016), 184–199.
- [TSY07] Yong Tu, Hua Sun, and Shi-Ming Yu, *Spatial autocorrelations and urban housing market segmentation*, *The Journal of Real Estate Finance and Economics* **34** (2007), 385–406.
- [TZH11] Geoffrey K. Turnbull and Velma Zahirovic-Herbert, *Why do vacant houses sell for less: Holding costs, bargaining power or stigma?*, *Real Estate Economics* **39** (2011), 19–43.
- [Ude10] C.E. Udechukwu, *The impact of lagoon water views on residential property values in nigeria*, *Lagos Journal of Environmental Studies* **7** (2010), 34–45.
- [VB08] Ioan Voicu and Vicki Been, *The effect of community gardens on neighboring property values*, *Real Estate Economics* **36** (2008), 241–283.
- [VP16] Athanasios Votsis and Adriaan Perrels, *Housing prices and the public disclosure of flood risk: A difference-in-differences analysis in finland*, *The Journal of Real Estate Finance and Economics* **53** (2016), 450–471.

- [WW08] Susan M. Wachter and Grace Wong, *What is a tree worth? green-city strategies, signaling and housing prices*, Real Estate Economics **36** (2008), 213–239.
- [WWB14] Scott Wentland, Bennie Waller, and Raymond Brastow, *Estimating the effect of crime risk on property values and time on market: Evidence from megan’s law in virginia*, Real Estate Economics **42** (2014), 223–251.
- [YCK12] Tun-Hsiang Yu, Seong-Hoon Cho, and Seung Gyu Kim, *Assessing the residential property tax revenue impact of a shopping center*, The Journal of Real Estate Finance and Economics **45** (2012), 604–621.
- [Zab16] Jeffrey Zabel, *A dynamic model of the housing market: The role of vacancies*, The Journal of Real Estate Finance and Economics **53** (2016), 368–391.
- [ZCKS14] Siqi Zheng, Jing Cao, Matthew E. Kahn, and Cong Sun, *Real estate valuation and cross-boundary air pollution externalities: Evidence from chinese cities*, The Journal of Real Estate Finance and Economics **48** (2014), 398–414.
- [ZFR11] Bing Zhu, Roland Fuss, and Nico B. Rottke, *The predictive power of anisotropic spatial correlation modeling in housing prices*, The Journal of Real Estate Finance and Economics **42** (2011), 542–565.
- [ZHG14] Velma Zahirovic-Herbert and Karen M. Gibler, *Historic district influence on house prices and marketing duration*, The Journal of Real Estate Finance and Economics **48** (2014), 112–131.
- [ZHT08] Velma Zahirovic-Herbert and Geoffrey K. Turnbull, *School quality, house prices and liquidity*, The Journal of Real Estate Finance and Economics **37** (2008), 113–130.
- [ZSG11] Jozef Zurada, Alan S. Levitan, and Jian Guan, *A comparison of regression and artificial intelligence methods in a mass appraisal context*, Journal of Real Estate Research **33** (2011), no. 3, 349–387.
- [ZZS08] Joachim Zietz, Emily Normal Zietz, and G. Stacy Sirmans, *Determinants of house prices: A quantile regression approach*, The Journal of Real Estate Finance and Economics **37** (2008), 317–333.