# Essays on Machine Learning in Applied Microeconomics

Jenny Ho

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Patrick Bajari, Chair

Yanqin Fan

Robert Halvorsen

Program Authorized to Offer Degree:
Department of Economics

University of Washington

**Abstract**

Essays on Machine Learning in Applied Microeconomics

Jenny Ho

Chair of the Supervisory Committee:
Professor Patrick Bajari
Department of Economics

Hedonic models are commonly used to recover the implicit prices of house attributes and local non-market public goods. Yet they are plagued by omitted variable bias when variables that are correlated with the attribute in question are unobservable. The increase in availability of big data and unstructured data in the form of text and images allow for a more extensive set of variables that are relevant to consumers to be included in hedonic methods. Unstructured data are high-dimensional and require machine learning methods that are robust to multicollinearity and irrelevant variables. They can also nest previous econometric methods. In this dissertation, I show that by controlling for more home attributes, bias is significantly reduced when estimating willingness-to-pay for environmental and urban amenities. I first estimate the effects of air pollution on house prices in Pennsylvania. By incorporating a rich home transaction dataset collected from Zillow, I reduce bias by more than half. Using a similar dataset, I then estimate how minimum lot-size zoning impacts home prices in Seattle. I nest a boundary discontinuity design within ensemble tree models such as random forest and gradient boosting and find that zoning is associated with 5% increase in home prices, a number significantly smaller than estimates when limited data and standard linear models are used. Last, I demonstrate how features can be extracted from curbside view images using computer vision tools. These features can be used to model curb appeal, a home attribute that has never been included in hedonic models but is of importance to consumers.

The combination of more data and machine learning tools leads to models that are more predictive as well as significant reduction in bias when estimating treatment effects.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I would like to give special thanks to my advisor Patrick Bajari and committee members Yanqin Fan and Robert Halvorsen. I would also like to thank Pasita Chaijaroen for helpful comments.

# DEDICATION

To Mom, Dad, and my sisters Amy and Connie.

Chapter 1

# MACHINE LEARNING FOR CAUSAL INFERENCE: AN APPLICATION TO AIR QUALITY IMPACTS ON HOUSE PRICES

**Abstract**

Hedonic models are commonly used to recover the implicit prices of house attributes and local nonmarket public goods such as environmental quality. Yet they are plagued by omitted variable bias when variables that are correlated with the attribute in question are unobservable. Typically, researchers have relied on fixed effects, instrumental variables, or quasi-randomness to control for this. However, these methods require strong underlying assumptions that are often a priori implausible. The increase in availability of big data and unstructured data in the form of text and images allow for a more extensive set of variables that are relevant to consumers to be included in hedonic methods. Unstructured data are high-dimensional and require machine learning methods that are robust to multicollinearity and irrelevant variables. I collect a rich and comprehensive dataset of property listings from Zillow.com and extract features from house descriptions and curbside view images using natural language and computer vision tools. I apply machine learning techniques to estimate the effects of air pollution on house prices in Pennsylvania. Coupled with the inclusion of more data, this approach nests previous methods to further reduce bias. My results show that omitting important variables can understate the negative effects of air pollution on house prices by more than half.

### *1.1 Introduction*

The hedonic model is typically used to estimate the market price of various attributes of heterogeneous goods. It has frequently been applied in the housing market to divide the price of a house into individual prices of house characteristics. It is one of the most widely accepted methods for estimating the tradeoffs between private goods and nonmarket amenities by treating public goods as house attributes. One of the main challenges in hedonic valuations of public goods is omitted variable bias, when variables unobservable to the econometrician are correlated with any of the included regressors. This causes biased coefficients and occasionally, flipped signs. Researchers typically rely on econometric methods such as fixed effects, instrumental variables, or quasi-randomness to obtain unbiased estimates. However, these methods require strong underlying assumptions that are a priori implausible.

With the increase in the availability of data from a wide array of public sources, bias can be reduced by controlling for previously omitted variables that are correlated with the included regressors. In this paper, I demonstrate how new sources of data can reduce bias and how machine learning methods can be used to analyze high-dimensional data. These approaches can nest previous methods and incorporate more data to further reduce bias. I apply these methods to a hedonic framework to answer the important question of how much households value air quality.

Using more data can strengthen econometric methods frequently used to control for omitted variable bias. For example, in the seminal paper by [11], boundary discontinuity design is used to estimate the value of school quality. Better schools tend to be located in better neighborhoods, causing biased coefficients when neighborhood attributes are omitted. However, homes on opposite sides of a school district boundary, but in close proximity to each other, are assumed to have similar unobservable neighborhood characteristics. By restricting comparisons to homes close to school district boundaries but on opposite sides, treatment effect estimates are unbiased if homes conditional on observables, vary discontinuously only by school quality. Boundary discontinuity design exploits the randomness close to a treatment

assignment boundary to obtain unbiased estimators.

However, the assumed randomness at the boundary fails to hold if there are discontinuities in unobservables that also influence price. As discussed in [7], households sort themselves and differ significantly by demographics immediately across district boundaries. This sorting is likely to cause correlation between better schools and local public amenities as well as housing quality even in close proximity to the boundaries. For instance, households with higher income tend to remodel their homes by upgrading appliances or having better construction quality. These features, which have previously been ignored in hedonic models, can now be captured by property descriptions and images as part of online listings. Controlling for these house attributes that are likely to be discontinuous at the boundary can further reduce bias under boundary discontinuity design.

Buying a home is unquestionably a high-dimensional problem. Homebuyers factor in a large number of property details before making a purchase: they look at design and quality of the house structure, interior appliances, neighborhood character, local amenities, school quality, curb appeal, etc. However, a majority of hedonic analyses solve a low-dimensional problem with a limited set of features. They do not realistically model the complexity of the decision and predict out-of-sample poorly. I more accurately model homebuyers' decisions by including a rich set of features from Zillow that include house listing descriptions and images. Information delivered in this format is both informative to a buyer and likely to be correlated with other home attributes causing bias when omitted in hedonic regressions. These types of data have traditionally been avoided in analyses largely due to a lack of ways to quantify and incorporate them into regressions, but are nevertheless, relevant to homebuyers.

I start with two motivating examples to demonstrate that there is pertinent information from listing descriptions and images that can explain price variation but has never been used in hedonic models. In figure 1.1, the curbside views of two homes that share the same basic characteristics are shown. They are in the same census tract with similar average school quality. The property on the left has 3 bedrooms, 2 bath, and 1900 square footage. The property on the right has 3 bedrooms, 1.5 bath, and 1700 square footage. Standard hedonic

Price: $96,406

Price: $118,304

Figure 1.1: Curb Appeal and Price

applications use only the attributes just listed as controls and then apply census tract fixed effects to control for unobservables. However, the sale prices of these two homes are very different: the one on the left sold for $96,406 while the one on the right sold for $118,304. A model that controls only for these attributes would predict similar prices for the two homes, if not a lower price for the home on the right, and be biased by almost 20%. Looking at the curbside views, a buyer would predict a higher price for the home on the right because it has higher curb appeal[1]. The house on the right has a nicely mowed lawn with sufficient green decorating the front of the house. These images can partially explain the price variation.

A similar example with listing descriptions is shown in table 1.1, again the two homes are in the same census tract and similar school districts. They both have 3 bedrooms, 2 bath, and are about 1800 square footage. However, from the description for house A, the home has a much older feel and "needs some TLC" even though both homes are about 70 years of age at time of sale. Home B indicates that it has been remodeled with new appliances and is near a park. Actual sale prices are $90,717 for home A and $122,488 for the home B. Again, most hedonic models would predict similar prices for the two homes. These two examples show

[1]Many articles from realtors suggest that adding flowers and having a neat front lawn boost curb appeal. Source:"8 Tips for Adding Curb Appeal and Value to Your Home" Houselogic

Table 1.1: Two Home Descriptions

| Price | Listing Description |
|---|---|
| Home A sold for $90,717 | "An old gem in the Park Plan.....Lovely brick 2 story with spanish tile roof. This home was a great family home for years but is ready for a fresh start. Needs some TLC, but has a great floor plan, plumbing/bthrms in all the right places, and tons of character and charm." |
| Home B sold for $122,488 | "Wonderful lot adjoins Satler Park, a Boro-owned nature reserve, located on a quiet st; totally & amp;tastefully remodeled-new K w/ stainless appliances, 2 new baths, refin.wood flrs, new plaster, updated windows & amp;roof, freshly painted; huge rear deck overlooks woods at rear." |

that there is information in the form of unstructured data that is important to homebuyers but has historically been ignored in hedonic applications. Including this information into hedonic models can undoubtedly improve house price models. In fact, research has shown that potential buyers spend up to 60% of their time looking at pictures, another 20% reading property descriptions, and the remaining 20% looking at specs (number of bathrooms, number of bedrooms, square footage, floors, etc.) when viewing online listings.[2] Previous papers have modeled house prices off information that buyers spend less than 20% of their time looking at online. I propose to incorporate all of this information.

This information is considered unstructured data, or data that are not organized in a specified format. Unstructured data are usually in the form of raw text, images, or video and can be transformed into structured data depending on the desired usage. A simple approach used in natural language processing is to transforms text into a matrix of word frequencies with each unique word defining a variable. This is called the bag-of-words model

[2]Source: Seiler, Michael, Poornima Madhavan, and Molly Liechty. "Toward an understanding of real estate homebuyer internet search behavior: an application of ocular tracking technology." Journal of Real Estate Research (2012).

and naturally creates high-dimensional data. Common phrases, or words that frequently show up together, create correlation across words resulting in multicollinearity. With property listings, there are typically select key words that are informative to buyers such as "TLC","Park", and"remodeled" in the example in table 1.1, but a majority of words are irrelevant and can be disregarded. These words are predictive of price, but when dealing with thousands of listing descriptions, knowing which words are relevant is a challenge. These issues can easily be dealt with using machine learning methods that perform variable selection based on how predictive variables are. Also, prediction accuracy is not affected by multicollinearity since these methods only select one variable from a group of highly correlated variables for prediction.

Machine learning methods have primarily been used for prediction but more recently, there have been theoretical discussions on how to adjust tree models for causal inference. Tree models group similar observations together in a way that minimizes prediction error. [2] partition trees to get heterogeneous treatment effects. [41] suggests that when given clear control and treatment groups, if the model is predictive, a way to simulate the counterfactual is to predict the outcome for the treated group using a model trained on the control group. The residual between the predicted outcome and the actual outcome is interpreted as the treatment effect. Using this approach, I show that these models are better at predicting house prices for the control group and thus yield better counterfactuals of the treated group.

This paper has two main contributions. First, I show how the increase in the availability of data can be used to reduce omitted variable bias in hedonic evaluations and yield more accurate models and treatment effects estimations. Bias in hedonic models come from correlation between omitted variables and included regressors. The size of this bias is proportional to the degree of the correlation from the omitted variables ([13]). Including more variables can lower the number of omitted attributes that are correlated with the included regressors and reduce the size of this bias. Various methods such as fixed effects, nearest neighbor matching, instrumental variables, and boundary discontinuity design have been used to control for omitted variable bias. However, they can significantly fail to provide

accurate estimates when important variables are unobserved or when there are not enough instruments. Zillow makes available a comprehensive set of property listings incorporating house characteristics, descriptions, and images. Using a randomly selected dataset on homes in Pennsylvania, I show that areas with higher pollution are also associated with homes of higher quality, better features, and interior appliances. Controlling for these attributes yields intuitive negative treatment effects and failing to control for them can understate the negative effects of air pollution, in some cases, by more than half.

Second, I show how the use of unstructured data in the form of text and images can improve models predictions and reduce bias. I demonstrate methods to turn highly unstructured features such as words and images into interpretable structured variables using natural language and computer vision tools. Text and image data tend to be dominated by irrelevant and highly correlated variables and require machine learning methods. Use of these methods has been popular in computer science, statistics, and fields such as biostatistics where it is difficult to know which genes are good predictors for specific diseases. However, few papers in economics have used these methods, let alone for the evaluation of environmental goods. I show how these methods can be used to analyze large datasets and conduct model selection in the context of causal inference.

This paper is in line with growing literature using natural language and image processing tools in economics. [22] use machine learning and natural language tools to estimate political affiliations from speeches. [28] use Yelp reviews to predict restaurant hygiene violations. I transform Zillow house descriptions into a bag-of-words, using straight-forward natural language parsing techniques. [25] use Google street view images to predict income levels in New York City . My use of curbside view images of homes to assess the curb appeal of homes and its influence on price fits into this genre of literature. Past studies have used unstructured data for prediction. This paper is the first to use text and image data to reduce bias in causal inference. The results show that home descriptions and curb appeal are correlated with air quality, and failure to control for them underestimates the negative effects of air pollution.

I use particle matter (PM) levels as a proxy for air pollution since it globally affects the health of the most people compared to any other pollutant. These are small elements in the atmosphere that can have adverse effects on health with particular risk to respiratory and cardiovascular systems[3]. The smaller the particle matter, the easier it can penetrate the lungs. I focus specifically on $PM_{10}$ levels, which are particles on the order of 10 micrometers or less in diameter. $PM_{10}$ originate primarily from industrial activities such as construction, motor vehicles, and road dust, but can also vary widely due to wind and other meteorological patterns. I find that a 1 unit decrease in pollution is associated with a $1300 decrease in house prices. Isolating the effects from just one pollutant can be difficult since economic activity causes many other pollutants such as ozone, nitrogen dioxide, and sulfur dioxide. Thus, my results should be interpreted as decreases in house prices from air pollution more generally.

The remainder of the paper is organized as follows. Section 1.2 describes the theoretical framework of hedonic models. Section 1.3 explains the machine learning methods and how average treatment effects can be estimated. Section 2.3 describes the data. Section 1.5 discusses how I process raw text and images. In section 1.6, I explain how I select the control and treatment groups. I compare the treatment effect estimates from OLS and machine learning in section 2.5. I conclude in section 1.8.

### *1.2 Hedonic Model*

I start with a standard hedonic model in a differentiated product market where price is a function of a vector of product attributes ([36]). The slope of the hedonic price function is interpreted as the marginal willingness to pay (MWTP) for an incremental change in a product attribute. I offer a brief overview of the model's application in environmental markets; for more detailed explanations, see [38].

The hedonic model can be applied to environmental markets where preferences for non-

[3]The World Health Organization air quality guidelines in 2005 documented such risks. http://apps.who.int/iris/bitstream/10665/69477/1/WHO_SDE_PHE_OEH_06.02_eng.pdf

market attributes are revealed through the tradeoffs households make in their housing decisions. In the housing market, locational amenities are treated as attributes that enter the house price function. Subsequently, the MWTP, or implicit price, for an incremental change in a nonmarket good can be estimated by exploiting the spatial heterogeneity across locations. When solving for the implicit price of a nonmarket good such as air quality, $z$, within a housing hedonic model, the most frequent functional form is

$$ln\, p_{ikt} = x'_{ikt}\beta + \theta z_{ikt} + \epsilon_{ikt}. \tag{1.1}$$

$p_{ikt}$ is the price of home $i$ in neighborhood $k$ at time $t$. Log price is a function of a vector of property attributes $x_{ikt} = (f_{ikt}, d_{ikt})$ which includes structural home attributes, $f_{ikt}$, and neighborhood demographic attributes, $d_{ikt}$. $\epsilon_{ikt} = u_{ik} + \xi_{ikt}$ is the error term with two components: individual and location specific unobservables, $u_{ik}$, and an idiosyncratic error term, $\xi_{ikt}$. Under this framework, households' optimal choices are such that their MWTP for an incremental change in an attribute is equal to the marginal price, or the slope of the price function, at that point. Therefore, $\theta$ can be interpreted as the MWTP for air quality. Equation 1.1 is frequently applied assuming $E[\epsilon_{ikt}|x_{ikt}] = 0$ holds to obtain unbiased estimators. When there are unobservable attributes in $u_{ik}$ that are correlated with any of the regressors, $x_{ikt}$ or $z_{ik}$, then $E[\epsilon_{ikt}|x_{ikt}] \neq 0$, and $\beta$ and $\theta$ estimates are biased.

Applications of this theoretical framework to nonmarket amenities are diverse; past studies have focused on water quality ([30]), cancer clusters ([17]), fracking wells ([32]), crime ([31]), air quality [14], and power plants ([18]). These papers have used various econometric approaches to control for omitted variable bias, but they are insufficient substitutes to directly controlling for omitted attributes.

Consider the case of air quality, where the usual culprit is that pollution is positively correlated with areas of higher employment and better urban amenities, often resulting in perverse positive treatment effect estimates. Most papers use an instrument for pollution ([14]), but fail to control for all other endogenous attributes that are correlated with unob-

servables. Consistent estimation requires one instrument for each endogenous variable, which is rarely available. An alternative approach is to use individual home fixed effects on repeat sales to control for omitted variable bias ([32]). This assumes that unobservables are time invariant, but this is unlikely in the housing market where sellers often remodel their homes right before sale. Remodeling is rarely captured in hedonic studies since it does not change the main features of a homes such as size or number of bedrooms, but it can significantly increase the price. Additionally, regressors are often collinear with fixed effects. While many papers have used a variety of approaches to reduce omitted variable bias when recovering the hedonic price function, none have taken the approach of directly controlling for omitted attributes using big data.

### *1.3 Methodology*

In this section, I provide an introduction to high-dimensional regression methods, regression trees, and tree-based machine learning techniques: gradient boosting and random forest. The methods in this section are chosen because they perform effective variable selection that predict out-of-sample well despite the presence of multicollinearity and irrelevant variables.

With large datasets, the most important variables are often known, but the importance of the remainder of the variables is unknown. Researchers have commonly relied on industry knowledge when selecting which variables are relevant for a particular outcome. Selecting sparse models with a small subset of variables stems from the limitations of econometric models that fail in high dimensions due to identification problems when there are more variables than observations or constrained computational power. While there is no perfect substitute for expert knowledge or market intuition, machine learning methods offer data driven approaches to model selection that can complement prior knowledge. Allowing for patterns in the data to guide model selection and a set of hyperparameters, or tuning parameters, chosen by the researcher to avoid overfit, these machine learning methods can provide insight on which variables are important and which ones are irrelevant.

### *1.3.1 Regression Methods for High-Dimensions*

The first group of methods - regression methods for high-dimensional data - perform regularization and shrinkage where the coefficients are reduced towards zero to reduce complexity. This is done by modifying the cost function in OLS to penalize additional variables in the model, or complexity. These methods also reduce the standard errors and are particularly useful in the presence of multicollinearity. I discuss one method from this family - Lasso (least absolute shrinkage and selection operator) - as it conducts variable selection; see [20] for discussion on alternative methods.

*Post-Lasso*

Lasso is a regularization method that induces variable selection by shrinking some model coefficients to zero ([39]). Setting a majority of coefficients to zero mediates model interpretation as well as prevents overfit. With OLS, $\hat{\beta}$ minimizes the residual sum of squares error (RSS). In Lasso, an L1 penalty term is added to the cost function in OLS. For a model with $y_i$ as the outcome for observation $i$ and a vector $x_i$ of $p$ controls, the optimal coefficients become

$$\hat{\beta}^{Lasso} = \underset{\beta \in R^p}{\arg\min}\, E[(y_i - x_i'\beta)^2] + \frac{\lambda}{n}||\beta||_1$$

The threshold $\lambda$ is a tuning parameter that can be chosen based on theoretical properties or via cross validation.[4] Lasso improves OLS in two ways: reducing the variance of predictions and simplifying the interpretation of the model by eliminating less influential variables. It effectively discards variables with small coefficients and variables with large coefficients that do not improve prediction. It also handles multicollinearity by randomly selecting

[4]Cross validation divides data into $k$ partitions and the model is trained on $k-1$ subsamples using the $k$th partition as an out-of-sample test set. This process is repeated $k$ times, each time using a different partition as the test set. The tuning parameter from the best out-of-sample prediction accuracy is chosen.

one variable from a group of correlated variables. This is a practical approach to account for multicollinearity as long as the treatment variable is not highly correlated with other regressors, which would result in omitted variable bias. [5]

Since Lasso estimates are nonlinear and non-differentiable functions, accurate standard errors cannot be computed. It also biases the estimated coefficients towards zero. [9] propose to run OLS using only the regressors with nonzero coefficients from the Lasso estimates to resolve these issues. Standard errors can easily be computed and estimated coefficients are less bias. The tuning parameter, $\lambda$ is a data-driven penalty term.[6] A large $\lambda$ will reduce variance but increase bias. An optimal choice will balance these two. The intuition behind their choice for $\lambda$ is that it is just large enough to tune out noise from irrelevant variables. This approach effectively uses Lasso only as a model selection step. The advantage of post-Lasso is that the final model enjoys the benefits of regularization and variable selection from Lasso while still yielding OLS estimates with standard errors.

*Post-Double-Selection*

In many cases, the econometrician is particularly interested in the treatment effect of one variable in the presence of many controls. Inference using post-Lasso can result in omitted variable bias when faced with treatment selection bias. When only controls that are important to the outcome variable are chosen, treatment effect estimates are bias if there are variables correlated with the treatment that are not selected. For causal inference, [10]

[5] In contrast, a L2 penalty under ridge regression can be used if it is believed that the true model includes a majority of the variables. This method does not conduct variable selection but will still shrink the coefficients towards zero. Elastic net provides a mixture of ridge and Lasso and in practice, works well with groups of highly correlated variables ([43]). Hence, some knowledge of the underlying structure of the variables is recommended when selecting an appropriate method. This knowledge can be hard to acquire when handling large datasets.

[6] $\lambda = 2c'\hat{\sigma}\Lambda(1-\alpha|X)$ with c=1.1 and $\alpha$=0.05, $\Lambda(1-\alpha|X)$ is the $(1-\alpha)$ quantile of $n||E_n[x_i\epsilon_i]/\sigma||_\infty$, and $\hat{\sigma}$ is a data driven estimate of $\sigma$.

propose a two-step approach to unbiased estimators. Consider a model as follows:

$$y_i = x_i'\beta_g + \theta z_i + \xi_i \tag{1.2}$$

$$z_i = x_i'\beta_m + \nu_i \tag{1.3}$$

where $E[\xi_i|z_i, x_i] = E[\nu_i|x_i] = 0$ for observation $i = 1, 2, ..., n$ and $z_i$ is a scalar treatment variable (air pollution in this case), and $\theta$ is the parameter of interest. A standard post-Lasso model would conduct variable selection on equation (2.2) to find variables, $x_y$, that predict $y$ well. However, the authors argue that this can lead to omitted variable bias when there are variables not in $x_y$ that influence $z$. They propose a post-double-selection method where variable selection via Lasso is also done on equation (2.3) to obtain $x_z$. This conducts variable selection in two steps. Then running least squares with $y$ on the union of $x_y$ and $x_z$ variables would obtain unbiased estimates of $\theta$. This is a way to handle treatment selection bias, similar to a propensity score matching method where estimation needs to control for pre-treatment bias. See [10] for more details.

### *1.3.2 Regression Tree Models*

Tree-methods are popular alternatives to Lasso-type methods. Unlike Lasso, they nonparametrically divide the input space into regions and automatically explore nonlinearities and interactions. They predict very well out-of-sample by averaging many trees together. Averaging across a large number of trees increases prediction accuracy at the cost of easy interpretations of all variables. However, I will show how they can still be used to isolate the effects of select treatment variables. More details on these methods can also be found in appendix A and [20] and [33].

*Gradient Boosting*

[21] introduces gradient boosting as an additive ensemble model where a weighted sum of tree models are added together. The trees are grown sequentially, unlike random forest where trees are grown separately and then averaged. At each iteration, a new tree is added to a weighted sum of all previous trees.

The model is initialized at the average of the outcomes, $\bar{y}$ for $N$ observations. At each subsequent iteration, the model parameters are chosen to lower RSS from the previous iteration with an additional tree and a subsample of the original data. Suppose there are $M$ iterations, then the model is defined as:

$$F_M(x) = \sum_{m=1}^{M} \nu f_m(x) \tag{1.4}$$

where $f_m(x)$ is the tree at iteration $m$ and $\nu$ is a shrinkage parameter to prevent overfit and to slow down the learning process to not allow one iteration to have too much influence. By setting $\nu$ to be a small number, usually between 0.01 and 0.1, this puts low weight on each additional tree and makes the final prediction less sensitive. $f_m(x)$ is usually chosen to be a stump where the input space is cut by one variable; the depth of the trees can be increased but performs best with few terminal nodes. Gradient boosting can handle missing values well because it selects cuts based only on observations that have data. This makes it ideal in high-dimensional datasets that are often plagued with missing values because it eliminates the need for arbitrary imputation without discarding observations.[7]

Understanding the effect of a single variable in the model can be difficult when predicted outcomes are a function of many trees. Partial plots are a useful tool for this purpose. They can provide insight on how a variable influences the outcome by graphing the predicted outcome $\hat{F}(x)$ as a function one of the inputs after factoring out the effects of all the other

[7] It is important to note that if there are systematic biases that cause the data to be missing, gradient boosting does not correct for it. The results of the model assume that the data are missing at random and output predictions based only observations with data.

variables. In other words, a partial plot for input variable $x_j$ would be a plot of

$$\hat{F}_{x_{-j}}(x_j) = \hat{F}(x_j|x_{-j}) \tag{1.5}$$

This is estimated from the training data with $N$ observations by averaging over the conditioned variables

$$\bar{F}_j(x_j) = \frac{1}{N}\sum_{i=1}^{N} \hat{F}(x_j, x_{i,-j}) \tag{1.6}$$

In general, partial dependence plots are the most useful when the output variable is strongly dependent on $x_j$ and when $x_j$ is not too correlated with other variables $x_{-j}$. Nevertheless, these plots can still give insights on how a variable influences the predicted outcome. I will use these plots to interpret the results of these models.

### *1.3.3 Average Treatment Effect Estimation*

While gradient boosting and random forest provide rankings of variable importance, they do not estimate direct effects of variables on the outcome. As [41] proposes, if a model is predictive for the control group, then it yields a model of the counterfactual world, one where the treatment group is not treated. The difference in the predicted and the actual can be interpreted as the treatment effect. This is one way to use predictive models for causal inference. Take $F(x)$ and train it on the control group. All splits and thresholds will be determined by the control group. Then feed the treatment group into this model and take the predictions as the counterfactual. The average treatment effect on the treated can then be written as

$$ATE = \frac{1}{N^{Treat}} \sum_{1}^{N^{Treat}} (Y^{Treat} - \hat{F}(x)) \tag{1.7}$$

where $\hat{F}(x)$ is the predicted outcome from the model trained on the control group, $N^{Treat}$ is the number of observations in the treated group, and $Y^{Treat}$ is the realized outcome or house price of the treated. This is the average of the residuals on the treated group.

Tree models match treatment group observations to control group observations directly

on attributes. Treated observations are matched to regions within the tree model based on attributes and assigned the average outcome of observations in that region. Counterfactual estimates are based, in boosting, on a weighted sum of trees, and in random forests, an average of trees. Treated observations are not matched to just one observation in the control group, but instead a group of observations [8]. The predicted residuals can then be regressed on other covariates to recover heterogeneity in treatment effects. I select the treatment and control groups by training a gradient boosting model on all the data and looking at the partial plots. The partial plot reveals how important different levels of pollution are to house prices and be discussed further in section 1.6.

## *1.4 Data*

### *1.4.1 Housing and Census Data*

A random sample of homes in Pennsylvania are collected from Zillow.com. All properties are single family homes with last sold dates between 2000-2012. Variables collected include structured and unstructured (text and images) characteristics from the home details page for each property. A summary of the available structural property details beyond the basic variables (number of bedrooms, number of bathrooms, square footage, and lot square footage) is in table 2.1. These attributes enter the data as dummy variables for each possible value that a feature category can take on. Not all houses list these variables, thus, if a feature is listed in a category, then it takes on zero for all other possible categories. But if no feature in a category is listed, then missing values are assigned to all categories. For example, if a home lists its roof type as "slated", then it takes on a zero for all other roof types. However, if no roof type is listed at all, then it receives missing values for all possible roof types. The treatment and assignment of missing values are important because gradient boosting is robust to them while random forest is not. This can affect the thresholds at which cuts are chosen within regression trees.

[8] This is similar to synthetic control methods ([1]) which compute counterfactuals by matching to a group

Table 1.2: Summary of Rich Structural Variables

| Feature category | Values |
|---|---|
| Appliances | dishwasher, dryer, freezer, garbage disposal, microwave, range oven, refrigerator, washer |
| Architecture type | bungalow, cape cod, colonial, contemporary, craftsman, french, georgian, loft, modern, ranch, spanish, split level, tudor, victorian |
| Basement type | finished, partial, unfinished |
| Cooling source | central, evaporative, geothermal, refrigeration, wall |
| External material | brick, cement/concrete, composition, metal, shingle, stone, stucco, vinyl, wood, wood products |
| Features | attic, barbecue, basketball court, cable ready, ceiling fan, deck, disability access, dock, double pane storm windows, elevator, fenced yard, furnished, garden, gated entry, greenhouse, high-speed internet, high-speed internet ready, hot tub/spa, intercom, jetted tub, lawn, mother-in-law, patio, pond, pool, porch, RV parking, sauna, security system, skylight, sports court, sprinkler system, vaulted ceiling, waterfront, wet bar, wired |
| Floor covering | carpet, concrete, hardwood, laminate, linoleum/vinyl, slate, softwood, tile |
| Heating type | baseboard, forced air, heat pump, radiant, stove, wall |
| Heating source | coal, electric, gas, oil, propane/butane, solar, wood pellet |
| Parking type | carport, garage, garage-attached, garage-detached, off-street, on-street |
| Roof type | asphalt, built up, composition, metal, shake shingle, slate, tile |
| Rooms | breakfast nook, dining room, family room, laundry room, library, master bath, mud room, office, pantry, recreation room, sun room, walk-in closet, workshop |
| View | city, mountain, park, territorial, water |

Unstructured characteristics lie in the form of text describing each home and the neighborhood and pictures of the interior and exterior of the homes posted by the sellers. Image processing is done only for curbside view images. Sellers post a variety of other images (e.g. bathroom, kitchen, bedroom, etc.) but the most consistent image content across listings is curbside view. A variable for the number of images posted in included as a structural variable. A thorough description of how these types of data are processed and used is in section 1.5. Properties with missing values for basic variables are dropped. The final dataset has 56,262 observations.

Each property was mapped to its closest elementary, middle, and high school, and data on the quality of those schools is from GreatSchools.org, where schools are rated on a scale of 1-10, with 10 being the highest, based on state standardized test performance compared to other schools in the same state. Separately, properties were mapped to census blocks and the corresponding neighborhood characteristics. This includes income, educational attainment, population size, and race. Census data was collected for 2000 and 2010.

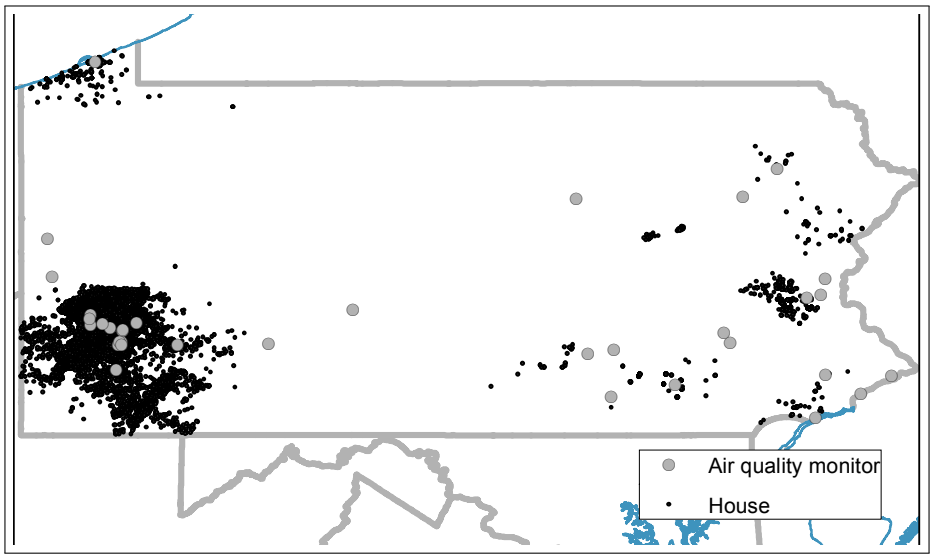
### *1.4.2 Air Quality Data*

Air quality data are collected from the US Environmental Protection Agency (EPA). The EPA regulates the pollutant closely as one of the six "criteria pollutants" via the Clean Air Act. The $PM_{10}$ concentration levels are measured in micrograms per cubic meter or $\mu g/m^3$. Annual averages are collected for years 2000-2012 from each air quality monitor in Pennsylvania. The $PM_{10}$ level for each home is calculated as a mean weighted by the inverse distance from all the monitors in Pennsylvania in its corresponding year of sale. Figure 1.2 shows the location of the properties in relation to the $PM_{10}$ monitors. This map confirms that monitors are placed in residential areas and that the monitors are reasonably close to the homes in the data for accurate pollution level estimates.

I also collect the mean of the daily 1-hour maximum for sulfur dioxide ($SO_2$) because this pollutant is highly correlated with $PM_{10}$. $SO_2$ levels are also in $\mu g/m^3$ and are assigned to

of control observations based on a set of matching variables

Figure 1.2: Map of Pennsylvania Air Quality Monitors and Homes




Note: This shows the homes in the data in relation to $PM_{10}$ monitors. This shows monitors from 2000-2012.

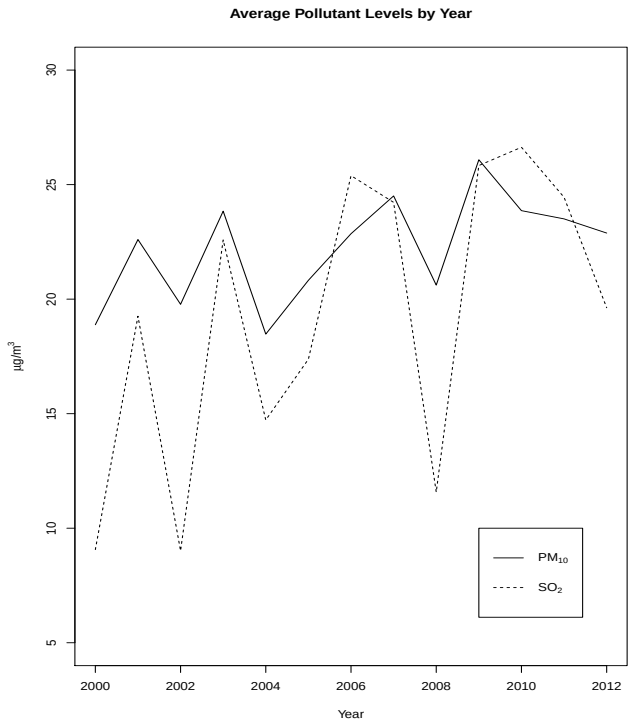
homes in the same way as $PM_{10}$ levels. Given the collinearity of the two pollutants, disentangling the effects can be problematic, and it is likely that my treatment effect estimates include the effects of both pollutants. A graph of the mean levels of the pollutants for the homes in each year are plotted in figure 1.3. Correlation across all the years is 0.77. Sources of $SO_2$ are largely from the burning of fossil fuels from power plants or industrial sites. It has similar effects on health as $PM_{10}$. Some studies indicate that it can react with other particles in the atmosphere creating more particle matter, which explains the high correlation between the two pollutants.

## *1.5 Processing Unstructured Features*

### *1.5.1 Estimating Curb Appeal*

In the data, 23,733 homes posted curbside images, accounting for roughly half of the observations. I use a combination of supervised and unsupervised learning for feature engineering,

Figure 1.3: Correlation Between $PM_{10}$ and $SO_2$ Across Years



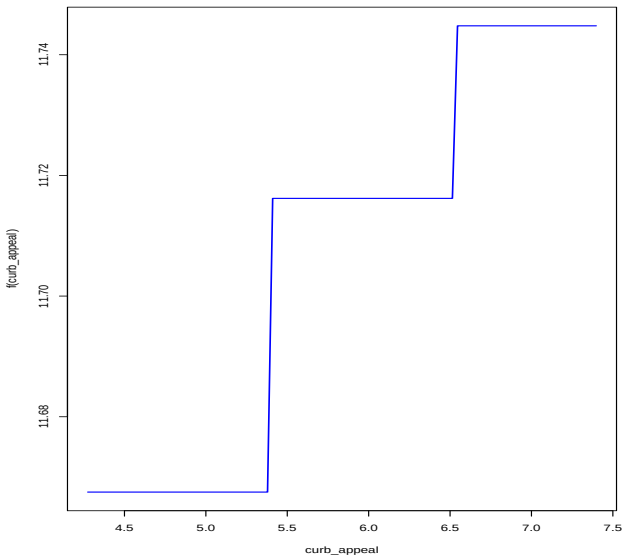

This graph plots the mean pollutant levels for homes in each year. It shows the high correlation between the two pollutants.

the process of creating summary variables that describe the data. Supervised learning allows for input values to be mapped to an output. Unsupervised learning entails using just input values that do not correspond to outputs. I extract 43 image features that on their own, are not informative and would be considered unsupervised learning since they effectively summarize the images (e.g. brightness, number of pixels, etc.). These can be used as controls on their own, but their effects on price are difficult to interpret. To provide a better understand of how images affect price, I use supervised learning and assign curb appeal ratings to the images using the raw features. See [27] for details. Reducing the dimensions of the raw features to curb appeal shows the relationship between the pictures posted and the price of

the house. To avoid limiting the effects of the image features to just one dimension - curb appeal, I use all the raw features as well as curb appeal ratings as house price controls. This adds 44 image variables to the final dataset.

In the gradient boosting model, curb appeal is the 34th most important variable out of 1,877 variables. One way to interpret how curb appeal affects the predicted price is to look at a dependence plot. I use a partial dependence plot of curb appeal to show an estimate of how log prices vary with curb appeal. Each jump in the graph corresponds to a split in the tree, and the predicted value is the average of the terminal nodes following that split. Figure 1.4 shows that the difference in prices between homes with curb appeal of less than 5 and home with curb appeal greater than 6 is roughly $6,000.

Figure 1.4: How Predicted Log Price Varies With Curb Appeal




Note: Each jump in the graph corresponds to a split in the tree, and the predicted value is the average of the terminal nodes following that split. This shows that the difference in prices between homes with curb appeal of less than 5 and home with curb appeal greater than 6 is roughly $6,000.

Table 1.3: Bag-of-words Representation

*"Beautiful wooded lot on quiet street, plumbed for bath in basement. Large high basement."*

| basement | bath | beautiful | high | large | lot | plumbed | wooded | quiet | street |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

### *1.5.2 Text Processing*

This section explains how property listing descriptions are processed. Aside from structured features, the hedonic attributes of a house can be defined by unstructured texts that an owner posts to describe a house. I apply a bag-of-words model where each paragraph is parsed into its unique words. Each distinct word is a count variable indicating the frequency of that word. I eliminate words with three or less letters and words with frequencies of less than 50. I also stem the words to group singular and plural forms of words. In the final dataset, each home has a set of words, or a bag-of-words, associated with it.

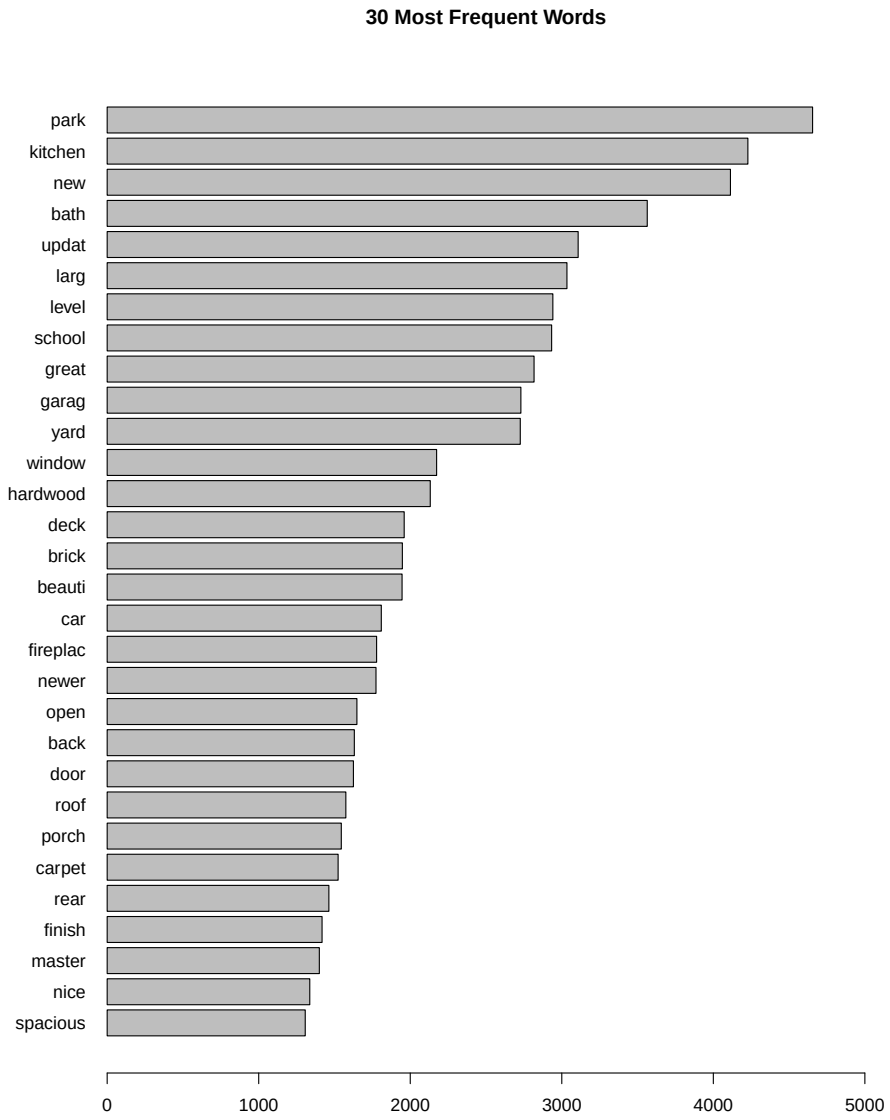
An example can be seen in table 1.3 where a sentence describing a home is represented as a bag-of-words with a column for each unique word. All words occur once except for "basement" which occurs twice. This approach ignores grammar and the ordering of words. In figure 1.5, I show the top 30 words and their frequencies. Many of these words are descriptive characteristics that would otherwise be unobservable: the presence of a "park" or having a "porch" While some of these words are subjective, such as "spacious" and "beautiful", and can be deemed marketing, there is evidence that a well written description can reduce time to sale. There have also been arguments that certain keywords can affect the sale price.[9] Including the text description can control for both unobserved neighborhood and physical features as well as possible marketing influence.

In a high-dimensional learning problem, the model is likely to depend only on some parts of an observation. For example, the information relevant to predicting house prices may lie

[9] In "Zillow Talk: The New Rules of Real Estate" by the CEO of Zillow, Spencer Rascoff and Stan Humphries look at how keywords in house description can influence the sale price.

in only a handful of its words. This makes variable selection useful when handling these types of data.

Figure 1.5: Common Words Used




Note: This shows the most frequent words that are used across the data in home descriptions. Words have been stemmed to group singular and plural forms of words.

### *1.6 Defining Treatment and Control Groups*

Since the treatment is continuous, finding appropriate treatment and control groups involves searching for thresholds in the price gradients in relation to air quality. Tree models cut variables at values that minimize the residual. If there is a range of values for a variable that is not predictive of the outcome variable, the algorithm will not partition the tree on those values. In other words, if there is a range of $PM_{10}$ levels that does not significantly affect house prices, there will not be a cut on that variable in those ranges. I exploit this property of tree models to define the treatment and control groups.

I use a gradient boosting model on the entire dataset using just structural variables (no bag-of-words or images features). This is because many home listings include the names of neighborhoods, which reveal the locations and are highly collinear with pollution levels. Including the names of neighborhoods can results in trees that never cut on pollution levels and potentially bias the partial plots. I display the $PM_{10}$ level partial plot in figure 1.6 which shows the cuts made across trees.[10] For $PM_{10}$ levels between 15 and 22 $\mu g/m^3$, there are no predicted changes in house prices. This cutoff is consistent with the World Health Organization's assessment of maximum level of $PM_{10}$ at 20 $\mu g/m^3$ that are safe for heart and lung cancer health. One note is that these plot can be sensitive to outliers and the initial jump in log price is due to noise as less than 1% of the observations have $PM_{10}$ levels less than 15 $\mu g/m^3$. Since $SO_2$ levels are highly correlated with $PM_{10}$ levels, I also use the partial plot for $SO_2$. I incorporate $SO_2$ levels to guarantee that my control and treatment groups are not confounded by an omitted pollutant that is highly correlated to the treatment. The control group is defined as homes with $PM_{10}$ levels between 15 and 22 $\mu g/m^3$ and $SO_2$ levels between 10 and 23 $\mu g/m^3$. The treatment group is defined as observations with $PM_{10}$ levels greater than 22 and $SO_2$ levels greater than 23 $\mu g/m^3$.

Since the partial plots do not show perfect causal relationships, I suggest two ways to check that this is an appropriate control group. The first one is to train two models: one

[10] Random forest yielded similar results

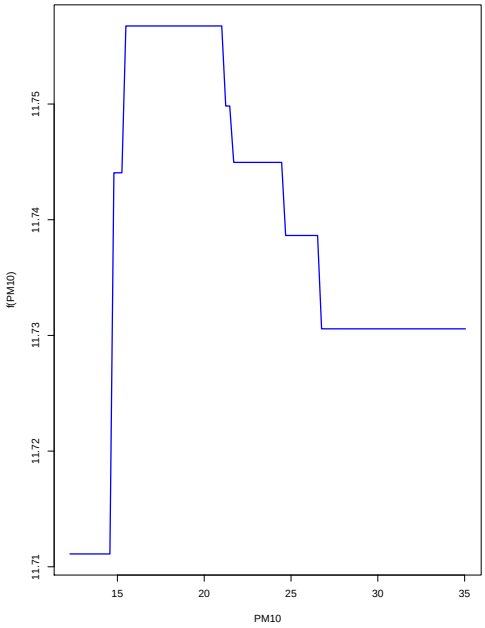



Figure 1.6: $PM_{10}$

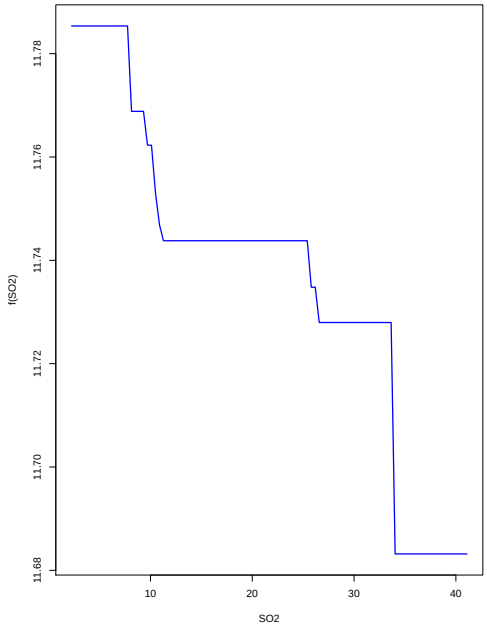



Figure 1.7: $SO_2$

Note: These graphs show the partial plots for the two pollutants and their influence on predicted house prices. The control group is defined as homes with $PM_{10}$ levels between 15 and 22 $\mu g/m^3$ and $SO_2$ levels between 10 and 23 $\mu g/m^3$. The treatment group is defined as observations with $PM_{10}$ levels greater than 22 and $SO_2$ levels greater than 23 $\mu g/m^3$.

using the control group and one using the treatment group and then comparing the list of important variables. I verify that $PM_{10}$ and $SO_2$ are not chosen as a variables in the control group, but are in the model using the treatment group. This means that the control group is effectively untreated.

The second way to check is similar in spirit to propensity score matching. An appropriate counterfactual model will control for covariates that could cause pretreatment bias since treatment assignment is not assumed to be random. Since both random forest and gradient boosting assign more importance to some variables, it is important to verify that any potential pretreatment variables are selected in the models. Finally, ideally, the two groups would

have similar distributions for these variables. However, if the two groups are not identical, what is important for appropriate estimates is for the control group to have distributions that have enough overlap with the distribution of the treated group on variables that are predictive of the treatment so that treated observations are matched to regions of the tree with similar pretreatment values.

I run models with $PM_{10}$ levels as the outcome variable and all other structural variables as controls. Figures 1.8 and 1.9 shows the important variables in random forest and gradient boosting. In table 1.4, a summary of these important pretreatment variables and basic house attributes are shown. The two group differer slightly in house structure and on almost all the census variables chosen, indicating the presence of sorting across demographic characteristics. Homes tend to be less expensive and in neighborhoods with lower income and less education on average. However, the control group has a wider distribution in almost these categories with enough overlap to appropriately match a majority of the treated observations to control observations.

Finally, I propose to see if pretreatment bias is present by dropping these important pretreatment variables and comparing the treatment effects to a model will all these variables. This will test if these variables make a difference in the average treatment effects. I find that dropping them results in very high negative treatment effect estimates indicating that these variables that are predictive of pollution levels also significantly affect price and are selected as important variables in the final models. Details on these results are in appendix B.

## *1.7 Results*

I first provide the coefficients from OLS and Lasso-type methods using the entire dataset (control and treatment groups). I report the coefficients on $PM_{10}$ and $SO_2$. Overall, these methods are unable to incorporate as many variables into the models as random forest and gradient boosting. Because of this, they understate the negative treatment effects.

I start with a limited and ignorant approach to estimating the coefficients on $PM_{10}$ and

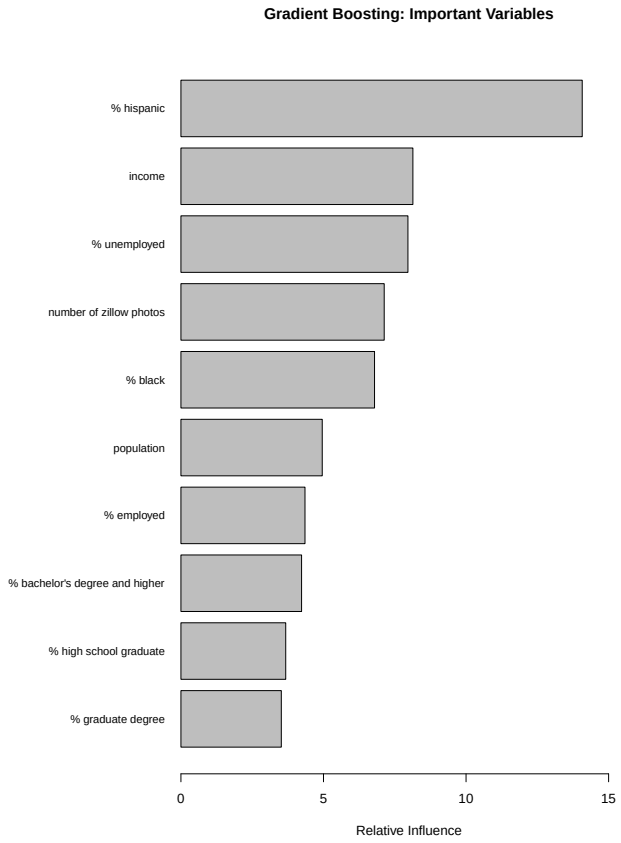
Table 1.4: Mean Summary Statistics

| | Control group* | | Treatment group** | |
|---|---|---|---|---|
| Variable | Mean | SD | Mean | SD |
| House characteristics | | | | |
| price ($2010) | 138,530.6 | (83,493.7) | 131,986.7 | (79,108.7) |
| age at sale (years) | 50.4 | (25.0) | 42.6 | (22.9) |
| square footage | 1,828.7 | (742.3) | 1,831.1 | (691.7) |
| bedrooms | 3.3 | (0.8) | 3.3 | (0.7) |
| bathrooms | 2.2 | (0.7) | 2.1 | (0.6) |
| basement sqft | 522.3 | (217.8) | 548.5 | (201.6) |
| parking spaces | 1.5 | (0.6) | 1.6 | (0.6) |
| % with fireplace | 1.0 | (0.2) | 1.0 | (0.1) |
| school rating (1-10) | 6.9 | (2.5) | 6.8 | (2.6) |
| distance to school (km) | 2.1 | (1.2) | 2.1 | (1.1) |
| Census characteristics | | | | |
| median income | 61,159.6 | (21,336.4) | 53,240.4 | (20,660.5) |
| % black | 6.4 | (11.9) | 5.9 | (12.1) |
| % unemployed | 5.5 | (3.2) | 3.5 | (2.8) |
| % employed | 64.1 | (6.1) | 61.8 | (6.0) |
| % high school graduate | 30.4 | (11.2) | 33.0 | (11.0) |
| % bachelors | 22.9 | (8.7) | 20.2 | (9.3) |
| % bachelors or higher | 38.1 | (17.7) | 32.4 | (17.4) |
| Observations | 18,999 | | 14,760 | |

* Observations with $SO_2$ between 10 and 23 and $PM_{10}$ levels between 10 $\mu/m^3$ and 22 $\mu/m^3$

** Observations with $SO_2$ greater than 23 and $PM_{10}$ levels greater than 22 $\mu/m^3$

Figure 1.8: Predicting $PM_{10}$ Levels with Boosting

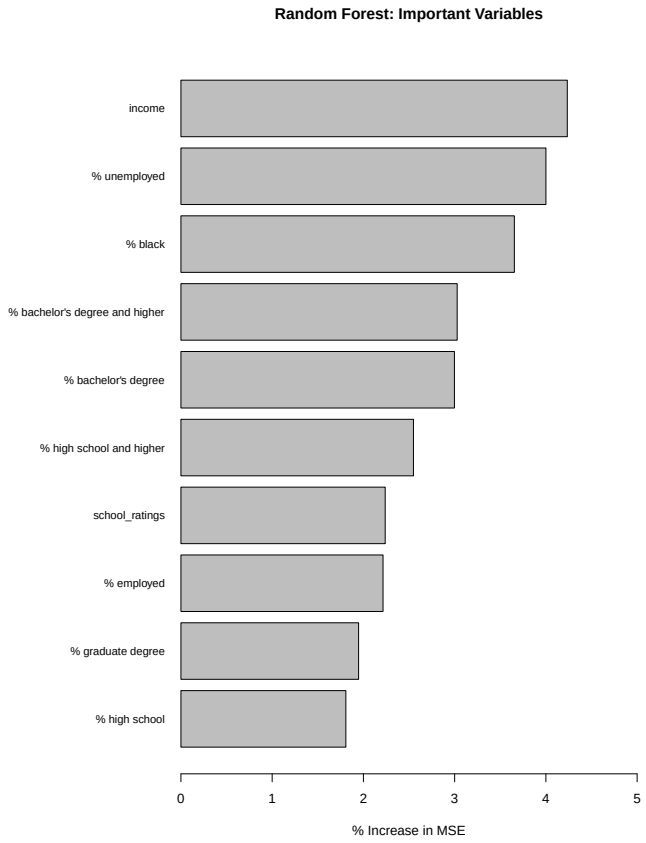



$SO_2$. Using OLS, I estimate equation the following equation.

$$lnp_{ikt} = x'_{ikt}\beta + \theta z_{ikt} + \mu_k + q_t + \epsilon_{ikt} \tag{1.8}$$

where $p_{ikt}$ is the price of home $i$ in neighborhood $k$ at time $t$; $x_{ikt}$ represents the vector of house and demographic attributes in a census block, $z_{ikt}$ is the pollution level, $\mu_k$ are census tract fixed effects, and $q_t$ are time fixed effects. I only include structured variables because of multicollinearity and the presence of a high number of irrelevant variables in OLS. Results for these three models are in table 1.5. All variables are included in post-Lasso and double-

Figure 1.9: Predicting $PM_{10}$ Levels with Random Forest




selection[11]. None of the coefficients estimated are statistically significant for $PM_{10}$. OLS and post-Lasso obtain negative signs but double selection has the perverse sign. Coefficients on $SO_2$ are statistically significant and negative and make sense but these models fail to disentangle the effects of the two pollutants. Post-Lasso and double selection impose a level of sparsity on the models that does not eliminate omitted variable bias when the true model has more variables. While post-Lasso obtains a negative coefficient, I will show that this is likely an underestimate of the negative effects of $PM_{10}$.

[11]Post-Lasso and post-double-selection can easily be implemented using the hdm package in R.

Table 1.5: Coefficients on Pollutants

| Pollutant | OLS | Post-Lasso | Double-selection |
|---|---|---|---|
| $PM_{10}$ | -0.00204 | -0.00098 | 0.00056 |
| | (0.00112) | (0.00114) | (0.00113) |
| $SO_2$ | -0.01257 | -0.01156 | -0.01197 |
| | (0.00084) | (0.00086) | (0.00085) |
| Variables | structured | all | all |

Notes: standard errors are in parenthesis

### *1.7.1 Model Accuracy*

Machine learning methods tend to overfit on the training data so looking at the root mean-squared error (RMSE) of an out-of-sample test set is useful to compare prediction accuracies across models. Having a model that predicts house prices accurately on the control group gives a good counterfactual prediction of the treatment group. Training only on $\frac{3}{4}$ sample of the control group, I compare three models: post-Lasso, gradient boosting, and random forest. Double selection requires inclusion of the treatment variable and is dropped in this comparison since the models are not trained on the treated group. Lasso methods return very sparse models with a majority of coefficients set to zero.[12] They require strong parametric assumptions and that the true model be driven by a small number of variables ([9]). These restrictions make them less ideal for situations where the true model has many variables or when prediction accuracy is preferred.

I start each model using basic variables which is defined as bedrooms, bathrooms, square footage, lot square footage, census block demographics variables (race, income, education, population, employment), and year and census tract fixed effects. These models easily handle categorical variables. Next, I add all structured variables to the models, this includes all

[12]Lasso-type methods, like most machine learning methods, rely on tuning parameters that determine the complexity of the model. They can be changed to include more variables but at the cost of statistical power. In practice, they are often determined using cross-validation.

Table 1.6: Model Accuracy Comparisons

| Model | # Variables selected** | Out of Sample RMSE | Standard Error of RMSE |
|---|---|---|---|
| Post-Lasso* | | | |
| basic | 15 | 0.386 | 0.0103 |
| all structured | 14 | 0.397 | 0.0010 |
| + pics | 14 | 0.397 | 0.0099 |
| + words | 21 | 0.397 | 0.0099 |
| Boosting | | | |
| basic | 15 | 0.378 | 0.0110 |
| all structured | 80 | 0.352 | 0.0107 |
| + pics | 113 | 0.353 | 0.0107 |
| + words | 153 | 0.348 | 0.0107 |
| Random Forest | | | |
| basic | 15 | 0.366 | 0.0105 |
| all structured | 171 | 0.361 | 0.0105 |
| + pics | 216 | 0.361 | 0.0105 |
| + words | 1893 | 0.361 | 0.0107 |

*Model with basic variable is a standard OLS. Lasso is performed on the remaining three model specifications.
**based on 171 structural variables, 44 picture variables, and 1677 word variables

features, appliances, exterior materials, etc. from table 2.1. Then I add in the 44 image features and last, the 1677 words. I report the number of variables selected, RMSE on the remaining $\frac{1}{4}$ out-of-sample test set, and the standard errors in table 1.6. Standard errors are calculated holding the model parameter from the training set fixed and then feeding a 50% subsample with replacement of the data into the model 10,000 times. [6] show that the three prediction models are asymptotically normal. Showing how many variables are selected is informative since Lasso and boosting are both effective variable selection models. Random forest conducts a form of variable selection by assigning variable importance ranking to variables, but it effectively includes all the variables with low weights given to irrelevant ones.

Use of these models requires some upfront tuning parameters. Random forest is run with 1000 trees and a random 1/3 of the variables possibly used for each split ($m = \frac{p}{3}$). I select

the minimum number of observations in a terminal node to be 50. Gradient boosting is run with 1000 trees, a shrinkage parameter of 0.1, minimum number of observations in a terminal node to be 100, and splits on 3 variables in each tree[13]. Boosting has a higher tendency to overfit than random forest because additional trees are cut based on the residual. Thus, I set a higher minimum number of observations in the terminal nodes. The optimal number of trees to predict on is selected through a K-fold cross validation with K=10, a common choice in the machine learning literature. Among the four models, boosting obtains the lowest error with random forest having the second lowest. The two tree-based models perform similarly but outperform post-Lasso by more than 10%.

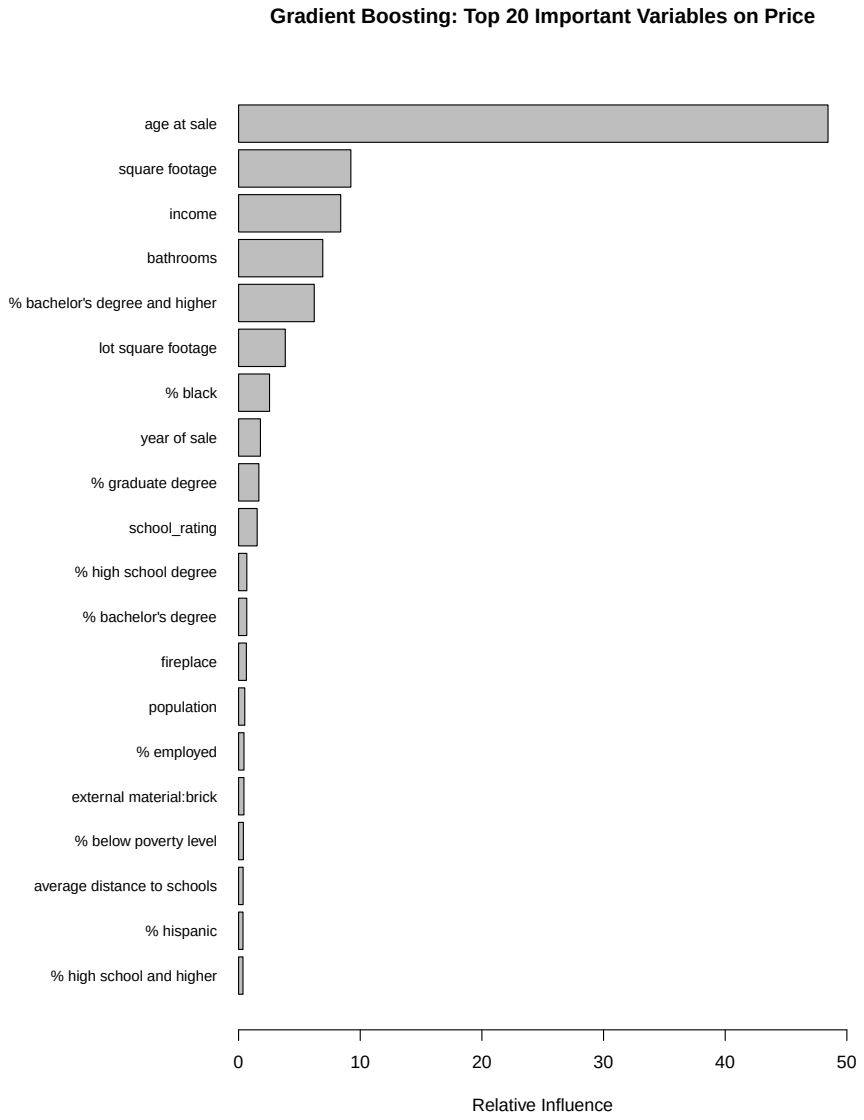
In general, adding more variable improves prediction accuracy. The inclusion of image features does not increase accuracy in any of the models. There are two possible explanations for this result: only half of the properties had a picture posted and curb appeal is likely highly correlated with other home attributes in the control group which will be already be captured in the models. Adding variables that are highly correlated with other variables in the model do not significantly improve prediction accuracy in ensemble trees. Nevertheless, since the image features are correlated with air quality, including them in the model is important when estimating treatment effect.

Figures 1.10 and 1.11 show the top 20 variables are important in each of the two tree-based models. Pretreatment variables (from figures 1.8 and 1.9) are indeed important to price in both models and are controlled for in the model. More discussion on this in appendix B. In random forest, variable importance is defined as how much the mean squared error (MSE) increases on a random out-of-sample test set (to avoid overfit) when the variable is removed from the model. The higher this percentage, the more important the variable is. For boosting models, this is called the relative influence of the variables and is measured by how much splitting on a variable improves the mean squared error. Figures 1.10 and 1.11 show the 20 most influential variables for the two models. The results are sparse in

[13]All these parameters in gradient boosting are selected based on cross validation R package Caret.

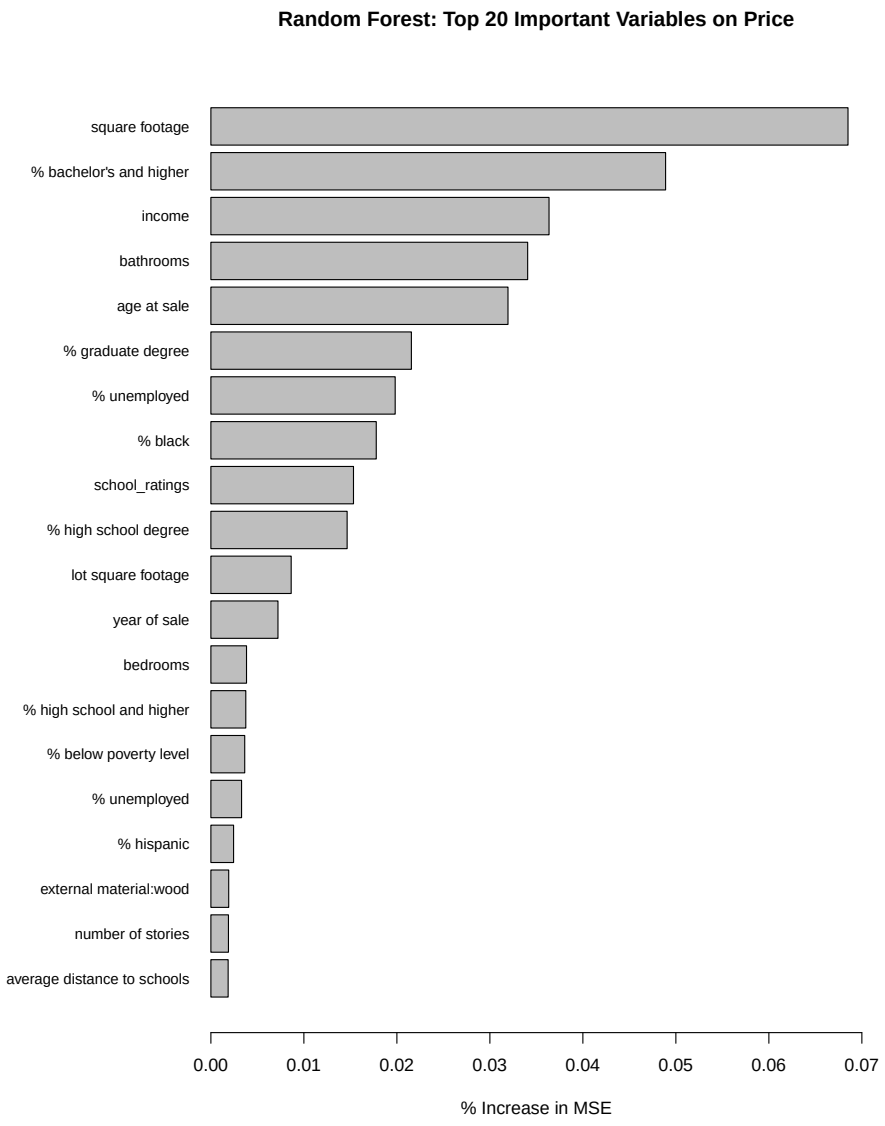
the sense that a select few variables drive a majority of the model. Random forest assigns more even weights to variables due to its decorrelating nature of randomly selecting a limited set of variables for each split. Post-Lasso has the lowest accuracy out of all the models,

Figure 1.10: Important Predictors for Price in Boosting




mainly due to the sparsity imposed on the model. It is likely that the true models depends

Figure 1.11: Important Predictors for Price in Random Forest




on more variables. Changing the tuning parameters (i.e lowering the penalty term $\lambda$) can increase the number of variables selected and slightly improve out of sample predictions. However, exploring interactions and nonlinearities in post-Lasso is still cumbersome since new variables must manually be created. The automation of these tasks makes gradient

boosting and random forest preferable in this context. Post-Lasso is a useful tool when there are no clearly defined control and treated group or if simultaneous interpretation of several variables is desirable.

There are several explanations for the overall lower prediction accuracy and the decrease in performance in random forest as more variables are added. The first reason is because a majority of the variables are words (1677 out of 1893) which are likely to be sparse with many zeros and irrelevant for prediction. The random forest algorithm cuts a tree using a random sample of 1/3 of the total number of variables at each iteration to prevent overfit. Cases where a majority of the remaining variables are words, the models are less predictive. This is reflected in the out-of-sample accuracy table and explains why the random forest does not improve with additional covariates as noted by [40]. This number of variables used in each tree can be increased but with many variables this significantly increases computation time and will increase overfit.

The second reason is that random forest does not handle missing variables. Many homes do not provide detailed listings, resulting in a lot of missing values. I replace missing values with zeros when necessary, but this artificially shifts the distribution of certain variables towards zero and prevents the addition of variables with many missing values from improving prediction. Imputation is a potential solution but many times this will assign the mean values to missings causing another type of bias[14].

The ability of gradient boosting to handle both of these issues as well as obtaining the highest prediction accuracy rate makes it the preferred model. Boosting makes cuts on variables using only those with non-missing and thus utilizes the data only where there is information. This allows the dataset to be larger because it avoids dropping observations that have missing values. This is particularly sensible for large datasets with many controls where a majority of observations have at leave one missing value.

---

[14]R offers a missForest package that imputes missing values. However this package increase computation time exponentially without much gain in prediction accuracy.

### *1.7.2 Treatment Effects*

Average treatment effect estimates are reported for varying levels of $PM_{10}$ in table 1.7 and table 1.8. Standard errors of predicted average treatment effects are consistent with those calculated for out of sample RMSE for model accuracy. Since gradient boosting is the most accurate model, I shows the changes in treatment effects as more variables are added to it in table 1.7. This is also shown in figure 1.12. The results show that the addition of more structured variables correct for the fact that homes in more polluted areas tend to have higher quality homes. This bias is evident in all three types of data added to the model: additional structured variables, images, and words. Controlling for these increase the magnitude of the negative treatment effects, and in most cases, more than double the magnitude.

I take the model with all the variables and compare them to the treatment effects from post-Lasso and gradient boosting. These results are shown in table 1.8. Most of the estimates are negative with a select few having a perverse sign, but these are close to zero and not statistically significant. Post-Lasso leads to very sparse models with few variables resulting in low prediction accuracy and likely overestimates of the negative effects of air pollution. The boosting model is the most intuitive with confidence bounds below zero and mean treatment effects increasingly negative with increasing levels of $PM_{10}$ for majority of the pollutant brackets. A graph of the average treatment effects with standard errors is plotted in figure 1.13. At mean house price of \$130,000 (2010 dollars), these estimates roughly correspond to a 1% decrease in price for every 1 $\mu g/m^3$ increase in $PM_{10}$ or about \$1300.

It is likely that this estimate also includes the effects of $SO_2$ since the two pollutants are highly correlated and it is difficult to find separate control and treatment groups that would disentangle the two effects. In an attempt to see the marginal effects of each, I regress the predicted treatment effects on the two pollutants for all three models and obtain statistically significant estimates for both pollutants. Results are in the following section in table 1.9.

Table 1.7: Boosting Model Average Treatment Effects by $PM_{10}$ Levels

| | $PM_{10}$ $\mu g/m^3$ levels | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variables | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Basic | 0.011 | -0.006 | 0.013 | -0.025 | 0.009 | -0.061 | -0.069 | -0.066 | -0.122 |
| | (0.010) | (0.010) | (0.010) | (0.010) | (0.010) | (0.013) | (0.014) | (0.019) | (0.027) |
| All structured | -0.017 | -0.039 | -0.038 | -0.050 | -0.015 | -0.062 | -0.064 | -0.075 | -0.099 |
| | (0.012) | (0.011) | (0.012) | (0.012) | (0.012) | (0.015) | (0.016) | (0.022) | (0.032) |
| +pic | -0.021 | -0.038 | -0.034 | -0.054 | -0.020 | -0.064 | -0.066 | -0.078 | -0.102 |
| | (0.012) | (0.011) | (0.012) | (0.012) | (0.012) | (0.015) | (0.016) | (0.022) | (0.032) |
| +words | -0.025 | -0.043 | -0.034 | -0.057 | -0.023 | -0.069 | -0.073 | -0.089 | -0.115 |
| | (0.012) | (0.011) | (0.012) | (0.012) | (0.012) | (0.015) | (0.016) | (0.022) | (0.032) |
| N | 1906 | 2575 | 2712 | 2398 | 2156 | 1342 | 1021 | 617 | 317 |

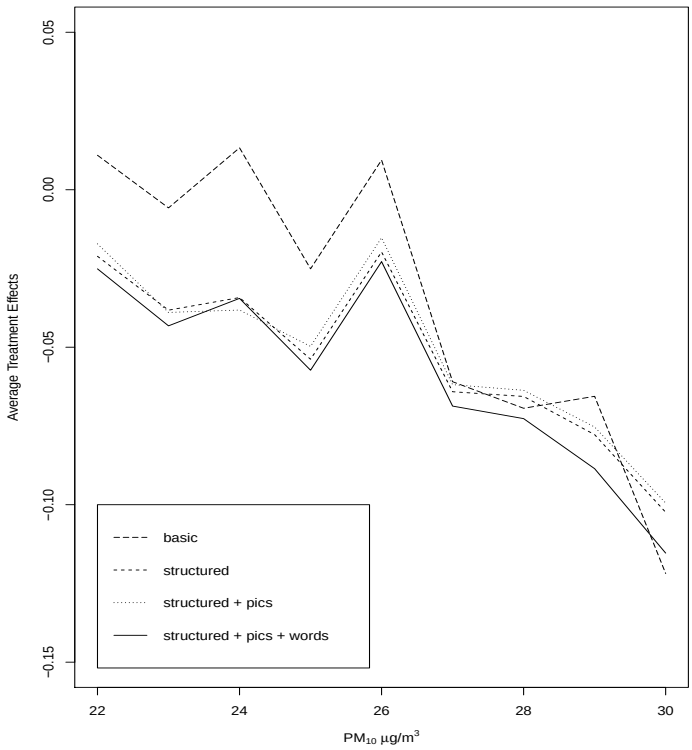
Note: Standard errors in parenthesis. Basic variables include bedrooms, bathrooms, square footage lot size, and census block level demographics.

Table 1.8: Model Comparison: Average Treatment Effects by $PM_{10}$ Levels

| | $PM_{10}$ $\mu g/m^3$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Post-Lasso | -0.041 | -0.069 | -0.047 | -0.023 | 0.019 | -0.054 | -0.007 | -0.027 | -0.016 |
| | (0.013) | (0.013) | (0.014) | (0.014) | (0.013) | (0.017) | (0.019) | (0.024) | (0.036) |
| Boosting | -0.025 | -0.043 | -0.034 | -0.057 | -0.023 | -0.069 | -0.073 | -0.089 | -0.115 |
| | (0.012) | (0.011) | (0.012) | (0.012) | (0.012) | (0.015) | (0.016) | (0.022) | (0.032) |
| Random Forest | 0.007 | -0.015 | -0.017 | -0.047 | -0.012 | -0.042 | -0.042 | -0.069 | -0.093 |
| | (0.012) | (0.012) | (0.012) | (0.012) | (0.012) | (0.016) | (0.017) | (0.023) | (0.032) |
| Observations | 1906 | 2575 | 2712 | 2398 | 2156 | 1342 | 1021 | 617 | 317 |

Note: Standard errors in parenthesis.

Figure 1.12: Gradient Boosting: Average Treatment Effects as Variables are Added

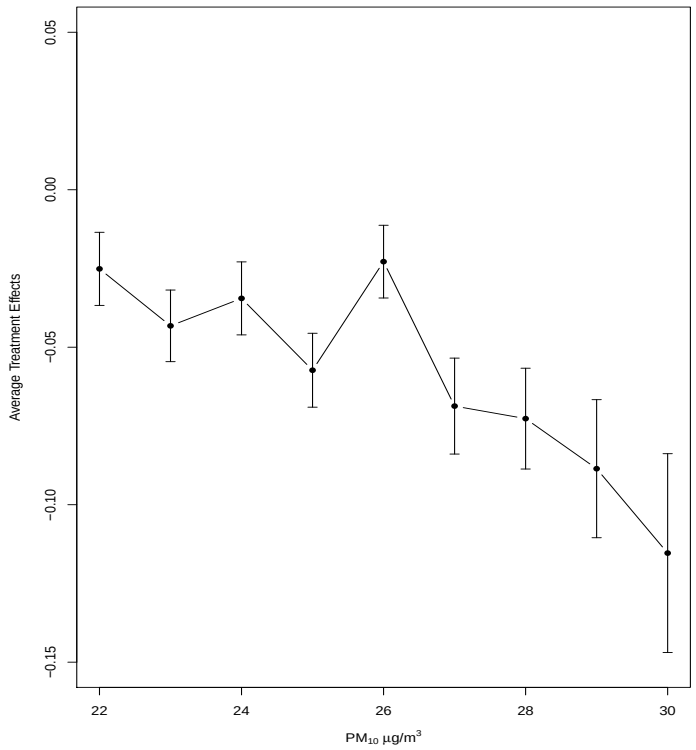



Note: This graph shows how the average treatment effects change as more variables are added to the model. Overall, adding more variables decreases the treatment effects.

### *1.7.3 Heterogeneous Treatment Effects*

In this section, I disentangle the effects of $PM_{10}$ and $SO_2$ and see if there are heterogeneous effects. I regress the residuals, which are interpreted as the treatment effects from each of the three models, on the two pollutants and census block demographics: race, education, income, and employment. Year fixed effects were also included. Results are in table 1.9. All signs are negative on both pollutants, but post-Lasso coefficients are not statistically significant. Gradient boosting and random forest have similar coefficients on the two pollutants and are both statistically significant. By more effectively disentangling the effects of the two

Figure 1.13: Gradient Boosting: Average Treatment Effects with Standard Errors




Note: This graph shows that, in general, average treatment effects decrease with increasing levels of pollutions and are statistically significant.

pollutants. these results are more reasonable than OLS and Lasso method coefficients shown in table 1.5. Taking the coefficients from the boosting and random forest models and using a mean house price of \$130k, an estimated range of MWTP for a 1 $\mu g/m^3$ reduction in $PM_{10}$ is \$650 to \$1,040 in 2010 dollars. Similarly, the MWTP range for $SO_2$ is \$520 to \$780. This is much higher than previous studies but as mentioned earlier, controlling for omitted variables doubles some of the average treatment effects as seem in table 1.7. The sum of the midpoints of the two ranges of MWTP is close to the \$1300 treatment effects using residuals found in section 1.7.2. These estimates are higher than past studies on the MWTP of $PM_{10}$

Table 1.9: Heterogeneous Treatment Effects

| dep var: treatment effect | Post-Lasso | Boosting | Random Forest |
|---|---|---|---|
| $PM_{10}$ | $-0.003$ | $-0.005^{**}$ | $-0.008^{***}$ |
| | (0.002) | (0.002) | (0.002) |
| $SO_2$ | $-0.004^{**}$ | $-0.006^{***}$ | $-0.004^{***}$ |
| | (0.002) | (0.002) | (0.001) |
| % in labor | $-0.001^{***}$ | $0.002^{***}$ | $-0.001^{***}$ |
| | (0.0003) | (0.0004) | (0.0003) |
| % high school degree | $0.001^{*}$ | $-0.001^{*}$ | $-0.001^{***}$ |
| | (0.0003) | (0.0004) | (0.0003) |
| % black | $-0.003^{***}$ | $0.001^{***}$ | $-0.001^{***}$ |
| | (0.0002) | (0.0002) | (0.0002) |
| income ($000s) | $-0.001^{***}$ | 0.0002 | $0.0003^{*}$ |
| | (0.0002) | (0.0002) | (0.0002) |
| Adjusted $R^2$ | 0.070 | 0.015 | 0.030 |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

and $SO_2$. [8] find MWTP of \$130 for 1 $\mu g/m^3$ reduction in $PM_{10}$. [14] find MWTP of \$191 for 1 $\mu g/m^3$ reduction. Before controlling for curb appeal and home description, average treatment effects are, either have the counterintuitive positive sign or significantly smaller in magnitude. Controlling for all these previously omitted attributes can explain the much higher MWTP.

In general, air pollution has a lower effect on homes in areas where there are more minorities but more significant effects on areas where people are more educated. The signs on income and percent in labor flip across the models and definitive conclusions cannot be made. Nevertheless, this has implications on pollution being located in areas where the residents are of lower socioeconomic status. The MWTP numbers could be higher if pollution is located in wealthier areas.

### *1.7.4 Discussion on Bias*

The results show that there are many features that are correlated with air pollution that bias the treatment effects when omitted, but it can be cumbersome to try to identify all these features. In high-dimensional datasets, it is impossible to understand the effect of each dimension independently. Machine learning models, especially ensemble tree models, can lack transparency when many trees are averaged together. However, I will show the directional impact (positive or negative lift) of a handful of important features that can give intuition for the results. Since most of the attributes in the data are binary, I calculate the estimated lift or increase associated with having the feature using the partial plots from the gradient boosting model.

I select 16 important variables and show their partial plots in figure 1.14. Word frequencies are not binary but the plots show that there are no additional effects to using these words more than once, so I report the lifts from using the word once. Central cooling is associated with a $2590 increase in price. Again, these are not interpreted as causal effects but can suggest the general direction of the predicted price used in the model. For example, these results do not propose that using the word "gorgeous" increases the price by $5000 but that homes described as "gorgeous" tends to sell at higher prices. Similarly, houses described as "nice" tend to sell for less. I compare the summary statistics for these 16 features in the control group and the treatment group in table 1.10. The table lists the variables in order of importance based on the gradient boosting model. Overall, the treated group has more variables of higher importance that have a positive lift. The treated group has less homes with the word "gorgeous" and "beauti," but these variables have lower importance and have less impact on the overall predicted price. Many descriptions include names of small neighborhoods that are not reported here but received heavy weights. This is due to the fact that homebuyers associate certain neighborhood names with specific characteristics that cannot be captured by any other variables. My results indicate that the net effect of all important words is positive for homes in the treated group. To check that these results are not just due

to better marketing, I also report the average number of features and appliances listed, and the number of words used in the home description for the two groups. There are no clear differences in these values between the two groups. Looking through the partial plots can provide intuition for the change in average treatment effects as more variables are added to the models. Nevertheless, its is near impossible to make sense of all nearly 2,000 variables.

Table 1.10: Mean Characteristics for Control and Treated Groups

| Variable | Control | Treatment | Lift ($) |
|---|---|---|---|
| % with external material: brick | 0.733 | 0.72 | 6180 |
| % with roof type: slate | 0.315 | 0.454 | 8830 |
| % with structure type: cape cod | 0.184 | 0.181 | -6120 |
| number of zillow photos | 10.342 | 12.607 | >0 |
| parking spaces | 1.62 | 1.647 | 2500/space |
| time since last remodel (years) | 31.452 | 35.397 | $< 0$ |
| % with central cooling | 0.998 | 0.999 | 2590 |
| % with appliance: dishwasher | 0.909 | 0.892 | 6120 |
| % with feature: ceiling fan | 0.584 | 0.645 | 1245 |
| % with structure type: colonial | 0.689 | 0.732 | 100 |
| % with feature: pool | 0.164 | 0.262 | 4660 |
| % with roof type: composition | 0.123 | 0.067 | -2100 |
| word count: "gorgeous" | 0.009 | 0.001 | 5000 |
| word count: "beauti" | 0.037 | 0.007 | 2500 |
| word count: "nice" | 0.025 | 0.006 | -2200 |
| % with feature: jetted tub | 0.138 | 0.043 | 2510 |
| number of features listed | 5.0 | 5.2 | |
| number of appliances listed | 4.9 | 5.1 | |
| number of words used | 12.0 | 11.3 | |

Note: Shown in order of importance. Missing values dropped in mean calculation.

### *1.8 Conclusion*

In the housing market, unobservable house and neighborhood attributes often result in omitted variable bias when estimating the market price of public goods. Previous methods used to control for this bias are insufficient substitutes to directly controlling for omitted attributes.

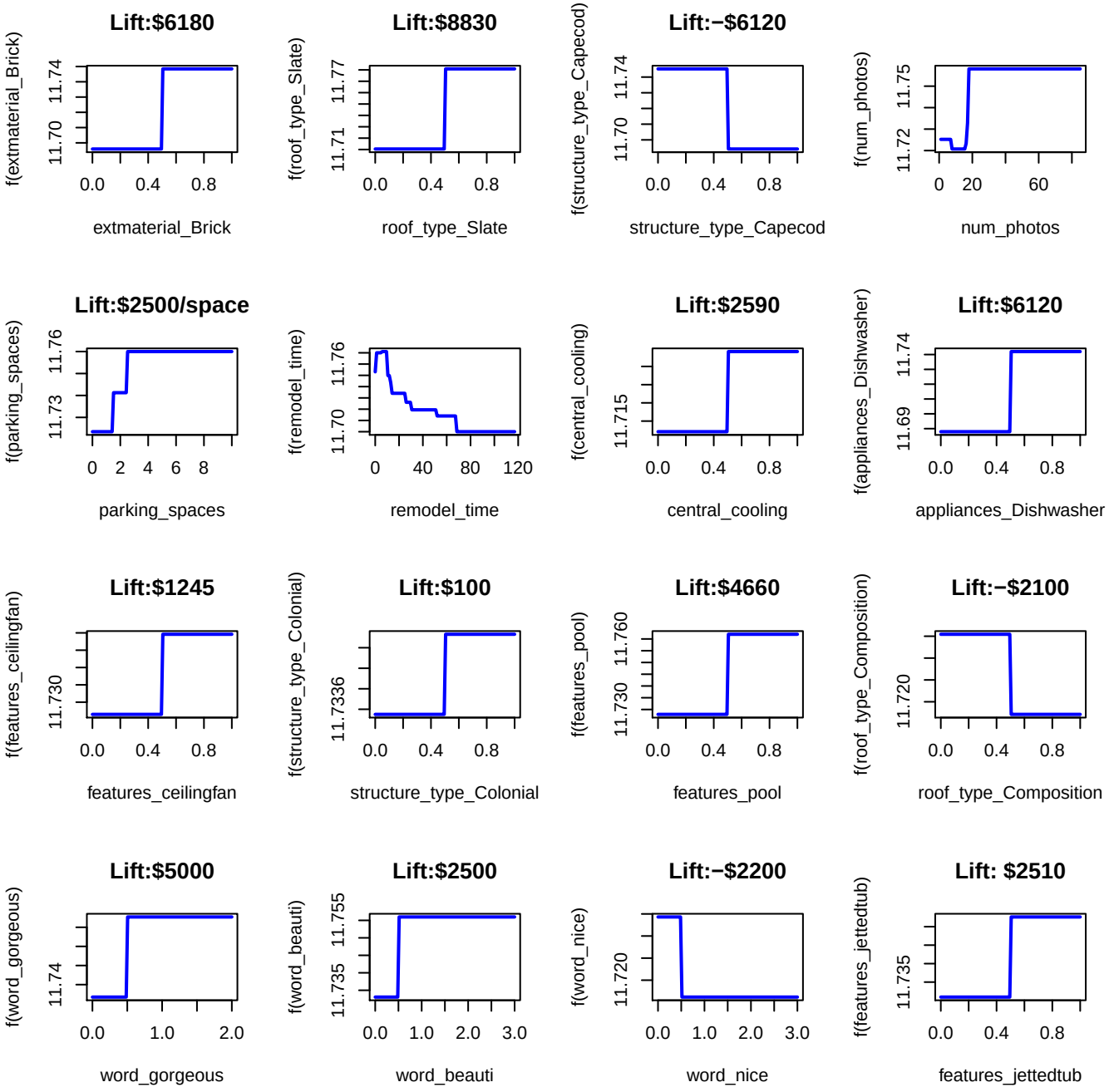
Data relevant to homebuyers that have historically been ignored in hedonic applications are mostly in the form of unstructured text and images found in online property listings such as those on Zillow.com. Home descriptions and pictures are important to consumers, but have not been included in hedonic models due to a lack of methods to conduct feature engineering and to handle multicollinearity common in high-dimensional data. The growth in the available of such data coupled with machine learning methods make possible the use of richer controls to reduce omitted variable bias. The inclusion of more variables can reduce bias and even strengthen methods such as fixed effects and regression discontinuity.

Machine learning techniques provide a medium for the effective use of such data by allowing for more flexible models that are robust to irrelevant variables and multicollinearity. They conduct variable selection and automatically explore nonlinearities and interaction across variables for effective model selection. My results show how nonlinear tree-models can significantly improve out-of-sample prediction accuracy and yield more accurate estimation of counterfactuals. I collect a unique and comprehensive dataset on house characteristics and integrate the econometrics of hedonic models and machine learning to estimate the implicit price of air quality. This paper demonstrates useful applications of machine learning techniques as well as tools to include raw text and images as input variables.

These results highlight the importance of controlling for a rich set of attributes when estimating the effects of air quality on house prices. Omitting them results in underestimates of the negative effects of air pollution by more than half. This emphasizes the important of using more data in any environmental and urban policy analysis. Policy implementations and analyses are innately high-dimensional problems. Policy makers need to incorporate a large number of factors when suggesting changes and predicting the effects. Thus, there can be significant implications on policy recommendations when better models that control for more attributes are used. While my data controls for a significantly greater number of variables compared to previous papers, it is undoubtedly incomplete and there is no way to control for everything. Nevertheless, this paper encourages researchers and policy makers to incorporate more data into their analysis as methods to accommodate these types of data

are becoming more prevalent.

Figure 1.14: Lift from House Features




Note: Partial plots for top predictors of house prices. The lift or predicted change in price in the model for having each features is shown.

Chapter 2

# THE EFFECTS OF ZONING REGULATIONS ON HOUSING AFFORDABILITY


## Abstract

Relaxing zoning regulations by allowing for denser housing can encourage more efficient land use and more affordable housing. One way to estimate the impact on home prices is to use a boundary discontinuity design. However, this approach fails if unobservable attributes that affect price are not random at the boundary. Using a high-dimensional Zillow dataset on house transactions in Seattle, I apply a variety of machine learning methods that are suitable for high-dimensional data. In this paper, I find that a large number of variables that have been ignored in previous studies are discontinuous at zoning boundaries. My results show that the single family zoning results in 5% increases in home prices, a number significantly smaller than estimates when limited data and standard linear models are used.

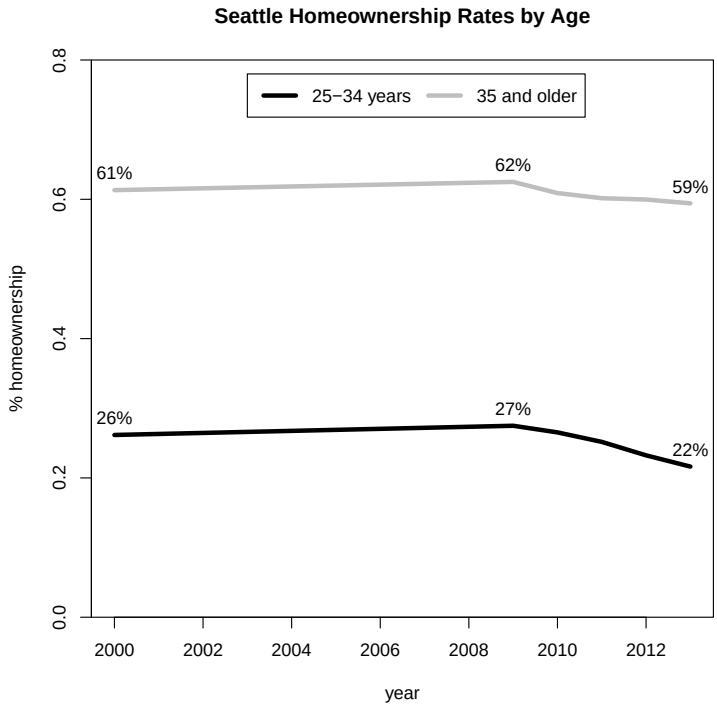
## *2.1 Introduction*

Obtaining affordable housing in major US cities has been a perpetual issue for the last 30 years. One factor contributing to rising prices is land-use regulation, particularly minimum lot-size zoning. Many studies look at how this restricts the supply of housing resulting in higher home prices. However, it has been less studied how zoning is associated with other home characteristics such as household characteristics and home quality which also significantly influence price. Households that wish to live in single family zones tend to be more focused on raising children and therefore strategically choose to live in locations close to public parks and better school systems. They might also be wealthier and invest in higher quality homes with upgraded interiors like remodels kitchens or bathrooms. In other words, zoning reflects a slew of housing and neighborhood amenities that can increase the price of homes separate from price increases purely from supply restrictions. Failing to account such outcomes overstates the estimated direct price impacts of zoning.

Seattle is an example of a city with burgeoning demand and was labeled by the Census Bureau as the fastest growing populous city from 2012 to 2013 at 2.8%. For individuals between 25 and 34 years old, the fraction of homeowners has been decreasing since 2008 as seen in figure 2.1. This can partially be attributed to the 2008 recession and stricter credit constraints but for the same period, the fraction of individuals older than 34 years that are homeowners decreased by much less. Figure 2.2 shows Seattle's median sale price for all homes since 1996. Higher prices will only further exacerbate homeownership rates. One potential reason for these dramatic increases is that homes in Seattle are heavily restricted by land use regulations such as minimum lot sizes causing stagnant housing supply at a time of rising demand. In this paper, I estimate the impact of such zoning regulations on house prices.

The main contribution in this paper lies in using a rich dataset from Zillow with high-dimensional controls to model house prices to reduce omitted variable bias. I apply a variety of machine learning methods that handle big data well. Adoption of these methods in applied

Figure 2.1: Home Ownership by Age




microeconomics is growing as large datasets have become more accessible. Due to the lack of rich data on house structures and local amenities, previous studies in urban economics have only used a handful of features to approximate homebuyers' decisions leading to omitted variable bias. Even when presented with many controls, researchers often rely on industry knowledge and select a small set of variables. Using machine learning methods allows for data driven ways to perform model selection in the presence of irrelevant variables. Variable selection in this manner cannot replace expert intuition of the underlying economic factors, but it can certainly complement it.

It has been well established in the literature that such zoning regulations increase the price of homes ([26] and [42]) due to supply restrictions, previous studies control for only a handful of house attributes. While zoning regulates the lot size of homes, this is highly correlated with many other housing attributes such as finished square footage as well more

Figure 2.2: Median Home Sale Prices


parking spots and the presence of a garden. All these additional attributes contribute to higher prices and failing to account for them would overestimate the direct impact of zoning on home prices. I exploit a boundary discontinuity design to estimate the treatment effects of minimum lot size zoning on housing prices using a high-dimensional set of housing data collected from Zillow. My results show that there are many attributes that are correlated with zoning regulations and properly controlling for them results in treatment effect estimates on home prices that are a third the magnitude at 5%. The main contribution of this paper is the use newly available big data and the application of machine learning techniques within a standard quasi-experimental tools, specifically regression discontinuity, to reduce omitted variable bias.

This paper is organized as follows. In section 2.2, I discuss the relevant literature. Section 2.3 explains the data. Section 2.4 explains the estimations methods and machine learning.

Section 2.5 is the results, and I conclude in section 2.6.

## *2.2 Relevant Literature*

This paper touches on two fields of literature: land-use regulation and inference using high-dimensional datasets. Many researchers have discussed how land use regulations impact development and housing using much smaller datasets. [26] and [42] tackled the question of the impact of minimum lot requirements. Glaeser et al. find that regulation decreases the issuance of new permits. This restricts supply causing what researchers call a "zoning tax" resulting in higher home prices. Zabel et al. estimate the impact on house prices by looking at the change in prices for homes before and afters changes in regulation and find increases on home prices of up to 20%. [24] also find that lack of affordability in the housing market is largely due to strict supply regulations as opposed to land scarcity.

Zoning regulations are heavily influenced by local politics since changes are determined by city council. City councilman are selected by residents in each district with each district having one representative. Thus, zoning regulation are largely set by the residents within those zones and can be used as a way to eliminate free riding. By increasing the price of homes, this ensures that neighborhood residents pay for their consumption of neighborhood amenities and school systems in the form of property taxes. Many papers have discussed the use of zoning as a way to keep neighborhoods homogeneous and restrict low income households from moving into richer areas ([5] and [26]).

This paper fits into a growing literature on using big data in urban economics. The increase of available datasets such as Yelp and Zillow can be used to better control for city characteristics and property characteristics. Nevertheless, controlling for all possible characteristics is infeasible and the use of natural experiments remains necessary for unbiased results. The use of boundary discontinuity design is well established starting with [11] looking at the difference in house prices across school zone boundaries. A similar approach was done in [7] to highlight the presence of households sorting themselves at the boundary. My approach in this paper can thus be seen as a demonstration of new big data techniques to

traditional quasi-natural experiments, specifically boundary discontinuity design.

## *2.3 Data*

### *2.3.1 Census Data*

Demographic data is collected from the American Community Survey. I match each house to a census tract. For each census tract, the median income, family size, estimated commute time, education levels, and race are included.

### *2.3.2 Crime Data*

Crime data in Seattle is made publicly available by the Seattle Police Department and can be downloaded from Socrata. The dataset includes the time and date of the crime, the type of crime, and longitude and latitude coordinates of the incident. I categorize the types of crimes into three groups: violent crime, property crime, and other. These groupings are consistent with those done by the Seattle Police Department. For each property, a circle with varying radii of 0.5km, 1km, 2km, 3km, 4km, 5km with the total number of violent, property and other crimes occurring within that radius that occurred in the 6 months before the date of purchase.

### *2.3.3 Housing Data*

I collect house transaction details from Zillow for 2010-2016. The home structure covariates are rich with more than 100 structural variables beyond the standard number of bedrooms, number bathrooms, finished square footage, and lot size. The full list features is shown in 2.1. In terms of unstructured variables, I include home descriptions as posted by sellers on Zillow. The descriptions are turned into a bag-of-words model. Each unique words becomes a variable taking on the frequency of that word in a specific property description as the value. This approach can highlight amenities that are not listed elsewhere such as access to local parks or grocery stores.

Table 2.1: Summary of Rich Structural Variables

| Feature category | Values |
|---|---|
| Architecture type | bungalow, cape cod, colonial, contemporary, craftsman, french, georgian, loft, modern, queen anne victorian, ranch rambler, santa fe pueblo style, spanish, split level, tudor |
| Features | attic, barbecue, basketball court, cable ready, ceiling fan, controlled access, deck, disability access, dock, doorman, double pane, fenced yard, fireplace, fitness, garden, gated entry, greenhouse, high speed internet ready, hot tub, intercom, jetted tub, lawn, mother-in-law, patio, pool, porch, sauna, security system, skylight, spa, sportscourt, sprinklers, storage, storm windows |
| Heating type | baseboard, forced air, heat pump, radiant, stove, wall |
| Heating source | coal, electric, gas, oil, propanebutane, solar, woodpellet |
| Rooms | breakfast nook, dining, family, laundry, library, master bath, mud, office, pantry, recreation, solarium atrium, sun, walk-in closet, workshop |
| Roof type | asphalt, built up, composition, metal, shake, shingle, slate, tile |
| Parking type | carport, garage-attached, garage-detached, off-street, on-street |
| View | city, mountain, park, territorial, water |

A unique feature of this data is that it includes details on local amenities such as access to transportation, grocery stores, school quality, etc. School quality data is from GreatSchools.org where schools are rated on a scale of 1-10 base on state standardized test performance compared to other schools in the same state. Walk score, transit score, and bike score are from walkscore.com. This service maps a location to its access to completing daily tasks via the three types of transportation.

My data has 17,130 transactions and includes homes on lowrise and single family zones. Lowrise homes sit on a minimum square foot lot size of 1600 sqft; they include rowhouses, townhomes, and cottages. Single family houses are detached homes with three minimum lot size zones: 5000, 7200, and 9600 sqft. Map details on zoning restriction boundaries is recovered from data.seattle.gov and can be seen in figure 2.4. I use the map to assign zones

to the geocoded addresses of homes in the data. A summary of the dataset based on the zone and minimum lot size is shown in table 2.2. The difference in the structural characteristics across zones is apparent. Homes on larger lots are more spacious with a higher number of bedrooms and interior square footage.

## *2.4 Estimation Methods*

### *2.4.1 Boundary Discontinuity Design*

There have not been significant land use regulation changes in Seattle, making it difficult to study the impact of zoning under the settings of a natural experiment. Sorting across minimum lot size zones lead to biased treatment effect estimations when just using zoning fixed effects since households with strong distaste for denser neighborhoods would strategically locate farther from zone boundaries. Even with a rich set of controls for neighborhood amenities, there are still significant factors such as noise, community culture, overall neighborhood attractiveness, etc. that are unobservable to econometricians. Using simple zoning fixed effects would still produce confounding treatment effect estimates. Furthermore, studies that attempt to use a different-in-difference approach require no time-varying changes. This requirement is rarely met as demographics are likely to change as zoning regulations change. An appropriate alternative is to use a boundary discontinuity design within a hedonic home price model to look at homes close to zoning boundaries but on opposite ends of the boundary. This controls for the omitted variables that are correlated with distance from the zoning boundary.

Evidence of the importance of these variables can be seen in figure 2.3. There is clear sorting with single family homes to the right of the boundary more expensive the farther they are. The opposite is true for lowrise homes to the left of the boundary. I find evidence that this is due to the fact that city centers are farther from these boundaries and households located there are willing to pay for the convenience. The boundary discontinuity regression takes this form

$$ln(P_{iht}) = \alpha_0 + X'_{iht}\beta + \gamma D_{ih} + \theta_h + \mu_{bh} + q_t + \epsilon_{iht} \tag{2.1}$$

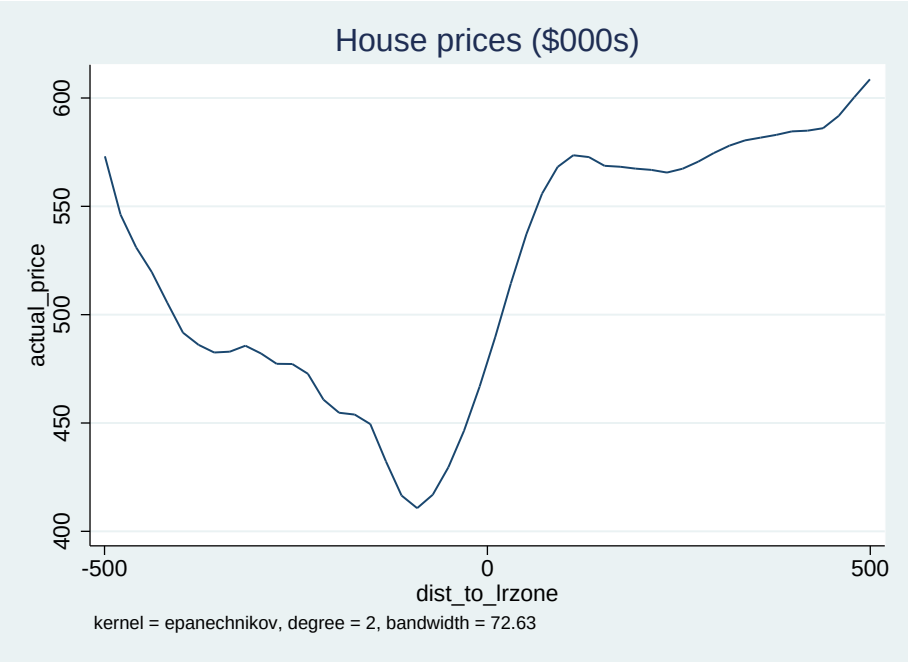
Table 2.2: Summary Statistics

| Variable | Lowrise 1600 | Single Family 5000 | Single Family 7200 |
|---|---|---|---|
| Price (in $1000s) | 464.14 | 620.71 | 551.03 |
| | (181.75) | (312.50) | (319.56) |
| Lot size sqft | 2493.49 | 5162.36 | 8139.26 |
| | (2417.20) | (1891.83) | (2400.70) |
| Finished sqft | 1603.63 | 2185.43 | 2200.49 |
| | (549.12) | (832.58) | (904.41) |
| Bedrooms | 2.85 | 3.31 | 3.37 |
| | (0.72) | (0.93) | (0.88) |
| Bathrooms | 2.40 | 2.14 | 2.15 |
| | (0.79) | (0.90) | (0.89) |
| % fireplace | 0.68 | 0.75 | 0.83 |
| | (0.47) | (0.43) | (0.38) |
| Transit score | 54.97 | 50.42 | 45.45 |
| | (8.03) | (8.37) | (6.69) |
| School rating | 7.52 | 7.63 | 7.21 |
| | (1.41) | (1.65) | (1.47) |
| % white | 0.72 | 0.77 | 0.77 |
| | (0.19) | (0.20) | (0.15) |
| % bachelors | 0.37 | 0.35 | 0.33 |
| | (0.11) | (0.10) | (0.08) |
| % graduate degree | 0.23 | 0.27 | 0.23 |
| | (0.11) | (0.13) | (0.12) |
| Median income | 67.03 | 89.97 | 80.98 |
| | (21.11) | (32.75) | (29.80) |
| Violent crimes within 1km | 129.41 | 86.28 | 73.59 |
| | (131.53) | (72.37) | (52.62) |
| Property crimes within 1km | 712.60 | 512.94 | 419.86 |
| | (580.10) | (379.17) | (292.05) |
| Observations | 1845 | 7215 | 1933 |

Notes: standard deviations in parenthesis

where $ln(P_{iht})$ is the log price of home $i$ in boundary $h$ at time $t$; $X_{iht}$ is a vector of structural home attributes, $D_{ih}$ is a binary zoning variable, $\theta_h$ is a matrix of boundary dummies, $\mu_{bh}$ represents neighborhood $b$ fixed effects that do not vary over time, and $q_t$ represents time fixed effects.

Figure 2.5: House Prices as Distance to Boundary




### *2.4.2 Double Machine Learning*

Use of high dimensional datasets in standard linear regression poses two main problems: 1) a high ratio of variables to observations, if not more, and 2) the presence of irrelevant variables. In fact, given a high number of irrelevant variables, appropriate model selection is a challenge. The econometricians may obtain expert knowledge on a subset of the variables and have intuition on the most important variables. However, the importance of a large set of variables as well as the parametric model is much more problematic. Machine learning methods can handle both of these problems. They allow for nonparametric estimations and conduct variable selection. This is particularly useful in regression discontinuity because the

analysis includes only observations within a boundary threshold leaving fewer observations in the final regression.

Typically, omitted variable bias can be reduced by using observations very close to the boundary but this is costly because it leads to less degrees of freedom when observations farther from the boundary are discarded leaving few observations. Machine learning makes it less costly to discard observations, first by conducting variable selection and by allowing for more controls that directly reduce omitted variable bias.

Nevertheless, there remains a potential caveat to machine learning approaches when estimating parameters for causal inference. Variables are selected based on how well they predict the outcome variable. Thus, variables that are highly correlated with the treatment variable might be omitted from the final analysis and this results in omitted variable bias. I use a method called double machine learning introduced by [15]. Consider the following simple model:

$$ln(P_i) = \beta d_i + g_0(x_i) + \xi_i \tag{2.2}$$

$$d_i = m_0(x_i) + \nu_i \tag{2.3}$$

where $E[\xi_i|d_i, x_i] = E[\nu_i|x_i] = 0$ for observation $i = 1, 2, ..., n$, $ln(P_i)$ is the log price for home $i$, $x_i$ is a vector of $p$ controls, $d_i$ is a scalar treatment variable, and $\beta$ is the parameter of interest. A standard prediction model would conduct variable selection on equation (2.2) to find variables, $x_y$, that predict $y$ well. However, this can lead to omitted variable bias when there are variables not in $x_y$ that influence $d$. By using (2.3) to obtain $x_d$ and including variables that are predictive for the treatment in the final model, I can obtain unbiased estimates of $\beta$. [15] utilize a variety of machine learning methods that are good for prediction and can avoid overfit by using cross-validation. The standard cross-validation approach is where one machine learning method is used and the sample is split and predictions are made using the same machine learning method. They propose to use a $k$ ensemble of machine learning methods and then selection the machine learning method that predicts best out of

sample.

## 2.5 Results

### 2.5.1 Treatment Effects and Model Comparison

In this section, I show the treatment effect estimates when a high-dimensional set of attributes are added. I compare the results of using a limited set of variables that are used in previous studies to using a set of high-dimensional controls collected from Zillow. My results demonstrate how insufficient the use of a limited set of property attributes is when looking at the impact of zoning on home prices. Specifically, treatment effect estimates are significantly smaller than those found in previous studies and are one third the magnitude of the estimates in Seattle when using a limited set of controls.

In table 2.3, I show the results of using a standard set of attributes that are common in previous studies ([26]and [42]). This includes number of bedrooms, number of bathrooms, finished square footage, number of floors, school quality, census tract variables (race, education, crime rates, and average transit time), neighborhood and boundary fixed effects, and year sold fixed effects. As I vary the distance to the boundary, the treatment estimates generally increase the farther away from the boundary, but at 0.2 miles, the difference in price is the greatest at 0.1614. Intuitively, this is likely due to an optimal mix of single family homes being far away from density and noise but still close to transit as this jump in treatment effects decreases once these variables are added to the model. These estimate are much larger than the estimates once I control for a rich set of attributes as seen in the following set of results.

I report the results of double machine learning [15] under a variety of machine learning methods (double selection lasso ([10]) , random forest, and boosting) in figures 2.4 and 2.5. It is clear flexible nonparametric tree methods yields more predictive models. I report the mean squared error (MSE) for predictions out of sample for price and the treatment variable, single family zoning. These methods are more predictive because they allow for more controls and interactions across variables in the models. Treatment effects are on log price and can

Table 2.3: OLS Comparison

| ATE on log(price) | 0.1 miles | 0.2 miles | 0.3 miles | All |
|---|---|---|---|---|
| basic | 0.1481 | 0.1614 | 0.1574 | 0.1570 |
| | (0.0088) | (0.0077) | (0.0075) | (0.0074) |
| + structured | 0.1358 | 0.1500 | 0.1469 | 0.1455 |
| | (0.0091) | (0.0080) | (0.0076) | (0.0075 ) |
| N | 7,136 | 10,678 | 12,518 | 14,036 |

Notes: standard errors are in parenthesis

be interpreted as percentage change in price. In figure 2.4, I use only structured variables. The results from adding the bag-of-words are reported in 2.5.

Boosting yields the lowest MSE predictions on test sets. It also gives the most intuitive results with treatment effect estimates increasing with distance from the boundary, indicating the validity of a BDD approach. The results show that even with a high-dimensional set of controls and machine learning methods, there is still the presence of omitted variable bias as evident in the increasing treatment effect estimates as observations get closer to the boundary. Thus, the use of big data and machine learning complement the use of a boundary discontinuity design to minimize bias. The final estimates in 2.5 from the boosting model indicate that the impact of zoning on home prices is about 4.7%. This is much smaller than the 7.6% treatment effect estimate from table 2.5 and significantly smaller than the estimates in table 2.3 for observations within 0.1 miles from the boundary.

### *2.5.2 Discussion*

Boundary discontinuity is used to control for omitted variable bias. However, when there are omitted variables that influence price and are also discontinuous at the boundary, treatment effect estimates remain biased. The assumed randomness at the boundary is violated. Controlling for these omitted attributes by including a high dimensional set of controls can further reduce bias but require the use of machine learning methods. Example of variables

Table 2.4: Model Comparison: Structured Variables

| ATE | 0.1 miles | 0.2 miles | 0.3 miles | All |
|---|---|---|---|---|
| Double-selection | 0.1243 | 0.1215 | 0.1266 | 0.1367 |
| | (0.0139) | (0.0117) | (0.0113) | (0.0110) |
| MSE$[Y\|X]$ | 0.3947 | 0.3687 | 0.3593 | 0.3519 |
| MSE$[D\|X]$ | 0.3719 | 0.3477 | 0.3352 | 0.3221 |
| Forest | 0.0515 | 0.0748 | 0.0722 | 0.0820 |
| | (0.0167) | (0.0147) | (0.0141) | (0.0140) |
| MSE$[Y\|X]$ | 0.3682 | 0.3474 | 0.3368 | 0.3317 |
| MSE$[D\|X]$ | 0.3244 | 0.2921 | 0.2767 | 0.2652 |
| Boosting | 0.0756 | 0.0877 | 0.0927 | 0.1025 |
| | (0.0165) | (0.0147) | (0.0153) | (0.0144) |
| MSE$[Y\|X]$ | 0.3520 | 0.3335 | 0.3231 | 0.3152 |
| MSE$[D\|X]$ | 0.3225 | 0.2853 | 0.2670 | 0.2575 |
| N | 7,136 | 10,678 | 12,518 | 14,036 |

Notes: standard errors are in parenthesis

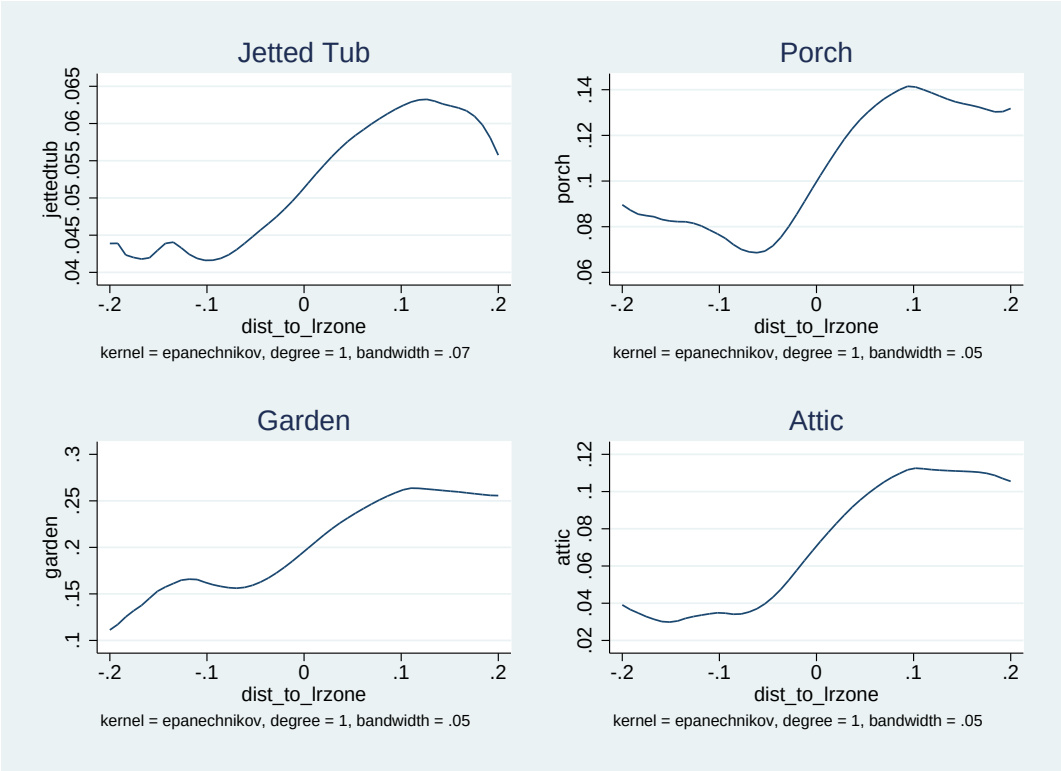
Table 2.5: Model Comparison: All Variables

| ATE | 0.1 miles | 0.2 miles | 0.3 miles | All |
|---|---|---|---|---|
| Double-selection | 0.1022 | 0.1003 | 0.1017 | 0.1110 |
| | (0.0162) | (0.0141) | (0.0136) | (0.0132) |
| MSE$[Y\|X]$ | 0.3761 | 0.3521 | 0.3428 | 0.3349 |
| MSE$[D\|X]$ | 0.3313 | 0.3051 | 0.2943 | 0.2831 |
| Forest | 0.0398 | 0.0579 | 0.0527 | 0.0594 |
| | (0.0188) | (0.0167) | (0.0157) | (0.0161) |
| MSE$[Y\|X]$ | 0.3714 | 0.3510 | 0.3401 | 0.3344 |
| MSE$[D\|X]$ | 0.3028 | 0.2714 | 0.2573 | 0.2457 |
| Boosting | 0.0472 | 0.0631 | 0.0739 | 0.0821 |
| | (0.0190) | (0.0166) | (0.0165) | (0.0158) |
| MSE$[Y\|X]$ | 0.3538 | 0.3321 | 0.3235 | 0.3149 |
| MSE$[D\|X]$ | 0.2913 | 0.2612 | 0.2493 | 0.2418 |
| N | 7,136 | 10,678 | 12,518 | 14,036 |

Notes: standard errors are in parenthesis

that influence price and are also discontinuous at the boundary are shown in figure 2.6.Previous studies control for number of bedrooms and finished square footage but not for the presence of a garden or attic which are clearly correlated with minimum lotsize and increase the price of a house.

Figure 2.6: Discontinuous at Boundary Variables




The interpretation of these results remains complex. Larger lot sizes lead to bigger homes possibly resulting in the sorting of households (households with more children live in bigger homes) as well as more property amenities such as an attic or a pool. Analyzing how zoning directly impacts these factors influence overall home prices is left for future research.

## *2.6 Conclusion*

This paper demonstrates how machine learning can be used to incorporate high-dimensional data and nest standard econometric methods such as boundary discontinuity design. These methods reduce the possibility of errors made by researchers when selecting a functional form. They also allow for a systematic way to identify which variables are highly correlated and select the ones that are most linked to treatment variables. Controlling for more attributes and using machine learning models reduce bias by more than half. Failing to control for these attributes results in an overestimation of the impact of minimum lot-size zoning. The results indicate that using boundary discontinuity design in combination with machine learning minimizes bias, and future studies asking similar urban policy questions would benefit from this approach.

Figure 2.3: Map of Seattle Homes

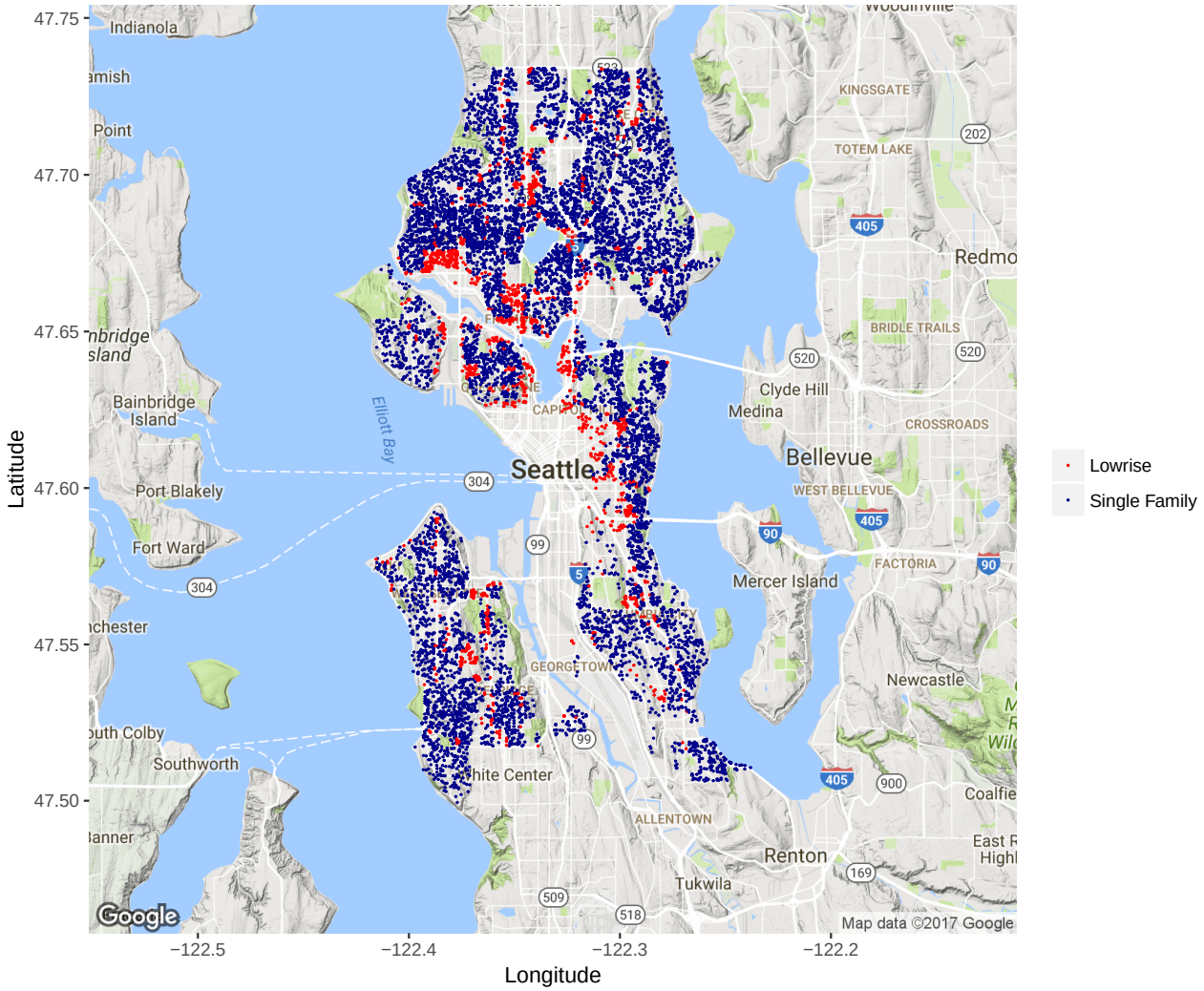
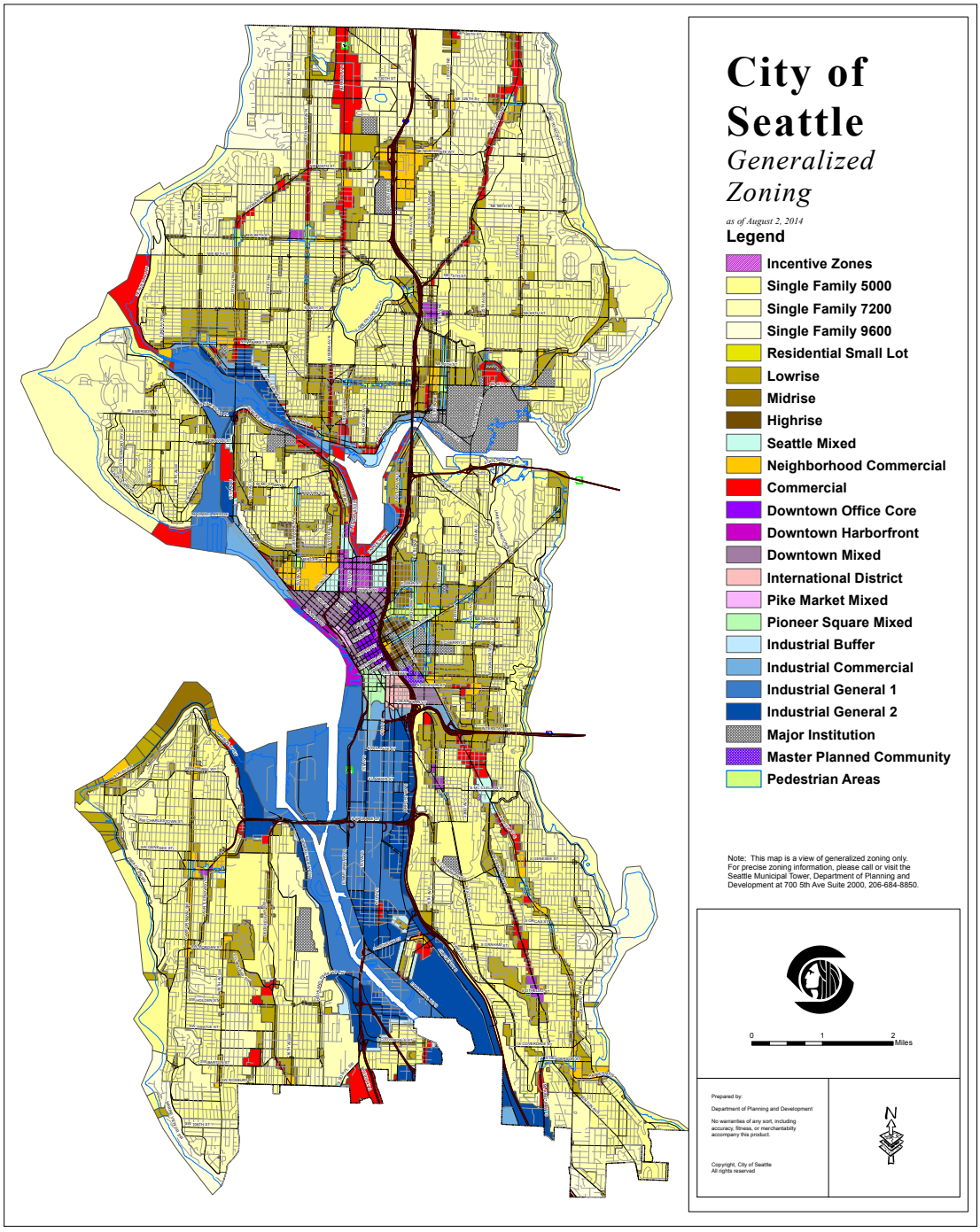


Figure 2.4: Seattle Zoning Map



# Chapter 3

# PICTURES AS REGRESSORS: ESTIMATING CURB APPEAL


## Abstract

When controlling for product attributes in hedonic models, economists typically use a narrow set of attributes. This limitation is largely due to a lack of processing tools or unobservable data leading to poor models or omitted variable bias. Within housing hedonics, the concept of curb appeal is a home attribute that is of importance to buyers. Yet, no previous study has controlled for it. In this paper, I apply computer vision tools to extract features from curbside view images of homes posted on Zillow.com. The features are used to model curb appeal, which can then be used in hedonic home price applications.

## *3.1 Introduction*

Curb appeal is one of the most important attributes of a home and is one of the first attributes of a home that buyers look at. However, in the housing hedonics literature, no study has attempted to control for the curb appeal of homes. This is of particular importance in policy analysis when omitted variable bias leads to biased treatment effect estimates. Omitted attributes lead to models that predict home prices poorly and omitted variable bias when curb appeal is correlated with any of the included regressors. The curb appeal of one particular home is generally highly correlated across homes within one neighborhood. Other neighborhood attributes such as household income, employment or school quality will be correlated with curb appeal. In this paper, I present an approach to transform pictures into regressors, specifically curbside view images into curb appeal ratings. The regressors can then be used as controls in policy analysis or predictive models. I use Amazon's Mechanical Turk - an online scalable marketplace for tasks that require human intelligence - to have workers rate the curb appeal of a home by looking at a picture. I then borrow computer vision techniques to extract image features to model curb appeal.

In this paper, I perform feature engineering on images to create variables to control for home attributes. I use a variety of tools from computer vision such as wavelets and segmentation to create informative variables, also known as feature engineering. While these technique for feature engineering are not new, this approach to transforming image features into controls within regressions is relatively new in economics. A similar approach is done with Google street view images can be used to predict income levels in New York City ([25]). [34] explore the relationship between urban appearance and socioeconomic status of inhabitants using street-level images.

The paper is organized as follows. In section 3.2, I explain how the training data is obtained. Section 3.3 details all the features that are extracted. Section 3.4 is the estimation model and the results. I conclude in 3.5.

### *3.2 Training the Data*

Data are collected from Zillow.com with a total of 23,733 homes posted curbside images. Since supervise learning requires labeled outcomes, I submit a training sample of 2,774 images to Amazon's Mechanical Turk to obtain human rated curb appeal for each image. The online marketplace is a popular tool for what are often simple tasks that require human intelligence. I recruit 44 Mechanical Turk workers to rate the curbside appeal of house images on a scale of 1 to 10, with 10 being the most appealing. Each photo was rated by 2 unique workers and the average of the two is used as the final rating. A model is built on this training set and used to predict the rating for all curbside view images.

### *3.3 Feature Extraction*

I use a combination of supervised and unsupervised learning for feature engineering, the process of creating summary variables that describe the data. Supervised learning allows for input values to be mapped to an output. Unsupervised learning entails using just input values that do not correspond to outputs. I extract a total of 43 image features that on their own, are not informative and would be considered unsupervised learning since they effectively summarize the images (e.g. brightness, number of pixels, etc.).

For feature extraction, I follow [16] to estimating the aesthetic appeal of an image but model the curbside appeal of a house instead. I collect 43 of the 56 image features introduced in their paper depending on the relevance to curbside appeal. Transforming images into data is straight forward as images are made up of matrices of pixels. See appendix C for more details. [1] The bulk of the features represent the textures and colors within segmented parts of an image.

Table 3.1 summarizes all the features. I first offer a brief introduction on wavelet analysis

[1]The purpose of Datta et al.'s paper is to model aesthetic appeal within images. The images in their data are diverse (objects, people, scenery, etc). They propose a set of features to capture familiarity of the content shown. This is not useful for curb appeal since these pictures are relatively homogeneous with all images being of the front of homes. I exclude these from my analysis.

Table 3.1: Summary of Image Features

| Features | Description |
|---|---|
| f1 | average pixel intensity |
| f2 | colorfulness |
| f3 | average saturation |
| f4 | average hue |
| f5-f7 | average hue, saturation, and value for the center 1/3 of the image |
| f8-f18 | Daubechies wavelet transform values to measure textures |
| f19 | $X + Y$ size of image |
| f20 | $X/Y$ aspect ratio |
| f21-f43 | segmentation |

and segmentation which account for a majority of the features and then discuss how these features can be used to model curb appeal. Feature engineering allows for diverse approaches and can be as complex or simple as the researcher desires. I demonstrate just one approach from the computer science literature appropriate for the purpose of gauging curb appeal.

Wavelets can be used to analyze textures within images to indicate what types of materials are present in the image (i.e. brick vs. concrete homes vs. grass ). Wavelets, more generally, smooth and capture information about signals, which for images correspond to the magnitude of color and intensity contrasts across pixels. These changes can detect the roughness or smoothness within an image and when analyzed by the patterns of changes, they can detect textures. For my analysis, I calculate the average changes in groups of pixels, also called wavelet coefficients, to capture the intensity of color changes in an image. In the context of images, horizontal as well as vertical changes in color intensity are captured. The pattern of these changes represent a texture. This provides a summary of the texture such as how much grass, brick, or sky is reflected in the photo. I summarize the textures within each image averaging the wavelet coefficients for hue, saturation, and value creating image features f8-f18. See [16] to get more details on how these variables were creates. I use a basic Haar wavelet. A more detailed and mathematical explanation on how wavelets are used in this

context is provided in appendix D. [37] also discuss this technique in detail.

Segmentation is the process of grouping or clustering similar pixels together, often to segregate objects within an image. Clustering is done based on colors and implemented using the K-Means algorithm. The ability for clusters to accurately summarize the data depends heavily on the number of clusters chosen. For example, if the true data have 3 clusters but K is chosen to be 2 then the final clusters would not be representative of the true structure of the data. The same is true if there are no true clusters in the data but K-means is still performed with K>1. To provide a more flexible analysis, I create an RGB histogram with 64 bins for each image and then choose K to be the number of peaks that surpass 500 pixels. Thus, K varies for each image as there is no reason to expect each image to have the same number of clusters. Next, I run K-Means on the pixels based on their colors using the chosen K and kept the 5 largest segments of each image to analyze. An example of this is shown in figure 3.1. The size (or number of pixels) and the average H, S, and V values for each segment define features f21-f43.

Figure 3.1: Segmentation Example

Original image After segmentation

K is selected to be 9 based on the number of peaks in a RGB color histogram. On the left is the original image. On the right is the image after segmentation.

### *3.4 Estimating Curb Appeal*

I select to use a gradient boosting regression model to predict curb appeal[2]. I use 1000 trees, apply 5-fold cross-validation and a shrinkage parameter of 0.01.

Figure 3.2: Predict Curb Appeal Out-of-Sample

Predicted rating of 3.7

Predicted rating of 7.4

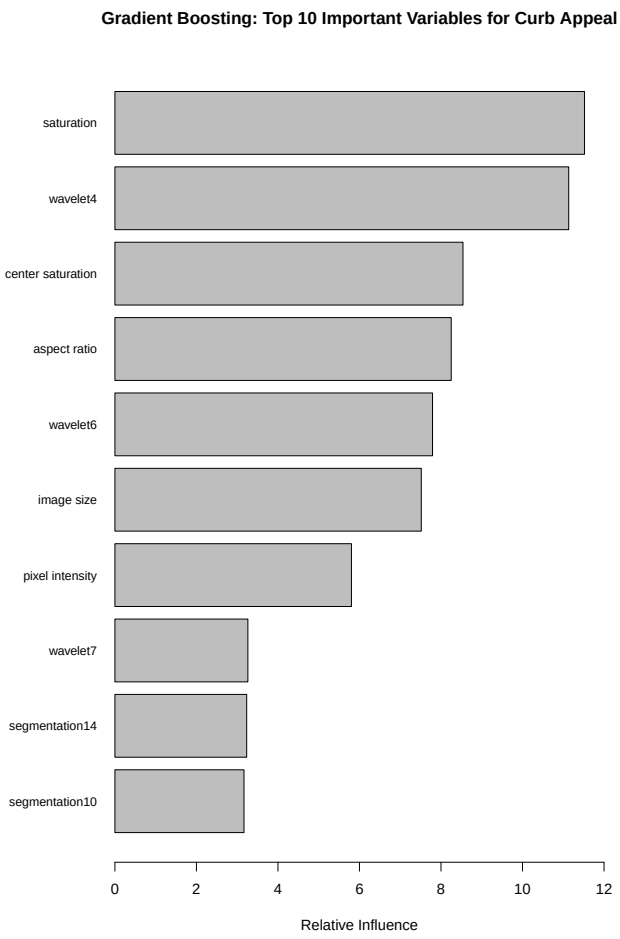
Predicted curb appeal ratings for two homes that are not in the training set. These results show that the model predicts out-of-sample reasonably well.

The top 10 most important predictors for curb appeal are shown in figure 3.3. Curb appeal reflect the quality of the photos with saturation, aspect ratio, and image size being some of the top predictors. One potential issue with this approach is that pictures of homes with low curb appeal can be manipulated (e.g. photoshopped) and incorrectly labeled with high curb appeal ratings by the gradient boosting model. However, if humans can detect the curb appeal of homes regardless of photo edits then the training set labeled via Mechanical

[2]Datta et al. use support vector machines to train the model. While support vector machines are strong classification models, they do not work well with high dimensional features. I also contrast the predicted values using a random forest model and obtain similar results.

Turk would control for this. Also, home sellers that upload poor quality photos would reveal attributes regarding the home such as poor maintenance, and curb appeal as these attributes are likely correlated.

Figure 3.3: Important Variables




A summary of the curb appeal rating distribution of the training set and the predicted set are in table 3.2. The predicted curb appeals are much more conservative than the training set since machine learning models often predict outliers poorly. Nevertheless, the model predicts well. Figure 3.2 shows the predicted curb appeal ratings for two homes that are not in the training set. I use predicted curb appeal ratings of the training images in the final

Table 3.2: Summary of Curb Appeal Ratings

| Images | Mean | Min | Max | N |
|---|---|---|---|---|
| Mechanical Turk | 6.379 (1.516) | 1.000 | 10.000 | 2,774 |
| Predicted data | 6.332 (0.475) | 3.662 | 7.730 | 23,733 |

Notes: Standard deviations in parenthesis

dataset to maintain consistency between the training and untrained set. In figure 3.4, I show a range of predicted curb appeal ratings. This spectrum of curb appeal ratings from the model suggest that the chosen image features explain curb appeal well and the predictions are not confounded by photo alterations.

Figure 3.4: Range of Predicted Curb Appeals

3.7 4.9 5.9 6.8 7.5

Shown above are a range of predicted curb appeal values

### *3.5 Conclusion*

Advertising and marketplaces have transitioned toward online platforms. The increase in online data fosters intensive use and analysis of data by researchers. For example, online marketplaces such as Craigslist and Amazon make available rich details on product attributes

in the form of descriptions, pictures, and videos to encourages consumers to make decisions based on information online. The shift towards online purchases and marketing also makes such data available to researchers. This data can be used to enhance policy analysis by reducing omitted variable bias and enhancing model prediction accuracies.

Accessibility of images via online sources has been the catalyst for the use of a variety of media data in economic analysis. This paper demonstrates a scalable way to incorporate unstructured data in the form of images into regressions. Tools such as Mechanical Turk make obtaining human training sets simple. This approach can be applied for future work for policy analysis to yield more predictive models and decrease omitted variable bias. Extracting qualitative features from images as done here can be modified to suit a broader genre of applications such as assessing the value of product images posted on eBay and Craigslist. Similar analyses can be done with product descriptions in those markets.

# BIBLIOGRAPHY

[1] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *The American Economic Review*, 93(1):113–132, 2003.

[2] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*, 2015.

[3] Patrick Bajari and C Lanier Benkard. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach, 2001.

[4] Patrick Bajari, Jane Cooley Fruehwirth, Kyoo Il Kim, and Christopher Timmins. A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution. *The American Economic Review*, 102(5):1898–1926, 2012.

[5] Patrick Bajari and Matthew E Kahn. Estimating hedonic models of consumer demand with an application to urban sprawl. *Hedonic methods in Housing markets*, pages 129–155, 2008.

[6] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Demand estimation with machine learning and model combination. Technical report, National Bureau of Economic Research, 2015.

[7] Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. Technical report, National Bureau of Economic Research, 2007.

[8] Patrick Bayer, Nathaniel Keohane, and Christopher Timmins. Migration and hedonic

valuation: The case of air quality. *Journal of Environmental Economics and Management*, 58(1):1–14, 2009.

[9] Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. 2009.

[10] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[11] Sandra E Black. Do better schools matter? parental valuation of elementary education. *Quarterly journal of economics*, pages 577–599, 1999.

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[14] Kenneth Y Chay and Michael Greenstone. Does air quality matter? evidence from the housing market. *Journal of Political Economy*, 113(2):376–424, 2005.

[15] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

[16] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301. Springer, 2006.

[17] Lucas W Davis. The effect of health risk on housing values: Evidence from a cancer cluster. *The American Economic Review*, 94(5):1693–1704, 2004.

[18] Lucas W Davis. The effect of power plants on local housing values and rents. *Review of Economics and Statistics*, 93(4):1391–1402, 2011.

[19] Dennis Epple. Hedonic prices and implicit markets: estimating demand and supply functions for differentiated products. *Journal of political economy*, 95(1):59–80, 1987.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[21] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[22] Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Measuring polarization in high-dimensional data: Method and application to congressional speech. 2015.

[23] Edward Glaeser and Joseph Gyourko. Zoning's steep price. *Regulation*, 25:24, 2002.

[24] Edward L Glaeser and Joseph Gyourko. The impact of zoning on housing affordability. Technical report, National Bureau of Economic Research, 2002.

[25] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. Big data and big cities: The promises and limitations of improved measures of urban life. Technical report, National Bureau of Economic Research, 2015.

[26] Edward L Glaeser and Bryce A Ward. The causes and consequences of land use regulation: Evidence from greater boston. *Journal of Urban Economics*, 65(3):265–278, 2009.

[27] Jenny Ho. Pictures as regressors: Estimating curb appeal. *Working Paper*, 2017.

[28] Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. Where not to eat? improving public policy by predicting hygiene inspections using online reviews. In *EMNLP*, pages 1443–1448. Citeseer, 2013.

[29] David S Lee and Thomas Lemieuxa. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

[30] Christopher G Leggett and Nancy E Bockstael. Evidence of the effects of water quality on residential land prices. *Journal of Environmental Economics and Management*, 39(2):121–144, 2000.

[31] Leigh Linden and Jonah E Rockoff. Estimates of the impact of crime risk on property values from megan's laws. *The American Economic Review*, 98(3):1103–1127, 2008.

[32] Lucija Muehlenbachs, Elisheba Spiller, and Christopher Timmins. The housing market impacts of shale gas development. *The American Economic Review*, 105(12):3633–3659, 2015.

[33] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] Nikhil Naik, Ramesh Raskar, and César A Hidalgo. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *The American Economic Review*, 106(5):128–132, 2016.

[35] Raymond B Palmquist. Property value models. *Handbook of environmental economics*, 2:763–819, 2005.

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

[37] P Scheunders, S Livens, G Van de Wouwer, P Vautrot, and D Van Dyck. Wavelet-based texture analysis. *International Journal on Computer Science and Information Management*, 1(2):22–34, 1998.

[38] V Kerry Smith and Carol CS Gilbert. The valuation of environmental risks using hedonic wage models. In *Horizontal Equity, Uncertainty, and Economic Well-Being*, pages 359–392. University of Chicago Press, 1985.

[39] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[40] Eugene Tuv, Alexander Borisov, George Runger, and Kari Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10(Jul):1341–1366, 2009.

[41] Hal R Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

[42] Jeffrey Zabel and Maurice Dalton. The impact of minimum lot size regulations on house prices in eastern massachusetts. *Regional Science and Urban Economics*, 41(6):571–583, 2011.

[43] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

## *Appendix*

### *A Machine Learning*

#### *Regression Trees*

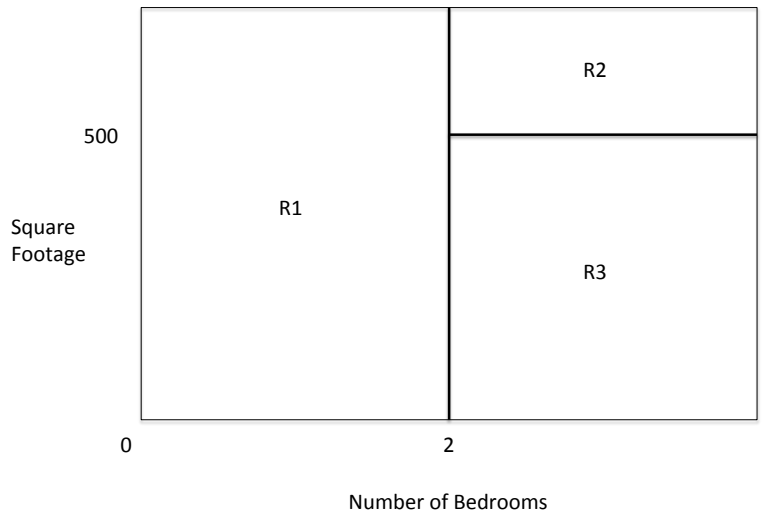
Regression trees split the variable input space into separate regions as a way to group observations that are similar together. They recursively divide the input space into a set of regions and then assign the average of a region's outcomes as the predicted value. An example of a simple tree with two splits is shown in figure A5. The first split variable is on the number of bedrooms at a threshold of 2. The second split is on square footage at a threshold of 500. In total, the tree splits the input space into 3 regions: one where the number of bedrooms is less than 2, one where the number of bedrooms is greater than two and square footage is less than 500, and one where the number of bedrooms is greater than two and square footage is greater than 500. Each time a split is made, the algorithm aims to minimize the RSS in the two regions it creates. This means that in the second split, splitting the input space at square footage of 500 minimizes the residual joint sum of squares in $R_2$ and $R_3$. When assigning variable importance, number of bedrooms gets more weight because it is selected as the first variable to split the input space on and therefore decreases RSS by more than if square footage is selected. A more complex tree using 1,000 observations from the data is in figure A6.

Assuming an outcome $y_i$ for observations $i = 1, ..., N$ with $p$ inputs, $x_i = (x_{i1}, x_{i2}, ....., x_{ip})$, a tree is formally defined as:

$$f(x) = \sum_{j=1}^{J} \gamma_j I(x \in R_j) \tag{1}$$

where $f(x)$ is the predicted outcome. The input space is divided into $J$ regions $R_j, j = 1, 2, ..., J$ with $I(\cdot)$ as an indicator function for whether $x$ is in a particular region $R_j$. $\gamma_j$ is a constant and chosen such that the RSS, $\sum(y_i - f(x_i))^2$, is minimized. The optimal $\gamma_j$ that minimizes this error is the average of the outcomes of all the observations within $R_j$. The tree model selects optimal cutoffs to split the input space. The variable and threshold are

Figure A5: Tree Split into Three Regions




Note: This is an example of a tree with 2 splits and 3 final regions, or terminal nodes. The first split is on number of bedrooms. The second split is on square footage but only for homes with more than 2 bedrooms.

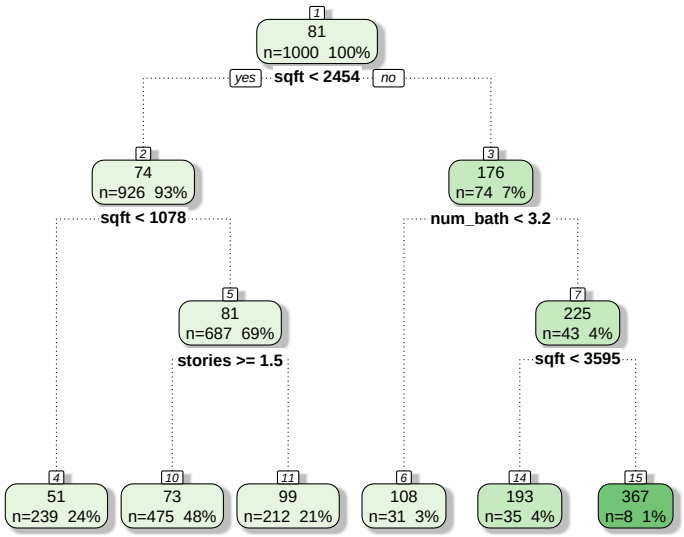
chosen such that the RSS in the two regions combined are minimized. Splits are done with a single variable at a chosen value threshold. For example, if variable $k$ is selected as the splitting variable at a threshold $s$, then a region would be split into two half-planes:

$$R_1(k, s) = \{x|x_k < s\} \text{ and } R_2(k, s) = \{x|x_k >= s\}. \tag{2}$$

Then splits are sequentially chosen on each of the separate regions and all the subsequent regions in the same way.[3] This means that at each split, the algorithm will search through all variables and thresholds that will minimize the RSS of two new regions. Multiple splits

[3]The algorithm follows a greedy approach in the sense that splits are done sequentially and once a split is made, there are no changes to previous splits. A split is made to minimize the sum of squares in two regions but does not guarantee that the RSS of all regions is minimized.

Figure A6: An example of a tree with 1,000 observations




on the same variable at different thresholds can be selected. The final regions are called leaves or terminal nodes. A chosen limit on the number of terminal nodes or minimum number of observations remaining in each region are usually set as stopping criteria, or tuning parameters. In general, the more terminal nodes and fewer observations left in each region, the lower the bias and the more accurate the fit on the training set.

Benefits of this approach are that data can be high-dimensional at no cost to the model and that nonlinear relationships and interactions across variables are automatically explored. It can easily handle multicollinearity since variables are selected based on their ability to predict the outcome. Thus, if two variable are highly correlated, the model will only select one of them for prediction, but including both does not lower the prediction accuracy. Incorporating fixed effects can easily be done by including additional dummy variables. They also have low bias and fit the in-sample observations well, but since they are very sensitive to the

training data, they tend to have high variance in predictions. The predictions across trees split on different data can vary widely.

Techniques such as bagging and bootstrapping of the data reduce this variance by using the average predictions from an ensemble of trees to reduce this variance. Bagging involves bootstrapping samples, forming a tree for each sample, and then averaging the predictions from a set of trees. Gradient boosting is an ensemble tree method that incorporates elements of bagging to reduce variance. I provide a summary of gradient boosting in the following section. Random forest is a similar method that predicts using the average of a set of trees. See appendix section A for explanation on random forest and appendix section A for more details on gradient boosting.

An example of a tree model for house price (in \$1,000) as a function of its attributes with 1,000 observations is shown in figure A6. The depth of the tree here is 3, the total number of nodes is 11, and the total number of terminal nodes is 6. The house attributes that split the data into two regions are listed with the numbers of observation in the regions as well as the proportion of the data. Notice that regions can be split on the same attribute multiple times. The algorithm splits until a specified number of observations are left in each terminal node. The terminal nodes in this tree define the final regions and the average house price for homes in that region is given in each terminal node.

*Gradient Boosting*

[21] introduces gradient boosting as an additive ensemble model where a weighted sum of base models are added together. Using tree models as base models, a new tree is added to a weighted sum of all previous trees at each iteration. I use a sum of squares loss function $L(y, F(x)) = -\frac{1}{2}(y - F(x))^2$, but a variety of different loss functions can be used.

The model is initialized to be the average of the outcomes, $\bar{y}$ for $N$ observations. At each subsequent iteration, the model parameters are chosen to correct the error from the previous

iteration. Suppose there are $M$ iterations, then the model is defined as:

$$F_M(x) = \sum_{m=0}^{M} f_m(x) \tag{3}$$

where

$$f_m(x) = -\rho_m h_m(x). \tag{4}$$

Equation (3) can be rewritten as

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{5}$$

where $h_m(x)$ is a regression tree model fit to the residuals, $g_m(x)$, from the previous iteration which is also equal to the gradient of the loss function. $\rho_m$ is the step size or weight on a model and chosen to minimize the loss at iteration $m$. In other words,

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h_m(x_i)) \tag{6}$$

and

$$g_m(x) = y - F_m(x) = \left[\frac{\partial L(y, F(x))}{\partial F(x)}\right]_{F=F_{m-1}} \tag{7}$$

This is similar to the gradient descent algorithm for a minimization problem with $-g_m$ as the negative gradient and $\rho_m$ as the "line search" along the gradient. This is why it is called a gradient (descent) boosting model. In practice, $h_m(x)$ is an approximation to the error, $g_m(x)$, and takes the form of a regression tree with $J$ cuts as defined in the section 1.3.2. The step by step algorithm is in table A3. For prediction models, gradient boosting can be computed with a shrinkage parameter, $\nu$, to prevent overfit. $\nu$ effectively shrinks the influence of each additional tree. Now, equation (5) is replaced with

$$F_m(x) = F_{m-1}(x) + \nu\rho_m h_m(x) \tag{8}$$

The gradient boosting model iteratively corrects the errors in the model and then applies a shrinkage parameter to the errors to prevent them from overfitting at the next iteration. This

prevents the errors in one iteration from having too much influence on the final predictions. In practice, the optimal model for out-of-bag sample predictions will depend on both the choice of $M$ and $\nu$. Greater values of $\nu$ require lower $M$ and vice versa. It is common to set $\nu = 0.1$. An additional way to mitigate overfit is to subsample the data at each iteration $m$. This reduces the variance of the predicted outcome.

Table A3: Gradient Boosting Algorithm

1. Initialize the model $F_0(x) = \underset{\gamma}{\operatorname{argmin}} L(y_i, \gamma)$
2. For m=1 to M do:
   (a) Compute
   $$g_m(x_i) = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x)}\right]_{F=F_{m-1}} \quad \text{for } i = 1, ..., N$$
   (b) Fit a regression tree, $h_m(x)$, to the errors, $g_m(x)$, with $J$ cuts
   (c) Compute
   $$\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h_m(x_i))$$
   (d) $F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$
3. The predicted outcome for observation x is $F_M(x)$

*Random Forest*

Random forest averages the predictions from a set of trees formed from bootstrapped samples to decrease the variance of the predictions. A tree is represented as $f(x)$ in equation 1 and has the same form as discussed earlier with the same parameters. A random forest with $B$ number of bootstrapped trees can be written as:

$$F(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x). \tag{9}$$

Random forest bootstraps the training data B number of times by randomly selecting $m$ variables from $p$ input variables each time before another split is made. Selecting only a subset of variables decorrelates the trees and prevents overfitting by decreasing the variance of the average, which is predicted value. With independently distributed variables, correlation

$\rho$, and variance $\sigma^2$, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{10}$$

As B increases, the second term goes to zero, leaving the first term. Thus, as the correlation, $\rho$, decreases, the entire term goes to zero. In practice, $m$ is usually set to $\frac{p}{3}$ for regression trees. Random forests can also reduce the variance of predictions by averaging bootstrapped samples of regression trees. [20] argue that random forest is computationally scalable and able to deal with irrelevant inputs.

Random forest bootstraps the training data B number of times and generates a tree for each iteration. B can be selected by plotting out-of-sample error rates at each iteration and stopping when this converges. Random forest makes a slight modification by restricting each split to a random set of $m$ variables from the $p$ input variables. This limitation prevents strong predictors from dominating all the trees. Common choices for $m$ are $\sqrt{p}$ or $\frac{p}{3}$ where $p$ is the number of variables. Selecting only a subset of variables for each split decorrelates the trees and prevents overfitting by decreasing the variance of the predicted values. This method is powerful in practice and is one of the most popular methods used for prediction. I offer the algorithm in table A4.

Table A4: Random Forest Algorithm

1. For b=1 to B:
   (a) Draw a bootstrapped sample of size N from the training sample.
   (b) Grow a regression tree $f_b$ using the bootstrapped sample. Stop when the minimum node size is obtained. At each terminal node,
      (i) Randomly select m variables from p total variables.
      (ii) Split on the best variable from the m variables selected
2. Output the ensemble of trees $\left\{f_b\right\}_1^B$.
3. The predicted outcome for an observation $x$ is $F(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$.

## *B Selection Bias*

Despite being chosen as top variables in the boosting model, gradient boosting assigns a lot of weight to the top 5 variables. Age at sale is five times more influential than the next three variables: square footage, income, and number of bathrooms. This can make it unclear whether or not the pretreatment variables are effectively controlling for both neighborhood characteristics, but more importantly, any selection bias that would affect the estimates.

One way to see how these variables affect average treatment effects is to run a model leaving these variables out. I report the results in table A5 using the gradient boosting model. The average treatment effects are significantly more negative than those in table 1.7; in some cases, almost double in magnitude. These estimates are as expected since neighborhoods with higher air pollution are also associated with areas with lower economics advantages which negatively affect house prices. This shows that these variables significantly affects the treatment effect estimates despite having much lower variable influence than the top variables in the gradient boosting model. The significant difference in average treatment effects also shows that there are not a lot of other variables that can serve as surrogates for demographics.

## *C Images as Data*

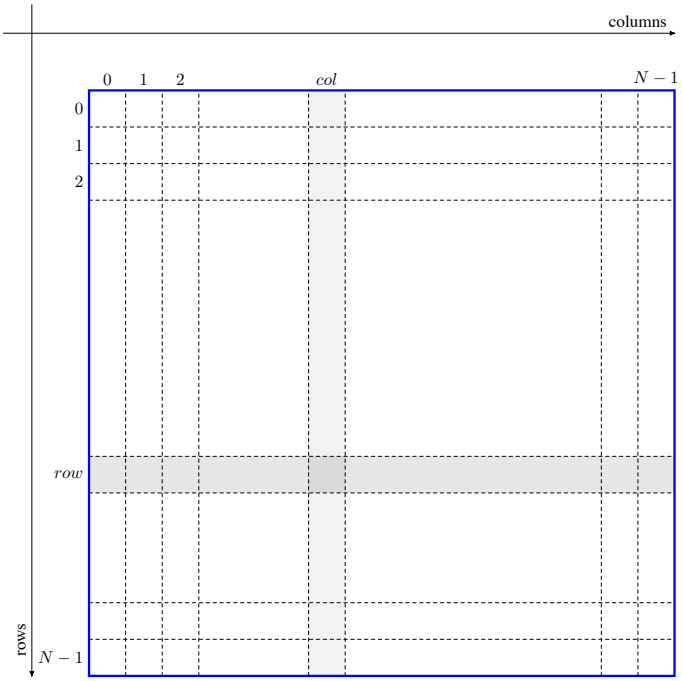
An image is an array of numbers that reflect the colors of the pixels. For a grayscale image with $X \times Y$ pixels, the corresponding array is $X \times Y$ where the value in each cell ranges from 0 - 255, or black to white, to indicate the color of the pixel. An example can be seen in figure A7. For a color image, the array is $X \times Y \times 3$ to reflect different values for each red, green, and blue (RGB) spectrum. Each pixel has coordinates $(x, y)$ and an $I(x, y)$ function that maps the pixel coordinates to the color value of that pixel. For a grayscale image, $I(\cdot)$ returns one value while an RGB image has 3 values for each pixel - one for each color component: $I_R(x, y)$, $I_G(x, y)$, and $I_B(x, y)$. I will be using a variety of cylindrical coordinate representations such as hue-saturation-value (HSV).

Table A5: Boosting Model Average Treatment Effects by $PM_{10}$ Levels Dropping Pretreatment Variables

| | $PM_{10}$ $\mu g/m^3$ levels | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variables | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Structured | 0.020 | -0.011 | -0.040 | -0.070 | -0.101 | -0.204 | -0.195 | -0.255 | -0.262 |
| | (0.013) | (0.012) | (0.012) | (0.013) | (0.013) | (0.016) | (0.018) | (0.023) | (0.035) |
| +pic | 0.022 | -0.013 | -0.037 | -0.063 | -0.092 | -0.203 | -0.191 | -0.249 | -0.263 |
| | (0.012) | (0.012) | (0.013) | (0.013) | (0.013) | (0.016) | (0.018) | (0.023) | (0.035) |
| +words | 0.018 | -0.019 | -0.039 | -0.060 | -0.092 | -0.200 | -0.179 | -0.241 | -0.250 |
| | (0.012) | (0.012) | (0.012) | (0.013) | (0.013) | (0.016) | (0.018) | (0.022) | (0.034) |
| N | 1906 | 2575 | 2712 | 2398 | 2156 | 1342 | 1021 | 617 | 317 |

Note: Standard errors in parenthesis.

Figure A7: Image as a Matrix




## D *Wavelets for Texture Analysis in Images*

For a signal $x(u)$, where $u$ is the pixel's x-coordinate, the Haar wavelet is defined by

$$\psi(u) = \begin{cases} -\frac{1}{2} & -1 \leq u < 0 \\ \frac{1}{2} & 0 \leq u < 1 \\ 0 & otherwise \end{cases}$$

The color intensity $x(u)$ across pixels is filtered by the wavelet since for all values less than $-1$ and 1, all the values are set to zero when the signal is multiplied, or convolved, by the wavelet. The usefulness of the wavelet for texture analysis is seem in the integral of the

convolved signal $\psi(u)x(u)$. The integral is defined as

$$\begin{aligned}\mathcal{W} &= \int_{-1}^{1} \psi(u)x(u)du \\ &= -\int_{-1}^{0} x(u)du + \int_{0}^{1} x(u)du\end{aligned}$$

$W$ is called a wavelet coefficient. Given this equation, it is clear that the wavelet coefficient is the difference between the averages of the two intervals. If $x(u)$ for $u < 0$ and $x(u)$ for $u > 0$ differ greatly, then the wavelet coefficient will be very high. In contrast, if neighboring pixels have the similar colors and then, the two sets of signals will be approximately the same and $W$ would be close to zero.