Accessibility or Amenities? Estimating the Value of Light Rail Transit

Mark Ponder
University of Minnesota

Veronica Postal
University of Minnesota

July 28, 2020

Abstract

This paper examines consumer marginal willingness-to-pay or the introduction of light rail transit in Minneapolis. We estimate the resulting change in local property prices to assess what share is attributable to the direct effect of improved access to public transit and what share is attributable to the increase in local amenities. After assembling a rich spatial dataset encompassing every residential property in Minneapolis and hundreds of thousands of businesses and neighborhood amenities, we use machine learning techniques to estimate a hedonic pricing surface. We extend the method of Boosted Smooth Trees introduced by Fonseca et al. (2014) to a high-dimensional dataset and to incorporate instrumental variables, allowing us to control for endogeneity in amenity changes. Our results indicate that the price of properties located within a half mile of a light rail station increased by around 11.3%. The direct impact of access to the light rail itself is estimated to increase local housing prices by 5.5%, while the estimated spillover due to changes in amenities is quantifiable at 5.8%.

1 Introduction

The large capital investment required to construct new mass transit projects, coupled with the long time horizon associated with this type of investment, means that estimating the demand for public transit has been a topic of interest for modern economics at least since McFadden (1974). The most common approach to measuring the demand for public transportation follows the hedonic pricing model of Rosen (1974), where changes in house prices after the introduction of a transit system are used to infer the marginal willingness-to-pay (MWTP) of local residents for access to the new system. This paper is closely related to the extensive branch of this literature that employs hedonic models to estimate the impact of opening new mass transit projects on the urban environment, in particular their effect on local housing prices.[1] Our specific application focuses on the introduction of the METRO Blue Line in Minneapolis, Minnesota.

Most event studies in the literature commonly employ a cross-sectional or difference-in-differences approach to estimate a single treatment effect arising from the introduction of a new mass transit project, usually finding an increase in house prices and rents for properties located close to new transit stations.[2] However, these techniques are unable to identify which effects are directly attributable to improved access to public transportation, and which effects arise indirectly from the transit system. Transit systems connect distant parts of urban areas, increasing the catchment area for local retail shops and other local businesses. This increase in demand can encourage

[1] A number of studies are available on the subject, spanning dozens of cities. These include Atlanta (Cervero (1994), Ihlanfeldt (2003), Immergluck (2009)), Buffalo (Hess and Almeida (2007)), Charlotte (Billings (2011)), Chicago (McDonald and Osuji (1995), McMillen and McDonald (2004)), Dallas (Clower et al. (2002), Nelson et al. (2015)), Hampton Roads (Wagner et al. (2017)), Huston (Pan (2013)), Miami (Gatzlaff and Smith (1993)), Los Angeles (Cervero and Duncan (2002)), Philadelphia (Kilpatrick et al. (2007)), Phoenix (Seo et al. (2014)), Portland (Dueker and Bianco (1999)), Sacramento (Rewers (2010)), Santa Clara County (Weinberger (2001)), San Diego (Duncan (2008)), Washington County (Knaap et al. (2001)), Washington DC (Damm et al. (1980), Grass (1992), Cervero (1994)). Outside the United States, there are studies focusing on Amsterdam (Debrezion et al. (2011)), Bogotá (Tsivanidis (2018)), Haifa (Portnov et al. (2009)), London (Gibbons and Machin (2005)), Manchester (Forrest et al. (1996)), Ottawa (Hewitt and Hewitt (2012)), Seoul (Bae et al. (2003)), Shangai (Pan and Zhang (2008)), Toronto (Dewees (1976), Bajic (1983)), among others.

[2] See Debrezion et al. (2007), Hess and Almeida (2007), and Mohammad et al. (2013) for more general surveys of the findings of this literature.

additional businesses, such as restaurants and entrainment, to enter near transit stations, creating a positive externality on nearby households, who will now have access to a higher number and larger variety of local businesses.

A few studies (Bowes and Ihlanfeldt, 2001; Zheng et al., 2016) have attempted to disentangle the direct and indirect effects of mass transit investments, but they typically assume that the level of amenities is independent of unobserved characteristics that impact housing prices, conditional on observed characteristics. However, this assumption is complicated by the existence of preference externalities. If residents tend to cluster based on shared unobserved preferences, each neighborhood will see a different composition of establishments entering in response to the introduction of a new transit system. The mix of new establishments will naturally be correlated with unobserved preferences and will therefore confound estimation. This paper decomposes the benefits of introducing a light rail system into the direct and indirect effects, while controlling for endogeneity. Using recent techniques from the machine learning literature, we then estimate heterogeneous effects for different types of neighborhoods, based on a large selection of covariates.

Formerly known as Hiawatha Line, the construction of the METRO Blue Line was first proposed by the Minnesota Department of Transportation in 1985, but it was not until January 2001 that construction began. The Blue Line started operations between a subset of 12 stations in June 2004, and full service started in November 2004. It connects downtown Minneapolis with its southern suburbs, counting 18 stations and spanning a total of 12 miles. Two studies have examined the impact of the Blue Line in Minneapolis on housing properties.[3] Goetz et al. (2010) published a comprehensive study focusing on the Blue Line's impact on property prices, housing investment and land use. They find modest price premiums (in the order of 3.8-4.0%) for single family homes located within a half mile of a station in South Minneapolis, with the net effect

[3]Two additional studies examine on the effect of the Blue Line on other outcomes: Ko and Cao (2013) focus on industrial and commercial properties values, and Hurst and West (2014) investigate the effect of the Blue Line on land-use changes.

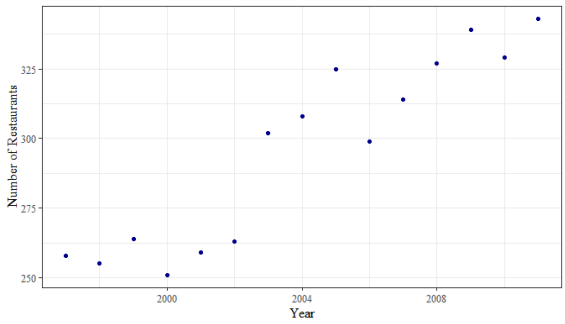
varying non-linearly as a function of distance. Pilgram and West (2018) use repeat sales to establish the effect of opening the Blue Line within a difference-in-difference setup. They find that single-family homes located within half a mile from a station in South Minneapolis experience a positive price premium (2.5-4%), but that the premium is diminishing over time, potentially as a result of the Great Recession.

However, even before construction was completed, neighborhoods surrounding Blue Line stations started seeing an uptick in the number of business being opened. For example, Figure 1 illustrates the number of restaurants, art, and entertainment establishments within one mile of a Blue Line station between 1997 and 2011. A significant increase in the number of these establishments is evident starting around 2003. This is in line with the findings of Berry and Waldfogel (2010) who find that restaurants and other businesses with high variable cost increase in number and diversity as market size increase. This large increase in local amenities likely had a sustained impact on house prices and consumer welfare in addition to the direct impact of improved access to public transit through the Blue Line. Goetz et al. (2010) and Pilgram and West (2018) are unable to establish whether price premiums occurring after the introduction of the Blue Line are to be attributed to the direct impact of the light rail on neighborhood accessibility by public transit, or whether they are due to a spillover effect on the entry of amenities. Moreover, the benefits of a light rail system are largely heterogeneous, with certain locations benefiting more than others, certain types of residents benefiting more than others, and houses closer to the station benefiting more than those further away. Employing a more flexible specification will allow our paper to identify which groups benefit the most from the introduction of a new transit system.

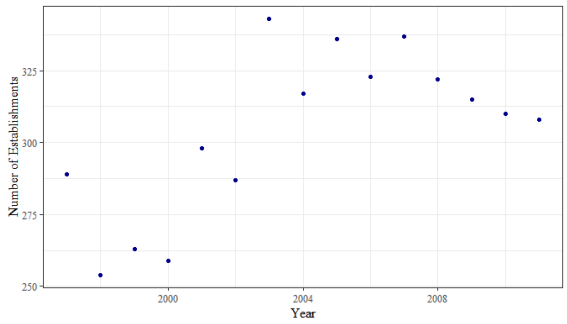
The closest paper to ours in terms of methodology is Ho (2016), which uses gradient boosting techniques to estimate the effects of air pollution on house prices. Our paper however employs a different empirical strategy to control for endogeneity. Ho (2016) follows Varian (2014) to identify which properties are unaffected by air pollution from a first-stage estimation and uses these observations as a control group. A second-stage estimation is then performed based on these ob-

Figure 1: Establishments within 1 mile of a Blue Line Station

(a) Restaurants



(b) Entertainment



servations and property prices are predicted for the treatment group. The difference between the predicted and realized prices is the estimated effect of air pollution, which is then regressed onto observed covariates to model the effect heterogeneity. We use a similar method to estimate the direct effect of the Blue Line introduction, where a predictive model does not need to control for endogeneity. Our approach to estimate the effect of new amenities follows instead Athey et al. (2019), who apply gradient boosting directly to local instrumental variable moment conditions. This approach side-steps the need to select a control group using first-stage estimates that are contaminated with endogeneity and allows us to estimate this spillover effect.

The results of our estimation routine using smooth trees show that the price of properties located within a half mile of a light rail station increased by around 11.3%. This total effect is in line with that estimated via DiD (10.4% in our preferred specification), and somewhat higher than the overall effect estimated by Goetz et al. (2010) and Pilgram and West (2018). This might be due at least in part to a different geographic focus, since we examined all neighborhoods along the path of the Blue Line, while other studies on the impact of the Blue Line focus exclusively on neighborhoods in Southern Minneapolis. The smooth trees estimation procedure also allows us to directly calculate the estimated spillover due to changes in amenities, quantifiable at 5.8%, while the direct impact of access to the light rail itself is estimated to increase local housing prices

by 5.5%. Thus over 51% of the overall appreciation in housing prices after the introduction of the Blue Line is attributable to an increase in the number of new amenities around light rail stations. The only comparable result in the literature is from Zheng et al. (2016) who found that the increase in neighborhood restaurant activities due to the introduction of a new subway station in Beijing captures 20 to 40% of the overall appreciation in home values. The discrepancy might be explained by the fact that we control for a far greater variety of businesses than Zheng et al. (2016), who focus exclusively on restaurants.

The rest of the paper will proceed as follows: Section 2 introduces the theoretical framework for our estimation, while Section 3 discusses the different possible approaches to estimating treatment effects, such as difference-in-differences and machine learning methods. Section 4 summarizes our data sources for housing values, neighborhood amenities and demographics, while Section 5 presents our results under the different estimated approaches, and Section 6 concludes.

2 Theoretical Framework

The starting point for our analysis is the hedonic model introduced by Rosen (1974) who proposes generating a pricing surface based on a vector of characteristics of the good of interest, in our application housing. Under certain regularity assumptions, the derivative of the pricing surface with respect to a given set of characteristics represents the consumers marginal willingness-to-pay for said characteristic.

Consider z, a vector describing the characteristics of a good (in our case, residential housing). The good has a market price which arises as an equilibrium object from the endogenous sorting of buyers and sellers, where the buyers' indifference curve and sellers' offer curve are tangent, conditional on the housing characteristics. Housing prices can thus be written in terms of the vec-

tor of housing characteristics z, $p(z)$. Let x represent the consumption bundle of all other goods. Then a consumer with utility function U solves the following utility maximization problem:

$$\max_{\{z_j\}} U(x, z_1, z_2, ..., z_J) \quad \text{subject to} \quad y = x + p(z)$$

where income y is measured in units of x. For each housing characteristic j, the consumer's first-order conditions are given by:

$$\frac{\partial U(y - p(z), z_1, z_2, ..., z_J)}{\partial z_i} = -U_x \frac{\partial p(z)}{\partial z_i} + U_{z_i} = 0$$

So that the consumer marginal willingness-to-pay for characteristic z_j can be written as:

$$\frac{\partial p(z)}{\partial z_i} = \frac{U_{z_i}}{U_x}$$

Several complications arise when estimating $p(z)$. First, the pricing surface is an equilibrium object and therefore can change over time, so it will not necessarily be the case that $p_{t-1}(z) = p_t(z) = \bar{p}(z)$. When using a before and after approach (such as difference-in-differences), the estimated effect is a combination of the marginal effect and the equilibrium response. This complicates the interpretation of the parameter estimates.[4] Dealing with a shifting pricing surface is beyond the scope of this paper, so we assume that $p_t(z) = \bar{p}(z)$. For a similar reason, Pakes (2003) points out that this specification is not appropriate in the presence of markups. In such a case, the hedonic pricing surface depends on both strategic incentives and the utility maximization of consumers. We therefore cannot interpret $\frac{\partial p(z)}{\partial z_i}$ as the true marginal willingness-to-pay, but will have to settle for a bound instead.

[4] See Taylor (2003) and Palmquist (2005) for a review of the empirical literature and various difficulties that arise when using a hedonic approach.

If, following the literature, we assume that houses are provided in a perfectly competitive environment so that there is an absence of strategic effects and our estimates could be interpreted as the true MWTP. Under this assumption, Bajari and Benkard (2005) propose a method to recover consumer preferences that only depends on the distribution of prices conditional on observed characteristics. This allows them to recover consumer preferences even when products are discrete and some characteristics are unobserved. This method however does not work well in our setting for several reasons. First, it places a substantial burden on the data and does not scale well to higher dimensions due to the curse-of-dimensionality when non-parametrically estimating density functions. Since one of the contributions of this paper is incorporating a large number of covariates in our estimation routine, this is a significant drawback. In general, hedonic pricing models are not reliable in the presence of unobserved product characteristics that determine the pricing surface. We assume that the issues of unobserved product characteristics is less impactful to our results than in other environments since we rely on a large number (almost two hundred) observed product characteristics in our estimation routine. Moreover, since each housing property can be considered as a different product, with a different bundle of housing characteristics, and there are about 38,930 observed transactions in our data it can be argued that the housing market approximates a continuum of products rather than a market for discrete products.

Hedonic pricing surfaces are commonly estimated in the literature using linear regression or log-linear regression. Using a linear specification forces the marginal willingness-to-pay be constant across households, an assumption that is difficult to maintain in practice. The log-linear specification allows the willingness-to-pay to vary across households but in a strict way. For instance,the log-linear specification has derivatives equal to:

$$\frac{\partial p}{\partial z_i} = \beta_{z_i} p$$

The MWTP is thus proportional to the housing price p and does not depend on the value of any other housing characteristic. This can be limiting in certain settings: consider the case of two

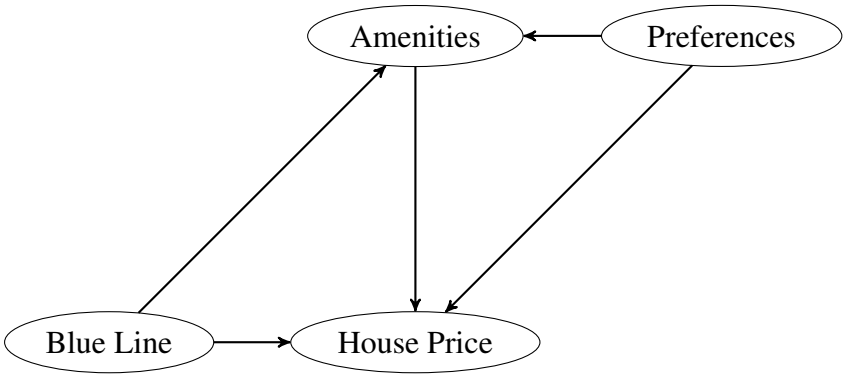
households that are the same distance from a light rail stop. If one of the houses is also near a bus stop, it might have a lower MWTP for being near the light rail stop as it can more easily substitute to a different type of public transportation. The log-linear specification does not allow for this type of interaction. We can ameliorate this issue by including high-order polynomial terms but this unfortunately results in an explosion of (often highly correlated) regressors. The is especially true when we have over-parameterized our regression to avoid omitted-variable bias, as is common in the literature. When the model is over-parameterized, researchers often resort to using some form of regularization, such as a ridge regression, LASSO regression, or an Elastic Net.

Ho (2016) proposes an innovative approach using gradient boosting with decision trees as weak learners to approximate the hedonic pricing surface. She compares this approach to using a Post-LASSO and finds that the Post-LASSO regression returns an overly sparse model that performs less reliably out of sample. Gradient boosting is better able to adapt to the local behavior of the pricing function, allowing the MWTP to vary based on the vector of housing characteristics. However, decision trees are locally constant, meaning that a hedonic pricing surface estimated in levels will have a zero derivative almost everywhere so that MWTP cannot be easily calculated. Ho (2016) deals with this by calculating the difference in predicted and actual sales price for discrete values of air pollution. The marginal willingness-to-pay can then be approximated by a finite difference across levels. However, for levels of pollution between these cutoffs the marginal willingness-to-pay is zero. Our approach to estimating the MWTP is inspired by Ho (2016) use of gradient boosting, and is outlined in more detail in the next section.

3 Estimating Treatment Effects

Early studies on the impact of mass transit projects focused on the estimation of hedonic pricing surface using a panel of housing sales. Typically, the (log) price is regressed on a set of covariates

Figure 2: Blue Line DAG



which include the distance to the nearest transit stop, d. The MWTP in this setting is given by:

$$\frac{\partial p_{it}}{\partial d_{it}} = \beta_d p_{it}$$

Because house prices are observed, consistent estimation of the MWTP is equivalent to consistently estimating β_d. This approach implicitly assumes that the covariates are orthogonal to the error term, but this is unlikely to hold when amenities are included as a determinant of house prices.

To see why, consider the causal relationship depicted in Figure 2. The introduction of the Blue Line has a direct impact on house prices, which is the MWTP for access to public transportation. However, it also causes an increase in local amenities, providing an indirect channel through which it again impacts home values. This channel is confounded by the presence of unobserved preferences which impact both the level of amenities available in a given neighborhood as well as house prices. For instance, more expensive restaurants might locate in wealthier neighborhoods or bars and more movie theaters might open up in neighborhoods that are relatively younger. It is theoretically possible to control for this omitted variable bias using standard regression methods by including additional demographic covariates, but this runs up against the bias-variance trade-off; researchers select a parsimonious specification to generate more precise estimates but

in doing so increase the chances of excluding relevant regressors. Our approach relies instead on machine learning methods to incorporate more covariates and uses instrumental variables to explicitly control for the endogeneity introduced by unobserved preferences.

3.1 Cross Sectional Methods and LASSO

The simplest approach to estimate the impact of the introduction of light rail transit is to use a time-varying cross sectional regression of home sales and estimate the coefficient with respect to distance to public transit.[5]This approach is valid under the assumption that the covariates included control for all channels through which unobserved preferences impact housing prices. Therefore, a simple regression of the distance to public transit, amenities, and exogenous covariates would provide valid estimates for each channel. Then the impact of public transit on amenities could be estimated to generate the desired decomposition. This method is easy to implement and the assumptions for valid causal identification are apparent. Further, it is easy to extend this approach to allow for nonlinear effects and interactions across covariates, allowing the research to specify a model with rich heterogeneous effects. Finally, if there are still concerns about endogeneity, implementing a cross-sectional approach with instruments is straightforward.

Of course, controlling for the impact of unobserved preferences on house prices requires knowing which exogenous covariates to condition on, which the researcher will not know a priori. The general strategy is then to include a large number of demographic and individual characteristics to avoid any omitted variable bias. However, the inclusion of irrelevant regressors or collinear regressors leads to higher variance estimates. Additionally, adding interactions and higher-order terms can quickly lead to a situation where $K \gg N$. For instance, the number of terms in a

[5]This approach is perhaps the most popular in the literature. See Bajic (1983), Gatzlaff and Smith (1993), Forrest et al. (1996) Bowes and Ihlanfeldt (2001), Cervero and Duncan (2002), Bae et al. (2003), Hess and Almeida (2007), Duncan (2008), Immergluck (2009), Portnov et al. (2009), Rewers (2010), Weinberger (2001), Debrezion et al. (2011) among others.

fully saturated model grows exponentially in K and can therefore dominate N even for a modest number of covariates. To avoid this, the researcher needs to determine which terms to include a priori, without guidance from the data. Machine learning techniques such as LASSO are effective at generating a parsimonious specification but tend to lead to overly sparse models. Additionally, they have a harder time to adapt to the nature of the data generating process.

3.2 Difference-in-Differences

Recognizing the limitations of a cross-sectional approach, recent papers have instead employed a difference-in-differences empirical strategy.[6]This strategy entails defining treatment and control groups by concentric circles around each transit station, treating the inner circle as the treatment group and the outer circle as the control group. The causal impact of the transit system may be recovered by looking at the difference in house prices between treatment and control group before and after the introduction of the light rail system, assuming that this difference would be constant absent any treatment. This strategy has several strengths. It is easy to implement and gives valid causal estimates if the underlying assumptions hold. The average treatment effect may be consistently estimated with few assumptions about functional form and the researcher is not required to control for all determinants of house prices, since it mitigates endogeneity concerns arising from omitted variable bias in traditional cross sectional hedonic models. Additionally, this method may be extended to allow for the estimation of continuous pricing surfaces, such as in Diamond and McQuade (2019).

Typically, these papers assign houses that are within a certain radius (usually 1 km or 0.5 miles) of a new station to a treatment group and use houses located further from the station as a control group. The pre-treatment period can be defined in several ways: before the system is announced,

[6]Among others, these include Baum-Snow and Kahn (2000), Gibbons and Machin (2005), Goetz et al. (2010), Billings (2011), Wagner et al. (2017), Pilgram and West (2018).

before construction beings, or before the system opens. Based on these definitions, a simple difference-in-differences estimator is implemented to produce an estimate of the total effect of public transportation on housing prices.

This approach has however several limitations. First, the definition of control and treatment group is not data driven and any contamination between the two can lead to inconsistent estimates. A similar problem exists in the definition of pre- and post-treatment periods. To ameliorate these issues, researchers typically test several different specifications to see how sensitive the results are to the definition of each group. The other drawback of this type of analysis is that the estimated average treatment effect is a combination of the direct effect from access to public transit and the indirect effect of amenity changes. To decompose these effects, we need a consistent estimate of the impact of transit on amenities and the impact of amenities on housing prices. Estimating the impact of the Blue Line on amenities is straightforward, but the existence of preference externalities and other confounding factors yields the estimates of amenities on house prices inconsistent. Because we consider a wide selection of amenities, a simple before and after approach will not identify each individual effects. Our proposed solution is to explicitly model all channels that affect housing prices and find relevant instruments to obtain causal identification.

3.3 Machine Learning

Varian (2014) proposes a method for estimating treatment effects given a well defined treatment and control group, and a predictive model. Assume the relationship of interest is given by:

$$p_{it}^{c} = f(X_{it}^{c}) + \varepsilon_{it}^{c}$$

for the control group, while for the treatment group we have:

$$p_{it}^t = f(X_{it}^t) + g(X_{it}^t) + \varepsilon_{it}^t$$

where p_{it} is the housing price (possibly in logs) and X_{it} is a vector of housing and neighborhood characteristics that determine house prices. Here, ε_{it} is a mean-zero shock that is potentially correlated with elements of X_{it}. The function $g(\cdot)$ captures the direct effect of public transportation on housing prices, and can potentially depend on a subset of the covariates X_{it}. The total effect is then given by:

$$TE = f(X_{it}^t) - f(X_{it}^c) + g(X_{it}^t)$$

If we train the data on the control group, we can approximate the conditional mean function as:

$$\hat{E}[p_{it}|X_{it}] \approx f(X_{it}) + E[\varepsilon_{it}|X_{it}]$$

Under the assumption that $\varepsilon_{it}^c|X_{it}^c \sim \varepsilon_{it}^t|X_{it}^t$, we have:

$$E[p_{it}^t - \hat{E}[p_{it}|X_{it}^t]|X_{it}^t] = g(X_{it})$$

allowing us to generate an estimate of the direct treatment effect.[7] Following Bajari and Benkard (2005), it is then possible to take the residual $r_{it} = p_{it}^t - \hat{E}[p_{it}|X_{it}^t]$ and regress it on the covariates X_{it} to uncover heterogeneous treatment effect resulting from the treatment. Unfortunately, following this approach does not allow us to recover the indirect treatment effect as well. The indirect effect is given by:

[7]This strategy is similar to the synthetic control method ((Abadie and Gardeazabal, 2003; Abadie et al., 2010)), which uses a weighted combination of observations in the control group in order to approximate the desired attributes in the treatment group in order to estimate a pricing function analogous to $f(X_{it})$.

$$IE = f(X_{it}^t) - f(X_{it}^c)$$

but we can only estimate the term:

$$Biased\ IE = (f(X_{it}^t) + E[\varepsilon_{it}|X_{it}^t]) - (f(X_{it}^c) + E[\varepsilon_{it}|X_{it}^c])$$

However, assuming the group definitions are valid, we can combine the estimates from a difference-in-differences approach and the matching approach to decompose the total treatment effect into an average direct effect and an average indirect effect. Given a well defined pre-treatment and post-treatment period, this method also provides a check on the definition of the control group. Before treatment occurs, it must be that:

$$E[p_{it}^c|X_{it}^c] = E[p_{it}^t|X_{it}^t]$$

so the residual from predicting outcomes in the pre-treatment treatment group using the control group should have mean zero, but will in general not be mean zero in the post-treatment period. If estimates using the control group are not mean zero, then the control group is not sufficiently similar to the treatment group to provide reliable estimates. Of course, this does not necessarily guarantee that the control group is valid, especially if the control group is contaminated by the treatment. In this case, we would expect the control group to do a decent job in predicting the treatment group because the control group should have been included in the treatment group to begin with.

3.4 Regression Trees with Gradient Boosting

To generate a predictive model of housing prices, we rely on recent advances in machine learning. Specifically, we use the gradient boosting method proposed in Friedman (2001). This approach builds up an estimate of $F(X_{it}) = f(X_{it}) + E[\varepsilon_{it}|X_{it}]$ by using functional gradient descent to iteratively improve the performance of regression function. The key idea is to take

$$\hat{F} = \text{argmin}_F E_{p,x}[L(p, F(X))]$$

where F is the function of interest, p is the dependent variable, and L is a loss function. Solving for F directly is infeasible, but we can use gradient descent to update an approximation in step m as:

$$F_m(X) = F_{m-1}(X) - \gamma_m \bigtriangledown_{F_{m-1}} L(p, F_{m-1}(X))$$

To reduce variance, we approximate the function $r = -\bigtriangledown_{F_{m-1}} L(p, F_{m-1}(X))$ with a weak learner h_m. The weak learners are chosen so that they have high bias and low variance. This means that an individual h_m does a poor job approximating a given function, but an ensemble of weak learners can provide an arbitrarily close approximation. Decision trees of this type are commonly used because they are flexible and can adapt well to the local structure of the function. The gradient r is then fit with the function h_m, and the estimator is updated according to:

$$F_m(X) = F_{m-1}(X) + \gamma_m \hat{h}_m(X)$$

The step size γ_m is estimated by regressing $\hat{h}_m(X)$ on r. It is common to take the loss function to be the quadratic loss $L(p_i, F(X_i)) = (p_i - F(X_i))^2$ and h_m to be decision trees with depths

ranging from 2 to 6 (Hastie et al., 2009).

Gradient boosting is easy to carry out in most modern statistical packages. Its ease of implementation makes it popular in the machine learning literature because it allows the researcher to select a parsimonious set of regressors whose selection is data driven. This method not only preforms better on the bias/variance trade-off than linear regression and LASSO, but it also yields better predictions than cross-sectional regression and LASSO approaches. Moreover, regression with gradient boosting naturally lends itself to estimating models with heterogeneous and nonlinear effects. On the other hand, this method is very computationally demanding, especially because cross-validation is necessary to choose the model's meta-parameters (such as the depth of the regression tree). Decision trees do not provide easily interpretable parameter coefficients and do not result in a smooth pricing surface. Furthermore, this method does not easily extend to estimation with instruments.[8]

As mentioned above, a viable approach to estimating the average indirect treatment effect involves taking the difference between our DID estimate and the direct effect estimated using gradient boosting. However, if we want to measure how the indirect treatment may vary heterogeneously, it is necessary to consistently estimate the impact of amenities on house prices. We next propose a method that allows the estimation of both direct and indirect treatment effects without needing to rely on comparing estimates from two different estimation routines. Additionally, this method yields estimates that have a derivative that is not zero almost everywhere so that the MWTP can be better approximated.

[8]For a recent approach at extending regression forests to the case of instrumental variables, see Athey et al. (2019).

3.5 Boosted Smooth Trees

Decision trees use local averaging, leading to function approximations that are step functions. As such, the approximation's derivative is zero almost everywhere. Because MWTP is based on the derivative of the hedonic pricing function, we prefer and approximation that is smooth. Following Fonseca et al. (2018), we can rewrite the decision tree weak learner as a linear regression on an indicator basis functions. Let J be the set of parent nodes and T be the set of terminal nodes. Then the decision tree can be written as:s

$$h_m(x_i) = \sum_{k \in T} \beta_k B_{J_k}(x_i; \theta_k)$$

where

$$B_{J_k}(x_i; \theta_k) = \prod_{j \in J} I(x_{s_j}; c_j)^{\frac{n_{kj}(1+n_{kj})}{2}} \left(1 - I(x_{s_j}; c_j)\right)^{(1-n_{kj})(1+n_{kj})}$$

and

$$I(x_{s_j}; c_j) = \begin{cases} 1 & \text{if } x_{s_j} \leq c_j \\ 0 & \text{otherwise} \end{cases}$$

and

$$n_{kj} = \begin{cases} -1 & \text{if the path of leaf k does not include the parent node j} \\ 0 & \text{if the path of leaf k includes the right-hand child of parent node j} \\ 1 & \text{if the path of leaf k includes the left-hand child of parent node j} \end{cases}$$

Note that $\sum_{k \in T} B_{J_k}(x_i; \theta_k) = 1$ and each observation x_i is mapped uniquely to some region of space. Fonseca et al. (2018) propose replacing the indicator $I(x_{s_j}; c_j)$ with a sigmoid function:

$$L(x_{s_j,i};\gamma_j,c_j) = \frac{1}{1+e^{-\gamma_j(x_{s_j,i}-c_j)}}$$

so that every point has a positive probability of being assigned to any terminal leaf. As the term γ_j increases, the model converges to a standard decision tree. Moderate values of γ_j smooth the estimates and the authors show that this allows for better estimation of the derivatives.

Unfortunately, this specification is far more computationally demanding than using a regression tree. The main issue is that the gradient boosting algorithm does not require us to actually construct the matrix $\{B_{J_k}(x_i;\theta_k)\}_k$ and regress it on r for each potential split. However, this step is unavoidable when using $L(\cdot)$ because testing a new split requires recalculating the choice probabilities for every leaf. This makes the Fonseca et al. (2018) algorithm, BooST, impractical for very large datasets. In Appendix B, we propose two refinements to the BooST algorithm to remove the runtime's quadratic dependence on the number of observations and to test all potential splits with a single pass through the data. This allows us to efficiently scale the algorithm to problems with several hundred covariates and have it run in a couple minutes, rather than a few days.

Another advantage of the linear regression formulation is that it is straightforward to incorporate instruments in the estimation routines. Assume that we have access to a set of instruments Z, such that local estimation equation holds:

$$E[Z'(p-F(X))|X] = 0$$

Then we can introduce the following loss function:

$$L(p,F) = (p-F(X))'P_Z(p-F(X))$$

and apply the gradient boosting algorithm with smooth trees. At each step m, we fit the residual:

$$r = -\gamma_m \bigtriangledown_{F_{m-1}} L(p, F_{m-1}(X))$$

with a weak learner $h_m(x)$ that is a smooth tree, using Z as a matrix of instruments. By construction, the residual is orthogonal to the matrix of instruments at each step of the estimation routine, resulting in a final estimator that satisfies the local moment condition for all values of X. We do not currently have a proof of consistency, but provide Monte Carlos Appendix B to justify this approach. Further, we note the similarity between this approach and that of Athey et al. (2019), which uses decision trees rather than smooth trees, but provides some theoretical guarantees of consistency.

This algorithm provides several advantages. First, it provides a smooth pricing surface for which derivatives can be easily calculated. Second, it allows us to choose relevant regressors in a data driven manner, akin to the standard gradient boosting algorithm. Finally, it allows us to instrument for amenities values, and therefore approximate the indirect effect of public transportation on house values. We turn next to a discussion of the instruments we use during estimation.

3.6 Instrumental Variables

Our principal concern is that the level of amenities is correlated with the error in the house price equation. The sign of this correlation is in general unknown. For instance, stores might locate in neighborhoods with more disposable income meaning that the level of amenities is positively correlated with the error term. We could correct for this by including neighborhood fixed effects. However, we might find that conditional on the neighborhood, amenities locate in areas with

lower rent, causing a negative correlation between their level and the unobserved error. Instead of including fixed effects for increasingly granular geographic regions, we instrument for the level of amenities using the growth in amenities in all neighborhoods excluding the location of interest. A firm's entry decision depends not only on the observed and unobserved characteristics of the neighborhood, but also on aggregate trends in demand and supply. For instance, a general rise in income will lead to more restaurants entering all markets and reflects shifts in demand that are uncorrelated with local unobservables. Similarly, citywide changes in the cost structure of firms will impact entry decisions, but will be orthogonal to local unobservables. Firm entry outside of the neighborhood will therefore be correlated with local entry but will be orthogonal to the unobservable error. This logic is similar to that of preference externality instumental variables (PEIV) and the instruments used in Fan (2013).[9]

If aggregate trends in demand and supply shifters impact individual house prices then these instruments would be invalidated. This would be the case if house prices increase dues to increases in wages or increase or asset prices. However, we include several covariates that control for aggregate trends in house prices, such as the Case-Shiller index for Minneapolis. The identifying assumption is that conditional on the observed city-wide covariates, our instruments are orthogonal to local unobservables.

4 Data

4.1 Housing Data

This analysis quantifies the effect of the construction of the Blue Line by examining its impact on the sale price of residential properties. The sale records for each property were collected from the City of Minneapolis Tax Assessor Office, along with basic property characteristics, such as the year of construction, the square footage, the number of stories, the number of bedrooms and

[9]For a discussion of PEIV, see Li et al. (2020).

bathrooms. An identifier number (PID) unique to each property allowed us to merge this information with Hennepin County records in order to geocode the location of each property. Geocoding allowed us to determine the distance of each property from the closest Blue Line stations, as well as other transit options and nearby amenities. The analysis focuses on sales occurring between 2002 and 2006, the two years before and after the introduction of the Blue Line in 2004. This yields a total of 38,930 individual transactions, after excluding foreclosures and other non-market sales.

Table 1: Summary Statistics of Housing Data

	Mean	St. Dev.	Min	p25	p50	p75	Max	N
Sale Price	224,801	108,139	11,000	158,500	200,988	263,500	779,737	38,930
Distance to BL	2.046	1.301	0.046	0.904	1.888	2.942	5.236	38,930
Year Built	1,938	32.105	1,900	1,913	1,926	1,955	2,006	38,922
Sq. Feet	2,020	808.906	224	1,514	1,978	2,450	4,996	38,930
# of Stories	1.457	0.463	1.000	1.000	1.200	2.000	5.000	38,561
# of Baths	1.764	0.775	0.000	1.000	2.000	2.000	6.000	38,598

4.2 Transportation Data

Information on the public transit system in Minneapolis was obtained from the Minnesota Geospatial Commons, which yielded a dataset containing the location of over 5,518 transit stops within the City of Minneapolis across 147 separate transit routes. The closest transit stop for each transit line was identified for each residential property in the sample. In order to reduce the dimensionality of the data, the closest stop along the major transit axes between Minneapolis and its suburbs was also identified (see Appendix A for details.)

4.3 Neighborhood Amenities

A list of amenities within 0.5 miles of each property was compiled using ReferenceUSA data on local businesses, updated for each year between 2002 and 2006. We were thus able to track new businesses openings, existing businesses changing locations and businesses closing down within the City of Minneapolis over this time period. NAICS codes were used to categorize of each business, in order to calculate the density of each type of amenity around each property. This exercise yielded 28 amenity categories, such as “Full-Service Restaurants” or “Museums, Historical Sites, and Similar Institutions", to be used in later analysis (see Appendix A for details.) Further information on the quality of the amenities in the neighborhood of each property was scraped from Yelp, in particular the average rating of shopping outlets and restaurants, as well as information on the distance to the closest educational institution (childcare centers, elementary schools, high schools and colleges) to each property.

4.4 Demographic Data

Demographic information for each Census Tract was downloaded from Social Explorer for the 1990 and 2000 Decennial Census, and the 2008 - 2012 American Community Survey (ACS). The key variables of interest include the demographic make up of each neighborhood (% white residents, % black residents, % female residents), educational attainment (% college graduates, % high school graduates), economic variables (median household income, % living in poverty, % receiving public assistance, % unemployed), the share of owner occupied units and of vacant units, information about means of transportation to work (% commuting by car, % commuting by public transit) and the average commute length.

5 Results

5.1 Difference-in-Difference

The standard approach to a problem such as this is using a difference-in-differences framework where outcomes of properties located within a certain radius from the closest Blue Line station are compared to those of properties located beyond this radius. Thus, properties located within a 0.5 miles radius from the closest Blue Line stop have been assigned to the treatment group, while properties located between 0.5 and 1 miles of the closest Blue Line stop were assigned to the

Table 2: Difference in Difference Regression Results, Log Sale Price

	(1)	(2)	(3)
VARIABLES	Log Sale Price	Log Sale Price	Log Sale Price
Treatment	0.0100	0.0120	0.0361***
	(0.0164)	(0.0154)	(0.0139)
Treatment * Post	0.126***	0.123***	0.104***
	(0.0193)	(0.0187)	(0.0168)
Year Built			0.00260***
			(0.000122)
Sq. Feet			0.000123***
			(7.59e-06)
# of Stories			-0.119***
			(0.0104)
# of Baths			0.161***
			(0.00729)
Post	0.145***		
	(0.0130)		
Constant	12.09***	12.07***	6.673***
	(0.0109)	(0.0219)	(0.239)
Observations	10,541	10,541	10,295
R-squared	0.061	0.068	0.248
Month Fixed Effects	No	Yes	Yes
Year Fixed Effects	No	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

control group. Estimation results for this technique are reported in Table 2. The first specification reports results for a DiD routine with no controls, the second specification adds year and month fixed effects, and the last specification controls for housing characteristics, such as the year of construction, square footage, number of bedrooms, bathrooms and stories.

All specifications display strongly significant coefficients for the interaction terms capturing the DiD effect. The impact of the Blue Line in treatment neighborhoods is estimated to increase housing prices between 10.4 and 12.6%. These results should however be interpreted with caution.

5.2 Gradient Boosting with Decision Trees (XGBoost)

To recover the direct effect of the Blue Line, we estimated the conditional mean of the pricing surface using gradient boosting with decision trees. Following our difference-in-differences specification, we separated the sample into control and treatment groups based on $\frac{1}{2}$ mile and 1 mile concentric circles, training the model on the control group. We used a 10% hold-out sample and cross-validated the model by searching for the set of parameters that minimized the out-of-sample mean-squared error. The parameters we searched over included the maximum tree depth, the number of iterations, the rate of convergence, and the ℓ_2 regularization weight.

Of the 198 covariates included in our analysis, 132 were estimated to have a non-zero effect. Table 3 shows the top 20 covariates, ranked by their contribution to the reduction in mean-squared error. Unsurprisingly, a houses total area is very predictive of the house price, as well as the number of bathrooms and the age of the house. The Case-Shiller Index was also highly predictive, showing the sensitivity of individual house prices to aggregate trends. One measure of amenities, the number of restaurants within a half-mile, was also strongly predictive of house prices. Several transit routes were significant as well, including the minimum distance to a route 12 bus and a

Table 3: Top 20 covariates by importance

Feature	Gain	Cover	Frequency
Building Area	0.120	0.081	0.072
Ground Floor Area	0.107	0.040	0.036
No. of Bathrooms	0.068	0.022	0.012
Case-Shiller Index	0.060	0.041	0.035
No. of Restaurants	0.040	0.002	0.004
Min. Dist. 12	0.039	0.002	0.003
Second Floor Area	0.037	0.035	0.031
Min. Dist. 4	0.035	0.013	0.010
Age of House	0.035	0.047	0.038
Percent College	0.022	0.009	0.006
Min. Dist. 32	0.022	0.009	0.005
No. of Bedrooms	0.020	0.016	0.013
Min. Dist. 46	0.020	0.020	0.013
Housing Stock	0.016	0.006	0.004
Min. Dist. 27	0.015	0.022	0.014
Average Commute	0.014	0.001	0.003
Min. Dist. 21	0.012	0.031	0.018
Min. Dist. 22	0.011	0.046	0.026
Percent High School	0.010	0.004	0.006
Finished Basement	0.010	0.022	0.015

Note: 132 of 198 covariates had nonzero gain.

route 4 bus. These routes connect downtown Minneapolis with the wealthier suburbs, and so it is unsuprising that they are important determinants of house prices. Finally, several demographics were significant including the percent of college graduates in a census block and the average commute time. An advantage of using this machine learning approach is that we can select relevant covariates in a data driven way without imposing our model be sparse. This allows for localized, non-linear interactions across a high number of covariates.

Table 4 reports the estimated direct effect using gradient boosting with regression trees. The Pre-Treatment column shows that the algorithm trained on the control group, that is, property sales occurring between 0.5 and 1 miles of a Blue Line station, is able to correctly predict sale prices in the treatment group before the introduction of the Blue Line. While mean residual is positive,

Table 4: XGBoost: Treatment Effect

Predicted Residual:	Pre-Treatment	Post-Treatment	Implied Spillover
Mean	0.005	0.071	0.033
Std Dev.	(0.01)	(0.01)	–

Note: The implied spillover is calculated as the difference between the post-treatment prediction of the direct impact of the Blue Line (0.071) and the overall treatment effect calculated via DiD in specification (3) of Table 2 (0.104).

it is not significantly different from zero. After the introduction of the Blue Line, Post-Treatment prices in the treatment group increase by 7.1% more than predicted, even though the algorithm accounts for the introduction of new amenities and for demographic shifts in treatment neighborhoods (see Section 4 for a list of control variables). Thus the XGBoost estimation routine implies that the direct effect of the Blue Line is an increase in sale prices of 7.1% for properties located within 0.5 miles of a station.

Comparing these results with those from the DiD regression we can obtain an approximate measure of the implied spillover effect arising from the introduction of the Blue Line. Our preferred (and most conservative) specification for the DiD results predicts prices will increase by 10.4% in treatment neighborhoods. This increase can be though of as the total effect arising from the introduction of the Blue Line, compounding both the direct effect of access to light rail transit itself and the effect of the amenities changing because of increased accessibility in treatment neighborhoods. The difference between these two estimates (3.3%) can be though of as the implied spillover effect, that is, the impact that amenities changing as a result of the introduction of the Blue Line have on sale prices.

Heterogeneous effects for different types of neighborhoods can be obtained by regressing the residuals from the Post-Treatment predictions presented in Table 4 on neighborhood attribute. Table 5 reports the results of such a regression on tract-level characteristics captured by the 2000 Census. Neighborhoods that before the introduction of the Blue Line had a higher share of white residents saw a significant increase in their home values after the transit line was introduced. An

Table 5: XGBoost: Heterogeneous Effects

Predicted Residual:	Intercept	Distance to BL	% White	% Driving	Median Income
Mean	0.24	0.09	0.33	-0.60	-0.012
Std Dev.	(0.07)	(0.08)	(0.07)	(0.09)	(0.006)

increase in the share of white residents by 10 percentage points translates to a 3% higher increase in house prices. Wealthier neighborhoods and neighborhoods where a greater share of residents commute by car saw less of a benefit from the Blue Lines introduction. This is unsurprising as these are neighborhoods where there is less benefit from having access to public transportation. Interestingly, properties located further from the Blue Line also tend to see an appreciation in sale prices in the Post-Treatment period, although this effect is not statistically significant.

5.3 Smooth Trees

Table 6 reports the estimation results using the Boosted Smooth Trees estimation routine with our proposed instruments. The Pre-Treatment column shows that the algorithm trained on the control group is again able to correctly predict sale prices in the treatment group, with the prediction residuals for sale prices in the treatment group clustering around zero. After the introduction of the Blue Line, Post-Treatment prices in the treatment group increase by 5.5% as a direct effect of the Blue Line on property prices. This algorithm also allows us to directly approximate the spillover. To do this, we hold the level of amenities fixed at their pre-Blue Line levels and predict what housing prices would have been after it was introduced and compare these results to the predicted values post-introduction. This give us an approximation of $E[f(X_{it}^t) - f(X_{it}^c)]$, and thus the indirect effect. The change in amenities are predicted to increase the sale prices of properties located in the treatment group by a further 5.8%, implying that the total effect of the Blue Line on property prices is around 11.3%, remarkably close to the DiD prediction reported in Table 2. Following a similar procedure without instruments found a spillover of 1.3%, meaning that results that do not account for endogeneity would be downward bias and would

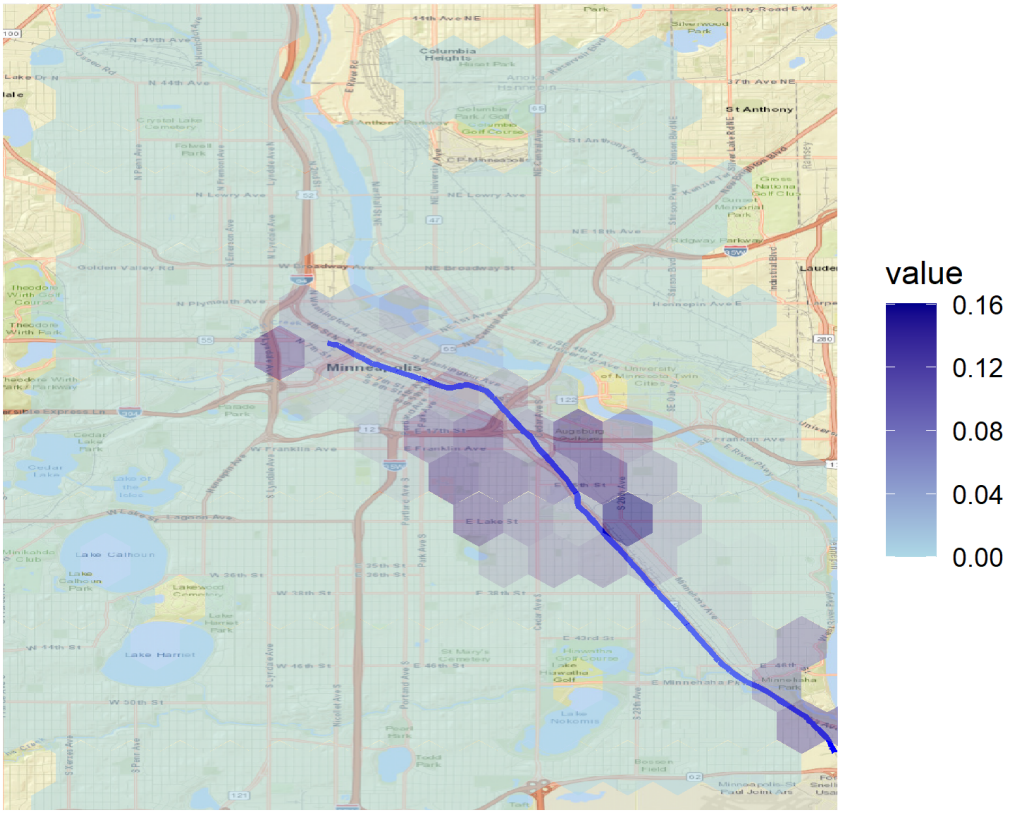
Table 6: Boosted Smooth Trees

Predicted Residual:	Pre-Treatment	Post-Treatment	Spillover
Mean	0.0009	0.0546	0.0584
Std Dev.	(0.0166)	(0.0198)	(0.0255)

tend to overstate the direct effect of the Blue Line relative to the indirect. With instruments, we find that the indirect effect accounts for over 50% of the total effect and is therefore an important channel through which public transportation impacts housing prices and consumer utility.

These effects are not homogenous and depend on where houses are located along the Blue Line. Figure 3 plots the treatment effect averaged across groups of houses along with the path of the Blue Line. The treatment effect is minimal for downtown, meaning that houses located in the

Figure 3: Predicted Change in Housing Prices



city-center did not see much of a pricing effect after the line was introduced. This could be for several reasons. First, downtown is already served by several bus lines, so there is less need for public transportation. Second, higher property prices might have discouraged business entry in the wake of the Blue Lines arrival. Houses just outside of the city-center benefited the most. This includes houses in gentrifying neighborhoods such as East Phillips and Corcoran. These neighborhoods benefited from having additional direct transportation to downtown, while also seeing a significant boom in local businesses.[10]

6 Conclusion

This paper applies recent advances in machine learning methods to investigate the impact that the construction of the METRO Blue Line had on housing prices and neighborhood amenities in Minneapolis. While many studies exist on the impact of mass transit on the urban environment, these studies generally do not decompose the overall impact of the introduction of a new mass transit system into direct and indirect effects. We apply a smooth tree learning algorithm to predict the direct and indirect effect of the introduction of the Blue Line. Our methodological contribution is a scalable algorithm for smooth tree boosting and a framework to incorporate instruments within this technique to control for endogeneity.

Our results show that that the price of properties located within 0.5 miles of a light rail station increased by around 11.3% compared to houses located further away. This can be thought of as the total impact of the Blue Line on local housing prices, encompassing both the direct benefit of improved access to public transit and the indirect benefit of an increase in the number neighborhood amenities. The direct impact of access to the light rail itself is estimated to increase local housing prices by 5.5%, while the spillover effect due to changes in amenities is quantifiable at

[10]Note: Need to add in the graph of local businesses.

5.8%. Thus, the majority of the overall appreciation in housing prices following the introduction of the Blue Line is not due to residents MWTP for public transit but is rather a spillover effect attributable to an increase in the number of amenities around light rail stations.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of californiaâĂŹs tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *American economic review 93*(1), 113–132.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Bae, C.-H. C., M.-J. Jun, and H. Park (2003). The impact of seoul's subway line 5 on residential property values. *Transport policy 10*(2), 85–94.

Bajari, P. and C. L. Benkard (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of political economy 113*(6), 1239–1276.

Bajic, V. (1983). The effects of a new subway line on housing prices in metropolitan toronto. *Urban studies 20*(2), 147–158.

Baum-Snow, N. and M. E. Kahn (2000). The effects of new public projects to expand urban rail transit. *Journal of Public Economics 77*(2), 241–263.

Berry, S. and J. Waldfogel (2010). Product quality and market size. *The Journal of Industrial Economics 58*(1), 1–31.

Billings, S. B. (2011). Estimating the value of a new transit option. *Regional Science and Urban Economics 41*(6), 525–536.

Bowes, D. R. and K. R. Ihlanfeldt (2001). Identifying the impacts of rail transit stations on residential property values. *Journal of Urban Economics 50*(1), 1–25.

Cervero, R. (1994). Rail transit and joint development: Land market impacts in washington, dc and atlanta. *Journal of the American Planning Association 60*(1), 83–94.

Cervero, R. and M. Duncan (2002). Land value impacts of rail transit services in los angeles county. *Report prepared for National Association of Realtors Urban Land Institute*.

Clower, T. L., B. L. Weinstein, et al. (2002). The impact of dallas (texas) area rapid transit light rail stations on taxable property valuations. *Australasian Journal of Regional Studies, The 8*(3), 389.

Damm, D., S. R. Lerman, E. Lerner-Lam, and J. Young (1980). Response of urban real estate values in anticipation of the washington metro. *Journal of Transport Economics and Policy*, 315–336.

Debrezion, G., E. Pels, and P. Rietveld (2007). The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics 35*(2), 161–180.

Debrezion, G., E. Pels, and P. Rietveld (2011). The impact of rail transport on real estate prices: an empirical analysis of the dutch housing market. *Urban Studies 48*(5), 997–1015.

Dewees, D. N. (1976). The effect of a subway on residential property values in toronto. *Journal of Urban Economics 3*(4), 357–369.

Diamond, R. and T. McQuade (2019). Who wants affordable housing in their backyard? an equilibrium analysis of low-income property development. *Journal of Political Economy 127*(3), 1063–1117.

Dueker, K. J. and M. J. Bianco (1999). Light-rail-transit impacts in portland: The first ten years. *Transportation Research Record 1685*(1), 171–180.

Duncan, M. (2008). Comparing rail transit capitalization benefits for single-family and condominium units in san diego, california. *Transportation Research Record 2067*(1), 120–130.

Fonseca, Y., M. Medeiros, G. Vasconcelos, and A. Veiga (2018). Boost: Boosting smooth trees for partial effect estimation in nonlinear regressions. *arXiv preprint arXiv:1808.03698*.

Forrest, D., J. Glen, and R. Ward (1996). The impact of a light rail system on the structure of house prices: a hedonic longitudinal study. *Journal of Transport Economics and Policy*, 15–29.

Gatzlaff, D. H. and M. T. Smith (1993). The impact of the miami metrorail on the value of residences near station locations. *Land Economics*, 54–66.

Gibbons, S. and S. Machin (2005). Valuing rail access using transport innovations. *Journal of urban Economics 57*(1), 148–169.

Goetz, E. G., K. Ko, A. Hagar, H. Ton, and J. Matson (2010). The hiawatha line: impacts on land use and residential housing value.

Grass, R. G. (1992). The estimation of residential property values around transit station sites in washington, dc. *Journal of Economics and Finance 16*(2), 139–146.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hess, D. B. and T. M. Almeida (2007). Impact of proximity to light rail rapid transit on station-area property values in buffalo, new york. *Urban studies 44*(5-6), 1041–1068.

Hewitt, C. M. and W. Hewitt (2012). The effect of proximity to urban rail on housing prices in ottawa. *Journal of Public Transportation 15*(4), 3.

Ho, J. (2016). Machine learning for causal inference: An application to air quality impacts on house prices.

Hurst, N. B. and S. E. West (2014). Public transit and urban redevelopment: The effect of light rail transit on land use in minneapolis, minnesota. *Regional Science and Urban Economics 46*, 57–72.

Ihlanfeldt, K. R. (2003). Rail transit and neighborhood crime: the case of atlanta, georgia. *Southern Economic Journal*, 273–294.

Immergluck, D. (2009). Large redevelopment initiatives, housing values and gentrification: the case of the atlanta beltline. *Urban Studies 46*(8), 1723–1745.

Kilpatrick, J., R. Throupe, J. Carruthers, and A. Krause (2007). The impact of transit corridors on residential property values. *Journal of Real Estate Research 29*(3), 303–320.

Knaap, G. J., C. Ding, and L. D. Hopkins (2001). Do plans matter? the effects of light rail plans on land values in station areas. *Journal of Planning Education and Research 21*(1), 32–39.

Ko, K. and X. J. Cao (2013). The impact of hiawatha light rail on commercial and industrial property values in minneapolis. *Journal of Public Transportation 16*(1), 3.

McDonald, J. F. and C. I. Osuji (1995). The effect of anticipated transportation improvement on residential land values. *Regional science and urban economics 25*(3), 261–278.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics 3*(4), 303–328.

McMillen, D. P. and J. McDonald (2004). Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. *Real Estate Economics 32*(3), 463–486.

Mohammad, S. I., D. J. Graham, P. C. Melo, and R. J. Anderson (2013). A meta-analysis of the impact of rail projects on land and property values. *Transportation Research Part A: Policy and Practice 50*, 158–170.

Nelson, A. C., D. Eskic, S. Hamidi, S. J. Petheram, R. Ewing, and J. H. Liu (2015). Office rent premiums with respect to light rail transit stations: Case study of dallas, texas, with implications for planning of transit-oriented development. *Transportation Research Record 2500*(1), 110–115.

Pakes, A. (2003). A reconsideration of hedonic price indexes with an application to pc's. *American Economic Review 93*(5), 1578–1596.

Palmquist, R. B. (2005). Property value models. *Handbook of environmental economics 2*, 763–819.

Pan, H. and M. Zhang (2008). Rail transit impacts on land use: Evidence from shanghai, china. *Transportation Research Record 2048*(1), 16–25.

Pan, Q. (2013). The impacts of an urban light rail system on residential property values: a case study of the houston metrorail transit line. *Transportation Planning and Technology 36*(2), 145–169.

Pilgram, C. A. and S. E. West (2018). Fading premiums: The effect of light rail on residential property values in minneapolis, minnesota. *Regional Science and Urban Economics 69*, 1–10.

Portnov, B., B. Genkin, and B. Barzilay (2009). Investigating the effect of train proximity on apartment prices: Haifa, israel as a case study. *Journal of Real Estate Research 31*(4), 371–395.

Rewers, J. M. (2010). *Identifying the impacts of light rail station location on residential property values in the city of Sacramento*. Ph. D. thesis.

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

Seo, K., A. Golub, and M. Kuby (2014). Combined impacts of highways and light rail transit on residential property values: a spatial hedonic price model for phoenix, arizona. *Journal of Transport Geography 41*, 53–62.

Taylor, L. O. (2003). The hedonic method. In *A primer on nonmarket valuation*, pp. 331–393. Springer.

Tsivanidis, N. (2018). The aggregate and distributional effects of urban transit infrastructure: Evidence from bogotáâĂŹs transmilenio. *Job Market Paper*.

Varian, H. R. (2014, May). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Wagner, G. A., T. Komarek, and J. Martin (2017). Is the light rail âĂIJtideâĂİ lifting property values? evidence from hampton roads, va. *Regional Science and Urban Economics 65*, 25–37.

Weinberger, R. R. (2001). Light rail proximity: Benefit or detriment in the case of santa clara county, california? *Transportation Research Record 1747*(1), 104–113.

Zheng, S., Y. Xu, X. Zhang, and R. Wang (2016). Transit development, consumer amenities and home values: Evidence from beijing's subway neighborhoods. *Journal of Housing Economics 33*, 22–33.

A Data Appendix

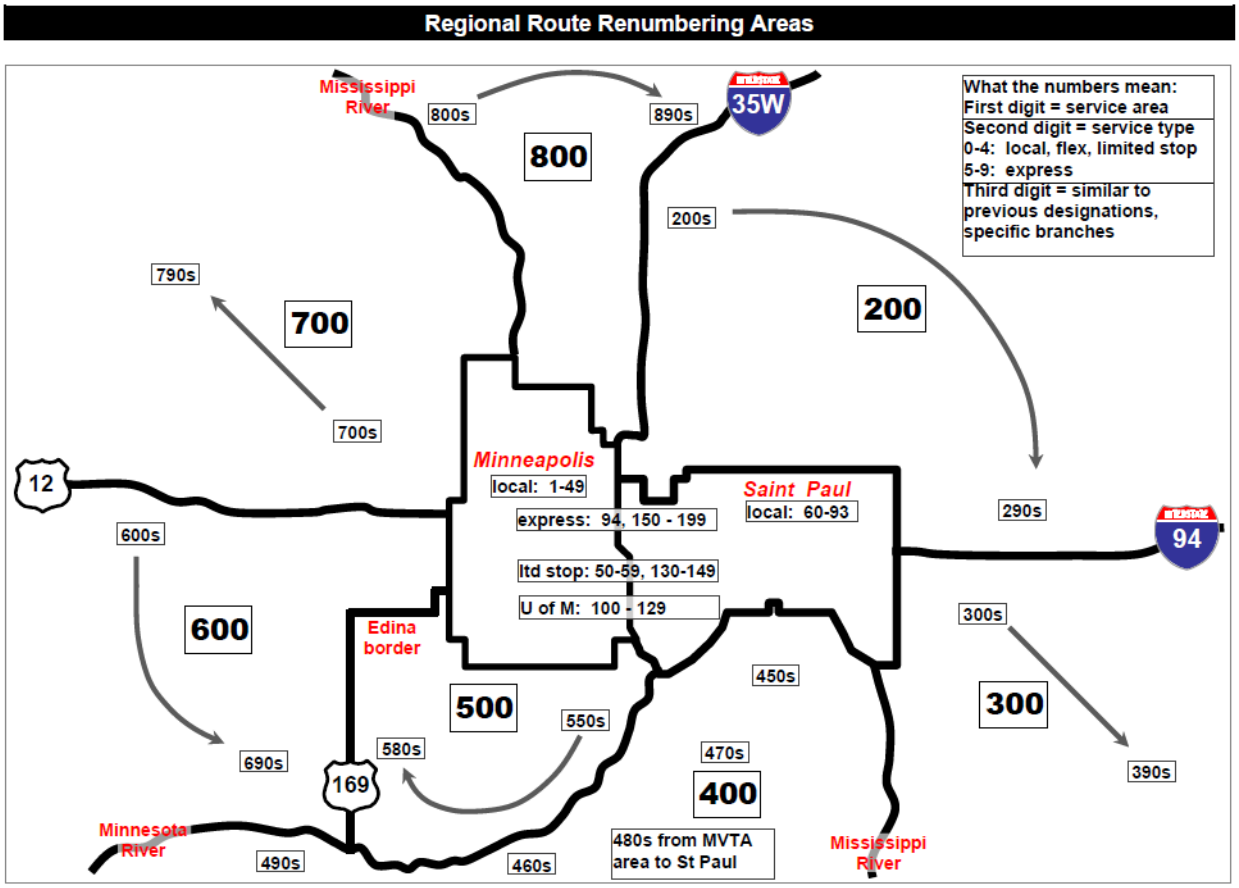
A.1 Transit Data

The full set of transit data includes information on the location of 5,518 transit stops across 147 transit routes. The closest stop to each of the downtown and local routes (route numbers 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 14, 17, 18, 19, 21, 22, 23, 25, 27, 30, 32, 39, 46, 53, 59, 133, 134, 135, 141, 146, 156) was calculated for each property in Minneapolis.

In order to reduce the dimensionality of the data, the closest stop along each of the major axes connecting Minneapolis to its suburbs was calculated as illustrated in Table 7, based on information obtained from the Twin Cities Metro Transit.

Table 7: Summarized Transit Variables

Variable	Direction	Route Number
Closest 200 Route	Roseville, White Bear Lake	250, 252, 261, 263, 264, 270, 272, 288.
Closest 300 Route	Woodbury	353, 355, 365, 375.
Closest 400 Route	Eagan	436, 446, 452, 460, 464, 465, 467, 470, 472, 475, 476, 477, 478, 479, 490, 491.
Closest 500 Route	Edina, Bloomington	515, 535, 552, 553, 554, 558, 578, 579, 587, 588, 589, 597.
Closest 600 Route	Minnetonka, Eden Praire	600, 612, 643, 645, 652, 663, 664, 667, 668, 670, 671, 672, 673, 677, 679, 690, 695, 697, 698, 699.
Closest 700 Route	Plymouth, Maple Grove, Brooklin Park	721, 724, 742, 747, 755, 756, 758, 760, 761, 762, 763, 764, 765, 766, 767, 768.
Closest 800 Route	Coon Rapids	824, 825, 850, 852, 854, 865, 887, 888.
Closest UofM Route	University of Minnesota	111, 113, 114, 115, 118, 121, 122, 129.
Closest Route to St Paul	St Paul	61, 67, 74, 94.



Source: Twin Cities Metro Transit

A.2 Amenities Data

Table 8 reports the NAICS codes and the average number of observations per year for the amenity categories included in our analysis obtained from ReferenceUSA.

Table 8: NAICS Codes of Neighborhood Amenities Variables

NAICS Category	Avg. Obs per Year	Description
22, 562	18,183	Utilities and Waste Management and Remediation Services
55	7,567	Management of Companies and Enterprises
61	145,603	Educational Services
92	176,390	Public Administration
442	39,264	Furniture and Home Furnishings Stores
443	42,862	Electronics and Appliance Stores
444	44,681	Building Material and Garden Equipment and Supplies Dealers
445	144,267	Food and Beverage Stores
446	47,312	Health and Personal Care Stores
447	18,993	Gasoline Stations
448	118,738	Clothing and Clothing Accessories Stores
451	60,489	Sporting Goods, Hobby, Musical Instrument, and Book Stores
452, 453	174,366	General Merchandise Stores, Miscellaneous Store Retailers
481	1,402	Air Transportation
541	2,576,613	Professional, Scientific, and Technical Services
621	1,115,182	Ambulatory Health Care Services
622	13,248	Hospitals
623	25,474	Nursing and Residential Care Facilities
711	75,620	Performing Arts, Spectator Sports, and Related Industries
712	20,727	Museums, Historical Sites, and Similar Institutions
713	37,170	Amusement, Gambling, and Recreation Industries
721	22,620	Accommodation
812	300,115	Personal and Laundry Services
813	376,453	Religious, Grantmaking, Civic, Professional, and Similar Organizations
722310, 722320	21,587	Food Service Contractors, Caterers
722410, 722515	73,817	Drinking Places (Alcoholic Beverages), Snack and Nonalcoholic Beverage Bars
722511	285,943	Full-Service Restaurants
722513, 722514	12,268	Limited-Service Restaurants, Cafeterias, Grill Buffets, and Buffets

A.3 Population Changes by Neighborhood

Table 9: Population Change in Census Tracts Adjacent to the Blue Line, 2000-2010

Census Tract	Neighborhood	Pop. 2000	Pop. 2010	Growth Rate
5901	Elliot Park	3,060	3,166	0.03
11998	Minnehaha	4,058	3,980	- 0.02
104400	Downtown West	1,499	2,097	0.40
104800	Cedar Riverside	7,551	8,094	0.07
105400	Elliot Park	3,416	3,527	0.03
106000	Ventura Village	3,462	3,339	- 0.04
106200	Seward	3,356	3,499	0.04
107400	Longfellow	1,713	1,726	0.01
107500	Longfellow/Seward	2,019	1,988	- 0.02
108600	Corcoran/Powderhorn Park	3,087	2,880	- 0.07
108700	Corcoran/Standish	3,550	3,274	- 0.08
108800	Howe/Longfellow	3,813	3,786	- 0.01
110200	Standish	3,518	3,522	0.00
110400	Hiawata/Howe	2,929	2,733	- 0.07
110500	Hiawata/Howe	4,438	4,694	0.06
111100	Ericsson	3,149	3,192	0.01
125900	East Phillips	4,147	4,269	0.03
126100	Downtown East/West	3,210	4,938	0.54
126200	North Loop	1,515	4,291	1.83

Source: US Cenus Bureau.

B Efficient Boosted Smooth Trees

There are two principal difficulties in using smooth trees for gradient boosting. First, for each split we test, we need to re-regress the residual on the matrix of leaf node probabilities. The time complexity of this regression is $O(C^2N)$, where C is the number of leaves and N the number of observations. If we test each observation as a splitting point, then the total time complexity is given by $O(C^2N^2)$. So smooth trees increase quadratically in the depth of the trees and the number of observations. Second, when using instruments, we need to form the product of the leaf probabilities with the instruments. The time complexity of this step is $O(KN)$, where K is the number of instruments. Repeating this multiplication N times yields an asymptotic rate of $O(KN^2)$. The purpose of this appendix is to proposed an algorithm that cuts these rates by a factor of N.

The key idea is to transform the problem so that we can update the gain by changing a single covariate at a time, eliminating the factor C^2. This is done in a manner analogous to updating a Kalman filter, where we use the bordering method and a pr-recalculated matrix inverse to perform the regression. We then use a sigmoid function that closely approximates the logit sigmoid but has the added property of being multiplicatively separable in its inputs. This allows us to efficiently calculate the instrument moments for any split in the data.

Let P_{t-1} be the matrix of choice probabilities as of step $t-1$. Note that each column of P_{t-1} represents a leaf of the smooth tree, with each row of $P_{t-1,j}$ being the probability that X_i ends up in leaf j. We want to test whether a branch is added to leaf j, such that

$$P_t = [P_{t-1,-j}, P_{t-1,j}L(X_i), P_{t-1,j}(1-L(X_i))]$$

Let P_z be the projection matrix for the instruments Z. Define the new regressors

$$\tilde{y} = P_z y$$

$$\tilde{P}_t = P_z P_t$$

$$\tilde{P}_{t-1} = P_z P_{t-1}$$

The residual from a ridge regression is given by

$$R(y, X, \lambda) = y'(I - (1-\lambda)X(X'X + \lambda I)^{-1}X')y$$

We accept this addition if it maximizes the gain

$$\begin{aligned} G(\tilde{y}, \tilde{P}_t, \tilde{P}_{t-1}, \lambda) &= \frac{1}{(1-\lambda)}\left(R(\tilde{y}, \tilde{P}_{t-1}, \lambda) - R(\tilde{y}, \tilde{P}_t, \lambda)\right) \\ &= \tilde{y}'\tilde{P}_t(\tilde{P}_t'\tilde{P}_t + \lambda I)^{-1}\tilde{P}_t'\tilde{y} - \tilde{y}'\tilde{P}_{t-1}(\tilde{P}_{t-1}'\tilde{P}_{t-1} + \lambda I)^{-1}\tilde{P}_{t-1}'\tilde{y} \end{aligned}$$

Redefine P_t as

$$P_t = [P_{t-1}, P_{t-1,j}L(X_i, c_t)]$$

and $\tilde{P}_t$ is constructed as before. This $\tilde{P}_t$ gives identical coefficients and residuals as the previous one, but only involves a single new regressor, rather than two. As we update $L(X_i, c_t)$, the term P_{t-1} stays fixed. For ease of notation, let $B = P_{t-1}$ and $A = P_{t-1,j}L(X_i, c_t)$. Then the term $(\tilde{P}_t'\tilde{P}_t + \lambda I)^{-1}$ can be written as the inverse of a symmetric block matrix

$$(\tilde{P}_t'\tilde{P}_t + \lambda I)^{-1} = \begin{bmatrix} B'B + \lambda I & B'A \\ A'B & A'A + \lambda \end{bmatrix}^{-1}$$

Here, A is a $N \times 1$ vector, so we can re-write this inverse using the bordering method. This states that the inverse of a a bordered matrix is given by

$$\begin{bmatrix} Q & \delta \\ \delta' & Z \end{bmatrix}^{-1} = \begin{bmatrix} Q^{-1} + \frac{Q^{-1}\delta\delta' Q^{-1}}{\mu} & -\frac{Q^{-1}\delta}{\mu} \\ -\frac{\delta' Q^{-1}}{\mu} & \frac{1}{\mu} \end{bmatrix}$$

where

$$\mu = Z - \delta' Q^{-1} \delta$$

Therefore, our inverse becomes

$$\begin{bmatrix} (B'B+\lambda I)^{-1} + \frac{(B'B+\lambda I)^{-1}B'AA'B(B'B+\lambda I)^{-1}}{\mu} & -\frac{(B'B+\lambda I)^{-1}B'A}{\mu} \\ -\frac{A'B(B'B+\lambda I)^{-1}}{\mu} & \frac{1}{\mu} \end{bmatrix}$$

with

$$\mu = A'A + \lambda - A'B(B'B+\lambda I)^{-1}B'A$$

The residual can be expressed as

$$\tilde{y}'\tilde{P}_t(\tilde{P}_t'\tilde{P}_t + \lambda I)^{-1}\tilde{P}_t'\tilde{y} =$$

$$y'B(B'B+\lambda I)^{-1}B'y + \frac{1}{\mu}y'B(B'B+\lambda I)^{-1}B'AA'B(B'B+\lambda I)^{-1}B'y$$

$$-\frac{2}{\mu}y'AA'B(B'B+\lambda I)^{-1}B'y$$

$$\frac{1}{\mu} y'AA'y$$

Note that

$$\tilde{y}'\tilde{P}_{t-1}(\tilde{P}'_{t-1}\tilde{P}_{t-1}+\lambda I)^{-1}\tilde{P}'_{t-1}\tilde{y} = y'B(B'B+\lambda I)^{-1}B'y$$

so these terms cancel. This leaves

$$G(\tilde{y},\tilde{P}_t,\tilde{P}_{t-1},\lambda) = \frac{1}{\mu}\left(y'B(B'B+\lambda I)^{-1}B'A - y'A\right)^2$$

which, conditional on A, can be calculated with three dot products ($A'B$, $y'B(B'B+\lambda I)^{-1}B'A$ and $A'A$) and one low-dimensional matrix multiplication that scales with the depth of the trees.

The principal question then is how quickly can we construct the matrix $P'_{t-1}A$ or $Z'A$. The semi-naive approach would be to update A in each step and calculate these products. I say semi-naive because this is necessary for most sigmoid functions, and is what greatly reduces the computational efficiency of smooth trees. An alternative procedure would be to use the following sigmoid

$$L(X_i,c_j) = \begin{cases} 1-\frac{1}{2}\frac{2^{c_j}}{2^{X_i}} & \text{if } c_j < X_i \\ \frac{1}{2}\frac{2^{X_i}}{2^{c_j}} & \text{if } c_j \geq X_i \end{cases}$$

Assume that the kth regressor, X_k, is sorted from smallest to largest, and that Z_k is sorted based on the ordering of X_k.[11] The goal is to test all elements of X_k to find the split that maximizes the gain. At iteration 1, we have

[11] This only needs to be done once at the start of the algorithm for all regressors.

$$A_1 = Z'(P_{t-1,j}L(X_k, X_1)) \tag{1}$$

$$= \sum_i Z_i P_{ij}^{t-1} \left(1 - \frac{1}{2}\frac{2^{X_{1k}}}{2^{X_{ik}}}\right) \tag{2}$$

This can be broken into two parts

$$\bar{Z}_1 = \sum_i Z_i P_{ij}^{t-1}$$

and

$$\tilde{Z}_1 = \frac{1}{2}\sum_i Z_i P_{ij}^{t-1}\frac{2^{X_{1k}}}{2^{X_{ik}}}$$

This define

$$Z_1^r = \bar{Z}_1 - \tilde{Z}_1$$

and

$$Z_1^l = 0$$

so that $A_1 = Z_1^l + Z_1^r$. The key idea is that going from $X_{m-1,k}$ to X_{mk} only involves updating according to

$$\bar{Z}_m = \bar{Z}_{m-1} - Z_{m-1}P_{m-1,j}^{t-1}$$

$$\tilde{Z}_m = \left(\tilde{Z}_{m-1} - \frac{1}{2} Z_{t-1} P_{1j}^{t-1} \right) \frac{2^{X_{m,k}}}{2^{X_{m-1,k}}}$$

$$Z_m^l = \left(Z_{m-1}^l + \frac{1}{2} Z_{m-1} P_{m-1,j}^{t-1} \right) \frac{2^{X_{m-1,k}}}{2^{X_{m,k}}}$$

and

$$A_m = \bar{Z}_m - \tilde{Z}_m + Z_m^l$$

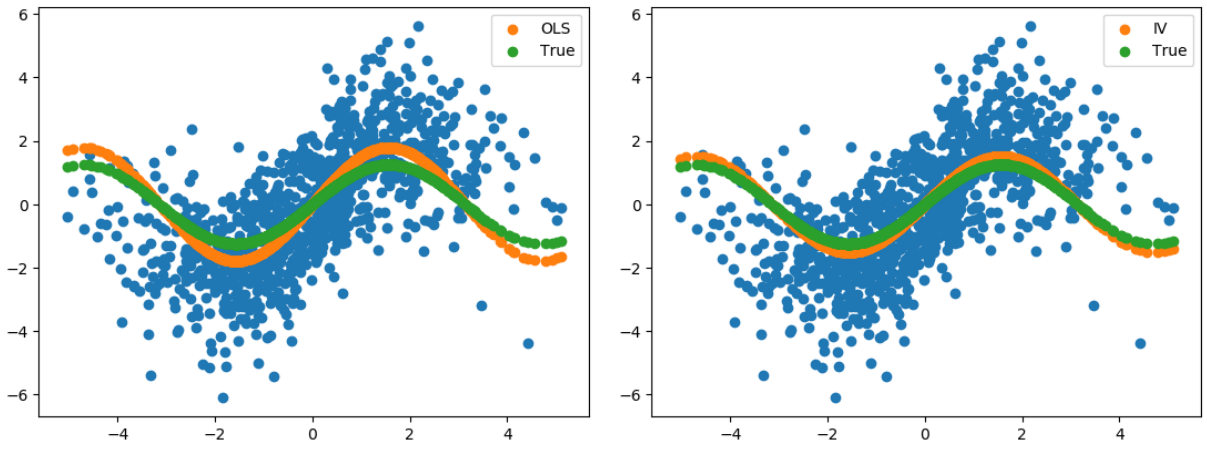
This allows us to calculate all potential A_m with a single pass through the data, reducing the time complexity of calculating A by a factor of N, from $O(N^2K)$ to $O(NK)$.

B.1 Monte Carlos

We used $N = 1{,}000$ observations with $Z_i, \varepsilon_i \sim N(0, 0.5^2)$ and $X_i = Z_i + \frac{1}{2}\varepsilon_i$. The dependent regressor is determined by the following nonlinear relationship

$$y_i = 1.25 \sin(X_i) + \varepsilon_i$$

We used a learning rate of $\gamma = 0.05$ and a minimum of 50 observations per node. Finally, we trained with $M = 350$ iterations and cross-validated to determine the optimal λ. Figure 4 plots the data and the parameter OLS and IV estimates. OLS tends to over-predict due to the positive correlation between the residual and the regressor. The instrumental variable estimates improve this somewhat but tend to still over-predict. Figure 5 shows the non-parametric estimates using



(a) OLS (b) IV

Figure 4: Parametric Estimation

Boosted Smooth Trees. There is close agreement in the range of -3.5 to 3.5, and divergence beyond that point. This is largely due to the lack of observations in the tails of the distribution. Using the above values, we repeated this exercise 100 times and calculate the RMSE for each sample. The mean RMSE was 0.09 and the standard deviation was 0.018.

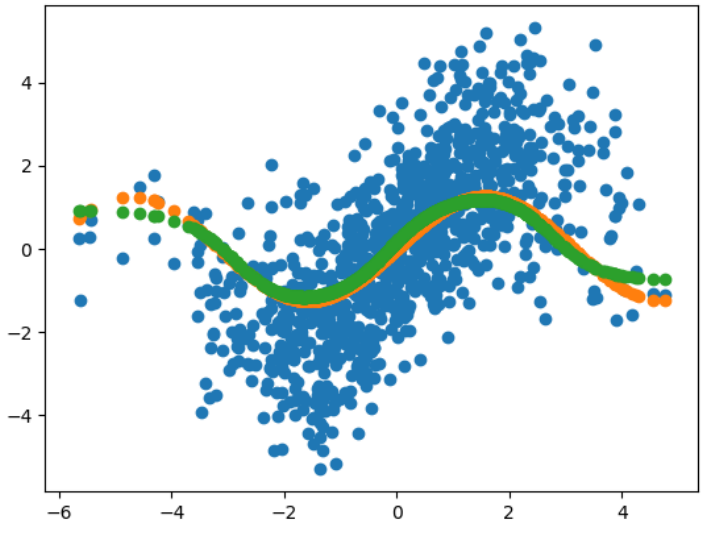


Figure 5: Endogenous Regressor