Essays in Urban Economics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Veronica Postal

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Thomas J. Holmes, Adviser

August, 2020
Acknowledgments

I am very grateful to my advisor, Thomas J. Holmes, and my committee members, and Amil Petrin, Joel Waldfogel and Doireann Fitzgerald, for their advice and support over the past six years. Their guidance has been invaluable in teaching me how to select, develop and present research projects. Additionally, my research has benefited from conversations with members of the University of Minnesota Applied Microeconomics workshop and fellow graduate students in the Department of Economics. I owe special thanks to Mark Ponder who co-authored the final chapter of my dissertation. Further, I acknowledge the financial support of the Graduate School of the University of Minnesota. Finally, I would like to thank the Department of Economics staff, especially Wendy Williamson.
Dedication

To my family, in particular my parents, Maurizio and Cosetta, unwavering support and encouragement over the years. To Mark and Emmy, for never leaving my side.
Abstract

This dissertation is comprised of three essays, each dealing with topics in Urban Economics and Applied Microeconomics.

In the first chapter, I use a dynamic discrete-continuous choice model to examine the decision to invest in home improvement. In each time period, households face the decision of whether to reoptimize their housing consumption, either through home improvement or by selling their property. The structural model is estimated using a uniquely rich micro-level dataset that encompasses each property in Minneapolis for a period of almost 20 years. I reconstruct the optimal investment policy and choice probabilities as a function of a property’s housing quality and neighborhood quality. Then, I explore a counterfactual scenario in which the cost of investment in home improvement is subsidized and show that such a policy would be effective in increasing the predicted level of investment. I find that policies aimed at encouraging home improvement can be a cost-effective tool to leverage private investment in housing renovation, and to promote urban revitalization without displacing the residents of low income neighborhoods.

In the second chapter, I investigate how the development of new apartment buildings can affect local property prices. On one hand, increasing the supply of available residential units is expected to lower the price of other housing options in a given area through a substitution effect. On the other hand, apartment building development
may produce aggregation economies and other spillovers increasing the desirability of a given neighborhood and in turn property prices. Estimating the net effect under a standard parametric framework is complicated by the non-linear interaction of geographic and temporal distance from the site of construction. I apply a new econometric technique developed by Diamond and McQuade (2019) to non-parametrically estimate the effect of apartment building construction on nearby residential property prices, transforming transaction prices into a price gradient that is a smooth function of time and distance and then numerically integrating over the estimated derivatives to measure changes in property prices. I find that property prices increase with distance from the site of a new development, suggesting that the substitution effect might be stronger than other potential spillovers, although the overall effect varies heterogeneously across different types of neighborhood.

In the third chapter, Mark Ponder and I examine the effect of the introduction of light rail transit in Minneapolis. We focus on decomposing the overall impact on 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 an indirect spillover effect through 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. (2018) to a high-dimensional dataset and to incorporate instrumental variables, allowing us to control for endogenous changes in amenities. 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%. 
Contents

1 Housing Choice, Home Improvement and Neighborhood Revital-
ization 1

1.1 Introduction 1

1.2 Literature Review 1

1.3 The MHFA Fix Up Loan Program 1

1.4 Data and Descriptive Facts 10

1.4.1 Housing Data 10

1.4.2 Descriptive Facts 14

1.4.3 Assumptions for Causal Inference 18

1.5 Model 22

1.5.1 State Variables 23

1.5.2 Utility 25

1.5.3 Value Function 27

1.5.4 Model Limitations 30

1.6 Estimation 32

1.6.1 Static Parameters 32

1.6.2 Structural Model Parameters 34

1.7 Results 38

1.7.1 Baseline Model 38

1.7.2 Policy Counterfactuals 43
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8 Simulation of Evolution of Aggregate Investment and Capitalization</td>
<td>49</td>
</tr>
<tr>
<td>1.9 Conclusion</td>
<td>55</td>
</tr>
</tbody>
</table>

2 The Spillover Effects of Rental Property Development  

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Introduction</td>
<td>57</td>
</tr>
<tr>
<td>2.2 Data</td>
<td>60</td>
</tr>
<tr>
<td>2.3 Exploratory Analysis</td>
<td>66</td>
</tr>
<tr>
<td>2.4 Housing Choice Model</td>
<td>71</td>
</tr>
<tr>
<td>2.5 Estimation</td>
<td>72</td>
</tr>
<tr>
<td>2.6 Results</td>
<td>76</td>
</tr>
<tr>
<td>2.6.1 Heterogeneous Effects</td>
<td>76</td>
</tr>
<tr>
<td>2.7 Conclusion</td>
<td>79</td>
</tr>
</tbody>
</table>

3 Accessibility or Amenities? Estimating the Value of Light Rail Transit  

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>81</td>
</tr>
<tr>
<td>3.2 Theoretical Framework</td>
<td>87</td>
</tr>
<tr>
<td>3.3 Estimating Treatment Effects</td>
<td>91</td>
</tr>
<tr>
<td>3.3.1 Cross Sectional Methods</td>
<td>92</td>
</tr>
<tr>
<td>3.3.2 Difference-in-Differences</td>
<td>93</td>
</tr>
<tr>
<td>3.3.3 Machine Learning</td>
<td>95</td>
</tr>
<tr>
<td>3.3.4 Boosted Smooth Trees</td>
<td>98</td>
</tr>
<tr>
<td>3.3.5 Instrumental Variables</td>
<td>101</td>
</tr>
<tr>
<td>3.4 Data</td>
<td>102</td>
</tr>
<tr>
<td>3.4.1 Housing Data</td>
<td>102</td>
</tr>
<tr>
<td>3.4.2 Transportation Data</td>
<td>103</td>
</tr>
<tr>
<td>3.4.3 Neighborhood Amenities</td>
<td>103</td>
</tr>
<tr>
<td>3.4.4 Demographic Data</td>
<td>104</td>
</tr>
<tr>
<td>3.5 Results</td>
<td>105</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Average Estimated Market Value and Average Sale Price by Year</td>
<td>13</td>
</tr>
<tr>
<td>1.2</td>
<td>Investment Probability, Sale Probability and Average Investment by ( q ) and ( P )</td>
<td>15</td>
</tr>
<tr>
<td>1.3</td>
<td>Investment Frequency by Investment Size</td>
<td>16</td>
</tr>
<tr>
<td>1.4</td>
<td>Investment Probability by Tenure Length</td>
<td>18</td>
</tr>
<tr>
<td>1.5</td>
<td>Joint Distribution of Investment Probability and Average Investment</td>
<td>19</td>
</tr>
<tr>
<td>1.6</td>
<td>Predicted Investment Probability</td>
<td>39</td>
</tr>
<tr>
<td>1.7</td>
<td>Predicted Optimal Investment</td>
<td>40</td>
</tr>
<tr>
<td>1.8</td>
<td>Model Fit: Empirical Data vs. Model Predictions</td>
<td>42</td>
</tr>
<tr>
<td>1.9</td>
<td>Change in Predicted Investment Probability and Average Investment</td>
<td>44</td>
</tr>
<tr>
<td>1.10</td>
<td>Change in Average Investment under Means Tested Policy</td>
<td>46</td>
</tr>
<tr>
<td>1.11</td>
<td>Model Predictions Under Alternative Policy Scenarios</td>
<td>47</td>
</tr>
<tr>
<td>1.12</td>
<td>Change in Aggregate Investment under Alternative Policy Scenarios</td>
<td>52</td>
</tr>
<tr>
<td>1.13</td>
<td>Cumulative Change in Neighborhood EMV under Alternative Policy Scenarios</td>
<td>53</td>
</tr>
<tr>
<td>1.14</td>
<td>Cumulative Change in Aggregation EMV under Alternative Policy Scenarios</td>
<td>54</td>
</tr>
<tr>
<td>2.1</td>
<td>New Construction in Minneapolis 1991-2016</td>
<td>61</td>
</tr>
<tr>
<td>2.2</td>
<td>Share of Rentals in Minneapolis 1991-2016</td>
<td>62</td>
</tr>
<tr>
<td>2.3</td>
<td>Apartment Trends in Minneapolis 1991-2016</td>
<td>63</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Summary Statistics</td>
<td>11</td>
</tr>
<tr>
<td>1.2</td>
<td>Common Building Permits</td>
<td>12</td>
</tr>
<tr>
<td>1.3</td>
<td>Summary Statistics for State Variables</td>
<td>14</td>
</tr>
<tr>
<td>1.4</td>
<td>Parameter Estimates for the Evolution of $P$</td>
<td>33</td>
</tr>
<tr>
<td>1.5</td>
<td>Parameter Estimates for the Evolution of $q$</td>
<td>34</td>
</tr>
<tr>
<td>1.6</td>
<td>Parameter Estimates for Structural Model</td>
<td>37</td>
</tr>
<tr>
<td>1.7</td>
<td>Total Investment Under Alternative Policy Scenarios</td>
<td>49</td>
</tr>
<tr>
<td>1.8</td>
<td>Parameter Estimates for the Evolution of $EMV$</td>
<td>51</td>
</tr>
<tr>
<td>2.1</td>
<td>Summary Statistics for Apartment Building Construction</td>
<td>63</td>
</tr>
<tr>
<td>2.2</td>
<td>Summary Statistics for Apartment Building Construction</td>
<td>66</td>
</tr>
<tr>
<td>2.3</td>
<td>Difference in Difference Regression Results, Log Sale Price</td>
<td>68</td>
</tr>
<tr>
<td>3.1</td>
<td>Summary Statistics of Housing Data</td>
<td>103</td>
</tr>
<tr>
<td>3.2</td>
<td>Difference in Difference Regression Results, Log Sale Price</td>
<td>105</td>
</tr>
<tr>
<td>3.3</td>
<td>Boosted Smooth Trees</td>
<td>107</td>
</tr>
<tr>
<td>A.1</td>
<td>Average Number of Returns Filed in Minneapolis, 2006-2016</td>
<td>125</td>
</tr>
<tr>
<td>A.2</td>
<td>Average Income by Median Neighborhood EMV</td>
<td>127</td>
</tr>
<tr>
<td>A.3</td>
<td>Minneapolis Mean and Median Income by Zip Code</td>
<td>128</td>
</tr>
<tr>
<td>A.4</td>
<td>Comparison of Estimates for Structural Model Parameters</td>
<td>131</td>
</tr>
<tr>
<td>A.5</td>
<td>Total Investment Under Alternative Policy Scenarios</td>
<td>135</td>
</tr>
</tbody>
</table>
Chapter 1

Housing Choice, Home Improvement and Neighborhood Revitalization

1.1 Introduction

Home improvement constitutes a substantial share of housing investment, netting just under $200 billion in 2018 and comprising over 35% of the overall spending on residential construction. This paper uses a uniquely rich micro-level dataset to estimate a dynamic structural model of housing choice that includes home improvement as a possible dimension to alter housing consumption. Even though households wishing to re-optimize their level of housing consumption can do so either by moving to a new property or by altering the characteristics of their current unit, the latter option is often ignored in the literature. Several studies in the housing demand literature examine housing choice within the framework of a structural model, but they generally omit home improvement as a possible dimension of housing consump-
For instance, Bayer et al. (2016) build a dynamic structural model of housing choice where households can alter their housing consumption only by moving to a new property in a new neighborhood. Most papers focusing specifically on home improvement, on the other hand, consist of reduced-form studies in a static setting, and thus they do not lend themselves to policy analysis. Yet, home improvement is strongly tied to the process of urban renewal, which is often a dimension of interest to policymakers.

Several studies have identified home improvement as a key dimension of neighborhood revitalization (DeGiovanì 1984, Helms 2003, Ellen and O’Regan 2011) and gentrification (Brueckner and Rosenthal 2009). It is also well-established that investment in housing renovation can generate positive externalities on the property values of nearby homes (DeGiovanì 1984, Ding et al. 2000, Ooi and Le 2013). These positive spillovers have been shown to persist even when housing rehabilitation is spurred by government programs (Simons et al. 2003, Edmiston 2012). Further, Rossi-Hansberg et al. (2010) argue that the housing externalities produced by neighborhood improvement programs can justify a role for government intervention, since they imply that the equilibrium allocation of investment differs from efficient outcomes. Moreover, Ioannides (2002) and Helms (2012) find evidence of the fact that housing renovation can create a positive feedback effect with neighborhood quality, and suggest that policies encouraging this type of investment are likely to be “spatially multiplied”. These findings suggest that government policies aimed at encouraging home improvement might be effective at leveraging private investment in housing renovation.

1 See for example Bishop (2008), Murphy (2018), Kennan and Walker (2011), Bayer et al. (2016) for some recent examples of dynamic structural models applied to the analysis of housing demand.
I contribute a structural model in which households are free to adjust their housing consumption either by moving to a new location or by altering their existing property through investment in home improvement. Households are assumed to be forward-looking and make their decisions in a finite, discrete-time setting. The household’s decision regarding the optimal amount to invest in home improvement is nested within the decision of what action to undertake, in a discrete-continuous choice framework. The estimation of this class of models can be challenging because the discrete choice creates kinks and non-concavities in the value functions that can in turn introduce discontinuities in the policy functions. I follow Iskhakov et al. (2017) and tackle this issue by introducing an idiosyncratic shock to the discrete choice problem; this technique facilitates the numerical solution of the model and allows for the use of maximum likelihood estimation to recover the structural model parameters.

I estimate the model using a panel of parcel-level observations encompassing every property in the City of Minneapolis for over two decades. This extraordinarily rich dataset includes each property’s sale history and assessed market value, property characteristics, and construction permits. Then, using the estimated model parameters, I explore a counterfactual scenario in which an existing program sponsored by the Minnesota Housing Finance Agency aimed at making home improvement more affordable for low-income households is expanded. I find that a program reducing the cost of home improvement by 10% would translate into a substantial increase in the average amount of investment in housing renovation, directly leveraging $2.42-$2.61 in private investment per dollar spent. I also find that this program would not encourage the displacement of the residents of low income neighborhoods. Further, the results of a simulation using the structural model parameters suggest that this positive effect on investment is likely to be compounded over time, with aggregate investment predicted to grow at an average of 2.18-2.25% faster as a result of a 10%
reduction in the cost of investment. However, since each dollar invested in home improvement is capitalized in housing values at a rate of $0.31 per dollar spent, the cost-effectiveness of the program falls below $1 if the metric of interest is capitalization in housing values rather than change in aggregate investment.

The rest of the paper is organized as follows. The next section reviews the relevant literature, while Section 1.3 introduces the Minnesota Fix Up Loan Program, a potential avenue for policy expansion. Section 1.4 presents the data and key descriptive facts, and the structural model is introduced in Section 1.5. Section 1.6 describes the estimation strategy, while Section 1.7 presents results and policy experiments and Section 1.8 introduces a simulation of the dynamic effects of their effects. Section 1.9 concludes.

1.2 Literature Review

This paper is closely related to several strands of literature, in particular to the literature on home improvement and housing renovation, and the literature on the effect of taxation on housing consumption.

It has been recognized at least since Mendelson (1977) that home improvement is a key dimension of housing consumption. Households that wish to alter their consumption of housing can move to a new property or they can change the existing property by investing in improving their current home. In this respect, framing household decisions as decisions between investing in home improvement or moving finds ample support in the home improvement literature (Scheer 1983, Montgomery 1992, Plant and Plaut 2010, among others) although it is rarely mentioned in more
general papers on housing consumption. Empirical findings support the view that home improvement and moving decisions belong in the same choice set. Potepan (1989) shows that the home improvement decision is sensitive to the cost of moving: when interest rates rise, the relative cost of moving increases, and households are more likely to invest to enhance their current home.

Most empirical studies in the home improvement literature try to examine what factors tend to increase the likelihood of home improvement. Household demographic characteristics seem to have an effect: households that are richer, whiter, or have high socioeconomic status tend to be more likely to invest in improving the quality of their property (Mendelsohn, 1977; Plaut and Plaut, 2010). The value of a home compared to that of its neighbors also seems to have an impact (Montgomery, 1992; Charles, 2013) although there is no consensus on its direction. Neighborhood attributes and social effects have also been shown to be important. Boehm and Ihlanfeldt (1986) find that, in addition to household characteristics and input costs, neighborhood quality is a key determinant of the likelihood of home improvement. Similarly, Helms (2003) finds that both housing characteristics and neighborhood attributes influence home improvement activity. The “spatial interdependence” of investment in housing quality has been confirmed by several subsequent studies, including Ioannides (2002), Helms (2012) and Munneke and Womack (2015).

In light of these findings, the model presented in this paper allows the decision to invest in home improvement to depend on both housing quality and neighborhood quality. While some authors, for instance Baker and Kaul (2002), note the importance of dynamic factors in the decision to invest in home renovation, to the best of my knowledge all studies in the home improvement literature so far have assumed a static environment. Since investment is an inherently forward-looking decision,
I build these insights into a dynamic structural model that allows households that wish to re-optimize their housing consumption to improve their current property or to move, and where housing quality and neighborhood quality can each have an impact on the investment decision.

Data limitations are likely one of the factors that has prevented this literature from progressing further. Only (Helms, 2003, 2012), Charles (2013) and Munneke and Womack (2015) have access to parcel-level data on construction permits or building characteristics; most authors in the literature rely instead on information gathered from the US Census Bureau, such as the American Housing Survey or the American Community Survey. I estimate my model using an uniquely extensive parcel-level dataset that encompasses every property unit in Minneapolis over more than two decades, including not only construction permits and building characteristics, but also sale history and estimated market values.

This paper is also connected with the rich literature that examines the role of taxation on housing choices. Laidler (1969), Aaron (1970), and Rosen (1979) are among the first authors to investigate how exemptions for mortgage interest and property taxes can affect housing demand, but primarily focus on the distributional implications of these provisions. More recent studies focusing on the distributional effects and welfare implications of the home mortgage interest deduction have been completed by Gervais (2002), Glaeser and Shapiro (2003), Poterba and Sinai (2008), and Hilber and Turner (2014), among others.

Mills (1987) and Poterba (1992) build theoretical models suggesting that tax incentives might affect housing consumption and tenure choice. Hanson (2012) uses variation in state level policies regarding home mortgage interest deductions to find
evidence of how these policies increase the size of purchased homes. This question has been addressed more recently through structural models (Chambers et al., 2009) that endogenize housing prices and rents (Flototto et al., 2016; Sommer and Sullivan, 2018) who find that the tax-exempt status of mortgage interest deductions affects housing consumption, especially at the top of the income distribution. Despite this robust literature on the effects of tax policy on housing consumption, none of these studies include home improvement as a potential dimension for housing consumption, nor do they examine the impact that taxation can have on the likelihood or magnitude of housing renovations.

A smaller set of papers looks at the effects of energy tax credits on home improvement, specifically energy-efficiency improvements, but examines this choice disjointedly from that of housing consumption. Dubin and Henson (1988) and Walsh (1989) do not find that energy tax credits have any effect on increasing the likelihood of households investing in energy conservation improvements, but more recent work (Hassett and Metcalf, 1995) finds a significant impact. Only Galster (1987) and Galster and Hesser (1988) look specifically at programs targeted at home improvement, using homeowners surveys to assess the impact of several locally administered programs aimed at encouraging home renovations in Minneapolis. These papers find that these programs were successful at increasing the likelihood of homeowners undertaking home enhancing investments and at leveraging private investment, but their estimates rely entirely on self-reported data by homeowners.

This paper uses parcel-level micro-data to estimate a structural model of housing consumption that allows for counterfactual analysis of how investment in home improvement would change under different tax policy regimes and incentive programs. The next section examines a potential avenue to implement such a policy.
1.3 The MHFA Fix Up Loan Program

This paper aims at exploring how policies subsidizing investment in home improvement might affect the likelihood and magnitude of home improvement. In general, government programs aimed at promoting home improvement tend to be limited in scope. The leading home improvement programs offered through the Department of Housing and Urban Development (HUD) primarily focus on facilitating loans to homeowners by insuring private lenders against losses arising from property improvement loans, and are not aimed at directly helping homeowners offsetting any costs associated with home improvement. For instance, the two most popular HUD programs for home improvement are the HUD Title 1 Property Improvement Loan Program and the FHA 203(k) Rehabilitation Mortgage Insurance Program. Both of these programs take the form of private loans backed by the government, without any direct financial assistance to homeowners.

There exist a few smaller-scale programs offered through the housing departments of local or county governments aimed at making home improvement more affordable, especially for low income households. For example, the Minnesota Housing Finance Agency (MHFA) offers the “Fix Up Loan Program” in order to help homeowners with repairs, remodels, and energy efficiency improvements, through a discounted fixed interest rate loan. This program has been set up with the objective of promoting affordable housing by making homeownership more accessible, preserving the housing stock, and supporting low income communities. The program is sponsored by the Minnesota Housing Finance Agency (MHFA) and is administered locally by the City of Minneapolis through its Community Planning & Economic Development (CPED) division, and is specifically aimed at encouraging the development and preservation of affordable housing.
Since this program is aimed at assisting low-income households meet their housing needs, the income of participating households may not exceed $141,000. Loans can range from $2,000 to $30,000, with repayment terms between one and 20 years. Appendix A.1 reports additional details of the MHFA Fix Up Loan Program available through the City of Minneapolis.

The loans’ interest rates depend on several factors, such as whether the property is already encumbered by an existing mortgage, by the proposed loan repayment term, and by the type of home improvements that homeowners wish to complete. For instance, the reference interest rate for a $30,000 secured 20 years loan is 5.38%.\footnote{These reference rates are reported in the Fix Up Loan application made available in June 2019 by the City of Minneapolis.} Assuming that these reference interest rates are representative of average interest rates within the program, households financing home improvement using a secured loan through the Fix Up Program have access to interest rates 0.5-1.40% lower than commercial rates\footnote{Bankrate.com reports that the average interest rate in June 2019 was 5.87% for a home equity loan, and 6.77% for a home equity line of credit (HELOC).}. This translates in a cost saving ranging between 3.96% and 10.60% for households participating in the program compared to commercial financing opportunities.

As it currently stands, the MHFA Fix Up Loan Program is relatively small. In recent years, an average of only 669 households in Minnesota have taken advantage of this opportunity, a negligible fraction of the state’s more than 2 million households.\footnote{The average enrollment is based on the figures reported for the Home Improvement Loan Program for 2016-2018, Minnesota Housing 2018 Program Assessment Report.} \footnote{The Minnesota State Demographic Center estimates a total of 2,221,628 households in the state in 2018.} This paper argues that this program would have the potential to promote investment in home improvement and the revitalization of depressed neighborhoods if it was made more widely available.
1.4 Data and Descriptive Facts

The main data source for this paper has been assembled from the City of Minneapolis Tax Assessor records. This is a uniquely rich dataset that includes extensive information on every property parcel located within the Minneapolis city limits, some 130,000 individual properties, spanning several decades.

1.4.1 Housing Data

One of the most thorny issues when thinking about housing consumption is that of accurately measuring housing quality and its evolution over time. One option would be to consider the sale price of a property as the market’s assessment of its quality. However, this method does not work well with a dynamic investment model, as households are assumed to make a decision regarding their property in each time period after observing the quality of their housing and neighborhood. Since most households do not sell their property every year, it is not possible to directly observe the sale price, and hence the quality of a property, on a yearly basis. In order obtain a time-varying measure of housing quality, the estimated market value (EMV) of each property was instead used as a starting point in this paper.

The Minneapolis Tax Assessor’s office is tasked with determining the market value of all taxable properties in Minneapolis every year. The main goal is to establish the taxable basis for each property in order to appropriately apportion property taxes. Several methods are employed in order to calculate a property’s EMV, which should correspond to the expected sale price of a property if placed on the open market. The most common of these methods is the “Sales Comparison Approach”, whereby a computerized appraisal program uses the sale prices of similar properties in the
Table 1.1: Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>St Dev</th>
<th>Min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sale Price</td>
<td>219,773</td>
<td>112,239</td>
<td>409</td>
<td>150,000</td>
<td>199,900</td>
<td>261,900</td>
<td>4,146,358</td>
<td>49,373</td>
</tr>
<tr>
<td>EMV</td>
<td>212,555</td>
<td>100,217</td>
<td>35,000</td>
<td>149,000</td>
<td>190,900</td>
<td>250,000</td>
<td>1,090,000</td>
<td>738,149</td>
</tr>
<tr>
<td>Year Built</td>
<td>1,930</td>
<td>18.839</td>
<td>1,900</td>
<td>1,916</td>
<td>1,926</td>
<td>1,946</td>
<td>2,017</td>
<td>738,084</td>
</tr>
<tr>
<td>Area</td>
<td>2,172</td>
<td>555</td>
<td>100</td>
<td>1,804</td>
<td>2,102</td>
<td>2,163</td>
<td>4,979</td>
<td>738,149</td>
</tr>
<tr>
<td>Baths</td>
<td>1.698</td>
<td>0.737</td>
<td>1.000</td>
<td>1.000</td>
<td>2.000</td>
<td>2.000</td>
<td>6.000</td>
<td>737,378</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>2.930</td>
<td>0.842</td>
<td>0.000</td>
<td>2.000</td>
<td>3.000</td>
<td>3.000</td>
<td>9.000</td>
<td>707,522</td>
</tr>
<tr>
<td>Stories</td>
<td>1.450</td>
<td>0.394</td>
<td>1.000</td>
<td>1.200</td>
<td>1.200</td>
<td>1.700</td>
<td>3.000</td>
<td>738,130</td>
</tr>
<tr>
<td>Sale</td>
<td>0.067</td>
<td>0.250</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>738,149</td>
</tr>
<tr>
<td>Any Investment</td>
<td>0.102</td>
<td>0.303</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>738,149</td>
</tr>
<tr>
<td>Total Investment</td>
<td>1477.355</td>
<td>10,256.99</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1,049,273</td>
<td>738,149</td>
</tr>
</tbody>
</table>

Note: The summary statistics reported here are based on the final sample used for estimation, and not for larger sample encompassing every property in Minneapolis. Properties are included in the sample only if they are single family houses registered for residential use and if they have been sold at least once between 1983 and 2016.

In order to find comparable sales, the Assessor’s office maintains meticulous records on each property’s sale history and physical characteristics. In addition, all building permits are tracked in order to account for any work done on the property between visits. Table 1.1 reports the summary statistics for the key variable in the sample. This process is repeated on an annual basis, as the Assessor is required to update all properties’ EMV in order to make sure they remain in line with market conditions and to account for changes to the property that may have occurred in the preceding year. The Assessor periodically visits each property to verify or gather new information, and to update the assessment in response to any changes to the property.

The dataset thus includes the EMV, sale prices, structure characteristics, and all building permits released to any property in the City of Minneapolis between 1999 and 2016. Sale records trace back to 1983, making it possible to establish how long each household has been living in a given property. Building permits are particularly informative because they not only include a description of the type of construction
Table 1.2: Common Building Permits

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roof</td>
<td>24,036</td>
<td>31.86</td>
</tr>
<tr>
<td>Window</td>
<td>16,535</td>
<td>21.92</td>
</tr>
<tr>
<td>Garage</td>
<td>14,792</td>
<td>19.61</td>
</tr>
<tr>
<td>Basement</td>
<td>10,478</td>
<td>13.89</td>
</tr>
<tr>
<td>Remodel</td>
<td>9,196</td>
<td>12.19</td>
</tr>
<tr>
<td>Bathroom</td>
<td>9,166</td>
<td>12.15</td>
</tr>
<tr>
<td>Repair</td>
<td>7,107</td>
<td>9.42</td>
</tr>
<tr>
<td>Insulation</td>
<td>7,097</td>
<td>9.41</td>
</tr>
<tr>
<td>Deck</td>
<td>7,006</td>
<td>9.29</td>
</tr>
<tr>
<td>Siding</td>
<td>5,487</td>
<td>7.27</td>
</tr>
<tr>
<td>Kitchen</td>
<td>5,178</td>
<td>6.86</td>
</tr>
<tr>
<td>Door</td>
<td>5,077</td>
<td>6.73</td>
</tr>
<tr>
<td>Addition</td>
<td>2,251</td>
<td>2.98</td>
</tr>
</tbody>
</table>

Note: This table was constructed by using text analysis on the declared scope of each building permit to assess the categories most commonly involved in a construction project. Categories are not exclusive.

project to be completed (see Table 1.2 for the most common categories), but also an estimate of the value of the proposed construction project. The latter variable can be used as a measure of the investment level in each property in a given year.

Since the focus of this paper is home improvement, new construction projects have been excluded from the sample. The sample is further restricted to single family homes that were sold at least once between 1983 and 2016, in order to accurately measure investment and tenure length of each household. This leaves a sample of 42,522 individual properties. While just over 10.2% of properties apply for a building permit in a given year, most properties (83.4%) apply for a permit at least once over the sample period, for a total of more than 97,000 separate building permits. The average value of a construction project for households that apply for a construction permit is of $14,462.
Figure 1.1: Average Estimated Market Value and Average Sale Price by Year

Note: This figure illustrates the average estimated market value (EMV) and the average sale prices for properties that was sold between 1999 and 2016.

Since it is not possible to observe the sale of each property in each period, the EMV is used as the basis to infer the quality of each property. The Minnesota Department of Revenue guidelines indicate that the EMV should fall between 90 and 105 percent of a property’s market value. Figure 1.1 illustrates the relationship between EMV and sale prices over the sample period.

In this analysis, housing quality is thought of in relative terms, that is, as the value of a property compared to that of its neighbors. Thus, the quality of property $i$ at time $t$, $q_{it}$, is measured as an index that compares property $i$’s EMV to the average EMV.

\footnote{This is because home improvement in condominiums and multi-family apartment buildings is often be carried out by entities separate from the household themselves, such as HOAs or building managers. Thus for these types of units it is often not possible to accurately measure individual households’ investment in home improvement.}
Table 1.3: Summary Statistics for State Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St Dev</th>
<th>Min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>209,070</td>
<td>85,704</td>
<td>65,000</td>
<td>157,000</td>
<td>189,750</td>
<td>246,000</td>
<td>597,250</td>
<td>738,149</td>
</tr>
<tr>
<td>q</td>
<td>1.015</td>
<td>.193</td>
<td>.534</td>
<td>.887</td>
<td>1.000</td>
<td>1.119</td>
<td>1.886</td>
<td>738,149</td>
</tr>
</tbody>
</table>

in the neighborhood where property \(i\) is located. Assuming property \(i\) is located in neighborhood \(j\), the EMV for the neighborhood, \(P_{jt}\), is calculated as the median EMV in a given Census block group in a given year.\(^8\)

Thus:

\[
EMV_{ijt} = q_{it} \cdot P_{jt}
\]  

(1.1)

Since both the EMV of the property and the median neighborhood market value \(P\) are observable, housing quality can then be calculated as an index \(q = EMV/P\). Table 1.3 reports summary statistics for \(q\) and \(P\), the state variables of the dynamic model.

1.4.2 Descriptive Facts

When examining data on investment in home improvement and housing renovation, a few salient features stand out. First, as illustrated in Figure 1.2, investment tends to be more likely to occur in homes that have a higher score on the home quality index and in homes that are located in neighborhoods with a higher median housing value. This might be as a result of selection - households that value higher levels of housing consumption choose to live in higher quality houses and in higher quality neighborhoods. These data features might also be driven in part by an income

\(^8\)Please see Appendix A.2 for a map illustrating the granularity of Census Block Groups within the City of Minneapolis.
Figure 1.2: Investment Probability, Sale Probability and Average Investment by $q$ and $P$

Note: The figures in the top row illustrate the empirical investment probability, i.e. the likelihood of observing a household choosing a positive level of investment in the data, as a function of the housing quality index $q$ (left), and of the median neighborhood price $P$ (right). The figures in the middle row illustrate the empirical sale probability, as a function of the housing quality index $q$ (left), and of the median neighborhood price $P$ (right). The figures in the bottom row illustrate the average investment amount conditional on deciding to invest as a function of the housing quality index $q$ (left) and of the median neighborhood price $P$ (right). Data have been graphed by applying a generalized additive model using cubic regression splines for smoothing. The shaded region around each curve plots standard errors.
Figure 1.3: Investment Frequency by Investment Size

Note: This figure illustrates the distribution of investment for households at the bottom and top quartiles of housing quality $q$ (left) and for households at the bottom and top quartiles of neighborhood quality $P$ (right).

... effect: households living in more expensive houses and in more expensive neighborhoods also tend to have a higher income, and can in turn, afford to improve their property more often.\footnote{The relationship between income and likelihood of home improvement is in line with empirical findings in the literature, such as Mendelson (1977), Plaut and Plaut (2010). Please see Table A.2 in Appendix A.3 for some illustrative evidence of the relationship between income and a neighborhood’s median EMV. Appendix A.3 examines more closely the role played by income.} Furthermore, it is important to note that the fact that households living in high quality homes are also more likely to invest is to some extent an endogenous aspect of this process, as households that are more likely to invest in home improvement will end up with houses that have better quality relative to that of their neighbors. On the other hand, the probability of a property being sold modestly declines as a function of housing and neighborhood quality.

Second, neighborhood quality is correlated with the average investment size among households that decide to invest in home improvement. Figure 1.2 shows that average investment tends to increase as a neighborhood’s median housing value and a property’s relative quality increases. In addition to the selection and income effects discussed above, this characteristic can also be partly attributed to the fact that...
when a house is more expensive, a larger investment might be required in order to
have a substantial impact on the property’s value, while smaller investments might
go further to improve the relative value of a cheaper property.

The relationship between housing quality and the average investment size is simi-
larly increasing, with higher levels of investment in houses at the top of the quality
distribution. It is possible that houses at the top of the housing quality distribution
are likely to display a higher level of average investment because of the same end-
dogenous process described above, whereby household with a higher propensity to
invest are most likely to live in high quality housing. Figure 1.3 shows that, while
most households that decide to invest in home improvement choose to invest less
than $10,000, households that are at the top of the distribution of housing quality
and neighborhood quality are more likely to undertake large projects.

Third, as illustrated in Figure 1.4, a household’s likelihood to invest in home im-
provement tends to decline as the number of years since the property was purchased
increases. This might partly due to the fact that making improvements to a prop-
erty soon after purchase allows households to enjoy higher housing consumption for
a longer time frame. This data feature underpins the need for a dynamic model in
order to accurately account for the evolution of household incentives over time.

Figure 1.5 highlights the importance of accounting for the interaction of housing
quality $q$ and of neighborhood prices $P$, as they tend to reinforce each other. The
likelihood of investment, as well as average investment, are highest in neighborhoods
with relatively high $q$ and $P$. For instance, houses at the bottom of the housing
quality distribution in the most expensive neighborhoods are as likely to invest as
houses at the top of the quality distribution in neighborhoods below the median
Note: This figure illustrates the empirical investment probability conditional on tenure length, i.e. the number of years since the property was purchased. Data have been graphed by applying a generalized additive model using cubic regression splines for smoothing. The shaded region around each curve plots standard errors.

These characteristics suggest building a dynamic model in which the likelihood of investing in home improvement and the average amount invested depend on both housing quality and neighborhood values. Section 1.5 introduces a model that attempts to capture these features.

### 1.4.3 Assumptions for Causal Inference

The key relationships underpinning the structural model are those between the state variables, housing quality and neighborhood quality, and the outcomes of interest,
Figure 1.5: Joint Distribution of Investment Probability and Average Investment

Note: The top panel of this figure illustrates the empirical investment probability, i.e. the likelihood of observing a household choosing a positive level of investment in the data, as a function of both the housing quality index \( q \) and of the median neighborhood price \( P \). Similarly, the bottom panel of the figure illustrates the average investment amount conditional on deciding to invest as a function of both the housing quality index \( q \) and of the median neighborhood price \( P \).
such as the likelihood and magnitude of investment in home improvement. Under an ideal experiment scenario, it would be possible to observe exogenous changes in \( q \) and \( P \) in order to establish their impact on home improvement. For example, housing values could be subject to an exogenous shock due to extreme weather conditions. To my knowledge however, no such shock has occurred on a sufficiently large scale in Minneapolis within the sample period. Moreover, the impact of such a shock might not be generalizable to other settings. Several assumptions are thus required in order to interpret changes in investment observed in the data as being caused by changes in housing quality or neighborhood quality.

In order to interpret changes in \( q \) and \( P \) as having a causal impact on investment, it is necessary to assume there are no omitted factors systematically correlated with the state variables that also have an impact on investment decisions. An example of potentially omitted factors in this application is that of heterogeneous preferences towards home improvement due to differing income levels. If micro-data on household income was available, it would be possible to exploit income shocks, for example shocks due to job changes or job losses, to isolate the role played by income separately from that of housing quality and neighborhood quality. Due to the lack of data availability, Appendix A.3 explores an extension of the baseline model of investment in home improvement to account for the effects of income by replicating the distribution of income in each neighborhood.

Measurement error is another confounding factor that needs to be ruled out in order to establish causal inference. The main difficulties measuring housing quality and neighborhood quality are outlined in Section 1.4.1. The accuracy of both quality measures ultimately relies on the accuracy of the EMV. As illustrated in Figure 1.1, EMV is closely correlated with sale prices but it is not a perfect predictor. Further-
more, while the data on construction permits used in this paper is, to my knowledge, the most extensive and complete dataset available on home improvement, construction permits do not perfectly capture all instances of home improvement. Smaller home improvement projects, for example painting, papering, tiling, etc., do not require a permit even though they contribute to the level of housing quality. Even in instances where a permit is required, it is likely that not all homeowners will obtain a construction permit, since applying for a permit requires paying a fee to the City of Minneapolis as well as at least one visit by an inspector. Lastly, the investment amount reported in the permit is likely to be an estimate of the total cost of the project rather than the actual cost.

In the absence of more complete data on home improvement, in order to obtain causal identification it is necessary to assume that the measurement error arising because of these issues is negligible compared to the overall level of investment in home improvement. In the case of home improvement work that does not require a permit, this assumption can be partially justified because these type of projects tend to be small in scale and are more likely to be a reflection of households preferences for decorating rather than having an objective impact on the level of housing quality. Projects that are required to be permitted are usually larger in scale, and often have to be legally completed by licensed contractors. Licensed contractors have stronger incentives to ensure that home improvement work is appropriately licensed by the City, since they risk losing their license otherwise. The incentives not to report the fair value of the construction project are mitigated by the fact that the City inspects all construction plans before approval and is expected to inspect the project one or more times.

Lastly, simultaneous causality has to be ruled out in order to ensure that changes
in $q$ and $P$ can be interpreted as a causal impact on investment. The structure of the data and the way in which property assessments are carried out are helpful in this regard. The assessed EMV for year $t$ is based on market characteristics in the previous year and finalized by the end of year $t - 1$ and thus are not affected by investment in home improvement in year $t$. While investment in year $t$ can affect housing quality in later periods, this channel is explicitly modeled in the structural model presented in the next section.

1.5 Model

Most decisions related to housing, and in particular any decision related to home improvement and housing renovation, are inherently dynamic in nature. As with any type of investment, investing in home improvement means foregoing current consumption in order to increase future consumption. Thus, it is crucial to examine the determinants of investment in home improvement within a dynamic framework. As recognized by Montgomery (1992) home improvement is only one of the avenues available to households who wish to re-optimize their consumption of the housing good. Selling the property and moving to a new dwelling is an option always available to households, and it would therefore be spurious to exclude this possibility from a model looking at housing investment.

Thus, in this model, household decisions are modeled in a dynamic setting with a finite number of time periods ($T = 35$). In each time period, households face a discrete choice: they can choose to do nothing and continue holding their property, enjoying the level of housing consumption it provides; they can invest in home improvement to increase their future housing quality; they can decide to sell their property and
move. When a household chooses to invest in home improvement, it will also have to choose the optimal investment amount. Thus each household faces a continuous decision nested within a discrete choice framework.

In each period, households observe the state of the world, summarized by the model’s state variables, and decide what action to undertake. Households make their decision based on quality of the neighborhood in which their property is located and the quality of their own property. Let \( d \) index the discrete choice options that the household can undertake in each time period:

\[
d = \begin{cases} 
1, & \text{do nothing;} \\
2, & \text{invest, choose optimal investment amount;} \\
3, & \text{sell.}
\end{cases}
\]

If a household decides to do nothing or to invest in the property, the same options will be available in the successive time period. However, if a household decides to sell the property, the household enters an absorbing state, and it will no longer have to make any choice in subsequent periods.

1.5.1 State Variables

The state variables influencing household decisions capture the key components of the housing good: housing quality and the average neighborhood quality. This modeling choice builds on Bayer et al. (2016) who develop a dynamic model of housing choice that includes both household-specific and neighborhood-specific characteris-
tics in the vector of observed state variables. Further empirical support is found in Montgomery (1992), Helms (2003) and Charles (2013), who identify building and neighborhood characteristics as key determinants of urban housing renovation.

At the beginning of each period, households observe the average housing quality in their neighborhood, $P$, and the housing quality of their own property $q$. Households have the option to invest in their property to increase their future housing quality but assume that they cannot directly influence their neighborhood quality. The evolution of the average neighborhood price $P$ is thus modeled as an AR(1) process independent of the household’s decisions:

$$P' = \rho_0 + \rho_1 P + \zeta$$

(1.2)

where the average neighborhood quality in the next time period, $P'$, depends on the current neighborhood quality, $P$, and on a randomly distributed shock, $\zeta$.

Housing quality is a relative measure, calculated with respect to the average level of housing quality in the neighborhood, similarly to Montgomery (1992). Thus, if the EMV captures the overall value of a house, it can be decomposed in a portion that depends on the neighborhood quality, $P$, and in an housing quality index, $q$, that captures the property’s quality level compared to the neighborhood average:

$$EMV = q \cdot P$$

(1.3)

Households are assumed to be atomistic – they do not expect that their actions will affect the average neighborhood property value, but they may decide to invest in their property in order to increase the future quality of their own housing. Future housing quality, $q'$, depends on current housing quality, $q$, and on how much households decide to invest in home improvement for their property. Let $i \in [0, \bar{i}]$ denote the
amount of investment undertaken in the current time period. Then:

\[ q' = \delta_1 q + \delta_2 i + \eta \]  \hspace{1cm} (1.4)

where \( \delta_1 \) captures the natural evolution of housing quality in the absence of any investment in the upkeep of the property, \( \delta_2 \) captures the marginal impact of investment on housing quality, and \( \eta \) is a normally distributed shock. Since the values of \( q' \) and \( P' \) are related, the shocks \( \zeta \) and \( \eta \) are assumed to be jointly distributed.

### 1.5.2 Utility

Households derive utility from living in a better house, either because it is located in a neighborhood with high housing quality, or because the house itself is of high quality. If the current housing quality is sub-optimal, investing in housing can increase a household’s future utility; this however comes at the cost of reducing its disposable income in the current period. Alternatively, the household can sell the property and use the sale proceeds to move to a new property. Hence, the flow utility for each household depends on the discrete decision, \( d \), the each household undertakes at the beginning of each period.

Households that decide to hold the property without doing any home improvement work in the current period do not face any investment cost. Their utility function is modeled as a Cobb-Douglas function that depends on the two dimensions of housing consumption, the current level of housing quality, \( q \), and the current level of neighborhood quality, \( P \):

\[ u(q, P \mid d = 1) = q^{\theta_1} P^{\theta_2} \]  \hspace{1cm} (1.5)
Notice that this functional form implies that housing quality and neighborhood quality are substitutes: households might be willing to trade a lower level of housing quality in exchange for a better neighborhood and vice versa.

The utility function of a household that chooses to invest in home improvement models the payoff from housing consumption analogously, except that the household must now face a disutility arising from the cost of investment. The cost of investment is modeled under an additive separability assumption, as is standard in the housing demand literature.\footnote{See for example Bayer et al. (2016).} In addition, a household choosing to invest incurs a fixed cost of investing, $FC_2$. This fixed cost can be thought of as a term capturing the psychological cost representing the potential disruption from remodeling, the value of time spent finding contractors, etc.

Hence, the utility function for households that choose to make an investment of size $i$ in housing renovation ($d = 2$) is given by:

$$
\begin{align*}
\begin{align*}
 u(q, P \mid d = 2) &= q^{\theta_1} P^{\theta_2} - \theta_3 (i)^{\theta_4} - FC_2 
\end{align*}
\end{align*}
\tag{1.6}
$$

In this baseline model, $i$ capture the entirety of the monetary cost of investment borne by the household. Any policy aimed at encouraging investment in home improvement by either directly or indirectly reducing the monetary cost of home improvement can be incorporated by introducing a $(1 - \tau)i$ term to mitigate the direct impact of investment on current utility.

Households that choose to sell their property ($d = 3$) can cash out on the value of their home and use the sale proceeds to move elsewhere, and thus derive utility from
their property characteristics at the time of sale:

\[ u(q, P|d = 3) = q^{\theta_1} P^{\theta_2} - \theta_5 (q \cdot P) - FC_3 \quad (1.7) \]

where \( \theta_5 (q \cdot P) \) captures the disutility generated by the costs of selling that are proportional to the housing value, such as realtor fees. Similarly to the previous case, households selling a property also face a fixed cost of selling, captured by the term \( FC_3 \), that can be interpreted as the psychological (or pecuniary) cost of moving to a new property in addition to the proportional component captured by \( \theta_5 \).

### 1.5.3 Value Function

Let \( V_t(q, P) \) be the expected discounted lifetime utility of owning a property of quality \( q \) in a neighborhood of quality \( P \). The household’s discrete choice problem can be expressed as:

\[ V_t(q, P) = \max \{ v_t(q, P, d) + \sigma \epsilon(d) : d = 1, 2, 3 \} \quad (1.8) \]

where \( v_t(q, P, d) \) is the choice specific value function, and \( d \) is the index that denotes each discrete decision. The error term \( \sigma \epsilon(d) \) is a random, choice-specific, taste shifter that is additively separable, i.i.d. and follows a Type I extreme value distribution with scale \( \sigma \). In addition to making the estimation of this model tractable by smoothing the primary and secondary kinks created in the value and policy functions by the discrete choice, the introduction of these idiosyncratic shocks has the advantage of yielding closed-form multinomial logit formulas for the discrete choice probabilities (Iskhakov et al., 2017).

The choice-specific value function conditional on not selling the property \((d = 1, 2)\) depends on the flow utility in the current time period and on the expected continuation value from holding the property in the next period. The continuation value

\[^{11}\text{An alternative specification for } u(q, P|d = 3) \text{ is explored in Appendix A.4}\]
is the discounted sum of the expected utility generated by the property over the remaining lifetime of the household, hence its dependence on $t$, accounting for the potential shocks to neighborhood quality and housing quality in the following periods.

The choice-specific value function for households that do nothing ($d = 1$) is thus given by:

$$v_t(q, P, d = 1) = q^{\theta_1} P^{\theta_2} + \beta \int EV_{t+1}^\sigma(q', P') f(d\eta, d\zeta)$$  \hspace{1cm} (1.9)

Households that choose to invest in home improvement in the current period ($d = 2$) need to choose the investment amount $i$ that maximizes their expected lifetime utility:

$$v_t(q, P, d = 2) = \max_{0 \leq i \leq \bar{i}} \left\{ q^{\theta_1} P^{\theta_2} - \theta_3 i^{\theta_4} - FC_2 + \beta \int EV_{t+1}^\sigma(q', P') f(d\eta, d\zeta) \right\}$$  \hspace{1cm} (1.10)

where the investment choice $i$ affects the evolution of housing quality in the next period, $q'$ through the process illustrated in Eq. (1.4).

Unlike households that choose to continue holding their property, households that decide to sell effectively leave the sample, since selling is an absorbing state:

$$v_t(q, P, d = 3) = q^{\theta_1} P^{\theta_2} - \theta_5 (q \cdot P) - FC_3 + \beta W_{t+1}(q, P)$$  \hspace{1cm} (1.11)

Thus, the value function after the household decides to sell the property can be thought of as a sum of discounted future payoffs, holding the value of the house, and hence the level of $q$ and $P$, fixed to the current period. This functional form can be interpreted as households receiving a payoff from the sale of their property that is proportional to the utility that can be generated by the existing property features – $q$ and $P$ – at the time of sale, since all households are assumed to value properties
based on their characteristics as summarized by \( q \) and \( P \). The choice specific value function then becomes:

\[
W_t(q, P) = q^{\theta_1}P^{\theta_2} - \theta_5(q \cdot P) - FC_3 + \beta W_{t+1}(q, P)
\]  

(1.12)

Households that decide to invest in home improvement then select the optimal investment amount given the investment cost and the potential for improvement in future housing quality. This choice implies an Euler equation of the following form:

\[
u_\iota(q, P | d = 2) + \beta \int u_i(q', P' | d' = 1)P(d' = 1 | q', P') + u_i(q, P' | d' = 2)P(d' = 2 | q', P')
+ u_i(q', P' | d' = 3)P(d' = 3 | q', P') f(d\eta, d\zeta) = 0
\]

(1.13)

where \( u_i(q, P | d) \) is the marginal utility with respect to investment in the current period and \( u_i(q', P' | d') \) is the marginal utility with respect to investment in the previous period. In this case:

- \( u_i(q, P | d = 2) = -\theta_5 q^{\theta_4 - 1} \)
- \( u_i(q', P' | d' = 1) = u_i(q', P' | d' = 2) = \delta_2 h^{\theta_1 - 1}P^{\theta_2} \)
- \( u_i(q', P' | d' = 3) = \delta_2 h^{\theta_1 - 1}P^{\theta_2} - \delta_5 L \)

Since the choice-specific taste shifter \( \epsilon(d) \) is an independently distributed random variable following the Type I extreme value distribution, the expected value function can be calculated using the log-sum formula first introduced by McFadden et al. (1973):

\[
EV_{t+1}^{\sigma_\epsilon}(q, P) = E \left[ \max \{ v_{t+1}(q, P, d) + \sigma_\epsilon \epsilon(d) \}_{d=1,2,3} \right]
\]

\[
= \sigma_\epsilon \log \left( \sum_{d=1}^{3} \exp (v_{t+1}(q, P, d)/\sigma_\epsilon(d)) \right)
\]

(1.14)

This distributional assumption also yields the following logit choice probabilities for
each of the discrete choice faced by the household:

\[ P_t(d \mid q,P) = \frac{\exp \left( \frac{v_t(q,P,d)}{\sigma} \right)}{\sum_{d=1}^{3} \exp \left( \frac{v_t(q,P,d)}{\sigma} \right)} \]  

\[ (1.15) \]

The model can be solved by backwards induction starting in the terminal period \( T \). Once the choice specific value functions \( v_T(q,P,d) \) are calculated, it is possible to recursively define a sequence of value functions \( \{V_t\} \) and Euler equations for the previous periods. This yields a policy function for investment \( \hat{i}_t(q,P) \) and a discrete choice policy function defined by the choice probabilities \( P_t(d \mid q,P) \) for each time period \( t \in 1,\ldots,T \).

### 1.5.4 Model Limitations

In an ideal scenario, it would be possible to predict the impact of a policy aimed at promoting home improvement through an investment subsidy by setting up an experiment in which households are randomly selected to receive such a policy. In lieu of such an experiment, exogenous variation in the cost of investment could be exploited in a quasi-experimental setting. However, the sources of variation in the cost of investment for which data is available, such as changes in interest rates, tend to be correlated with other factors, such as changing housing prices and income shocks, that make causal inference difficult.

In order to examine policy counterfactuals under the structural framework presented above, household decisions are assumed to be primarily determined by the value of the state variables \( q \) and \( P \), a household’s age, and on the parameters regulating the utility functions and cost function. It is of course possible that there exist some unobserved sources of heterogeneity that have an impact on household decisions to
invest or sell their property beyond what is specified above.

While the model includes a random, choice-specific taste shifter $\sigma \epsilon(d)$ that aims to capture any unobservable factors having an impact on the household’s discrete choices, this shock is assumed to be i.i.d., and thus independent of past choices of the same household. Hence, the model as it stands cannot rationalize household tastes that are persistent over time, such as whether some households are more inclined towards home improvement than others. Thus, a necessary assumption underlying this model is that there are no unobserved factors having an impact on household decisions that are persistent over time.

If household tastes toward home improvement vary systematically because of some omitted, but measurable, factor, the model could be expanded to account for these variables. For example, Appendix A.3 explores how the model can be augmented to account for differences in household’s income levels. In this case, the main obstacle to specifying a richer model is the absence of suitable micro-data encompassing both household’s decisions regarding home improvement as well as other characteristics, such as detailed demographic information.

In the baseline model presented here, the role of the state variables $q$ and $P$ an household’s decisions should be thus interpreted with caution. In the presence of persistent taste differences, households with stronger preferences toward housing consumption would endogenously select into properties with higher levels of $q$ and $P$. Then, model estimates would not only pick up the direct impact of these variables, but $q$ and $P$ would effectively function as proxies, partially capturing the effect of these unobservable taste differences between households. In this case, the model might be able to accurately predict households’ propensity to invest based on observed levels of $q$.
and $P$, and the one-off effects of a policy aimed at promoting home improvement, such as that explored in Section [1.7] would be accurately estimated. However, the dynamic effects of such a policy, examined in Appendix [1.8], would be less reliable, as the model’s predictions of how a household would respond to changing levels of neighborhood quality, $P$, might be inaccurate.

1.6 Estimation

The key model parameters are estimated in two steps. The first step estimates the static parameters regulating the evolution of the states using a linear regression model. The second step estimates the parameters indexing the utility functions and cost functions of the model using maximum likelihood estimation.

1.6.1 Static Parameters

The two key states in this model, neighborhood quality $P$ and housing quality $q$, are approximated from the data based on each property’s EMV compared to the median neighborhood EMV, following the relationship detailed in Eq. (1.1). The median neighborhood value $P$ and housing quality $q$, are assumed to evolve in a static setting. That is, their value in the current period is thought of as depending exclusively on what happened in the previous period and on a random shock drawn from a multivariate normal distribution.

The evolution of the median neighborhood value, $P_t$, is modeled as an AR(1) process.
Table 1.4: Parameter Estimates for the Evolution of $P$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{t+1}$</td>
<td>$0.986^{***}$</td>
<td>(4587.74)</td>
</tr>
<tr>
<td>$P_t$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>$1.109^{***}$</td>
<td>(231.84)</td>
</tr>
<tr>
<td>$N$</td>
<td>701,899</td>
<td></td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Variable $P$ measured in $\$10,000.$

with parameters $\rho_0$ and $\rho_1$ regulating its evolution:

$$P_{t+1} = \rho_0 + \rho_1 P_t + \zeta_t \quad (1.16)$$

Analogously, future housing quality, $q_{t+1}$, depends on current housing quality, $q_t$ and on how much households decide to invest in their property in the current period. Let $i_t \in [0, \bar{i}]$ denote the amount of investment, measured in dollars, undertaken in a given year in each property. Then:

$$q_{t+1} = \delta_1 q_t + \delta_2 i_t + \eta_t \quad (1.17)$$

where $\delta_1$ is a parameter to be estimated capturing the relationship between current and past housing quality, $\delta_2$ captures the impact of investment on housing quality.

The random shocks $\zeta$ and $\eta$ follow the multivariate normal distribution:

$$\begin{bmatrix} \zeta \\ \eta \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$
Table 1.5: Parameter Estimates for the Evolution of $q$

<table>
<thead>
<tr>
<th></th>
<th>$q_{t+1}$</th>
<th>$q_t$</th>
<th>0.998***</th>
<th>(17788.17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_t$</td>
<td>0.0086***</td>
<td>(147.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>702,087</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: Variable $i$ measured in $\$10,000$.

The parameters $\rho_0, \rho_1$ regulating the neighborhood evolution process can be consistently estimated by OLS regression. The results of this estimation are reported in Table 1.4. The parameters $\delta_1$ and $\delta_2$ regulating the process of evolution of housing quality can also be consistently estimated by OLS regression. Estimation results are reported in Table 1.5.

The variance-covariance matrix is computed from the residuals of these regressions, using the maximum likelihood estimator of $\Sigma$:

$$
\hat{\Sigma} = \begin{bmatrix}
\hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\
\hat{\Sigma}_{21} & \hat{\Sigma}_{22}
\end{bmatrix} = \begin{bmatrix}
2.292 & 0.002 \\
-0.002 & 0.002
\end{bmatrix}
$$

1.6.2 Structural Model Parameters

The parameters for the dynamic model can be estimated using a nested fixed point routine following Rust (1987). In general, nesting a continuous choice within a dis-
crete choice in a discrete-continuous setting can cause issues that can make solving
and estimating this class of models problematic. The optimal discrete choice for the
household is based on the comparison of each choice-specific value function; if the
optimal choice changes as a function of the state variables, the value function can
form a kink or become non-convex. Both these issues can affect the Euler equations
of the continuous choice, in turn causing nonlinearities in the policy function.
Moreover, the kinks can propagate to previous periods through the expected utility
function, compounding these issues over time. However, Iskhakov et al. (2017) show
that the introduction of a random i.i.d. Type I extreme value shock, as in Eq. (1.8),
has a smoothing effect on the value function, mitigating the kinks introduced by the
discrete choice and sidestepping the associated nonlinearities in the policy function.
Further, they show that the model’s structural parameters can be accurately estimated via maximum likelihood.

Let the data be a panel \( \{q_{it}^d, P_{it}^d, i_{it}^d, d_{it}^d\}_{i=1,...,N, t=1,...,T} \) containing observations on
housing quality, neighborhood quality, investment amount and household decisions.
Assuming the investment amount is measured with error, the discrepancy between
the observed investment \( i_{it}^d \) and that predicted by the policy function \( i(q_{it}^d, P_{it}^d|\theta, d_{it}^d = 2) \) is given by:

\[
i_{it}^d = i(q_{it}^d, P_{it}^d|\theta, d_{it}^d = 2) + \xi_{it}
\]

where \( \xi_{it} \) is assumed to be a random error term that follows the log-normal distribution variance \( \sigma^2_\xi \). Applying logs yields a linear approximation to the discrepancy
between observed and predicted investment:

\[
\log(\xi_{it}) = \log(i_{it}^d) - \log \left( i(q_{it}^d, P_{it}^d|\theta, d_{it}^d = 2) \right)
\]

where \( \log \xi_{it} \sim N(0, \sigma^2_\xi) \), i.i.d., \( \forall i, t \). This distribution has been chosen because
investment in home improvement cannot be negative and its distribution is right
tailed. Assuming that the choice specific shocks \( \epsilon(d) \) are independent of the discrepancy between predicted and observed investment, \( \log \xi_{it} \), the likelihood contribution of household \( i \) at time \( t \) is given by:

\[
\ell_{it}(\theta, \sigma_{\xi}) = P_t(d_{it}^d = 1 \mid \theta, q_{it}, P_{it}, s_{it}) I(d_{it}^d = 1) P_t(d_{it}^d = 2 \mid \theta, q_{it}, P_{it}, s_{it}) I(d_{it}^d = 2) P_t(d_{it}^d = 3 \mid \theta, q_{it}, P_{it}, s_{it}) I(d_{it}^d = 3) \left( \frac{\phi(\log \xi_{it}(\theta)/\sigma_{\xi})}{\sigma_{\xi}} \right) I(d_{it}^d = 2)
\]

The joint log likelihood function can then be written as \( \mathcal{L}(\theta, \sigma_{\xi}) = \sum_i \sum_t \log \ell_{it}(\theta, \sigma_{\xi}) \). Parameter estimates can then be obtained by maximum likelihood estimation. Results of this estimation routine are reported in Table 1.6.

The positive relationship between the likelihood of investment and housing and neighborhood quality, and that between average investment and housing and neighborhood quality help pinning down the parameters \( \theta_1 \) and \( \theta_2 \), regulating the impact of housing consumption in the utility function for each of the three discrete choices. The parameters regulating the disutility generated by the cost of investment are largely governed by the investment trends illustrated in Figure 1.3. Most households in the data tend to invest small amounts, and large investments are relatively rare, suggesting that the cost of investment generates some negative utility impact \( (\theta_3 > 0) \), and that there exist some nonlinearities \( (\theta_4 > 1) \) in the effect of investment on utility, whereby larger investments have an increasingly large impact.

The parameters capturing the fixed cost of investment, \( FC_2 \) and the fixed cost of selling \( FC_3 \) are pinned down by the relative difference between the likelihood of selling and investing compared to the decision to do nothing. The parameter \( \theta_5 \) is instead pinned down by the difference between the likelihood of selling and doing nothing as a function of the states \( q \) and \( P \).
Table 1.6: Parameter Estimates for Structural Model

<table>
<thead>
<tr>
<th></th>
<th>Parameter Estimates</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utility Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housing Quality</td>
<td>$\theta_1$</td>
<td>1.657</td>
</tr>
<tr>
<td>Neighborhood Quality</td>
<td>$\theta_2$</td>
<td>0.433</td>
</tr>
<tr>
<td><strong>Investment Cost Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Cost of Investment</td>
<td>$\theta_3$</td>
<td>0.041</td>
</tr>
<tr>
<td>Curvature on Investment</td>
<td>$\theta_4$</td>
<td>1.649</td>
</tr>
<tr>
<td>Fixed Cost of Investing</td>
<td>$FC_2$</td>
<td>78.889</td>
</tr>
<tr>
<td><strong>Sale Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Sale Cost</td>
<td>$\theta_5$</td>
<td>0.010</td>
</tr>
<tr>
<td>Fixed Cost of Selling</td>
<td>$FC_3$</td>
<td>-1.029</td>
</tr>
<tr>
<td><strong>Variances</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taste Shock</td>
<td>$\sigma_\epsilon$</td>
<td>37.370</td>
</tr>
<tr>
<td>Measurement Error</td>
<td>$\sigma_\xi$</td>
<td>1.164</td>
</tr>
</tbody>
</table>

A few interesting results emerge from this estimation. First, the model suggests there are increasing returns to consumption of the housing good. This is consistent with the empirical trends illustrated in Figure 1.2 and Figure 1.5, in which both the likelihood of investment and the average investment size increase as housing quality, $q$, and neighborhood prices, $P$, increase. While the investment cost parameters $\theta_3$ and $\theta_4$ have the expected magnitude and sign, the fixed cost of investment, $FC_2$, is large, although not precisely estimated. This might be due to the fact that households decide to invest relatively rarely – on average only 10.2% of households decide to invest in home improvement in a given year.

While the linear sale cost parameter $\theta_5$ indicates that there is some disutility of selling that is proportional to the value of a property, the sign on the fixed cost of selling, $FC_3$, is the opposite of what was expected. This might be a facet of the fact that people sell their properties somewhat more often than the model can rationalize. Neither of the parameters that capture sale costs are significant, which might
be due to the fact that households sell their property relatively infrequently in the data – in any given year, only 6.7% of households decide to sell their property on average.

1.7 Results

The next subsection illustrates the results of the structural model estimated using the City of Minneapolis data by comparing how the model performs vis-à-vis the key empirical trends in the data. The second subsection then performs a policy experiment examining a counterfactual scenario in which the MHFA Fix Up Loan Program is scaled up to be made available to a larger number of households.

1.7.1 Baseline Model

Using the parameters reported in Table 1.6, we can reconstruct the optimal investment decisions and the choice probabilities predicted by the model. Figure 1.6 illustrates the model’s prediction for the probability of investing in home improvement \(d = 2\), as a function of a property’s housing quality, \(q\), and neighborhood quality, \(P\).

Households with a higher level of housing quality than their neighbors and households that live in neighborhoods that on average have more expensive properties are more likely to invest. The two effects reinforce each other, with households that live in the highest quality houses in the highest quality neighborhoods being the most likely to invest. Conversely, households that live in the worst quality houses in the worst quality neighborhoods are the least likely to invest.
Figure 1.6: Predicted Investment Probability

Note: This figure illustrates the predicted probability of investing in home improvement as a function of housing quality, $q$, and neighborhood values, $P$. 
Figure 1.7: Predicted Optimal Investment

Note: This figure illustrates the optimal level of investment predicted by the model as a function of housing quality, \( q \) and neighborhood values, \( P \), conditional on deciding to invest.
While the difference in the predicted likelihood to invest is not large — varying between 10.2% and 11.3% — the optimal investment level changes significantly as a function of the state variables. Again, optimal investment tends to increase monotonically as a function of the two state variables, housing quality and neighborhood quality. As illustrated in Figure 1.7, the model predicts that the optimal investment for households with the highest housing quality and neighborhood quality is almost 15 times larger than that for households with the lowest housing quality in the least expensive neighborhoods — $22,000 and $1,500 respectively.

Figure 1.8 illustrates the model predictions vis-à-vis the empirical trends in the data by plotting how the model predictions change with respect to each state variable, while holding the omitted one at its median level. The model correctly predicts that the amount invested is the variable changing most significantly as housing quality and neighborhood values increase, with the optimal investment amount increasing as both \( q \) and \( P \) increase. As in its empirical counterpart, the model predicts that the likelihood of investment is increasing as a function of both housing values and neighborhood values, however the change is somewhat less pronounced than in the data. Lastly, the predicted likelihood of selling is decreasing in housing quality and slightly increasing with neighborhood values.

All in all, the model preforms remarkably well at reproducing the main trends observed in the data, with the larger discrepancies between predicted and empirical choices being at the lower end of the distribution of housing and neighborhood quality. This might be due to the influence that unobserved factors, such as income, can play on the optimal decision. Appendix A.3 explores refinements of the model that take into consideration income effects.
Note: This figure illustrates predicted discrete decision probabilities and the optimal investment policy compared to their empirical counterparts in the data. The top row depicts the decision to invest, the middle row depicts the decision to sell a property, and the bottom row depicts the investment amount, as a function of the housing quality index $q$ (left) and of the median neighborhood price $P$ (right). The empirical data have been graphed by applying a generalized additive model using cubic regression splines for smoothing. The shaded region around the curve plots standard errors. The values predicted by the model have been calculated as a weighted average with respect to the tenure length, and plotted holding the omitted dimension at its median value.
1.7.2 Policy Counterfactuals

In this section, I will examine how scaling up the MHFA Fix Up Loan Program would affect investment in home improvement and other outcomes of interest. For simplicity, I will first examine a scenario in which all households in Minneapolis are automatically enrolled to participate in the program. Then I examine how a modification of this policy might be implemented in order to better achieve specific policy goals, such as the revitalization of low income neighborhoods.

In the first counterfactual scenario, it is assumed that all households in the sample are able to save 10\% on the cost of their investment by financing a loan through a Fix Up Secured Loan rather than through commercial financing opportunities. This can be thought of as a reduction $\tau = 10\%$ in the cost of investment for the household. This counterfactual is based on a simplifying assumption, namely that all households investing in home improvement are financing their investment through borrowing. While this assumption is undoubtedly not tenable in practice, it is adequate to examine the key outcome of interest, that is, how would investment in home improvement change in response to a policy aimed at reducing the cost of investment directly borne by the household, be it through lending at a discounted rate to through a direct subsidy.

Figure 1.9 illustrates the model’s prediction for the counterfactual scenario in which all households in Minneapolis see their cost of investment reduced by 10\%, potentially through an expansion of the MHFA Program. The results show that the program would not significantly change the likelihood of investment but it would encourage larger amounts of investment in home improvement. The predicted likelihood of investment exhibits only a modest increase – the net change in the decision to invest
Figure 1.9: Change in Predicted Investment Probability and Average Investment

Note: The top panel of this figure illustrates the change in the optimal investment amount as a function of both the housing quality index \( q \) and of the median neighborhood price \( P \). The bottom panel of this figure illustrates the change in the predicted likelihood of investing as a function of both the housing quality index \( q \) and of the median neighborhood price \( P \). This counterfactual scenario assumes that every household received a 10% discount on the cost of investment through the MHFA Fix Up Loan Program.
is predicted to be between 0.01\% and 0.2\% – with the largest predicted increase for households at the top of the distribution of housing quality and neighborhood quality.

On the other hand, the proposed policy would have a considerable impact on the optimal amount of investment (Figure 1.9), with the optimal investment amount increasing significantly for all households in response to a decrease in investment cost. The increase is largest in absolute terms for households with the highest levels of $q$ and $P$, where the additional investment generated by this policy exceeds $5,000. For household with the lowest levels of housing quality and in the least valuable neighborhoods, optimal investment increases by around $1,000; while the predicted increase in the optimal level of investment is the lowest for these households, the policy is still predicted to generate a non-negligible relative change (+66.6\%) in the optimal investment level for household with the lowest levels of housing quality and in the least valuable neighborhoods.

The counterfactual results illustrated in Figure 1.9 suggest that an expansion of the MHFA Fix Up Loan Program extended to all households might effectively be a regressive policy. Households living in more expensive neighborhoods and higher quality housing are more likely to invest in home improvement in the first place, and thus would be more likely to receive a subsidy to finance their investment. Additionally, these households are the ones making the largest investments before the introduction of the policy and the ones that are predicted to experience the largest increase in optimal investment under the proposed policy framework, and thus would effectively receive a larger subsidy to offset the cost of their investment. If the objective of the social planner were to boost investment in home improvement in depressed neighborhoods (and indeed one of the key objectives of the Minnesota Housing Agency is to “strengthen disinvested communities”), then this policy is not
Figure 1.10: Change in Average Investment under Means Tested Policy

Note: This figure illustrates the change in the optimal investment amount as a function of both the housing quality index $q$ and the median neighborhood price $P$ under a policy in which only households earning less than $75,000 a year are eligible for the MHFA Fix Up Loan Program. Likely to achieve the desired outcome.

Figure 1.10 illustrates the predicted effect on optimal investment of a modified version of the MHFA Fix Up Loan Program available exclusively to households earning less than $75,000 per year.\textsuperscript{12} This calculation uses IRS data to estimate how many households would fall under this income threshold in each neighborhood and uses this probability to calculate the expected change in investment under this means-tested policy compared to the baseline model.\textsuperscript{13} The results of this exercise suggest that in such a scenario, households at the top of the distribution of housing quality would still see their optimal investment increase the most. However, differences across diff-

\textsuperscript{12}In its current version, the MHFA Fix Up Loan Program is already a means-tested policy, although the limit on household income is rather high at $141,000 for loans administered by the City of Minneapolis. However, this income limit would not substantially change the results reported in Figure 1.9 as only a relatively small percentage of households in Minneapolis earn more than $141,000 per year. See Table A.3 and Figure A.3.1 in Appendix A.3 for more details.

\textsuperscript{13}See Appendix A.3 for data description.
Figure 1.11: Model Predictions Under Alternative Policy Scenarios

Note: This figure illustrates predicted optimal investment policy (top row) and probability of selling (bottom row) under different policy scenarios. The blue line depicts the prediction for the baseline model, while the red lines depict the model prediction under a policy allowing households to reduce investment cost by 10%. Solid red lines represent the prediction if this policy is applied to all households, and dashed red lines represent the prediction for a means-tested policy, available only to households earning under $75,000 per year. Predictions are based on 2016 data.

Different levels of neighborhood quality would be substantially mitigated. These results should be interpreted with caution – the underlying assumption is that households with similar levels of housing quality and neighborhood quality would exhibit a similar propensity to invest, regardless of income differences. Appendix A.3 explores potential avenues to relax this assumption.

Figure 1.11 compares the model predictions under the alternative scenarios discussed.
above: the baseline model (no investment subsidy), a MHFA program in which all households participate (a $\tau = 10\%$ reduction in investment cost) and a means-tested program open to all households earning less than $75,000 per year (a $\tau = 10\%$ reduction in investment cost for qualifying households only). Both subsidy programs are predicted to be effective at increasing the expected level of investment in home improvement. The means-tested policy performs well at increasing the optimal amount of investment especially at the bottom of the distribution of neighborhood quality, where a larger percentage of households would be eligible for the subsidy.

Moreover, the model predicts that neither of these policies would significantly affect the likelihood of a household selling their property. This suggests that policies reducing the cost of investment might be an effective tool to encourage neighborhood revitalization without encouraging the displacement of current residents, and that means-testing these policies might be an effective tool to target the revitalization of poorer neighborhoods.

Table 1.7 summarizes the effect of these policies based on 2016 data for Minneapolis. The baseline model predicts $34.5$ million in aggregate investment in home improvement. Granting all households a $10\%$ subsidy on the cost of investment is predicted to increase the overall investment level by $10.7$ million at a cost of $4.5$ million for the government – thus investment would increase by $\$2.42$ for each dollar spent.\footnote{For comparison, Galster (1987) and Galster and Hesser (1988), also using Minneapolis data, estimate that each dollar spent on policies aimed at encouraging home rehabilitation resulted in an additional $\$0.34$-$1.62$ spent on home improvement.} 

A means-tested policy subsidizing $10\%$ of the cost of investing in home improvement, available only to households earning less than $75,000 per year is instead predicted to increase the overall level of investment by $\$7.5$ million while costing $\$2.9$ million,
Table 1.7: Total Investment Under Alternative Policy Scenarios

<table>
<thead>
<tr>
<th>Model</th>
<th>Total Investment</th>
<th>Total Cost</th>
<th>Change in Investment</th>
<th>Returns per Dollar</th>
</tr>
</thead>
<tbody>
<tr>
<td>No policy ($\tau = 0$)</td>
<td>$34,795,751</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$)</td>
<td>$45,930,716</td>
<td>$4,593,072</td>
<td>$11,134,965</td>
<td>$2.42</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$), Means-Tested</td>
<td>$42,286,104</td>
<td>$2,869,570</td>
<td>$7,490,353</td>
<td>$2.61</td>
</tr>
</tbody>
</table>

Note: This table was constructed by calculating and aggregating the predicted level of investment in home improvement for each household under alternative policy scenarios, using 2016 data for Minneapolis.

thus raising the impact of each dollar spent to $2.61. Section 1.8 explores the dynamic effects of this policy, while Appendices A.3 and A.4 examine the robustness of this finding to alternative model specifications.

1.8 Simulation of Evolution of Aggregate Investment and Capitalization

This section examines the dynamics of policies aimed at increasing investment in home improvement and their capitalization into housing values by simulating their evolution over time. As previously shown in Figure 1.8, both the likelihood of investment and optimal investment increase as neighborhood quality increases. Under a policy framework that successfully promotes investment in home improvement, neighborhood values can be expected to experience a more rapid appreciation, which in turn can make investment in home improvement in later periods more likely. This self-reinforcing process is the main source of dynamic effects in the model.

The starting point for the simulation outlined below is the assessed EMV – and hence $q$ and $P$ – for single family homes in Minneapolis in 2016. Each household
is assigned a probability of investing, a probability of selling and an optimal investment amount as predicted by the model under each policy framework, based on that household’s housing quality, neighborhood quality and tenure length. In each time period, a random number is drawn to determine which discrete action the household will undertake. Then, a log-normally distributed random number is drawn to simulate $\xi$, the multiplicative investment error term, in order to select an investment level for households deciding to invest.

The estimated market value of each property is then updated according to:

$$EMV' = \gamma_0 + \gamma_1 EMV + \gamma_2 i + \nu$$

(1.20)

where $\nu \sim N(0, \sigma_\nu)$. The parameters of this equation are estimated via OLS regression. Results of this estimation routine are reported in Table 1.8.

This modeling choice is based on the underlying assumption that households are atomistic in their actions. In the baseline model, this assumption means that households internalize the fact that their investment in home improvement can affect the evolution of their future housing quality $q'$, but not that of their future neighborhood value $P'$. In the simulation however, the actions of each household affect not only that household’s property, but are also aggregated at the neighborhood level. Thus, the total investment in each neighborhood affects the future value of that neighborhood through the evolution of each property’s EMV.

A normally distributed random number is drawn to simulate the evolution of the EMV to the next period, where all households age by one year, and properties that were sold in the previous period are assigned a tenure length of one. Neighborhood values are updated by recalculating the median EMV in each neighborhood, and
Table 1.8: Parameter Estimates for the Evolution of $EMV$

<table>
<thead>
<tr>
<th></th>
<th>$EMV_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$EMV_t$</td>
<td>0.990***</td>
</tr>
<tr>
<td></td>
<td>(4250.59)</td>
</tr>
<tr>
<td>$i_t$</td>
<td>0.313***</td>
</tr>
<tr>
<td></td>
<td>(130.36)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.002***</td>
</tr>
<tr>
<td></td>
<td>(186.91)</td>
</tr>
<tr>
<td>$N$</td>
<td>701,899</td>
</tr>
<tr>
<td>$\sigma_\nu$</td>
<td>1.906</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Variable $EMV$ measured in $\$10,000$.

housing quality indexes are updated according to the relationship $q = EMV/P$. Other factors, including the fraction of people earning less than $75,000$ per year, are assumed to remain constant over time. This process is repeated over five time periods, and the simulation is repeated for $N = 1,000$ iterations. The averaged results are reported below.

Figure [1.12] illustrates how aggregate investment in home improvement in Minneapolis would change over time under a policy subsidizing $\tau = 10\%$ of the investment cost compared to the baseline scenario with no policy ($\tau = 0$). The simulation shows that, in the first year of this policy being in place, the total amount of investment in home improvement in Minneapolis would increase by $21.9$ million if all households were to take advantage of this policy, and by $14.7$ million under a means tested policy targeted at households earning less than $75,000$ per year. The “excess” in aggregate investment is predicted to increase in each subsequent period by an average of 2.18% for the policy open to all households, and by 2.25% for the means
Figure 1.12: Change in Aggregate Investment under Alternative Policy Scenarios

Note: This figure reports the results of a simulation in which investment is subsidized ($\tau = 10\%$) for a period of five years. The column on the left reports results for a policy accessible by all households, and the column on the right reports the results for a means-tested policy for which only households earning less than $75,000 are eligible. The top panel illustrates the change in aggregate investment compared to a baseline scenario with no subsidy for home improvement ($\tau = 0$), while the bottom panel reports the predicted cost of the policy.

tested policy, compared to the baseline scenario with no investment subsidy. This can be interpreted as the spillover effect of the policy, that is, the increase in aggregate investment arising from the fact that housing quality and neighborhood quality are higher than in a scenario with no policy ($\tau = 0$).

\footnote{The predicted change in investment is higher in this simulation than in the results reported in Table 1.7. This is because the simulation accounts for the investment “error”, that is, for households investing an amount different from the optimum. The structure of this error is elucidated in Eq (1.18) and (1.19) and the error variance $\sigma_\xi$ is estimated with the rest of the structural model parameters, as reported in Table 1.6.}
Figure 1.13: Cumulative Change in Neighborhood EMV under Alternative Policy Scenarios

Note: This figure reports the results of a simulation in which investment is subsidized ($\tau = 10\%$) for a period of five years. The column on the left reports results for a policy accessible by all households, and the column on the right reports the results for a means-tested policy for which only households earning less than $75,000 are eligible.

The proposed policy is predicted to cost between $9.0 and $9.8 million per year if open to all households and between $5.6 and $6.1 million per year if means-tested, with a return per dollar spent ranging between $2.44 and $2.63.

Figures 1.13 and 1.14 illustrate the simulation results of how property values, captured by the EMV, are forecast to evolve under the proposed policy scenarios. Figure 1.13 reports results for two sample neighborhoods, Como (Block Group 3) and Lynnhurst (Block Group 2). These neighborhoods are illustrative of the dynamics at play in low-income and high-income neighborhoods. Como’s median property value starts out around $182,000 in the baseline year of the simulation, and 75.0% of households earn less than $75,000 per year. Lynnhurst is much wealthier, with its median property valued at $611,000 and with only 41.8% of households earning less than $75,000 per year. Both neighborhoods encompass a total of 222 households.
Figure 1.14: Cumulative Change in Aggregate EMV under Alternative Policy Scenarios

Note: This figure reports the results of a simulation in which investment is subsidized ($\tau = 10\%$) for a period of five years. The column on the left reports results for a policy accessible by all households, and the column on the right reports the results for a means-tested policy for which only households earning less than $75,000 are eligible. The top panel illustrates the cumulative change in aggregate property values (EMV) compared to a baseline scenario with no subsidy for home improvement ($\tau = 0$), while the bottom panel reports the cumulative predicted cost of the policy.

Under a policy environment where all households have access to a 10% investment subsidy, the neighborhood in Lynnhurst adds the largest total amount of housing equity to the neighborhood, over $292,000 after five years. The less wealthy neighborhood in Como, on the other hand, adds $163,000 to the aggregate neighborhood EMV over the same time period, compared to a scenario with no policy aimed at subsidizing investment. Under a policy framework where the investment subsidy is
targeted towards all households earning less than $75,000 per year, the two neighborhoods’ aggregate EMV is predicted to increase by a similar amount, between $137,000 and $161,000 after a five year period, compared to a baseline scenario with no policy in place. Thus, means-testing the policy could be an effective strategy to mitigate the differences in aggregate investment capitalization between the two neighborhoods.

Figure 1.14 reports analogous results for the City of Minneapolis as a whole. After five years, the aggregate EMV for Minneapolis is predicted to increase by $34.6 million under a 10% investment subsidy, compared to a baseline scenario with no subsidy. Under a means-tested policy, the total increase in housing values is predicted to be of $23.4 million. Interestingly, if the metric used to assess the success of the policy is capitalization into market values, rather than overall investment, the returns per dollar spent are lower than those shown in Table 1.7. As reported in Table 1.8, each dollar invested in home improvement increases the estimated market value by about $0.31. The cumulative cost of the investment subsidy policy after five years is close to $47.0 million and $29.4 million respectively. Thus, if measured by change in housing values rather than by the change in aggregate investment, the policy’s return can be quantified at $0.74 per dollar if accessible to all households, and at $0.80 per dollar when means-tested.

1.9 Conclusion

Investment in home improvement is a relatively understudied dimension of housing consumption that accounts for a substantial share of overall residential investment. This paper examines the decision to invest in home improvement in the context of a
dynamic structural model in which households can choose to reoptimize their housing consumption depending on their current level of housing quality and neighborhood quality. In each time period, households face a discrete choice, deciding whether or not they wish to do any home improvement work or to sell their property, and a continuous choice to select the optimal amount of investment.

The model is estimated using a rich micro-level dataset that encompasses each property in Minneapolis for a period of almost 20 years. Using the structural model parameters, I reconstruct the optimal investment policy and choice probabilities as a function of a property’s housing quality and neighborhood quality. Then, I explore a counterfactual scenario in which the cost of investment in home improvement is subsidized and show that such a policy would be effective at increasing the predicted level of investment. Results suggest that means-testing this policy could be a cost-effective tool to promote neighborhood revitalization, increasing predicted investment by $2.61 per dollar spent. Further, the results of a simulation using the structural model parameters suggest that this positive effect on investment is likely to be compounded over time.
Chapter 2

The Spillover Effects of Rental Property Development

2.1 Introduction

Professionally managed apartment buildings have become increasingly common in urban areas across the country over the last few years, changing the urban landscape and attracting an increasing share of the population, especially among the younger generations. While until 1990s properties specifically built for rental purposes tended smaller in scale, usually duplexes or triplexes, in the last two decades large, professionally managed apartment buildings have become the norm. In the aftermath of the Great Recession, the growth in the number of rental apartments as far outstripped that of any other type of residential property, and now represents the largest share of new unit construction in Minneapolis (Figure 2.1).

This paper contributes an empirical analysis of the net effect associated with new
apartment building developments on neighboring properties and how these effects can vary across different types of neighborhood. In particular, I am at assessing whether the positive effect on prices stemming from potential agglomeration benefits outweighs the negative impact on prices associated with an increase in the local supply for housing.

The construction of large-scale apartment building developments has obvious repercussions not only the rental property market, but can also have spillover effects the values of nearby properties, increasing urban density and altering the demographic composition of a neighborhood. On the one hand, if rental apartments and owner occupied units are substitutes, increasing the supply of available residential units can be expected to decrease housing prices in a given area. On the other hand, increasing urban density produces aggregation economies and other spillovers that might increase the desirability - and in turn the housing prices - of a given neighborhood.

Several papers in the literature support the idea that with increasing neighborhood density, more amenities become sustainable, increasing the number of services, entertainment and dining options available in a given area. Glaeser et al. (2001) argue that city density creates agglomeration effects not only in production, but also in consumption, benefiting especially the service industry. Berry and Waldfogel (2010) show that the number of restaurants increases proportionally to market size, and that larger markets tend to accommodate more high quality restaurants even in per capita terms. Further, Carlino and Saiz (2008) and Rappaport (2008) argue that more residents might be attracted to living in urban areas that are more amenity-rich, potentially giving rise to a virtuous circle between amenities and density. Moreover, turning empty lots, old warehouses and abandoned buildings in new apartment developments can have a direct positive impact on the price of nearby property (Han)
in addition to boosting local tax revenue, increasing funding to services such as public transit and the local school district.

I collected an extensive dataset spanning each of the 129,859 property parcels in Minneapolis from the City of Minneapolis Tax Assessor Office, which was then geocoded using Hennepin County data. The resulting dataset includes the transaction history of each property, as well as structural characteristics. I use detailed information about construction permits to identify construction events - that is the construction or conversions of new apartment building developments. I identify 185 construction events and examine their impact on surrounding properties using both a traditional difference-in-differences approach and a non-parametric framework.

Thus, I employ a new econometric technique developed by Diamond and McQuade (2019) to estimate the impact of apartment building construction in a non-parametric setting where treatment is a smooth function of distance. The authors develop this method to examine the repercussions of Low Income Housing Tax Credit (LIHTC) developments on the surrounding properties. The extension of this framework to estimate the effect of newly constructed apartment building is immediate. This estimation technique transforms data on property prices to data on the derivative of house prices with respect to distance from the site of a new housing development, building upon the statistics literature and in particular Charnigo et al. (2011) and Charnigo and Srinivasan (2015).

I focus on properties located within a one mile radius of a new development, that were sold within ±3 years from the start of the construction. I find that on average, properties located close to a newly developed apartment building tend to see the smallest gains in property values compared to the rest of their neighborhood after
construction is completed. However, as time goes by, this price effect is mitigated, with less significant within-neighborhood differences three years after construction is completed. This suggests that there are potentially non-linear interactions in the interplay of time since construction and distance from construction that are difficult to capture with a standard parametric approach. Moreover, my findings suggest that the impact of new apartment building developments can vary heterogeneously with respect to neighborhood characteristics. In particular, properties located in close proximity to a new apartment building in relatively less affluent neighborhoods seem to appreciate more than surrounding properties.

In Section 3.4 I present the data used for the analysis. Section 2.3 presents the reduced form results that examine the effect of new developments on neighboring property values using a standard difference-in-differences technique. I then proceed in replicating the setup developed by Diamond and McQuade (2019) to the case of rental apartment developments. Section 2.4 introduces the model of housing choice underpinning the econometric technique they develop, and Section 2.5 describes the details of the estimation process, outlining the steps for the empirical derivative estimation, kernel smoothing and numerical integration. Last, estimation results are presented in Section 3.5.

2.2 Data

This analysis uses data collected from the City of Minneapolis Tax Assessor Office to quantify the effects of the recent boom in construction of apartment buildings on nearby property prices. Large-scale apartment buildings have become an increasingly common feature of the urban landscape. While only 22 new apartment
Figure 2.1: New Construction in Minneapolis 1991-2016

Note: This figure illustrates the number of new apartment buildings and the number of new units built in each year between 1991 and 2016 in the City of Minneapolis. The total number of new apartment buildings (top row) is calculated as the number of newly constructed multifamily buildings containing at least six units. The total number rental apartments (middle row) is calculated as the total number of units within each newly constructed apartment building. The total number of owner-occupied units (bottom row) is calculated as the total number of new single family homes, condominiums, townhouses, duplexes, and triplexes built each year.
buildings were constructed in Minneapolis during the 1990s, their number increased to 65 in the 2000s, and to 91 between 2011 and 2016. This trend is perhaps more evident when examined in terms of the total number of newly constructed apartments: while only 1,765 new apartments were built in the 1990s, 4,462 new apartments were built in the 2000s and 9,262 built between 2011 and 2016 within Minneapolis city limits. By contrast, the construction of new residential units primarily intended for ownership, such as condominiums, single family homes, and townhouses, peaked during the housing bubble in the mid 2000s, with 5,689 units built between year 2001 and 2010, and stalled thereafter. Since 2010, only 630 new owner-occupied units have been constructed in Minneapolis.\footnote{Henceforth, residential units primarily intended for ownership, including condominiums, single family homes, townhouses, duplexes and triplexes will be referred to as “owner-occupied units.” Technically, these units can be rented by their owners, but they are usually more likely to be owner-occupied (Figure 2.2).}

The construction of rental apartments has consistently outstripped the construction of owner-occupied units, such as single family homes and condominiums, starting

\begin{figure}
\centering
\caption{Share of Rentals in Minneapolis 1991-2016}
\includegraphics[width=\textwidth]{figure22.png}
\end{figure}

\textit{Note:} This figure illustrates the share of single family homes, condominiums, townhouses, duplexes, and triplexes with an active rental license in each year between 1991 and 2016 in the City of Minneapolis.
Concomitantly, apartment buildings have grown larger and more luxurious. The average number of apartments per building has nearly doubled since 1991, and the inflation-adjusted value of each apartment has increased by over 50% over the same time period (Figure 2.3).

The data used in this analysis was compiled from the City of Minneapolis Tax Assessor Office, whose primary objective is to estimate property values yearly in order to levy property taxes. The main dataset includes a panel of property valuations that starts in 1988 for each of the 129,859 property parcels in the City of Minneapolis,

Table 2.1: Summary Statistics for Apartment Building Construction

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St Dev</th>
<th>Min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Units</td>
<td>83.724</td>
<td>74.884</td>
<td>6</td>
<td>30</td>
<td>59</td>
<td>119</td>
<td>354</td>
<td>185</td>
</tr>
<tr>
<td>Average Unit Value</td>
<td>115,372</td>
<td>80,889</td>
<td>5,188</td>
<td>60,417</td>
<td>96,344</td>
<td>149,929</td>
<td>600,000</td>
<td>159</td>
</tr>
</tbody>
</table>
basic structural information, such as the number of bedrooms, bathrooms and year of construction for each property, land use codes defining whether the parcel is for residential or commercial use, housing type, the sale history starting in 1986, recording transaction price as well as the name of any present and past owner, the rental history starting in 1991, and any business licenses or construction permits issued after 1999. This information was merged with Hennepin County records to obtain the GIS location and lot size of each parcel, which allows for the identification of the precise geographic location of each property.

The City of Minneapolis Tax Assessor records show that 185 new multi-unit apartment buildings constructed between 1991 and 2016, with an average of 83.7 apartments per building and an average unit value of $115,372. Summary statistics are reported in Table 2.1, while Figure 2.4 maps the 185 separate locations of each construction event.

Following Bajari and Benkard (2005), I examine residential property sales within a radius of one mile of the construction of each new apartment building occurring within an interval of ± 5 years from the start of construction, as the neighborhood surrounding each construction event can be viewed as a continuous surface of equilibrium prices. This yields a total of 133,519 transactions taking place between 1986 to 2016. Summary statistics are reported in Table 2.2.

---

These criteria are somewhat more restrictive than those of Diamond and McQuade (2019), who use a radius of 1.5 miles around each LIHTC development. However, I feel this is appropriate since Minneapolis is much more densely populated than the average county in Diamond and McQuade (2019) analysis.
Figure 2.4: Construction Events within City of Minneapolis, 1991-2016

Note: This figure illustrates the locations of the 185 new multi-unit (6+) apartment building constructed in Minneapolis between 1991 and 2016.
Table 2.2: Summary Statistics for Apartment Building Construction

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>St Dev</th>
<th>Min</th>
<th>p25</th>
<th>p50</th>
<th>p75</th>
<th>Max</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sale Price</td>
<td>207,628</td>
<td>180,386</td>
<td>20,000</td>
<td>109,900</td>
<td>170,000</td>
<td>246,000</td>
<td>6,300,000</td>
<td>133,519</td>
</tr>
<tr>
<td>N. of Sales</td>
<td>3.034</td>
<td>1.391</td>
<td>1.000</td>
<td>2.000</td>
<td>3.000</td>
<td>4.000</td>
<td>6.000</td>
<td>133,519</td>
</tr>
<tr>
<td>Condominium</td>
<td>0.169</td>
<td>0.375</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>133,437</td>
</tr>
<tr>
<td>Duplex/Triplex</td>
<td>0.101</td>
<td>0.302</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>133,437</td>
</tr>
<tr>
<td>Single Family</td>
<td>0.703</td>
<td>0.457</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>133,437</td>
</tr>
<tr>
<td>Townhouse</td>
<td>0.027</td>
<td>0.162</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>133,437</td>
</tr>
<tr>
<td>Square Footage</td>
<td>2,120</td>
<td>931</td>
<td>1.000</td>
<td>1,584</td>
<td>2,024</td>
<td>2,512</td>
<td>18,259</td>
<td>133,415</td>
</tr>
<tr>
<td>N. of Stories</td>
<td>1.466</td>
<td>0.468</td>
<td>1.000</td>
<td>1.000</td>
<td>1.200</td>
<td>2.000</td>
<td>5.000</td>
<td>132,500</td>
</tr>
<tr>
<td>N. of Baths</td>
<td>1.828</td>
<td>0.857</td>
<td>1.000</td>
<td>1.000</td>
<td>2.000</td>
<td>2.000</td>
<td>12.000</td>
<td>132,042</td>
</tr>
<tr>
<td>N. of Bedrooms</td>
<td>3.011</td>
<td>1.096</td>
<td>0.000</td>
<td>2.000</td>
<td>3.000</td>
<td>4.000</td>
<td>15.000</td>
<td>108,436</td>
</tr>
</tbody>
</table>

2.3 Exploratory Analysis

The goal of this analysis is to quantify the net effect of constructing a new apartment building on the prices of other properties in the same neighborhood. Standard microeconomic theory suggests that if owning and renting an apartment are substitutes, increasing the supply of apartments should decrease the price of apartments and that of substitute goods, that is, housing units primarily intended for ownership, such as condominiums, townhouses and single-family homes.

However, the construction of large apartment buildings can also create several externalities within a neighborhood. In the short term, the disruption caused by construction can have detrimental effects on neighborhood property prices. In the long term, an increase in congestion and traffic can make the neighborhood less desirable. On the other hand, increasing the number of residents within a given neighborhood can produce agglomeration economies with positive spillover effects on property prices. More amenities, such as retail stores, restaurants and coffee shops can be supported in a neighborhood when population density increases. Moreover, an increase in the aggregate property tax revenue can translate in an increase in the available public
goods, such as higher funding for the local school district and a greater incentive to improve public transit access in the neighborhood. If rental apartments and condominiums are not close substitutes, it is possible that the effects of these agglomeration economies will outweigh the negative effect on prices of an increase in the supply of residential units. Thus, the impact of the construction of apartment buildings on nearby property values is ultimately an empirical question.

Here, I examine the effects of constructing an apartment building on neighboring property prices by viewing new construction “events” as a treatment, and comparing the change in residential property prices in the three years before and after construction in treated and untreated areas. Endogeneity is obviously a concern, as developers’ decisions to build in a given neighborhood is certainly correlated with local housing trends that affects property prices. However, once the decision is made to build in a specific neighborhood, developers are constrained to few possible locations due to land availability. While it is true that in general the location of the site of new apartment building is not random, the key assumption is that, were it not for the construction of a new apartment building, properties located in the same neighborhood would have followed the same price trend. Under this assumption, it is then possible to examine the effects of construction with traditional econometric methods, such as a difference-differences method. In this application, a neighborhood is defined as the one mile radius around a new apartment building.

The main focus of the analysis is whether distance from a new development might account for within neighborhood, within year price variation, conditional on the set of control variables. I include variables controlling housing characteristics, such as square footage, the number of stories, the number of bedrooms and bathrooms, building age and building type (condominiums, duplexes/triplexes, single family homes.
Table 2.3: Difference in Difference Regression Results, Log Sale Price

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>-0.00876**</td>
<td>-0.00841**</td>
<td>-0.00810*</td>
<td>-0.00780*</td>
<td>-0.00666</td>
<td>-0.00627</td>
<td>-0.00603*</td>
<td>-0.00586**</td>
<td>-0.00600</td>
<td>-0.0237***</td>
</tr>
<tr>
<td></td>
<td>(0.00465)</td>
<td>(0.00467)</td>
<td>(0.00472)</td>
<td>(0.00479)</td>
<td>(0.00482)</td>
<td>(0.00484)</td>
<td>(0.0049)</td>
<td>(0.00509)</td>
<td>(0.0052)</td>
<td>(0.00570)</td>
</tr>
<tr>
<td>Distance Dummy</td>
<td>-0.00865**</td>
<td>-0.00828**</td>
<td>-0.00811**</td>
<td>-0.00784**</td>
<td>-0.00688</td>
<td>-0.00664</td>
<td>-0.00638**</td>
<td>-0.00619**</td>
<td>-0.00600</td>
<td>-0.0237***</td>
</tr>
<tr>
<td></td>
<td>(0.00288)</td>
<td>(0.00289)</td>
<td>(0.00289)</td>
<td>(0.00289)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
</tr>
<tr>
<td>Distance Dummy * Post</td>
<td>-0.00866**</td>
<td>-0.00829**</td>
<td>-0.00813**</td>
<td>-0.00784**</td>
<td>-0.00688</td>
<td>-0.00664</td>
<td>-0.00638**</td>
<td>-0.00619**</td>
<td>-0.00600</td>
<td>-0.0237***</td>
</tr>
<tr>
<td></td>
<td>(0.00288)</td>
<td>(0.00289)</td>
<td>(0.00289)</td>
<td>(0.00289)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
<td>(0.00288)</td>
</tr>
<tr>
<td>Distance</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
<td>0.000317***</td>
</tr>
<tr>
<td>sq, Feet</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
<td>(12.98-06)</td>
</tr>
<tr>
<td># of Stories</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
<td>0.155***</td>
</tr>
<tr>
<td></td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
<td>(0.00447)</td>
</tr>
<tr>
<td># of Bathrooms</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
<td>0.000121**</td>
</tr>
<tr>
<td></td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
<td>(0.00011)</td>
</tr>
<tr>
<td>Building Age</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
<td>-0.000316***</td>
</tr>
<tr>
<td></td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
<td>(0.00306)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
<td>0.00122***</td>
</tr>
<tr>
<td></td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
<td>(0.00031)</td>
</tr>
<tr>
<td>Observations</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
<td>188,353</td>
</tr>
<tr>
<td>Lognormal</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
<td>0.649</td>
</tr>
<tr>
<td>Building Code Fixed Effect</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Year Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Month Fixed Effects</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
| Construction Event Fixed Effects| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes | ** Note:** In specification (1), the distance dummy takes value of 1 if the property is located within 0.1 miles from a construction event. In specification (2), the distance dummy takes value of 1 if the property is located within 0.2 miles from a construction event, and so on through specification (9).
In specifications (1)-(9) I use a set of dummies to examine how the impact of the construction of a new apartment building varies by distance. For example, in specification (1) the “Distance Dummy” variable takes the value of 1 if a property is located within 0.1 miles from the location of the construction, and 0 if the property is located between 0.1 and 1 miles from the site of the new apartment building. In specification (2) the threshold is 0.2 miles, and so on through specification (9), where the threshold is 0.9 miles. This is akin to drawing a series of concentric circles around the location of construction and comparing prices between properties located in the inner circle and in the outer circle for varying radii of the inner circle.

For convenience, the coefficients on the interaction effect between the distance dummy and the post-construction dummy, capturing the difference-in-differences estimated
effect, are graphed in Figure 2.5. In general, being closer to the site of a new apartment building has a negative effect on housing prices, with an negative and significant interaction term in specifications (2) through (6). Thus, being within 0.6 miles of a newly built apartment building has a negative impact on prices. This effect is most pronounced for properties within 0.2 miles, whose prices are predicted to decrease 2.7%, and decreases thereafter.

In the last specification of Table 2.3, I display the results of a continuous effect difference-in-differences set-up. The key variables here are distance from the new development, and its interaction with a dummy variable taking value of 1 if the sale occurred after the construction. I find that house prices increase by about 2.0% with each additional mile from the new construction. Since the coefficient on the post-construction dummy is negative (-2.4%), taken together these results suggest that the net impact of the construction of a new apartment building on local property prices is negative, with nearby properties seeing a larger price decrease than properties located further away.

Thus, the results from reduced-form estimation indicate that the negative impact on price stemming from an increase in the supply of housing outweighs the potential positive effects resulting from agglomeration economies and other spillovers, and that rental apartments and units intended primarily for ownership are indeed substitute goods. There is some evidence (Figure 2.5) of more complex interactions between sale prices, temporal distance and geographical distance that are hard to capture in a traditional linear regression setting. The non-parametric estimation method presented next allows for a more flexible estimation approach to attempt to capture these interactions.
2.4 Housing Choice Model

This analysis is based on a model of housing choice analogous to the hedonic model presented in Diamond and McQuade (2019) which in turn builds on the hedonic model introduced by Rosen (1974). Let each housing property $j$ be characterized by a vector $(r_j, X_j, Y_j, \xi_j)$, where $r_j$ denotes the distance from the site of a new apartment building, $X_j$ is a vector of attributes specific to property $j$, $Y_j$ is a vector of neighborhood characteristics, and $\xi_j$ is a vector of characteristics observable to prospective buyers but not to the econometrician. Prospective home buyers view properties differing along any dimension of $(r_j, X_j, Y_j, \xi_j)$ as distinct goods. This assumption allows for the identification of the heterogeneous effects arising from construction taking place at different times or in different types of neighborhood.

Home buyers are assumed to maximize their utility with respect to the set of attributes of property $j$ and a composite consumption good $c$, subject to their income $y_i$:

$$\max U_i(r_j, X_j, Y_j, \xi_j, c) \text{ subject to } p_t(r, X, Y, \xi) + c \leq y_i$$

(2.1)

Note that some of the attributes of property $j$, such as square footage, lot size and distance form a new apartment complex, naturally form a continuous choice set. Thus, at least for these continuous choice variables, home buyers will be able to select the optimum bundle of characteristics such that its marginal utility equals its marginal cost. This is also the case for distance from a new development, $r_j$, which in turn implies:

$$\frac{\partial U(r_j^*, X_j^*, Y_j^*, \xi_j^*, c^*)}{\partial r_j} = \frac{\partial p_t(r_j^*, X_j^*, Y_j^*, \xi_j^*)}{\partial r_j}$$

(2.2)

Under standard regularity assumptions, Bajari and Benkard (2005) show that, in equilibrium, the hedonic model implies that there exists a unique and continuous
price surface, which can non-parametrically be estimated to recover the consumers' marginal preferences with respect to the housing market.

2.5 Estimation

Under the hedonic framework outlined above, each transaction price can be viewed as a single equilibrium price point on a price surface, varying continuously with respect to the non-discrete attributes in \((r_j, X_j, Y_j, \xi_j)\). The impact of apartment building construction can be investigated by examining all housing sales occurring within a one mile radius of a new apartment building development, which is henceforth designated as neighborhood \(l\). Each location within neighborhood \(l\) can be identified by its polar coordinates, the radial distance \(r\), the angle measured in radians, \(\theta\), and the amount of time since the start of construction, \(\tau\).

Under the hedonic framework in which individual house prices reflect both house and neighborhood characteristics, the triple \((r, \theta, \tau)\) forms a unique mapping to the set of available characteristics \((X, \xi)\) in neighborhood \(l\):

\[
(X, \xi) = g_l(r, \theta, \tau)
\]  

(2.3)

Hence each transaction price can be viewed as a function of the house and neighborhood characteristics unique to location \((r, \theta)\) and time \(\tau\):

\[
p_{jt} = p_t(r, X, Y, \xi) = p_t(r, g_l(r, \theta, \tau), Y)
\]  

(2.4)

Assuming that the impact of new construction on a given property is additively separable with respect to the other attributes of that property, prices can be decomposed
as:

\[
\log p_t(r, X, Y, \xi) = \tilde{m}_Y(r, \tau) + h_t(g_t(r, \theta, \tau)) \tag{2.5}
\]

where \( \tilde{m}_Y(r, \tau) \) is a derivative capturing the change in house prices with respect to the distance from the site of a new apartment building, while \( h_t(g_t(r, \theta, \tau)) \) accounts for the impact of other housing attributes. The impact of construction is allowed to vary with respect to geographic distance, measured by the number of miles \( r \) from the development, and temporal distance, measured as the number of years \( \tau \) transpired since the inception of construction.

In order to non-parametrically estimate the pricing surface in each neighborhood \( l \), it is necessary to calculate the empirical derivative at the site of each housing transaction, indexed by \( \iota \). In an ideal setup, this could be achieved by comparing the prices of identical properties differing only in their distance from \( \iota \), holding everything else constant. In other words, the angle \( \theta \) and the time at which the transaction occurs, \( t \), would be held constant, allowing only the distance from \( \iota \), \( (r_p - r_\iota) \) to vary. However, housing transactions do not occur with a sufficient frequency as to allow holding \( \theta \) and \( t \) fixed. The next-best option is then comparing the prices of properties “sufficiently close” to \( \iota \) in the \( \theta \) and \( t \) dimension, where “sufficiently close” is precisely defined below in Eq. (2.6).

The set of eligible locations to compute the derivative at location \( \iota \) is given by:

\[
L_{r,\iota} := \left\{ p \in \{1, ..., n\} : \frac{(t_p - t_\iota)^2}{(r_p - r_\iota)^2} < \vartheta^t, \frac{(\theta_p - \theta_\iota)^2}{(r_p - r_\iota)^2} < \vartheta^\theta \right\} \tag{2.6}
\]

In this application, the bow-tie width in years is chosen to be \( \vartheta^t = 1.6 \), while the bow-tie width in terms of distance perpendicular to \( r \) is set as \( \vartheta^\theta = 0.4 \). Note that
this specification implies an acceptable threshold for angular distance $(\theta_p - \theta_i)$ that is increasing in the radial distance $(r_p - r_i)$ from $i$; similarly, the acceptable threshold for temporal distance $(t_p - t_i)$ is the wider, the larger the radial distance $(r_p - r_i)$. Thus, the eligible set forms a three-dimensional bow-tie centered around the transaction point $i$. Figure B.1 reproduces an illustration from Diamond and McQuade (2019) to help visualizing the set $L_{r,i}$ in two-dimensions.

Once the eligible set of transactions around $i$ has been identified, the derivative capturing the change in transaction prices with respect to distance from the new development, $r$, is computed. Transactions in the bow-tie are divided in two subsets, where set $b$ encompasses all transactions in the triangular section closer to the new building development than $i$, and set $a$ denote the transactions in the bow-tie that lie further away than $i$ from the new development. The transactions in each set are then ranked with respect to their distance from $i$, so that $a(1, i, r)$ and $b(1, i, r)$ will be the closest pair of observations to $i$, where $r_{b(1,i,r)} \leq r_i \leq r_{a(1,i,r)}$. Similarly, the pair $a(2, i, r)$ and $b(2, i, r)$ will comprise the next closest observations to $i$, so that $r_{b(2,i,r)} < r_{b(1,i,r)} < r_i < r_{a(1,i,r)} < r_{a(2,i,r)}$, and so on.

I select at most $\kappa = 5$ pairs of transactions within each bow-tie, although I allow the possible number of transaction pairs to fall below $\kappa$, so that in practice $\kappa_n \leq \kappa$. The empirical derivative is a weighted average of the price difference of each of the $\kappa_n$ pairs:

$$\tilde{Y}_{i,l} = \sum_{k=1}^{\kappa_n} w_k \frac{\log p_{a(k,i,r)} - \log p_{b(k,i,r)}}{r_{a(k,i,r)} - r_{b(k,i,r)}}$$

(2.7)

where the weights $w_k$ are such that observation pairs lying closer to $i$ are given higher weight than observation pairs lying further away:

$$w_k = \frac{(\kappa_n - k) + 1}{\kappa_n(\kappa_n + 1)/2}$$

(2.8)
Since the focus of this analysis is on changes in property prices before and after the construction of a new apartment building, using post-treatment observations to compute the derivative at pre-treatment locations \( \tau < 0 \), and vice versa, is to be avoided. Thus, for each location two derivatives, one using exclusively pre-treatment data \( \hat{Y}_{l,-\tau} \), and one using only post-treatment data \( \hat{Y}_{l,\tau} \) are calculated.

Once an estimate for the pre- and post-treatment \( \hat{Y}_{l} \) is available, a Nadaraya-Watson kernel estimator is used to smooth the empirical derivative of transaction prices across different locations \((r,t)\):

\[
\hat{\Phi}_l(r,t) = \frac{n^{-1} \sum_{i=1}^{n} K_h((r,t) - (r_{i},t_{i})) \hat{Y}_{l,i}}{n^{-1} \sum_{i=1}^{n} K_h((r,t) - (r_{i},t_{i}))}
\]

(2.9)

where:

\[
K_h((r,t) - (r_{i},t_{i})) = \frac{1}{h_{r,n}h_{t,n}} K\left(\frac{r-r_{i}}{h_{r,n}}, \frac{t-t_{i}}{h_{t,n}}\right)
\]

(2.10)

and \(K(\cdot, \cdot)\) is the multiplicative multivariate Epidanokov kernel function in two dimensions, with bandwidths \(h_{r,n}, h_{t,n}\), defined as

\[
K(u_1, u_2) = \left(\frac{3}{4}\right)^2 (1-u_1^2)_{1[|u_1| \leq 1]} (1-u_2^2)_{1[|u_2| \leq 1]}
\]

(2.11)

where the first argument is \(u_1 = \frac{r-r_{i}}{h_{r,n}}\) and the second is \(u_2 = \frac{t-t_{i}}{h_{t,n}}\). The bandwidth for kernel smoothing in miles is \(h_{r,n} = 0.3\), while bandwidth for kernel smoothing in years is \(h_{t,n} = 5\). The before and after kernel estimates are then averaged across the \(N_Y = 185\) apartment building construction events to finally estimate the price gradient:

\[
\frac{\partial \hat{m}_Y(r, \tau)}{\partial r} = \frac{1}{N_Y} \sum_{l \in Y} \left[ \hat{\Phi}_l(r, T_l + \tau) - \hat{\Phi}_l(r, T_l - 1) \right]
\]

(2.12)

Diamond and McQuade (2019) generalize a consistency result from Charnigo et al. (2011) and Charnigo and Srinivasan (2015), showing that for nonparametric derivatives where data is randomly observed, the difference \(\hat{\Phi}_l(r, T_l + \tau) - \hat{\Phi}_l(r, T_l - 1)\)
converges in probability to \( \frac{\partial \hat{m}(r,r)}{\partial r} - \frac{\partial \hat{m}(r,-1)}{\partial r} \).

2.6 Results

I obtain price effects by numerically integrating the estimated derivative defined in Eq. (2.12) with respect to the distance from a new development, \( r \), to obtain an estimate of the change in the log of property prices. Figure 2.6 illustrates the estimated impact of the construction of an apartment building on the price level of nearby properties.

Immediately after the completion of construction, a clear trend is visible, whereby property prices seem to increase significantly the further they are from the site of (future) construction. In subsequent years, this trend is somewhat less pronounced, with smaller gains in property values for properties located away from construction. Notably, properties located within 0.2 miles of a new apartment building development see the least increase in their property prices. This might be explained by the disruptive effect that a large construction site can have on its immediate neighborhood. Moreover, it is possible that property prices might react more strongly once a large number of vacant apartments hits the market, and that this effect might weaken as the vacancy rate decreases in later years.

2.6.1 Heterogeneous Effects

Next, I examine possible heterogeneous effects by dividing the 185 event neighborhoods into a group of above-average property value neighborhoods and a group of
Figure 2.6: Average Price Impact of New Apartment Building Construction

Kernel smoothed estimates of the log of residential property prices using Nadaraya-Watson function with Epanechnikov kernel. Estimates integrate over the estimated derivatives to measure log price change.

below-average value neighborhoods. I do so by ranking neighborhoods according to their average transaction price, and subdividing them in an above- and below-mean samples, where the mean neighborhood as an average sale price of $207,628.

Figure 2.7 illustrates the estimated effect in above-median neighborhoods. These more affluent neighborhoods display similar trends to the overall average. Property prices increase with distance from the site of a new apartment building development after the construction of an apartment building, with the trend being slightly less marked 2 or more years after construction has completed. Properties located within 0.2 miles of the new development again see the smallest gain in property values com-
Kernel smoothed estimates of the log of residential property prices using Nadaraya-Watson function with Epanechnikov kernel. Estimates integrate over the estimated derivatives to measure log price change. I sort my 185 neighborhoods into two groups, one with high average property prices and one with low average property prices. The cutoff used is a $207,628 average sale price, equal to the median neighborhood average sale price.

Figure 2.7: Price Impact in Neighborhood Above Median Sale Value

Figure 2.8 illustrates instead the estimated effect in below-median neighborhoods. In these neighborhoods the trend of increasing prices as distance from the site of a new apartment building is much less pronounced. Moreover, in these relatively less affluent neighborhoods, locations in close proximity to the site of the new apartment building do not see the same dip in property value as more affluent neighborhoods. This suggests that the amenity value produced by the construction of a new apartment building in less prosperous neighborhoods might outweigh the negative impact...
2.7 Conclusion

In this paper, examines the effect of new apartment building construction on neighboring property prices. From a theoretical standpoint, the impact of the development of new apartment buildings on local property prices is not clear a priori. On one hand, a new apartment building can negatively affect local property prices by increasing
the supply of available residential units in a given area. Further negative spillovers such as construction noise and an increase in traffic and congestion in the neighborhood could further compound this effect. On the other hand, the development of a large-scale apartment building might produce aggregation economies and other positive spillovers that might increase the desirability of a given neighborhood and in turn property prices.

Estimating the net effect under a standard parametric framework is complicated by the non-linear interaction of geographic and temporal distance from the construction. I employ a non-parametric approach that follows Diamond and McQuade (2019) to estimate the net effect of apartment building construction on nearby residential property prices. Overall, I find that properties in close proximity to a new apartment building tend to see the smallest gains in property prices compared to their neighborhood after construction is completed, although this effect might lessen over time. I find some evidence that these effects might vary heterogeneously across different types of neighborhood, with properties in close proximity to the site of a new apartment development being relatively better off in less affluent neighborhoods.
Chapter 3

Accessibility or Amenities?
Estimating the Value of Light Rail Transit

3.1 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. 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. However, these studies do not identify which effects are directly attributable to improved access to public transportation, and which effects arise indirectly from the transit system. These indirect effects arise from the fact that 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 additional businesses, such as restaurants and entertainment, 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.

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), Houston (Pan, 2013), Miami (Catanzarro and Smith, 1993), Los Angeles (Cervero and Duncan, 2002), Philadelphia (Kilpatrick et al., 2007), Phoenix (Seo et al., 2014), Portland (Duerer 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; Greer, 1992), Cervero, 1994). Outside the United States, there are studies focusing on Amsterdam (Debrezion et al., 2011), Beijing (Zheng et al., 2016), Bogotá (Tsianikis, 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), Shanghai (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.
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 the local market 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 adapts recent techniques from the machine learning literature to decompose the benefits of introducing a light rail system into direct and indirect effects. Furthermore, we estimate how both the direct and indirect effects can vary heterogeneously across different types of neighborhoods. Our approach allows us to identify relevant features given a large selection of covariates in a data-driven manner, while simultaneously incorporating instrumental variables to control for endogeneity.

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. 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 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

\footnote{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.}
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 3.1 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 increases. The large increase in local amenities in response to the introduction of the Blue Line likely had a sustained impact on house prices and consumer welfare in addition to the direct impact of improved access to public transit. 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 access to public transit, or whether they are due to a spillover effect of the entry of new 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 each station benefiting more than those further away. Employing a more flexible specification allows 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
Figure 3.1: Establishments within 1 mile of a Blue Line Station

(a) Restaurants

(b) Entertainment
(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 observations 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 how the estimated effect varies heterogeneously. 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 the spillover effect.

The results of our estimation routine using Boosted 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 on this topic focus exclusively on neighborhoods in Southern Minneapolis. The Boosted 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 3.2 introduces the theoretical framework for our estimation, while Section 3.3 discusses the different possible approaches to estimating treatment effects, such as difference-in-differences and machine learning methods. Section 3.4 summarizes our data sources for housing values, neighborhood amenities and demographics, while Section 3.5 presents our results under the different estimated approaches, and Section 3.6 concludes.

3.2 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 characteristics.

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 vector 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 $p_{t-1}(z) = p_t(z) = \hat{p}(z)$. 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.\footnote{See Taylor (2003) and Palmquist (2005) for a review of the empirical literature and various difficulties that arise when using a hedonic approach.} Dealing with a shifting pricing surface is beyond the scope of this paper, so we assume that $p_t(z) = \hat{p}(z)$. For similar reasons, 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 could not interpret $\frac{\partial p(z)}{\partial z_i}$ as the true marginal willingness-to-pay, but rather as a bound on it.
If, following the literature, we assumed that houses are provided in a perfectly competitive environment so that there is an absence of strategic effects, our estimates could then 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 approach allows for the recovery of 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. Moreover, hedonic pricing models tend not to be reliable in the presence of unobserved product characteristics that determine the pricing surface. We assume that the issue of unobserved product characteristics is less impactful to our results than in other environments since we rely on a large number (almost two hundred) of observed product characteristics in our estimation routine. Another reason we depart from Bajari and Benkard (2005) is that each individual 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:
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. This 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.

3.3 Estimating Treatment Effects

Figure 3.2: Blue Line DAG

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

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

Because house prices are observed, consistent estimation of the MWTP is equivalent to consistently estimating $\beta_d$. This approach implicitly assumes that the covariates are orthogonal to the error term, but this is unlikely to be true when amenities are included as a determinant of house prices. To see why, consider the causal relationship depicted in Figure 3.2. 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.

3.3.1 Cross Sectional Methods

The simplest approach to estimating 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. 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 researcher 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 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 local nature of the data generating process.

### 3.3.2 Difference-in-Differences

Recognizing the limitations of a cross-sectional approach, recent papers have instead employed a difference-in-differences empirical strategy. This strategy entails defining treatment and control groups using concentric circles around each transit station,

---

\(^6\) See for example Zheng et al. (2016).
treating the inner circle as the treatment group and the outer circle as the control group. 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, 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.

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 some of the endogeneity concerns arising from omitted variable bias in traditional cross sectional hedonic models. Additionally, this method can be extended to allow for the estimation of continuous pricing surfaces, such as in Diamond and McQuade (2019).

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

\footnote{Among others, these include \cite{Baum-Snow_2000}, \cite{Gibbons_2005}, \cite{Goetz_2010}, \cite{Hillings_2011}, \cite{Wagner_2017}, \cite{Pilgram_2018}.}
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 can once again make 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.

3.3.3 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:

\[ p_{it}^c = f(X_{it}^c) + \epsilon_{it}^c \]

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

\[ p_{it}^t = f(X_{it}^t) + g(X_{it}^t) + \epsilon_{it}^t \]

where \( p_{it} \) is the housing price (possibly measured in logs) and \( X_{it} \) is a vector of housing and neighborhood characteristics that determine house prices. Here, \( \epsilon_{it} \) 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 ($TE$) 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:

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

Under the assumption that $\epsilon_{it}^c|X_{it}^c \sim \epsilon_{it}^t|X_{it}^t$, 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. 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. Several different predictive models have been used in the literature, including regression using LASSO and gradient boosting with regressions trees. We implement the latter approach in Appendix C.2. Unfortunately, following this approach does not allow us to explicitly recover the indirect treatment effect as well. The indirect effect is given by:

---

8This 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 $f(X_{it})$. 

96
\[ IE = f(X_{it}^t) - f(X_{it}^c) \]

but we can only estimate the term:

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

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 perform well at predicting the treatment group because the control group should have been included in the treatment group to begin with.
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.

3.3.4 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:

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

where:

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

and

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

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

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

\[ L(x_{sj,i}; \gamma_j, c_j) = \frac{1}{1 + e^{-\gamma_j(x_{sj,i} - c_j)}} \]

so that every point has a positive probability of being assigned to any terminal leaf. As the term \( \gamma_j \) increases, the model converges to a standard decision tree. Moderate values of \( \gamma_j \) 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 \( \{B_{jk}(x_i; \theta_k)\}_k \) 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 C.3 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:

\[
r = -\gamma_m \nabla F_{m-1} L(p, F_{m-1}(X))
\]

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 in Appendix C.3 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.

3.3.5 Instrumental Variables

Our principal concern is that the level of amenities is correlated with the unobserved error term in the house price equation. The sign of this correlation is in general unknown. For instance, retail 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 rental costs, causing a negative correlation between their level and the unobserved error term. 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 orthog-
onal 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 instrumental variables (PEIV) and the instruments used in Fan (2013)\footnote{For a discussion of PEIV, see Li et al. (2020).}

If aggregate trends in demand and supply shifters also impact individual house prices then these instruments would be invalidated. This would be the case if house prices increase due 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.

3.4 Data

3.4.1 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 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 station, 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 3.1: Summary Statistics of Housing Data

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

3.4.2 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 C.1 for details.)

3.4.3 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 individual 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 C.1 for details.) To accomplish this, we calculated the number of businesses within a 0.5 miles radius of each property for each category. Further information on the quality of the amenities in the neighborhood of each individual 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.

3.4.4 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.
3.5 Results

3.5.1 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

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

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1
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 control group. Estimation results for this technique are reported in Table 3.2. 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 some caution for the reasons discussed in Section 3.3.

3.5.2 Boosted Smooth Trees

Table 3.3 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 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 $E[f(X_{it}^n) - f(X_{it}^n)]$, and thus the indirect effect. The change in amenities are predicted to increase the sale prices of properties located in the treatment group by
Table 3.3: Boosted Smooth Trees

<table>
<thead>
<tr>
<th></th>
<th>Pre-Treatment</th>
<th>Post-Treatment</th>
<th>Spillover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0009</td>
<td>0.0546</td>
<td>0.0584</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>(0.0166)</td>
<td>(0.0198)</td>
<td>(0.0255)</td>
</tr>
</tbody>
</table>

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 3.2. Following a similar procedure without instruments found a spillover of 1.3%, meaning that results that do not account for endogeneity would be downwardly biased and would tend to overstate the direct effect of the Blue Line relative to its indirect effect. With instruments, we find that the indirect effect accounts for over 51% of the total effect and is therefore an important channel through which public transportation impacts housing prices and consumer welfare.

These effects are not homogeneous and depend on where houses are located along the Blue Line. Figure 3.3 plots the direct treatment effect averaged across groups of houses along the path of the Blue Line. The direct treatment effect is lowest for the suburbs in Southern Minneapolis and certain parts of the downtown area, meaning that houses located in these neighborhoods did not see much of a pricing effect after the Blue Line was introduced as a result of better access to public transportation. This could be due to the fact that the downtown area is already served by several bus lines and the fact that the southern suburbs have a higher rate of drivers, so there is less need for public transportation. 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. The indirect treatment effect (Figure 3.4) is instead highest in the downtown area,
Figure 3.3: Predicted Change in Housing Prices, Direct Effect
Figure 3.4: Predicted Change in Housing Prices, Indirect Effect
which saw the largest increase in the entry of new amenities around the introduction of the Blue Line, while the suburbs in Southern Minneapolis were relatively unaffected by this channel.

3.6 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 Boosted 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 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 5.8%. Thus, just over half 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.
Bibliography


Clower, T. L., B. L. Weinstein, et al. (2002). The impact of dallas (texas) area rapid


Appendix A

Appendix to Chapter 1

A.1 MHFA Fix Up Loan Program Details

Figure A.1: Minnesota Housing Fix Up Loan Program Details

Note: APR amounts reported in the figure above are based on a 5.375% interest payable over 20 years for a $30,000 Fix Up secured loan; a 6% interest payable over 10 years for a $15,000 Fix Up unsecured loan; a 4.99% interest payable over 10 years for a $15,000 energy conservation loan (secured or unsecured). Source: City of Minneapolis.
A.2 Maps

Figure A.2: Minneapolis 2010 Census Tract and Block Group with Neighborhood Boundaries

Legend
- 2010 Census Tract
- 2010 Census BlockGroup
- Neighborhood

Note: This figure illustrates the boundaries for the Census tracts and for the Census block groups within the Minneapolis city limits. Source: City of Minneapolis
Figure A.3: Zip Codes in the City of Minneapolis

Note: This figure illustrates the boundaries for each ZIP code within the Minneapolis city limits. Source: City of Minneapolis
A.3 Robustness Checks: Incorporating Income

The structural model illustrated thus far is a powerful tool to examine housing consumption and the decision to invest in home improvement. With some additional assumptions, this model can be further refined to account for the role played by income in housing consumption choices. While income is an unobserved source of heterogeneity in the baseline model, its distribution within each neighborhood can be used to estimate choice probabilities and optimal investment in the context of a mixed logit model.

A.3.1 Data

The dataset on housing values and investment in home improvement obtained from the City of Minneapolis Tax Assessor is augmented using the Statistics of Income (SOI) data from the IRS. The SOI data reports the total number of income tax returns filed with the IRS each year, categorized according to the ZIP code provided on the taxpayer return. Since individual-level information on the income status of property owners is not available, this data is used to simulate the impact of income on the decision to invest in a property by approximating the income distribution of each neighborhood.

The key variable of interest is the number of returns filed with the IRS each year, that is, a count of all Forms 1040, 1040A and 1040EZ at each income level. The SOI statistics provide information on the number of returns that fall in each of the six adjusted gross income categories:

- $1 to $25,000
$25,000 to $50,000

$50,000 to $75,000

$75,000 to $100,000

$100,000 to $200,000

$200,000 or more

Table A.1: Average Number of Returns Filed in Minneapolis, 2006-2016

<table>
<thead>
<tr>
<th></th>
<th>$1 to $25,000</th>
<th>$25,000 to $50,000</th>
<th>$50,000 to $75,000</th>
<th>$75,000 to $100,000</th>
<th>$100,000 to $200,000</th>
<th>$200,000 or more</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Returns</td>
<td>94,011</td>
<td>68,154</td>
<td>34,844</td>
<td>19,189</td>
<td>23,698</td>
<td>8,846</td>
<td>248,743</td>
</tr>
<tr>
<td>%</td>
<td>37.79</td>
<td>27.40</td>
<td>14.01</td>
<td>7.71</td>
<td>9.53</td>
<td>3.56</td>
<td>100</td>
</tr>
</tbody>
</table>

The data is available for the 2006-2016 period for each of the 23 ZIP codes within the Minneapolis city limits. Table A.1 reports the average number of returns filed in Minneapolis within each adjusted gross income category. As reported in Table A.2, in general neighborhoods that have a higher median estimated market value \( P \) tend to also have a higher average income. Further information disaggregated by ZIP code is reported in Table A.3.

\(^1\) As illustrated by Figure A.3 in Appendix A.2 the City of Minneapolis reports 26 ZIP codes within its boundaries. There are however only 23 ZIP codes that have been matched with the housing data: 55401, 55402, 55403, 55404, 55405, 55406, 55407, 55408, 55409, 55410, 55411, 55412, 55413, 55414, 55415, 55416, 55417, 55418, 55419, 55421, 55423, 55430, 55454. This discrepancy is due to the fact that the IRS does not separately report information for ZIP Codes with less than 100 returns, identified as a single building ZIP code, or identified as a nonresidential ZIP code, as is the case for ZIP codes 55487 and 55488, and to the fact that there are no single family homes located within a given ZIP code, as is the case for ZIP code 55455.
Figure A.4: Income Distribution by Zip Code

Note: This figure illustrates the distribution of tax returns within each adjusted gross income category for each ZIP code within the Minneapolis city limits. The data is averaged over the 2006-2016 period.
Table A.2: Average Income by Median Neighborhood EMV

<table>
<thead>
<tr>
<th>$P$ Range</th>
<th>Average Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$100,000 to $200,000</td>
<td>$45,437</td>
</tr>
<tr>
<td>$$200,000 to $300,000</td>
<td>$62,525</td>
</tr>
<tr>
<td>$$300,000 to $400,000</td>
<td>$85,047</td>
</tr>
</tbody>
</table>

Note: This table was constructed by calculating the median EMV ($P$) for each zip code in the data. After subdividing $P$ in three category, the average zip code income was calculated based on a weighted average of IRS data over the 2006-2016 period.

A.3.2 Model

Income can be incorporated in the baseline model with minimal modifications. The component of utility generated by the consumption of the housing good, $q^{\theta_1}P^{\theta_2}$, does not need to be modeled to depend explicitly on income, as this term is common across all potential household choices. The only component of the choice-specific utility function that needs to be adjusted is the disutility generated by the cost of investment, which now becomes a function of the income level. Thus, the disutility of investment is now modeled to depend not only on the overall dollar amount spent on investment, $i$, but also on the household income, $y$.

The household’s utility function can then be written as:

$$ u(q, P|d) = \begin{cases} 
q^{\theta_1}P^{\theta_2}, & \text{for } d = 1; \\
q^{\theta_1}P^{\theta_2} - \theta_3 \frac{q^\gamma}{y^\delta} - FC_2, & \text{for } d = 2; \\
q^{\theta_1}P^{\theta_2} - \theta_5 (q \cdot P) - FC_3, & \text{for } d = 3.
\end{cases} $$

127
Table A.3: Minneapolis Mean and Median Income by Zip Code

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>55454</td>
<td>2,739</td>
<td>1to25k</td>
<td>22,262</td>
<td>2,980</td>
<td>1to25k</td>
<td>26,930</td>
</tr>
<tr>
<td>55411</td>
<td>9,195</td>
<td>1to25k</td>
<td>28,952</td>
<td>12,120</td>
<td>1to25k</td>
<td>33,197</td>
</tr>
<tr>
<td>55404</td>
<td>9,979</td>
<td>1to25k</td>
<td>29,626</td>
<td>12,060</td>
<td>25to50k</td>
<td>37,168</td>
</tr>
<tr>
<td>55430</td>
<td>9,844</td>
<td>25to50k</td>
<td>37,570</td>
<td>11,310</td>
<td>25to50k</td>
<td>41,099</td>
</tr>
<tr>
<td>55412</td>
<td>9,415</td>
<td>25to50k</td>
<td>36,310</td>
<td>10,680</td>
<td>25to50k</td>
<td>41,302</td>
</tr>
<tr>
<td>55444</td>
<td>8,085</td>
<td>1to25k</td>
<td>38,824</td>
<td>10,540</td>
<td>1to25k</td>
<td>47,275</td>
</tr>
<tr>
<td>55421</td>
<td>13,025</td>
<td>25to50k</td>
<td>41,585</td>
<td>14,720</td>
<td>25to50k</td>
<td>48,174</td>
</tr>
<tr>
<td>55407</td>
<td>15,321</td>
<td>25to50k</td>
<td>38,706</td>
<td>18,600</td>
<td>25to50k</td>
<td>48,595</td>
</tr>
<tr>
<td>55408</td>
<td>14,444</td>
<td>25to50k</td>
<td>39,337</td>
<td>18,490</td>
<td>25to50k</td>
<td>52,726</td>
</tr>
<tr>
<td>55423</td>
<td>18,172</td>
<td>25to50k</td>
<td>45,624</td>
<td>19,700</td>
<td>25to50k</td>
<td>56,534</td>
</tr>
<tr>
<td>55413</td>
<td>5,848</td>
<td>25to50k</td>
<td>41,109</td>
<td>7,230</td>
<td>25to50k</td>
<td>57,379</td>
</tr>
<tr>
<td>55418</td>
<td>15,293</td>
<td>25to50k</td>
<td>46,387</td>
<td>16,720</td>
<td>25to50k</td>
<td>61,184</td>
</tr>
<tr>
<td>55406</td>
<td>17,181</td>
<td>25to50k</td>
<td>46,295</td>
<td>18,540</td>
<td>25to50k</td>
<td>63,047</td>
</tr>
<tr>
<td>55403</td>
<td>8,572</td>
<td>25to50k</td>
<td>54,821</td>
<td>9,710</td>
<td>25to50k</td>
<td>67,212</td>
</tr>
<tr>
<td>55409</td>
<td>5,750</td>
<td>25to50k</td>
<td>52,709</td>
<td>6,280</td>
<td>25to50k</td>
<td>67,691</td>
</tr>
<tr>
<td>55405</td>
<td>7,769</td>
<td>25to50k</td>
<td>54,421</td>
<td>8,470</td>
<td>25to50k</td>
<td>67,934</td>
</tr>
<tr>
<td>55417</td>
<td>12,879</td>
<td>25to50k</td>
<td>57,576</td>
<td>13,510</td>
<td>50to75k</td>
<td>76,701</td>
</tr>
<tr>
<td>55416</td>
<td>16,413</td>
<td>25to50k</td>
<td>71,552</td>
<td>18,450</td>
<td>50to75k</td>
<td>88,974</td>
</tr>
<tr>
<td>55419</td>
<td>13,130</td>
<td>25to50k</td>
<td>75,710</td>
<td>13,960</td>
<td>50to75k</td>
<td>96,472</td>
</tr>
<tr>
<td>55401</td>
<td>4,111</td>
<td>25to50k</td>
<td>74,264</td>
<td>7,010</td>
<td>50to75k</td>
<td>99,183</td>
</tr>
<tr>
<td>55415</td>
<td>666</td>
<td>25to50k</td>
<td>59,745</td>
<td>1,900</td>
<td>50to75k</td>
<td>104,763</td>
</tr>
<tr>
<td>55402</td>
<td>954</td>
<td>50to75k</td>
<td>96,067</td>
<td>860</td>
<td>50to75k</td>
<td>112,442</td>
</tr>
<tr>
<td>55410</td>
<td>10,136</td>
<td>50to75k</td>
<td>80,678</td>
<td>10,340</td>
<td>50to75k</td>
<td>113,740</td>
</tr>
</tbody>
</table>

Note: This table summarizes the IRS data reported for 2006 and 2016 for each ZIP code within the Minneapolis city limits. The total number of tax returns, adjusted gross income category for the median return and mean income are reported for each ZIP code. The mean income for a ZIP code is calculated taking the average value of each income bracket, and constructing a weighted average using the number of returns as weights.
where $\theta$ is a parameter to translate the impact of income into utility term. Thus, the disutility of investing $i$ dollars in home improvement can now be written as $\theta \frac{i^3 y^2}{y^6}$. Everything else being the same, households with a higher income level tend to be less sensitive to the investment cost, while the fixed term $FC_2$ still captures the negative utility impact arising from the decision to invest.

A.3.3 Estimation

Let the data be a panel \( \{q_{dit}, P_{dit}, i_{dit}, q_{dit}, \pi_{yjt}\}_{i=1,...,N, \; t=1,...,T, \; y=1,...,Y} \) containing observations on housing quality, neighborhood quality, investment amount and investment decisions, and on the income distribution of each neighborhood. Assuming the investment amount is measured with error, the discrepancy between the observed investment $i_{dit}$ and that predicted by the policy function $i(q_{dit}, P_{dit}, \pi_{yjt} | \theta, d_{dit} = 2)$ is given by:

\[
i_{dit}^d = i(q_{dit}, P_{dit}, \pi_{yjt} | \theta, d_{dit} = 2)\xi_{it}
\]

where $\xi_{it}$ is assumed to be a random error term that follows the log-normal distribution with variance $\sigma^2_{\xi}$. Applying logs yields a linear approximation to the discrepancy between observed and predicted investment:

\[
\log(\xi_{it}) = \log(i_{dit}^d) - \log \left( i(q_{dit}, P_{dit}, \pi_{yjt} | \theta, d_{dit} = 2) \right)
\]

where $\log \xi_{it} \sim N(0, \sigma^2_{\xi})$, i.i.d., $\forall i, t$.

Two assumptions are needed at this point in order to ensure that this model can be estimated via maximum likelihood. First, the choice specific shocks $\epsilon(d)$ need to be
independent of the discrepancy between predicted and observed investment, $\log \xi_{it}$, as in the baseline model. The second assumption concerns the relationship between income and housing quality: while the joint distribution of income and neighborhood quality is observed and can thus be explicitly modeled in the likelihood function, the relationship between housing quality and income is not observed. In the absence of micro-data on the joint distribution of income and housing characteristics, an attempt can be made at estimating this model by assuming that housing quality is independently distributed with respect to income within each neighborhood. Hence, under this assumption, richer households are more likely to be located in more expensive neighborhoods, but within each neighborhood they are not more likely to select a more expensive house.

The likelihood contribution of household $i$ at time $t$ is then given by:

$$
\ell_{ijt}(\theta, \sigma_\xi) = \int P_t \left( d_{dit} = 1 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 1\} P_t \left( d_{dit} = 2 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 2\} P_t \left( d_{dit} = 3 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 3\} \left( \phi(\log \xi_{dit}(\theta)/\sigma_\xi) \right) I\{d_{dit} = 2\} f(y) dy
$$

where $f(y)$ represents the probability density of income $y$.

Since the IRS data on income is structured in six discrete categories, this equation can be re-written as:

$$
\ell_{ijt}(\theta, \sigma_\xi) = \sum_y \pi_{y_{jit}} P_t \left( d_{dit} = 1 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 1\} P_t \left( d_{dit} = 2 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 2\} P_t \left( d_{dit} = 3 \mid \theta, q_{dit}, P_{dit}^{d}, y_{jit}, s_{dit} \right) I\{d_{dit} = 3\} \left( \phi(\log \xi_{dit}(\theta)/\sigma_\xi) \right) I\{d_{dit} = 2\}
$$

Thus, the overall likelihood contribution for each observation is a weighted average of the likelihood contribution for a model solved for a specific income value $y_{jit}$, using
$\pi_{yjt}$, the probability of that income value being observed in that ZIP code, as the weight. Hence the model is solved six times, once for each income bracket, using the median bracket value as input value for $y_{jt}$.

Table A.4 reports the structural model parameters estimated via maximum likelihood compared to those of the baseline model.

Table A.4: Comparison of Estimates for Structural Model Parameters

<table>
<thead>
<tr>
<th></th>
<th>Baseline Model</th>
<th>With Income Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimates</td>
<td>Standard Errors</td>
</tr>
<tr>
<td><strong>Utility Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housing Quality</td>
<td>$\theta_1$</td>
<td>1.657 (0.694)</td>
</tr>
<tr>
<td>Neighborhood Quality</td>
<td>$\theta_2$</td>
<td>0.433 (0.104)</td>
</tr>
<tr>
<td><strong>Investment Cost Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Cost of Investment</td>
<td>$\theta_3$</td>
<td>0.041 (0.009)</td>
</tr>
<tr>
<td>Curvature on Investment</td>
<td>$\theta_4$</td>
<td>1.649 (0.413)</td>
</tr>
<tr>
<td>Curvature on Income</td>
<td>$\theta_6$</td>
<td>- -</td>
</tr>
<tr>
<td>Fixed Cost of Investing</td>
<td>$FC_2$</td>
<td>78.889 (139.752)</td>
</tr>
<tr>
<td><strong>Sale Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Sale Cost</td>
<td>$\theta_5$</td>
<td>0.010 (0.185)</td>
</tr>
<tr>
<td>Fixed Cost of Selling</td>
<td>$FC_3$</td>
<td>-1.029 (20.360)</td>
</tr>
<tr>
<td><strong>Variances</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tase Shock</td>
<td>$\sigma_\epsilon$</td>
<td>37.370 (32.111)</td>
</tr>
<tr>
<td>Measurement Error</td>
<td>$\sigma_\xi$</td>
<td>1.164 (1.885)</td>
</tr>
</tbody>
</table>
Figure A.5: Predictions of Baseline Model and Model with Income Effects

Note: This figure illustrates the predicted probability of investing in home improvement (top row), the probability of selling (middle row) and the optimal amount of investment (bottom row) as a function of housing quality, $q$ and neighborhood values, $P$. The line in red depicts the predictions of the baseline model, while the blue lines depict the predictions of the model with income for different income levels.
A.3.4 Results

Figure A.5 illustrates how the model’s predictions change once income is incorporated into the baseline model of housing consumption. The predicted likelihood of investment for the model with income effects is slightly lower at all levels of housing quality $q$ and neighborhood quality $P$, but does not display much variation across different income levels. The largest discrepancy between the two models is in the predicted probability of selling, which is lower than in the baseline model for all income levels, across almost the entire distribution of $q$ and $P$. The optimal investment amount predicted by the policy function is predicted to change quite dramatically based on the level of household income. Households with a higher level of income $y$ are predicted to have a much higher level of optimal investment for all levels of housing quality and neighborhood quality. In this specification, the optimal amount of investment remains increasing as housing quality and neighborhood quality increase, with the relationship being somewhat more pronounced for the top income brackets.

The policy predictions for the model with for income effects are illustrated in Figure A.6. The effects of a $\tau = 10\%$ subsidy available to all households are magnified once the role of income is introduced into the model. Optimal investment is predicted to increase by over $25,000$ for households living in neighborhoods at the top of the quality distribution, and by close to $17,000$ for households living in houses at the top of the quality distribution. The predicted effect of the means-tested policy is more modest compared to that of the baseline model (Figure 1.11), with smaller increases in optimal investments at all levels of $q$ and $P$. The difference between the optimal

---

2 Thus the model is solved for each of the following values of $y$: $12,500$, $37,500$, $62,500$, $87,500$, $150,000$, $280,000$. The value for the top income bracket, defined by the IRS as “$200,000 or more” is taken to be $1.4$ times the lower bound of the bracket. This approach is standard in the literature, see for example Lemieux (2006).
Figure A.6: Model Predictions Under Alternative Policy Scenarios

Note: This figure illustrates predicted optimal investment policy (top row) and probability of selling (bottom row) under different policy scenarios. The blue line depicts the prediction for the baseline model, while the red lines depict the model prediction under a policy allowing households to reduce investment cost by 10%. Solid red lines represent the prediction if this policy is applied to all households, and dashed red lines represent the prediction for a means-tested policy, available only to households earning under $75,000 per year. Predictions are based on 2016 data.

Investment prediction for a subsidy available to all households and a means-tested policy is at its minimum for households living neighborhoods at the lower end of the quality distribution, and increases as neighborhood values increase. Similarly to the baseline model’s predictions, the likelihood of selling a property remains unchanged or sensibly decreases under a policy aimed at subsidizing investment in home improvement.
<table>
<thead>
<tr>
<th>Model</th>
<th>Total Investment</th>
<th>Total Cost</th>
<th>Change in Investment</th>
<th>Returns per Dollar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No policy (τ = 0)</td>
<td>$34,795,751</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Policy (τ = 10%)</td>
<td>$45,930,716</td>
<td>$4,593,072</td>
<td>$11,134,965</td>
<td>$2.42</td>
</tr>
<tr>
<td>Policy (τ = 10%), Means-Tested</td>
<td>$42,286,104</td>
<td>$2,869,570</td>
<td>$7,490,353</td>
<td>$2.61</td>
</tr>
<tr>
<td><strong>With Income Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No policy (τ = 0)</td>
<td>$62,319,590</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Policy (τ = 10%)</td>
<td>$98,277,808</td>
<td>$9,827,781</td>
<td>$35,958,218</td>
<td>$3.66</td>
</tr>
<tr>
<td>Policy (τ = 10%), Means-Tested</td>
<td>$69,121,404</td>
<td>$4,425,931</td>
<td>$6,801,814</td>
<td>$1.54</td>
</tr>
</tbody>
</table>

Note: This table was constructed by calculating and aggregating the predicted level of investment in home improvement for each household under alternative policy scenarios, using 2016 data for Minneapolis.

Under a policy framework where all households receive a subsidy to mitigate the cost of investing in home improvement, richer households are predicted to have the largest response, while poorer households are predicted to have a more moderate reaction. Table A.5 illustrates the predicted impact of these policies on aggregate investment and compares them to the baseline model’s predictions. A subsidy available to all households is predicted to have a higher return per dollar spent ($3.66) than a means-tested policy ($1.54).

These results suggest that if the main objective of this subsidy were to allocate scarce resources to promote the revitalization of low income neighborhoods, a means-tested policy might be more effective at promoting investment in home improvement in low income neighborhoods without encouraging the displacement of existing residents. However, if the objective were to simply increase aggregate investment, a policy open to all households would yield the highest return per dollar.

Comparing the predictions of the model with income effects to those of the baseline
model, there is some noticeable polarization in the results. Once the effects of income are accounted for, richer households are predicted to have a higher propensity to invest (both in larger amounts and in terms of likelihood), and thus the overall impact of a subsidy available to all households is magnified. On the other hand, households with a lower level of income are now predicted to have a lower propensity to invest than in the baseline model, and hence the impact of a means-tested policy is mitigated. However, the predicted return per dollar for both policies remains above 100%, and thus a subsidy aimed at defraying the cost of home improvement is still predicted to be successful at leveraging private investment in housing renovation.

A.4 Robustness Checks: Specification of Sale Decision

As an additional test to assess the results’ sensitivity to model misspecification, here we examine a variation of the model with an alternative utility function specification for households that decide to sell their property. The results of this exercise show that the specification in the baseline model yields a more conservative estimate of the effects of a policy subsidizing investment in home improvement.

A.4.1 Model

In the baseline model, the choice-specific utility function $u(q, P | d)$ was comprised of a common component $q^{\theta_1} P^{\theta_2}$ aimed at capturing the utility that households received from the consumption of a housing good with quality $q$ in a neighborhood of value $P$. Thus, the utility function of households deciding to sell their property was specified as $u(q, P | d = 3) = q^{\theta_1} P^{\theta_2} - \theta_5(q \cdot P) - FC_3$, where the term $\theta_5(q \cdot P)$ captures the costs of selling proportional to the housing value and $FC_3$ accounts
for the fixed cost of selling. This functional form can be interpreted as households receiving a payoff from the sale of their property that is proportional to the utility that can be generated by the existing property features, $q$ and $P$, at the time of sale.

Alternatively, the utility accruing to households that decide to sell their property can be written as a discounted present value of the EMV of the property at the time of sale, since $EMV = q \cdot P$. The household’s utility function can then be written as:

$$u(q, P | d) = \begin{cases} 
q^{\theta_1} P^{\theta_2}, & \text{for } d = 1; \\
q^{\theta_1} P^{\theta_2} - \theta_3 i^{\theta_4} - FC_2, & \text{for } d = 2; \\
\theta_7 (q \cdot P)^{\theta_8} - FC_3, & \text{for } d = 3.
\end{cases}$$

where the parameters $\theta_7$ and $\theta_8$ are added to translate the value of the property into utility terms. The choice specific value function for households that decide to sell their property can then be written as:

$$W_t(q, P) = \theta_7 (q \cdot P)^{\theta_8} - FC_3 + \beta W_{t+1}(q, P)$$

### A.4.2 Estimation

The structural model parameters can be estimated via maximum likelihood by specifying a log-likelihood function analogous to that outlined in Section 1.6. The results of this estimation routine are reported in Table A.6.
Table A.6: Comparison of Estimates for Structural Model Parameters

| Parameter | Baseline Model Estimates | Standard Errors | Updated u(q, P|d = 3) Parameter Estimates | Standard Errors |
|-----------|--------------------------|-----------------|----------------------------------------|----------------|
| **Utility Parameters** | | | | |
| Housing Quality θ₁ | 1.657 (0.694) | | 1.277 (0.286) | |
| Neighborhood Quality θ₂ | 0.433 (0.104) | | 0.267 (0.071) | |
| **Investment Cost Parameters** | | | | |
| Linear Cost of Investment θ₃ | 0.041 (0.009) | | 0.019 (0.002) | |
| Curvature on Investment θ₄ | 1.649 (0.413) | | 1.412 (0.682) | |
| Fixed Cost of Investing FC₂ | 78.889 (139.752) | | 118.540 (188.136) | |
| **Sale Parameters** | | | | |
| Linear Sale Cost θ₅ | 0.010 (0.185) | | - | - |
| Linear Sale Price θ₇ | - | - | 2.308 (211.304) | |
| Curvature on Sale Price θ₈ | - | - | -0.236 (62.318) | |
| Fixed Cost of Selling FC₃ | -1.029 (20.360) | | -1.581 (314.164) | |
| **Variance** | | | | |
| Tase Shock σₑ | 37.370 (32.111) | | 56.247 (16.915) | |
| Measurement Error σₑ | 1.164 (1.885) | | 1.166 (1.898) | |

A.4.3 Results

Figure A.7 illustrates a comparison of the two models’ predicted choice probabilities and optimal investment. The predictions for the two models are remarkably similar, especially so with respect to the probability of investing and optimal investment as a function of neighborhood quality, where the predicted policies overlap. While there are relatively minor discrepancies in the predicted probability of selling and optimal investment as a function of housing quality, these differences tend to occur at the extremes of the distribution of q and P, and thus affect a relatively small number of observations.

Figure A.8 illustrates the predicted effects of a τ = 10% subsidy aimed at defraying the cost of investing in home improvement. A subsidy available to all households
Figure A.7: Predictions of Baseline Model and Model with Income Effects

Note: This figure illustrates the predicted probability of investing in home improvement (top row), the probability of selling (middle row) and the optimal amount of investment (bottom row) as a function of housing quality, $q$ and neighborhood values, $P$. The line in red depicts the predictions of the baseline model, while the line in blue depicts the predictions of the model with an alternative specification for $u(q, P|d = 3)$. 

139
Note: This figure illustrates predicted optimal investment policy (top row) and probability of selling (bottom row) under different policy scenarios. The blue line depicts the prediction for the baseline model, while the red lines depict the model prediction under a policy allowing households to reduce investment cost by 10%. Solid red lines represent the prediction if this policy is applied to all households, and dashed red lines represent the prediction for a means-tested policy, available only to households earning under $75,000 per year. Predictions are based on 2016 data.

would have a significant impact on the predicted level of investment and a somewhat more pronounced effect than that predicted by the baseline model (see Figure 1.11). Means-testing this policy, making the subsidy available only to households earning less than $75,000 per year, is again predicted to perform well at increasing the optimal amount of investment at the bottom of the distribution of neighborhood quality, where a larger percentage of households would be eligible for the subsidy. The predicted likelihood of selling a property is not predicted to change under either policy scenario, suggesting that these policies would not promote investment while
Table A.7: Total Investment Under Alternative Policy Scenarios

<table>
<thead>
<tr>
<th>Model</th>
<th>Total Investment</th>
<th>Total Cost</th>
<th>Change in Investment</th>
<th>Returns per Dollar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No policy ($\tau = 0$)</td>
<td>$34,795,751</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$)</td>
<td>$45,930,716</td>
<td>$4,593,072</td>
<td>$11,134,965</td>
<td>$2.42</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$), Means-Tested</td>
<td>$42,286,104</td>
<td>$2,869,570</td>
<td>$7,490,353</td>
<td>$2.61</td>
</tr>
<tr>
<td>**Updated $u(q, P</td>
<td>d = 3)$**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No policy ($\tau = 0$)</td>
<td>$33,151,574</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$)</td>
<td>$47,755,489</td>
<td>$4,775,549</td>
<td>$14,603,916</td>
<td>$3.06</td>
</tr>
<tr>
<td>Policy ($\tau = 10%$), Means-Tested</td>
<td>$42,999,379</td>
<td>$2,924,650</td>
<td>$9,847,805</td>
<td>$3.37</td>
</tr>
</tbody>
</table>

Note: This table was constructed by calculating and aggregating the predicted level of investment in home improvement for each household under alternative policy scenarios, using 2016 data for Minneapolis.

Table A.7 summarizes the policy impacts predicted by the model under the alternative specification of $u(q, P|d = 3)$ compared to the baseline model. Under the alternative specification, households are more responsive to changes in the marginal cost of investing, and thus a $\tau = 10\%$ subsidy is predicted to have a greater impact on aggregate investment under either policy scenario. The returns per dollar spent are as high as $3.06 for a subsidy available to all households, and $3.37 for a means-tested policy, compared to $2.42 and $2.61 respectively under the baseline model. Thus, while the two specifications yield similar predictions, the baseline model represents a more conservative estimate of the effects of a policy aimed at encouraging investment in home improvement.
Appendix B

Appendix to Chapter 2

Figure B.1: Bowtie Threshold for Empirical Derivative Calculation

Note: At the center of the circle is a construction "event", and all properties within the circle constitute neighborhood $l$. The empirical derivative is calculated at the location of each housing transaction $\iota$, marked by "X" in the figure, by matching houses on the proximal and distal side of the transaction according to their distance from $\iota$. Three house pairs ($\kappa_n = 3$) are displayed in the picture. A weighted average of the price difference constitutes my empirical derivative at location $\iota$ in neighborhood $l$, $Y_{\iota,l}$. Source: Diamond and McQuade (2019)
Appendix C

Appendix to Chapter 3

C.1 Data Appendix

C.1.1 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 C.1 based on information obtained from the Twin Cities Metro Transit.
### Table C.1: Summarized Transit Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Direction</th>
<th>Route Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closest 400 Route</td>
<td>Eagan</td>
<td>515, 535, 552, 553, 554, 558, 578, 579, 587, 588, 589, 597.</td>
</tr>
<tr>
<td>Closest 600 Route</td>
<td>Minnetonka, Eden Prairie</td>
<td>721, 724, 742, 747, 755, 756, 758, 760, 761, 762, 763, 764, 765, 766, 767, 768.</td>
</tr>
<tr>
<td>Closest 700 Route</td>
<td>Plymouth, Maple Grove, Brook Park</td>
<td>821, 825, 850, 852, 854, 860, 887, 888.</td>
</tr>
<tr>
<td>Closest 800 Route</td>
<td>Coon Rapids</td>
<td>111, 113, 114, 115, 118, 121, 122, 129.</td>
</tr>
<tr>
<td>Closest UofM Route</td>
<td>University of Minnesota</td>
<td>61, 67, 74, 94.</td>
</tr>
</tbody>
</table>

Source: Twin Cities Metro Transit

---

![Regional Route Renumbering Areas](image-url)
### C.1.2 Amenities Data

Table C.2 reports the NAICS codes and the average number of observations per year for the amenity categories included in our analysis obtained from ReferenceUSA.

<table>
<thead>
<tr>
<th>NAICS Category</th>
<th>Avg. Obs per Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>22, 562</td>
<td>1,948</td>
<td>Utilities and Waste Management and Remediation Services</td>
</tr>
<tr>
<td>442</td>
<td>4,838</td>
<td>Furniture and Home Furnishings Stores</td>
</tr>
<tr>
<td>443</td>
<td>5,662</td>
<td>Electronics and Appliance Stores</td>
</tr>
<tr>
<td>444</td>
<td>7,271</td>
<td>Building Material and Garden Equipment and Supplies Dealers</td>
</tr>
<tr>
<td>445</td>
<td>13,065</td>
<td>Food and Beverage Stores</td>
</tr>
<tr>
<td>446</td>
<td>6,713</td>
<td>Health and Personal Care Stores</td>
</tr>
<tr>
<td>447</td>
<td>2,830</td>
<td>Gasoline Stations</td>
</tr>
<tr>
<td>448</td>
<td>10,931</td>
<td>Clothing and Clothing Accessories Stores</td>
</tr>
<tr>
<td>451</td>
<td>7,061</td>
<td>Sporting Goods, Hobby, Musical Instrument, and Book Stores</td>
</tr>
<tr>
<td>452, 453</td>
<td>18,007</td>
<td>General Merchandise Stores, Miscellaneous Store Retailers</td>
</tr>
<tr>
<td>481</td>
<td>532</td>
<td>Air Transportation</td>
</tr>
<tr>
<td>541</td>
<td>162,156</td>
<td>Professional, Scientific, and Technical Services</td>
</tr>
<tr>
<td>55</td>
<td>888</td>
<td>Management of Companies and Enterprises</td>
</tr>
<tr>
<td>61</td>
<td>17,909</td>
<td>Educational Services</td>
</tr>
<tr>
<td>621</td>
<td>130,663</td>
<td>Ambulatory Health Care Services</td>
</tr>
<tr>
<td>622</td>
<td>1,496</td>
<td>Hospitals</td>
</tr>
<tr>
<td>623</td>
<td>4,231</td>
<td>Nursing and Residential Care Facilities</td>
</tr>
<tr>
<td>711</td>
<td>6,040</td>
<td>Performing Arts, Spectator Sports, and Related Industries</td>
</tr>
<tr>
<td>712</td>
<td>2,077</td>
<td>Museums, Historical Sites, and Similar Institutions</td>
</tr>
<tr>
<td>713</td>
<td>4,982</td>
<td>Amusement, Gambling, and Recreation Industries</td>
</tr>
<tr>
<td>721</td>
<td>2,899</td>
<td>Accommodation</td>
</tr>
<tr>
<td>722310, 722320</td>
<td>2,003</td>
<td>Food Service Contractors, Caterers</td>
</tr>
<tr>
<td>722410, 722515</td>
<td>6,019</td>
<td>Drinking Places (Alcoholic Beverages), Snack and Nonalcoholic Beverage Bars</td>
</tr>
<tr>
<td>722511</td>
<td>26,799</td>
<td>Full-Service Restaurants</td>
</tr>
<tr>
<td>722513, 722514</td>
<td>847</td>
<td>Limited-Service Restaurants, Cafeterias, Grill Buffets, and Buffets</td>
</tr>
<tr>
<td>812</td>
<td>27,770</td>
<td>Personal and Laundry Services</td>
</tr>
<tr>
<td>813</td>
<td>34,529</td>
<td>Religious, Grantmaking, Civic, Professional, and Similar Organizations</td>
</tr>
<tr>
<td>92</td>
<td>12,029</td>
<td>Public Administration</td>
</tr>
</tbody>
</table>
### C.1.3 Population Changes by Neighborhood

Table C.3: Population Change in Census Tracts Adjacent to the Blue Line, 2000-2010

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

Source: US Census Bureau.

### C.2 Gradient Boosting with Regression Trees

An alternative possible avenue generate a predictive model of housing prices is to use the gradient boosting method proposed in [Friedman (2001)](https://www-stat.stanford.edu/~jhf/ftp/gbc0210.pdf). This approach builds up an estimate of \( F(X_{it}) = f(X_{it}) + E[\epsilon_{it}|X_{it}] \) by using functional gradient descent to iteratively improve the performance of regression function. The key idea is to take:

\[
\hat{F} = \arg\min_F E_{p,x} [L(p, F(X))]
\]
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:

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

To reduce variance, we approximate the function $r = -\nabla L(p, F_{m-1}(X))$ 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 $\gamma_m$ 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.\footnote{For a recent approach at extending regression forests to the case of instrumental variables, see Athey et al. (2019).}

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.

\section*{C.2.1 Results Using 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 treatment and control groups using concentric circles with a radius of 0.5 miles and 1 mile respectively, 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 $\ell_2$ regularization weight.
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. Of the 198 covariates included in our analysis, 132 were estimated to have a non-zero effect. Table C.4 shows the top 20 covariates, ranked by their contribution to the reduction in mean-squared error. Unsurprisingly, the total area of a property 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. Sev-

<table>
<thead>
<tr>
<th>Feature</th>
<th>Gain</th>
<th>Cover</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building Area</td>
<td>0.120</td>
<td>0.081</td>
<td>0.072</td>
</tr>
<tr>
<td>Ground Floor Area</td>
<td>0.107</td>
<td>0.040</td>
<td>0.036</td>
</tr>
<tr>
<td>No. of Bathrooms</td>
<td>0.068</td>
<td>0.022</td>
<td>0.012</td>
</tr>
<tr>
<td>Case-Shiller Index</td>
<td>0.060</td>
<td>0.041</td>
<td>0.035</td>
</tr>
<tr>
<td>No. of Restaurants (within 0.5 miles)</td>
<td>0.040</td>
<td>0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>Minimum Distance Route 12</td>
<td>0.039</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>Second Floor Area</td>
<td>0.037</td>
<td>0.035</td>
<td>0.031</td>
</tr>
<tr>
<td>Minimum Distance Route 4</td>
<td>0.035</td>
<td>0.013</td>
<td>0.010</td>
</tr>
<tr>
<td>Age of House</td>
<td>0.035</td>
<td>0.047</td>
<td>0.038</td>
</tr>
<tr>
<td>Percent College</td>
<td>0.022</td>
<td>0.009</td>
<td>0.006</td>
</tr>
<tr>
<td>Minimum Distance Route 32</td>
<td>0.022</td>
<td>0.009</td>
<td>0.005</td>
</tr>
<tr>
<td>No. of Bedrooms</td>
<td>0.020</td>
<td>0.016</td>
<td>0.013</td>
</tr>
<tr>
<td>Minimum Distance Route 46</td>
<td>0.020</td>
<td>0.020</td>
<td>0.013</td>
</tr>
<tr>
<td>Housing Stock</td>
<td>0.016</td>
<td>0.006</td>
<td>0.004</td>
</tr>
<tr>
<td>Minimum Distance Route 27</td>
<td>0.015</td>
<td>0.022</td>
<td>0.014</td>
</tr>
<tr>
<td>Average Commute</td>
<td>0.014</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Minimum Distance Route 21</td>
<td>0.012</td>
<td>0.031</td>
<td>0.018</td>
</tr>
<tr>
<td>Minimum Distance Route 22</td>
<td>0.011</td>
<td>0.046</td>
<td>0.026</td>
</tr>
<tr>
<td>Percent High School</td>
<td>0.010</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>Finished Basement</td>
<td>0.010</td>
<td>0.022</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Note: 132 of 198 covariates had nonzero gain.
eral transit routes were significant as well, including the minimum distance to bus stops along routes 12 and 4. These routes connect downtown Minneapolis with its wealthier suburbs, and so it is unsurprising that they are important determinants of house prices. Finally, several demographic characteristics were significant, including the percent of college graduates in a census block and the average commute time.

Table C.5 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, 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 3.4 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

<table>
<thead>
<tr>
<th>Predicted Residual:</th>
<th>Pre-Treatment</th>
<th>Post-Treatment</th>
<th>Implied Spillover</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.005</td>
<td>0.071</td>
<td>0.033</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>-</td>
</tr>
</tbody>
</table>

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 3.2 (0.104).
predicts prices will increase by 10.4% in treatment neighborhoods. This increase can be thought 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 interpreted 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 (Table C.5) on neighborhood attributes. This type of regression allows us to explore how the direct effect of the Blue Line varies with respect to neighborhood characteristics, although it does not allow us to do the same for the indirect effect. Table C.6 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 increase in the share of white residents by 10 percentage points translates to a 3.3% 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 Line introduction. This is unsurprising as these neighborhoods can be expected to have a lower MWTP for 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.
<table>
<thead>
<tr>
<th>Predicted Residual:</th>
<th>Intercept</th>
<th>Distance to BL</th>
<th>% White</th>
<th>% Driving</th>
<th>Median Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.24</td>
<td>0.09</td>
<td>0.33</td>
<td>-0.60</td>
<td>-0.012</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>(0.07)</td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.09)</td>
<td>(0.006)</td>
</tr>
</tbody>
</table>

### C.3 Efficient Boosted Smooth Trees

There are two main difficulties to using smooth trees for gradient boosting. First, for each branch split we test, we need to re-regress the residual on the matrix of leaf node probabilities. The time complexity of this regression is $O(C^2 N)$, 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 $O(C^2 N^2)$. 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 $O(K N)$, where $K$ is the number of instruments. Repeating this multiplication $N$ times yields an asymptotic rate of $O(K N^2)$. The purpose of this appendix is to propose 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 which 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:

$$
G(\tilde{y}, \tilde{P}_t, \tilde{P}_{t-1}, \lambda) = \frac{1}{(1 - \lambda)} \left( R(\tilde{y}, \tilde{P}_{t-1}, \lambda) - R(\tilde{y}, \tilde{P}_t, \lambda) \right)
$$

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

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

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

where

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

Therefore, our inverse becomes:

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

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

The residual can be expressed as:

\[ \tilde{y} \tilde{P}_l(\tilde{P}_l \tilde{P}_l + \lambda I)^{-1} \tilde{P}_l \tilde{y} = \]

\[ y'B'(B'B + \lambda I)^{-1}B'y + \frac{1}{\mu} y'B'(B'B + \lambda I)^{-1}B'A'A'B'(B'B + \lambda I)^{-1}B'y \]

\[ -\frac{2}{\mu} y'A'A'B'(B'B + \lambda I)^{-1}B'y \]

\[ \frac{1}{\mu} y'A'y \]

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:

\[ G(\tilde{y}, \tilde{P}_l, \tilde{P}_{t-1}, \lambda) = \frac{1}{\mu} (y'B'(B'B + \lambda I)^{-1}B'A - y'A)^2 \]

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. This approach is 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:

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

Assume that the \( k \)th regressor, \( X_k \), is sorted from smallest to largest, and that \( Z_k \) is sorted based on the ordering of \( X_k \). The goal is to test all elements of \( X_k \) to find the split that maximizes the gain. At iteration 1, we have:

\[
A_1 = Z'(P'_{t-1,j}L(X_k, X_1)) 
= \sum_i Z_i P'_{ij} \left( 1 - \frac{1}{2} \frac{X_{1k}}{2^{X_{ik}}} \right) 
\]

This can be broken into two parts:

\[
\tilde{Z}_1 = \sum_i Z_i P'_{ij}^{-1}
\]

and

\[
\tilde{Z}_1 = \frac{1}{2} \sum_i Z_i P'_{ij}^{-1} \frac{2^{X_{ik}}}{2^{X_{ik}}}
\]

\footnote{This only needs to be done once at the start of the algorithm for all regressors.}
This defines:

\[ Z_1^r = \tilde{Z}_1 - \check{Z}_1 \]

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:

\[ \tilde{Z}_m = \tilde{Z}_{m-1} - Z_{m-1} P_{m-1,j}^{t-1} \]

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

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

and

\[ A_m = \tilde{Z}_m - \tilde{Z}_m + Z_m^l \]

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 $O(NK)$.

C.3.1 Monte Carlos

Since we do not currently have a proof for consistency of the Boosted Smooth Trees estimation routine, we provide a small Monte Carlo simulation to demonstrate the efficacy of our method. We used $N = 1,000$ observations with $Z_i, \epsilon_i \sim N(0, 0.5^2)$ and $X_i = Z_i + \frac{1}{2} \epsilon_i$ in order to simulate the effects of endogeneity. The dependent regressor is determined by the following nonlinear relationship:

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

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 $\lambda$. Figure C.1 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 C.2 shows the non-parametric estimates using 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 C.1: Parametric Estimation

(a) OLS

(b) IV
Figure C.2: Endogenous Regressor