# Housing Markets, Subsidies and the Economic Effects of Infrastructure Investments

A Dissertation
Presented to
The Academic Faculty

By

Juan Pablo Uribe

BA, Universidad de Los Andes, 2011
MA, Universidad de Los Andes, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in the Department of Economics at Brown University

Providence, Rhode Island
May 2022

This dissertation by Juan Pablo Uribe is accepted in its present form
by the Department of Economics as satisfying the
dissertation requirements for the degree of Doctor of Philosophy.

Date ____________ ______________________________

Andrew Foster, Advisor

Recommended to the Graduate Council

Date ____________ ______________________________

Matthew Turner, Advisor

Date ____________ ______________________________

Jesse Shapiro, Advisor

Date ____________ ______________________________

Kenneth Chay, Reader

Approved by the Graduate Council

Date ____________ ______________________________

Andrew G. Campbell,

Dean of the Graduate School

## CURRICULUM VITAE

Juan Pablo Uribe was born and raised in Bogotá, Colombia. Prior to Brown, where started his Ph.D in 2015, he worked as a consultant investigating recent inequality trends in Latin America at the World Bank and the Inter-American Development Bank (IDB). He received a bachelor's in 2011 and master's degree in 2013 both from Universidad de Los Andes in Bogotá, Colombia.

# ACKNOWLEDGMENTS

TBD

# Introduction

The central topic of this dissertation is understanding how large-scale government intervention can shape markets and the human behavior that drives them. The different chapters of this thesis investigate the behavioral and market responses to government programs designed to increase access to affordable housing, utilities and internet.

The first chapter investigates how the Colombian housing market responds to a subsidy scheme designed to increase homeownership and increase new housing construction for low income households. Housing affordability is a pressing issue across the globe. In the past several decades, many governments adopted market-oriented strategies to encourage homeownership. Common strategies include down payment assistance, low interest rates, and subsidies to developers to build affordable housing. The assumption is that a market-oriented strategy is more efficient than direct government intervention. While policy-makers adopt these approaches on that premise, little is known about how the market actually responds to these interventions. It is difficult to create accurate models that can isolate the policy's effects on the housing market and the human behavior that drives it. This is due to several empirical and theoretical challenges. I found an ideal setting in Colombian housing policy to address these challenges. The policy is designed to promote housing for low-income families, but its design is similar to some first-time home buyer programs in the United States and Europe. In this setting, I am able to evaluate the housing market response to a housing price-capped policy. This price cap allows me to study the behavioral responses induced by the policy. Because the policy offers subsidies and tax incentives to both developers and households, I am able to adapt a model to untangle the supply and demand responses to this policy. The massive expansion of this policy over my study period allows me to provide compelling evidence that the behavioral responses I observe are actually caused by the policy.

Using unique data containing the universe of new construction projects and administrative records from the Ministry of Housing, I find bunching at the price cutoff. Additionally, as the subsidy increases its eligibility and size, the market response is more pronounced. The market share of units sold at the price cutoff increased from one percent to seven percent of the market. I employ techniques used in the bunching literature to estimate the distribution that would exist in the absence of the subsidy. Using this distribution, I find that the price cutoff combined with the tax incentive and subsidies is distorting the incentives of developers and households, who build and buy smaller housing units to comply with the maximum price that defines eligibility. They build or buy units up to 30 percent smaller to benefit from the policy scheme.

I introduce a housing equilibrium model that allows for product differentiation and agent heterogeneity and a novel identification strategy to recover the model primitives. The model rationalizes the market response. The identification relies on the behavioural responses induced by the subsidy. I use marginal conditions and the estimated behavioural effects to estimate the parameters describing the cost and utility functions of the model. The economic model and estimated parameters allow for an evaluation of the policy's welfare effects. The model also allows for the study of potential impacts of alternative policies such as removing the price threshold, imposing a minimum size limit, or removing tax benefits.

Focusing on the agents responding to the policy, I find that there are some efficiency losses associated with the subsidies and that the government expenditure seems to be higher than the benefits to households, even excluding the efficiency costs. On the developer side, removing the tax incentives will generate a provisioning problem, particularly by the end of my study period. Developers would have a drop in their profits that would make them worse off than in a scenario without subsidies, which could lead to an excess demand for low-cost housing. However, these government expenditure is considerable an the government ends up subsidizing many developers that would produce low cost housing even in the absense of the subsidy scheme. The results of these paper shows that the policy design could be improve to be more effective.

The paper makes important methodological contributions to the bunching and hedonic equilibrium models literature. It also provides new empirical and theoretical insights on a first-order question. The findings of this paper inform the design of an effective housing policy, which is fundamental to providing affordable housing.

The second chapter studies the effects of a location-based redistribution policy on the housing market in Bogotá, Colombia. The policy was designed to ease the financial burden of utility payments for low-income homeowners and renters. In Bogotá, like many cities in the developing world, it is difficult to identify low-income residents, because many of them do not participate in formal labor markets. In some cases, there are no pay stubs or tax returns to assist in identifying the people who would benefit from welfare programs. To address this challenge, policymakers in Bogotá chose to use neighborhood quality as a proxy for income. Each neighborhood in the city was assessed and given a neighborhood quality index. Residents of neighborhoods with high quality scores pay above the market rate for utilities and neighborhoods with low quality scores are subsidized and therefore pay less. While it was designed as a redistributive welfare program, its design induced unintended effects on the housing market. Using a Regression Discontinuity Design, I find that the difference in subsidy levels across neighborhoods induces new construction in highly subsidized areas. As predicted by the theory, housing prices are higher in these areas, which ends up canceling out the utility subsidy for renters.

This research is important because it provides rigorous empirical evidence to support the urban economist mantra *subsidize people, not places*. This type of evidence is required to inform the renewed interest in using location as a targeting tool to implement redistributive policies and welfare programs. Understanding the interaction between housing markets and redistributive policies has implications in many settings. Similar research designs may be used to understand location-based policies like school redistricting, the placement of green energy infrastructure, and the designation of communities as economic Opportunity Zones. More generally, my work also underscores the importance of accounting for potential unintended consequences when evaluating the potential of a policy.

In the last chapter, my coauthor and I explore a governmental investment in internet expansion's impact on educational outcomes. We use variation over time and an instrumental variable approach exploiting the costliness of extending the existing internet infrastructure to connect new areas. We then identify the causal impacts of internet access on test scores. A natural assumption would be that internet expansion is correlated with improved educational outcomes. The existing literature on this topic does not find this to be the case. In our setting, we find the same. However, when we disaggregate educational outcomes of students by achievement, we find that outcomes of the lowest performing students actually do improve. This work underscores the importance of not just interrogating popular beliefs and assumptions, but also common findings in the literature.

Underpinning my thesis is a desire to understand the effects of public investments in infrastructure and be able to do a welfare analysis of its economic impacts. Across various large-scale investments and subsidy programs, I find behavioral responses induced by those policies that may lead to unintended consequences that sometimes mitigate the intended effects and others amplified the welfare effects.

# Chapter 1

# Equilibrium Effects of Housing Subsidies: Evidence from a Policy Notch in Colombia

**Abstract**

This paper studies how the housing market in Colombia responds to policies that aim to increase homeownership among low-income households. Private sector developers received tax incentives to build houses priced below a cutoff, and households received subsidies to buy houses below the same cutoff. To benefit from the policy, households change their housing consumption and spend less on housing, bunching at the cutoff. To rationalize this response, the paper models an equilibrium between heterogeneous developers building differentiated housing and heterogeneous households buying them. The model is estimated with the moments recovered by comparing the distribution with bunching to a counterfactual distribution estimated using the techniques from the bunching literature. I use the model and estimated parameters to evaluate the policy. I calculate the efficiency cost induced by the notched subsidy scheme and I show that without supply-side incentives, developers may exit the market; their profits would be up to 14 percent lower. However, the existence of these tax incentives artificially increases the profits of developers who would build low-cost housing even in their absence.

## 1.1 Introduction

Many governments invest significant resources through subsidies or tax incentives to provide housing solutions for low-income households and promote homeownership. Approaches such as the mortgage interest deduction (MID) aim to encourage homeownership through tax incentives. However, they raise concerns because they primarily benefit the rich and there is little evidence that they increase homeownership (E. Glaeser & Shapiro, 2003; Gruber et al., 2021; OECD, 2021a). Alternative strategies include downpayment assistance, subsidized interest rates, and subsidies to developers to build affordable housing. We know little about the market effects of these alternative approaches, which face the same concerns as MID tax incentives. To understand the effectiveness of these policies, we need to know the market responses, who benefits from them, and the potential inefficiencies of the subsidies.

Studying the market response and recovering the structural parameters of a model that allows a welfare evaluation of these policies is challenging. Valid counterfactuals to a housing market without subsidies are rare. Additionally, a reduced form analysis would be insufficient to understand the effectiveness of the policies. We need to disentangle demand and supply responses and have a model to interpret them. Identifying market effects in markets with heterogeneous agents and differentiated products, such as the housing market, is particularly challenging.[1] A model that does not allow product differentiation could not account for changes in the type of housing built and consumed, which could be a relevant response to the government incentives.

Hedonic equilibrium models are a common approach to model differentiated product markets and have been widely used to model housing markets. However, existing identification approaches for estimating structural parameters rely on strong assump-

[1] Zoutman, Gavrilova, and Hopland (2018) shows that a single tax or subsidy can help to identify supply and demand responses in a market with homogeneous goods and agents.Implementing this approach to a market with differentiated products will require multiple instruments.

tions.[2] Few papers actually attempt to estimate structural parameters. Most of those that do focus on estimating hedonic regressions that provide equilibrium marginal willingness to pay (MWTP) for different housing characteristics (Greenstone, 2017). Although MWTP estimates can be informative, they do not allow for non-marginal policy evaluations or counterfactual policy analysis.

This paper studies the social housing policy in Colombia, which combines subsidies and tax incentives for developers and households buying and building low-cost housing. The policy design allows me to overcome the empirical challenges associated with the evaluation of housing subsidies. Low-cost housing is defined using a market price cut-off of 135 times the monthly minimum wages (roughly $USD\,40,000$). This cutoff introduces notches, or discontinuous incentives, on both the supply and demand sides. This triggers bunching at the cutoff. I use this notch and the variation of the notch size overtime to provide evidence of the market response to these subsidies. I propose a model that rationalizes the observed equilibrium, and I integrate the bunching and hedonic equilibrium literature to propose a method to identify and estimate the structural parameters of the model.[3] The model and estimated parameters are used to evaluate the effectiveness of the Colombian policy design and to evaluate counterfactual policies.

Between 2006-18, the policy expanded, doubling the subsidy amount and the number of households receiving it. To show the market response, I combine data from a

[2]There are three main identification approaches in these types of models: (i) Excluded instruments and variation across markets (Epple, 1987; Brown & Rosen, 1982; Wooldridge, 2010), (ii) Functional forms and inversion methods (Bajari & Benkard, 2005; Yinger, 2015; Bishop & Timmins, 2019), (iii) Non parametric identification and single index reduction (Ekeland, Heckman, & Nesheim, 2004; Heckman, Matzkin, & Nesheim, 2010; Chernozhukov, Galichon, Henry, & Pass, 2021; Epple, Quintero, & Sieg, 2020). For more details see Chernozhukov et al. (2021). An approach that integrates the hedonic insights into a discrete choice framework is Bayer, Ferreira, and McMillan (2007) or Anagol, Ferreira, and Rexer (2021).

[3]This is not the first paper suggesting to use bunching to identify hedonic or sorting equilibrium models. Kuminoff, Smith, and Timmins (2013, p.1009) wrote: "Equilibrium sorting models provide the means to implement both the original Blinder and Rosen (1985) idea and the Saez (2010) test and extend them for policies that target public goods or other amenities that affect agents differently." However, I am not aware of any paper that actually implements this approach.

construction census containing the universe of new housing developments between 2006 and 2018 and administrative records for the subsidies awarded from the Ministry of Housing.

I show strong evidence of bunching at the cutoff. Following the bunching literature, I estimate a counterfactual distribution of market shares by price to recover the behavioral responses induced by the subsidy. The households that change their housing consumption to receive the subsidy spend up to 85 percent less in housing to take advantage of the subsidy. Given the equilibrium prices, this is translated to a housing unit up to 90 percent smaller. Using the variation in the subsidy over time, I show that increasing government expenditure on the policy increases the share of units sold at the cutoff defining low-cost housing. The fact that the bunching amount increases as the generosity of the subsidies increases demonstrates that Colombia's social housing policy matters a lot and may provide credible identification of market structure.

During my study period, an interest rate subsidy was introduced, the downpayment subsidy increased, and eligibility expanded. Households received around 13 percent of the price of the house at the cutoff in 2006 and around 24 percent in 2018. As a result, the excess mass, or bunching, sold at the price cutoff increases from around 3 percent of the market share around 2006 to about 16 percent by 2018. I provide suggestive evidence that housing characteristics and in particular, housing size drive the behavioral responses resulting in the bunching.

To rationalize the observed equilibrium responses, I introduce and estimate a competitive housing market equilibrium model. The model includes the policy-induced notch, to a hedonic – or sorting – equilibrium model.[4] Households are heterogeneous in income, developers in productivity level, and housing in size. I use the model to show how the notch creates incentives for developers and households to bunch at the thresh-

[4] For a review of the general approach, see Kuminoff et al. (2013) or Greenstone (2017). For recent applications, see Epple et al. (2020) and Chernozhukov et al. (2021).

old. Like in the observed equilibrium, buyers and developers in the model change the type of units they buy and build to take advantage of subsidies, and consequently the equilibrium density has bunching at the cufoff point.

I propose an identification strategy based on a two-step procedure suggested by S. Rosen (1974). The first step follows the standard practices in the literature to estimate the implicit price function for housing size and to use the reduced-form estimates from the first part of the paper. The main innovation of my paper is in the second step. In the second step, I use the discontinuity and estimated behavioral responses and adapt the identification strategy proposed in the literature using notches to estimate the structural parameters. By comparing the counterfactual distribution with the observed distribution, I learn about the trade-offs between developers and households. Using the model, I show that there is a marginal buncher who is indifferent to receiving the subsidy but consuming less housing and not receiving the subsidy and consuming their optimal housing type. This insight of the model allows me to use the counterfactual distribution to recover two points on the same indifference curves for the marginal buncher.[5] Using the parameters of the first step and the marginal buncher indifference condition, I can estimate the shape of the utility and cost functions. Because the policy scheme has one subsidy targeted to developers and one for households using the same cutoff, I have three different prices. I have the market price, the price received by developers and the price paid by households. This allows me to use the same logic that I used to estimate the demand parameters to estimate the parameters describing developers' marginal costs.[6]

[5] Best, Cloyne, Ilzetzki, and Kleven (2019) use the same identification idea to estimate the intertemporal elasticity of substitution from the behavioral responses induced by notches in the interest rates for loan refinancing. Other examples are Einav, Finkelstein, and Schrimpf (2015) and Z. Chen, Liu, Suárez Serrato, and Xu (2021) or Kleven and Waseem (2013). Bertanha, McCallum, and Seegert (2021) and Blomquist, Newey, Kumar, and Liang (2017) discuss how in contrast with changes in the slope, or kinks, notches allow to recover structural parameters.

[6] To estimate the shape of the indifference curve and offer curve, I impose functional forms for the utility function and a cost function. The utility function is a CES utility function depending on consumption on housing and consumption on other goods, and the cost function depends on housing size and

The model and estimated parameters are used to evaluate how marginally subsidized households and developers benefit from the subsidy scheme. On the demand side, I compare the utility levels of the marginally subsidized households in two counterfactual scenarios. In the first counterfactual scenario, households do not get subsidies. Marginally subsidized households, that reduce their housing consumption to benefit from the subsidies could be better of if they receive the money without a restriction on the cost of the house. I calculate this welfare loss associated with the policy design. Quantifying this is relevant to assess if a notched policy design is better than a linear subsidy. H. S. Rosen (1985) show than depending on the elasticity of substitution, notched policy designs may be more effective than linear incentives at targeting subsidies. My structural parameters suggest an elasticity of substitution between housing and consumption of other goods of around higher than one. Therefore, housing and consumption of other goods are gross substitutes.

I compare the observed equilibrium with a counterfactual scenario with subsidized households but without tax refunds for developers. Developers would be worse off. Between 2006 and 2009, the profits for marginally subsidized developers would be 5 percent lower, and by 2016, after the subsidy's expansion, their profits would be 14 percent lower. The marginally subsidized developers have higher marginal costs when producing at the price cutoff. They are competing with more productive developers, which have profits when building low-cost housing even without subsidies. A price increase, desirable for both developer types, will lead to non-eligibility for the subsidies. Because of the price cap, marginally subsidized developers need tax incentives to build low-cost housing, and without them, the market may face a rationing problem. The existence of these tax incentives can prevent the exit of the marginally subsidized developers, however they artificially increase the profits in more than 5 pecent for developers that would produce low-cost housing even in the absence of these

the number of units built. I observed equilibrium relationships non parametrically.

incentives.

### *Contributions and Related Literature*

I make several methodological and empirical contributions. My first contribution is to provide additional evidence of bunching in the housing market, which is a relatively unexplored setting. Carozzi, Hilber, and Yu (2020) provide evidence of bunching in response to a similar housing policy in the United Kingdom. McMillen and Singh (2020) show that apartment rents cluster at values near the fair market rent in Los Angeles, California. There is also evidence of bunching in the density of mortgages with notches in the interest rate schedule.[7] This paper complements the bunching evidence around a price cutoff in the housing market.

The main contribution of this paper is it provides a method to recover structural parameters using bunching responses in a market equilibrium of a vertically differentiated product. The paper offers a new framework to use the observed bunching responses to do a welfare analysis of housing policies. The approach proposed in this paper can be applied to other settings with policy interventions with discontinuous incentives that cause bunching (e.g., Carozzi et al., 2020; McMillen & Singh, 2020). The proposed method complements the approaches that use notches and bunching moments for identification by providing the same identification principle to recover model primitives in the sorting or hedonic models. I use the bunching evidence and moments to estimate the structural parameters of economic models, as done recently in other settings by Einav et al. (2015) for the drug market, Best et al. (2019) for the mortgage market, and Z. Chen et al. (2021) for incentives for research and development in China.[8] This paper brings this relatively novel approach to the housing market

[7] For example, DeFusco and Paciorek (2017) use these bunching responses to estimate the interest rate elasticity of mortgage demand. Best and Kleven (2017); Kopczuk and Munroe (2015); Slemrod, Weber, and Shan (2017) report housing transaction bunching responses around notches in transaction costs.

[8] In contrast to this approach, alternative approaches implemented, for example, by Saez (2010), Chetty et al. (2011), or Chetty, Friedman, and Saez (2013) use the bunching moments to derive reduced

literature.

This paper makes important methodological contributions to the bunching and hedonic equilibrium models literature, but it also provides new empirical and theoretical insights into a first-order question. The findings of this paper can inform the design of housing policies aimed at providing affordable housing. The model presented in this paper allows me to estimate the welfare effects on developers and households, which has direct implications for policy design. While other papers investigate the effects of housing programs on households, my study contributes to the literature by also investigating the effects of these housing programs on developers.[9] My setting and approach also allow me to understand the effect of these programs on the housing market itself. My findings suggest that the policy affects behavior on both sides of the market. The policy incentives shape the type of housing that is built and sold, which has implications for how the city grows and develops. This is particularly relevant in a world of increasing urbanization. Furthermore, the affordability crisis in many developed and developing countries highlights the importance of effective housing policies. It also informs the debate about the effectiveness of housing programs compared to other social assistance programs such as unconditional cash transfers (Olsen, 2003; Olsen & Zabel, 2015).

The paper has three parts. The first part introduces the reduced-form analysis. In

form elasticities and use them as sufficient statistics for welfare analysis. See Kleven (2016) for a review of the literature using bunching. Some recent applications include studies on minimum wage (Cengiz, Dube, Lindner, & Zipperer, 2019; Harasztosi & Lindner, 2019; Jales, 2018), overpay hours (Goff, 2021; Bachas & Soto, 2018; Abel, Dey, & Gabe, n.d.), marriage market (Persson, 2020), Crime (Goncalves & Mello, 2021) among others.

[9]Many papers study housing market policies implemented in the United States. For example, Baum-Snow and Marion (2009), Soltas (2021) and Sinai and Waldfogel (2005) study the LIHTC, Collinson and Ganong (2018), McMillen and Singh (2020) study housing vouchers and Gruber et al. (2021), E. Glaeser and Shapiro (2003) study mortgage interest deductions (MID). Olsen (2003) and Olsen and Zabel (2015) compares different approaches. H. S. Rosen (1985); Poterba (1992); Galiani, Murphy, and Pantano (2015); Quigley (1982); Geyer (2017)carry out incidence and welfare analysis on housing policies.In addition to housing subsidies, there is literature on alternative approaches to affordable housing including public housing (Kumar, 2021; Franklin, 2019; van Dijk, 2019), rent control (E. L. Glaeser & Luttmer, 2003; Autor et al., 2014; Diamond et al., 2019), maximum permitted construction (Anagol et al., 2021). OECD (2021b) describes the different approaches implemented around the world to promote affordable housing.

the next section, I present the Colombian housing policy, institutional context and the discontinuities created by the subsidy scheme. Section 1.3 presents the housing market data and provides reduced-form evidence of the housing market response. The second part of the paper contains the housing equilibrium model and identification strategy. Section 1.4, introduces the model, section 1.5, presents the identification strategy. The third part, presented in section 1.7, shows the estimates for the structural parameters, the policy counterfactuals, and welfare analysis.

## 1.2 Institutional Context and Data

This section introduces the Colombian housing policy, describes the subsidy expansion and shows how the discontinuity creates incentives to bunch at the price cutoff.

### 1.2.1 Colombian Housing Policy

*Institutional context.* Colombian housing policy aims to provide a decent home and suitable living, reduce housing deficits, and achieve the dream of being a country of homeowners.[10] Since the 1990s, Colombia and other Latin American countries have changed their approach, moving from state-provided housing to a market-oriented solution based on subsidies.[11] This policy approach aims to incentivize the purchase and construction of low-cost housing through subsidies to households and developers. On the demand side, there are two main policy tools: 1) mortgage assistance through a downpayment subsidy and 2) a subsidized interest rate. On the supply side, the

[10]The first and second goals are based on Article 51 of the Colombian Constitution. The goal of being a country of homeowners appears in the country's last three National Development Plans (see, for example p104 of the National Development Plan for 2002–06).

[11]For example, in a 1993 report the World Bank said that "housing policy making must thus move away from its previously narrow focus on a limited engagement of government in the direct production of low-cost housing." World Bank Group (1993, p.1) Following these recommendations, many Latin American countries, including Chile and Colombia, abandoned the construction of public housing and implemented a market-oriented approach called ABC (from Spanish, *Ahorro*-Savings,*Bonos*-Bonds, *Creditos*-Credit) (A. Gilbert, 2014; Cohen, Carrizosa, & Gutman, 2019).

policy tool is a tax refund for developers who build low-cost housing.[12]

*Low-cost housing definition.* The policy design is heavily based on the definition of low-cost housing, which is a unit with a market price below an arbitrary threshold $\underline{P} = 135$ times the monthly minimum wage ($mMW$).[13]This arbitrary threshold is the same for all cities, and changes over time are associated with changes in the minimum wage.[14]The subsidies apply only to households and developers who buy and build low-cost housing. There is an additional definition creating a similar discontinuity at a lower price cutoff. Housing units below $70 \times mMW$ (around $US\$20,000$) classify as priority low-cost housing. This cutoff defines eligibility for some subsidies for the extreme poor and those affected by forced displacement or natural disasters. Between 2012 and 2015 these subsidies included 100 thousand free housing units. This paper focuses mostly on the subsidies targeting the population buying low-cost housing units.

*Subsidies over-time.* During my study period, the demand side subsidies increased in generosity and were modified. The interest rate subsidy was introduced, the subsidy amount increased, and individuals in the informal sector became eligible. I use these changes to show how the housing market responds to changes in subsidies. I divide my study period into four sub-periods corresponding to the distinct set of policies available. The four periods are 1) **2006-08** downpayment subsidy available only to formal employees, 2) **2009-11** downpayment subsidy and interest rate subsidy available

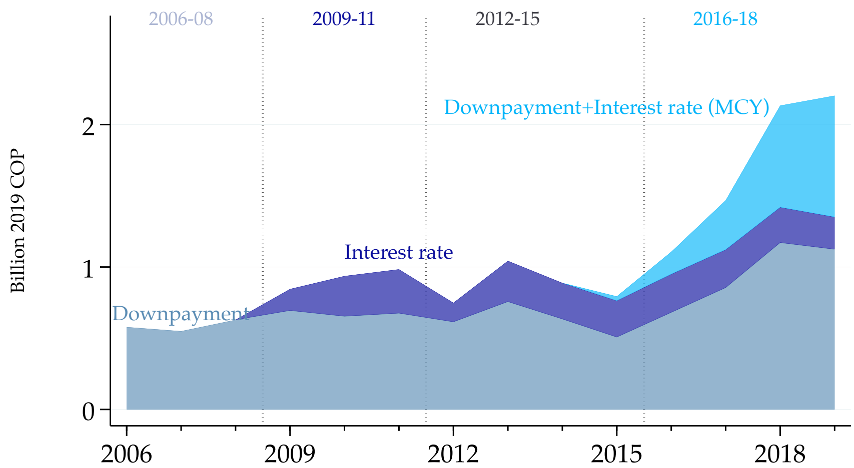
[12]The demand subsidies are similar to many *first time buyers programs* in the United States and some housing policies in the United Kingdom, such as *help to buy* (Carozzi et al., 2020). The supply subsidy could be compared to the Low Income Housing Tax Credit (LIHTC) in the United States (Baum-Snow & Marion, 2009). However, in contrast to LIHTC, the units built are not rental units, they are units to be occupied by the owner.

[13]In Colombia the minimum wage is adjusted every year based on the inflation, productivity growth and an agreement between different representatives of the different economic sectors. Appendix Figure 1.A.1 shows the evolution of the minimum wage an inflation during my study period.

[14]This price limit is set by the government's National Development Plan. It was the same from 1997 until 2019. With law 1467 of 2019, the it increased to 150 $mMW$ for the five largest cities (including the metropolitan areas) and remained the same in the other cities.

only to formal employees. 3) **2012-15** unstable period with rapid changes in interest rate subsidy and the existence of programs targeted at the extreme poor.[15] 4) **2016-18** An increased downpayment subsidy for formal employees and interest rate subsidies. Additionally, the program *Mi Casa Ya* is available to all households with earnings of 4 mMW or less and automatically includes the downpayment and the interest rate subsidy.

*Figure 1.1:* Total Government Expenditure on Demand Subsidies over Time




SOURCE: Administrative records from the Ministry of Housing. Appendix 1.A provides more details about the data.
NOTE: This Figure shows the evolution of total government expenditure by type of subsidy. The down-payments are the subsidies awarded to employees affiliated to family funds. The interest rate represents the total amount paid by the government to the banks corresponding to the interest rates payments. I assigned the total amount to the year of the subsidy assignment. I calculated this amount using the administrative data containing detailed information on each loan. Mi Casa YA corresponds to the payments for the interest rate and the downpayment subsidy. Figure 1.A.2 shows the number of assigned subsidies over time.

*Subsidy expansion.* The government expenditure on these subsidies doubled during my study period. Figure 1.1 shows the total government expenditure from 2006 until 2018. The **gray blue** area shows the expenditure on downpayment subsidies. The expenditures were stable until 2015, when the subsidy's size increased. The **dark**

[15] Including the provision of *100,000 free housing units* and the country's primary mortgage downpayment subsidy program for the vulnerable population (VIPA). For more details, see Camacho, Caputo, and Sanchez (2020) and A. Gilbert (2014)

**blue** area shows the total government expenditure on the subsidized interest rate. The number of households that received this subsidy was stable over time, but government expenditure decreased slightly due to the lower interest rate.[16] The **light blue** area shows the expenditure related to the *Mi Casa Ya* program, which provides downpayment assistance and covers the interest rate discount.

*Supply subsidy–value-added tax (VAT) refund.* To encourage developers to build low-cost housing, the government introduced a VAT refund. Developers get up to 4 percent of the sale price of each unit as a refund for taxes paid on construction materials. This subsidy, was introduced in 1995, a couple of years after the beginning of the downpayment subsidies.[17]

*Comparison to other subsidies.* As a reference point, the Colombian conditional cash transfer, *Familias en Acción*, benefited almost 4 million people with an expenditure of 3 billion COP in 2019 (DNP, 2018), while 100,000 of these housing subsidies cost 2 billion COP plus the tax benefits for developers. Becoming homeowners can confer substantial benefits to low income households and is therefore a legitimate goal. However, few households benefit from the subsidies to become homeowners, raising the question of whether it could be better to target a wider population with other subsidies like rent vouchers or unconditional cash transfers. For this reason, it is crucial to

[16]To obtain the government expenditure, I calculate the total savings on mortgage payments induced by the discount at the interest rate. I calculate the monthly payments of each loan using the administrative records for the subsidy and the formula for monthly payments on a mortgage, $L_{monthly} = L \cdot \kappa(i,n)$ with $\kappa(i,n) = \frac{i}{12} \cdot \left(1 + \frac{i}{12}\right)^{12 \cdot n} / \left(\left(1 + \frac{i_h}{12}\right)^{12 \cdot n} - 1\right)$. Where $i$ is the interest rate, $i_{subsidy}$ is the interest rate discount, $n$ is the loan term in years, $L$ is the loan amount. The government pays the difference in the amount paid by households ($L \cdot \kappa(i_\tau, n)$, with $i_\tau = i - i_{subsidy}$) and the amount received by the bank ($L \cdot \kappa(i_\tau, n)$). In particular, $\tau^i = \sum_{t=1}^{84} L_{month}(i,n)(i,n) - L \cdot \kappa(i_\tau, n)$, the sum of monthly payments for seven years, the period during which the subsidy applies. Figure 1.A.6 shows the loan terms by unit price and Figure 1.A.7 shows the market interest rate and the interest rate that households pay.

[17]This policy instrument was first introduced in the 1995. Even-thought it has been regulated by different laws and acts, for example, Law 1607 of 2012 or Act 2924 of 2013 (Camacol (2016) p.25.), it has always had the same incentive capped at 4 percent of the value of each unit.

understand the effects on the housing market and the welfare gains associated with these subsidies.[18]

### 1.2.2 The Notch

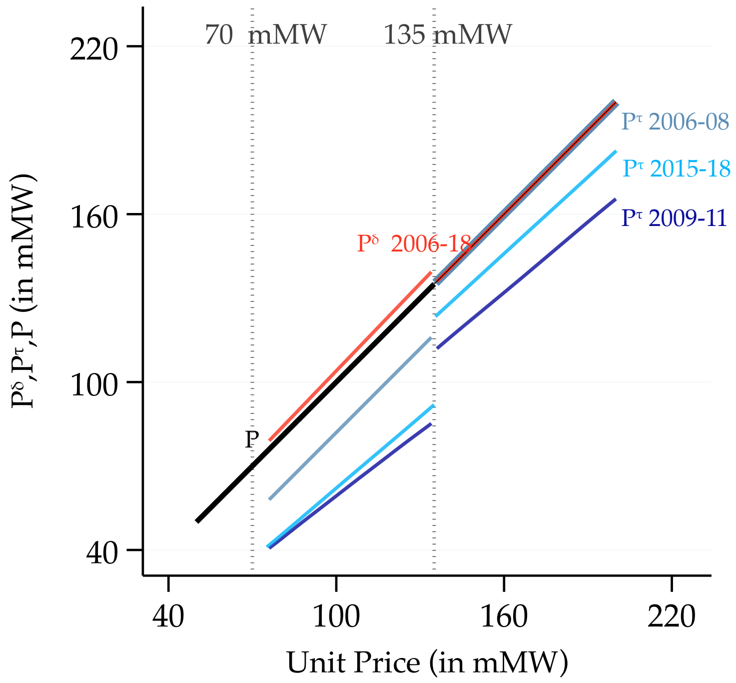
The combination of the arbitrary definition of low-cost housing and supply and demand subsidies creates discontinuous incentives or notches around the low-cost housing cutoff point.

*Relevant prices.* The subsidy scheme creates three different prices: transaction or market price, $\mathbf{P}$; developer price, $\mathbf{P}^{\delta}$, or the price per unit that developers receive after including tax refunds; and household price, $\mathbf{P}^{\tau}$, or the price households pay net of subsidies. Therefore, $P^{\delta} = P \cdot (1 + \delta)$ and $P^{\tau} = P - \tau$. Where, $\delta = 4$ percent is the tax refund rate, and $\tau = \tau^{m} + \tau^{i}$ is the total amount of money paid by the government for a housing unit. $\tau^{m}$ represents the downpayment subsidy and $\tau^{i}$ is the interest rate subsidy. The downpayment assistance is a fixed amount, independent of the housing price. The interest rate subsidy is related to the housing price to the extent that they depend on the size of the mortgage. There is an interest rate subsidy above the low-cost housing cutoff, but there is a jump in the subsidy at that cutoff. For example, in the 2016-18 period, the subsidy goes from 4pp for a house with a price below 135 $mMW$ to 2.5 pp above that cutoff. I use administrative records to calculate the government expenditure on interest rate payments using a typical mortgage at each price level (see details in Appendix 1.A).

*The notch.* The demand notch is the difference between the blue lines and black line in Figure 1.2. A household buying a housing unit below the cutoff qualifies for more subsidies; the blue lines are below the black lines because of the housing subsidies. The gap between the black line and the blue line is the money paid by the government

[18]See Pattillo (2013) for discussion about housing as a commodity vs a right

*Figure 1.2:* The Notch




NOTE: This figure compares the **market price P**, the price received by developers $P^{\delta}$, and the price households pay net of subsidies (in blue). The x-axis represents the market price **P**, and the y-axis represents the price received by developers $\mathbf{P}^{\delta}$ or paid by households $\mathbf{P}^{\tau}$. The 45-degree black line represents the market price. The three different blue lines correspond to the three subsidy schemes available during the study period,$P^{\tau}$ **2006-08**,$P^{\tau}$ **2009-11** and $P^{\tau}$ **2016-18**. The price paid by households is $P - \tau^m - \tau^i$, $\tau^m$ is a transfer from the government for the downpayment and does not depend on the price of the house. $\tau^i$ are the savings in interest rate payments, because this depends on the mortgage; it is calculated by taking a typical mortgage at each market price using administrative records from the Ministry of Housing.

$\tau$. The supply notch is the difference between the red and black lines. Figure 1.2 illustrates how the subsidy scheme creates incentives for developers and households to build and buy housing units with a price at or below the cutoff. By reducing the housing consumption from above to below the cutoff, households and developers have a discontinuous jump in the price they pay or receive.

*Notch over time.* The notch on the demand side increases over time. The **gray blue** line shows the household price when only the downpayment subsidy was available between 2006-08. Before the government introduced the interest rate subsidy, households buying a unit priced above the cutoff paid the full price. In the figure, the black and blue lines coincide above the cutoff. The **dark blue** line shows the price paid by a household that gets the downpayment subsidy and the interest rate subsidy during 2009-11. The interest rate is also available if households get a unit above the price cutoff, but the discount is smaller. The two subsidies combined increase the discontinuity or notch at the cutoff. The **light blue** line shows the price paid by households who received the two subsidies after the *Mi Casa Ya* program was introduced and the increase in the downpayment subsidy. During this period, there was a drop in the interest rate and therefore the interest rate subsidy was lower. This explains why the price paid by households below the cutoff was similar during 2009-11 and 2016-18, even if the downpayment subsidy was higher. It also explains why the price in 2016-18 was lower above the cutoff. Despite these changes in the interest rate, the notch increased during this period. There were many changes during the period between 2012 and 2015. In addition to the 100,000 free housing units priced at $70 \times mMW$ or below the interest rate, the subsidy changed many times. For completeness, I include this period when presenting the data and results; however, I see it as a transition period and therefore pay little attention to it.

*Notch size.* Table 1.1 shows the size of the jump at the cutoff during the study period and the number of assigned subsidies for each program. Around 45,000 households received the downpayment subsidy each year, with slight variation across years, and around 22,000 households received the interest rate subsidy. Households can get both supports, but they have to apply separately to each program. Each year, around 17,000 households receive the subsidy from the *Mi Casa Ya* program, which grants both subsidies.

*Table 1.1:* Notch and number of subsidies by period

| | Notch *(in mMW )* | | | Subsidies *(in thousand)* | | |
|---|---|---|---|---|---|---|
| | $\tau^M$ | $\tau^i$ | $\tau$ | *downpayment* | *interest rate* | *Mi Casa Ya* |
| 2006-2008 | 18.0 | . | 18.0 | 47.1 | . | . |
| 2009-2011 | 20.0 | 5.85 | 25.9 | 46.4 | 16.7 | . |
| 2012-2015 | 19.9 | 9.55 | 29.5 | 41.1 | 22.2 | . |
| 2016-2018 | 25.3 | 7.24 | 32.6 | 44.5 | 23.4 | 16.8 |

NOTE: This table shows the size of the notch in figure and by period and differentiating the discount coming from the interest rate subsidy and the discount from the downpayment assistance. It also shows the number of subsidies (in thousands) assigned to each type of program by year, downpayment, interest rate, and the two together with *Mi Casa Ya*. The value for each period is the average number. Figures 1.A.7 and 1.A.6 shows the loan terms and interest rate over time.

### 1.2.3 Data

In addition to the administrative records for the subsidies that I presented above, the main analysis of the paper is based on a census of all new construction projects. This subsection introduces that census.

*Data source.* The data are from a monthly census, called *Coordenadas Urbanas*, collected by the Colombian Chamber of Construction-CAMACOL and containing all new construction units built in 126 Colombian municipalities between 2006 and 2018.[19] The unit of observation is a housing unit type. For example, there may be three different apartment types in a housing development such as studios, one-bedrooms, and two-bedrooms. I observe the price and characteristics of each of them. I observe all housing development projects of at least 300 square meters of construction. The census excludes small, single-family homes and informal housing. It does not contain information on resales of existing housing units. Although this is a limitation of the

[19]Not all cities have information starting in 2006, the census expanded its coverage over time.

data, the subsidies apply only to new housing, so the data covers the directly affected part of the market.

*General characteristics.* The data contain detailed information of the house such as the unit size; location, including the exact latitude and longitude coordinates; number of rooms; quality of appliances; estrato, which is an index summarizing neighborhood quality; and *developer and project characteristics*, like firm tax identifier and the number of units built in each project. The data also include detailed characteristics of the housing development, including the number of parking spots, the number of towers built, the lot size and an indicator function equal to 1 if the units are apartments and 0 if they are single family units, among other details. Finally, I observe the *sale price* at different stages of the construction process. To ease the comparison, I take the price at the beginning of the construction of the project. All prices are in 2019 COP or $mMW$. In Colombia, there is a national $mMW$, which is adjusted every year based on inflation (see Figure 1.A.1). In most of the analysis, I express the price in $mMW$ to make it comparable with the price cutoff defining low-cost housing units.

## 1.3 Housing Market Responses

Section 1.2 shows how the Colombian social housing policy design creates incentives for households to bunch at the cutoff. In this section, I show the response of the housing market to those incentives. There is clear bunching at the price cutoff, and the bunching gets bigger as the generosity of the subsidies increases.

### 1.3.1 Bunching in Observed Market Outcomes.

*Bunching around the price limit.* Figure 1.2 shows how the subsidy scheme creates incentives for households and developers to buy and build housing units priced at or below the cutoff. The data allow me to differentiate the product choice from the number of units that developers build. I leverage this advantage of the data in the model presented in section 1.4.

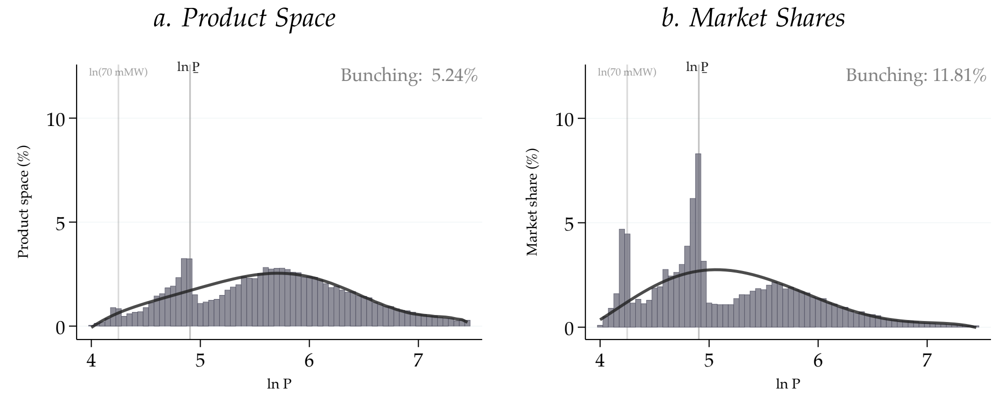
Figure 1.3 shows the market response to these subsidies. Figure 1.3a shows the distribution of the product space by unit price (i.e., type of units built). Figure 1.3b shows the distribution of market shares by unit price for all years and cities in the data. The difference between the two distributions is explained by the fact that developers who build cheaper housing units usually build more units. Apartment buildings often have cheaper housing than projects with single-family homes.

The figure shows a sharp and clear excess mass, or bunching, around the price cutoff defining low-cost housing. For the whole sample, about 12 percent of the market share moves from above the cutoff to below the cutoff.This response is the result of the notched policy design. From this figure it is not possible to determine if the policy induced new units to be built. However, there is a reallocation of the type of new construction. The subsidy moved people and developers from above to below the cutoff. In the absence of the policy, they would buy and produce more expensive housing, but they change their behavior to take advantage of the subsidy. From this, Figure 1.3 shows why a naive policy evaluation comparing the number of units to the left and to the right of the cutoff would be misleading. In this comparison, the treated group would be "inflated" and the control group "deflated" by households that modify their consumption, but in both scenarios buy a house. Therefore, accounting for this type of response becomes essential to understand the effect of the policy.

*Counterfactual.* The solid line in Figure 1.3 represents the counterfactual distribution. That is, the distribution of housing units in the absence of the subsidy. To construct the counterfactual distribution, I follow the standard techniques from the bunching literature (Kleven, 2016). The idea behind the estimation of the counterfactual distribution is to fit a flexible polynomial to fit the observed distributions and include dummies

*Figure 1.3:* Bunching around the Low-Cost Housing Price Limit

*All Data 2006-18*



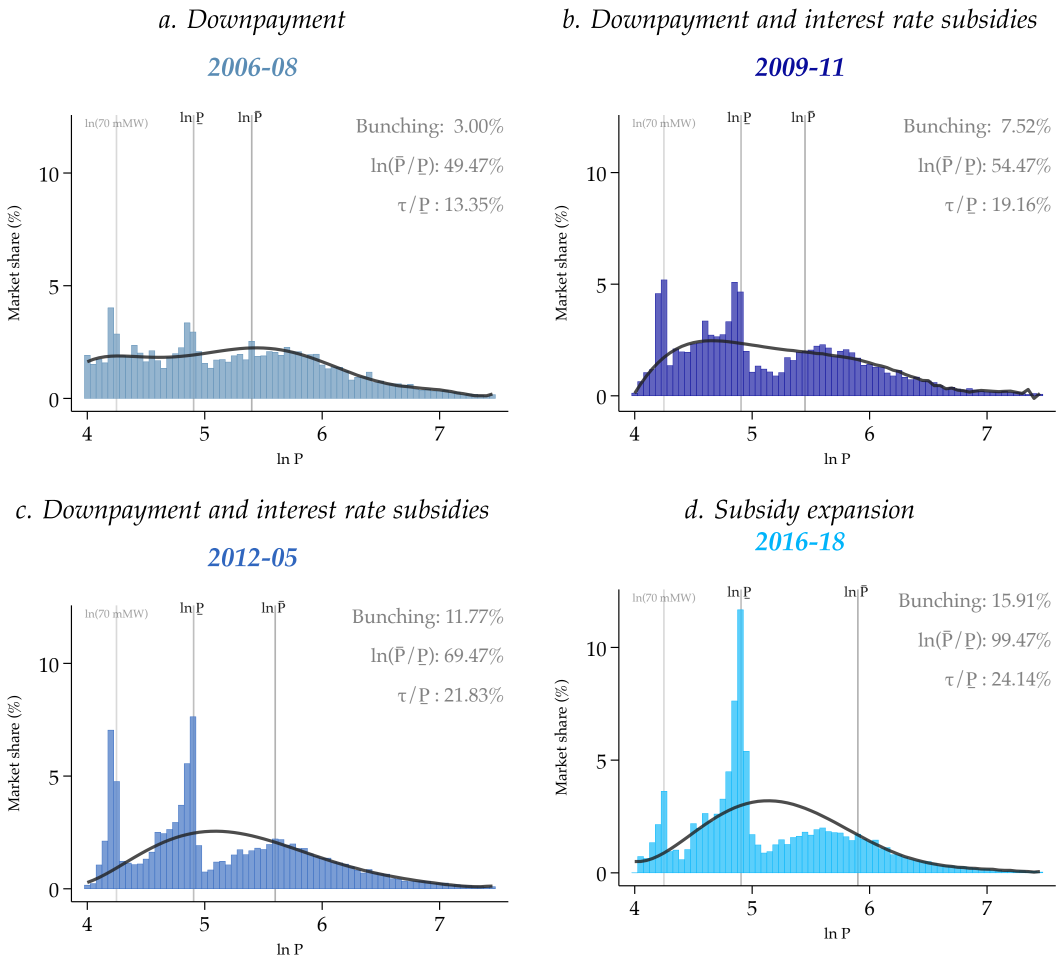

Note: This figure shows the distribution or the market share of housing units by sale price (expressed in logs (mMW). The vertical lines are the cutoffs defining *low-cost housing* $\underline{P}$ = 135 mMW and *priority low-cost housing* 70 mMW. The figure shows all the units from 2006 to 2018 in all the cities.

around the discontinuity. The distribution is the prediction of the flexible polynomial once we exclude the bins around the cutoff. To select the parameters related to the estimation, I follow a similar approach to Diamond and Persson (2016) and Z. Chen et al. (2021). For a given bin size, I chose the degree of the polynomials and the number of bins to exclude to match the missing and excess mass. The details of the estimation approach are in Appendix 1.B.

### 1.3.2 Bunching Over Time and Counterfactual Distribution

An advantage of my setting is that I can see how the housing market responds to changes and increases in demand subsidies. The subsidy scheme evolved during my study period, and the demand notch almost doubled. Figure 1.4 shows the distributions of market share over time. This figure also shows the size of the notch, as the percentage of the price of a house at the cutoff (i.e., $\frac{\tau}{\underline{P}}$) and the maximum change in housing consumption (i.e., $ln(\frac{\overline{P}}{\underline{P}})$. The relationship between these two magnitudes give me a reduced-form semi-elasticity, relating how much households are willing to change their housing consumption to get the subsidy.

*Figure 1.4:* Bunching over time




Note: This figure shows the distribution or the market share of housing units by sale price (expressed in log of mMW). The lines are the cutoffs defining *low-cost housing* $\underline{P}$ = 135 mMW and *priority low-cost housing* 70 mMW. The additional lines shows the point, $\overline{P}$, where the counterfactual and observed distribution coincide again after the cutoff. The figure shows the different periods for all available cities.

*Bunching over time results.* Figure 1.4 shows these magnitudes based on comparisons between the observed distribution and the calculated counterfactual distributions. The figure provides compelling evidence that the housing market responds to the subsidy scheme. At the beginning of my period, when only the downpayment for formal employees was available, households reduced their housing consumption by up to 50 percent to receive a subsidy of 13 percent of the value of the house. The semi-elasticity is 3.85 . Bunching in that period is 3 percent. This number is equivalent to the market share that changes its behavior to take advantage of the subsidy. In 2009-11, when the interest rate subsidy appeared, the notch jumped to 19.2 percent of the house price at the cutoff. Consequently, the share of households responding in this dimension is . In 2012-15 there is a big bunching point at the cut-off point of 70 $mMW$. This corresponds to the program of 100 thousand free housing units for the most vulnerable, which took place during that period (A. Gilbert, 2014; Camacho et al., 2020). The notch also increased, which triggered a larger share of the market to modify housing consumption to take advantage of the subsidy. In 2016, when the program *Mi Casa Ya* was introduced, households could receive almost a quarter of the value of the house if they reduced their consumption to qualify for the subsidy. Up to percent of households modify their behavior in this way. The semi-elasticity for this period is 3.85 .

### 1.3.3 Bunching as an Equilibrium Response

To be able to learn something about the market structure from these reduced-form evidence, it is crucial to know the specific responses of developers and households that lead to the observed market equilibrium.

*Market adjustment on housing characteristics.* The mechanism explored in this paper is that households and developers adjust the characteristics of the housing units they buy and produce to take advantage of subsidies. There are two main reasons for taking this approach. First, the construction sector is perceived as highly competitive and developers have no incentive to build larger units when, for the same price, households would buy smaller units.[20] Second, Figure 1.5 provides suggestive evidence that the subsidy scheme affects the characteristics of the housing stock.

*Alternative explanations.* There could be at least three explanations that explain this equilibrium. First, a change in housing characteristics is the main explanation explored in this paper. In this approach, households and developers change the type of housing they are consuming and building. This would include less quality, smaller, or fewer amenities. Second, a pure behavioral response. In the literature, the bunching is usually explained by reporting (Chetty et al., 2013) or relabeling (Z. Chen et al., 2021). However, as I show in the next section, it can also be explained by real market response like changes in characteristics. In the institutional setting studied in this paper, there are many agents with competing interests; households, banks, developers and the government. This makes a simple reporting response costly and less likely.[21] For example, banks do not have an incentive to under-report the price of the house for the mortgage. A third explanation could be a response only in prices. Households and developers buy and produce the same type of housing, but developers reduce the price for units above but close to the cutoff, allowing households to get the subsidy. This explanation would require an high degree of market power. Although the last two explanations are plausible, a detailed investigation of them is beyond the scope of this paper. The explanation explored in this paper rationalizes the observed equilibrium.

*Housing characteristics.* It is difficult to summarize housing units into a single variable, as they differ in many dimensions. However, focusing on a single characteristic makes the analysis more tractable. My data have exact size, which allows me to use this fea-

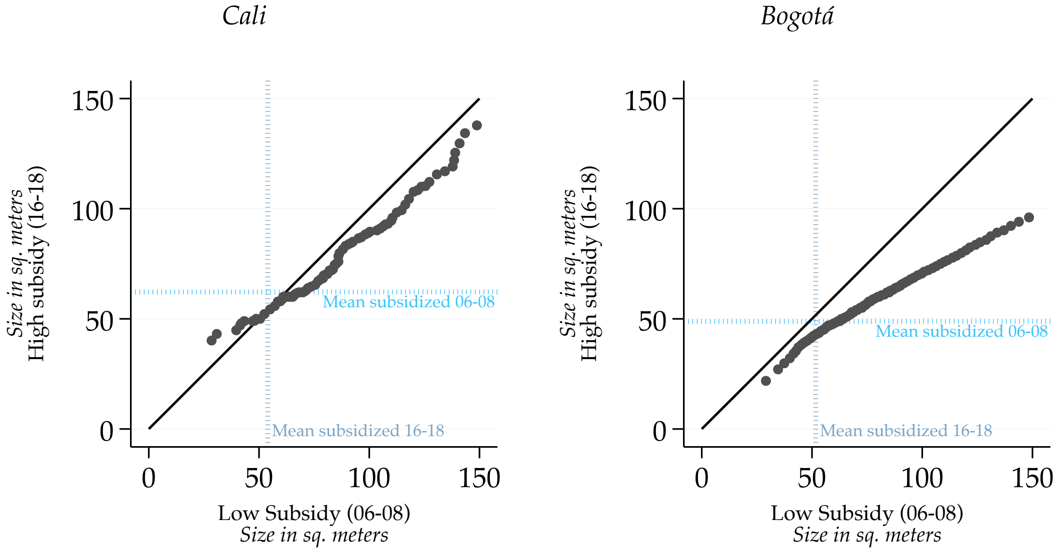
[20] This argument will be clearer when I introduce the model in the next section. The argument applies for any characteristics that imply any cost for developers.

[21] Anecdotal evidence may suggest that some housing units sold in expensive neighborhoods as low-cost housing show miss-reporting (Metrocuadrado, 2022; radio, 2022). However, in many cases those houses do not have appliances when they are sold or are extremely small (20 square meters). In that sense, this type of response falls in the category of market response and not miss-reporting.

ture to analyze the structure of the housing market. Choosing size as the main variable in the analysis has several advantages. First, it is a concrete feature of the apartment itself, as opposed to the amenities of the neighborhood. Size is easy for builders to adjust in response to government policy, than neighborhood amenities. Second, the detail of the data allows to control for an unusually large set of variables that may be correlated with size, including apartment and building characteristics, neighborhood quality (estrato, exact location), and structural characteristics of the house such as number of rooms, if there is a porch, number of bathrooms, among others. Third, size has the strongest reduced-form association with price. Fifth, it allows estimation of a continuous implicit price function which is important for the modeling approach considered in this paper; 5) has a monotonic relationship with price, unlikely to be the case for other continuous variables (exact location). Finally, other characteristics such as quality are not detailed enough to allow for as plausible an analysis of overall market structure. Lastly, as Figure 1.5 shows, the market seems to respond in unit size.

*Size response.* Figure 1.5 shows that with the increase in subsidies, the housing size distribution is affected. Specifically, it is affected around the median size of the subsidized housing. Figure 1.5 has quantile-to-quantile plots for housing size at the beginning and end of the study period for two different cities (Figure 1.C.10 in the appendix shows more cities). During this period, the notch induced by the policy increased from 18 $mMW$ to 33 $mMW$. If the distribution of housing size did not change from the beginning to the end of the study period, the black dots would be on the 45-degree line. The blue dotted lines show the average size of a subsidized house. These figures suggest a change in the size distribution around the average subsidized unit. Changes in housing characteristics, particularly size, can explain the increase in bunching from 2006-08 to 2016-18.

*Figure 1.5:* Quantile-to-Quantile Plots of Housing Size: Low versus High Subsidy Periods




Note: This figure shows the quantile-to-quantile plots for observed housing size in square meters for two representative cities, Cali and Bogotá. The y-axis shows the size at the end of the period, when subsidies are high, and the x-axis shows the size at the beginning of the period, when subsidies are low. The dotted vertical and horizontal lines show the average size of subsidized units. The dots represent the same quantiles in both years. If there are no changes in housing size, they would be on the 45-degree line. Instead, the figure shows how there are changes in size at the quantiles near the average subsidized house.

*Section takeaways.* This section provides compelling evidence that the Colombian housing market responded to the discontinuous incentives generated by the subsidy scheme. The suggestive evidence supports the view taken in this paper that these responses come from changes in housing characteristics and, in particular, households and developers buying and building smaller housing units to take advantage of the subsidy. How does the supply side adjust to this change? How is the equilibrium price set? Is the subsidy to developers necessary to prevent housing rationing? Are there any inefficiency gains or welfare losses associated with the subsidy scheme? The purpose of the remaining sections of this paper is to be able to address these questions.

## 1.4 Competitive Housing Market Equilibrium Model

This section introduces a housing market equilibrium model. There are three main objectives for the model. First, provide a framework for rationalizing the observed equilibrium and understanding the economic behavior driving the market equilibrium. Second, it describes the equilibrium conditions for the model and the role of the hedonic price function. Third, it motivates a novel identification approach to recover the behavioral parameters of households and developers.

### 1.4.1 Model Setup

The proposed model introduces the discontinuous incentives produced by the Colombian subsidy scheme into a standard hedonic equilibrium model or sorting model.[22]

***Housing.*** Housing is a vertically differentiated product characterized as a continuous variable $h$. In this case, all units are standard units that differ only in how large they are, therefore $h$ represents the size of the house in square meters. The price of the housing unit $P$ depends on the size $h$, and is described by the implicit price for the size $P(h)$, which can be nonlinear.

***Households.*** Households looking to buy a *new* housing unit are indexed by $i$, are heterogeneous in their wealth level $Y_i \sim F_Y$.[23] Households decide how much housing to buy, $h_i$ and how much to consume of other goods, $C_i$, to optimize utility $U(C_i, h_i; \theta)$, where $\theta$ is a preference parameter to be estimated.

[22]This is a canonical version of a model with heterogeneous households and developers buying housing units of different sizes. For ease of exposition, I simplify it by assuming that a single variable describes the housing. For some examples of these types of models, without a notch in the budget set, see S. Rosen (1974), Epple (1987), Ekeland et al. (2004), Bajari and Benkard (2005), Heckman et al. (2010), Epple et al. (2020) or Chernozhukov et al. (2021). The literature based on this models is summarized by Kuminoff et al. (2013) and Greenstone (2017). For a survey of the empirical applications see Palmquist (2006).

[23]I call $Y_i$ wealth for simplicity. It is a measure containing wealth, assets and their returns, transfers, income, etc. $F_Y$ is the $cdf$ describing the wealth distribution

***Developers.*** Developers are indexed by $j$ and heterogeneous in their productivity $A_j \sim G_A$. They decide what type of product they want to build. In particular, the size of housing units, $h_j$, to maximize profits. The number of units, $Q_j$, is determined exogenously by the function $Q\left(h_j\right)$. The difference between the choice of units they build and the type of units they build is behind the differences between the distribution of market shares in Figure 1.3. They face construction costs $B(h_j, Q\left(h_j\right); \beta)$[24] where $\beta$ characterizes the cost function and is the supply parameter to be estimated.

*Simplifying assumptions.* I introduce three simplifying assumptions. First, I assume that the market is perfectly competitive,that is, developers cannot individually affect prices and $P(h)$ is independent of $Q$. Second, developers only choose the unit size they build. They follow a unit supply function that is exogenous and differentiable $Q = Q^S(h)$. The number of units does not need to be predetermined since apartment size is an endogenous choice, but the allocation of property to developers is predetermined. This is, for a given lot, households need to decide the size of the units, and the regulation framework and construction and technological constraints will determine how many units they can build. Third, construction costs depend on $Q(h)$, $h$, and productivity levels, that is, $B = B\left(Q(h), h, A_j; \beta\right)$. The last two simplifying assumptions make it straightforward to specify functional forms for the profit function and to offer curves. Allowing for a completely endogenous choice of $Q$ could be a better characterization, but obtaining a functional form for the offer curve, which is essential in the identification approach, is highly dependent on particular functional forms. Relaxing this assumption and allowing for imperfect competition is feasible, but beyond the scope of this paper.

[24] The cost function $B\left(Q, h, A_j; \boldsymbol{\beta}\right)$ is derived from minimizing the production constraints related to producing $Q$ units with characteristics $h$. $A_j$ reflects underlying variables in the cost minimization, that is, factor prices and production function parameters. Different values of $A$ express different factor prices or productivity among developers. For a discussion, see S. Rosen (1974, p.43)

*Equilibrium.* When households decide the type of units they buy, they choose the developer type from which to buy, and vice versa. Then, the equilibrium is an implicit price making the densities of the housing units demanded and produced match.

### 1.4.2 Optimal choices

#### *Prices*

Section 1.2 explained that given the subsidy scheme, there are three relevant prices. They are the market, household and developer price.

$$\text{Market:} \qquad P(h) \tag{1}$$

$$\text{Household:} \qquad P^{\tau}(h,\tau) = P(h) - \tau \cdot \mathbb{1}\left[P(h) \leq \underline{P}\right] \tag{2}$$

$$\text{Developer:} \qquad P^{\delta}(h,\delta) = P(h)\left(1 + \delta \cdot \mathbb{1}\left[P(h) \leq \underline{P}\right]\right) \tag{3}$$

*size threshold* Note that given the price function $P(h)$, there is a maximum size that households can buy to qualify for the subsidy. This is the size threshold;

$$\underline{h} = P^{-1}\left(\underline{P}\right) \tag{4}$$

*Differences in prices.* A household buying a low-cost house pays a price $P^{\tau}(h,\tau)$ instead of $P(h)$, and developers who build low-cost houses can get back the VAT taxes paid for the construction materials. The reimbursement of VAT taxes cannot exceed a value $\delta = 4$ percent of the value of the house. In other settings where the price can increase and the limit is set in terms of size, market equilibrium could be achieved by increasing the price and $\delta$ would represent a premium to build low-cost housing. The price function $P(h)$ can be a continuous and differentiable function for all $h \in \mathcal{H}$, but the developer and the household price functions, $P^{\delta}(h,\delta)$, and $P^{\tau}(h,\tau)$, are not differentiable at $\underline{P}$.

### *Decision Problem*

*Households.* A household $i \in N$ maximizes its utility given its level of wealth $Y_i$. It solves the following optimization problem:

$$\begin{aligned} \max_{h,C} \quad & U\left(h, C; \theta\right) \\ \text{subject to:} \quad & Y_i = P^{\tau}\left(h, \tau\right) + C, \\ & h \geq 0. \end{aligned}$$

*Bid functions (or indifference curves).* $\varphi_D\left(h, Y, \bar{U}; \theta\right)$, represent all the combinations of prices $P$ and unit size $h$ that provide the same level of utility $\bar{U}$ to a household with $Y = Y_i$. Therefore, $\varphi_D$ is such that

$$\bar{U} = U\left(h, Y_i - \varphi_D; \theta\right) \tag{5}$$

*Developers.* Developer's profits $\pi\left(Q, h, A_j\right)$ are determined by the total revenue minus costs.

$$\begin{aligned} \max_{h} \quad & \pi\left(Q, h.A_j\right) \\ \text{subject to:} \quad & \pi = Q \cdot P^{\delta}\left(h, \delta\right) - B\left(Q, h, A_j; \beta\right) \\ & Q = Q\left(h\right) \end{aligned}$$

*Offer function (or iso-profits)* The offer function represents the indifference surface for all possible combinations of prices and size $h$ providing the same profits. $\varphi_j^S$ represents the price that developers are willing to accept at different unit sizes to obtain the same level of profits $\bar{\pi}_j$. To define the offer function, I replace the developers' price, $P^{\delta}\left(h, \delta\right)$, by $\varphi_j^s$, profits by $\bar{\pi}$, and solve for $\varphi_j^s$,

$$\varphi_j^s = \frac{B\left(Q^s\left(h\right), A_j; \beta\right) + \bar{\pi}}{Q^s\left(h\right)} \tag{6}$$

### ***Tangency Conditions***

*Households.* On the demand side, households choose their housing size $h$ to maximize their utility. Due to the notch in the budget set, the standard *tangency conditions* do not correspond to the optimal choice for all households. I define the *tangency conditions*,[25]

$$\frac{\partial P\left(h\right)}{\partial h} = \frac{\frac{\partial U\left(h, C; \theta\right)}{\partial h}}{\frac{\partial U\left(h, C; \theta\right)}{\partial C}} \tag{7}$$

Assuming that equation 7 has a unique solution and using the budget constraint, $P^\tau\left(h, \tau\right) - Y_i = C_i$, we can *solve* for $h^*$, the choice of housing satisfying *tangency conditions*.[26]

$$h^*\left(Y_i, \tau; \theta, \underline{P}\right) = \begin{cases} h\left(Y_i + \tau; \theta\right) & \text{if } P\left(h\right) \leq \underline{P} \\ h\left(Y_i; \theta\right) & \text{if } \underline{P} < P\left(h\right) \end{cases} \tag{8}$$

*Developers.* On the supply side, the design that satisfies the optimality conditions $h^*\left(A_j, \beta\right)$ for a given price function $P\left(h\right)$ is achieved when developers maximize profits subject to the developer's price being equal to the offer curve $P^\delta = \varphi^s$. The unit size that satisfies the tangency conditions $h^*\left(A_j, \beta\right)$ and the optimal profits $\bar{\pi}\left(A_j, \beta\right)$

[25]This follows by defining a Lagrangian and taking first-order conditions with respect to $h$ and $C$ and taking the ratio. I assume that the composite good has a price $p_c = 1$

[26]It has been discussed in the literature that a sufficient condition for this to hold is to assume a Spence-Mirrlees type single crossing condition. See for example, Heckman et al. (2010, p.1573) or Kuminoff et al. (2013) for an overview.

are achieved when the price and offer curves are tangent.

$$\frac{\partial \varphi^s\left(h, A_j;\beta,\bar{\pi}\right)}{\partial h} = \begin{cases} \dfrac{\partial P\left(h\right)}{\partial h}\cdot\left(1+\delta\right) & \text{if } P\left(h\right) \leq \underline{P} \\ \\ \dfrac{\partial P\left(h\right)}{\partial h} & \text{if } \underline{P} < P\left(h\right) \end{cases} \tag{9}$$

We can solve 9 for $h$, and obtain an expression for the tangency conditions,

$$h^*\left(A_j,\delta;\beta,\underline{P}\right) = \begin{cases} h\left(A_j,\delta;\beta\right) & \text{if } P\left(h\right) \leq \underline{P} \\ h\left(A_j;\beta\right) & \text{if } \underline{P} < P\left(h\right) \end{cases} \tag{10}$$

### 1.4.3 Marginal Bunchers and Optimizer Types

The individual level demand and supply do not correspond to optimality conditions in this setting because there is a subset of households for which it is optimal to sacrifice housing consumption to obtain the subsidy. For developers, it is also beneficial to produce a smaller housing unit to benefit from the tax refund. There are three types of households and developers; *always-takers*, *marginally subsidized*, and *never-takers*. To define them, I use two key agent types; *marginal buncher* and *threshold optimizer* for both households and developers. The marginal buncher agents are a key component of the empirical approach of this paper; they define the identification approach presented in Section 1.5. They are indifferent to changing their behavior and receiving the subsidies or not changing their behavior and buy and produce the housing unit satisfying the optimality conditions.

Marginal Buncher Household: $Y_i = \overline{Y}$

$$h^*\left(\overline{Y},\tau;\theta,\underline{P}\right) = \overline{h} \iff U\left(\overline{Y} - P^{\tau}\left(\overline{h},\tau;\right),\overline{h};\theta\right) = U\left(\overline{Y} - P^{\tau}\left(\underline{h},\tau\right),\underline{h} : \theta\right) \tag{11}$$

Marginal Buncher Developer: $A_j = \overline{A}$

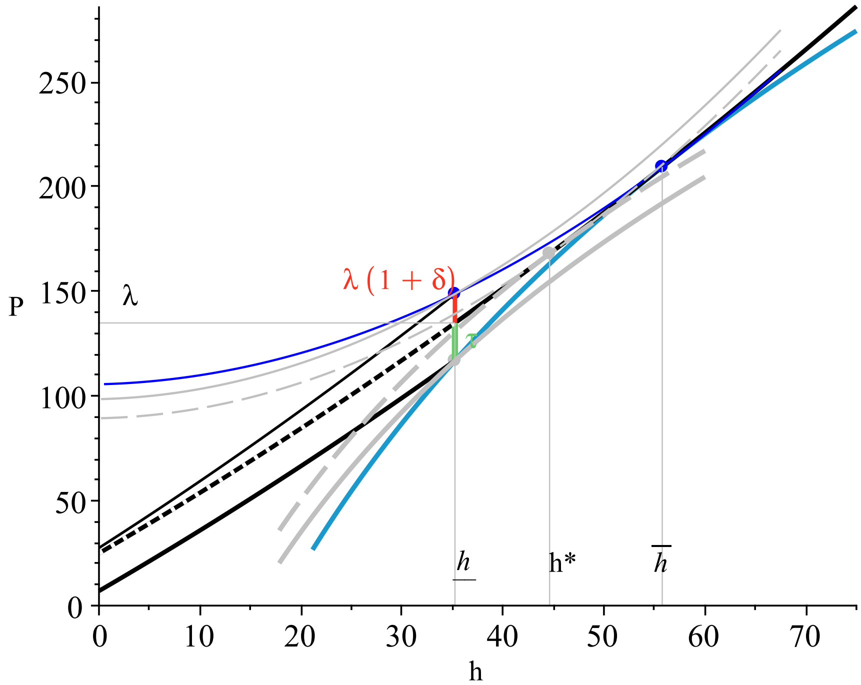
$$h^*\left(A_j, \delta; \beta\right) = \overline{h} \iff \pi\left(Q(\underline{h}; \alpha), \overline{A}; \delta\right) = \pi\left(Q(\overline{h}; \alpha), \overline{A}; \delta\right) \tag{12}$$

*Threshold optimizer* are the households and developers optimizing at $P^{-1}\left(\underline{P}\right) = \underline{h}$. They are have wealth and productivity $Y_i = \underline{Y}$ and $A_j = \underline{A}$ respectively.

*Individual-Level Supply and Demand.* The demand and supply function will be different for the three groups of agents. The *always-takers* with $Y_i \in (0, \underline{Y})$ and $A_j \in (0, \underline{A})$, receive subsidies and optimize at the tangency point. The *Marginally subsidized* with $\underline{h}$ $Y_i \in (\underline{Y}, \overline{Y})$ and $A_j \in (\underline{A}, \overline{A})$, are the ones that bunch at the cutoff. The policy design induces a change in their behavior but they consume and produce less housing than the optimality conditions would suggest. An important component of the welfare analysis in Section 1.7 is to calculate the size of this efficiency cost. The *never-takers* $Y_i > \overline{Y}$ and $A_j > \overline{A}$ do not find it beneficial to modify their behavior to take advantage of the subsidy.

*Figure explanation.* Figure 1.6 shows an example of the equilibrium choices of developers and households. The price function is the envelope of the offer curves when developers produce their optimal unit size and the assigned number of units. The figure shows a representative marginal buncher household and developer. It also shows in gray marginally subsidized households and developers, which are the agents that change their behavior to take advantage of the subsidy. A developer type $A_j$ matches with a household type $Y_i$ in terms of their optimal choice of $h$ when the dashed lines meet. However this is not an equilibrium choice because both developers and households can be better off if they reduce size $h$. Figure 1.D.11a, shows the case of subsidized households and developers. Below $\bar{h}$, developers receive $P\left(1+\delta\right)$ and households pay $P - \tau$. Developers and households increase their utility and profits as a result. The marginal bunching agents are indifferent between getting the subsidy or not. The identification approach in this paper relies on these agents and therefore the main identification strategy is conveyed in Figure 1.6. The idea is that the bunching in the observed equilibrium distribution allows me to recover $\overline{h}$. Therefore, I can observe two points, $\underline{h}$ and $\overline{h}$, on the same indifference curves and offer functions and recover their shape. Figure 1.D.11 shows the optimal choices for other types of developers and types of households.

*Figure 1.6:* Marginally Subsidized and Marginal Buncher Agents' Choices




NOTE: This figure shows the optimal choices for the marginal buncher household and developer. The figures presents the intuition for the identification idea. The gray offer and bid functions represent the indifference curves for the marginally subsidized agents. These are the ones who can increase their profits or utility by increasing or reducing h to take advantage of the subsidy and tax incentives. The demand and supply functions are defined as follows:

$$h^{D} = \begin{cases} h^{*}\left(Y_i, \tau; \theta, \underline{P}\right) & \text{if } Y_i \leq \underline{Y} \\ \underline{h} & \text{if } Y_i \in \left(\underline{Y}, \overline{Y}\right) \\ h^{*}\left(Y_i, \tau; \theta, \underline{P}\right) & \text{if } \overline{Y} \leq Y_i \end{cases} \qquad h^{S} = \begin{cases} h^{*}\left(A_j, \delta; \beta, \underline{P}\right) & \text{if } A_j \leq \underline{A} \\ \underline{h} & \text{if } A_i \in \left(\underline{A}, \overline{A}\right) \\ h^{*}\left(A_j, \delta; \beta, \underline{P}\right) & \text{if } \overline{A} \leq A_j \end{cases}$$

### 1.4.4 Market-Level Supply and Demand

The market level demand and supply is defined by the individual demand and supply represented in Figure 1.6 and the distribution of wealth and productivity. The approach to derive the market-level supply and demand is to use the optimality conditions and the distributions $F_Y$ and $G_A$ and a change of variable formula.[27]

*Graphical Representation* Figure 1.7c shows the product space or developer density, and the exogenous unit supply function. Figure 1.7a shows an example of the equilibrium density when $f_Y$ and $g_A$ follow a log-normal distribution. The equilibrium price makes the product of the functions in figures 1.7c and 1.7b to match the demand density in Figure 1.7c. The observed density function suggests that the market equilibrium has a discontinuous density and that this stylized model can explain the observed equilibrium represented in Figure 1.4.

*Productivity and Income Mapping to Housing Size.* Households and developers only differ in wealth $Y_i$, and productivity $A_j$. If $h^*\left(Y_i, \tau; \theta, \underline{P}\right)$ is strictly monotone, there is a one to one mapping between $Y_i$ and $A_j$ and the optimality conditions.[28]

$$Y_i = \tilde{Y}\left(h, \tau; \theta, \underline{P}\right) = h^{*-1}\left(h_i, \tau; \theta, \underline{P}\right) \tag{13}$$

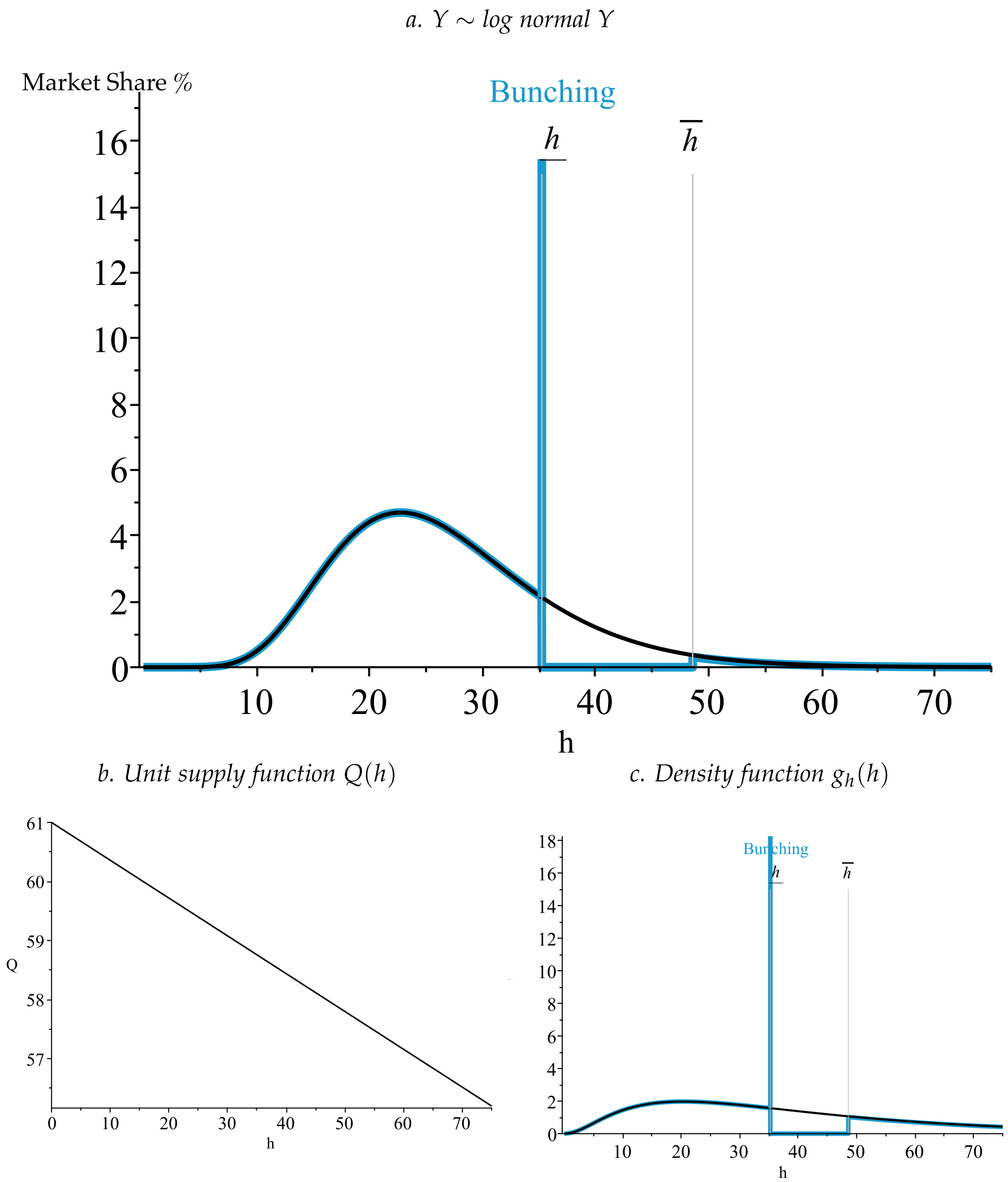
$$A_j = \tilde{A}\left(h; \beta, \delta\right) = h^{*-1}\big(A_j, Q\left(h\right); \beta, \delta, \underline{P}\big) \tag{14}$$

*From distribution of income and productivity to a size distribution.* The share of households and developers choosing $h$ determines the market-level demand and supply

[27] Heckman et al. (2010, p.1571) derives the market demand and supply densities in this way for the case without a notch.

[28] This mapping from housing consumption to income is a consequence of the assumption $\theta_i = \theta\, \forall i$. If I allow heterogeneity in $\theta$, the same demand for housing $h$ can come from different combinations of $Y_i$, $\theta_i$.

*Figure 1.7:* Equilibrium Density, Developer's Choice Density and the Unit Supply Function




Note: This figure shows the equilibrium market share or distribution of units by standard unit size for a given income density $f_y$ following a log-normal distribution. The bottom 2 figures show the share of developers choosing to build at each unit size and the unit supply function.

densities. Using equations 13 and 14, we can get the distribution of market shares and the product space that satisfies the optimality conditions of the market.

$$f_{h^*} = \begin{cases} f_Y\left(\tilde{Y}\left(h, \tau \neq 0; \theta, \underline{P}\right)\right) \frac{\mathrm{d}}{\mathrm{d}h} \tilde{Y}\left(h, \tau \neq 0; \theta, \underline{P}\right) & \text{if } h < \underline{h} \\ \\ f_Y\left(\tilde{Y}\left(h, \tau = 0; \theta, \underline{P}\right)\right) \frac{\mathrm{d}}{\mathrm{d}h} \tilde{Y}\left(h, \tau = 0; \theta, \underline{P}\right) & \text{if } \underline{h} < h \end{cases} \tag{15}$$

$$g_{h^*} = \begin{cases} g_A\left(\tilde{A}\left(h; \beta, \delta \neq 0\right)\right) \dfrac{\mathrm{d}\tilde{A}\left(h; \beta, \delta \neq 0\right)}{\mathrm{d}h} & \text{if } h < \underline{h} \\ \\ g_A\left(\tilde{A}\left(h; \beta, \delta = 0\right)\right) \dfrac{\mathrm{d}\tilde{A}\left(h; \beta, \delta = 0\right)}{\mathrm{d}h} & \text{if } \underline{h} < h \end{cases} \tag{16}$$

***Densities***

The distributions $f_{h^*}$ and $g_{h^*}$ and the demand and supply functions, $h^{\mathrm{D}}\left(Y_i; \tau, \theta, \underline{P}\right)$ and $h^{\mathrm{S}}\left(A_j, \delta; \beta, \underline{P}\right)$, allow to derive a the market-level demand density function, $D_h\left(h; \tau, \theta, \underline{P}\right)$, and the market-level supply function $S_h\left(h, \beta, \delta\right)$.

*Aggregate Demand density.* The demand for housing at the size limit $\underline{h}$ contains the demand for the *threshold maximizing households*, $f_{h^*}\left(\underline{h}; \tau, \theta\right)$, and *the marginally subsidized households* $\int_{\underline{h}}^{\overline{h}} f_{h^*}\left(h; \tau, \theta, \underline{P}\right) \, \mathrm{d}h$. Finally, there is no demand for housing units with $h \in \left(\underline{h}, \overline{h}\right)$.

$$D_h = \begin{cases} f_{h^*}\left(h;\tau,\theta,\underline{P}\right)\ \mathrm{d}h & \text{if } h < \underline{h} \\ \\ f_{h^*}\left(\underline{h};\tau,\theta,\underline{P}\right)\ \mathrm{d}h \\ \quad + \int_{\underline{h}}^{\overline{h}} f_{h^*}\left(h;\tau,\theta,\underline{P}\right)\ \mathrm{d}h & \text{if } \underline{h} = h \\ \\ 0 & \text{if } \in \left(\underline{h},\overline{h}\right) \\ \\ f_h^*\left(h;\tau,\theta,\underline{P}\right)\ \mathrm{d}h & \text{if } \overline{h} \le h \end{cases} \quad S_h = \begin{cases} g_{h^*}\left(h;\beta,\delta\right)\cdot Q\left(h\right) & \text{if } h < \underline{h} \\ \\ g_{h^*}\left(h;\beta,\delta\right)\cdot Q\left(h\right) \\ \quad + \int_{\underline{h}}^{\overline{h}} g_{h^*}\left(h;\beta,\delta\right)\ \mathrm{d}h\cdot Q\left(h\right) & \text{if } \underline{h} = h \\ \\ 0 & \text{if } \in \left(\underline{h},\overline{h}\right) \\ \\ g_{h^*}\left(h;\beta,\delta\right)\ \mathrm{d}h\cdot Q\left(h\right)\left(h\right) & \text{if } \overline{h} \le h \end{cases} \tag{17}$$

Given the hedonic price function $P(h) = P$ we can use a change of variable formula to get the market distribution in terms of price analogous to Figures 1.3 and 1.4.

### 1.4.5 Market Equilibrium

The housing market achieves an equilibrium $E$ when, a given price scheme $P\left(h\right)$, market-level demand and supply are equal for all values of $h$:

$$E = \left\{P\left(h\right) \in \mathcal{P} : D\left(h;\tau,\theta,\underline{P}\right) = S\left(h;A_j\right)\forall h \in \mathcal{H}\right\} \tag{18}$$

$D\left(h;\tau,\theta,\underline{P}\right)$ to be equal. The equilibrium price function allows the match between types of households and developers that clears the market.

*Existence of hedonic equilibrium.* The existence of a hedonic equilibrium has received comparatively less attention than the identification of this type of model. S. Rosen (1974) and Epple (1987) show that under some specified utility functions, cost functions, and distributions for the unobserved heterogeneity, a closed-form solution for the equilibrium price function exists. Heckman et al. (2010) explicitly describe how the equilibrium price function depends on the distributions of observable characteristics of firms and workers. Ekeland (2010) shows an existence proof and provides a particular example of an equilibrium. Moreover, Bajari and Benkard (2005) prove that in equilibrium, the price of a differentiated product will be a function of its characteristics if the utility is continuously differentiable, monotonic in numeraire, and Lipschitz continuous. Using some particular functional forms, the model presented in this paper can have an analytical solution.[29]

## 1.5 Identification and Estimation

The behavioral parameters to estimate are $\theta$, $\beta$, which describe the curvature of the bid and offer curves. This section explains the identification and estimation of those parameters.

### 1.5.1 Identification of the Structural Parameters

*Marginal Buncher indifference conditions.* The identification argument in this paper follows the approach used by Best et al. (2019) and, more generally, the one suggested by Bertanha et al. (2021) and Blomquist et al. (2021). The idea is that the existence of the marginal buncher allows observing two points in the same bid and offer function.[30] Following the hedonic literature, the identification and estimation approach follows a two-step procedure. In the first step, I use the analysis in Sections 1.4 and 1.3, to obtain the hedonic price function, the notch and the behavioral responses. In the second step, I use the marginal bunching condition to solve for the two parameters of interest from

[29] A particular example is available upon request.

[30] Best et al. (2019) proposition 1 and Bertanha et al. (2021) Theorem 1 both prove identification using the same identification idea as in this paper. The identification in my setting follows the same principal conditional on consistent estimates of the first step. Bertanha et al. (2021) describe the identification assumptions under which we can recover the structural parameters from the observed bunching. They argue that notches allow for the identification of elasticities, whereas kinks need additional assumptions about the unobserved heterogeneity. Blomquist et al. (2021) show the conditions under which elasticities can be identified under notches and kinks. They illustrate their approach using Saez (2010) setting. In contrast to Blomquist et al. (2021) who assume the $pdf$ of heterogeneity is monotone, Bertanha et al. (2021) derive partial identification bounds by assuming the $pdf$ has a bounded slope. Using censored regression models, covariates, and semi-parametric assumptions on the distribution of heterogeneity, they provide point estimation for kink points.

equations 11 and 12.

The estimation approach follows a two-step procedure. The first step estimates the hedonic price function at $\overline{h}$, $P\left(\overline{h}\right)$, the three prices at $\underline{h}$, $P^{\tau}\left(\underline{h}\right)$,$P\left(\underline{h}\right)$, and $P^{\delta}\left(\underline{h}\right)$, and the unit supply function $Q\left(h\right)$ at $\underline{h}$ and $\overline{h}$, $Q\left(\underline{h}\right)$, and $Q\left(\overline{h}\right)$. The second step uses these estimates and the marginal buncher condition to recover structural parameters $\beta$ and $\theta$ as the solutions of the marginal buncher equations. I do not observe $\overline{Y}$ and $\overline{A}$, but I use the fact that, given the assumptions I impose in this paper, there is a one-to-one mapping between $h$ and $Y$, and $A$, see equations 13, 14. This allows me to express $\overline{Y}$ and $\overline{A}$ in terms of observable characteristics. This requires estimates for $\frac{\partial P(h)}{\partial h}|_{h=\overline{h}} = p\left(\overline{h}\right)$ and $\frac{\partial Q(h)}{\partial h}|_{h=\overline{h}} = q\left(\overline{h}\right)$. Table 1.2 shows the functional forms that I use to recover $\theta$ and $\beta$ and the elements that I need to estimate in the first step.

The two unobservable objects are the parameters that describe the utility and cost functions $\boldsymbol{\beta}$ and $\theta$. All values summarized in Table 1.2 panel D, can be estimated. The parameter $\vartheta$ is not directly observed and it is assumed to be $\vartheta = \frac{1}{2}$. The identification of $\boldsymbol{\beta}$, $\theta$ is achieved by solving two equations with two unknowns. The two equations are the ones in Table 1.2 panel A, after replacing panels B and C.

The existence of a marginal buncher allows me to address the main challenge of causal inference; observing the same agent at two different states of the world. Using the insights from the bunching literature, we estimate a counterfactual distribution that allows us to observe those to points in the data. Section 1.5.2 describes the estimation approach for the values described in Table 1.2 panel D, and explains how I apply this intuition to my data.

*Table 1.2:* Functional Form and Identification Equations

| | |
|---|---|
| *A. Marginal Buncher Condition* | |
| Household | $V_D = U\left(\overline{Y} - \overline{P}, \overline{h}; \theta\right) - U\left(\overline{Y} - \underline{P}^{\tau}, \underline{h}; \theta\right) = 0$ |
| Developer | $V_S = \pi\left(\overline{Q}, \overline{A}, \overline{P}; \beta\right) - \pi\left(\underline{Q}, \overline{A}, \underline{P}^{\delta}; \beta\right) = 0$ |
| *B. Functional Forms* | |
| Utility | $U = \left[(1-\vartheta)\cdot C^{\theta} + \vartheta \cdot h^{\theta}\right]^{\frac{1}{\theta}}$ |
| Cost | $B = A_j \cdot Q \cdot h^{\beta}$ |
| *C. Optimality Conditions* | |
| Income | $\overline{Y} = \overline{P} - \left(\frac{\vartheta h^{\theta-1}}{\overline{p}\,(\vartheta - 1)}\right)^{\dfrac{1}{\theta - 1}}$ |
| Productivity | $\overline{A} = \dfrac{\left(\overline{P}\cdot\overline{q} + \overline{p}\cdot\overline{Q}\right)\overline{h}^{(1-\beta)}}{\overline{q}\cdot\overline{h} + \overline{Q}\cdot\beta}$ |
| *D. First Step Estimates* | |
| Marginal buncher thresholds | $\underline{h} = P^{-1}\left(\underline{P}\right)$ and $\underline{h} = P^{-1}\left(\underline{P}\right)$ |
| Hedonic price | at $\underline{h}$: $\underline{P}^{\tau} = P\left(\underline{h}\right) - \tau, \underline{P} = P\left(\underline{h}\right)$ $\underline{P}^{\delta} = P\left(\underline{h}\right)\cdot(1+\delta)$ |
| | at $\overline{h}$: $\overline{P} = P\left(\overline{h}\right), \overline{p} = \frac{\partial P(h)}{\partial h}\|_{h=\overline{h}}$ |
| Unit Supply Function | at $\underline{h}$: $\underline{Q} = Q\left(\underline{h}\right)$ |
| | at $\overline{h}$: $\overline{Q} = Q\left(\overline{h}\right), \overline{q} = \frac{\partial Q(h)}{\partial h}\|_{h=\overline{h}}$ |

NOTE: This Table summarizes the functional forms used for the estimation of $\beta$ and $\theta$. $\vartheta$ is assumed to be $\frac{1}{2}$, but section 1.8 shows sensitivity to different numbers. This parameter corresponds to the share of expenditure on housing.

### 1.5.2 Estimation

*Observed Equilibrium.* Figure 1.8 shows the joint densities of unit size and market price for all cities around 2006 when the subsidy notch on the demand side was small and around 2016 when the subsidy was twice as big. In each market, heterogeneous agents buy and sell different housing units. The same money may buy larger housing units in separate submarkets; therefore, agents cluster at different housing sizes for which the sale price is at or below the cutoff point. This figure would be the analog of figures 1.6 1.7a, if the only characteristic of a house was size. However, there are other characteristics such as neighborhood quality or structural parameters such as the number of rooms or the availability of extra space such as a studio or a porch. Therefore, to apply the framework to the data, I need to reduce the characteristics of the house into a single characteristic and consistently estimate the implicit or hedonic price of size.

### 1.5.3 Hedonic Price for Housing Size

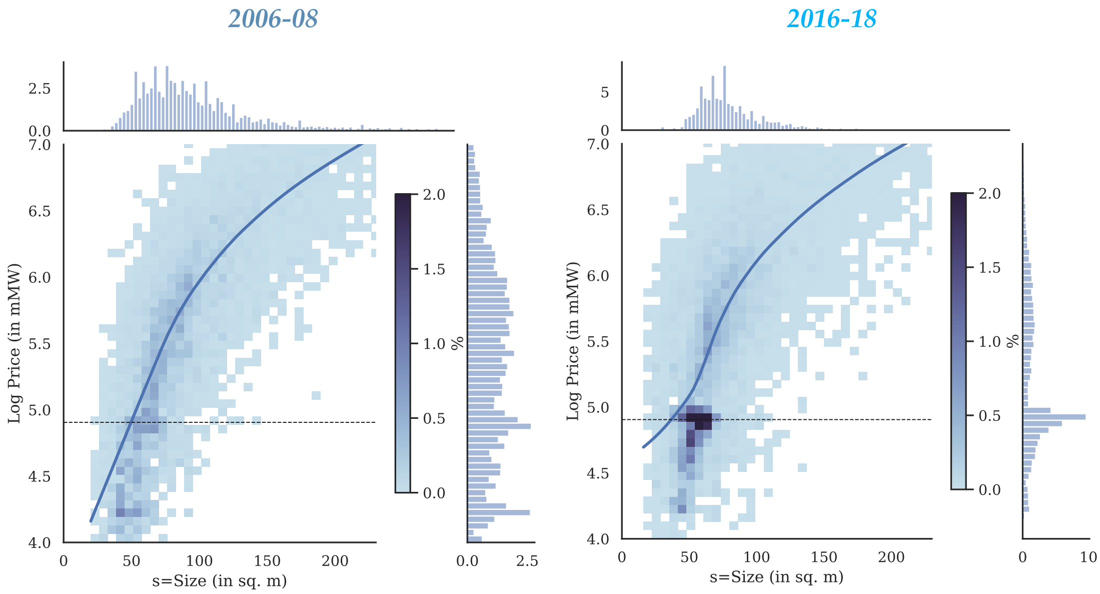
The estimation of the hedonic price function, subsidy notches, and the bunching analysis from section 1.3 are the two key components of the identification and estimation approach of this paper. The hedonic price function plays two roles. First, it allows to recover the function $P(h)$ and marginal willingness to pay for the size of the house $p(h)$ contitional on the other observed characteristics. Second, it allows me to use the bunching on the total price described in Section 1.3 to recover $\underline{h}$ and $\overline{h}$.

The solid line in Figure 1.8 shows that the non-parametric bivariate relationship between price and size is positive. This pattern follows the expected positive relationship and suggests that it could be nonlinear. However, this unconditional relationship may not represent the marginal equilibrium willingness to pay for housing size. There could be observable and unobservable characteristics that affect size and price, creating bias. I follow common practice in the hedonic literature to estimate the equilibrium implicit–or hedonic– price of housing size.[31] Equation 19 represents a general specification for the price function. Where $s_{ltc}$ is the size of the house, $X_{ltc}$ is a vector containing all other characteristics of the house, and $\omega_{ltc}$ represents the residual containing unobserved characteristics.

$$P_{ltc} = P(h_{ltc}) + \Gamma' X_{ltc} + \varepsilon_{ltc} \tag{19}$$

[31] Bishop and Timmins (2019), Bajari and Benkard (2005), Epple et al. (2020) or Bajari, Fruehwirth, Kim, and Timmins (2012)

*Figure 1.8:* Observed Market Equilibrium




NOTE: This figure shows the joint and marginal densities for housing size (x-axis) and price (y-axis). Darker dots inside the graph represent a higher market share. The figure contains all available cities in each period, and all the different unit types, that is, single-family homes, multifamily homes, condos, two bedrooms, one bedroom, and so forth. The solid line represents the non-linear relationship between housing size and price (using lowess).

I observe a type of unit $l$ in the city $c$ at time $t$. I assume that the housing price is additive and separable in the size of the house $s_{ltc}$, the observable characteristics are included in $X_{ltc}$, and $\varepsilon_{ltc}$ represents the unobserved characteristics. $X_{ltc}$, includes location, quality, number of rooms and neighborhood quality index (*estratos*),[32] among others. $P(.)$ is the implicit price function for the size of the housing. I follow M. D. Cattaneo, Crump, Farrell, and Feng (2019b) and M. D. Cattaneo, Crump, Farrell, and Feng (2019a) to estimate the function $P(h)$ non parametrically. Their approach also allows me to estimate $p(h)$ non parametrically.[33]

To estimate the implicit or hedonic price, I rely on independence conditional on ob-

[32] The *estratos* are codes from 1 to 6. They summarize the quality of the block, for more details, see Uribe (2021)

[33] An alternative estimation method Robinson (1988). We could also use a parametric approximation.

servable characteristics:[34]

$$\mathrm{E}\left(h_{ltc}|X_{ltc},\varepsilon_{ltc}\right)=0 \tag{20}$$

*Independence conditional on observables.* It is common to rely on conditional independence to recover the implicit price function of a certain characteristics. In my setting, I observe a rich and unique set of controls. This includes the exact location of the unit and general characteristics of the house, including the number of rooms and the neighborhood quality index. The assumption of conditional independence can be problematic in many settings. For example, Chay and Greenstone (2005) show that using a hedonic model to recover the marginal willingness to pay for air quality without using instruments generates biased results. Omitted variables could generate a bias in the current setting. However, I present two facts that are reassuring. First, in contrast to air quality, the hedonic regression does not show the opposite of the expected sign. Second, when I include characteristics, such as an indicator function equal to one, if the a house has an extra bathroom, or a studio or a porch, the magnitudes of the coefficients do not change. This type of characteristic is potentially unobserved by the econometrician in other settings, as it is related to size; so, it is reassuring that including it does not affect the size of the coefficients. However, this does not rule out the fact that other omitted variables could bias the results. For example, if changes in price generate the bunching with no change in size, the error term could be correlated with size, particularly for observations around the price cutoff.

[34] Bajari and Benkard (2005) propose three different identification assumptions; i) Independence conditional on observables, ii) Option packages and iii) instruments. My setting and data allows an implementation of each of the three identification approaches. However, the results presented in this paper rely on the first condition.

### 1.5.4 Unit Supply Function Notch

How do developers respond? One of the principal objectives of the economic model is to address this question in more detail. Developers built more housing units when they built smaller housing units. One advantage of the data is that I observe the number of units built by unit type; therefore, I can get empirical estimates of the trade-off between unit size and the number of units and account for it in the model. I follow a similar strategy that I use to estimate the hedonic regression to estimate this relationship $Q(h)$ nonparametrically.

$$Q_{ltc} = Q(h_{ltc}) + \Omega' X_{ltc} + \epsilon^{Q}_{ltc} \tag{21}$$

Like in the case of the hedonic regression estimation, I rely on independence conditional on observables.

$$\mathrm{E}\left(h_{ltc} | X_{ltc}, \epsilon_{ltc}\right) = 0 \tag{22}$$

I estimate $Q(h)$, nonparametrically using the approach proposed by M. D. Cattaneo et al. (2019b) and M. D. Cattaneo et al. (2019a). This approach also allows me to estimate the derivatives $q(h)$ that I required in the estimation of the structural parameters. In the set of controls, I include detail characteristics of each project such as number of towers, location of the lot, etc. The assumption of conditional independence is plausible in the sense that after controlling by all characteristics, the relationship between $h$ and $Q$ is given exogenously by regulatory constraints or the existent technology.

### 1.5.5 Marginal Bunching Thresholds:

The relevant observable characteristics described in panel D of Table 1.2, need to be evaluated at $\underline{h}$ and $\overline{h}$, which are not directly observable but can be recovered from the data using the estimates of the hedonic price function $\hat{P}(h)$ and the values of $\underline{P}$ and $\overline{P}$ recovered in Section 1.3.

$$\underline{h} = \hat{P}^{-1}\left(\underline{\hat{P}}\right) \text{ and } \overline{h} = \hat{P}^{-1}\left(\overline{\hat{P}}\right) \tag{23}$$

The implicit assumption is that I am comparing standard unit housings that only vary in size. The size distribution that would mirror the bunched price distribution is the distribution of the standardized size, not the observed size distribution. See Appendix 1.D.1 for an example with a parametric specification for $P(h)$.

### 1.5.6 Missing Mass: Model vs. Data

The model predicts a missing demand and supply for housing units between $\underline{h}$ and $\overline{h}$. However, in my setting, I only observe a partial missing mass in the distribution. This partial missing mass is common in bunching analysis using notches (Best et al., 2019; Kleven & Waseem, 2013). This is usually attributed to at least two potential factors, optimization frictions or heterogeneity in the behavioral parameters $\theta$. Some households may not be aware of the subsidies, or the application costs may be too high. In my setting, there are a limited number of subsidies and not all eligible households receive it. It is also the case that some households receive the downpayment and the interest rate subsidy, but others get only one of the two. This means that the notch may vary between individuals due to different types of frictions. Moreover, households that are eligible may not see the benefits because living in a low-cost housing unit could create stigma and households may have a large dis-utility related to that.

There may be a preference heterogeneity across cities of family size. In this case, Best et al. (2019) suggests that the behavioral response can be interpreted as the average marginal response.

## 1.6 Results

This section presents the main estimation results. First, the equilibrium characterization, corresponding to the first step of the two-step estimation procedure is presented. Second, the structural estimates corresponding to the second step of the proposed estimation approach is presented.

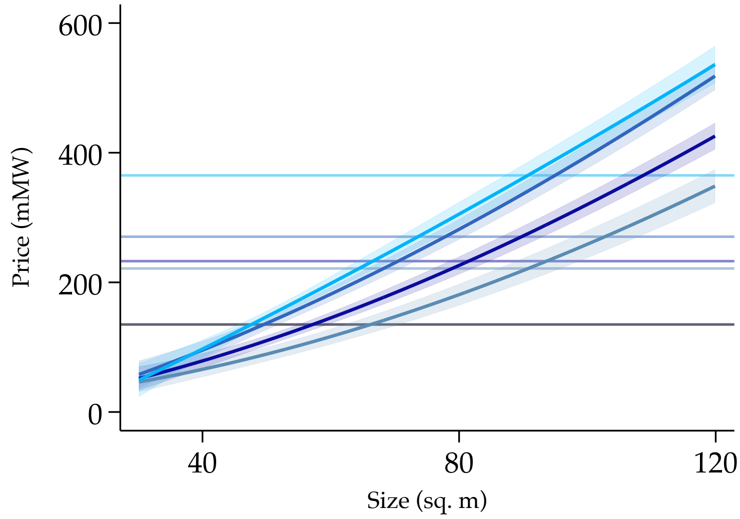
### 1.6.1 First Step: Equilibrium Characterization

*Describing observed equilibrium prices.* Figure 1.9 illustrates the estimated implicit price function $\hat{P}(h)$ and the marginal willingness to pay $\hat{p}(h)$ by size. The figure shows a change in the equilibrium price scheme. It is not possible to know if this change is only associated to the policy changes as other general demographic and economic factors changed during the same time period. Over my study period, housing became more expensive but particularly above the policy cutoff. The figure also shows that accounting for non-linearities in the estimation of $P(h)$. Note that in contrast to figure 1.8, I show the prices in levels and not logs. In terms of the marginal price for size, Figure 1.9b shows a difference, particularly around area of the marginally subsidized households.

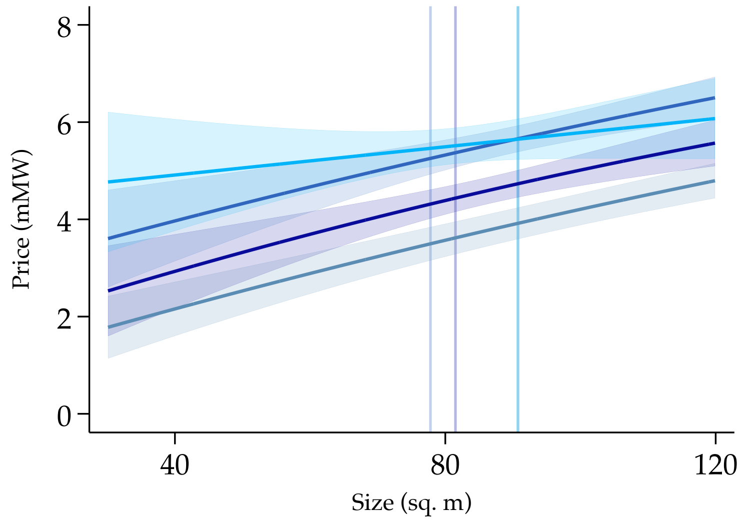
*Marginal bunching thresholds.* The vertical lines show the value of $\overline{P}$ for the different periods and $\underline{P}$. We can see in this figure how the estimation of the hedonic price function allows us to recover the marginal buncher thresholds in terms of size. This figure shows that the marginal buncher is is willing to cut the size of the housing he buys almost in half to take advantage of the subsidy, which represents up to a 25 percent of the value of housing at the cutoff. This figure also shows that the equilibrium size that you can buy with the 135 $mMW$ decreases overtime. In 2006, you could buy a house of around 66 square meters whereas in 2016 with the same money you can only buy a house of 47 square meters.

*Figure 1.9:* Hedonic Price for Housing Size

*a.* $P(h)$



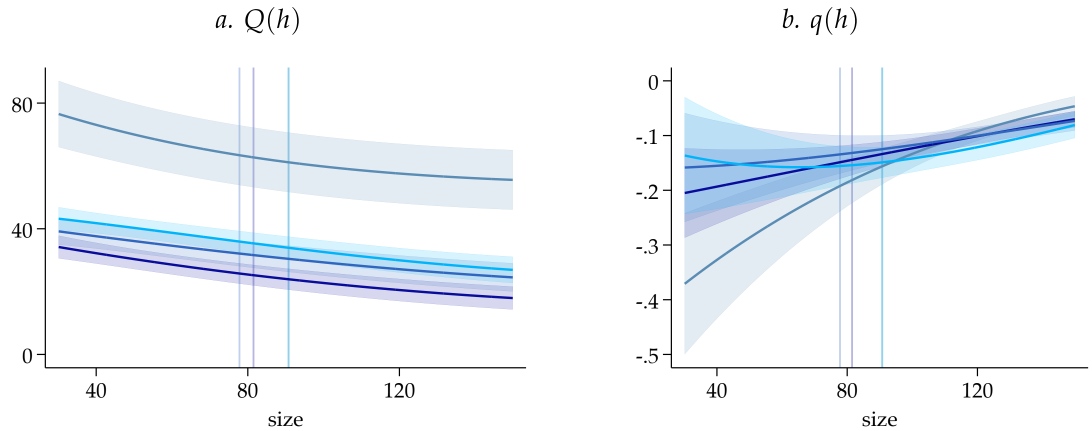

*b.* $p(h)$




NOTE: This figure shows $\hat{P}(h)|X$, and $\hat{p}(h)|X$ where $X$ includes number of bathrooms, number of rooms, an indicator equal to 1 if the unit is a building, location; dummy variables equal to one if the unit has a porch, studio, storage unit, dressing room, service room, dining and living room, fireplace, kitchen, clothes areas, patio; location coordinates interacted with town fixed effects and metropolitan area fixed effects, lot size, number of building blocks, apartments per floor, number of floors, total parking spots, and number of building units. To estimate these figures I use the approach outlined in M. D. Cattaneo et al. (2019a).

*The three prices.* Figures 1.2 and Table 1.1 together with the estimates in Figure 1.9 allow me to recover the developer's price $P^{\delta}(h)$ and the household's price $P^{\tau}$.

*Figure 1.10:* Unit Supply Function




NOTE: This figure shows the bin scatter for the number of units and for unit size after controlling for observable characteristics. In this figure, I use the same controls as in Figure 1.9. This figure includes the observations for all years and all cities.

*Unit Supply Function.* Figure 1.10 shows the unit supply function adjusted for the characteristics of the unit and the project using observations from all cities available in each period. The figure shows a negative relationship between unit size and the number of units, which is intuitive. Developers face a trade-off between building more but smaller units and fewer but larger units. There was a decrease in the number of units built at all levels. I do not have a clear explanation of this phenomenon, but it could be associated with lower availability of land, increases in the cost of building high or the fact that for the first years, the census covered mostly the main metropolitan areas whereas later years started to include smaller cities.

### 1.6.2 Second Step: Estimation of $\theta$, and $\beta$

Using the functional forms and estimates for the values in panel D of Table 1.2 and presented in section 1.6.1, I can solve for $\theta$ and $\beta$. The marginal buncher functions do not have a closed-form solution; therefore, I use numerical methods to find the values of $\theta$ and $\beta$. I present the estimates separately for each subperiod with specific subsidy schemes.

*Structural Parameters.* Figure 1.11 illustrates the equilibrium of the housing market and the preferences of households and the technology of developers using the estimated parameters presented in Table 1.3. The parameter $\sigma = 1/(1-\theta)$ represents the constant elasticity of substitution for the specified utility function (CES). It represents how the relative consumption of housing varies when the relative price changes.

*Households' parameters.* The elasticity of substitution estimates was around 1.2 at the beginning of the period and increased substantially to 3.8 at the end of the period. This could be explained by the introduction of the Mi Casa Ya program later in the period. Under this program, subsidies became available to informal employees and applicants automatically received both the downpayment subsidy and interest rate subsidy. The estimated parameters are similar across years, which is reassuring considering that these are economic fundamentals and, therefore, very unlikely to drastically fluctuate over time. The increase in the estimated parameter at the end of the period is likely given the changes in the policy and the fact that informal employees were now eligible.

An elasticity of 1 corresponds to a Cobb-Douglas elasticity. Therefore, my estimates suggest that a Cobb-Douglas utility function would not be a bad representation, but would be imprecise, particularly at the end of my period. A negative value of $\theta$ corresponds to an elasticity of substitution $\sigma$ less than one, which means that housing and other goods are gross complements. If $\theta$ is positive, the elasticity of substitution is greater than one, and the housing and consumption of other goods would be gross substitutes.

Bayer et al. (2007) present an approach that integrates the hedonic insights into a discrete choice framework. As pointed out by Yinger (2015), their approach implicitly assumes a linear utility function, which violates the strict quasi-concavity postulate. In other approaches in the urban economics literature, the utility function is assumed to be Cobb-Douglas. In my setting, I allow for a less restrictive functional form, but my estimates suggest that the Cobb-Douglas utility function would be a close approximation in some cases but not always.

*Figure 1.11:* Equilibrium Choices using the estimated parameters

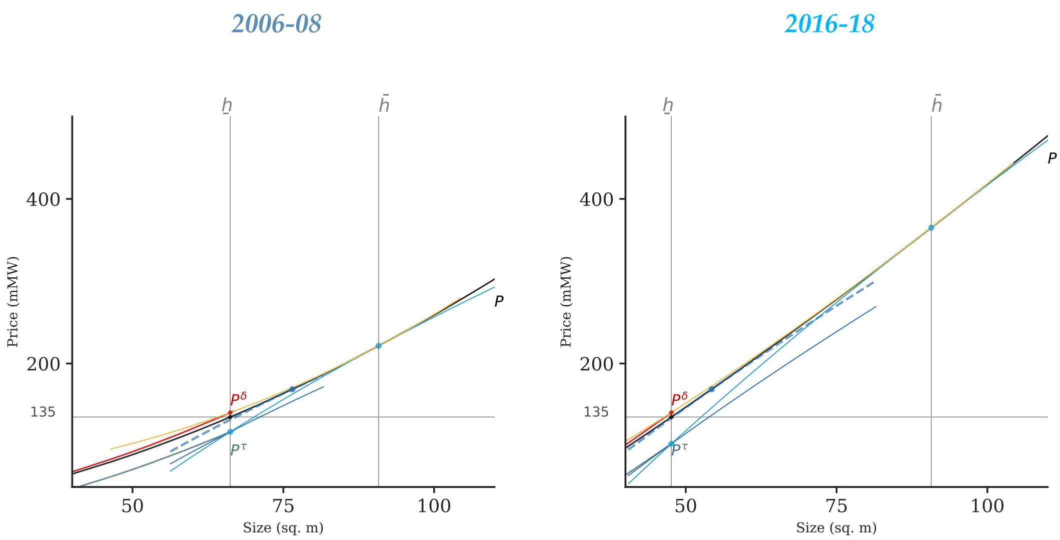



*Table 1.3:* Structural parameters

| | 06-08 | 09-11 | 12-15 | 16-18 |
|---|---|---|---|---|
| $\beta$ | 2.34 | 2.03 | 1.65 | 1.29 |
| $\theta$ | 0.55 | 0.40 | 0.55 | 0.74 |
| $\sigma$ | 2.23 | 1.68 | 2.22 | 3.88 |

NOTE: This figure uses the estimated parameters presented in Table 1.3 and creates the empirical analog of Figure 1.6 for the marginally subsidized households and developers. The figure represents the equilibrium choices and bid and offer functions estimated at the beginning and end of the study period. The elasticity of substitution implied by the CES utility function is $\sigma = 1/(1-\theta)$. $\vartheta$ is assumed to be 0.5.

*Developers' parameters.* On the developer side, the estimated parameter $\beta$, does not change much overtime. In the first period, $\beta$ is 2.34 , this decreases to around 1.26

in the following periods. This change means that the costs of building bigger houses decreased over time. It is hard to compare these estimates to the literature, as the paper that use an hedonic approach to estimate the housing market usually takes the supply function as given and does not allow for heterogeneity (Bishop & Timmins, 2019) or do not allow for product differentiation (Saiz, 2010). There are not many papers that consider the developer's decisions regarding how many units to build and which unit to build separately.

## 1.7 Welfare and Policy Evaluation

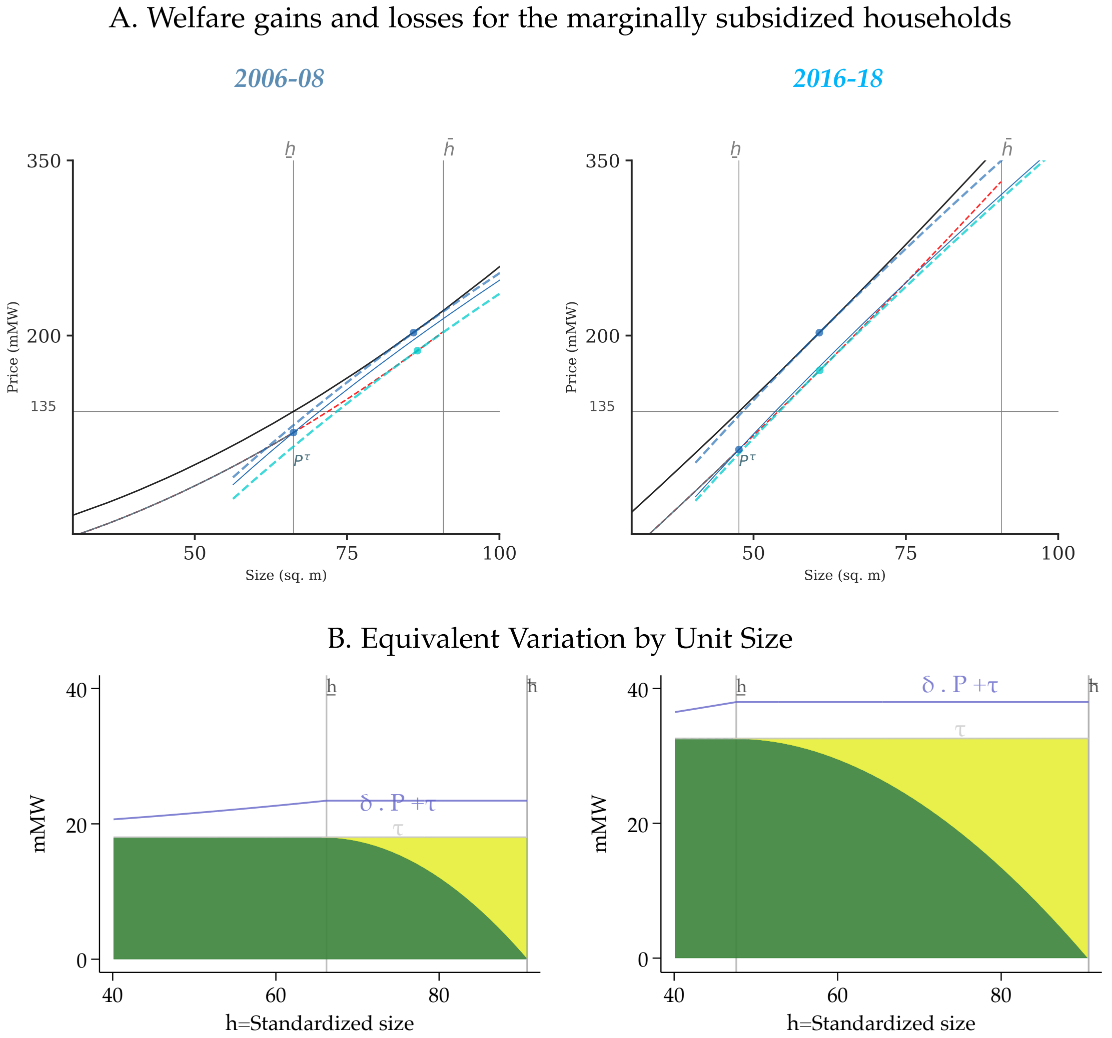
A goal of this paper is to evaluate the effectiveness of the policy scheme implemented in Colombia. The framework and estimated parameters presented in this paper allow for a different type of counterfactual policy evaluation and allows for an assessment of how much households and developers benefit from these policies. In this section, I illustrate the potential of this framework as a policy evaluation tool. I use it towards two aims. First, I compare the how much the government spends on these subsidies to how much the beneficiary households are willing to pay to increase their utility in an equivalent magnitude. The focus on the developer side also allows me to evaluate the efficiency loss induced by the notched subsidy scheme (Blinder & Rosen, 1985). Second, I explore the role of the subsidy on the supply side. I show what happens if the supply side subsidy is removed, and demonstrate that the notch incentive designed in Colombia requires this subsidy to prevent a shortage problem. However, once we account for this additional government expenditure, it is not clear that households value the subsidy enough to justify the government expenditure in these type of policies. I close the section with a discussion of other type of policies that could be evaluated, like the effect of a quality or size limit on the subsidized units.

### 1.7.1 Benefits and Efficiency Losses: the households' perspective

To evaluate the benefits of the subsidy, I focus on the effect of marginally subsidized households. Because the market response is a change in housing characteristics, the price limit creates efficiency losses. If the response was a pure price reduction without changes in characteristics, the policy would induce a transfer of welfare from developers to households. Figure 1.11 illustrates how a representative marginally subsidized household benefits from the subsidy. Households reduce their expenditure on housing to obtain the subsidy. By doing this, the household reaches a higher utility level, the indifference curve moves to the right from the dashed line to the solid blue line. Without the existence of the price limit and if the marginally subsidized household gets the subsidy without being forced to reduce its consumption, a household could increase its utility even more, as illustrated in the graph. This means the notched scheme introduces an inefficiency. However, from a targeting perspective, this type of inefficiency could be justified. Without the price cutoff,richer households could receive the subsidy to buy expensive units which undermines the objective of the program. That is to provide opportunities for low income households to become homeowners.

Blinder and Rosen (1985) shows examples where notches can be preferred to alternatives targeting approaches such as slope changes. In that paper we explore under what circumstances a notch scheme is preferable to a conventional linear subsidy. They show that the elasticity of substitution matters when comparing different targeting approaches. They show that a lower elasticity of substitution increases the efficiency of notches. Based on the elasticities estimated in this paper, it seams that the effectiveness of these subsidy scheme is fading over time as the elasticity of substitution is getting bigger. Moreover, I show that the efficiency losses (yellow area in 1.12) is larger at the end of the period. The dollar amount households would pay for the increase in utility is around the same value of the notch $\tau$, however, once we include the extra cost

*Figure 1.12:* Welfare Gains and Efficiency Losses for the Marginally Subsidized Households




Note: Panel A illustrates the changes in utility for a representative marginally subsidized household.Panel B shows how much households are willing to pay for their increase in utility (evaluated at $\underline{h}$ for households at different levels. The green are are the welfare gains and the yellow area represents the efficiency losses induced by the notch scheme. )

induced by the tax refunds to developers we can see that the households do not value the changes in housing units as much as the governement expenditure in a per unit basis. To calculate the total losses and gains we could multiply the share of buncher households by size and calculate their welfare gains and benefit losses.

***What happens in equilibrium.***

An advantage of this paper is that it allows me to think about market equilibrium. For example, in Section 1.7.1 I showed the efficiency losses that would arise if there is a notched scheme that could be reduced with a linear subsidy. What happens in that case with the developers that bunched? Under that scenario, there would not be any need to have tax incentives for developers. To show how these are very relevant in the type of subsidy scheme in Colombia, the next section explores what happens if the tax incentives are removed.

### 1.7.2 The effect of Removing Developers' Tax Incentives

An important policy debate related to housing policy is whether the use of tax incentives for developers is an effective redistributive tool. In the USA there is the Low Income Housing Tax Credit (LIHTC) which is intended to produce. These types of subsidies could be ineffective or very expensive. They could also benefit developers more than households which would maket them hard to justify. Soltas (2021) shows that these types of subsidies could be very expensive as they force to build low-income housing in expensive areas. Sinai and Waldfogel (2005) shows that Tenant-based housing programs, such as Section 8 Certificates and Vouchers, are more effective than project-based programs such as developers' subsidies. The type of developers subsidy implemented by the Colombian government is a little different, however, as it coexists with a demand-side subsidy targetted at households. The existence of these subsidies is an active policy debate in Colombia. I use the framework developed in this paper to show that these subsidies are required to avoid a shortage problem in a priced-capped policy scheme like the one implemented in Colombia.

In 2021, under the need for tax reforms, there was a policy proposal to remove these subsidies. However, developers actively opposed them, claiming that this would create a shortage problem.

*"If these items are repealed, in Valle del Cauca we would go from having an offer of low-cost housing and sales of 23,000 homes, average year, to one of sales of 4,600 homes"* [35]

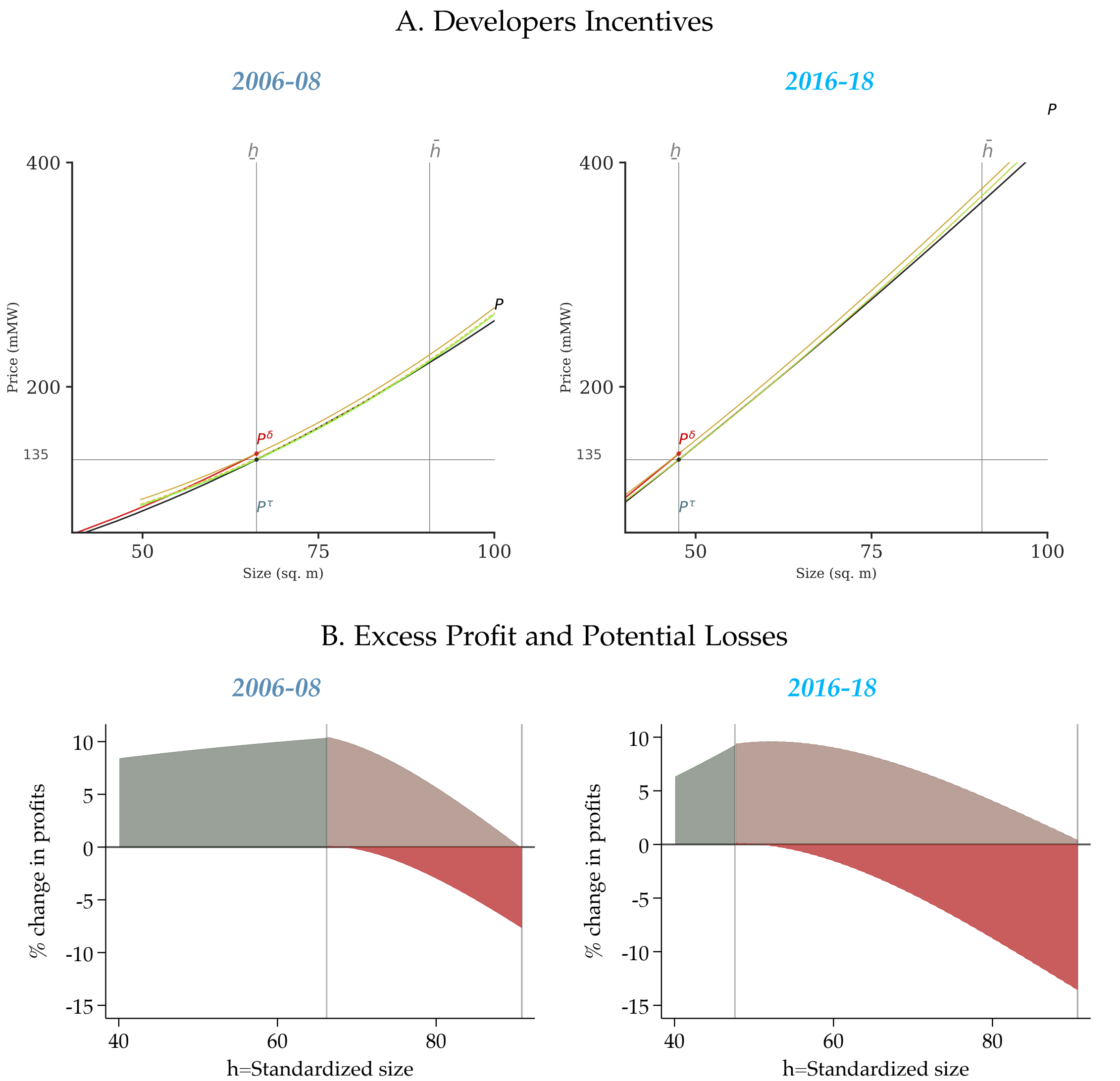
The framework developed in this paper shows that this could be the case. Figure 1.13 ilustrates the role of these tax incentives. Without the tax incentive, and under the existence of the price cap, developers that would produce housing units of size $\underline{h}$ and $\overline{h}$ would face no demand or a reduced demand. They build cheaper housing units to keep supply the households that changed the type of housing they buy. However, because they would be building units usually build by more productive developers they would have to reduce there profits or leave the market. In this sense the tax incentives guarantees that developers do not leave the market. The dashed yellow line, represents the equilibrium in the absence of the subsidies, the solid yellow line is the observed response and the green line represents the iso-profits if developers stay in the market but they cannot increase the price.

Figure 1.13 shows these responses at the beginning of my study period and at the end. The top panel illustrates the decision choice of a representative marginally subsidized developer under the observed scenario and two counterfactual scenarios, changing the type of housing to satisfied the subsidy induced demand at $h$ and the scenario under no subsidies. The figure in panel B, shows what would be the changes in profits for households producing different housing units if they do not get the subsidy and reduce their consumption. The figure show that at the begining of the period the losses would have been around 5 percent. However, at the end the losses could be up to 15 percent of the profits they would have in the absence of the subsidy scheme. The gray-red area is the increase in profits that they receive instead. The gray would produce low cost housing even in the absence of the subsidy. This analysis shows that the tax incentives may in fact prevent the exit of some developers and avoid a shortage

[35]source: El Tiempo (2021)

problem. However, by doing this they artificially increase the profits of developers that would build low-cost housing even in the absence of the subsidy and they make the potential exiters better off than in the absence of the subsidy.

*Figure 1.13:* Developer Response to Tax Incentives




NOTE: Panel A shows the incentives of a marginally subsidized developer if the tax incentives are removed. Panel B) shows the developers if there is not tax incentives as a percentage of their current profits (in red) and the induced excess profits to prevent the exit of those developers (in Green)

## 1.8 Robustness and Sensitivity Analysis

This section presents sensitivity analysis for the bunching and structural estimates.

### 1.8.1 Bunching Estimates and Structural Parameters

The estimation of the structural parameters in this paper rely heavily on the estimation of the counterfactual distribution of market shares. These estimates depends on the selection of different parameters. The bandwidth for the bins, the number of omitted bins to the right and the left of the cutoff and the degree of the polynomial. To select these parameters, I fix a bandwidth and select the the excluded number of bins to minimize the difference between the excess mass and missing mass. Figures 1.B.8 and 1.B.9 in the appendix show the bunching analysis forthree different bin size and 2 different criteria to select the excluded bins and polynomial degree. Table **??** shows the sensitivity of the structural estimates to different approches to estimate the bunching and the upper limit for the marginal buncher $\overline{P}$. It also shows sensitivity to different values of $\vartheta$ which represents the share of income devoted to housing. Overall the parameters are relatively stable to different approaches to estimate bunching. Regarding the share of income devoted to housing. The table shows that different values of $\vartheta$ does not affect the value of the elasticity of substitution. As expected, the elasticity drops as the share of consumption devoted to housing falls.

*Table 1.4:* Structural Parameters Using Different Estimation Approaches for $P$ and Different Values for Consumption Shares $\vartheta$labeltab:main$_{t}able_{s}ens$

| | $\vartheta = 60$ | $\vartheta = 50$ | $\vartheta = 40$ | $\vartheta = 30$ | |
|---|---|---|---|---|---|
| | $\sigma$ | | | | $\beta$ |
| 06-08 | | | | | |
| $bw = 0.05$ (at $\underline{P}$) | 1.53 | 1.32 | 1.19 | 1.09 | 2.70 |
| $bw = 0.05$ (all P) | 0.80 | 0.72 | 0.66 | 0.61 | 5.00 |
| $bw = 0.03$ (around $\underline{P}$) | 2.52 | 2.23 | 2.08 | 1.99 | 2.34 |
| $bw = 0.07$ (around $\underline{P}$) | 2.34 | 2.07 | 1.91 | 1.81 | 2.35 |
| 09-11 | | | | | |
| $bw = 0.05$ (at $\underline{P}$) | 1.13 | 1.00 | 0.90 | 0.82 | 2.51 |
| $bw = 0.05$ (all P) | 1.32 | 1.18 | 1.08 | 0.99 | 1.13 |
| $bw = 0.03$ (around $\underline{P}$) | 1.78 | 1.57 | 1.43 | 1.33 | 2.08 |
| $bw = 0.07$ (around $\underline{P}$) | 1.96 | 1.73 | 1.59 | 1.49 | 2.00 |
| 12-15 | | | | | |
| $bw = 0.05$ (at $\underline{P}$) | 2.06 | 1.85 | 1.72 | 1.63 | 1.69 |
| $bw = 0.05$ (all P) | 2.64 | 2.46 | 2.37 | 2.32 | 1.25 |
| $bw = 0.03$ (around $\underline{P}$) | 2.27 | 2.06 | 1.94 | 1.86 | 1.67 |
| $bw = 0.07$ (around $\underline{P}$) | 2.43 | 2.22 | 2.11 | 2.04 | 1.65 |
| 16-18 | | | | | |
| $bw = 0.05$ (at $\underline{P}$) | 2.47 | 2.28 | 2.17 | 2.11 | 1.32 |
| $bw = 0.05$ (all P) | 3.64 | 3.54 | 3.51 | 3.50 | 1.29 |
| $bw = 0.03$ (around $\underline{P}$) | 1.69 | 1.50 | 1.38 | 1.29 | 1.40 |
| $bw = 0.07$ (around $\underline{P}$) | 3.64 | 3.54 | 3.51 | 3.50 | 1.29 |

## 1.9 Conclusions

This paper presents compelling evidence of the market responding to subsidies. I rely on detailed data on the universe of new housing, data on subsidies to both households and developers, the policy cutoff inducing discontinuous incentives, and the variation of the subsidy over time. I use the behavioral responses induced by the subsidy and introduce a novel identification approach to estimate a hedonic housing market equilibrium with heterogeneous agents and housing that rationalizes the observed responses. The model-guided estimation approach translates the bunching reflecting the behavioral responses and the reduced form estimates into parameters of both household utility and developers' production function. I use those estimates to illustrate the type of welfare analysis that the estimation approach allows.

I find that households and developers changed their housing consumption to take advantage of the policy. The price cap, which could be important if the response does not induce a change in housing consumption, induced welfare losses. Households would have been better off if they received the subsidy without reducing their housing consumption. The welfare analysis also suggested that in a world with developer heterogeneity, subsidizing the demand side of the market may be insufficient. Developers need to be compensated to produce low-cost housing, which they can produce but at a higher marginal cost. The type of welfare analysis allowed by this approach goes beyond the examples presented here. Because I recovered the income and productivity levels of households and developers together with parameters describing their preferences and costs, the approach allow for the evaluation of different housing policies. The method could apply to other markets with vertical differentiation and price caps, such as labor markets. In this case, the policy induced a change in the type of housing bought and built. The housing stock accumulated smaller housing units purchased by households who would prefer bigger houses. Considering that housing is a durable asset that affects urban structure and city planning, this could translate into significant consequences for cities reaching a suboptimal equilibrium. The findings of this paper suggest that a careful evaluation of the market structure matters for effective policy design. Understanding how the policy affects the housing market's incentives is crucial to understanding how the observed equilibrium outcomes inform us about the effects of the policy.

# Appendix

## 1.A Colombian Housing Policy: Additional Details

### 1.A.1 Minimum wage and Inflation

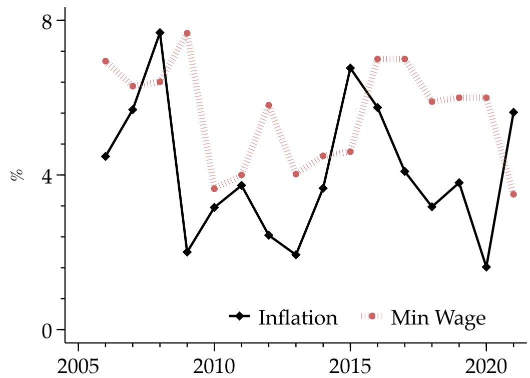
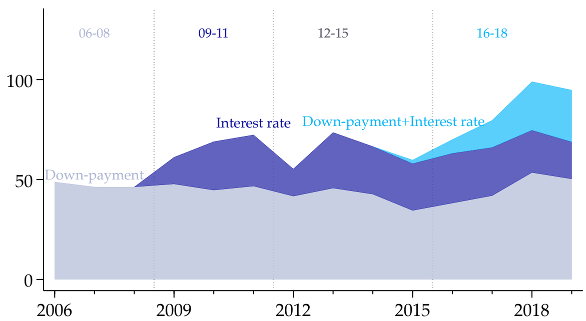



*Figure 1.A.1:* Inflation and Minimum Wage Over-Time

### 1.A.2 Number of Subsidies over Time

*Figure 1.A.2:* Total number of Subsidies Over Time




Source: Minvivienda, FRECH

### 1.A.3 Details on Subsidies

*downpayment subsidy.* The down payment subsidy was introduced at the beginning of the nineties and is available to formal employees who contribute to the family compensation funds.[36] The **gray blue** area in Figure 1.1 shows the number of subsidies and total government expenditure from 2006 to 2019. The number of downpayment subsidies to formal employees was more or less stable during the study period, but the government spending increased in 2015 due to an increase in the size of the subsidy. Only formal households earning less than four times the minimum wage (mMW) are eligible for the subsidy, and the subsidy can only be used to buy a low-cost housing units.

*Interest rate subsidy.* In 2009, the government introduced a program to subsidize mortgage' interest rates. This program, called *FRECH*, started as a program to incentivize economic growth after the crisis, but it became a permanent policy. In contrast to the downpayment subsidies, interest rate subsidies were also available to households buying housing units above the $135\times\ mMW$ threshold. However, the subsidy is larger if the households buy a low-cost housing unit, that is, the price is less than $135 \times\ mMW$. If a household receives the subsidy, the government pays the bank the corresponding amount during the first seven years of the loan. Three different schemes existed during the study period, but in all the schemes, there was a discontinuity in the subsidy at the cutoff defining low-cost housing. The **dark blue** area in Figure 1.1 shows the number of subsidies and total government expenditure from 2011 to 2019. The subsidies were more or less stable over time; around 20,000 households received this subsidy. This subsidy represents lower government expenditure and expenditure has slightly decreased overtime partly due to lower interest rates.

In contrast to the downpayment subsidy, interest rate subsidies were also available to

36

households buying housing units above the 135×m-MW threshold. However, there is a notch at 135×m-MW. Figure 1.A.3 shows the interest rate subsidies for the all the house price ranges.

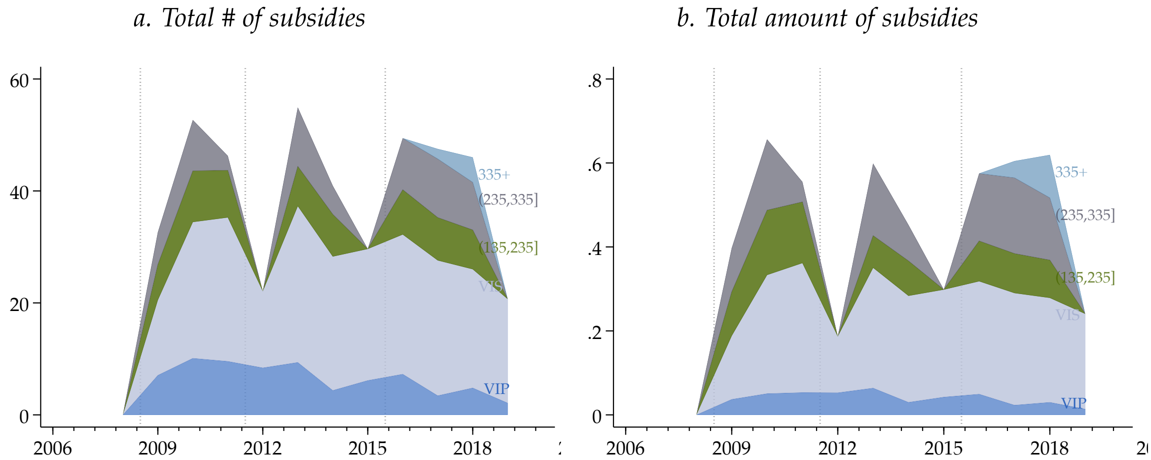



*Figure 1.A.3:* FRECH subsidies for different price ranges

*The subsidy expansion–Mi Casa Ya.* In 2015, the government doubled the effort and introduced a new program *Mi Casa Ya*, (My House Now). Before this program was introduced, the down payment subsidy was only available to formal employees-contributing to the family compensation funds.[37] This program extended the coverage of the downpayment to noncontributing households. The households participating in this program get the downpayment subsidy and interest rate subsidy automatically with a single application. The **light blue** area in Figure 1.1 shows the number of subsidies and government expenditure, which is the sum of the downpayment and the total expenditures with the interest rate discount. The figure shows that the increase in the number of subsidies and government expenditure that started in 2015 was mainly driven by the introduction of this program and the increase in the down

[37]In theory informal household could get access to housing subsidies. However, *fonvivienda*, the institution in charge of these subsidies, assigned mostly to vulnerable populations. The vulnerable populations are displaced by armed conflict and affected by natural disasters. I include the equivalent plots for those subsidies in the Appendix 1.A.

payment subsidy to formal employees.

*Supply subsidy–value added tax (VAT) tax refund.* A couple of years after the demand subsidies were introduced, to encourage developers to build low-cost housing, the government introduced a VAT tax refund. Developers get up to 4 percent of the sale price of each unit in the refund of taxes paid on construction materials. I include this subsidy in the analysis. Accounting for this subsidy introduces discontinous incentives on the supply side.

*Other subsidies.* The Colombian housing policy includes other subsidies excluded from the main analysis of this paper.[38] These are mainly subsidies to disadvantaged populations. These subsidies exist to follow a constitutional mandate to provide housing to people affected by forced displacement and environmental disasters. They are for cheaper housing units and households in extreme poverty. These subsidies can be used to buy *priority low-cost housing*, which is housing units with a market price of $70 \times mMW$ or less. The approach of using subsidies as an incentive to promote construction and purchase of housing units was mostly ineffective to provide this type of housing. As a result, in 2014, a program to build 100'000 free housing units was launched. The goal was to satisfy the constitutional mandate and provide housing to the disadvantaged population that was neglected by the previous policy approaches.A. G. Gilbert (2014) describes this program, *100 mil viviendas gratis*, and evaluates its potential effectiveness. Camacho et al. (2020) study the effect of this conditional transfer on the economic outcomes of the receiving households. The appendix Figure **??** shows the evolution of those subsidies.[39] The program of 100 thousand free housing units occurred between 2012-2015. There is a program for rural housing and subsidies for the military that I ignore in this paper.

38

39

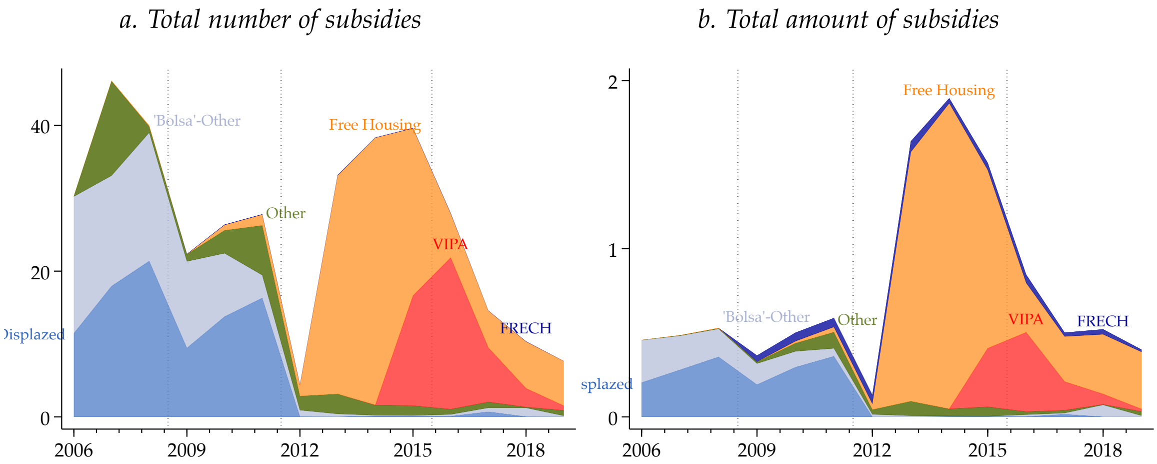


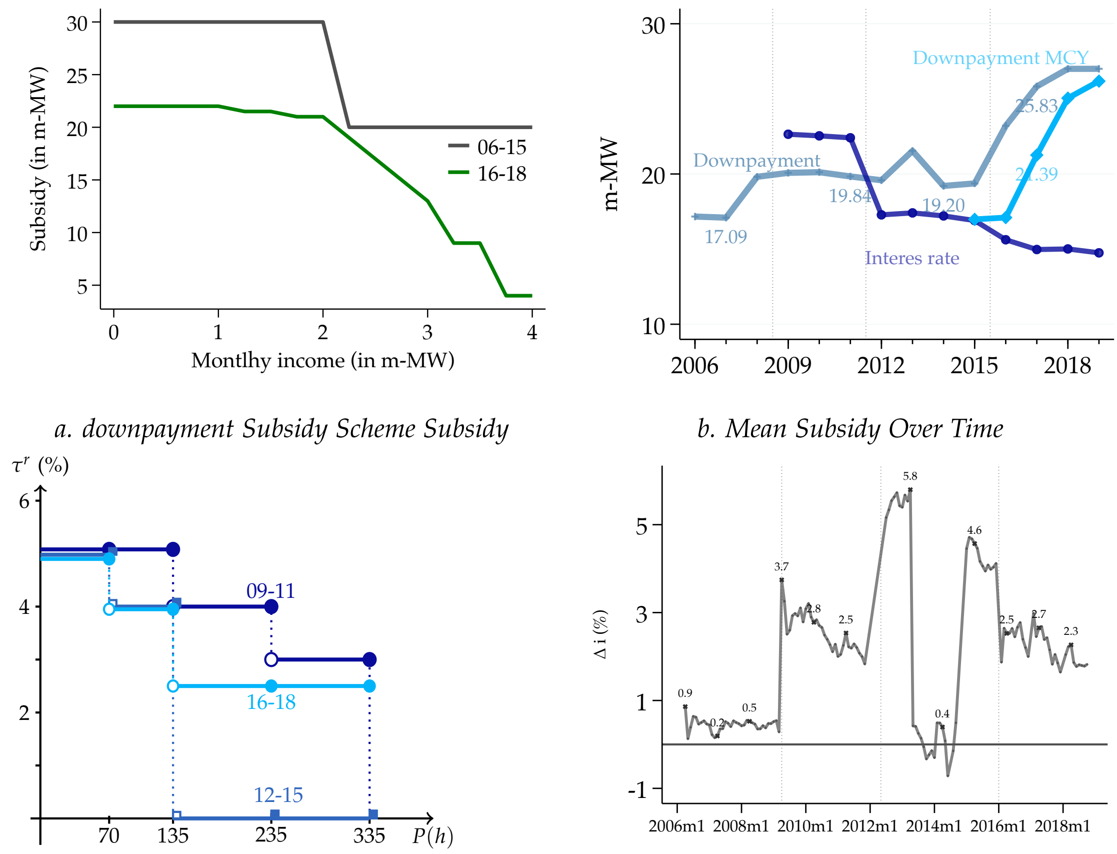

*Figure 1.A.4:* Subsidies for the Vulnerable Population. Housing Units Priced Below 70 mMW.

*Subsidy scheme.* Figure 1.A.5c shows the subsidy scheme for the interest rate. Three different schemes existed during my study period. Each scheme is represented in the figure by a different line. The x-axis is the monthly minimum wage and the y axis is the discount in the interest rate. If a household gets the subsidy, the government pays the bank the corresponding amount during the first seven years of the loan.

*Targeting instruments.* The authorities use two different tools to determine eligibility; the households' income and the total price of the housing unit. A unit can be subsidized only if the market price is below the low-cost housing threshold, 135 times the monthly minimum wage (m-MW). This arbitrary threshold is the same for all cities.[40] Regarding income, only households earning below four times the monthly minimum wage can get the subsidy. Figure 1.A.5a shows the subsidy scheme. Before 2015, the subsidy was decreasing on income, and the maximum possible subsidy was $22 \times$ m-MW. In 2016 the generosity increased, the limit increase to $30 \times$ m-MW for individuals with income below $2 \times$ m-MW and $20 \times$ m-MW for individuals with

40

income between $2 - 4 \times$ m-MW. As the Figure 1.1 shows, the increase in the limit is reflected in higher government expenditure. Figure 1.A.5b the average subsidy during my study period. We can see that the average subsidies were about 20 percent before 2015 where the mean subsidy is about 26 percent.




*a. downpayment Subsidy Scheme Subsidy*

*b. Mean Subsidy Over Time*

*c. Interest rate subsidy scheme*

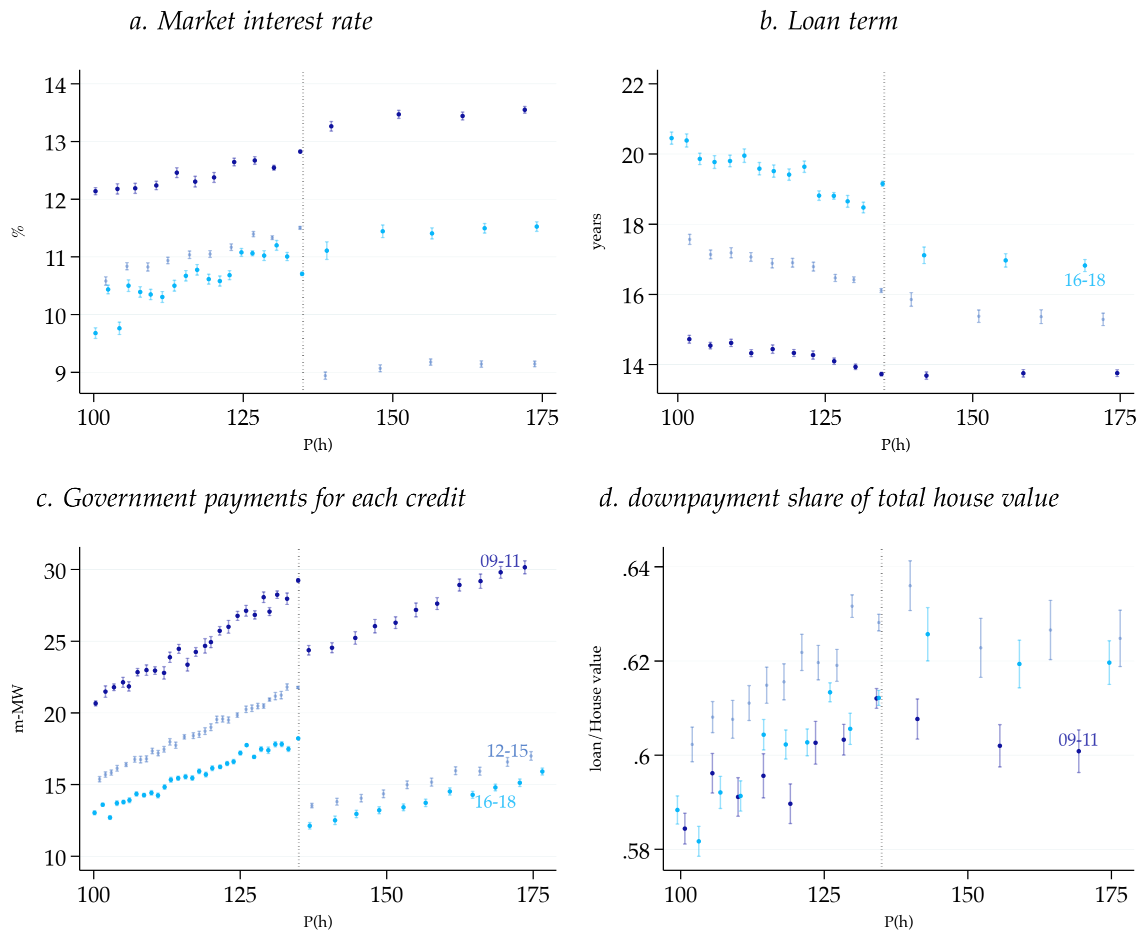
*d. Differences in Market interest rate i* $\tau^r$

*Figure 1.A.5:* Subsidy Scheme and Observed Mean Differences

NOTE: This figure shows the subsidy scheme and the evolution overtime of the subsidies for the interest rate and downpayment subsidy.

## 1.A.4 Mortgage terms:

*Figure 1.A.6:* Loan Terms by House Prices




NOTE: This figure shows the subsidy scheme and the evolution overtime of the subsidies for the interest rate and downpayment subsidy.

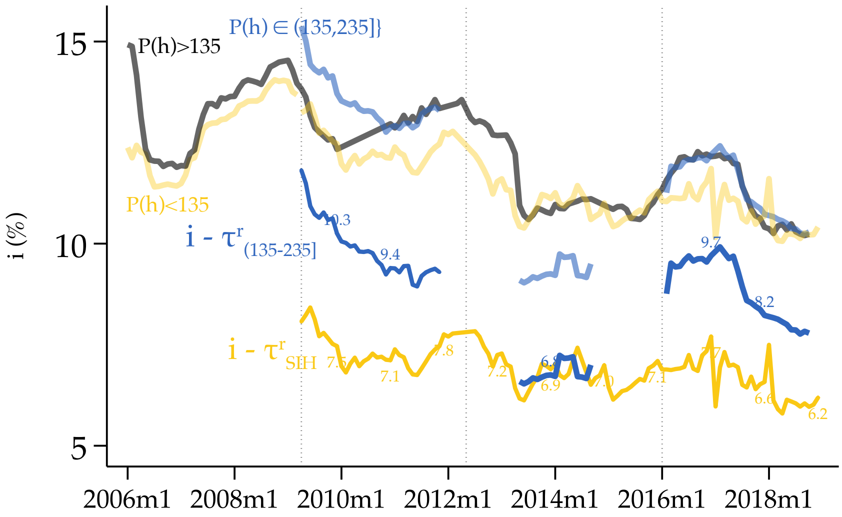



*Figure 1.A.7:* Market interest rate $i$ and subsidy $\tau^r$

NOTE: This figure shows the interest rates with and without the subsidy over time.

*Observed differences in monthly payments.* I use the administrative records of these subsidies and administrative records on all loans to check that subsidies are reflected in the lower interest rates paid by households. The administrative records for the subsidies contain relevant information about the mortgages. It has the market interest rate $i$, the loan $L$, the term $n$, the discount in the interest rate $\tau^r$, and the house price $P$. The administrative records for all loans contain less detailed information, but I observe the interest rate of each loan and the average loan amount. I use the loans for housing, which have an indicator variable equal to 1 if the house is low-cost housing and 0 otherwise.

## 1.B Bunching Estimation and More Bunching Results

### *Counterfactual Distribution Estimation*

In contrast with Figure 1.3, in this section, I present the market shares by standardized unit size, and not unit price, and by period. By doing that, I can interpret the changes induced by the subsidies as changes in the size of a standard unit. Developers and households build and purchase smaller houses to take advantage of the subsidy.[41]

Define $T(h)$ to be the difference between the observed $(f_{h^*})$ and counterfactual densities $(f_{h_0})$,

$$T(h) = f_{h^*} - f_{h_0} \tag{24}$$

and $\Delta h$ as the maximum change agents made to take advantage of the subsidy.

$$\Delta h = \overline{h} - \underline{h} \tag{25}$$

*Intuition.* The counterfactual is the distribution that would exist in the subsidy's absence. I calculate it fitting a flexible polynomial to the observed density and excluding the observations close to the cutoff. The differences between the counterfactual distribution and the observed distribution reflect the behavioral responses to the subsidy scheme's discontinuous incentives.

*Estimation.* To estimate $f_{h^*}$ and $f_{h_0}$, I rely on standard techniques from the bunching literature. To estimate the empirical distribution $\hat{f}_{h^*}$, I calculate the share of units in each bin $h_b$ of size $2 \cdot \epsilon$,

$$h_b = \frac{1}{N} \sum_{l=1}^{N} \mathbb{1}\left[h_l \in (b - \epsilon, b + \epsilon)\right] \tag{26}$$

[41] I assume all changes are in terms of the housing size of a standard unit. Although I acknowledge this is a strong simplifying assumption, it makes the analysis tractable. The setup and economic framework introduced in section 1.4 can be the basis to extend the analysis to separate changes in multiple characteristics, such as location or quality, and to include the possibility of changes prices in response to a setting with developer frictions or imperfect competition.

The estimated observed equilibrium distribution is

$$\hat{f}_{h^*}(h) = h_b$$

To estimate the counterfactual distribution, $\hat{f}_{h_0}$, I predict the observed values for $h_b$ using a flexible polynomial, $l(h_b) = \sum_{p=0}^{T} \iota_p h_b^p$ and excluding a region around the cutoff The function $o(h_b; L, H)$ includes all the indicator variables for the bins between $L$ and $H$, the lower and the upper bound,respectively, of the excluded area. $o(h_b; L, H) = \sum_{k=L}^{H} \mathbb{1}[h_k = h_b] h_b$

$$h_b = l(h_b) + o(h_b; L, H) + v_b \tag{27}$$

*Counterfactual distribution.* The counterfactual distribution is the predicted density using only the flexible polynomial.

$$\hat{f}_{h_0} = \hat{l}(h_b) = \sum_{p=0}^{T} \hat{\iota}_p h_b^p \tag{28}$$

*Bunching.* Using the estimated distributions, I can get an expression for bunching or excess mass at $\underline{h}$, and calculate the maximum behavioral change induced by the subsidy $\Delta h$:

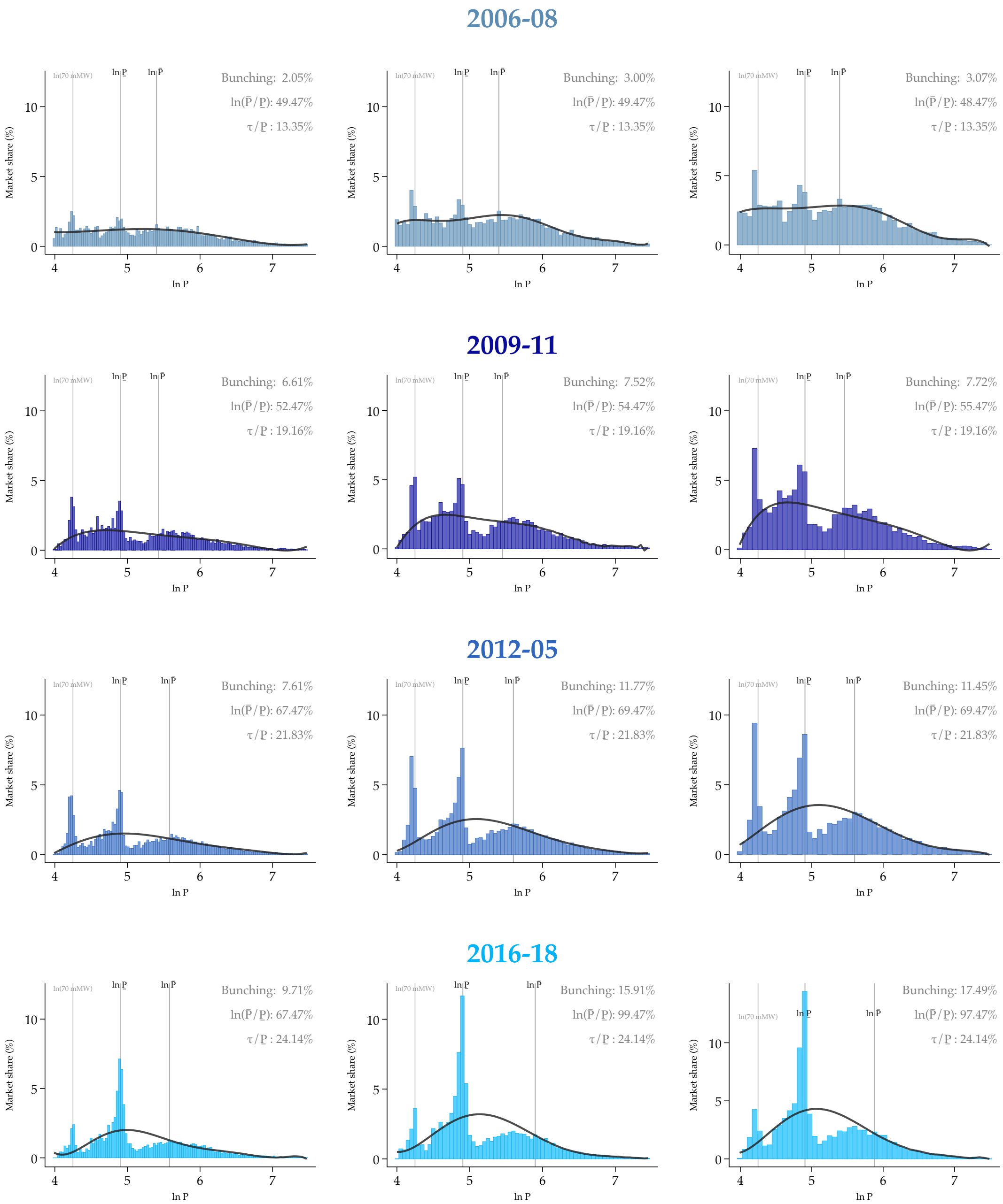
$$\hat{T}(\underline{h}) = \hat{f}_{h^*}(\underline{h}) - \hat{f}_{h_0}(\underline{h}) \tag{29}$$

Equation 29 is the difference between the observed distribution and the counterfactual distribution at the discontinuity point, $\underline{h}$, and it represents the share of individuals who would consume $h \in (\underline{h}, \overline{h})$ in the absence of the subsidy, but consume $h$ in a subsidy scenario.

*Maximum behavioral response.* The maximum behavioral response, $\overline{h}$, is obtained when the counterfactual and observed distributions coincide:

$$\overline{h} = min[h : h > \underline{h} \text{ and } \hat{f}_{h_0}(\underline{h}) - \hat{f}(h_b) = 0]$$

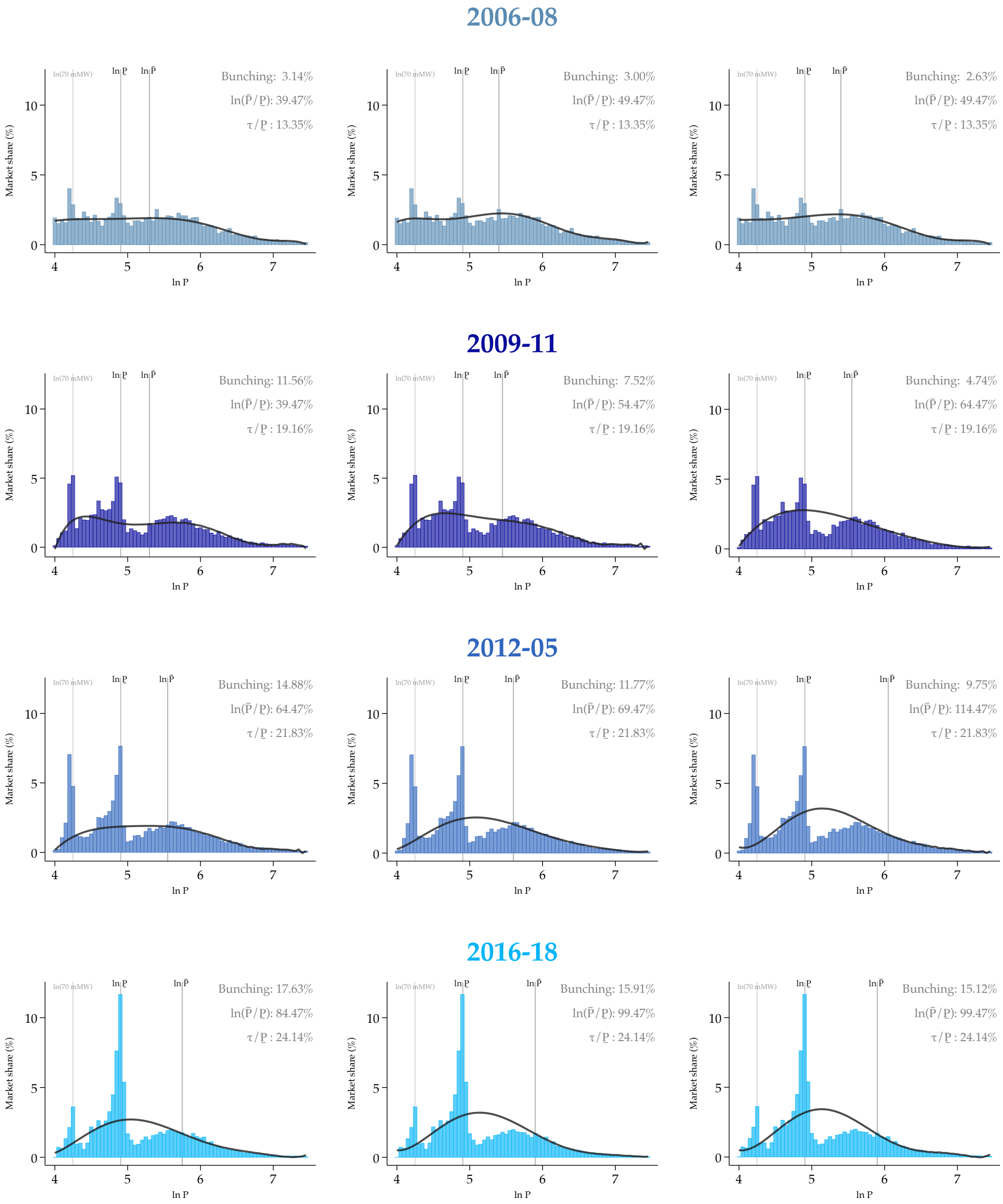
*Figure 1.B.8:* Bunching Over Time Using Different Binsize




*1) binsize0.03* *2) binsize* 0.05* *3) binsize 0.07*

NOTE: This figure shows the distribution or the market share of housing units by sale price (expressed in log of mMW). The lines are the cutoffs defining *low-cost housing* $\underline{P} = 135$ mMW and *priority low-cost housing* 70 mMW. The additional lines shows the point, $\overline{P}$, where the counterfactual and observed distribution coincide again after the cutoff. The figure shows for the different period for all available cities.

*Figure 1.B.9:* Bunching Over Time Using Different Criteria of Missing=excess mass to Select Estimation Parameters

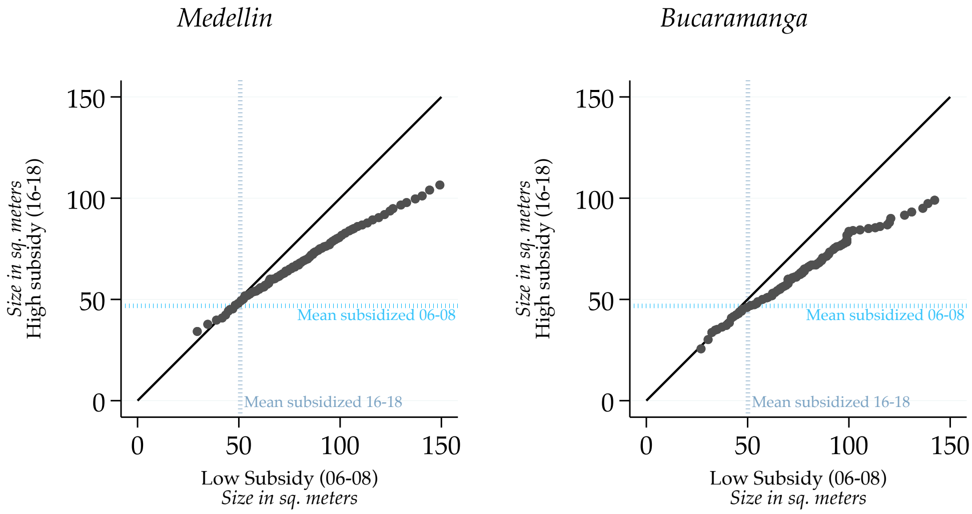



*1) At $\underline{P}$* *2)* Around $\underline{P}$* *3) All distribution*

Note: This figure shows the distribution or the market share of housing units by sale price (expressed in log of mMW). The lines are the cutoffs defining *low-cost housing* $\underline{P}$ = 135 mMW and *priority low-cost housing* 70 mMW. The additional lines shows the point, $\overline{P}$, where the counterfactual and observed distribution coincide again after the cutoff. The figure shows for the different period for all available cities.

## 1.C Response on size: Additional cities.

*Figure 1.C.10:* Quantile-to-Quantile Plots of Housing Size: Low versus High Subsidy Periods



## 1.D Model Appendix:

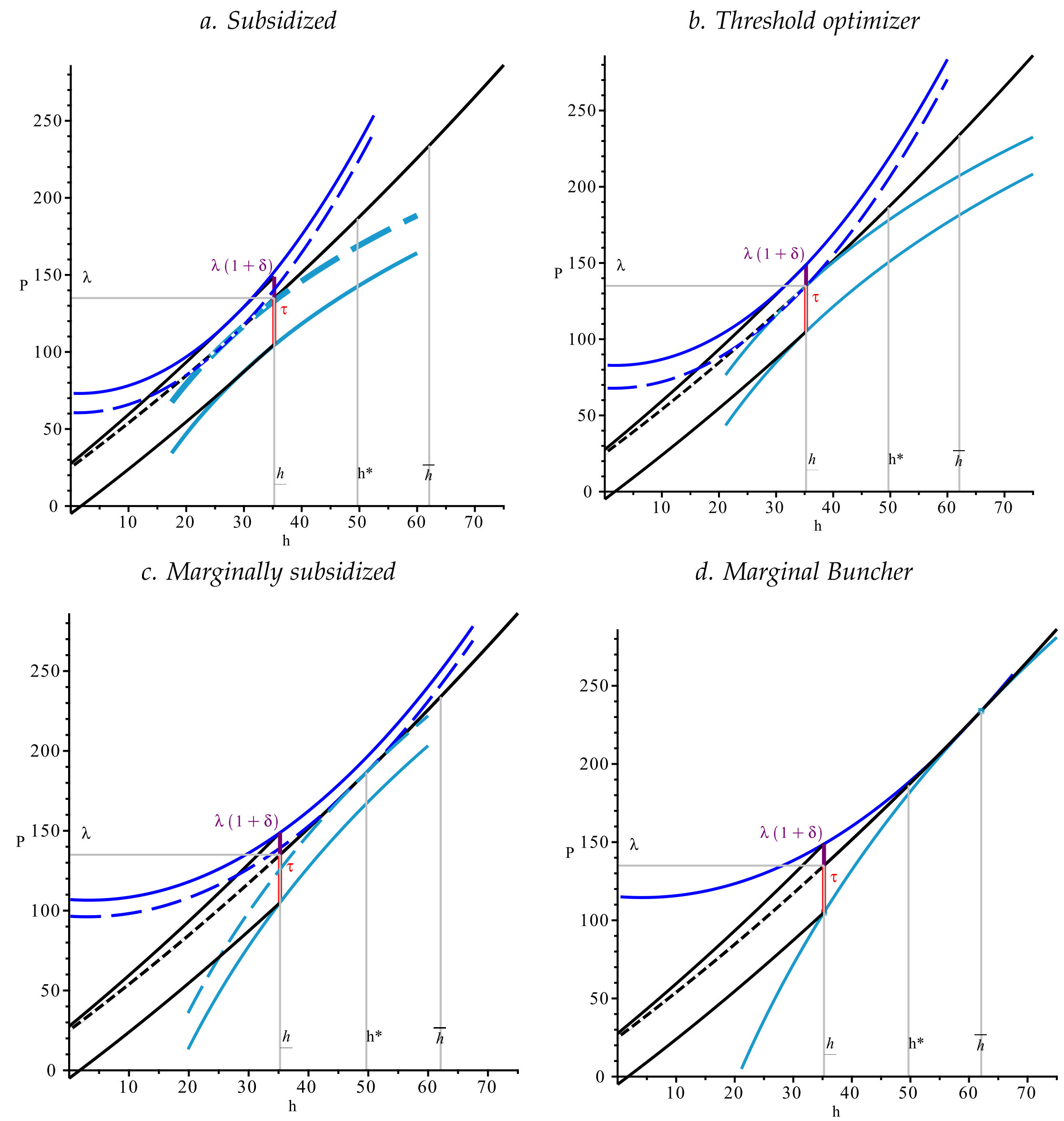



*Figure 1.D.11:* Graphical representation of equilibrium choices

The offer function functional form is: $\varphi_j^S = \frac{\bar{\pi}_j \cdot h^2}{\beta_1} + A_j$. This results from a cost function $C\left(Q(h), A_j\right) = A_j * Q(h)$, and $Q(h) = \frac{\alpha}{h^2}$

### 1.D.1 Standardized Housing and Unit Size

To make all the housing units comparable, I use the hedonic price function to standardize all housing units. In particular, I use the estimates of equation 19 to convert all housing units into a standard unit.

This hedonic price estimation decomposes the unit price into observed and unobserved characteristics. The standardized housing size, which I call $h$, is the size of a housing unit with average characteristics that will cost the same as the observed price.

$$\rho\left(h_{ltc}\right)+\Gamma'\bar{X}+\bar{\omega}=\rho\left(s_{ltc}\right)+\Gamma'X_{ltc}+\omega_{ltc} \tag{30}$$

$\bar{X}$,are the means of the observable characteristics and $\bar{\omega}$ equals the average residual. Solving for $h$ in the equation 30, I get the following measure of the standardized size measure:

$$h_{ltc}=\rho^{-1}\big(\rho\left(s_{ltc}\right)+\Gamma'\left(X_{ltc}-\bar{X}_{ltc}\right)+\left(\omega_{ltc}-\bar{\omega}_{ltc}\right)\big) \tag{31}$$

Intuitively, this means that if a house is more expensive because it has certain amenities or more bathrooms, I convert this characteristic into the equivalent square meters that the household could get if they had a standard house.

In my application, I standardize the units in a way that $\bar{P}=\rho\left(\bar{s}_{ltc}\right)+\Gamma'\bar{X}_{ltc}+\bar{\omega}_{ltc}$ is the observed average price for the average house. For the implicit price function, I use a parametric approximation $\rho(s)=\rho_0+\rho_1\cdot s+\rho_2\cdot s^2$.

Figure 1.9 shows the functional form of the estimated price function for the four different periods. The Figure 1.9 shows that the implicit price function has become steeper overtime.

*Implied Maximum Size of a Standard Subsidized Unit*

0.32

$$135 \times mMW = \lambda = \rho\left(\underline{h}\right)$$

Given a particular assumed functional form,

$$\underline{h} = \frac{-\rho_1 + \sqrt{\rho_1^2 - 4 \cdot \rho_2 \cdot \left(\bar{P} - \lambda\right)}}{2 \cdot \rho_2}, \tag{32}$$

In Figure 1.9, $\underline{h}$ corresponds to the value of h at which the implicit price intersects the price cutoff (gray horizontal line).

# Chapter 2

# The Effect of Location Based Subsidies on the Housing Market

***Abstract***

*This paper estimates the effect of a location-based price subsidy of utilities and property taxes on the housing market in Bogotá, Colombia. In Bogotá, neighborhood blocks are divided into 6 subsidy codes using arbitrary cutoffs on a block quality score. I use the discontinuity introduced by the cutoffs in a Regression Discontinuity Design to study the effect of the subsidy on the housing market. I find that blocks receiving a higher subsidy have newer houses - implying more construction in such areas. I also find a capitalization of the subsidy into housing prices: properties in areas receiving a small subsidy are cheaper than those receiving a bigger subsidy. These results suggest that a careful evaluation of location-based subsidies' incidence needs to consider potential capitalization into the housing market and other unintended effects such as new construction or renovations.*

# 1 Introduction

A challenge for social assistance programs and redistributive policies is to find the poor. It is particularly challenging in settings with high informality levels, where income-based, means-tested transfers do not get to the right people.[42] An alternative targeting tool is geographic location, such as neighborhoods, states, or cities. This targeting approach is particularly appealing to developing countries, where self-employment and economic informality are prevalent. Yet, location-based subsidies could affect people's decisions about where to live, and this could affect the housing market. We need to know how the housing market is affected before evaluating the efficacy of place-based redistributive policies. Standard urban economics theory would predict the subsidy's capitalization into the housing prices (i.e., housing prices in the subsidized areas are higher).[43] If the subsidized areas become more expensive, the subsidy benefiting the landowners could affect tenants. The subsidy can also attract more people and new construction, affecting city growth. The significance of these effects is an empirical question and the main purpose of this paper.

I use a location-based subsidy scheme in Bogotá, Colombia and I quantify the effect of the subsidies on the housing market. Bogotá uses the characteristics of neighborhoods to target subsidies on public utility services (i.e. water, gas, electricity, phone, and internet), and to charge differential rates on property taxes. Each city block receives a subsidy code based on neighborhood block quality scores. There are six subsidy

[42] See Hanna and Olken (2018) for a summary of the literature studying these proxy-mean tests to target subsidies.

[43] For example, E. L. Glaeser and Gottlieb (2008) suggest that "if high prices and low amenities offset high wages in a spatial equilibrium, there is nothing particularly equitable about taking money from rich places and giving it to poor places." In a recent paper Gaubert, Kline, and Yagan (2020) reconsider this view and suggest that using location as a targeting tool can be a successful redistributive tool under certain conditions. The empirical evidence is mixed. J. Chen, Glaeser, and Wessel (2019) study if the housing market in the USA responds to the Opportunity Zone program. They rule out price impacts bigger than 1.3 %. Lutz (2015) uses a school finance reform in New Hampshire and finds that lower property taxes generate new residential construction. He finds a higher response in areas with a higher elasticity of housing supply. Places with low elasticity of housing supply experienced an increase in prices. The property taxes are capitalized into housing prices.

codes called *estratos*. Houses in estrato 1,2 and 3 receive subsidies, those in estrato 4 pay the market price and those in estratos 5 and 6 pay a tax. Neighborhood score cutoffs are arbitrarily defined, making Bogotá an ideal setting to study subsidy effects. Bogotá's scheme allows me to use quasi-experimental variation that is rare in this type of policy.

In particular, I use the discontinuities introduced by the cutoffs defining the subsidy level, to study the effect of the subsidy scheme on the housing market. I focus on houses receiving a high subsidy (*estrato* 2) and a low subsidy (*estrato* 3). These two codes concentrate 57 percent of all housing units, which allows me get enough blocks to implement my Research Design. Additionally, the models for valuation of housing prices, are common across these two estratos,[44] and both *estratos* receive subsidies. By focusing on these two estratos, I minimize the potential effect of a behavioral response to avoid a stigma associated with receiving the subsidy and being labeled as poor.

I collect different sources of administrative data to replicate the neighborhood block quality score and to measure characteristics of the housing market. The neighborhood block quality score and *estratos* come from the Stratification Census. Using the replicated score, I correctly classify 99.9 percent of the city blocks in 1997, when the assignment method was first introduced. I use the Cadaster Census to construct measure of property size, quality and construction timing. The quality measure is a score that summarizes the interior characteristics of the property (i.e bathroom size, type of floor and materials of the kitchen). For construction timing, I observe the date when the structure was built in each lot. Having the construction date allows me to see the effect of the subsidy scheme on new construction. For land and structure price, I observe market price appraisals for each unit. These appraisals are the base for the

[44]The officials use a different model for codes 1,2,3 and 4,5,6 and a different model for single-family units and multi-family units

assessed values used for property tax purpose.

I find that the probability of new construction is 43 percent higher in the heavily subsidized areas,that pay around 30 percent less in utilities and property taxes. Consistently, the average age of the units is lower. Additionally, there is some evidence of a higher quality housing and no clear differences in property size. The heavily subsidized areas are more expensive. Most of the increase in value is explained by the age of the property, the quality and the increase in land prices. I can not reject the null hypothesis that difference in prices is equal to a benchmark calculation of full capitalization.

This paper contributes to the literature on capitalization of subsidies and taxes into house prices, and more generally the literature studying the effect of policies on the housing market.Hilber (2017) highlights that empirical evidence for capitalization subsidies and taxes into housing prices is limited. On the other hand, the body of literature studying effect of different policies and regulations on the housing market is large but concentrated in the USA.[45] As mentioned by Gyourko and Molloy (2015), while some papers investigate land regulations in England, Spain and Shanghai, the research for the developing world is scarce. This paper contributes by studying a different type of policy in a developing country using a novel Research Design.

I also contribute to the literature studying the effectiveness of location-based policies. There is little evidence of the efficacy of location-based subsidies, particularly outside of the USA. Kline and Moretti (2014) present a framework to study location-based

[45]For example,Bayer et al. (2007), Black (1999) show that houses in districts with better schools are more expensive. Chay and Greenstone (2005) find that less polluted counties have higher property prices. J. Chen et al. (2019) study how Opportunity Zone Program on housing prices. There is evidence of housing market responses to policies into particular locations. Land prices in regulated areas are higher. Turner, Haughwout, and van der Klaauw (2014) find that land regulation has an impact on land prices in the USA. Lutz (2015) uses a reform in school finances in New Hampshire and finds that lower taxes is associated with new construction. It also increases in property values in places where the elasticity of housing supply is lower. Haan and Simmler (2018) study the incidence of subsidies for wind energy on agricultural land prices in Germany.

policies and suggest that the effects of these type of policies are not well-studied and that it is important to identify who is benefiting from them. Normally the literature on location-based subsidies focuses on firms and efficiency losses. However there could be redistributive reasons to support location-based policies. Gaubert et al. (2020) propose a framework to study the equity-efficiency trade-offs when targeting policies to location and not individuals based on their income. They present some conditions under which location based subsidies could be effective redistribution tools.[46] In this paper, I provide empirical evidence on the effect of a subsidy on the housing market. I argue that the housing market effects should be included when evaluating the potential welfare gains of using location as a targeting tool.

Finally, I contribute to the literature studying the *estratos* as a targeting tool for subsidies in Colombia.The public utilities subsidies assigned using the estratos are an important social assistance program. They represent 0.27% of Colombia GDP. Despite the big proportion of national expenditure required to fund this subsidy, and the potential forgone revenue in property taxes,little is known about its cost effectiveness.[47] There are two conflicting papers that address a similar question. Medina and Morales (2007) use a boundary discontinuity design and find evidence of capitalization. Gallego, Montoya, and Sepúlveda (2016) use the cutoff points from the score for Bogotá and find higher prices in not subsidized areas. These conflicting results suggest that the discussion remains unresolved. In contrast to the analysis by Gallego et al. (2016), I recover the assignment score for all the blocks in the city. I contribute to this discussion, by refining the existing analysis and providing evidence of the effects on new construction and other housing outcomes. As suggested in other settings by

[46] In particular, location-based subsidies could be effective if less skilled households are concentrated in Distressed areas, if few households are indifferent between targeted and not targeted locations, if productivity differences across areas are small, or if the marginal utility of consumption declines slowly with income.

[47] The other large social assistance program, the Colombian Conditional Cash Transfer, *Familias en Acción* , is 0.35 percent of the GDP. Excluding the utility subsidies, the total expenditure on social assistance programs was 0.49 of the GDP in 2014. (Harker, Lustig, Martínez, & Melendez, 2016).

McRae (2015), Casas (2014) and Meléndez (2004), I also find that the subsidies are generating distortions in the housing market. These distortions could offset the primary redistributive goal.

The rest of the paper proceeds as follows. I first introduce the targeting tool and the subsidies in the institutional background section. Then I explain the assignment rule of the estratos and how I use it in my Identification Strategy. Then, I explain the data, show the results and present some robustness exercises. Finally, I give some concluding remarks.

# 2 Institutional background

## 2.1 Targeting Tool- the *estratos*, and the Subsidies

In 1994, law 142, mandated all municipalities to use a standard method to classify the houses into six different subsidy codes- the *estratos*. Each residential block in a Colombian city has an *estrato*. The *estratos* comprise a coding system designed to reflect the quality and urban characteristics of each block. In this paper, I focus on Bogotá, which implemented the new stratification method in 1997.[48] Figure 2.1a shows the distribution of the codes in Bogotá.

The purpose of the estratos is to create a targeting tool to materialize the criteria of solidarity and income redistribution contemplated in the tariff regime for public utility services.[49] The providers of public utilities such as electricity, water, sewage, etc., and property taxes authorities use the *estratos* to charge different tariffs. The public services have a cross-subsidy price scheme. Depending on the estrato of the house, people pay a subsidized price, pay the *market price*, or a tax. Production costs

[48]The city subsidizes public services since 1983,but the targeting tool to subsidize utilities was different. (Decreto Distrital 1140, 1983.)

[49]See law 142 of 1994 . This law regulates the provision of public services in the country.

determine the regulated *market price*.[50] People living in properties located in estrato 5 and 6 pay more than the *market price*, those in estrato 4 pay the market price, and those in estratos 1,2 or 3 pay less. Figure 2.1b shows the electricity price scheme. The y-axis represents the price by *kwh* relative to the market price. The x-axis shows the different estratos. In 2017, houses in *estrato* 4 paid a market price of 40.84$ $COP$ per *kwh*. In the figure, this is normalized to be 1. Households living in estrato 1 receive a 66 percent price discount, paying 44 percent of the *market price*. Those in estrato 2 receive a 45 percent discount, and those in estrato 3 receive a 15 percent discount. The higher-numbered estratos and public funds subsidize this discount. The households in estratos 5 and 6 pay 20 percent more than the *market price*.

Other policies rely on this targeting tool to charge differential rates. For example, in Bogotá, the property tax varies on two dimensions, the value of the property, and the estrato of the property. Figure 2.1c shows the property tax scheme for three different property values. I selected the prices to show the important breaks in house values, and to make the figure comparable to figure 2.1b, I show the property tax relative to that of properties with the same value but in *estrato* 4. In general, the differential tax rate is similar to the electricity price scheme. The property tax rate for houses in estratos 1 and 2 is around 62% of the rate for houses of the same value in estrato 4. If the house costs less than 43 million Colombian Pesos (COP), they do not pay taxes. Depending on their price, houses in estrato 3 pay between 81 percent and 84 percent of what they would pay if they were in estrato 4. Houses in estratos 5 and 6 pay between 10 and 30 percent more than what they would pay in estrato 4.[51]

[50]The market price is a regulated price that considers the production cost and a reasonable margin of benefits for the provider.

[51]In 2016, there was a reform to the tax code. Starting in January 2017, the property tax only depends on the value of the house. Therefore, the estratos do not longer serve as a way to target differential rates for the property tax. Other subsidies, like differential rates for public higher education institutions use the estratos as a targeting tool. Additionally, the estratos are important in terms of signaling. It is a very salient characteristic and plays an important role with social status.

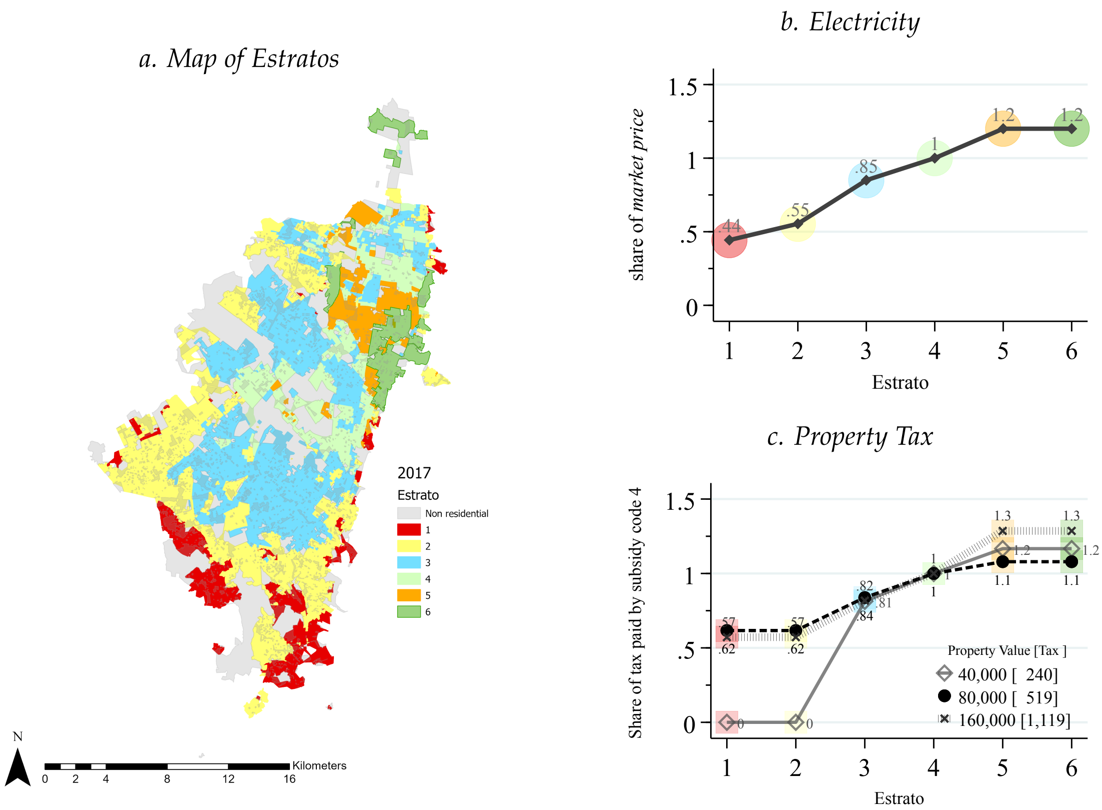



*Figure 2.1:* The Estratos and the Subsidies in Bogotá

NOTE: Figure 2.1a shows the map Map of Bogota in 2017. Each color represents one of the 6 different subsidy codes-*estratos*. The figure 2.1b shows the electricity price scheme. The y-axis represents the price by kwh relative to the market price. The x-axis shows the different estratos. The market price is regulated and set as a function of the production costs. In 2017 the price was 440.84 $COP/kwh$. All of the households start paying the market price for each $kwh$ consumed above the basic consumption of 130 $kwh$. The price discount only applies for consumption below 130$kwh$. For every extra $kwh$ consumed above 130$kwh$, subsidized households pay the market price. The price scheme for the other utility services is very similar. The price scheme is similar for the other utilities (gas, water, sewage, cleaning services, phone, and internet). The figure 2.1c shows the property tax rate relative to the rate a house of the same value would pay if located in subsidy code 4. The x-axis represents the estratos. The y-axis represents the share of tax a house in estrato 4 would pay. Each line in the figure represents the tax scheme for a given property value. The property value and the tax rate for *estrato* 4 are at the bottom right of the figure in brackets. *Sources:* Enel-Codensa Energy Rates and Article 2 of the agreement 105 of 2003

## 2.2 How big is the subsidy?

To compare the subsidy scheme with housing prices, I convert the monthly subsidies on utilities and property taxes into a stock variable.[52] I calculate the average subsidy in each estrato using the aggregate records from the public services regulator (SUI, 2020). I use the annual subsidy on electricity, water, gas, cleaning services, and the sewage system in 2011. For the property tax, I calculate the mean property tax rate, using the Cadaster appraisals and applying the tax formula. Figure 2.2a shows the Net Present Value (NPV) of the subsidies and property taxes. As a reference point, I also show the average housing price. The figure shows that the subsidies are important relative to the housing prices. Figure 2.2b shows the difference in the property tax and subsidies between estratos. This difference is a benchmark for a complete capitalization of the subsidy.[53] This *naive* comparison shows the opposite of a capitalization effect. For example, the difference between estratos 2 and 3 is 18.8 million COP, around 35 percent of the average house price in *estrato 2*. Despite the fact that they pay more on utilities, and property taxes, the average price of a house in estrato 3 is also higher. However, average houses and blocks in each estrato are not necessarily comparable; see appendix figure 2.18 as an example. The purpose of this paper is precisely to find a valid counterfactual to make a valid comparison of houses and blocks that are comparable but are classified in a different estrato. When I do that, the pattern shown in this figure is reversed, and I cannot reject a full capitalization of the subsidy. Houses in comparable houses in estrato 2 are more expensive than those in estrato 3.

[52] To do that, I use the net present value of an annual over a period of 30 years. I use as a discount factor, $\beta$, the interest rates for a 17 year national bond, $\beta = 0.0365$. $NPV(x) = \sum_{t=1}^{30} \frac{x}{(1+\beta)^t}$

[53] An alternative way to analyse the size of the subsidy is to take the monthly savings or taxes payed because of the subsidy scheme. The households living in estrato 2 received on average 87 thousand COP per month, and households in estrato 3 around 25 thousand COP per month. Households in the best neighbourhoods pay a tax of around one million COP per year. Appendix Figure 2.13a shows the average subsidy or contribution for the average household in each subsidy code in Bogotá in 2011. Appendix Figure 2.13b shows the subsidy relative to the average income in each subsidy code. The subsidy represents around 8 percent of the annual income of households living in code 1 and less than 2 percent of the income of households in code 3.

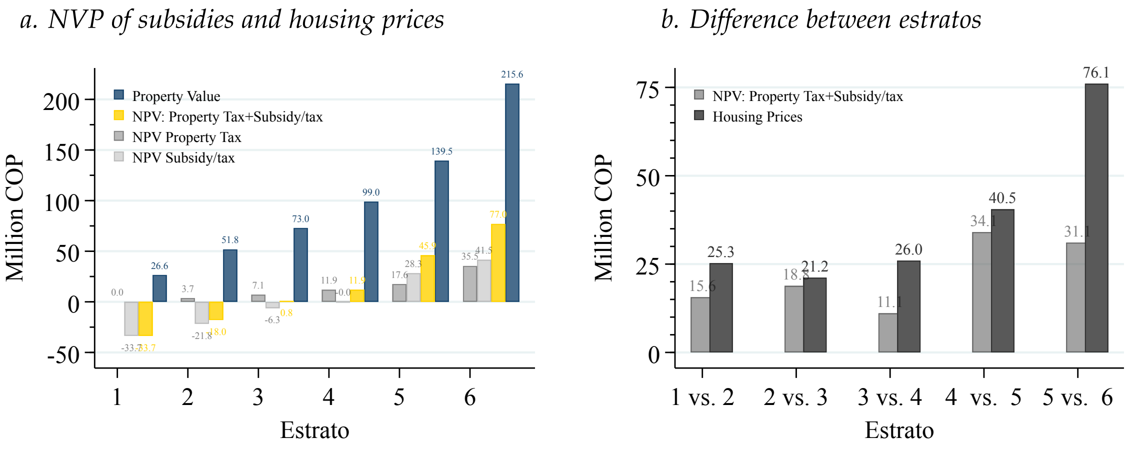



*Figure 2.2:* Subsidies and Housing Prices

Note: The value of utilities to calculate the NPV is the average yearly payments for a household in each *estrato*. I include expenditures in electricity, water, gas, cleaning services, and the sewage system in 2011. The data comes from SUI (2020). To calculate the property taxes, I take the average house appraisal in each *estrato* and I apply the tax formula. The property value is the average appraisals for each *estrato*.

# 3 Data

To study the subsidy scheme's effect on the housing market, I need to recover the estratos' assignment formula and observe housing characteristics, including housing prices. For this, I rely on administrative data. The two main data sources are the stratification census and the Cadaster census for Bogotá. Additionally, I use population census and some historical and recent shapefiles to create a constant geographic unit over time. In this section, I describe the different data sources.

## 3.1 Housing Characteristics

The *district cadaster* has a census of all the properties in the city. The authorities use the census to estimate the housing prices for each property.[54] With the census information, I construct variables to study the effect on new construction, house characteristics, and

[54]The 2017 and 2018 censuses are publicly available for a subset of variables. I only observe the appraisal and construction date for each unit for 2011. That is why my analysis uses 2011. I accessed the restricted data trough the CEDE data center, at the Universidad de Los Andes.

housing prices.

To analyze the subsidy's impact on the construction of new units, I test if the subsidized areas have a higher concentration of new properties. I use the construction date, and create an indicator variable for housing units built after the introduction of the stratification system. I also evaluate the effect on the average age of the unit's structure in each block.

Individuals could make quality or size upgrades in their units if the elasticity of substitution between consumption in public services and housing quality is negative. To study the effect on construction size, I use the construction's reported size in square meters. For the quality, I use the physical characteristics collected in the cadaster census. The data has measures for the kitchen's quality, the floors, and other detailed features of the houses. My quality measure is the *puntaje catastral*, a score summarizing the overall quality of the interior of the house.[55]

## 3.2 Housing Prices

The other outcome of interest is housing and land prices. Unlike the US and the UK, Bogota does not have a systematic data collection system of transactions to use as my main price data. However, I can observe the appraisals and assessments of the houses for all the properties in Bogota. This measure has some advantages and some disadvantages. The main advantage is that I have a price measure for all the properties using the same criteria. Ideally, I would like to observe transaction prices for all the existing units in the city. Transactions prices will give me the market valuation of the property. However, the decision of selling a house could be affected by the subsidy, and therefore a sample of transactions could be bias. Also, only a subset of units is

[55] Note that this is a different index that the block quality score use in the assignment of the estratos. First, it is constructed for a unit and not a block level. Second, it measures the interior quality in contrast with the block quality score's exterior characteristics. For a complete description of the score, see IGAC instructions document.

sold simultaneously, which creates power issues for identification in my setting.

I have overcome this challenge by having appraisal for all the properties. The district authorities try to get accurate market value for each housing unit for tax purposes. The city is widely recognized for its quality on the appraisals (Tsivanidis, 2018). In their effort to get market prices, they appraise a subset of properties and combine it with transaction data, the cadaster census, and neighborhood amenities. They calculate a property price for all the city's properties. This creates the problem that for many of the units I observe, the property price is model-based. This could generate concerns that I may be capturing some mechanical effects of the valuation formula or the appraiser. Additionally, it is usually the case that the valuations and assessed prices are below the market price.[56] With those caveats in mind, I use these appraisals as my housing prices measures.

The appraisals have several steps. First, the district acquires market prices for a subset of properties. These prices are transaction prices from property sales and individual appraisals by district appraisers.[57] Second, a statistical model relates these market values to a set of observable characteristics in the cadaster census and some amenities. The model includes variables like quality, age, and the type of construction. It also considers land use, location, general nearby amenities, like available parks, and access to public transportation, and estimates of land prices. The model includes 48 different attributes (see models in *Registro, Distrital Acuerdo Número 657* (2016)). Multi-family units under the horizontal property regime and the single-family units owning the lot are estimated in separate models. Estratos 1,2 and 3 and estratos 4,5 and 6 have two independent models. The combination of the two types of properties and the two

[56] Details for the calculation of the estratos are in *Registro, Distrital Acuerdo Número 657* (2016). This document includes an example of the statistical models used in the appraisals, p.3 Annex 2. For Bogota's case, the assessed value represents, on average, 80 percent of the commercial price (González, 2014). By law all the parcels are surveyed on around 60 variables.

[57] The process include officials to pose as potential buyers in order to negotiate a sales price under the premise of a cash payment and professional assessments for at least one property in the more than 16,000 *homogenous zones* (Lozano-Gracia and Anselin (2012), Ruiz and Vallejo (2015),Tsivanidis (2018))

groups leads to 4 different models. The main comparison in the paper uses blocks with subsidy codes 2 and 3. I separate multi-family units from single-family units to avoid mechanical effects of the statistical model. Once the district cadaster has the appraisals, they estimate the *avalúo catastral*, the assessed value for tax purposes.

Using the statistical model and the census information, the houses without an appraisal get an estimated market price for the property $\hat{p}_{m^2}^{catastro}$. This estimate can be used to calculate the structure value by square meter as the residual of the value of the property minus the lot value.[58]

The land price per square meter $\hat{p}_{m^2}^{land}$ is a block-level variable. To calculate the land price, the officials consider sales, offers, transactions, or leases of real estate, in addition to appraisals made by the district cadaster. It is an input in the appraisal models. Using information from the housing market, the cadaster office assigns a value per sq meter of land to each city block.

These appraisals are the base for the assessed value for property tax purposes. The appraisal has a land price component and a structure price component. For properties in subsidy code 1,2, 3, the assessed value, the *avalúo catastral*, corresponds to 70 percent of the land appraisal and 60 percent of the structure. The houses in subsidy code 4 will have an assessment equal to 85 percent of the land appraisal and 65 percent of the structure appraisal. The houses' assessment in subsidy codes 5 and 6 is 85 percent of the land value and 75 percent of the structure.[59] I use the appraisals, and not the assessed value used for property taxes, to estimate the effect of the subsidy scheme on the housing market.

[58] The price is defined as, $P_{m^2}^{structure} = \frac{size_{m^2}^{property} \times \hat{p}_{m^2}^{catastro} - size_{m^2}^{lot} \times \hat{p}_{m^2}^{land}}{size_{m^2}^{property}}$ if Single Family Units and $\frac{size_{m^2}^{unit} \times \hat{p}_{m^2}^{catastro} - size_{m^2}^{lot} \times \hat{p}_{m^2}^{land} \times \frac{size_{m^2}^{unit}}{size_{m^2}^{building}}}{size_{m^2}^{unit}}$ if Multi Family Units

[59] The full description of the assessment formula can be found the annex 2 of the Acuerdo Número 657

## 3.3 Stratificaton Census

To implement the Research Design I need the assignment score. Because the exact score and cutoffs are only available to the department that process the raw data, I use the raw data, and the score's formula, to replicate the assignment score and recover the cutoff points. The data is collected by the District Department of Planning in the Stratification Census.[60] I use the 2009 Stratification Census. This census defines the estratos in 2011 the year I observe the main outcomes.

The composition of estratos in the city has been relatively stable. Of the 36'985 existing in 1997 only 4.2 percent had a change in *estrato* between 1997 and 2009. Only 80 blocks had a change between 2009 and 2017. Appendix table 2.7 shows the changes in the estratos in these years.

To ensure the accuracy of my replications, I compare the prediction using my formula and the observed estratos for all of the available years of raw data. For 1997, I correctly predict 99.9 percent of all the blocks. For 2009, I correctly classify 99.4 percent of the blocks. Appendix Table 2.5 shows the percentage of correctly predicted blocks by estrato for all the years with available information. The drop in accuracy for later years is consistent with internal revisions of the assignments. In fact, a technical committee was created to deal with appeals in the estrato assignments. Even with this committee, however, very few blocks are revised and assigned to a different estrato.[61]

[60] I have complete access to different updates of the stratification census.The district planning office collects and updates the information every 2 to 5 years. The city had seven updates in the last 20 years.The updates were in 1999, 2002 2004 2007, 2009, 2013 2017 and 2019.The form used to collect the information is in the appendix figure 2.17.

[61] For example,Departamento Administrativo de Planeación Distrital (2004) document the changes from 1997 until 2004 (p35-p41). Up to 2004 680 blocks got a reassignment of the estrato. Note that the number of estratos that I cannot correctly predict in 2009,245, is below this number. The figure 2.20 in the appendix shows the number of requests for reassignment and the number of accepted reassignments. The figure clearly shows how only a low percentages of requests get approved.

## 3.4 Score Replication

The model to assign the estratos uses two main inputs; a code that summarizes the urban context of each block – the *habitat zones*, and a composite index score that summarizes the quality of each block.

The *Habitat Zones* are urban areas with a similar set of characteristics (e.g., Residential with low density, Industrial, etc.). The type of roads, the topography, the availability of public services like schools and parks, the current use of the land and type of urban development, among others, define the zones.[62] There are 12 habitat zones, and each Colombian city is sub-divided into one of these zones. For Bogotá in particular, some zones are split in two; *zone* $x$ $(-)$ and *zone* $x$ $(+)$. Thus, the city has a total of 20 different zones, coded from 1 to 20. Each block receives a code based on its urban context characteristics. A lower code indicates a lower quality of urban context. The codes 18-20, are for institutional use or green space. In this paper, I focus the analysis in habitat zones with codes 7-10 corresponding to blocks in estratos 2 and 3. Those Habitat Zones correspond to Industrial Use, Consolidated progressive development (-) and (+), and Commercial (-). The coding of the different blocks did not change in the last 20 years.[63] Land-use regulations are not associated directly with the habitat zones; In Bogotá the only practical use is for the assignment of the estratos. Figure 3.3a shows the distribution of the zones in the city. Appendix Table 2.6 describes in detail the characteristics defining each zone.

The quality score summarizes 7 characteristics of the block. i) the type of access road,

[62] In addition to the habitat zones there are two other relevant zoning definitions in Bogotá: The Physical Homogeneous Zone and the Geo economic zones. They are the geographical spaces determined from Physical Homogeneous Zones with similar unit values in terms of their price, according to the conditions of the real estate market.

[63] For example in an study by Econometria (1999) they conclude: "[...] the zoning used to carry out the stratification procedure responded to urban and socio-economic concepts of the environment (habitat zones), evaluated with a high degree of subjectivity and not to a zoning based on cadastral information, such as it was the alternative methodology" note 7 chapter 3 in (Sepulveda, Lopez Camacho, & Gallego, 2014)

ii) the type of sidewalk, iii) the existence of a front yard, iv) the parking type, v) the front of the houses, vi) the type of roof, and vii) predominant materials of the houses in the block. The unit of observation is a residential block-side. Each block-side receives a categorical value for each of the seven variables. For example, the variable predominant materials of the houses in the block has five categories; 1-Precarious materials, 2-basic materials, 3-unpainted low-quality bricks, 4-painted low-quality bricks, and 5-polished bricks or veneer.

To create the quality score, a log-rank transformation known as a *savage score* converts the categorical variables into a continuous index. The savage score has four main steps. First, construct a single value for each variable and each block. The value is the average of the categorical values of the different block's sides (normally a block has 4 side, but there are some exceptions). For example, if a block has precarious materials (code 1) in two sides,and basic materials (code 2) on the other two sides, the block will have a value of 1.5 for the variable predominant materials of a block. Second, rank the blocks using their value on each variable. $r_i^k$ is the rank of block $i$ and $k = \{1, ..., 7\}$ is an index for each of the 7 variables. Third, use the ranks $r_i^k$ each variable $k$ receives a savage score $h_i^k$.

$$h_i^k = \left( \sum_{j=N-r_i+1}^{N} \frac{1}{j} \right) - 1.$$

Finally, the 7 $h_i^k$ savage scores are added to create a quality score, $score_i$ for each block. The score is a sum of the savage score for each block, $score_i = \sum_{k=1}^{7} h_i^k$.[64] Figure 3.3b represents the distribution of the score in Bogotá.

[64] I use the formula described in SDP (2004) p.65

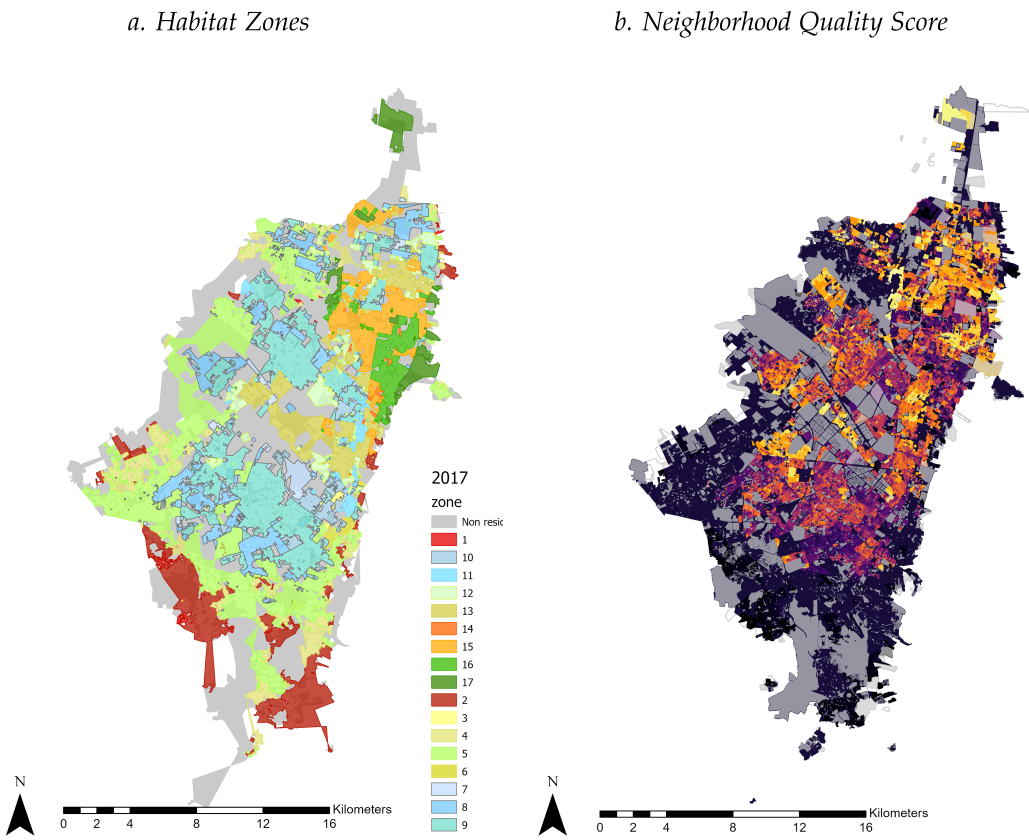



*Figure 3.3:* Distribution of Main Assignment Variables in Bogotá

NOTE: Panel A shows the distribution of the habitat zones in the city. Panel B shows the distribution of the quality score in the city. The light colors represent better quality areas and the dark colors lower quality areas.

The habitat zones and quality score are the two main inputs to define the estratos. Once the score is calculated, it is combined with the habitat zones. Neighborhood blocks are divided among the 6 estratos. The cutoffs are defined by the methodology Dalenius-Hodges, which tries to maximize the variance among groups and minimize the variance within groups.[65] The National Planning Department receives the raw

[65]Bogotá uses an adaptation called Bi-variate Dalenius-Hodges. This adaptation was implemented to treat the habitat zone independently and not as an additional value in the quality score.

data and processes the information. They construct the quality score, and assign an estrato to each block. They provide the estratos without making the composite index or the cutoffs public. The public service providers receive the information and use the codes to charge the differential price rate.

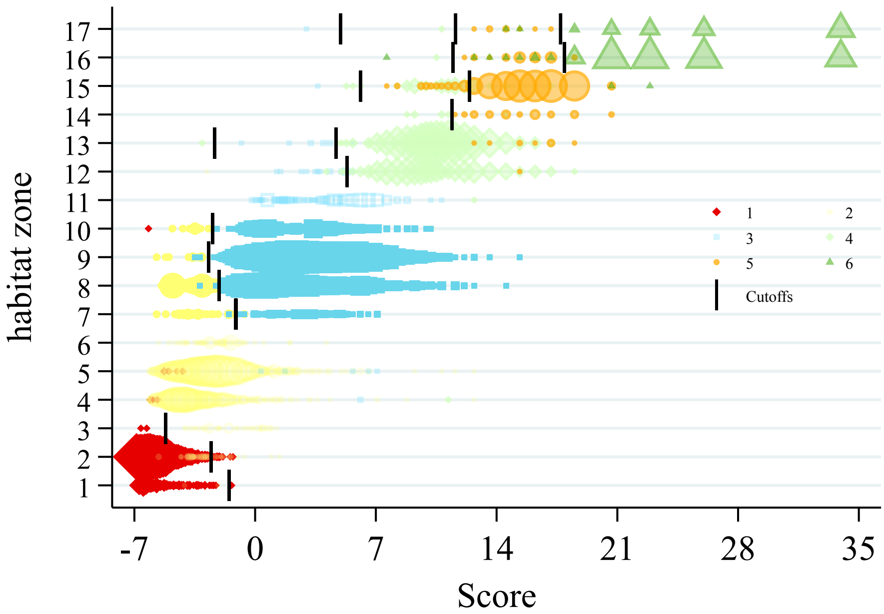



*Figure 3.4:* Subsidy Codes Based on Habitat Zones and the Score

NOTE: The figure presents the distribution of estratos by habitat zone and block quality score. A different color represents each estrato. The darker blue and yellow are the focus of the analysis.

I combine the score with the *habitat zones* to infer the location of the cutoffs. To define the cutoffs, I try to emulate the matrix employed in the bi-variate Dalenius-Hodge methodology. Within each Habitat Zones and the quality score, I define the cutoff as the value that maximizes the probability that I correctly predict the observed distribution of estratos. The figure 3.4 shows the scores (x-axis) and the *Habitat Zones* (y-axis) for 2009. The six different colors represent the six *estratos*. The black solid lines are the cutoff that I assign. The symbols are weighted to represent the number of blocks with a particular score and zone. I highlight the areas that are the main focus of the analysis (i.e. habitat zones 7-10). The blue squares represent blocks in *estrato* 3, and the yellow circles represent blocks in *estrato* 2. The figure shows that only a few habitat zones have more than one estrato. I use those zones in my identification strategy. The distribution of scores and estratos are similar in the other years with available data (see appendix figures 2.19).

## 3.5 Baseline Characteristics and Constant Geographic Unit

Some of the blocks change the geographic administrative codes over time. This makes it difficult to track a particular location at different points in time. To do that, I identify which codes correspond to the 2017 administrative codes based on the location and assign that code to all of the years based in a particular location over time. Additionally, there are two different administrative codes, one used by the district authorities and the other one used by the National Department of Statistics (DANE). To check for balance in preexisting characteristics I use data from the 1993 population Census. I am able to identify the exact block for each observation. However, the code for the block is not the same used in the stratification census. To create a crosswalk between the two blocks I use the location of based on a shapefile and assign the 2017 administrative codes based on the location. Like I did to create the constant geographic unit over time. For more details see appendix.

## 3.6 Main Sample of analysis

The main analysis of this paper, focuses on estratos 2 and 3. There are several reasons. First, these two groups represent an important share of the residential properties in the city. In 2011 Bogotá had 559,328 residential properties in estrato 2 and 654,136 in estrato 3, corresponding to 61 percent of all the residential properties in the city. Second, the subsidy for houses in estrato 2 and 3 represents an important share of household income and housing prices (see figure 2.2a and Appendix Figure 2.13b and

2.13a ). Third, the appraisals for those two codes are calculated with the same statistical model. I also avoid getting effects mainly explain by the difference in appraisal methods. This will be the case if I compare *estratos* 3 and 4.[66] Fourth, individuals may want to avoid the stigma associated with receiving the subsidy, the fact that the two groups receive a subsidy reduces the role of this mechanism.

The quasi-experimental variation I exploit in my research design is at the block level.Therefore, I create block-level variables and perform the analysis at the block level.[67] I restrict the analysis to residential properties because the subsidy scheme does not apply to non-residential properties. I restrict the analysis to residential properties because the subsidy scheme does not apply to non-residential properties.

# 4 Research Design

Studying the incidence of a location-based subsidy is challenging for at least two reasons. The targeted areas are particular, and the policy can induce sorting. Usually, the targeted locations are under-performing areas; they are poorer, have more crime, or are more polluted. Thus, a counterfactual that allows studying the incidence and effectiveness of the policy is hard to find. In contrast with other settings, randomly assigning the treatment to some areas is not easy, and is rarely implemented. Second, the policies can encourage people to move to take advantage of the subsidy. For example, people in the US decide where to live based on the school quality of each town (Bayer et al. (2007) and Black (1999)). A careful evaluation of location-based policies has to consider this sorting and the potential effect on the housing market. A comprehensive assessment of these subsidies has to account for any impact of the policies on

[66]The officials use different model for estratos 1,2,3 and 4,5,6 and for single and multifamily homes. In section 3 I explain more in detail.

[67]The results are very similar when I use individual level variables but to be more transparent about the source of variation I am using, and to have more conservative SE, I keep the block level specification as my main specification.

the housing market.[68]

Bogotá's particular targeting tool, the estratos, allows me to address these challenges using a quasi-experimental research design. As explained in section 3.4, Bogotá uses a uses a score summarizing the quality of the block and the urban context characteristics to assign the estratos. These two characteristics and arbitrary cutoffs separate the city into the six estratos. I use these cutoffs to apply a Regression Discontinuity Design and evaluate the effect of the subsidy scheme on the housing market. In this section, I explain how I use the features of the assignment in my Research Design and my estimation approach.

Figure 3.4 shows the cutoffs that define the estratos of different blocks within a habitat zone. The figure shows that there is not a single discontinuity, and a standard uni-dimensional RDD does not directly apply. There are different estratos and different cutoffs. In this paper, I am focusing on comparing the houses in estratos 2 and 3. I can reduce the discontinuities in the habitat zones 7-10 to a single discontinuity. Choi and Lee (2018b) among others show that the standard RDD will work under the assumption that the effect is the same in all the discontinuities.[69] Figure 4.5 shows the discontinuities at the cutoff for the particular areas where I focus my analysis (i.e habitat zones 7-10 which are highlighted in figure 3.4). I normalize the score in order to have the cutoff at 0 in each habitat zone. This figure shows a clear discontinuity in the probability of being treated at the cutoff. I use this discrete jump in the subsidy level to identify the effect of the subsidy on the housing market.

[68]for a detailed discussion of the difficulties of evaluating location base policies see Kline and Moretti (2014)

[69]Choi and Lee (2018a) Reardon and Joseph P. (2012) Wong, Steiner, and Cook (2013) Choi and Lee (2018b) Papay, Willett, and Murnane (2011)

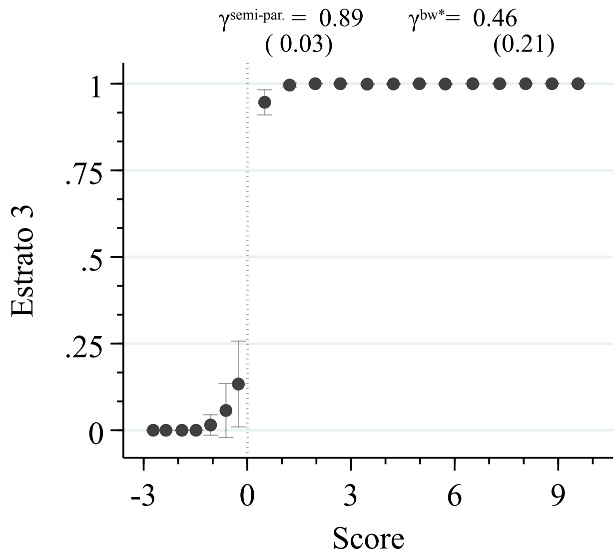



*Figure 4.5:* First Stage (habitat zones 7-10)

NOTE: This Figure represents the discontinuity in probability of being in estrato 3 introduced by the arbitrary cutoffs. The dots are local average for equally spaced bins. The number of bins minimizes the integrated mean square error (IMSE) on each side. The point estimates from 3 different estimation methods are at the top. $\gamma^{parm}$ and $\gamma^{semipar}$ use different approaches to estimate $h\left(S_{izt}\right)$ in the model: $\mathbb{1}_{[S_i=3]} = \alpha + \gamma_{2,3}\mathbb{1}_{\left[S_{izt}\geq\delta_2^3\right]} + h\left(S_{izt}\right) + \varepsilon_{izt}$.$\gamma^{semipar}$ estimates $h(S_{izt})$ non parametrically using a partially linear model (Robinson, 1988), and $\gamma^{parm}$ uses a parametric approximation using a polynomial of degree 1. $\gamma^{bw^*} = \lim_{s_{izt}\to\delta^{3}_{2}{}^{+}} \mathrm{E}\left(y_{izt}|S_{izt}=s_{izt}\right) - \lim_{s_{izt}\to\delta^{3,2-}} \mathrm{E}\left(y_{izt}|S_{izt}=s_{izt}\right)$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 7,8,9 and 10. I drop the observations with a score lower than -3 or higher than 10. I normalize the score to have the cutoff at 0 for all the habitat zones. .

The arbitrary cutoffs and the associated discontinuities in the subsidy levels allow me to identify the causal effect of the subsidy. I analyze how different outcomes $Y_i$, related to the housing market are affected by the subsidy scheme. The unit of study is a block $i$. Each block has two potential outcomes $Y_i\left(\mathbb{1}_{[S_i=2]}\right)$ and $Y_i\left(\mathbb{1}_{[S_i=3]}\right)$. I am interested in studying the difference between the potential outcomes under treatment, belonging to estrato 3, $\mathbb{1}_{[S_i=3]}$, and control belonging to estrato 2, $\mathbb{1}_{[S_i=2]}$. The parameter of interest, $\theta_i^{2,3}$ is therefore the difference between these two potential outcomes.

$$\theta_i^{2,3} = Y_i\left(\mathbb{1}_{[S=3]}\right) - Y_i\left(\mathbb{1}_{[S=2]}\right) \tag{33}$$

To estimate the causal effect of the subsidy scheme, I use the discontinuous shift in the subsidy level at an arbitrary cutoff in the block quality score. The idea behind this approach is that the only discrete change at the cutoff is the estrato level. Comparing blocks to the left and the right of the cutoff will allow me to identify $\theta_{(2,3)}$ under the assumption that the blocks on the left of the cutoff are a good counterfactual to what would had happen to the blocks on the right if they received a higher subsidy. The effect of the subsidy on the housing market will be identified by the discontinuous mean shift at the cutoff, if two assumptions are satisfied. First, when the estratos were assigned 1997, all the observable characteristics not involved in the assignement of the estratos and the unobservable characteristics should be continuous at cutoff. Second, the distribution of blocks around the cutoff should be the same at both sides of the cutoff, i.e., there is not manipulation around the cutoff. I provide evidence that these two assumptions are credible. First, in Figure 4.6, I present observable characteristics before the estratos were assigned. Second, In Figure 4.7, I show the distribution of the score for 2009 and a McCrary Manipulation test.

Figure 4.6 A and figure 4.6 B shows characteristics of the households (number of kids, number of people in the house, type of employment and education level) and housing characteristics (House with brick walls,Building,House with cement floor, number of rooms) by block quality score in 1993. I use this information to compare the blocks before the assignment of the policy. The baseline characteristics of the blocks seem to be balanced. With the exception of the number of rooms, there is not evidence of a discontinuous jump around the cutoff. To summarize the information, figure 4.6 C) shows the prediction on the main outcomes variables (i.e. age, and probability of building built after 1997, and prices) using the individual and housing characteristics in 1993. There is a small apparent discontinuity that is minor compared to the mean shift I will show in the next section. The small discontinuity in prices goes is the opposite of the discontinuity I observe in the prices in 2011. In addition to this exercise, In my empirical analysis I include this baseline characteristics and the results are unchanged. The fact that the variables are balanced in the period before the assignment of the estratos, gives me confidence on my research design.

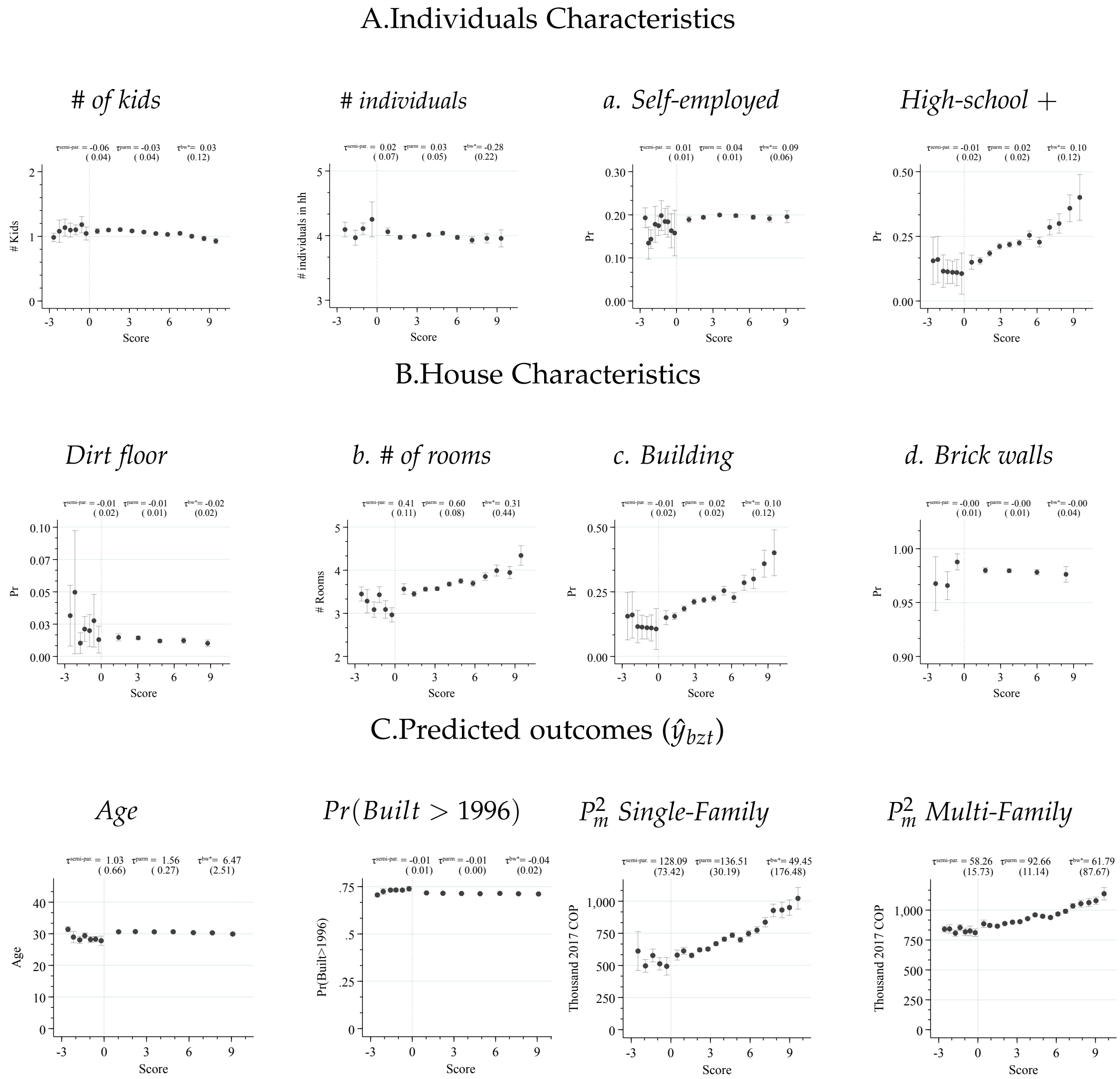



*Figure 4.6:* 1993 Balance

NOTE: The independent variables, $\hat{y}_{izt}$ is the prediction of an outcome $y_{izt}$ using 1993 demographic and housing characteristics (i.e. $y_{izt} = x_{izt}^{93}\beta + \varepsilon_{izt}$ and $\hat{y}_{izt} = x_{izt}^{93}\prime\hat{\beta}$ . The 1993 demographic characteristics are the number of individuals, the number of kids in a house, the educational level, and the type of education. The house characteristics are the type of floor, walls, the number of rooms, and the type of house (i.e., buildings, houses). The running variables uses the 1997 score.

Figure 4.7 shows the distribution of the block quality score (panel A) and the McCrary Manipulation test (panel B). The distribution blocks by the blocks quality does not have any apparent bunching at the cutoff. Given that the unit of observation is the block and the exact cutoff follows an objective methodology to classify the blocks, this is unsurprising. However, the cutoff is located at a point right before the slope's increase in the bell shape. This is consistent with the fact that the cutoff minimize the variance within groups and maximize the variance among groups (Delano-Hodges methodology). Because this results from an algorithm following and objective method, this should not be a major concern. The McCrary Manipulation test confirms that there is not manipulation around the cutoff. I cannot reject the null hypothesis that the distributions on both sides of the cutoff are the same (p-value is 0.55).

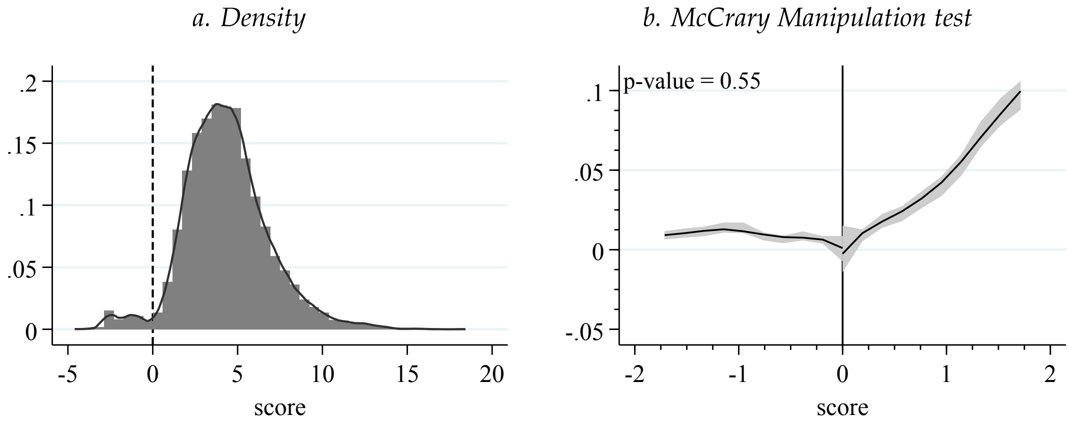



*Figure 4.7:* Score Distribution

Note: This Figure shows the distribution of the score for 2009. (left panel) and the McCrary Manipulation test (right panel). $H_0$: The density is the same on the right and on the left. The p-value of the test is on the top of the right panel. I cannot reject the null hypothesis. To implement the manipulation test I follow M. Cattaneo, Jansson, and Ma (2018).

## 4.1 Estimation

To estimate the parameter of interest I estimate the following model.

$$Y_{izt} = \alpha + \theta_{2,3}\mathbb{1}_{[S_izt=3]} + k\left(S_{izt}\right) + \Gamma' X_{izt} + \varepsilon_{izt} \tag{34}$$

Where $Y_{izt}$ is an outcome of interest for block $i$ in habitat zone $z$ in year $t$. $S_{izt}$ is the score assigned to each block, $\mathbb{1}_{[S_{it}=3]}$ is an indicator variable equal to one if the block belongs to estrato 3 and 0 otherwise, $k\left(S_{izt}\right)$ is a control function that captures the relationship between the block quality score used in the assignment and the outcome of interest $Y_{izt}$. $X_{izt}$ is a vector of observable characteristics, and $\varepsilon_{bzt}$ are the unobserved characteristics affecting $Y_{izt}$. The parameter of interest is $\theta_{2,3}$. The main challenge to estimate the effect of the subsidy scheme in the housing market is to recover the control function $k\left(S_{izt}\right)$. I estimate the control function $k\left(S_{izt}\right)$ non para-metrically using a partially linear model as in Robinson (1988). In this approach I use all the available information and I assuming that $k(S_{izt})$ is a smooth functions that do not change at the cutoff. I prefer this approach to an approach like M. Cattaneo, Titiunik, and Vazquez-Bare (2019) where I just use data only around the cutoff. M. Cattaneo et al. (2019) choose an optimal bandwith to estimate the effect using only observations around the cutoff. This approach, may be more accurate in other settings with more observations around the cutoff. In my case I do not have many observations and this lead to the estimation using the optimal bandwidth to include in some cases only 12 blocks. However, for completeness, I present estimates using both estimation approaches. As a robustness check I also estimate the control functions parametrically using a polynomial on $S_{izt}$.

The discontinuity, I am exploiting is not a sharp discontinuity, as Figure 4.5 shows. Thus, I apply the standard fuzzy RD framework. I use the assignment rule as an instrument for the treatment. $\gamma_2^3$ is the cutoff for blocks 2 and 3 and $\mathbb{1}_{\left[S_{izt}\geq\gamma_2^3\right]}$ is an indicator variable equal to one if the score is higher than $\gamma_2^3$ and zero otherwise.

$$\textit{First Stage}\text{: } \mathbb{1}_{[S_i=3]} = \alpha + \delta_{2,3}\mathbb{1}_{\left[S_{izt}\geq\gamma_2^3\right]} + h\left(S_{izt}\right) + \epsilon_{izt}$$

$$\textit{Reduced Form}\text{:} Y_{izt} = \beta + \tau_{2,3}\mathbb{1}_{\left[S_{izt}\geq\gamma_2^3\right]} + g\left(S_{izt}\right) + \varepsilon_{izt}$$

*The parameter of interest*: $\theta_{2,3} = \frac{\tau_{2,3}}{\delta_{2,3}}$

I use a partially linear model Robinson (1988) to estimate $g\left(s_{bzt}\right)$ non para-metrically and $\mathbb{1}_{\left[s_{bzt} \geq \gamma_2^3\right]}$.[70] I complement the analysis using non parametric approach following M. Cattaneo, Idrobo, and Titiunik (2018). I use local polynomial and I calculate the optimal bandwidth for both sides of the cutoff. I adjust the weights using a triangular kernel, giving more weight to the observations close to the cutoff. To get an estimate of $\theta_{2,3}$ I get not parametric approximation to the expected value of the outcome conditional on the score $S_{izt}$ on the right and on the left of the cutoff, $g\left(S_{izt}\right) = g^1\left(S_{izt}|S_{izt} < \gamma_2^3\right) + g^2\left(S_{izt}|S_{izt} > \gamma_2^3\right)$

$$\theta_{2,3} = \frac{\lim_{s_{izt} \to \gamma_2^{3+}} \mathrm{E}\left(Y_{izt}|S_{izt} = s_{izt}\right) - \lim_{s_{izt} \to \gamma_2^{3-}} \mathrm{E}\left(Y_{izt}|S_{izt} = s_{izt}\right)}{\lim_{s_{izt} \to \gamma_2^{3+}} \mathrm{E}\left(\mathbb{1}_{\left[S_{izt} \geq \tau_2^3\right]}|S_{izt} = s_{izt}\right) - \lim_{s_{izt} \to \gamma_2^{3-}} \mathrm{E}\left(\mathbb{1}_{\left[S_{izt} \geq \tau_2^3\right]}|S_{izt} = s_{izt}\right)} \tag{35}$$

# 5 Results

Figure 5.8 shows the share of buildings in estratos 2 and 3 built before and after the new stratification system was introduced in 1997. The x axis shows the score in 2011 and the y axis shows the percentage of blocks. The data reflects the housing stock in 2011. This figure suggests that after 1997, more construction was concentrated below the cutoff dividing estratos 2 and 3. The figure shows the cumulative share of housing stock by construction year. The two dashed lines represent all the blocks in estratos 2 and 3. The solid lines represent the blocks near the cutoff. When I compare all the blocks in estrato 2 and estrato 3, there is no apparent increase in units built after 1997. When I consider only the blocks around the cutoff, the subsidized blocks seem to have a higher concentration of new construction. More than 50 percent of the properties

[70]The parametric approach use in the robustness analysis I use a polynomial of degree three to approximate $g\left(S_{izt}\right)$. $g\left(S_{izt}\right) = S_{izt} + S_{izt}^2 + S_{izt}^3$.I fix the shape of the polynomial to be the same on both sides. The estimate for the reduced form $\gamma_2^3$, is the mean shift around the cutoff.

in estrato 2 with neighborhood quality scores close to the cutoff were built after 1997. This is suggestive evidence that developers and landowners built disproportionately more units in the blocks in estrato 2 comparable in terms of quality to the blocks in estrato 3. Therefore the subsidies seemed to induce new construction.

To test this hypothesis further, I use the research design to test whether the subsidy scheme affects the structure's age and the share of units in each block built after 1997.

*a. PDF of units by score and build timing* *b. CDF by construction date*

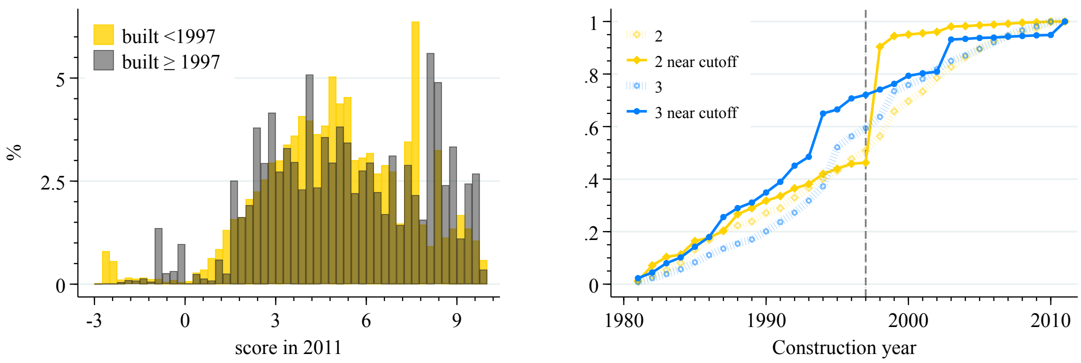



*Figure 5.8:* Stock of buildings in 2011 by construction date

## 5.1 Property Characteristics

The top graphs of Figure 5.9 show bin scatter plots for the variables related to the housing characteristics. Figure 5.9a shows the result for the age of the housing units, Figure 5.9b, the probability of having new construction, Figure 5.9c, the quality of units, and Figure 5.9d the average size of each unit. Following M. Cattaneo, Idrobo, and Titiunik (2018), I select equally spaced bins that minimize the integrated mean square error (IMSE) to the left and right of the cutoff. The bin scatters allow to visually explore any discontinuous shift at the cutoff without imposing structure on the data. The dashed line illustrate the identification approach where I impose some structure on the data. The dashed line in each figure shows the control functions $k(s_{izt})$. The idea of this identification approach is that in the absence of the subsidy scheme, the outcomes will follow $k(s_{izt})$. Therefore, the counterfactual to a world without subsidies is described by the control function. Thus the subsidy scheme's effect is the difference between the bins, and the control functions estimated with the model in equation 34.

The alternative is to estimate the effect using only the observations close to the cutoff. The idea is that the effect is the difference between the projection of the bin on the left and the right of the cutoff. The estimation uses the data around each dot to get a projection that intersects the cutoff. The point estimates for the two different approaches are at the top of the figure. These estimates are equivalent to the reduced form in a fuzzy RD design.

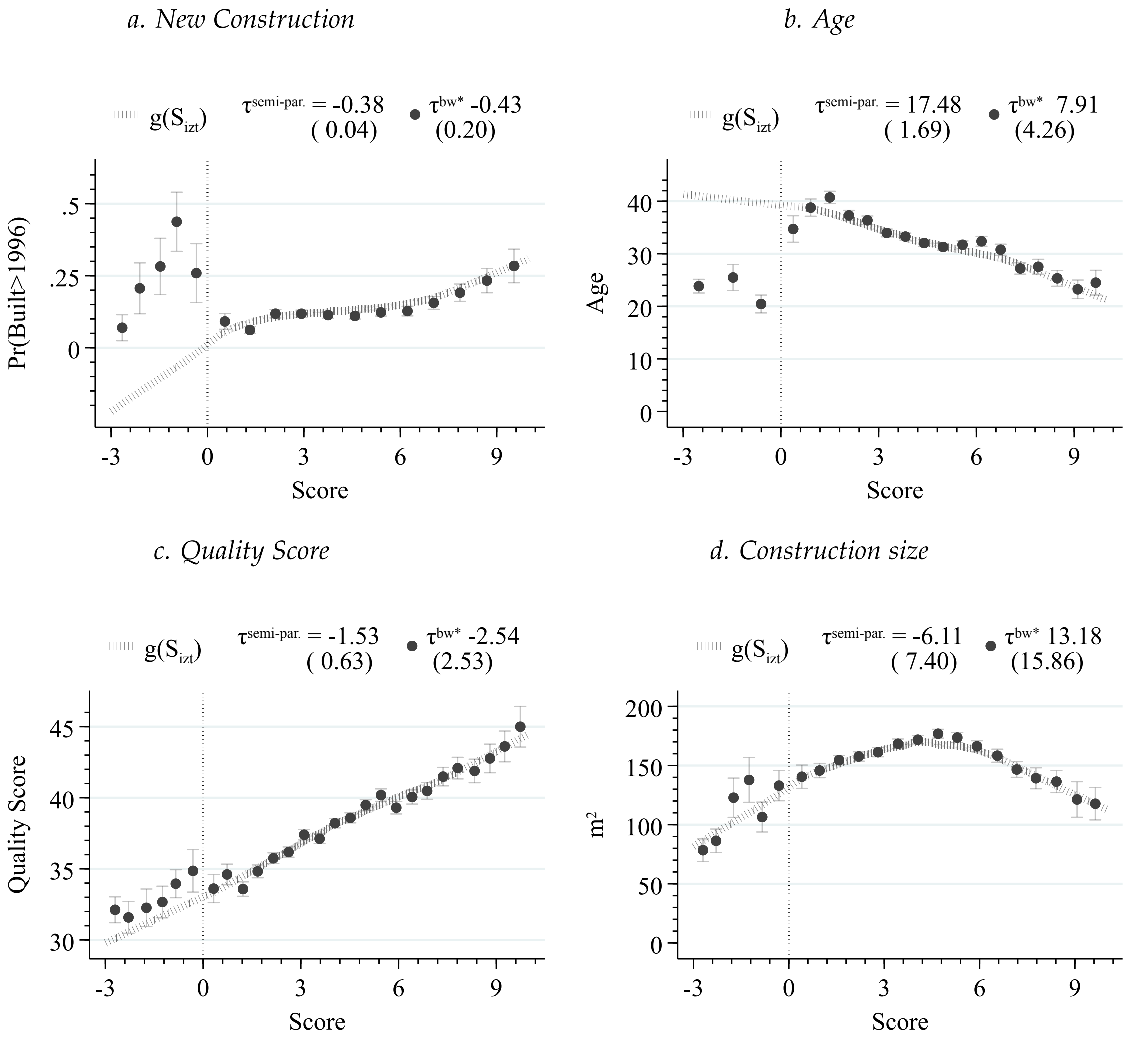



*Figure 5.9:* Property Characteristics

NOTE: This figure represents the reduced form. The dots are local average for equally spaced bins. The number of bins minimizes the integrated mean square error (IMSE) on each side. The point estimates from 3 different estimation methods are at the top. $\tau^{parm}$ and $\tau^{semipar}$ use different approaches to estimate $g\left(S_{izt}\right)$ in the model: $Y_{izt} = \alpha + \tau_{2,3}\mathbb{1}_{\left[S_{izt} \geq \delta_2^3\right]} + g\left(S_{izt}\right) + \varepsilon_{izt}$. $\tau^{semipar}$ estimates $g(S_{izt})$ non parametrically using a partially linear model (Robinson, 1988), and $\tau^{parm}$ uses a parametric approximation using a polynomial of degree 3. $\tau^{bw^*} = \lim_{s_{izt} \to \delta_2^{3+}} \mathrm{E}\left(y_{izt} | S_{izt} = s_{izt}\right) - \lim_{s_{izt} \to \delta_2^{3-}} \mathrm{E}\left(y_{izt} | S_{izt} = s_{izt}\right)$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 7,8,9 and 10. I drop the observations with a score lower than -3 or higher than 10.

Figures 5.9a and 5.9b show a clear mean shift at the cutoff, suggesting that the subsidy scheme affects how the housing market evolves in Bogotá. The blocks in the heavily subsidized areas (i.e., estrato 2) are newer and have a higher probability of being built after the assignment methodology for the estratos started in 1997. There is a smaller and less obvious jump at the cutoff for housing quality. Depending on the estimation method, the point estimate for the reduced form mean shift on average age of the units is between 7.9 and 20.8 years, and the mean shift on new construction is between 0.21 and 0.43 percent. The point estimate for the mean shift in the quality score is between $-1.5$ and $-2.5$, depending on the estimation method. For house size, there is no clear visual discontinuity.

The evidence presented to this point shows the effect of the assignment rule and not of being part of a different estrato and therefore paying different prices for utilities. To estimate the effect of the estratos in the housing market, I use the assignment rule as an instrument for being in estrato 3. The effect of the estrato is given by the reduced form estimates presented in figure 5.9 divided by the first stage presented in figure 4.5. The estimates are in table 6.3. I estimate the effects with the two different estimation methods, and I include some additional controls in columns 2, 3, and 5,6, respectively. In columns 2 and 5, I control by the habitat zone, the share of single-family units in a block, and an indicator variable equal to 1 if the block had a change in the estrato since the new stratification method started in 1997. In Columns 3 and 6, I control for the characteristics of each block in 1993. Each row presents the results for a different variable. For the partially linear model, I estimate the standard errors using bootstrap (100 repetitions). For the other non parametric approach, I use the 2sls robust standard errors.

The results are consistent with the visual description presented in the figures. Heavily subsidized blocks observe higher construction levels, which is reflected in a lower average age of its units. This conclusion does not depend on the inclusion of different controls or the estimation approach. The coefficients change very little when I include controls, and the point estimates are similar across all the different estimation approaches. Because the non-parametric estimates only consider information close to the cutoff, the estimates are less precise, and the standard errors are bigger. If the control function describing the relationship between the block quality score and the outcomes of interest is a valid counterfactual, considering information farther away from the cutoff is reasonable, it allows for more precise estimates. In the figures, the difference between the non-parametric coefficients and the other models was higher than in the table, where the point estimates are very similar across specifications.

Being in estrato 2 and getting a higher subsidy causes a decrease in the block's units' average age. The units in estrato 2 are between 15.80 and 22.08 years younger because they receive a higher subsidy. The blocks in estrato 3 have a 43 percent lower probability of having properties built after the introduction of the new stratification regime. The non-parametric estimation, the quality effect is not distinguishable from 0. For the partially linear model, the effect fluctuates between -1.72 and -2.04. The results for the size of the property suggest that there is not a difference in the size of properties caused by the subsidy scheme. When I include preexisting characteristics as additional controls, the coefficients are statistically different from 0 but economically small.

These results suggest that individuals and developers are responding to the subsidy by building new units in the areas where it is cheaper to live. There is some evidence that individuals may increase the quality of the units or build better quality units. This is consistent with individuals having a negative elasticity of substitution between consumption of utilities and housing quality; individuals that save in utilities consume better quality housing. Also, the new construction induced by the subsidy may be of better quality. There is an apparent effect of an increase in property size,which suggests that the new units are a bit smaller. These results suggest that the way the city, particularly the housing market, evolves is affected by the subsidy scheme. If the locations with higher subsidies are worse, this could be a sub-optimal equilibrium; the city growth is skewed towards its lower quality areas and worst neighborhoods.

*Table 5.1:* THE EFFECT ON CONSTRUCTION AND HOUSING CHARACTERISTICS.

| | Semipar | | | Non-parametric | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pr(Built>1996) | -0.43*** | -0.43*** | -0.43*** | -0.39+ | -0.61* | -0.54* |
| | (0.03) | (0.09) | (0.01) | (0.24) | (0.29) | (0.21) |
| Age | 19.64*** | 22.08*** | 22.08*** | 15.80*** | 18.16** | 10.76*** |
| | (0.27) | (0.71) | (1.44) | (4.73) | (6.83) | (3.11) |
| Quality Score | -1.72*** | -2.04*** | -2.04*** | -3.94+ | -4.97 | -3.94* |
| | (0.37) | (0.22) | (0.33) | (2.18) | (3.78) | (1.98) |
| Size $m^2$ | -6.87 | 10.27 | 10.27*** | 16.56 | 45.44 | 29.55* |
| | (5.97) | (8.29) | (0.96) | (18.85) | (40.35) | (14.30) |
| Controls | No | Yes | Yes | No | Yes | Yes |
| Char. 1993 | No | No | Yes | No | No | Yes |

NOTE: This table presents the estimates $\theta$ using three different estimation methods with and without controls.The estimates are the IV estimates using the assignment rule as an instrument for a block belonging to estrato 3. Each row is the estimate for a different variable. Columns 1 to 4 use different approaches to estimate $k\left(S_{izt}\right)$ in the model $y_{izt} = \alpha + \theta_{2,3}\mathbb{1}_{\left[S_{izt}\geq\delta_2^3\right]} + k\left(S_{izt}\right) + \beta X + \varepsilon_{izt}$. Columns 1 and 2 use a partially linear model Robinson (1988) to estimate $k(S_{izt})$. Columns 5 and 6 estimate $\theta$ using non parametric approach $\theta = \lim_{s_{izt}\to\gamma_2^{3+}} \lim \mathrm{E}\left(y_{izt}|S_{izt} = s_{izt}\right) - \lim_{s_{izt}\to\gamma_2^{3-}} \mathrm{E}\left(y_{izt}|S_{izt} = s_{izt}\right)$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth.The first stage and reduced form are represented in figures 4.5 and 5.9. The controls included are a dummy equal to 1 if the block changed the estrato since 1997, habitat zones, and the share of blocks that are multi family units. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 7,8,9 and 10. I drop the observations with a score lower than -3 or higher than 10.

Significance level +10%,* 5%,** 1%***0.1%

## 5.2 Property Values

Does the subsidy scheme affect housing prices? Or does it only affect the housing market through new construction? This subsection investigates the impact on prices. As explained in section III., the appraisals use different models for multi-family and single-family units. The appraisals also separately estimate land and structure prices.

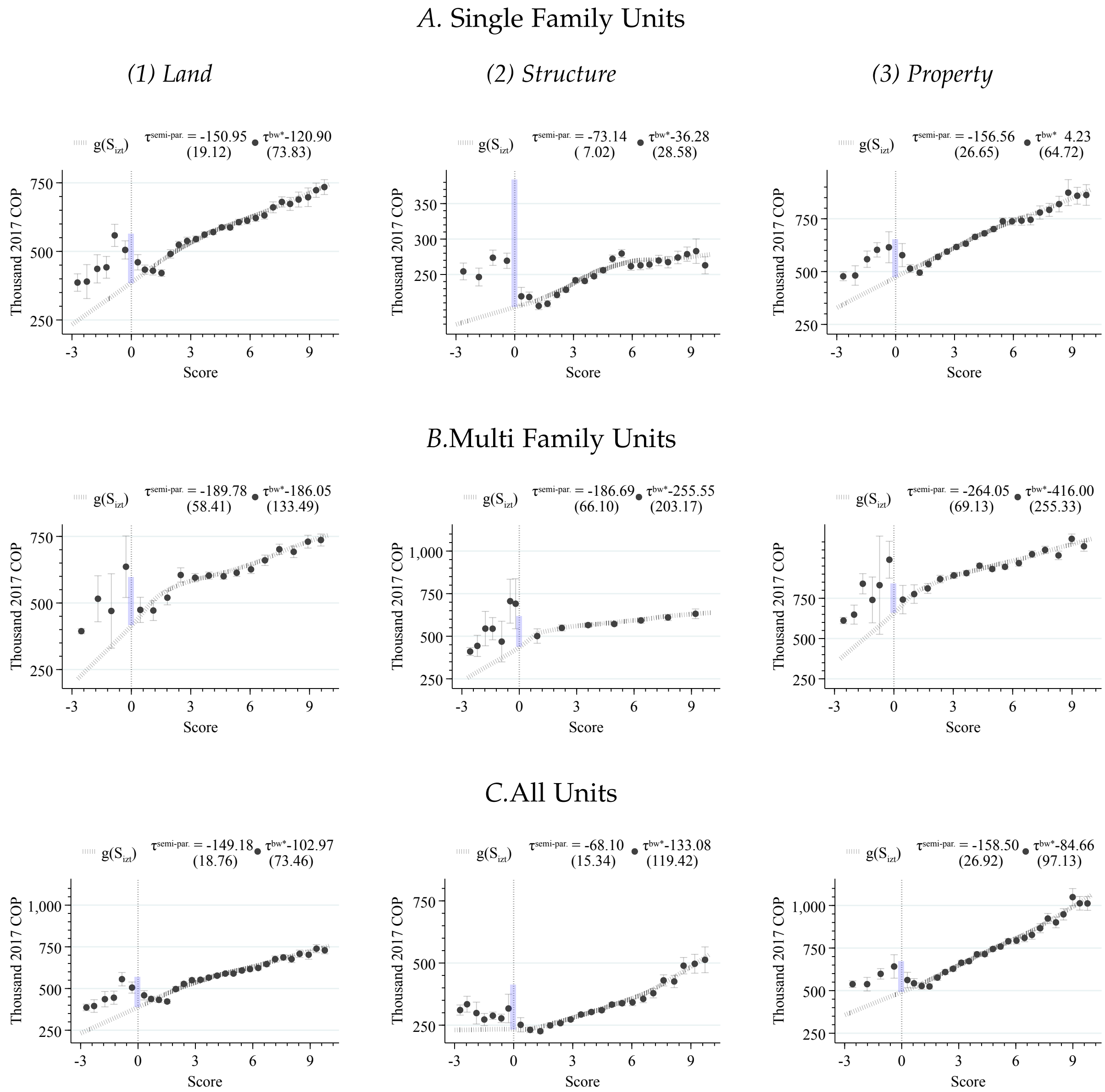



*Figure 5.10:* Housing Prices

Based on standard urban theory, the savings in utilities and property taxes should be offset by higher housing prices. All individuals have constant utility in a spatial equilibrium, and housing prices adjust to generate that equilibrium. Everything else equal, people would prefer to live in locations with higher subsidies. This higher demand to live in subsidized areas will cause housing prices to increase until everybody has the same utility, and nobody has incentives to move. This section tests this key assumption of spatial equilibrium models. My empirical setting provides a unique opportunity to perform such a test.

To set a capitalization benchmark, I take the net present value of the monthly expenditure on property taxes and utilities in each estrato. The difference between estratos 2 and 3 is 180 Thousand $COP/m^2$ per square meter and is the benchmark for a full capitalization in my empirical exercise. This is the number presented in section 2, divided by the average size of all the houses in estrato 3 and 2 (100 $m^2$).

In contrast with the case of housing characteristics, for housing prices, I only focus on blocks in habitat zones 8 and 9. To pool the four habitat zones into a single discontinuity, the assumption of constant treatment across habitat zone has to hold. For housing prices in habitat zones 7 and 10, this is not a reasonable assumption. Those zones have only few blocks with multi-family units, and housing prices are not comparable to those in habitat zones 8 and 9.[71]

Figure 5.10 shows the figures for the different measures of housing prices and the different types of units. Panels A and B show the results for multi and single-family units. Column 1 shows the land prices, column 2 structure prices, and column 3 overall property prices. All the prices are in thousands of 2017 COP per square meter. For completeness, the panel 3 presents the results for all the units. The point estimates

[71] The results on housing characteristics are similar when I use only habitat zone 8 and 9 (See Appendix table 2.4). I decided to present the results with all the habitat zones in the body of the paper to have more observations and get more precise estimates.

for the different estimation methods are at the top of each figure. As with figure 5.9, I include the control function as dashed line. In this case, the control functions represent a price gradient in the absence of the subsidy scheme.

The figures show a clear discontinuous mean shift for the multi-family units. The land prices show a mean shift between 158.6 Thousand COP/m2 and 189.7 Thousand COP/m2 depending on the estimation method. The single-family units have a less clear mean shift at the cutoff when we look at the bin scatters. Yet, the graphs imposing more structure in figure 5.10 suggest a mean shift at the cutoff for single families. In each of the graphs in figure 5.10, I add a line with the size of the benchmark capitalization. The difference between the bins and the control functions represents the effect of the subsidy under my preferred estimation approach. The figure shows an apparent effect that is close to the full capitalization benchmark. The non-parametric approach estimates are noisy, but except for total property value for single-family units (Column C panel 2), the estimates have the same sign. Their confidence intervals include the estimates from the other approaches. Figure 5.10 shows that the slope of the land gradient is steeper than the slope of the structure for both the multi and single-family units. The property structure gradient is relatively flat, suggesting that the structure's price does not vary much from the neighborhood block quality score.

Table 6.3 shows the estimates of the subsidy scheme's effect on the housing market when I instrument being in estrato 3 with the assignment rule. Columns 1 and 5 present the results for the different estimation methods without including controls. Columns 2 and 5 include basic controls, habitat zone fixed effects, the share of multi-family units, and an indicator variable equal to one if the block changed the subsidy code since implementing the stratification method 1997. Each row presents the point estimate and standard error for land, structure, and total property value per square meter. I include the p-value of a test for a full capitalization of the subsidy scheme. The null hypothesis is that the point estimate is equal to the full capitalization benchmark. i.e $\hat{\theta} = 180$. The housing characteristics analysis showed that the subsidy scheme affected the properties' age and possibly quality. These two characteristics could explain the differences in the value of the structure. Newer houses and better quality houses are usually more expensive. To test this hypothesis, columns 4 and 8 include the housing characteristics as additional controls. In particular, I control by the mean interior quality score, size, and age of the units in each block and the share of units built after 1997.

*Table 5.2:* Fuzzy RD estimate for appraisals

| | Semipar | | | | Non-parametric | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| | A.Mult-Family Units | | | | | | | |
| Land | -213.24$^{*}$ | -215.85 | -171.53$^{+}$ | -107.44 | -453.05 | -388.23 | 419.70 | 19.12 |
| | (107.77) | (138.12) | (100.12) | (78.62) | (408.10) | (324.65) | (765.80) | (163.96) |
| p-value | 0.76 | 0.80 | 0.93 | 0.36 | 0.50 | 0.52 | 0.43 | 0.22 |
| Structure | -209.76$^{+}$ | -230.84$^{+}$ | -197.26 | 6.20 | -658.88 | -518.64 | 478.16 | 28.10 |
| | (117.96) | (127.19) | (126.97) | (64.85) | (662.48) | (446.57) | (1,252.39) | (224.30) |
| p-value | 0.80 | 0.69 | 0.89 | 0.00 | 0.47 | 0.45 | 0.60 | 0.35 |
| Property | -296.69 | -318.96 | -257.83$^{+}$ | -91.25 | -847.15 | -966.25 | 2,106.98 | -224.02 |
| | (197.18) | (210.86) | (140.31) | (115.71) | (719.56) | (725.20) | (3,132.02) | (268.62) |
| p-value | 0.55 | 0.51 | 0.58 | 0.44 | 0.35 | 0.28 | 0.47 | 0.87 |
| | B.Single-Family Units | | | | | | | |
| Land | -169.61$^{***}$ | -161.52$^{***}$ | -176.34$^{***}$ | -114.76$^{***}$ | -169.71 | -119.45 | -240.70 | 106.37 |
| | (15.89) | (16.18) | (21.18) | (17.03) | (160.30) | (116.03) | (343.32) | (150.20) |
| p-value | 0.51 | 0.25 | 0.86 | 0.00 | 0.95 | 0.60 | 0.86 | 0.06 |
| Structure | -82.18$^{***}$ | -78.66$^{***}$ | -65.35$^{***}$ | -16.42$^{*}$ | -96.49 | -67.19 | -6.44 | -12.65 |
| | (9.35) | (5.73) | (8.55) | (6.54) | (66.85) | (44.88) | (34.37) | (35.77) |
| p-value | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.01 | 0.00 | 0.00 |
| Property | -175.91$^{***}$ | -166.33$^{***}$ | -171.92$^{***}$ | -119.33$^{***}$ | -44.42 | 81.00 | 45.48 | 391.54 |
| | (30.36) | (27.79) | (37.08) | (31.59) | (139.36) | (123.52) | (303.01) | (268.22) |
| p-value | 0.89 | 0.62 | 0.83 | 0.05 | 0.33 | 0.03 | 0.46 | 0.03 |
| | C.All Units | | | | | | | |
| Land | -167.62$^{***}$ | -166.48$^{***}$ | -176.78$^{***}$ | -99.31$^{***}$ | -122.79 | -162.99 | -38.06 | 59.56 |
| | (22.65) | (20.34) | (22.16) | (18.03) | (118.60) | (139.07) | (105.06) | (122.98) |
| p-value | 0.58 | 0.51 | 0.88 | 0.00 | 0.63 | 0.90 | 0.18 | 0.05 |
| Structure | -76.52$^{***}$ | -122.02$^{***}$ | -102.75$^{***}$ | -4.93 | -112.11 | -206.70 | -223.89 | -27.03 |
| | (16.88) | (8.86) | (12.35) | (7.85) | (78.85) | (166.68) | (422.83) | (36.29) |
| p-value | 0.00 | 0.00 | 0.00 | 0.00 | 0.39 | 0.87 | 0.92 | 0.00 |
| Property | -178.09$^{***}$ | -212.99$^{***}$ | -214.58$^{***}$ | -122.17$^{**}$ | -193.20 | -235.26 | -84.43 | -94.97 |
| | (27.80) | (23.31) | (29.03) | (38.66) | (177.38) | (189.88) | (119.01) | (97.28) |
| p-value | 0.95 | 0.16 | 0.23 | 0.13 | 0.94 | 0.77 | 0.42 | 0.38 |
| Controls | No | Yes | Yes | Yes | No | Yes | Yes | Yes |
| 1993 Char. | No | No | Yes | Yes | No | No | Yes | Yes |
| House Char. | No | No | No | Yes | No | No | No | Yes |

NOTE: This table presents the estimates for the RD design $\theta$ using three different estimation methods with and without controls.The estimates are the IV estimates using the assignment rule as an instrument for a block belonging to estrato 3. Each row is the estimate for a different variable. Columns 1 to 4 use different approaches to estimate $k(S_{izt})$ in the model $y_{izt} = \alpha + \theta_{2,3}\mathbb{1}_{[S_{izt}\geq\delta_2^3]} + k(S_{izt}) + \beta X + \varepsilon_{izt}$. Columns 1 and 2 use a partially linear model (Robinson, 1988) to estimate $k(S_{izt})$. Columns 5 and 6 estimate $\theta usingnonparametricapproach$ $\lim_{s_{izt}\to\delta_2^{3}+} \mathrm{E}(y_{izt}|S_{izt}=s_{izt}) - \lim_{s_{izt}\to\delta_2^{3}-} \mathrm{E}(y_{izt}|S_{izt}=s_{izt})$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth. The first stage and reduced form are represented in figures 4.5 and **??**. The controls included are a dummy equal to 1 if the block changed the estrato since 1997, habitat zones, and the share of blocks that are multi family units. House Char. controls includes age, quality score size and the share of houses built after 1997. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 8,9. I do not include the habitat zones 7 and 10 because they do not have enough block with multi family homes in estrato 2. I drop the observations with a score lower than -3 or higher than 10.

This table shows that the housing market seems to capitalize the savings caused by the subsidy scheme. Independent of the estimation method, the land price is indistinguishable from the multifamily units' full capitalization benchmark. The point estimates for single-family units when I do not include controls are the same for the partially linear and non-parametric approaches; $-169$ Thousand $COP/m^2$. It is a bit higher in the parametric approach. Once I include controls for housing characteristics, the partially linear model's point estimate is 98.40, about half of the full capitalization benchmark. I cannot reject the null hypothesis that they are the same for any of the estimation approaches. If I pool all the units ( in panel C), I get a similar result. Consistent with a spatial equilibrium prediction, these estimates suggest that the housing market capitalizes the subsidy into land prices.

In terms of structure, the mean shift in age and quality score seems to explain the discontinuity in structure prices apparent in figures 5.10. The point estimate is around $-20$ Thousand $COP/m^2$ for all the models using the partially linear and parametric approaches once I include the controls. The non-parametric approach estimates are very noisy, particularly for multi family units. I cannot reject the null hypothesis that they are equal to the full capitalization in all the cases. The only exception is structure when I introduce controls for the quality characteristics in the single-family units. In this case, the point estimate is $-6.44$ Thousand $COP/m^2$. This result suggests that the age and quality effect's market price is close to the full capitalization effect.

The total property effect for all the units has a point estimate of -127, -151, and -84 in each estimation method after I control for the housing characteristics. The point estimates are similar when I estimate separately for multi and single-family units. I cannot reject the null hypothesis that the housing market capitalizes on the subsidy in all the cases and estimation methods. This is the case after I control for housing characteristics and the effect of new housing described previously.

Two main results emerge. First,the city growth is being affected by the subsidy scheme. The units in the heavily subsidized areas are newer, and the share of buildings built after the stratification methodology started is larger. Second the subsidy is offset by an increase in housing prices. House characteristics and construction timing explain most of the observed mean shifts in the price of the structures. But the land values and property values in estrato 2 are higher, even after controlling by housing characteristics. I cannot reject the hypothesis that the effect is equal to the full capitalization.

# 6 Robustness

The control function $g(s)$ in my setting is an object of interest, It shows the relationship between neighborhood quality and the outcome of interest. Under the assumption that this function is the same in both sides of the cutoffs, the partially linear model and a parametric approach are more efficient at estimating the causal effect of the subsidy scheme because I use all the available data, instead of using only the data around the cutoff. Even without relying on the full sample and the meaning of the control function, I can use the fact that the blocks to the right and left of the cutoff have similar probability of being treated to identify the effect of the subsidy. The optimal bandwidth estimation relies more heavily on this assumption and the control function is not relevant. This comes at the cost of using only the information around the cutoff. To check the sensitive on the non parametric approach to the selection of the band with. I compare my estimates with the estimates using different bandwidths.

Figure 6.11, shows the estimates for age, probability of being built after 1997, interior quality score, and the different measures of property values. The x-axis represents the bandwidth; the y-axis is the point estimate. The dashed line is the parametric approach, the dotted line is the partially linear model, and the dot is the optimal bandwidth estimate. The solid line with the shaded area is the estimates for the different bandwidth. The figure shows that the estimates for housing characteristics do not depend on the estimation method. The point estimates of the partially linear model and the parametric approach are always in the confidence interval of the estimates using a bandwidth. The figure makes the trade-off between bias and precision of the estimates. Using a bandwidth, I am more agnostic about the functional form of the control function. However, this comes at an efficiency cost, because I only use the information around the cutoff. The figures for the housing values illustrate this point clearly. With a narrow bandwidth, the point estimates have large confidence intervals and include the point estimates of the other less agnostic approaches. When I increase the bandwidth, I gain precision on my estimates, and the bandwidth estimates converge towards the partially linear model and the parametric estimates. This gives me some confidence in those approaches. While this may not be the case in some RDD applications, in this particular case, using the full sample and estimating the control function seems reasonable and allows for point estimates more precisely estimated.

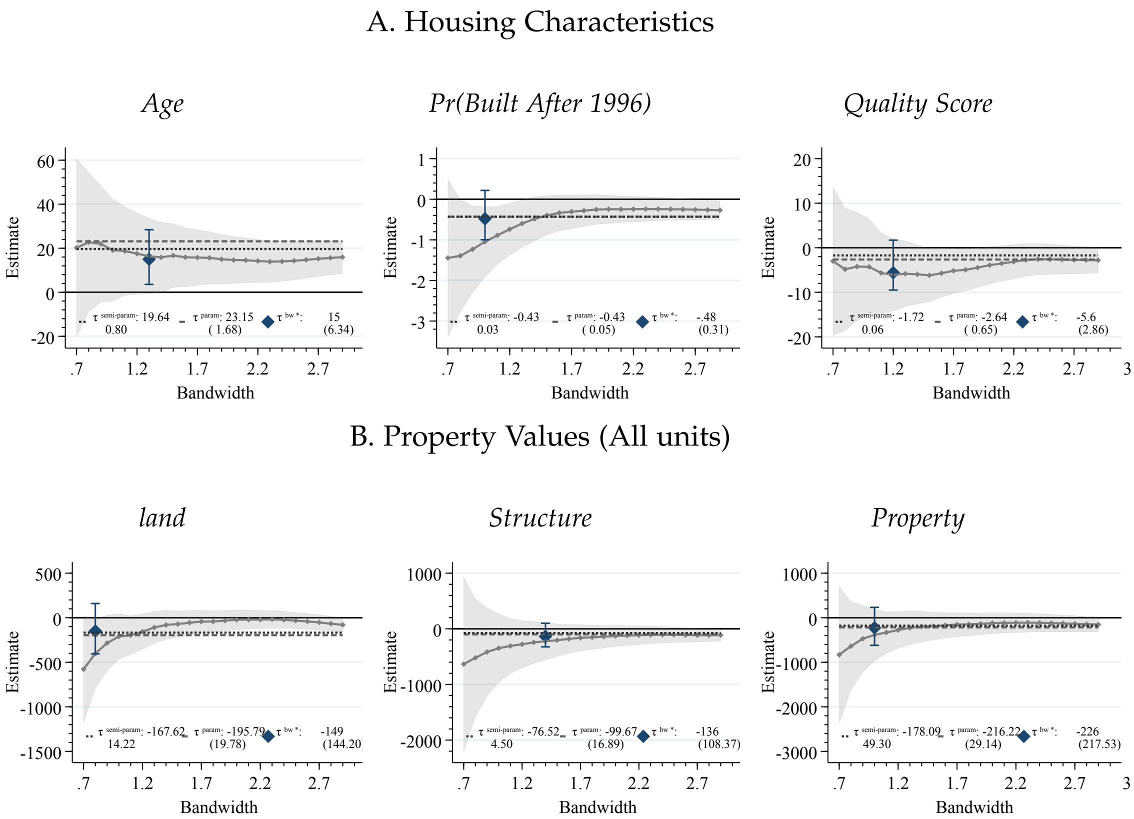



*Figure 6.11:* Estimates Using Different Bandwidths

*Table 6.3:* THE EFFECT ON CONSTRUCTION AND HOUSING CHARACTERISTICS.

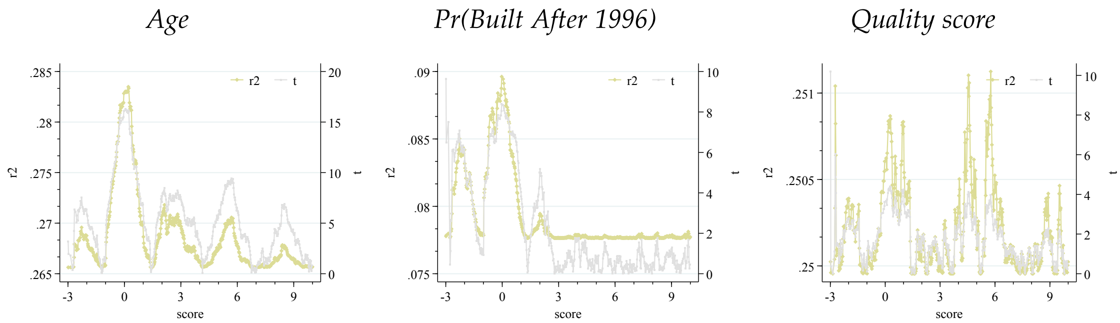
| | (1) | (2) | (3) |
|---|---|---|---|
| Pr(Built>1996) | -0.43*** | -0.42*** | -0.40*** |
| | (0.05) | (0.05) | (0.05) |
| Age | 23.15*** | 25.03*** | 19.71*** |
| | (1.68) | (1.57) | (1.49) |
| Quality Score | -2.64*** | -2.31*** | -3.85*** |
| | (0.65) | (0.64) | (0.63) |
| Size $m^2$ | -23.25** | -8.10 | 0.51 |
| | (7.83) | (7.62) | (7.56) |
| Controls | No | Yes | |

NOTE: This table presents the estimates $\theta$ using three different estimation methods with and without controls.The estimates are the IV estimates using the assignment rule as an instrument for a block belonging to estrato 3. Each row is the estimate for a different variable. Columns 1 to 4 use different approaches to estimate $k(S_{izt})$ in the model $y_{izt} = \alpha + \theta_{2,3}\mathbb{1}_{[S_{izt} \geq \delta_2^3]} + k(S_{izt}) + \beta X + \varepsilon_{izt}$. Columns 1 and 2 use a partially linear model Robinson (1988) to estimate $k(S_{izt})$. Columns 5 and 6 estimate $\theta$ using non parametric approach $\theta = \lim_{s_{izt} \to \gamma_2^{3+}} \lim \mathrm{E}(y_{izt}|S_{izt} = s_{izt}) - \lim_{s_{izt} \to \gamma_2^{3-}} \mathrm{E}(y_{izt}|S_{izt} = s_{izt})$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth.The first stage and reduced form are represented in figures 4.5 and 5.9. The controls included are a dummy equal to 1 if the block changed the estrato since 1997, habitat zones, and the share of blocks that are multi family units. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 7,8,9 and 10. I drop the observations with a score lower than -3 or higher than 10.

Significance level $^{+}$10%,$^{*}$ 5%,$^{**}$ 1%$^{***}$0.1%

The second robustness exercise is to check that the mean shift I observed happens only at the cutoff. To check this, I plot the $R^2$ and t-statistic using different hypothetical cutoff points. If the biggest t-statistic and $R^2$ is at the cutoff, this is a good indicator that the mean shift is not an artificial jump unrelated to the subsidy scheme. I use the parametric approach, including the basic set of controls. Figure 6.12 shows the results for the select set of outcomes. There is an important spike on the t-stat and $R^2$ in all the outcomes but on the interior quality score. This gives less confidence to that result. Also, it is noteworthy that for the case of land, there seems to be another mean shift around 2.5. This may reduce the confidence in the land price estimates.

A. Housing Characteristics




B. Housing Values (All Units)

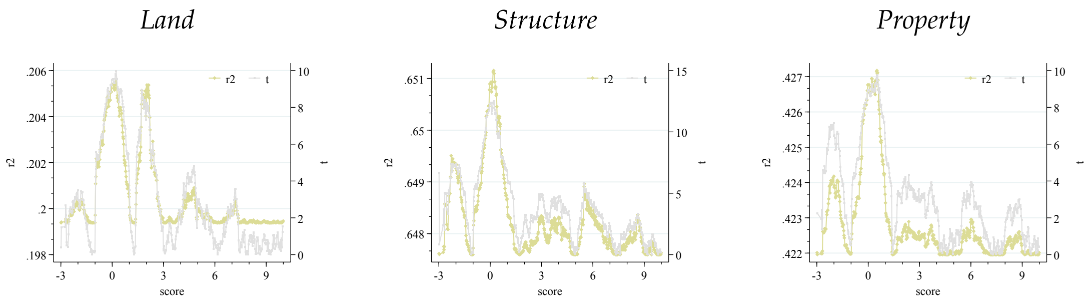



*Figure 6.12:* $R^2$ and $t$ at different cutoffs

## 7 Concluding remarks

My findings suggest that using a location to target redistribution policies has an impact on the housing market. There is any apparent effect on the new construction and capitalization of the subsidy into the housing market. Blocks in areas receiving a higher subsidy, have housing units 30 years more modern on average and have 43 percent more chances of having units built after the new targeting tool was introduced. The houses in blocks of estrato 2 capitalize on the differences in utility prices and property taxes. The NPV of the differences in costs is 180 thousand $COP/m^2$, and my preferred estimate for land prices is 89 and 127 thousand $COP/m^2$ for total property value.Under all the specifications, I cannot reject the null hypothesis of full capitalization.

These results have several implications that need to be a considered when using location as a targeting tool. First, the intended redistributive purposes can fail if the subsidy ends up being transferred to property owners in targeted areas. In the case of Bogotá, around 50 percent of households rent their houses in the neighborhoods in estrato 2 and 3. If the increase in housing prices get charged into monthly rents, the renters do not benefit from the transfers. In Colombia,renters usually pay the utilities bill. Moreover, if developers and not individual families drive the new construction, the subsidy scheme could be developers and not poor people.

Additionally, these types of schemes may generate inefficient city growth. If individuals and developers build new units to take advantage of the subsidy and not to get closer to job opportunities or market access, the city may grow inefficiently. People may decide to live far away from jobs and spend more time commuting, to take advantage of the subsidy. Developments driven by the subsidy will likely persist even if the subsidy scheme disappears. Finally, the transfer system may prevent some areas from experiencing dynamic improvements if the resident fears to lose the subsidy status. A careful analysis of those potential implications and the costs of this paper's findings needs to be considered when evaluating the efficacy of using a location as a targeting tool. This is an area of future research.

Recent research suggests that neighborhoods are critical for determining future income (Chetty, Hendren, and Katz (2016), Chetty, Friedman, Hendren, Jones, and Porter (2018)). A possible policy implication of these studies is to use location characteristics to target redistributive policies. A location-based transfer affects people's decisions about where to live, and therefore it may affect the housing market. These unintended effects can offset the intended impact of the location-based policy. Gaubert et al. (2020) suggests that location-based subsidy could be a useful tool for redistribution. I provide some evidence of the unintended consequences of using location. The costs of these consequences should be included in a framework analyzing the efficacy of location-based subsidies.

# Appendix

## 2.1 Appendix:

*a. Monthly Subsidy and Property Tax* *b. ... as a % of HH expenditure*


*Figure 2.13:* Subsidy Size

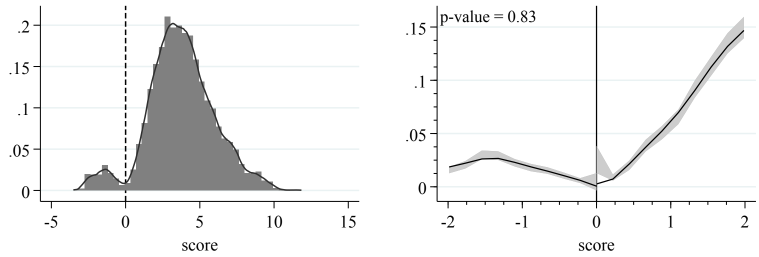
Note: The subsidy is the monthly average for each subsidy code. The property tax is the estimated payment of a house of average size assessed at an average price for $m^2$ divided by 12 within each code. As reference point, the GDP per capita in Colombia in 2011 was 13.4 million COP and the the average exchange rate for 2011 was 1848.17 COP pesos for a Dollar.

### 2.1.1 Baseline Characteristics

#### *No manipulation assumption (individual years)*

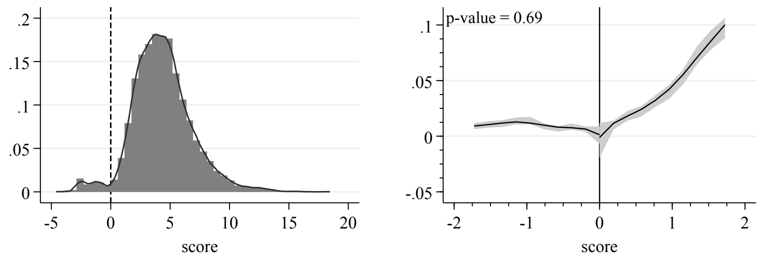
1997

*Density* *McCrary Manipulation test*




2013

*Density* *McCrary Manipulation test*




2017

*Density* *McCrary Manipulation test*

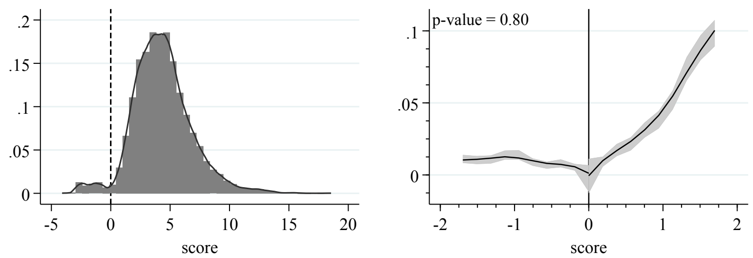



*Figure 2.14:* Score Distribution

Note: $H_0$: The density is the same on the right and on the left. The p-value of the test is on the top of panel b). I cannot reject the null hypothesis. To implement the manipulation test I follow (M. Cattaneo, Jansson, & Ma, 2018)

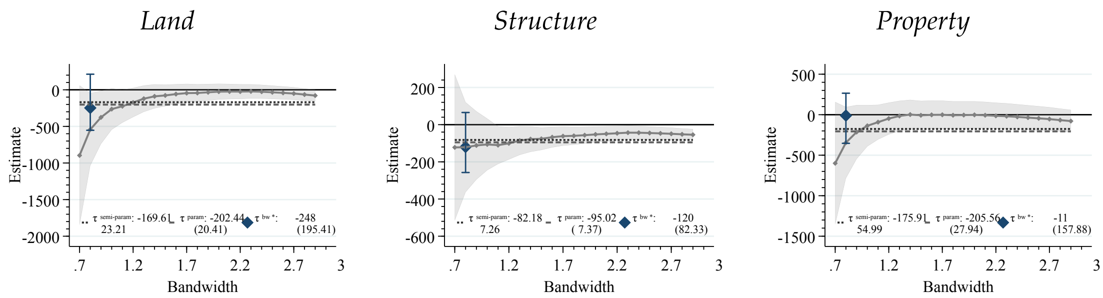
*Table 2.4:* THE EFFECT ON CONSTRUCTION AND HOUSING CHARACTERISTICS. (HABITAT ZONES 8 AND 9)

| | Semipar | | Parametric | | Non-parametric | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pr(Built>1996) | -0.49*** | -0.49*** | -0.48*** | -0.48*** | -0.68+ | -1.05+ |
| | (0.05) | (0.08) | (0.05) | (0.05) | (0.40) | (0.61) |
| Age | 22.36*** | 24.47*** | 25.41*** | 27.17*** | 17.45*** | 27.09* |
| | (1.58) | (1.33) | (1.59) | (1.49) | (5.30) | (11.86) |
| Quality Score | -1.60* | -1.82*** | -2.77*** | -2.36*** | -3.55+ | -4.62 |
| | (0.65) | (0.52) | (0.66) | (0.65) | (2.01) | (4.52) |
| Size $m^2$ | 2.65 | 21.22*** | -12.86+ | 3.04 | 15.60 | 68.15 |
| | (6.69) | (6.24) | (7.01) | (6.97) | (19.50) | (68.94) |
| Controls | No | Yes | No | Yes | No | Yes |

NOTE: This table presents the estimates $\theta$ using three different estimation methods with and without controls.The estimates are the IV estimates using the assignment rule as an instrument for a block belonging to estrato 3. Each row is the estimate for a different variable. Columns 1 to 4 use different approaches to estimate $k(S_{izt})$ in the model $y_{izt} = \alpha + \theta_{2,3}\mathbb{1}_{[S_{izt} \geq \delta_2^3]} + k(S_{izt}) + \beta X + \varepsilon_{izt}$. Columns 1 and 2 use a partially linear model Robinson (1988) to estimate $k(S_{izt})$. Columns 5 and 6 estimate $\theta$ using non parametric approach $\theta = \lim_{s_{izt} \to \gamma_2^{3+}} \lim \mathrm{E}(y_{izt}|S_{izt} = s_{izt}) - \lim_{s_{izt} \to \gamma_2^{3-}} \mathrm{E}(y_{izt}|S_{izt} = s_{izt})$. In this approach I use the method proposed by M. Cattaneo, Idrobo, and Titiunik (2018) to select the optimal bandwidth.The first stage and reduced form are represented in figures 4.5 and 5.9. The controls included are a dummy equal to 1 if the block changed the estrato since 1997, habitat zones, and the share of blocks that are multi family units. The numbers in parenthesis are robust standard errors. The figure contains all the blocks in estratos 2 and 3 in habitat zones 7,8,9 and 10. I drop the observations with a score lower than -3 or higher than 10.

Significance level $^+$10%,$^*$ 5%,$^{**}$ 1%$^{***}$0.1%

## Single Family Units




## Multi-Family Units

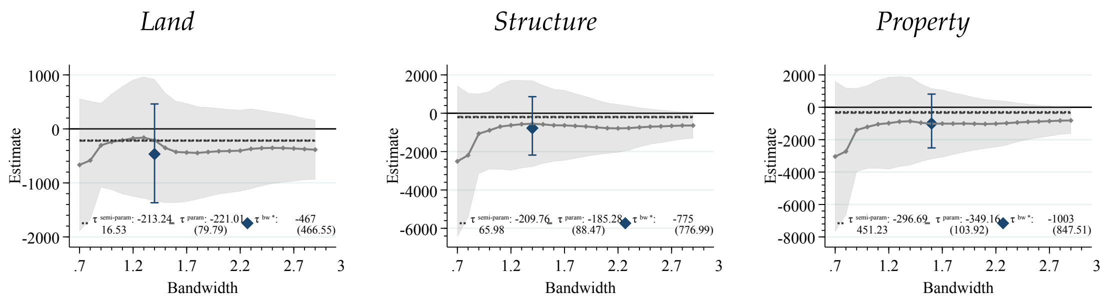
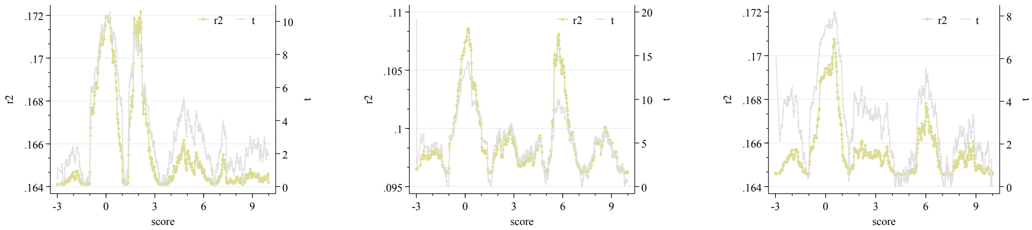



*Figure 2.15:* Using Different Bandwidths

## Single Family

*Land* *Structure* *Property*




## Multi-Family

*Land* *Structure* *Property*

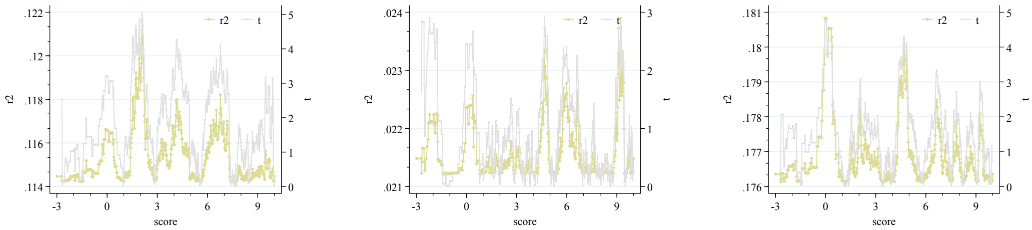



*Figure 2.16:* $R^2$ and $t$ at different cutoffs

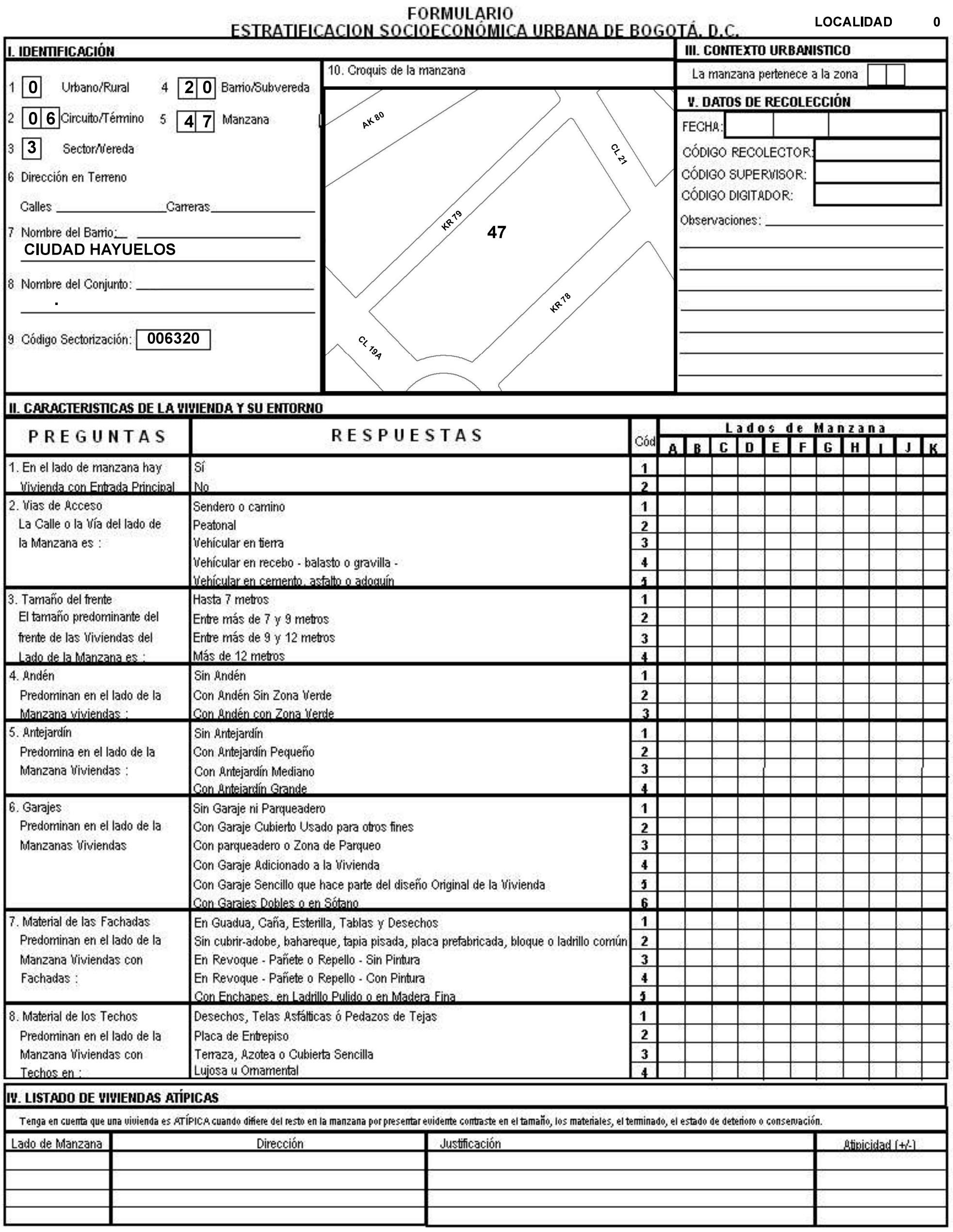
### 2.1.2 Estratos Assignment Details

*Table 2.5:* Percentage of Blocks Correctly Predicted

| | 1997 | 1999 | 2009 | 2013 | 2017 | All |
|---|---|---|---|---|---|---|
| Estrato 1 | 99.87 | 99.91 | 99.61 | 99.56 | 99.50 | 99.69 |
| Estrato 2 | 99.89 | 99.91 | 99.54 | 99.47 | 99.41 | 99.65 |
| Estrato 3 | 99.98 | 99.99 | 99.86 | 99.78 | 99.74 | 99.87 |
| Estrato 4 | 99.74 | 99.91 | 98.91 | 98.18 | 98.21 | 98.99 |
| Estrato 5 | 98.03 | 98.09 | 94.06 | 93.67 | 93.27 | 95.45 |
| Estrato 6 | 99.52 | 98.75 | 98.11 | 97.97 | 97.82 | 98.45 |
| All | 99.85 | 99.86 | 99.44 | 99.32 | 99.25 | 99.55 |
| Observations | 36985 | 37686 | 37089 | 37096 | 36258 | 185114 |

Note:

## 2.1.3 How Information is Collected



FORMULARIO
ESTRATIFICACION SOCIOECONÓMICA URBANA DE BOGOTÁ, D.C.

LOCALIDAD 0

**I. IDENTIFICACIÓN**

1 [0] Urbano/Rural
4 [2 0] Barrio/Subvereda
2 [0 6] Circuito/Término
5 [4 7] Manzana
3 [3] Sector/Vereda
6 Dirección en Terreno
Calles ________ Carreras ________
7 Nombre del Barrio: CIUDAD HAYUELOS
8 Nombre del Conjunto: ________
9 Código Sectorización: 006320

10. Croquis de la manzana



**III. CONTEXTO URBANISTICO**

La manzana pertenece a la zona

**V. DATOS DE RECOLECCIÓN**

FECHA:
CÓDIGO RECOLECTOR:
CÓDIGO SUPERVISOR:
CÓDIGO DIGITADOR:
Observaciones:

**II. CARACTERISTICAS DE LA VIVIENDA Y SU ENTORNO**

| PREGUNTAS | RESPUESTAS | Cód | Lados de Manzana | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | B | C | D | E | F | G | H | I | J | K |
| 1. En el lado de manzana hay Vivienda con Entrada Principal | Sí | 1 | | | | | | | | | | | |
| | No | 2 | | | | | | | | | | | |
| 2. Vias de Acceso La Calle o la Vía del lado de la Manzana es : | Sendero o camino | 1 | | | | | | | | | | | |
| | Peatonal | 2 | | | | | | | | | | | |
| | Vehícular en tierra | 3 | | | | | | | | | | | |
| | Vehícular en recebo - balasto o gravilla - | 4 | | | | | | | | | | | |
| | Vehícular en cemento, asfalto o adoquín | 5 | | | | | | | | | | | |
| 3. Tamaño del frente El tamaño predominante del frente de las Viviendas del Lado de la Manzana es : | Hasta 7 metros | 1 | | | | | | | | | | | |
| | Entre más de 7 y 9 metros | 2 | | | | | | | | | | | |
| | Entre más de 9 y 12 metros | 3 | | | | | | | | | | | |
| | Más de 12 metros | 4 | | | | | | | | | | | |
| 4. Andén Predominan en el lado de la Manzana viviendas : | Sin Andén | 1 | | | | | | | | | | | |
| | Con Andén Sin Zona Verde | 2 | | | | | | | | | | | |
| | Con Andén con Zona Verde | 3 | | | | | | | | | | | |
| 5. Antejardín Predomina en el lado de la Manzana Viviendas : | Sin Antejardín | 1 | | | | | | | | | | | |
| | Con Antejardín Pequeño | 2 | | | | | | | | | | | |
| | Con Antejardín Mediano | 3 | | | | | | | | | | | |
| | Con Antejardín Grande | 4 | | | | | | | | | | | |
| 6. Garajes Predominan en el lado de la Manzanas Viviendas | Sin Garaje ni Parqueadero | 1 | | | | | | | | | | | |
| | Con Garaje Cubierto Usado para otros fines | 2 | | | | | | | | | | | |
| | Con parqueadero o Zona de Parqueo | 3 | | | | | | | | | | | |
| | Con Garaje Adicionado a la Vivienda | 4 | | | | | | | | | | | |
| | Con Garaje Sencillo que hace parte del diseño Original de la Vivienda | 5 | | | | | | | | | | | |
| | Con Garajes Dobles o en Sótano | 6 | | | | | | | | | | | |
| 7. Material de las Fachadas Predominan en el lado de la Manzana Viviendas con Fachadas : | En Guadua, Caña, Esterilla, Tablas y Desechos | 1 | | | | | | | | | | | |
| | Sin cubrir-adobe, bahareque, tapia pisada, placa prefabricada, bloque o ladrillo común | 2 | | | | | | | | | | | |
| | En Revoque - Pañete o Repello - Sin Pintura | 3 | | | | | | | | | | | |
| | En Revoque - Pañete o Repello - Con Pintura | 4 | | | | | | | | | | | |
| | Con Enchapes, en Ladrillo Pulido o en Madera Fina | 5 | | | | | | | | | | | |
| 8. Material de los Techos Predominan en el lado de la Manzana Viviendas con Techos en : | Desechos, Telas Asfálticas ó Pedazos de Tejas | 1 | | | | | | | | | | | |
| | Placa de Entrepiso | 2 | | | | | | | | | | | |
| | Terraza, Azotea o Cubierta Sencilla | 3 | | | | | | | | | | | |
| | Lujosa u Ornamental | 4 | | | | | | | | | | | |

**IV. LISTADO DE VIVIENDAS ATÍPICAS**

Tenga en cuenta que una vivienda es ATÍPICA cuando difiere del resto en la manzana por presentar evidente contraste en el tamaño, los materiales, el terminado, el estado de deterioro o conservación.

| Lado de Manzana | Dirección | Justificación | Atipicidad (+/-) |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |

*Figure 2.17:* Stratification Census Form

Note: This Figure shows the form used to collect the information required to assign the *estratos*

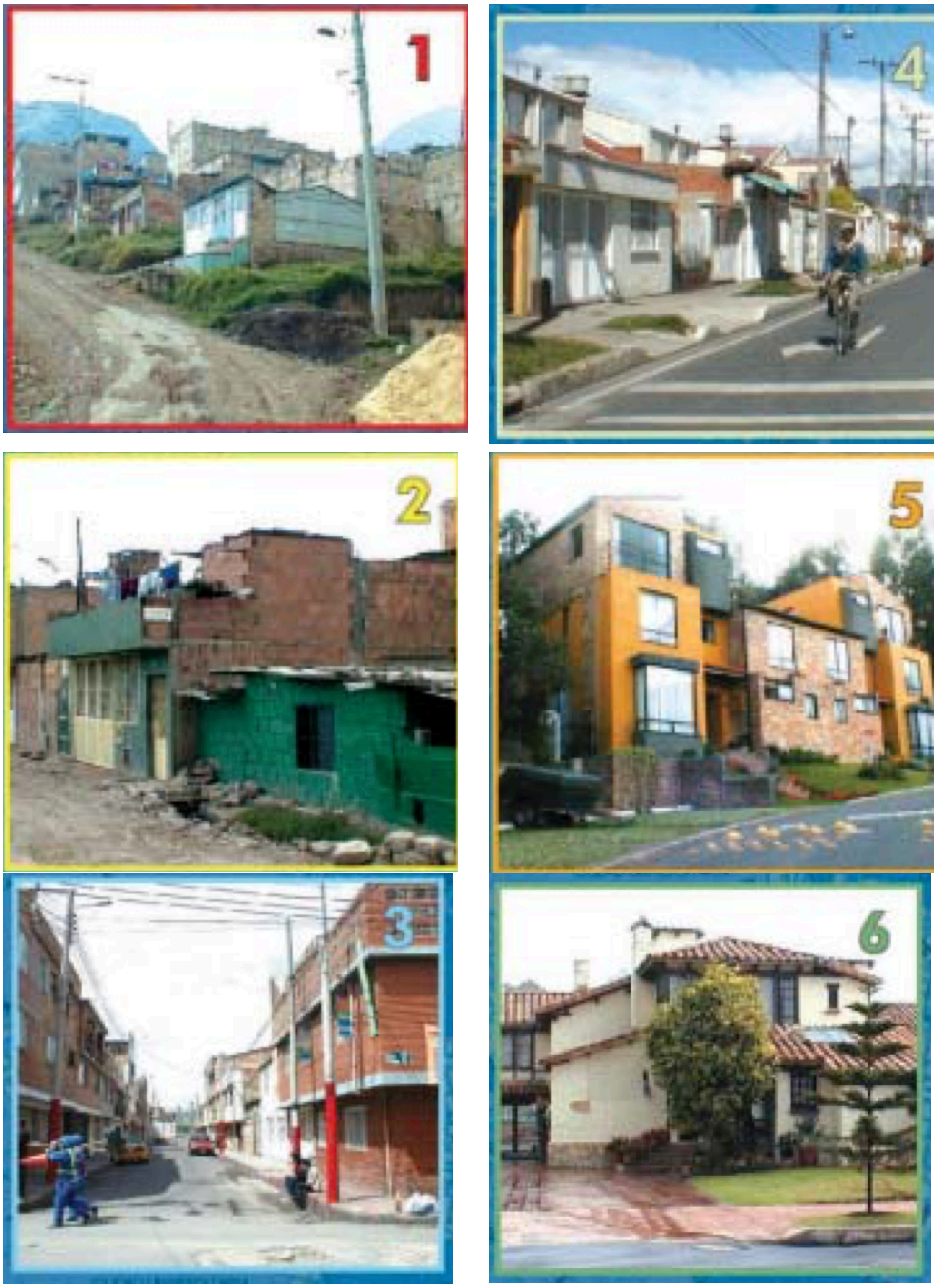



*Figure 2.18:* Example Units by estrato

Note:

### 2.1.4 Defining the Habitat Zones

*Table 2.6:* Habitat zones.

| General zone | Habitat zone | Description |
|---|---|---|
| 1. Poverty | 1<br>2 | Blocks with lacking planning. Fragile and short-lived materials characterize the houses. The houses usually do not have spaces specifically designed for cooking, bathing, or washing clothes. These houses have high housing density. Public areas, such as platforms and roads, are limited. In their immediate surroundings typically have black waters, garbage, and animals. They present problems of natural risk (i.g. landslides, floods, marshy lands, etc.) |
| 2.Tolerance zone | 3 | Blocks characterized by high delinquency, prostitution, distribution, or consumption of drugs. |
| 3. Unconsolidated development | 4<br>5 | Blocks with unfinished houses. They can be in blackwork or with uncovered or unpainted facades and temporary roofs, whose planks will eventually. Homes built in stages over the years.<br>Blocks with scattered buildings or sides of the block with an abundant presence of lots without closing or without a known destination; These characteristics make these spaces appear in the process of consolidation. |
| 4.Urban deterioration | 6 | Blocks in the historic center of the city. The facade looks old (manifested by aged ceilings, broken eaves, deteriorated facades, raised paintings, doors, corroded windows, etc.). The immediate surroundings show narrow streets and sidewalks, without front yards or green areas. We observe the presence of economic establishments of a different use (shops, mechanic workshops, canteens, or restaurants) that degrade the quality of homes. In some cities, the tenants are in these areas; it refers to large houses where several families separated in independent pieces with sanitary, kitchen, and patio clothing services. |
| 5.Industrial | 7 | Blocks with "factories" - destined for the massive production of goods. Surrounding areas have warehouses, premises for the sale of food, and the permanent traffic of trucks. The waste and noise, characteristic of the factories, pollute the environment. |

| | | |
|---|---|---|
| 6.Consolidated progressive development | 8<br>9 | Blocks with self-construction homes. They express the culmination of progressive development. For this reason, the landscape is heterogeneous or architectural diverse. The buildings occupy the space on each side of the block in a continuous manner, in such a way that its urban structure can be considered consolidated and definitive. It can include social housing, finished, and built-in series. |
| 7. Commercial | 10<br>11 | Blocks with mainly commercial buildings; initially, many of them were housing units but were refurbished for shops. Usually, homes are on the upper floors or in the interior part of the buildings. This zone is the CBD; in small cities, it is located around the plaza or along the main street. In some cities, these areas exist in the traditional historical center but could be in other places, such as neighborhoods and high vehicular traffic lanes. The public space is minimal, and the congestion of customers degrades the quality of the homes. |
| 8.Intermediate residential | 12<br>13 | Block with finished homes in residential neighborhoods. The immediate surroundings have wide public spaces, streets in good condition, green areas, and low density of commercial establishments. |
| 9.Commercial compatible | 14 | Blocks with residential homes and access to amenities such as gyms and saloons, boutiques or luxury goods stores, cigar stores, warehouses with decorated showcases, florists, etc.<br>The immediate environment have main roads with high vehicular traffic and alternate roads with less traffic. |
| 10.Exclusive residential | 15<br>16 | Blocks with residential buildings with modern designs, green areas, special systems of private surveillance, and almost no presence of economic establishments. |
| 11.Low density residential | 17 | Blocks with residential homes with the architectural design of their homes and salient decoration (i.g. fountains, gardens, lighting systems, etc.). It is characterized by big houses, mansions, or majestic buildings. In some cases, the roads are exclusive for residents and visitors and have private services. |
| 12.Institutional | 18<br>19<br>20 | Non-residential blocks. This zone includes blocks with institutional use, those with lots and others without homes, and those dedicated to green areas.. |

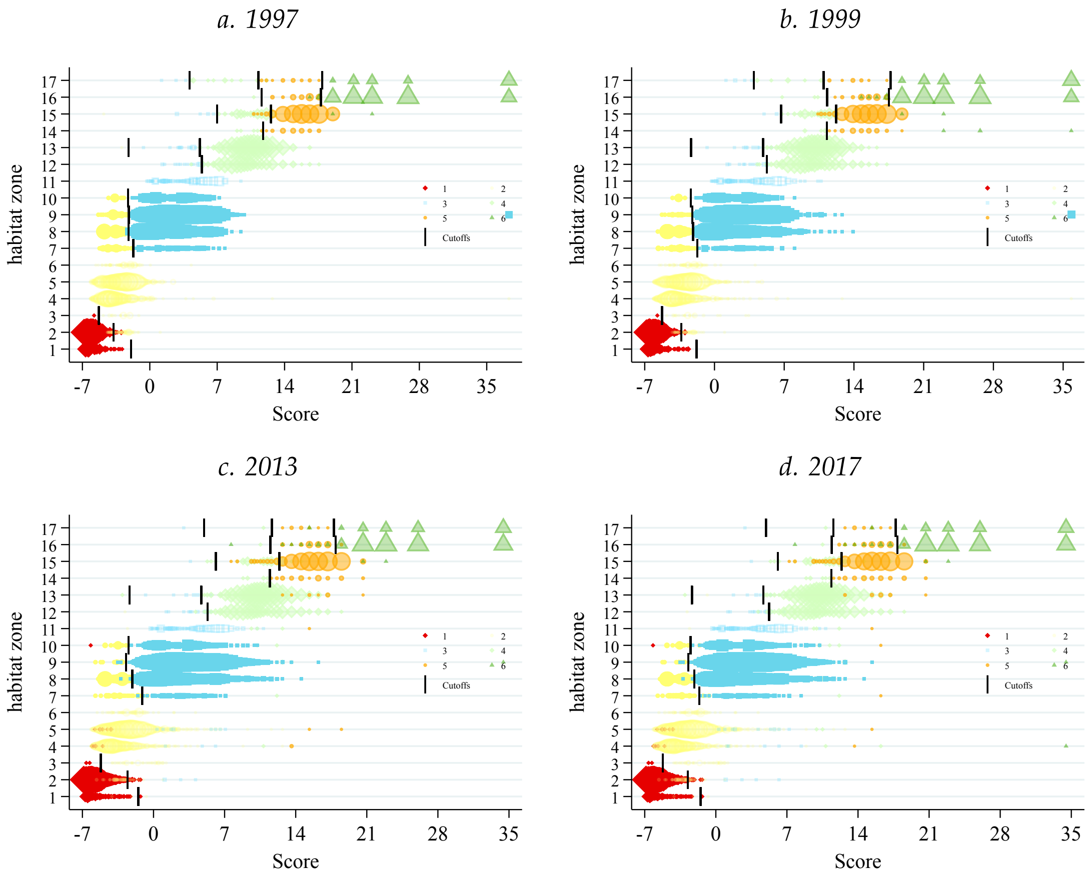



*Figure 2.19:* Subsidy Codes Based on Habitat Zones and the Score

NOTE:

The proposed cutoffs are the dark solid lines. Each dot in the figure represents one or more blocks with a given score and zone. The symbols are weighted to represent the number of blocks with a particular score and zone. The six different colors represent the six subsidy codes. For example, the green cross are blocks belonging to subsidy code 3. There are three main takeaways from this figure. First, the habitat zones play a more important role than the score at determining the different codes. Second, the worst zones (lower numbers) have lower scores than the best areas (higher numbers). However, there is some overlap in the index between some areas. Third, only a few zones have more than one code. For my empirical strategy I use the discontinuities


*Figure 2.20:* Reassignment of *Estratos*
Note: *source:* (Departamento Administrativo de Planeación Distrital, 2004)

introduced by the cutoffs.

### 2.1.5 Data Details

#### Constant Geographic Unit

There are two geographic codes for the blocks in Bogota. The Census Block Code from the National Geo-static Framework (*MGN*-from the Spanish spelling) (DANE, 2018) and the the cadaster block code (*ManCodigo* (Igac, 2013)). The National Adminisrative Department of Statistics (*DANE*-from the Spanish spelling) uses the The *MGN* code, and the district cadaster and the secretary of district planning use *ManCodigo* code to locate each property and to produce the stratification census.

The *MGN* code starts in 1993 with that year's population census. The codes have no significant changes over time, therefore we can track units over time. This is not the case for the cadaster block code, *ManCodigo*. Some blocks in the same geographic location change the *ManCodigo* over time. Therefore, to follow the same location over time, we need to define a time-invariant code for each location and unit of observation. I use the shapefiles of the Stratification Census to create such a code.[72] The shapefiles with the codes corresponding to the stratification census for the all the updates, allow me to know the exact geographic location of each unit independently of the code assigned to the unit in each year. The appendix figure 2.21 shows the changes in the residential city boundary. The appendix figure 2.21a shows the changes in residential areas in the city since 1997. The appendix figure 2.21b shows the places with codes that existed in 1997 but not in 2017. This figure makes it clear that the changes in codes are not only related to a change in residential city boundaries. Therefore, a time invariant geographic code is essential to analyze the data.

I create a stable geographic unit over time and create a crosswalk between the the two types of codes (MGN and ManCodigo). To do that, I fix the 2017 geographic codes, *ManCodigo*, and assign that code to each unit of observation based on the location obtained with the corresponding shapefile. The crosswalk between *MGN* and *ManCodigo* allows me to use the 1993 Population Census, to check for the balance in the characteristics of each block before the policy implementation. I create the crosswalk using the shapefiles for each type of code and doing a geographic merge, similar to what I did for the time-invariant code.[73]

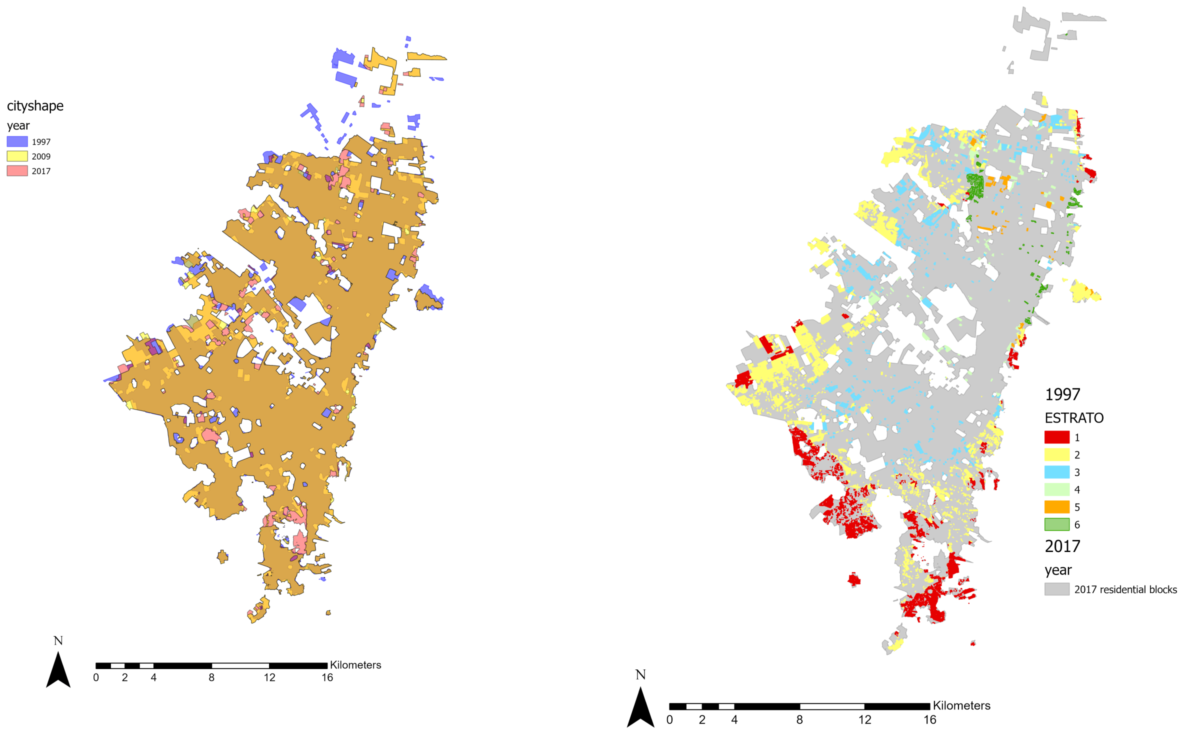
[72] The secretary of district planning provided me those shape files.

[73] For the MGN I only have the 2005 shape file. In total I geolocate an important fraction of the blocks. Aliaga-Linares and Alvarez-Rivadulla (2010) does a similar exercises.

*Figure 2.21:* Changes in codes and city boundaries

*a. Changes in city boundaries*

*b. Code changes between 1997 and 2017*




Note: The figure 2.21b shows the blocks that "exit" from 1997 to 2017 if we do not adjust the administrative code changes. Figure **??** shows the residential blocks for the different years where and update of the *stratification census* took place

### *Changes over time*

Because I have a time-invariant geographic code, I can check the changes in locations for each unit over time. Appendix Table 2.7 shows the changes in the estratos overtime. The top panel shows the changes between 2009 and 1997. Out of the 36,985 blocks in 1997, 32,425 have the same estrato in the two periods, 1,006 have a lower estrato in 2009 and 533 have a higher estrato, 3,125 blocks starting being defined as residential blocks and 3,021 were removed.

*Table 2.7:* Changes in the Subsidy Codes Over Time

| | 2009-1997 | | | | | | |
|---|---|---|---|---|---|---|---|
| | Same | Down | Up | New | Exit | Total 1997 | Total 2009 |
| 1 | 5,249 | . | 114 | 1,063 | 1,023 | 6,386 | 6,426 |
| 2 | 12,881 | 289 | 316 | 1,360 | 1,361 | 14,847 | 14,846 |
| 3 | 10,674 | 477 | 69 | 436 | 455 | 11,675 | 11,656 |
| 4 | 2,029 | 181 | 28 | 151 | 45 | 2,283 | 2,389 |
| 5 | 877 | 25 | 6 | 46 | 55 | 963 | 954 |
| 6 | 715 | 34 | . | 69 | 82 | 831 | 818 |
| All | 32,425 | 1,006 | 533 | 3,125 | 3,021 | 36,985 | 37,089 |
| | 2017-2009 | | | | | | |
| | Same | Down | Up | New | Exit | Total 2009 | Total 2017 |
| 1 | 6,375 | . | 35 | 50 | 193 | 6,603 | 6,460 |
| 2 | 14,268 | 4 | 18 | 43 | 548 | 14,838 | 14,333 |
| 3 | 11,379 | 4 | 12 | 34 | 188 | 11,583 | 11,429 |
| 4 | 2,266 | 2 | 3 | 11 | 22 | 2,293 | 2,282 |
| 5 | 969 | 2 | . | 5 | 6 | 977 | 976 |
| 6 | 769 | . | . | 9 | 26 | 795 | 778 |
| All | 36,026 | 12 | 68 | 152 | 983 | 37,089 | 36,258 |

.

# Chapter 3

# Internet Expansion and School Performance: Evidence from Colombia

coauthored with Aaron Weisbrod

***Abstract***

*We study the impact of a large public-policy push expanding internet access on secondary school test scores in Colombia. We use an instrumental variable approach exploiting the costliness of extending the existing internet infrastructure to connect new areas to identify causal impacts. We find that this internet expansion did have an effect on math test scores, which was concentrated in the bottom third of the test score distribution. Our estimates suggest that every* 10% *increase in the number of test-takers with access to the internet resulted in a 0.06 SD increase in math scores for this group. We find no significant effects on language test scores. We also present evidence that this expansion did not result in increased rates of test-taking students working or increased family incomes during this period.*

# 1 Introduction

The World Bank estimates that the percent of the world population using the internet has increased by more than seven times since 2000 CE. Internet infrastructure is expanding across the world, connecting populations in both high- and low-income countries to the global information economy. We have seen the positive impacts of these expansions on local labor markets, but we know relatively little about its impact on education outcomes, especially in low- and middle-income contexts.[74] This paper is the first to our knowledge quantifying the impact of these internet rollouts on education outcomes. We study a public policy push in Colombia (Vive Digital) that encouraged internet providers to expand their infrastructure to bring access to previously unconnected areas. This is an important and common context because infrastructure investments remain very expensive and government intervention is often necessary to push this access into poorer and more rural areas. This is especially salient for lower-income countries. It is therefore important to not only quantify the impacts of increasing this internet access on education outcomes, but also present the beneficial effects of these government-led expansion policies.

We study the impact of Colombia's internet rollout on the test score distributions of over 5,000 individual secondary schools. Our estimates show that increasing the percent of schools' students that have access to the internet had a significant impact on mathematics test scores, particularly at the bottom of the test score distribution. The estimates suggest that for every 10% increase in the percent of students with access to the internet, it increased a school's mathematic test scores by 0.06 standard deviations in the bottom third of their test score distribution. We did not detect any similar impact on language test scores. This shows that schools that experienced the largest increases in internet access for their students during this time also experienced

[74] Some examples are: Forman, Goldfarb, and Greenstein (2012), Akerman, Gaarder, and Mogstad (2015), Hjort and Poulsen (2019).

significant catch-up of their worst performing mathematics students, with their bottom third of test scores rising in the national distribution.

We identify causal impacts by borrowing an instrumental variable approach commonly used in the electrification literature. We employ an instrument that exploits the variation in cost of expanding this internet infrastructure associated with distance from the existing infrastructure. Our data provides us with a number of important controls for education attainment and conditional on our controls and fixed effects, we rely on a conditional independence assumption for our instrument to deliver identification. This approach allows us to focus on the variation in the supply-side of internet provision while removing the other sources of variation associated with the non-randomness of how separate municipalities may have been prioritized for connection.

We also present additional evidence that precludes various different channels through which our observed impact may be occurring and argue that this impact arises from increases in the marginal productivity of the educational investments. In other words, we believe that the internet is making it easier for this impacted subset of students to generate human capital. We show that this increase in internet access does not change students' propensity to engage in the labor force while studying. This is important because it rules out cases where students may be dropping out of school or entering school at higher rates as a result of this internet expansion, which would potentially lead our results to be driven by a change in the sample of test takers instead of actual learning on behalf of the test takers. We also find no impact of this internet expansion on self-reported family income, which rules out any income effects on the household's education decision problem. These estimated effects are consistent with the internet acting as an educational input that directly increases the marginal productivity of educational investments for these students.

This has important implications for policies aimed at expanding internet access into areas that are typically viewed as unprofitable. This is especially relevant for low- and middle-income countries, which share a common pattern of more modern and connected urban centers and then poorer and less developed rural areas. Our estimates show that closing the digital gap may also be an effective tool for reducing the achievement gap within schools. It seems that these policies do not just provide opportunities through the local economy, as has been documented in previous studies on the labor market effects, but also through increased educational outcomes. We also show that these internet access increases coincide with various other increases that would be consistent with a more modernizing wave affecting these areas, such as increased ownership of various home appliances. We include these changes in durable ownership as additional controls in our specification and find very similar results.

The rest of the paper is structured as follows. Section 2 presents the relevant literature, including various interventions that are focused on previous information and communications technology (ICT) interventions, papers studying internet rollouts on the labor market, and papers that have used a similar identification strategy to study the impacts of electrification. Section 3 gives a brief history of the Vive Digital Phase 1 policy and the relevant facts for our analysis. Section 4 details the sources and construction of the data that we use for our analysis. Section 5 outlines our identification strategy and presents our estimating equations. Section 6 presents our estimates of the impact of increasing internet access on test scores and on other variables that are important for the interpretation of our results. Section 7 concludes.

## 2 Literature

Technology-based learning interventions have enormous potential to improve education outcomes, especially in low- and middle-income countries. Many researchers have built up a significant body of evidence showing this potential, primarily based upon randomized controlled trials in a variety on countries and contexts. For example, Banerjee, Cole, Duflo, and Linden (2007) and Linden (2008) both studied a randomized intervention in India allowing students to spend two hours a week using computer-based learning software, which found significant positive impacts on mathematics test scores. Similar studies, such as Mo et al. (2013) in China and Carrillo, Onofa, and Ponce (2011) in Ecuador found positive effects of computer-based learning aids on test scores. It is worth noting however that some other papers studying similar programs have failed to find significant effects on test score outcomes, such as Goolsbee and Guryan (2006) in USA, Barrera-Osorio and Linden (2009) in Colombia, and Cristia, Ibarrarán, Cueto, Santiago, and Severín (2017) in Peru. These studies all show that digitally based instructional technologies can be important inputs into the education production function.

One of the attractive implications of using internet access as an education-enhancing technology is the ability for students and teachers to access information and instruction targeted to students' current level of knowledge. This is especially important for students who have fallen behind and may lack the prerequisite knowledge to make sense of grade-appropriate lesson plans. One randomized trial that is particularly relevant for our estimates is from Muralidharan, Singh, and Ganimian (2019) in India, which studied a cohort of students where this is the case. They study the impact of exposure to a software (Mindspark) that personalizes the instruction to students based upon their baseline knowledge. They found that once this instruction is targeted and level-appropriate, then these students showed significant increases in their independently administered test scores across their entire distribution of starting proficiency. These students were behind however, and the only detectable increase in the annual school math exams occurred for the students who scored highly in the Mindspark software, and were therefore more likely to be receiving grade-appropriate content.[75] Another relevant example comes from Carrillo et al. (2011), which found positive test score impacts that were disproportionately concentrated at the top of their test score distribution. While these impacts are concentrated at the opposite end of the test score distribution than our estimated impacts, it is another important example of heterogenous effects of these interventions across the distribution.

Whereas the previous papers are studying small-scale randomized interventions that provide ICT technology for education, this paper focuses on a country-wide rollout aimed at increasing internet access. Previous research has found that increased internet access can have significant impacts on labor markets, however these studies have focused on employment rates and the labor force. Some examples include Forman et al. (2012) on wage growth in USA, Akerman et al. (2015) on labor markets in Norway, and Hjort and Poulsen (2019) on labor markets in various Sub-Saharan African economies. All of these papers have found significant impacts on labor market outcomes, but to the best of our knowledge, our paper is one of the first to evaluate one of these rollout's impact on test score distributions.

This paper borrows its instrumental variable strategy from the electrification literature. These papers have used the fact that electricity infrastructure is expensive and that certain geographic features provide variation in the cost of expanding electricity provision. This has led these studies to construct instruments for electricity access from geographic features that make this provision more expensive. One example is the average gradient of the topography within a region, which would imply rugged terrain and make tower construction more expensive, has been used by both Dinkleman (2011) in South Africa and Grogan and Sadanand (2013) in Nicaragua. Other papers have used distance to either generation capacity, especially hydroelectric generation, or the existing electricity grid as instruments for access as the longer the distance, the more expensive it is to expand the grid. Some examples of papers using these strate-

[75] This is discussed more fully in relation to our estimates in the results section.

gies are Grogan (2016) in Colombia, van de Walle, Ravallion, Mendiratta, and Koolwal (2017) in India, Grogan (2018) in Guatemala, Lipscomb, Mobarak, and Barham (2013) in Brazil, and Squires (2015b) in Honduras. This is the first paper to take such instruments from the electrification literature and apply them to study the impacts of increasing internet access.

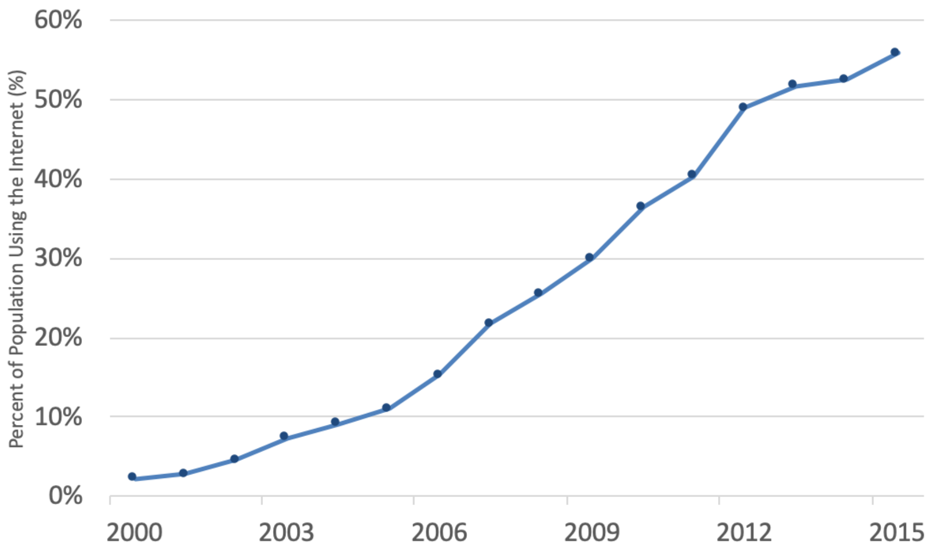
One cannot study the impact of increased internet access on education outcomes in a vacuum. Critically, there exist opportunity costs for education investments, and time spent making these investments comes at the cost of other opportunities, such as either working or leisure. This becomes especially relevant because there is good reason to expect that increasing internet access would not only increase the marginal productivity of educational investments, but would also increase the productivity of the other activities that compete for a child's time. One example is laid out in the papers mentioned above, where increased internet access may increase the wages that can be earned in the labor market. An example of substitution towards leisure comes from Malamud and Pop-Eleches (2011) in Romania, where distributing home computers to students found negative impacts on test scores, which were attenuated in the cases with stricter parental monitoring of use.

Many models have formalized household decision-making on education investments. Two examples come from both Shah and Steinberg (2017) and Bau, Rotemberg, Shah, and Steinberg (2020) in India. In these cases, a household will apportion a child's time to investing in their education and other activities such that their marginal productivities are equalized. This household decision implies that despite its likely positive impact on the marginal productivity of investing in education, this may actually reduce a household's investment in education if the internet expansion's impact on the marginal productivity of labor or leisure is sufficiently large. The relative strength of these changes in marginal products will determine both the magnitude and direction of students' responses. It is also possible that the relative magnitudes of these effects will differ by context. The body of work in the electrification literature provides an example of this, where both the magnitudes and signs of increased electricity access on education outcomes have varied. Jimenez (2017) provides a meta-analysis of these studies noting this dispersion in results and Squires (2015a) notes an example in Honduras where electrification actually reduced educational attainment, with the primary channel being substitution out of school towards labor.

## 3 Vive Digital Phase 1

Colombia made large gains in internet connectivity between 2000 and 2009 (See Figure 3.1 below). However, many gains during this period were concentrated in specific areas of the country, especially in large, urban areas. By 2010, despite these large gains, much of the country remained unconnected to Colombia's internet infrastructure. Less than 20% of Colombia's approximately 1100 municipalities were connected to the fiber optic network. Many barriers previously prevented this expansion of the internet from these concentrated areas, including the high costs of expanding the internet infrastructure. These high costs are also exacerbated by the mountainous geography of Colombia and the distance between urban centers. Colombia's limited financial capacity also precluded the State itself from making these investments (Vega, 2013, p.112).

*Figure 3.1:* Percent of Population in Colombia Using the Internet: 2000 – 2015

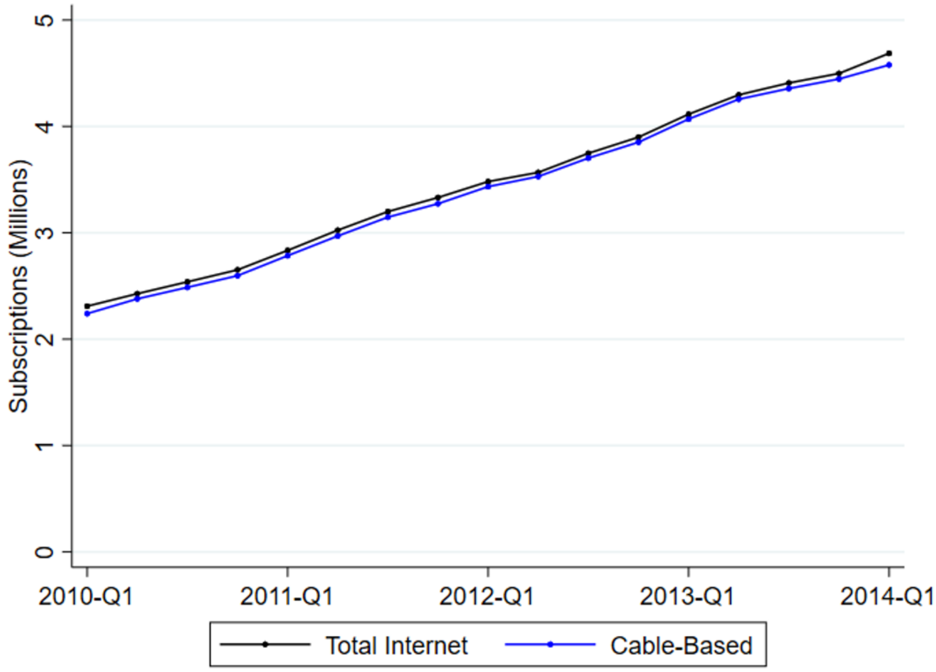



SOURCE: World Bank Development Indicators. Series: "Individuals using the Internet (% of population)". Accessed 29 May 2020.

We study Colombia's public policy push called "Vive Digital", which increased internet access into previously underserved areas throughout the country. Phase 1 of this program occurred between 2010 and 2014. In 2009, Colombia passed a series of laws that mandated expanding internet access as a national priority. A new government was elected in 2010 and with it came a new minister of ICT, Diego Molano Vega, who oversaw the creation and implementation of this plan.

Vive Digital Phase 1 had three main objectives focused on increasing internet access and usage. These were to (i) triple the number of municipalities connected to internet infrastructure, (ii) connect 50% of Colombia's microenterprises and small and medium enterprises (SMEs) and 50% of homes, and (iii) quadruple the total number of internet connections to 8.8 million (Vega, 2013, p.112). One of the four major strategies for achieving these goals included expanding the internet infrastructure to narrow the rural-urban divide in internet access. During this period, the primary mode of internet provision was through cable-based technologies. These included DSL, cable, and fiber optic technologies, which accounted for approximately 97% of connections.[76] Therefore, this infrastructure expansion involved expanding these cable-based networks into municipalities which remained previously unconnected, and indeed a large public focus remained on laying new fiber optic cables.

*Figure 3.2:* Total Number of Quarterly Subscriptions: Q1 2010 – Q1 2014




SOURCE: Data from Colombia's Communications Regulatory Commission (CRC). These numbers represent a stock of subscriptions in each period, not a flow.

This program was very successful and resulted in a large increase in total connections and the number of municipalities connected. Figure 3.2 below shows the quarterly increase in total connections within Colombia over the Vive Digital Phase 1 period. Indeed, by 2012 alone, the Colombian government had been successful in laying over 15,000km of new fiber-optic cable and installed multiple additional submarine internet cables. According to a 2013 report by the Ministry of ICT, over 250 additional

[76] Source: Data from Colombia's Communications Regulatory Commission (CRC). Note that this is the same data presented in Figure 2.

municipalities became connected to fiber optic cables between 2010 and 2012, and the government noted that it was on track to connect at least another 225 by 2014 (Vega, 2013, p.112). This increase constitutes a large, supply-side driven increase in internet access as a result of this government policy.

# 4 Data

Answering our research question requires very specific data. Studying school-level test distributions requires more data than is typical of a standard randomized trial as we must observe enough test-scores such that we have both (i) enough data within each school to observe a distribution, and (ii) enough schools to be able to compare the changes between them. We also require that the test scores are standardized and comparable both between schools and across time. Detailed information about the students who are taking these tests, which schools they attend, and where these schools can be located are also necessary for constructing control and geographic variables. Finally, we also require a measure of how many students have access to the internet and which municipalities have the necessary infrastructure to make that possible.

We bring together various different data sources to construct a school-level dataset that includes all of these elements. This section will cover the three major sources that we use in turn. Our test scores are based upon Colombia's SB 11 test score data, which is where we draw our outcome variables (e.g. test scores), information about the students, and measure of internet penetration. The second source includes data on locations for individual schools drawn from multiple sources within the Colombia government. The final source comes from Colombia's ICT regulator (The CRC) and includes data on which municipalities were connected to the internet. We focus on the years 2009 and 2015 as these years encapsulate the Vive Digital Phase 1 policy period and 2009 is the first year where there exists full data on our internet measure.

## 4.1 Saber 11 Testing Data

We use Colombia's Saber 11 (SB 11) standardized test data to measure academic achievement. This test is a national standardized test given twice a year for students finishing high school and is administered by Colombia's Institute for Education Evaluation (ICFES). It is a secondary school exit exam and is required for entry into higher education, although it is possible to take the exam at a later date after graduation. The exam covers many different subjects and we focus on two subjects that are consistent through the years and common in the literature: mathematics and language.

One of the advantages of this testing data is the very high test-taking rate, with over 90% of eligible students sitting the exam. Riehl, Saavedra, and Urquiola (2016) also note that the government uses these test scores to evaluate high schools, so the government tries to encourage this test taking rate to be as high as possible.[77] This is an important detail because it minimizes a common concern with these standardized tests, which is the selection problem of who chooses to write the test. This is especially salient for our paper as any such selection problem is likely to disproportionately censor a specific side of the distribution, so an observed increase in the bottom of a test score distribution may actually be generated from an exit of the lowest-scoring students from the distribution. Therefore, the very high writing rate of this exam alleviates many of these selection concerns.

The SB11 test score data includes a full census of every test-taker. Before taking the test, students complete a detailed socioeconomic survey about their individual and household characteristics. There is some variation in the set of questions asked, but they are also consistent enough to construct a set of comparable variables over years. As a result, this data contains approximately 560,000 test scores for each year, and

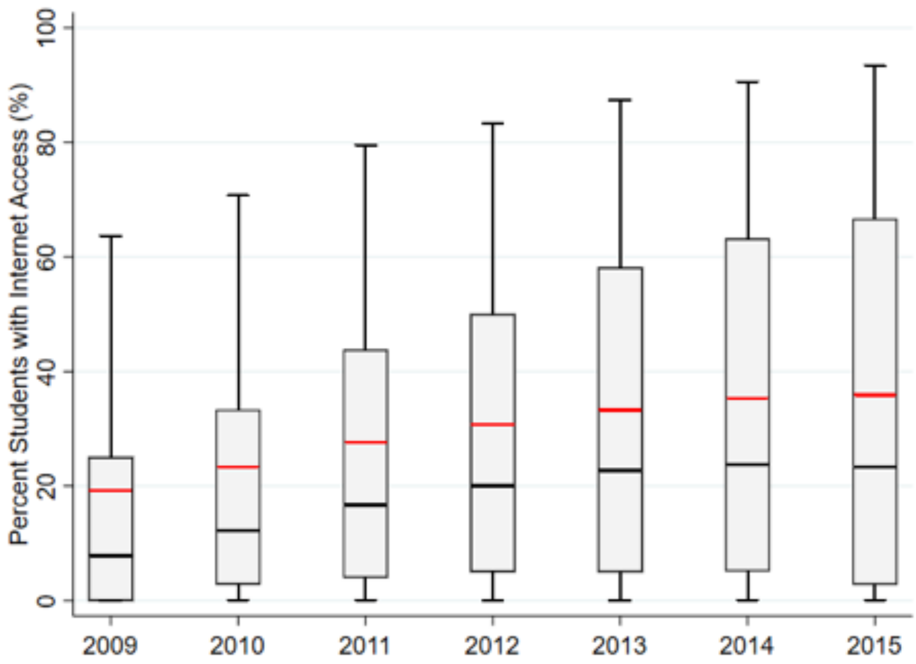
[77] This is noted by Riehl et al. (2016) as they document this from personal communications with ICFES and Angrist, Bettinger, and Kremer (2006, p.110) as they used the SB11 as well.

includes detailed information on each student. Each test score is standardized to the number of standard deviations from the mean score for each year. This data provides our primary outcomes variables, including test scores, family incomes and student employment, socio-economic controls, and our primary treatment variable, specifically if the student has access to the internet at home.

We use this test score data to construct annual data at the school level, which is our primary unit of analysis. Each student in the data is assigned to a specific education institution by a code assigned by Colombia's National Administrative Department of Statistics (DANE), hereafter referred to as a DANE code. We use this code to construct our data at the DANE code level and to assign each code to individual municipalities and respective province (departmentos). We exclude the DANE codes for institutions that are within Colombia's five largest cities.

We calculate school-level test score distribution statistics, including the mean values, scores at different percentiles, and the differences between the 90th and 10th percentile scores. We also construct school-level average measure for student characteristics. For example, individual indicators for if a student lives in an urban or a rural area are converted into the percentage of students who live in a rural area for a given year. Finally, we construct our final treatment variable, which is the percent of students in any given year that have access to the internet at home. Figure 3 below shows the distribution of these values across our sample of schools for each year. It shows the large increases in the distribution of these values over the policy period.

*Figure 3.3:* Distribution of School-level Values for Percent of Students with Internet Access (2009 – 2015)




Source: SB11 Data for 2009 - 2015. The bars show the 25th to 75th percentiles. The whiskers cover the 10th and 90th percentiles. Outliers are excluded. The red (top) line within the bar represent the mean value for the sample of schools. The black (bottom) line within the bar represent the median value.

## 4.2 School Location Data

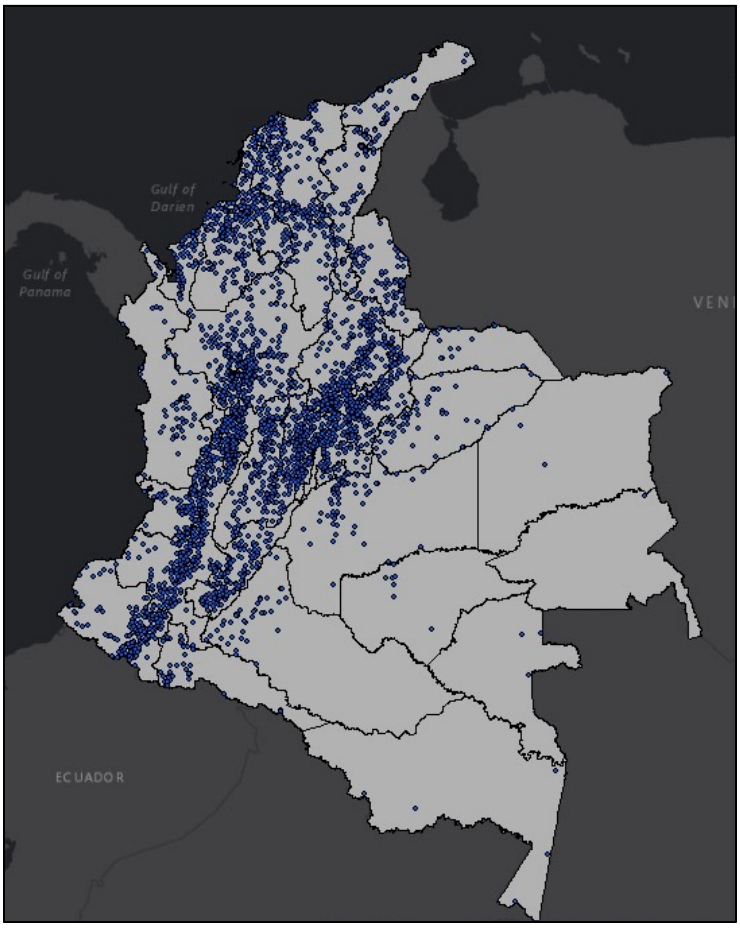
We use data from both Colombia's Ministry of Education and DANE to locate the schools in our data. Both of these sources use DANE codes for educational institutions as well, so we match between the SB11 data and location data based upon these codes. The Ministry of Education provides detailed latitude and longitude data for a subset of schools in Colombia. The data from DANE provides approximate locations for these DANE codes. The most common is a town code, which when combined with shape files from DANE, allows us to construct a centroid for the towns where the schools are located. In a few extra cases for rural schools, the data allows us to locate the school in the centroid of its vereda, which is a subset of the individual municipality.[78]

We begin by matching schools based upon the exact location data using the Ministry of Education Data. DANE codes describe an individual educational institution, so in some cases there are multiple locations associated with each code. An example of where this could occur would be if two affiliated schools were in nearby towns. In this case, we took their average coordinates and excluded any code where any of the original coordinates were more than 10km from the calculated average. We then matched the remaining schools with the DANE data, using the centroids of the towns or if the towns were unavailable, the vereadas where possible. This produced a final sample of 5,205 located schools with test score data out of the total 5,744 school codes that we observe in both 2009 and 2015, which constitutes a 91% success rate.[79]Figure 3.4 below shows this full sample of located schools.

[78]For reference, there are approximately 33,425 unique veredas in Colombia according to GIS data from DANE. They constitute a very granular spatial measurement.

[79]We match 3,445 DANE codes using the Ministry of Education data. We then match a further 1,691 using the town codes from the DANE data. Finally, we match another 69 schools based upon their veredas. This gives 5,205 in total.

*Figure 3.4:* Complete set of Located Schools in Data




Source: Colombian Ministry of Education and DANE data. All locations are plotted in ArcMap using standard base layers and shape files from DANE. The boundaries are for individual provinces. All GIS work is the authors'.

## 4.3 CRC Subscription Data

Colombia's Communications Regulatory Commission (CRC) also provides data on quarterly internet subscriptions. This dataset runs from the first quarter of 2010 (January-March) until the first quarter of 2014. It includes the total number of internet subscriptions at the municipality level, which can be further broken down by type (e.g. satellite or DSL) and provider. In our data, we make the distinction between cable-based" and "non-cable-based" internet provision. The former requires a physical cable connection and therefore requires necessary infrastructure to be extended in a network, such as a fiber optic connection for example. The "non-cable-based" systems do not necessarily require this, such as a satellite connection. This dataset is the primary source for the construction of our instrument. It allows us to see the number of active internet connections within each municipality, and therefore, the number of municipalities that meet any given threshold for the number of active internet connections.

# 5 Estimation Strategy

## 5.1 Identification Strategy

One challenge of studying a nationwide internet rollout is that the expansion of the network infrastructure is very unlikely to be randomly determined. There could be many reasons why certain municipalities may have been prioritized for connection as part of Vive Digital and if these reasons are correlated with education outcomes, then this would introduce bias into our estimates. Some examples may be changes in political influence of certain areas or a priority of the government to connect schools to internet connections. Vive Digital had a goal of connecting as many of Colombia's municipalities as possible over this period and so there were likely multiple factors driving the order in which municipalities became connected. Therefore, there will be multiples sources of the variation of the internet expansion over this period, including potentially confounding variation discussed above, but also variation in the cost of the expansion given the original existing infrastructure.

The places that have access to the internet may differ from those that do not in many respects that are relevant for education outcomes, and our empirical strategy addresses this issue in two ways. First, we first-difference the variables at the school level.

With this, we remove any time-invariant component at the school and municipal level. The first differences control for any school-specific geographic components and any common shocks, such as the introduction of additional submarine internet cables in Colombia. Second, we use the cost of the internet expansion as our source of our variation to identify the impact of increasing internet access on test scores. We use the fact that the vast majority of internet subscriptions ($> 97\%$) in Colombia are "cable-based" subscriptions and that the cost of expanding this infrastructure is costly and critically, that this cost increases with distance. This is especially the case in this context given Colombia's mountainous topography. We therefore use the distance of an individual school from the stock of internet infrastructure at the start of our period as an instrument for the change in internet access over this period. The logic of this instrument is the same as the many papers studying electrification in that we are focusing on part of the expansion that is driven by geographic factors and cost-minimization on the supply side.

The validity of this instrument relies on two main assumptions. The first is that this distance is indeed related to the cost of expanding this infrastructure and therefore does affect changes in internet access (relevance).The second is that while this distance does use the cost variation in expanding internet access, it only affects the change in test scores though this change in internet access. Our primary identification assumption becomes that, conditional on our socio-economic controls, within-province variation in distance to initial network infrastructure only affects test score changes through its impact on changes in internet access. In the context of a rapid expansion of internet coverage and infrastructure promoted by the national government, as it is the case of Colombia in our study period, the variation in this expansion is likely driven by two factors. The first is cost-minimization on the supply side, and this is the variation that our instrument uses to identify the impact of the change in internet access on test scores. The second factor is general policy concerns, which may drive the government to prioritize certain areas over another for this expansion beyond simple cost-minimization. This second source of variation is the variation in the changes in internet access that we aim to remove through the use of our instrument.

It is worth reiterating that we are instrumenting for the first-difference of the change in the percent of test takers that have internet at home, not the level. This is the common application of these geographic instruments in the electrification literature. van de Walle et al. (2017) lay out a particularly good discussion of these instruments and lay out the central argument very well when they write: "judgments on the plausibility of the identification strategy must also depend on what other control variables are used, given that the estimator is making a conditional independence assumptio" (van de Walle et al., 2017, p.397). The quality of our data has allowed us to include many important determinants of education investment choices and the likely demand for internet. This will critically control for the initial population characteristics and any changes in the makeup of the people who live in these areas that could drive the trajectories of internet take-up and educational outcomes. Given our controls and the short time frame, we believe that our identification assumptions are defensible.

## 5.2 Constructing the Distance Instrument

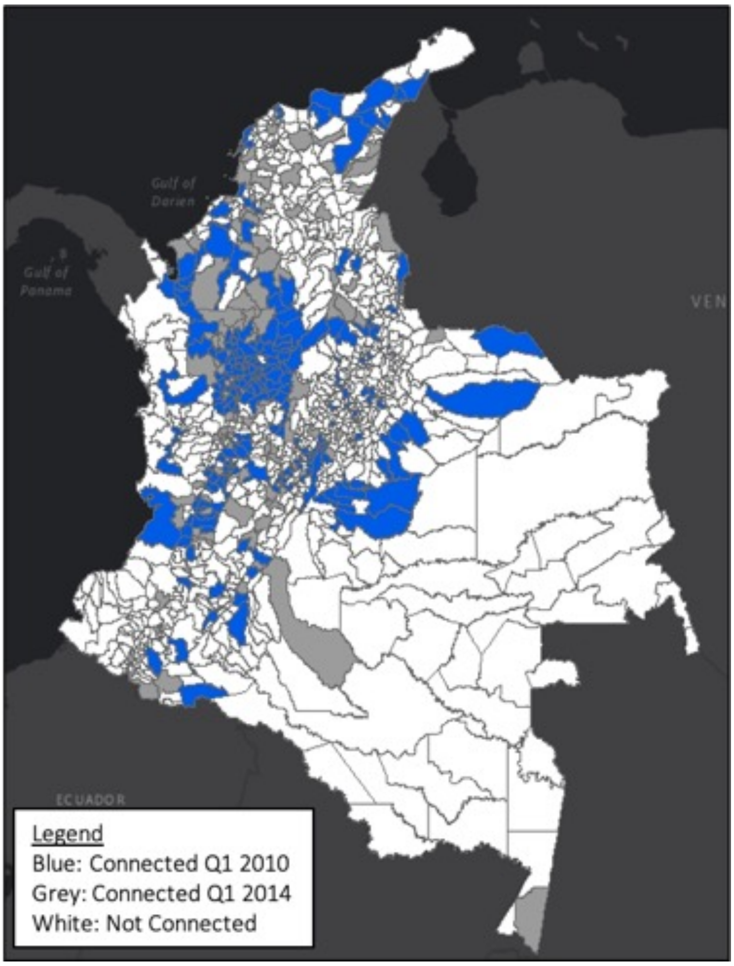
The first step in constructing our instrument is to determine which areas of the country were connected to the internet infrastructure at the beginning of 2010. We use the data from the CRC to determine how many cable-based internet connections existed within each municipality for each quarter. We first determine how many subscriptions are active in each municipality. We do not want to dismiss a municipality as unconnected just because it is sparsely populated however, so we convert these number of subscriptions into the number of subscriptions as a percentage of the total projected population for 2010.[80]

We categorize each municipality as connected if the number of "cable-based" subscriptions is greater than or equal to 1% of the projected population. This is preferable to simply marking a municipality as connected if it records any cable-based connections for two reasons. The first is that many municipalities record having less than 12 connections and there exists large bunching of municipalities at a single connection. If these are marked as connected, we end up with a very large number of connected municipalities. Additionally, when we compare the municipalities that are classified as connected in Q1 2010 and Q1 2014 using the rule of marking them as connected if they have at least a single connection, there are 49 municipalities that change from connected to unconnected over time, which is an undesirable property of this rule. We believe that is likely due to some measurement error. Using the 1% rule leaves only one municipality that changes from connected to unconnected, which makes it a more plausible indicator for if a municipality was actually connected to the internet infrastructure.

We use this measure of connectedness to construct a map of the municipalities that were plausibly connected to Colombia's cable-based internet infrastructure in Q1 2010. Figure 5 below shows the set of these municipalities. We also identify the municipalities that change from unconnected in Q1 2010 to connected by Q1 2014. We calculate the shortest distance to the nearest "connected" municipality in Figure 3.5 for each school in in Figure 3.4. This represents our measure of the linear distance of each located school to the nearest municipality connected to the starting cable-based internet infrastructure.

[80]This uses population projections from DANE based upon Colombia's 2005 national census. It is available at: DANE website (Accessed 02 March 2020)

*Figure 3.5:* Set of Municipalities Marked as "Connected" in Q1 2010 and Q1 2014




NOTES:The boundaries are for individual municipalities. Municipalities in blue were marked as connected by our 1% definition in Q1 2010. Municipalities in grey were not marked as connected in 2010, but became marked as connected by Q1 2014. Municipalities in white are not marked as connected in either period. Sources: Colombian CRC. All GIS work is the authors'.

## 5.3 Estimating Equation

We estimate the impact that changes in the percent of test-takers in a school that have internet access have on the change in test scores using 2SLS on our first-differenced data. Once again, we instrument for this change in internet access with the distance of each school to the municipalities that we marked as connected for Q1 2010. The estimating equations are:

$$\Delta y_{ip} = \beta_0 + \beta_1 \Delta Inet_{ip} + \theta_{ip} + \lambda_p + \varepsilon_{ip} \tag{36}$$

With the first stage given by:

$$\Delta Inet_{ip} = \delta_0 + \delta_1 Z_{ip} + \vartheta \Delta X_{ip} + \varphi_{ip} + \lambda_p + \mu_{ip} \tag{37}$$

In the above equations, an individual school is denoted with the subscript $i$, the province that the school is in is denoted by $p$, and first-differenced variables are preceded by $\Delta$. The first-differences are constructed using the difference between the 2015 and 2009 values. For variables that are not first-differenced, they are either time-invariant or use the starting level value for 2009.

There are a few outcome variables that we will examine, each denoted by $\Delta y_{ip}$. An example here would be the change in the mean mathematics test score for a school. Our primary treatment variable is the change in a school's percent of test takers with access to the internet at home, denoted by $\Delta Inet_{ip}$.This is instrumented by our measure of the shortest distance to the nearest 2010 "connected" municipality, denoted by $Z_{ip}$.We also include a series of demand-side controls that are important indicators of a student's academic performance and also may affect the probability that a family invests in internet access. These include if students live in a rural or urban area, their birth year, the child's sex, and the education level of each of the parents. We include the starting levels of these controls, denoted by $X_{ip}$,and their first differences, denoted by $\Delta_{ip}$.Provincial fixed effects are denoted by $\lambda_p$. Finally, idiosyncratic error terms is denoted by $\varepsilon_{ip}$ and $\mu_{ip}$ for each of the estimating equations.

# 6 Results

We divide our results into three sections. The first section presents our estimates of the impact of increasing home internet access on test scores. The second section confirms that we cannot detect any impact of increased internet access on family incomes or the students' propensity to work, which are two primary alternate channels by which the internet may affect education choices. Finally, we present evidence that increased internet access has a positive effect on the purchase of other durable goods, which remains a puzzling outcome, and discuss its implications for our results. The standard errors are clustered by province in all of this section's regression tables.

The change in the percent of test takers with internet access at home is instrumented by the shortest distance to our set of connected municipalities shown in Figure 3.5 above. The first assumption that we verify is the relevance assumption. Table 3.1 below presents the estimates analogous to Equation 2 in the previous section. It also includes the relevant F-statistics for our instrument. Once we include provincial fixed effects, our reported F-statistics more than satisfy any general rules of thumb regarding weak instruments. Table 3.1 reports multiple specifications, including various combinations of controls, with column 5 representing our full and preferred regression estimates and fully corresponding to Equation 2. The Column 5 estimate implies that for every 100km a school is away from the boundary of its nearest connected municipality, its increase in the percentage of students with internet access decreased by 12.7% once you control for the province and demand-side variables. It is worth noting that, given Colombia's largely mountainous topography, we would expect even modest distances to pose a significant cost barrier to expanding existing infrastructure.

*Table 3.1:* First Stage: Regression of Instrument and Controls on Changes in Internet Access

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Distance to Connected | -0.062** | -0.203*** | -0.174*** | -0.152*** | -0.127*** |
| | (0.026) | (0.057) | (0.046) | (0.030) | (0.024) |
| Observations | 5205 | 5205 | 5198 | 5200 | 5198 |
| Province FE | no | yes | yes | yes | yes |
| FD Controls | no | no | yes | no | yes |
| Level Controls | no | no | no | yes | yes |
| Fstat | 5.548 | 12.794 | 14.160 | 25.171 | 29.164 |
| $R^2$ | 0.024 | 0.121 | 0.236 | 0.305 | 0.396 |

Standard Errors Clustered by Province in Parentheses

$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

## 6.1 Primary Results: Test Score Outcomes

Our preferred estimates do not show any detectable impact of increasing internet access on mean test scores for either language or mathematics. Table 3.2 below presents our estimates of impact of increasing the percent of the test takers with internet access on schools' mean test scores. Both the estimates for the impacts on mathematics and language scores start off as positive and statistically significant, but cease to be statistically significant once all of our controls are added. Although we cannot statistically distinguish the impact on mathematics test scores from zero, it remains large and positive. On the other hand, the impact on language scores becomes very close to zero and insignificant.

*Table 3.2:* Impacts on Mean Mathematics and Language Test Scores

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | A.Math Score | | |
| FD: Pct Internet | 0.00555*** | 0.00543*** | 0.00529*** | 0.00314 |
| | (0.00153) | (0.00162) | (0.00188) | (0.00200) |
| | | B.Language Score | | |
| FD: Pct Internet | 0.00581*** | 0.00533*** | 0.00350** | 0.00070 |
| | (0.00151) | (0.00194) | (0.00167) | (0.00294) |
| Observations | 5205 | 5198 | 5200 | 5198 |
| Province FE | yes | yes | yes | yes |
| FD Controls | no | yes | no | yes |
| Level Controls | no | no | yes | yes |

Standard errors clustered by Province in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

One immediate question from the results presented in Table 3.2 is if the lack of significant results is because the increase in internet access had no impact on test scores. This would generate such a result at the mean if there existed a homogenous and negligible impact across the entire test score distribution. However, it might also be that the increase in internet access only had a significant impact for a subset of the test score distribution, which may be significant for that group but not sufficiently strong to generate an effect at the mean. An example of where this may occur would be if the internet was only a useful study tool for those students who have fallen behind, which would manifest in only an impact in the bottom of the distribution. We examine the difference between the $90^{th}$ and $10^{th}$ percentile test scores to distinguish between these two cases. In the first case, where there is a homogenous negligible effect across the whole test score distribution, we would expect the variance of scores to remain nearly constant. In the second case, on the other hand, we would expect to see a change in the variance of the distribution if there exists heterogeneous impacts along the distribution. This is assuming that there is some pattern to this heterogeneity consistent with most decision-making models. If there is heterogeneity that is simply noise, then this may not necessarily be the case. If there is some monotonic or at least more smooth profile of the impact along the test score distribution, then this would be the case however.

Our estimates show that there is indeed a variance reduction in the mathematics test score distribution for schools that experienced a larger increase in internet access. Table 3.3 below presents our estimates of the impact of increased internet access on the 90-10 percentile spreads for both mathematics and language test scores. Our estimated impact on this spread implies that for every 10% increase in the percent of test-takers that have access to the internet, the difference between the $90^{th}$ and $10^{th}$ percentile test score in a school decreases by 0.057 standard deviations. This implies that of the two cases outlined above, the second case is likely the better fit for mathematics test scores. Combining a positive estimated impact on mean test scores and a tightening of the distribution implies that the gains were larger for those at the bottom of the test score distribution.

*Table 3.3:* Impacts on Mean Mathematics and Language Test Scores

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | A.Math Score | | | |
| FD: Pct Internet | -0.00318 | -0.00422** | -0.00458** | -0.00566** |
| | (0.00201) | (0.00213) | (0.00232) | (0.00245) |
| | B.Language Score | | | |
| FD: Pct Internet | 0.00678*** | 0.00686*** | 0.00343 | 0.00353 |
| | (0.00233) | (0.00255) | (0.00288) | (0.00306) |
| Observations | 5205 | 5198 | 5200 | 5198 |
| Prov FE | yes | yes | yes | yes |
| FD Controls | no | yes | no | yes |
| Level Controls | no | no | yes | yes |

Standard errors clustered by province in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

There was no detectable impact on the 90-10 percentile spread for language test scores. It remains positive and loses significance with the addition of more controls, similarly to the impact on mean math scores present above. Table 5 and Figure 7 below will both present information on what is happening to the distribution in this case. Despite this positive spread overall, when viewed in concert with the negligible impacts on mean language scores, we conclude that no real story of impact is emerging for language test scores. It appears that the case of language test scores is much more similar to the first of the two cases outlined above and that increased internet access appears to not have impacted any segment of the distribution much at all.

Finally, we show estimated impacts along different points in the mathematics test score distribution that explain this variance reduction. This is important because a variance reduction may indicate either a relative gain at the bottom of the test score distribution or a relative decline at the top of the distribution. We do so by taking different percentile scores for each school and year and using the changes in those percentile scores as our outcome variable. This provides estimated impacts of increased internet access on different points in the test score distribution. Table 3.4 and Table 3.5 below present the estimated impacts of internet access on specific mathematics and language test score percentiles respectively. Figure 6 and Figure 7 visualizes the analogous preferred estimates from Table 3.4 and Table 3.5. We focus on the 10-30-50-70-90 percentiles, but our results are qualitatively similar if we use different percentiles, such as 20-40-60-80.

Our results do show statistically significant and large, positive impacts of increased internet access on the $10^{th}$ and $30^{th}$ percentile scores for mathematics. These results imply that for every 10% increase in the proportion of test-takers with internet access, the mathematics test scores increase by 0.056 and 0.059 standard deviations for the $10^{th}$ and $30^{th}$ percentiles respectively. This shows is a large impact in the bottom third of the test score distribution with a decline in impact further towards the top of the distribution and with a quite small and insignificant impact by the 70th percentile. The heterogeneous impacts across the test score distribution explains the positive, but insignificant impact on mean mathematics scores. Reassuringly, it appears that increasing internet access had large, positive effects on the bottom of the test score distribution, and while having little to no effect, did not have any negative consequences for test scores at the top of the distribution.

Our estimates for the impact on language test scores tell a different story. There remains no detectable impact on any specific percentile of the language test score distribution. When combined with our previous results that found no detectable impact on the mean or 90-10 percentile spread of the language test score distribution, this strengthens our interpretation that increasing internet access has simply had a negligible effect on language test scores. There is a little bit of a shape to the estimates of the individual percentile scores, but they all remain close to zero and both statistically and economically insignificant.

*Table 3.4:* Impacts on $10^{th}$, $30^{th}$, $50^{th}$, $70^{th}$, and $90^{th}$ Percentile Mathematics Test Scores

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | p10 | p30 | p50 | p70 | p90 | p10 | p30 |
| FD: Pct Internet | 0.00700*** | 0.00756*** | 0.00575*** | 0.00463*** | 0.00382** | 0.00562** | 0.00589** |
| | (0.00208) | (0.00178) | (0.00150) | (0.00141) | (0.00176) | (0.00244) | (0.00245) |
| Observations | 5205 | 5205 | 5205 | 5205 | 5205 | 5198 | 5198 |
| Pctile | 10 | 30 | 50 | 70 | 90 | 10 | 30 |
| ProvFE | Y | Y | Y | Y | Y | Y | Y |
| FDControls | N | N | N | N | N | Y | Y |
| LvlControls | N | N | N | N | N | Y | Y |

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

*Table 3.5:* Impacts on $10^{th}$, $30^{th}$, $50^{th}$, $70^{th}$, and $90^{th}$ Percentile Language Test Scores

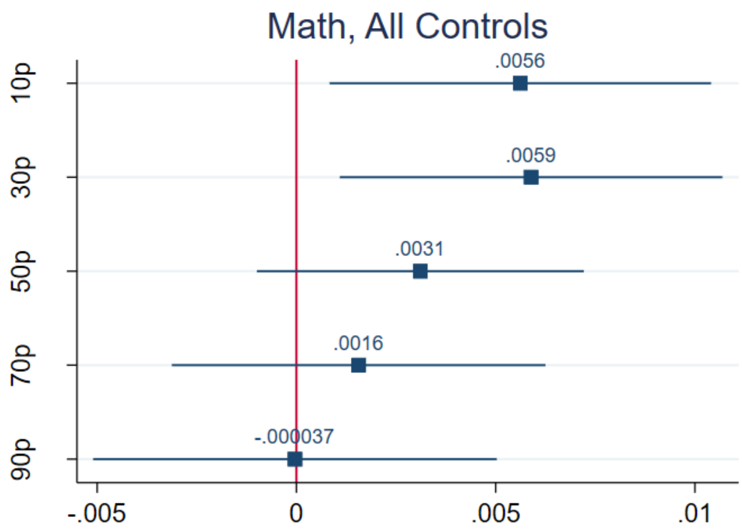
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| | p10 | p30 | p50 | p70 | p90 | p10 | p30 |
| FD: Pct Internet | 0.00141 | 0.00391*** | 0.00636*** | 0.00759*** | 0.00818*** | -0.00255 | -0.00124 |
| | (0.00140) | (0.00144) | (0.00128) | (0.00186) | (0.00288) | (0.00280) | (0.00265) |
| Observations | 5205 | 5205 | 5205 | 5205 | 5205 | 5198 | 5198 |
| Pctile | 10 | 30 | 50 | 70 | 90 | 10 | 30 |
| ProvFE | Y | Y | Y | Y | Y | Y | Y |
| FDControls | N | N | N | N | N | Y | Y |
| LvlControls | N | N | N | N | N | Y | Y |

Standard errors in parentheses

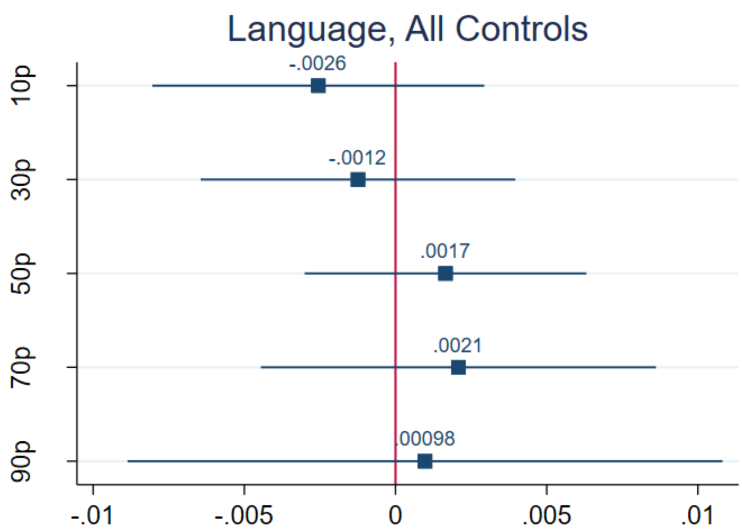
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

*Figure 3.6:* Preferred Coefficient Estimates for $10^{th}$, $30^{th}$, $50^{th}$, $70^{th}$, and $90^{th}$

*a. Mathematics Test Scores*




*b. Language Test Scores*




NOTES:These results are analogous to Columns 6-10 in Table 3.4 below. Coefficients and 95% confidence intervals are presented. The y-axis shows the corresponding percentile for each estimate.

Increasing the percent of students with access to the internet had a small, but significant impact on lower performing students over this period. Our estimates, which suggest a 10% increase in internet access increases test scores by 0.06 SD in the bottom third if the mathematics test score distribution, remain small but are also not implausibly so given the estimated impacts of other educational interventions. For example,Angrist, Bettinger, Bloom, King, and Kremer (2002) evaluated a voucher lottery program in Colombia that subsidized the cost of attending private school and found that winners increased their test scores by 0.2 standard deviations. Agüero and Beleche (2013) found that increasing the days of instruction in a school year in Mexico increased test scores by 0.04-0.07 SD per 10 days. (Bellei, 2009) found that moving from half day to full day instruction in Chile increased test scores by 0.05-0.07 SD in mathematics and 0.00-0.12 SD in Language. These are all examples of education interventions in nearby middle-income countries that produced impacts similar in magnitude to what we have observed in the bottom of the mathematics test score distribution.

These estimates imply that closing the digital gap between areas within Colombia will also go some way to close education achievement gaps. These test scores are all standardized, so these estimated increases at the bottom of the mathematics test score distribution for these schools imply that these students are actually increasing in their placement in the national distribution. Therefore, schools that experienced the largest increases in internet access also saw the largest increases in the national ranking of their mathematics test scores, specifically for those students in the bottom of the individual schools' test score distributions. This shows that equalizing internet access through expanding internet infrastructure (or likely through any other means) can lead to catching up of education outcomes relative to the whole country.

Given the high cost of extending physical infrastructure, this is far from a cost-effective method of increasing education outcomes. Our estimates imply that a 10% increase in internet access produces similar estimates to some increases in school days (Mexico) or daily instruction hours (Chile), but it was only on a subset of the test score distribution in our case instead of the mean, further limiting the scope of the benefits. Beyond this comparison, there are many interventions that have improved test scores at a significantly lower cost. For example, a monetary bonus program based upon teacher performance in India studied by Muralidharan and Sundararaman (2011) increased test scores by at least 0.15 SD at the cost of only $4 per year on average.[81] However, we do view this catchup induced by increasing internet access as an additional positive effect associated with the inevitable expansion of internet infrastructure in many middle- and low-income countries.

## 6.2 Alternative Channels: Propensity to Work and Family Incomes

There are two alternative channels that could generate changes in education outcomes beside a direct impact on the marginal return to studying that are important to consider. As discussed in the literature section, a general model of household decision making will show households weighing the return on educational investments against the opportunity costs, including the other potential productive uses of childrens' time. Both of these channels arise from the well-established positive effects of internet access on the labor market. The first is that increases in labor productivity may lead work to displace school and study time if the impact on labor productivity is sufficiently large to dominate the positive effects on studying. The second is the effect on adults' labor productivity and ultimately family income. In the case that a household views investing in a child's education as a normal good, then we would also expect an increase in internet access to generate positive income effects. It is worth noting that both of these forces likely work against each other with respect to their effects on educational investments, but they remain important factors in the decision making process.

[81]The $4 per student number is calculated by de Ree, Muralidharan, Pradhan, and Rogers (2017, p.1031) in their discussion of the paper.

This is particularly important given our estimated impacts concentrating in the lower portion of the mathematics test score distribution. An alternative explanation that could generate these results is if the internet induced more students to drop out and enter the labor force and if these marginal students were disproportionately located in the bottom of these test score distribution. This would lead the exit of the lowest test scores as opposed to increases in scores that were at the bottom of the distribution to potentially drive our estimated increases. It is, however, worth noting that these students need not be concentrated in the bottom of the distribution, and in some cases, students with the highest levels of starting human capital may be more likely to be drawn into the labor force. Where these affected marginal students are located within the test score distribution will depend on the complementarity between the starting level of human capital and the internet's effects on the marginal productivity of education investments and labor. It certainly can be the case however that these marginal students are located in the bottom of the distribution, and it is important to confirm that these alternative channels are not driving our results.

We present estimates in Table 3.6 below with two different outcome variables to test for the presence of these alternative channels. The first is the change in self-reported family income. This variable is collected from the test taker as part of the SB 11 testing data. It asks the student to choose certain bands of their parents' income relative to the minimum wage.[82] This is an imperfect measure, both because it is structured within bands and is self-reported by the student. However, it is the only straightforward measure of family income available to us for this test. The second is the changes in the percent of students that report also working a job. As our sample is for test-takers, this will only capture a change in the percentage of students that both work and attend school and will not capture any students that drop out entirely. However, we think it

[82]For example, one of the choices is if their parents earn between 3 times and 4 times the minimum wage. We convert this into an index with a value of 1 if the parents earn between 1 and 2 times the minimum wage, 2 if the parents earn between 2 and 3 times the minimum wage, etc.

reasonable to expect that if labor force opportunities are increased sufficiently to draw some marginal students to drop out, we would also be very likely to see an increase in the students who also work while studying. While certainly possible, we find it unlikely that we would observe students dropping out to work in the labor market and not also see some rise in students attempting to do both.

We find no evidence that this internet expansion led to a significant increase in family incomes or in the propensity for children to work. Our estimates of the impact of increased internet access on both outcome variables are small and far from significant. This is reassuring and we take this as evidence that our results in the previous section are likely occurring through the direct effect of increased internet access on studying productivity. We are examining a relatively short time period and our results do not preclude any longer term impacts of increased internet access on incomes and labor market outcomes, but it appears that in this case, any such impacts were not immediately occurring.

*Table 3.6:* Impacts on Percent of Students that Work Self-Reported Family Income

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | Work | | |
| FD: Pct Internet | 0.02809 | 0.01522 | -0.10162 | -0.02015 |
| | (0.06384) | (0.06015) | (0.10802) | (0.08307) |
| | | Family Income | | |
| FD: Pct Internet | 0.00360* | 0.00170 | 0.00527* | 0.00276 |
| | (0.00202) | (0.00205) | (0.00290) | (0.00284) |
| Observations | 5196 | 5189 | 5191 | 5189 |
| Province FE | yes | yes | yes | yes |
| FD Controls | no | yes | no | yes |
| Level Controls | no | no | yes | yes |

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

## 6.3 Impact on Durable Good Purchases

We present evidence that this increased internet access is also related to the increase in the purchase of various durable household goods during this time. We include various different outcome variables for each of the following goods: (i) washing machine, (ii) microwave, (iii) oven, (iv) car, (v) DVD player, and (vi) a modern floor. Each of the variables are based on a binary question asked to the test taker as part of the SB 11 testing data. Therefore, these variables represent the change in the percent of test takers for an individual school and year who indicate that their household does own this item.

Table 3.7 below presents the results of the impact of increasing internet access of the change in the percent of test takers' households that own these durable goods. With a single exception, it appears that the increase in internet access is associated with large and significant increases in households' acquisition of these other durable goods. This is to some extent a puzzling result. One interpretation of these results is that increased access to the internet also made individuals more likely to purchase these goods. An explanation along these lines could be internet access provides more information about these goods and investments and perhaps makes them more easily accessible. While certainly possible, the magnitudes of our estimates may suggest some alternative explanation.

*Table 3.7:* Impacts on Ownership of Durable Goods

| | (1) Washer | (2) Microwave | (3) Oven | (4) Car | (5) DVD | (6) Floor |
|---|---|---|---|---|---|---|
| FD: Pct Internet | 0.29586** | 0.28063*** | 0.36819*** | 0.29532*** | 0.25136 | 0.27487*** |
| | (0.12267) | (0.03816) | (0.07511) | (0.04698) | (0.17706) | (0.05933) |
| Observations | 5198 | 5198 | 5198 | 5198 | 5198 | 5198 |
| ProvFE | Y | Y | Y | Y | Y | Y |
| FDControls | Y | Y | Y | Y | Y | Y |
| LvlControls | Y | Y | Y | Y | Y | Y |

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The finding that using this variation in distance from our starting infrastructure is leading changes in internet access to also coincide with changes in other durable ownership may suggest that something important is occurring along this geographic dimension. However, our previous results show that it is not the case that these schools that experience the largest increases in internet access are also seeing increases in the family income of the students. Similarly, it is unlikely that these differences are driven by migration or a change in the makeup of the population as we control for other key demand variables, including rurality and parents' education. One explanation is that these areas are further from the traditional core, urban areas in Colombia may be experiencing a general push towards modernization. There may have been other policy-led efforts or natural catch-up of these more removed areas that lead these two increases (internet and durables) to coincide. However, given that this does not coincide with income increases or changes in the makeup of the population, these changes in durable ownership may also be driven by supply-side increases.

The extent to which this is concerning depends on the extent that this modernizing force is also important for education outcomes, holding constant incomes and our included controls. While we think it plausible that our identifying assumptions may remain unviolated, we also redo the analysis including all of the variables for the changes in the rates of durable ownership from Table 3.7 as additional controls. In this case, our estimates remain largely unchanged and qualitatively similar. Table 3.8 below presents these results for the impact on mean scores and 90-10 percentile spreads for both mathematics and language scores, both with and without the addition of the changes in durables ownership for easy comparison. It also includes the impact on the propensity to work and self-reported family income analogous to Table 6 with the changes in durables ownership also included. Our results remain also very similar for the impact on the percentile of test scores for both mathematics and language.

*Table 3.8:* Impacts on Ownership of Durable Goods

| | (1) | (2) | (3) | (4) | |
|---|---|---|---|---|---|
| | mathmeanFD2015 | mathmeanFD2015 | langmeanFD2015 | langmeanFD2015 | mat |
| FD: Pct Internet | 0.00314 | 0.00390 | 0.00070 | 0.00069 | |
| | (0.00200) | (0.00249) | (0.00294) | (0.00355) | |
| Observations | 5198 | 5198 | 5198 | 5198 | |
| ProvFE | Y | Y | Y | Y | |
| FDControls | Y | Y | Y | Y | |
| LvlControls | Y | Y | Y | Y | |
| Durables | N | Y | N | Y | |

Standard errors in parentheses

$^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

## 7 Conclusion

This paper shows that increasing internet access can have significant impacts on test scores. We studied the internet infrastructure rollout as part of the Vive Digital Phase 1 program in Colombia to identify the impacts of increasing internet access on high school test scores. Our estimates suggest that increasing internet access had a negligible impact on language test scores, but did have a significant impact on certain segments of the mathematics test score distribution. Specifically, we did not find a statistically significant impact on mean mathematics test scores, but it did have a significant impact on the bottom third of the distribution. Our estimates suggest that for every 10% increase in the percent of test takers that have access to the internet at home, there is a 0.06 SD increase in a school's mathematics test scores in the bottom third of the distribution.

This internet infrastructure expansion was the result of a large, government-directed policy push. We are not arguing that this constitutes a cost-effective method of increasing test scores as there are various randomized interventions that have delivered substantially larger test score gains for a fraction of the cost of large infrastructure investments. Instead, these test score gains should be viewed as another previously undocumented positive byproduct of these internet expansions. We know that increasing the necessary internet infrastructure to areas that were previously unconnected has modernizing effects on the labor market and our estimates now show that these effects also have effects on education outcomes. The inevitable rollout of internet access across middle- and low-income countries not only provides an opportunity to modernize these local economies, but also to generate some catch-up in education outcomes and this constitutes another newly documented and important impact of encouraging these internet expansions.

# Bibliography

## References


Abel, J. R., Dey, I., & Gabe, T. M. (n.d.). Productivity and the density of human capital*. *Journal of Regional Science*, *52*(4), 562-586. link

Agüero, J. M., & Beleche, T. (2013). Test-mex: Estimating the effects of school year length on student performance in mexico. *Journal of Development Economics*, *103*, 353-361. link

Akerman, A., Gaarder, I., & Mogstad, M. (2015, November). The Skill Complementarity of Broadband Internet. *The Quarterly Journal of Economics*, *130*(4), 1781–1824. 2019-11-20link

Aliaga-Linares, L., & Alvarez-Rivadulla, M. J. (2010). Segregación residencial en bogotá a través del tiempo y diferentes escalas. *Documento de Trabajo de Lincoln Institute of Land Policy. Recuperado de http://bit. ly/19h8zP7*.

Anagol, S., Ferreira, F. V., & Rexer, J. M. (2021, October). *Estimating the economic value of zoning reform* (Working Paper No. 29440). National Bureau of Economic Research. link

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002, December). Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *American Economic Review*, *92*(5), 1535–1558. link

Angrist, J., Bettinger, E., & Kremer, M. (2006, June). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia. *American Economic Review*, *96*(3), 847-862. link

Autor, D. H., Palmer, C. J., & Pathak, P. A. (2014). Housing market spillovers: Evidence from the end of rent control in cambridge, massachusetts. *Journal of Political Economy*, *122*(3), 661-717. link

Bachas, P., & Soto, M. (2018). *Not(ch) your average tax system: Corporate taxation under weak enforcement*. The World Bank. link

Bajari, P., & Benkard, C. L. (2005). Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of Political Economy*, *113*(6), 1239-1276. link

Bajari, P., Fruehwirth, J. C., Kim, K. i., & Timmins, C. (2012, May). A rational expectations approach to hedonic price regressions with time-varying unobserved product attributes: The price of pollution. *American Economic Review*, *102*(5), 1898-1926. link

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007, 08). Remedying Education: Evidence from Two Randomized Experiments in India*. *The Quarterly Journal of Economics*, *122*(3), 1235-1264. link

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in colombia*. The World Bank. link

Bau, N., Rotemberg, M., Shah, M., & Steinberg, B. (2020, May). *Human capital investment in the presence of child labor* (Working Paper No. 27241). National Bureau of Economic Research. link

Baum-Snow, N., & Marion, J. (2009). The effects of low income housing tax credit developments on neighborhoods. *Journal of Public Economics*, *93*(5), 654 - 666.

link
Bayer, P., Ferreira, F., & McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, *115*(4), 588-638. link
Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? results from a natural experiment in chile. *Economics of Education Review*, *28*(5), 629-640. link
Bertanha, M., McCallum, A. H., & Seegert, N. (2021). *Better bunching, nicer notching.* link
Best, M. C., Cloyne, J. S., Ilzetzki, E., & Kleven, H. J. (2019, 05). Estimating the Elasticity of Intertemporal Substitution Using Mortgage Notches. *The Review of Economic Studies*, *87*(2), 656-690. link
Best, M. C., & Kleven, H. J. (2017, 06). Housing Market Responses to Transaction Taxes: Evidence From Notches and Stimulus in the U.K. *The Review of Economic Studies*, *85*(1), 157-193. link
Bishop, K. C., & Timmins, C. (2019). Estimating the marginal willingness to pay function without instrumental variables. *Journal of Urban Economics*, *109*, 66-83. link
Black, S. E. (1999). Do better schools matter? parental valuation of elementary education*. *The Quarterly Journal of Economics*, *114*(2), 577-599. link
Blinder, A. S., & Rosen, H. S. (1985, September). Notches. *American Economic Review*, *75*(4), 736-747. link
Blomquist, S., Newey, W., Kumar, A., & Liang, C.-Y. (2017, December). *On bunching and identification of the taxable income elasticity* (Working Paper No. 24136). National Bureau of Economic Research. link
Blomquist, S., Newey, W. K., Kumar, A., & Liang, C.-Y. (2021). On bunching and identification of the taxable income elasticity. *Journal of Political Economy*, *129*(8), 000–000.
Brown, J. N., & Rosen, H. S. (1982). On the estimation of structural hedonic price models. *Econometrica*, *50*(3), 765–768. link
Camacho, A., Caputo, J., & Sanchez, F. (2020). *Un nuevo comienzo: impacto del programa vivienda gratuita sobre las condiciones de vida de los hogares beneficiarios* (Tech. Rep.).
Camacol. (2016). *Actividad edificadora y contexto tributario* (Tech. Rep.). link (Mimeo)
Carozzi, F., Hilber, C., & Yu, X. (2020). *On the economic impacts of mortgage credit expansion policies: Evidence from help to buy* [CEP Discussion Paper No 1681]. link
Carrillo, P. E., Onofa, M., & Ponce, J. (2011). Information technology and student achievement: Evidence from a randomized experiment in ecuador.
Casas, C. (2014, October). *Subsidies to Electricity Consumption and Housing Demand in Bogotá* (Borradores de Economia No. 847). Banco de la Republica de Colombia. link
Cattaneo, M., Idrobo, N., & Titiunik, R. (2018). *A practical introduction to regression discontinuity designs: Volume i.* Cambrigde University Press, forthcoming.
Cattaneo, M., Jansson, M., & Ma, X. (2018, November). *Simple Local Polynomial Density Estimators* (Paper). arXiv.org. 2019-05-10link

Cattaneo, M., Titiunik, R., & Vazquez-Bare, G. (2019). The Regression Discontinuity Design. , 36.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., & Feng, Y. (2019a). *Binscatter regressions.* arXiv. link

Cattaneo, M. D., Crump, R. K., Farrell, M. H., & Feng, Y. (2019b). *On binscatter.* arXiv. link

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019, 05). The Effect of Minimum Wages on Low-Wage Jobs*. *The Quarterly Journal of Economics*, *134*(3), 1405-1454. link

Chay, K. Y., & Greenstone, M. (2005). Does air quality matter? evidence from the housing market. *Journal of Political Economy*, *113*(2), 376–424. link

Chen, J., Glaeser, E. L., & Wessel, D. (2019, December). *The (non-) effect of opportunity zones on housing prices* (Working Paper No. 26587). National Bureau of Economic Research. link

Chen, Z., Liu, Z., Suárez Serrato, J. C., & Xu, D. Y. (2021, July). Notching rd investment with corporate income tax cuts in china. *American Economic Review*, *111*(7), 2065-2100. link

Chernozhukov, V., Galichon, A., Henry, M., & Pass, B. (2021). *Identification of hedonic equilibrium and nonseparable simultaneous equations.*

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R., & Porter, S. R. (2018, October). *The opportunity atlas: Mapping the childhood roots of social mobility* (Working Paper No. 25147). National Bureau of Economic Research. link

Chetty, R., Friedman, J. N., Olsen, T., & Pistaferri, L. (2011, 05). Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records *. *The Quarterly Journal of Economics*, *126*(2), 749-804. link

Chetty, R., Friedman, J. N., & Saez, E. (2013, December). Using differences in knowledge across neighborhoods to uncover the impacts of the eitc on earnings. *American Economic Review*, *103*(7), 2683-2721. link

Chetty, R., Hendren, N., & Katz, L. F. (2016, April). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, *106*(4), 855-902. link

Choi, J.-y., & Lee, M.-j. (2018a, January). Minimum distance estimator for sharp regression discontinuity with multiple running variables. *Economics Letters*, *162*, 10–14. 2019-04-01link

Choi, J.-y., & Lee, M.-j. (2018b, July). Regression Discontinuity with Multiple Running Variables Allowing Partial Effects. *Political Analysis*, *26*(3), 258–274. 2019-03-28link

Cohen, M., Carrizosa, M., & Gutman, M. (2019). *Urban policy in latin america: Towards the sustainable development goals?* Routledge.

Collinson, R., & Ganong, P. (2018, May). How do changes in housing voucher design affect rent and neighborhood quality? *American Economic Journal: Economic Policy*, *10*(2), 62-89. link

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2017, July). Technology and child development: Evidence from the one laptop per child program.

*American Economic Journal: Applied Economics*, *9*(3), 295-320. link
DANE. (2018). *Manual De Uso Del Marco Geoestadístico Nacional En El Proceso Estadístico.* 2020-06-12link
DeFusco, A. A., & Paciorek, A. (2017, February). The interest rate elasticity of mortgage demand: Evidence from bunching at the conforming loan limit. *American Economic Journal: Economic Policy*, *9*(1), 210-40. link
Departamento Administrativo de Planeación Distrital. (2004). *La estratificación en Bogotá D.C. y estudios relacionados, 1983-2004*. Bogotá: Alcaldía Mayor de Bogotá D.C., Departamento Administrativo Planeación Distrital. (OCLC: 221308728)
de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2017, 11). Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*. *The Quarterly Journal of Economics*, *133*(2), 993-1039. link
Diamond, R., McQuade, T., & Qian, F. (2019, September). The effects of rent control expansion on tenants, landlords, and inequality: Evidence from san francisco. *American Economic Review*, *109*(9), 3365-94. link
Diamond, R., & Persson, P. (2016, April). *The long-term consequences of teacher discretion in grading of high-stakes tests* (Working Paper No. 22207). National Bureau of Economic Research. link
Dinkleman, T. (2011). The Effects of Rural Electrification on Employment: New Evidence from South Africa. *The American Economic Review*, *101*(7), 3078–3108. link
DNP. (2018). *Sistema unificado de inversiones y finanzas publicas - suifp* (Working Paper). link
Econometria. (1999). *Revisión de la metodología de estratificación socioeconómica de Santa Fe de Bogotá. Bogotá.*
Einav, L., Finkelstein, A., & Schrimpf, P. (2015, 02). The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D *. *The Quarterly Journal of Economics*, *130*(2), 841-899. link
Ekeland, I. (2010, 02). Existence, uniqueness and efficiency of equilibrium in hedonic markets with multidimensional types. *Economic Theory*, *42*(2), 275-315. link (Copyright - Springer-Verlag 2010; Document feature - ; Equations; Last updated - 2021-02-16)
Ekeland, I., Heckman, J. J., & Nesheim, L. (2004). Identification and estimation of hedonic models. *Journal of Political Economy*, *112*(S1), S60-S109. link
Epple, D. (1987). Hedonic prices and implicit markets: Estimating demand and supply functions for differentiated products. *Journal of Political Economy*, *95*(1), 59-80. link
Epple, D., Quintero, L., & Sieg, H. (2020). A new approach to estimating equilibrium models for metropolitan housing markets. *Journal of Political Economy*, *128*(3), 948-983. link
Forman, C., Goldfarb, A., & Greenstein, S. (2012, February). The internet and local wages: A puzzle. *American Economic Review*, *102*(1), 556-75. link
Franklin, S. (2019). *The demand for government housing: Evidence from lotteries for 200,000 homes in ethiopia.* 2020-09-15link (Library Catalog: www.dropbox.com)

Galiani, S., Murphy, A., & Pantano, J. (2015, November). Estimating neighborhood choice models: Lessons from a housing assistance experiment. *American Economic Review*, *105*(11), 3385-3415. link

Gallego, J., Montoya, S., & Sepúlveda, C. (2016). Effect of socio-economic stratification on house value in bogotá.

Gaubert, C., Kline, P., & Yagan, D. (2020). *Place-based redistribution* (Paper). Berkeley. 2020-07-21link

Geyer, J. (2017). Housing demand and neighborhood choice with housing vouchers. *Journal of Urban Economics*, *99*, 48-61. link

Gilbert, A. (2014, June). *Housing policy in Colombia.* 2020-05-10link (Library Catalog: www.taylorfrancis.com Pages: 282-296 Publisher: Routledge)

Gilbert, A. G. (2014). Free housing for the poor: An effective way to address poverty? *Habitat International*, *41*, 253 - 261. link

Glaeser, E., & Shapiro, J. M. (2003). The benefits of the home mortgage interest deduction. *Tax Policy and the Economy*, *17*, 37-82. link

Glaeser, E. L., & Gottlieb, J. D. (2008, October). *The economics of place-making policies* (Working Paper No. 14373). National Bureau of Economic Research. link

Glaeser, E. L., & Luttmer, E. F. P. (2003, September). The misallocation of housing under rent control. *American Economic Review*, *93*(4), 1027-1046. link

Goff, L. (2021). *Treatment Effects in Bunching Designs: The Impact of the Federal Overtime Rule on Hours* (Tech. Rep.).

Goncalves, F., & Mello, S. (2021). A few bad apples? racial bias in policing. *American Economic Review*. link

González, J. I. (2014). *Aglomeraciones y condiciones de vida en Bogotá.* 2019-05-14link

Goolsbee, A., & Guryan, J. (2006, 05). The Impact of Internet Subsidies in Public Schools. *The Review of Economics and Statistics*, *88*(2), 336-347. link

Greenstone, M. (2017). The continuing impact of sherwin rosen's "hedonic prices and implicit markets: Product differentiation in pure competition". *Journal of Political Economy*, *125*(6), 1891-1902. link

Grogan, L. (2016). Household electrification, fertility and employment: Evidence from hydroelectric dam construction in Colombia. *Journal of Human Capital*, *10*(1), 109–158. link

Grogan, L. (2018, November). Time use impacts of rural electrification: Longitudinal evidence from Guatemala. *Journal of Development Economics*, *135*, 304–317. link

Grogan, L., & Sadanand, A. (2013, March). Rural Electrification and Employment in Poor Countries: Evidence from Nicaragua. *World Development*, *43*, 252–265. link

Gruber, J., Jensen, A., & Kleven, H. (2021, May). Do people respond to the mortgage interest deduction? quasi-experimental evidence from denmark. *American Economic Journal: Economic Policy*, *13*(2), 273-303. link

Gyourko, J., & Molloy, R. (2015). Chapter 19 - regulation and housing supply. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of regional and urban economics* (Vol. 5, p. 1289 - 1337). Elsevier. link

Haan, P., & Simmler, M. (2018). Wind electricity subsidies — a windfall for landowners? evidence from a feed-in tariff in germany. *Journal of Public Economics*, *159*, 16-32. link

Hanna, R., & Olken, B. A. (2018, November). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, *32*(4), 201-26. link

Harasztosi, P., & Lindner, A. (2019, August). Who pays for the minimum wage? *American Economic Review*, *109*(8), 2693-2727. link

Harker, A., Lustig, N., Martínez, V., & Melendez, M. (2016). *The impact of taxes and transfers on inequality and poverty in colombia* (Tech. Rep.). CEQ Working Paper 24. CEQ Institute, Center for Inter-American Policy and ....

Heckman, J. J., Matzkin, R. L., & Nesheim, L. (2010). Nonparametric identification and estimation of nonadditive hedonic models. *Econometrica*, *78*(5), 1569-1591. link

Hilber, C. A. L. (2017). The economic implications of house price capitalization: A synthesis. *Real Estate Economics*, *45*(2), 301-339. link

Hjort, J., & Poulsen, J. (2019, March). The Arrival of Fast Internet and Employment in Africa. *American Economic Review*, *109*(3), 1032–1079. 2019-11-20link

Igac, I. G. A. C. (2013, September). *Resolución 829 De 26 De Septiembre De 2013.* 2020-06-12link (Library Catalog: legal.legis.com.co)

Jales, H. (2018). Estimating the effects of the minimum wage in a developing country: A density discontinuity design approach. *Journal of Applied Econometrics*, *33*(1), 29-51. link

Jimenez, R. (2017). *Development Effects of Rural Electrification* (Policy Brief No. IDB-PB-261). Inter-American Development Bank. link

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, *8*(1), 435-464. link

Kleven, H. J., & Waseem, M. (2013, 04). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan *. *The Quarterly Journal of Economics*, *128*(2), 669-723. link

Kline, P., & Moretti, E. (2014). People, places, and public policy: Some simple welfare economics of local economic development programs. *Annual Review of Economics*, *6*(1), 629-662. link

Kopczuk, W., & Munroe, D. (2015, May). Mansion tax: The effect of transfer taxes on the residential real estate market. *American Economic Journal: Economic Policy*, *7*(2), 214-57. link

Kumar, T. (2021). The housing quality, income, and human capital effects of subsidized homes in urban india. *Journal of Development Economics*, *153*, 102738. link

Kuminoff, N. V., Smith, V. K., & Timmins, C. (2013, December). The new economics of equilibrium sorting and policy evaluation using housing markets. *Journal of Economic Literature*, *51*(4), 1007-62. link

Linden, L. L. (2008). *Complement or substitute?: The effect of technology on student achievement in india*. Citeseer.

Lipscomb, M., Mobarak, M., & Barham, T. (2013, April). Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil. *American Economic Journal: Applied Economics*, *5*(2), 200–231. link

Lozano-Gracia, N., & Anselin, L. (2012). Is the price right?: Assessing estimates of cadastral values for bogotá, colombia. *Regional Science Policy Practice*, *4*(4), 495-508. link

Lutz, B. (2015, February). Quasi-experimental evidence on the connection between property taxes and residential capital investment. *American Economic Journal: Economic Policy*, *7*(1), 300-330. link

Malamud, O., & Pop-Eleches, C. (2011, 05). Home Computer Use and the Development of Human Capital *. *The Quarterly Journal of Economics*, *126*(2), 987-1027. link

McMillen, D., & Singh, R. (2020). Fair market rent and the distribution of rents in los angeles. , *80*, 103397. 2020-09-19link

McRae, S. (2015, January). Infrastructure quality and the subsidy trap. *American Economic Review*, *105*(1), 35-66. link

Medina, C., & Morales, L. (2007). Stratification and public utility services in colombia: Subsidies to households or distortion of housing prices? [with comments]. *Economía*, *7*(2), 41–99. link

Meléndez, M. (2004, August). Subsidios al consumo de los servicios públicos en Colombia. ¿Hacia dónde nos movemos? 2019-03-08link

Metrocuadrado. (2022). *¿cuánto cuesta arreglar un apartamento en obra gris?* 2022-04-16link

Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M., & Rozelle, S. (2013). Can one-to-one computing narrow the digital divide and the educational gap in china? the case of beijing migrant schools. *World Development*, *46*, 14-29. link

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019, April). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, *109*(4), 1426-60. link

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from india. *Journal of Political Economy*, *119*(1), 39-77. link

OECD. (2021a, 01). *Building for a better tomorrow: Policies to make housing more affordable. oecd employment, labour and social affairs policy briefs.* link

OECD. (2021b). *Ph1.2 housing policy objectives and obstacles* (Working Paper). link

Olsen, E. O. (2003, January). Housing programs for low-income households [Book]. In *Means-tested transfer programs in the united states* (p. 365-442). University of Chicago Press. link

Olsen, E. O., & Zabel, J. E. (2015). Chapter 14 - us housing policy. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of regional and urban economics* (Vol. 5, p. 887 - 986). Elsevier. link

Palmquist, R. B. (2006). Property value models. In K. G. Mäler & J. R. Vincent (Eds.), *Handbook of environmental economics* (1st ed., Vol. 2, p. 763-819). Elsevier. link

Papay, J. P., Willett, J. B., & Murnane, R. J. (2011, April). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, *161*(2), 203–207. 2019-05-05link

Pattillo, M. (2013). Housing: Commodity versus right. *Annual Review of Sociology*, *39*(1), 509-531. link

Persson, P. (2020). Social insurance and the marriage market. *Journal of Political Economy*, *128*(1), 252-300. link

Poterba, J. M. (1992). Taxation and housing: Old questions, new answers. *The American Economic Review*, *82*(2), 237–242. link

Quigley, J. M. (1982). Nonlinear budget constraints and consumer demand: An application to public programs for residential housing. *Journal of Urban Economics*, *12*(2), 177 - 201. link

radio, U. (2022). *En bogotá se están vendiendo "viviendas de interés social" en estrato 6 y con un metro cuadrado mayor a los 9 millones de pesos: Diego laserna, concejal.* 2022-04-16link

Reardon, S. F., & Joseph P., R. (2012). Regression Discontinuity Designs With Multiple Rating-Score Variables. *Journal of Research on Educational Effectiveness*. 2019-03-27link

*Registro, distrital acuerdo número 657* (No. Annex 2). (2016). 2020-09-26link

Riehl, E., Saavedra, J. E., & Urquiola, M. (2016). *Learning and earning: An approximation to college value added in two dimensions* (Tech. Rep.). National Bureau of Economic Research.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954. link

Rosen, H. S. (1985). Chapter 7 housing subsidies: Effects on housing decisions, efficiency, and equity. In (Vol. 1, p. 375 - 420). Elsevier. link

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34-55. link

Ruiz, F., & Vallejo, G. (2015). *Using land registration as a tool to generate municipal revenue: Lessons from bogota.* link

Saez, E. (2010). Do taxpayers bunch at kink points? *American economic Journal: economic policy*, *2*(3), 180–212.

Saiz, A. (2010). The geographic determinants of housing supply*. *The Quarterly Journal of Economics*, *125*(3), 1253-1296. link

SDP. (2004). *Bogotá, D.C., y su Estratificación Socioeconómica.* 2019-03-08link

Sepulveda, C. E., Lopez Camacho, D., & Gallego, J. M. (2014). *Los límites de la estratificación: en busca de alternativas.* 2019-05-02link

Shah, M., & Steinberg, B. M. (2017). Drought of opportunities: Contemporaneous and long-term impacts of rainfall shocks on human capital. *Journal of Political Economy*, *125*(2), 527-561. link

Sinai, T., & Waldfogel, J. (2005). Do low-income housing subsidies increase the occupied housing stock? *Journal of Public Economics*, *89*(11), 2137 - 2164. link

Slemrod, J., Weber, C., & Shan, H. (2017). The behavioral response to housing transfer taxes: Evidence from a notched change in d.c. policy. *Journal of Urban Economics*, *100*, 137 - 153. link

Soltas, E. (2021). The price of inclusion: Evidence from housing developer behavior.. link

Squires, T. (2015a). The impact of access to electricity on education: evidence from honduras. *Job Market Paper, Brown University*.

Squires, T. (2015b). *The Impact of Access to Electricity on Education: Evidence from Honduras* (Economics Dissertation, Brown University, RI, USA). link

SUI, s. (2020). *Datos abiertos | superservicios - superintendencia de servicios públicos domiciliarios.* 2020-09-23link

Tsivanidis, N. (2018). *The aggregate and distributional effects of urban transit infrastructure: Evidence from bogotá's transmilenio* (Tech. Rep.). Working Paper, University of Chicago Booth School of Business.

Turner, M. A., Haughwout, A., & van der Klaauw, W. (2014). Land use regulation and welfare. *Econometrica*, *82*(4), 1341-1403. link

Uribe, J. P. (2021). ***The effect of location based subsidies on the housing market***.

van de Walle, D., Ravallion, M., Mendiratta, V., & Koolwal, G. (2017, June). Long-term Gains from Electrification in Rural India. *The World Bank Economic Review*, *31*(2), 385–411. link

van Dijk, W. (2019). The socio-economic consequences of housing assistance. *University of Chicago Kenneth C. Griffin Department of Economics job market paper, 0–46 i–xi*, *36*.

Vega, D. M. (2013). Colombia's digital agenda: Successes and the challenges ahead. *The Global Information Technology Report*, *Chapter 2.1*, 111–117. link

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). *Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods.* 2019-03-28link

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

World Bank Group, P. P. (1993). *Housing : enabling markets to work (english)* (Working Paper No. 27704). link

Yinger, J. (2015). Hedonic markets and sorting equilibria: Bid-function envelopes for public services and neighborhood amenities. *Journal of Urban Economics*, *86*, 9-25. link

Zoutman, F. T., Gavrilova, E., & Hopland, A. O. (2018). Estimating both supply and demand elasticities using variation in a single tax rate. *Econometrica*, *86*(2), 763-771. link