Essays on Applied Econometrics

Felipe Carneiro de Figueredo

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Rachel Heath, Chair

Alan Griffith

Denis O'Dea

Program Authorized to Offer Degree:
Department of Economics

University of Washington

Abstract

Essays on Applied Econometrics

Felipe Carneiro de Figueredo

Chair of the Supervisory Committee:
Rachel Heath
Deparment of Economics

This dissertation consists of three essays in applied econometrics that explore experimental, quasi-experimental, and observational methodologies to study political behavior, infrastructure-driven labor market change, and price elasticity of demand estimation.

The first chapter examines the causal impact of spatial proximity on legislative behavior by exploiting the randomized allocation of offices in the Brazilian Chamber of Deputies. The analysis finds that legislators assigned to neighboring offices are significantly more likely to vote alike in contested decisions—an effect amplified when at least one legislator is a policy expert, such as a committee member. The results provide empirical support for cue-taking theories of legislative decision-making, suggesting that informal physical proximity enhances the influence of expertise, particularly in closely divided votes.

The second chapter evaluates the labor market impacts of Brazil's broadband expansion policy using a regression discontinuity design (RDD) on about 2,000 municipalities covered by microwave radio technology. The findings reveal heterogeneous effects: while the policy stimulates job creation among low-educated workers and in commerce-related sectors, it simultaneously reduces hours worked and wages, especially for highly educated individuals, skilled occupations, and women in services. These findings highlight the inherent trade-offs of digital infrastructure policies, wherein the same technological improvements that foster inclusiveness and job growth can also precipitate labor market disruptions, possibly through automation and substitution effects.

The third chapter applies a new approach to estimating price elasticity of demand by combining double machine learning (DML) with multimodal embeddings derived from product descriptions and images. This combination offers two key advantages. First, the embeddings capture rich, high-dimensional signals of product quality that are often unobserved but crucial in shaping both prices and consumer demand. By leveraging visual and textual features, the method provides a

data-driven way to control for latent quality differences across products. Second, DML allows for flexible, machine-learning-based estimation of complex relationships—such as how price and demand depend on covariates—while still delivering valid causal estimates. Together, these tools offer a powerful solution to the endogeneity problem in demand estimation, substantially reducing bias from unobserved quality.

Together, these essays demonstrate how causal inference can be empirically applied to diverse data environments to uncover the mechanisms shaping legislative peer effects, labor market outcomes, and consumer behavior.

TABLE OF CONTENTS

Page

List of Figures . . . iii

Chapter 1: The Impact of Office Proximity in Legislative Decision-Making: Evidence from Brazil . . . 1
1.1 Introduction . . . 1
1.2 Literature Review . . . 6
1.3 Institutional Background . . . 8
1.4 Empirical Strategy . . . 13
1.5 Empirical Results . . . 20
1.6 Conclusion . . . 38

Chapter 2: Fast Internet and Labor Market Impact: Evidence from Differential Broadband Rollout in Brazil . . . 41
2.1 Introduction . . . 41
2.2 Background . . . 43
2.3 Empirical Model . . . 49
2.4 Results . . . 58
2.5 Conclusion . . . 78

Chapter 3: Double Machine Learning for Price Elasticity Estimation: Leveraging Unstructured Data from the Steam Digital Store . . . 79
3.1 Introduction . . . 79
3.2 Background and Literature Review . . . 82
3.3 Data . . . 89
3.4 Empirical Model: DML with Multimodal Embeddings . . . 97
3.5 Results and Interpretation . . . 101
3.6 Limitations . . . 107
3.7 Conclusion . . . 108

Bibliography . . . 109

Appendix A: The Impact of Office Proximity in Legislative Decision-Making: Evidence from Brazil . . . 119

Appendix B: Fast Internet and Labor Market Impact: Evidence from Differential Broadband Rollout in Brazil . . . 130

Appendix C: Double Machine Learning for Price Elasticity Estimation: Leveraging Unstructured Data from the Steam Digital Store 135

LIST OF FIGURES

Figure Number Page

1.1 Number of Roll Calls per Day - 55th and 56th legislature . . . 14
1.2 Number of Speeches per Day - 55th and 56th legislature . . . 15
1.3 Neighboring Locations . . . 17
1.4 Heterogeneity analysis on roll calls' topics . . . 29
1.5 Heterogeneity analysis on legislators' characteristics . . . 31

2.1 Rule of Assignment and Distribution of Municipalities . . . 45
2.2 Backhaul Technologies - 2011. *Source*:Henriksen et al. (2022) . . . 47
2.3 Histogram of Population Size . . . 51
2.4 Backhaul Capacity vs Population Size . . . 52
2.5 Density Plots of Manipulation Tests: Radio . . . 55
2.6 Density Plots of Manipulation Tests: Fiber . . . 55
2.7 Effects on Fast Internet on Internet Use - Radio . . . 61
2.8 Effects on Fast Internet on Internet Use - Fiber . . . 61
2.9 Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) - Radio . . . 64
2.10 Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) per Educational Level - Radio . . . 68
2.11 Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) per Skill Level - Radio . . . 71
2.12 Effects of Broadband Expansion on Municipal Labor Market Outcomes - Radio Technology, 2011–2015 . . . 77

3.1 Price Elasticity by Game Genre . . . 103
3.2 Price Elasticity vs Average Price by Game Genre . . . 105
3.3 Price Elasticity vs Sample Size by Game Genre . . . 106

A.1 Drawing for one of the offices . . . 119
A.2 Office Plan - Annex III . . . 120
A.3 Office Plan - Annex IV . . . 121
A.4 Distribution of Votes - 55th and 56th legislatures . . . 123
A.5 CCJC . . . 126
A.6 CFT . . . 127
A.7 CCULT . . . 127
A.8 Randomization Inference for causal estimate (sample: 5 p.p. result margin) . . . 128
A.9 Randomization Inference for Interaction of Office Neighbors and Committee. . . . 129

B.1 Net Firm Entry 134

ACKNOWLEDGEMENTS

I sincerely thank everyone who supported me throughout the journey of completing this Ph.D.

First, I would like to thank my advisor, Rachel, for her invaluable guidance and sharp insights that shaped the direction of this research. I also wish to express my gratitude to the members of my dissertation committee — Alan, John, Isabelle, and Dennis — for their thoughtful feedback and generous time. Their comments greatly improved the quality of this work.

A special thanks goes to my colleagues and friends at the Department of Economics, who provided both intellectual companionship and moral support. The countless discussions, coffee breaks, and moments of shared frustration and laughter were an essential part of this experience. My thanks also go to the staff of the Department of Economics for their continued assistance throughout my time here. I also wish to express my gratitude to Eduardo at the Department of Spanish and Portuguese, whose generosity and confidence greatly supported my academic experience.

I am profoundly thankful to my family for their unwavering support and belief in me. To my parents, Arliene and Liberato, and my sister, Fernanda — thank you for your love, sacrifices, and encouragement. I also extend my heartfelt gratitude to my extended family — grandparents, uncles, aunts, and cousins — for their constant support. Lastly, to the friends I was lucky to meet at University of Washington and in Seattle — thank you for being there providing encouragement when I needed it most.

DEDICATION

To my grandfather, Chico (in memoriam)

Chapter 1

THE IMPACT OF OFFICE PROXIMITY IN LEGISLATIVE DECISION-MAKING: EVIDENCE FROM BRAZIL

1.1 Introduction

Economists are increasingly acknowledging the important role of social interactions in shaping individual behaviors and how individuals value decisions and their outcomes. There is evidence of this influence in many different domains, for example, education (Sacerdote, 2001; Zimmerman, 2003), labor markets (Cingano and Rosolia, 2012; Zenou, 2008), welfare programs (Duflo and Saez, 2003), crime (Glaeser et al., 1996), youth behavior (Case and Katz, 1991; Evans et al., 1992; Kawaguchi, 2004), and demand for housing quality (Patacchini and Venanzoni, 2014). Social interactions have also played an important role in the legislative system, as interactions and interpersonal relationships among legislators across different social structures can significantly influence legislative behavior and policy outcomes (Caldeira and Patterson, 1987; Fowler, 2006).

Spatial proximity within the congress represents an important underlying social structure, as it influences the patterns of social interactions of legislators and their decision-making. Closer spatial proximity, such as through seating arrangements and office locations, is associated with stronger social connections and influence (Alquezar-Yus and Amer-Mestre, 2024; Darmofal et al., 2023; Ferber and Pugliese, 2000; Harmon et al., 2019; Lowe and Jo, 2024; Masket, 2008; Rogowski and Sinclair, 2012; Saia, 2018). Past studies have not been able to clearly determine whether these relationships are causal, as in many settings legislators may endogenously choose to locate near certain peers. For example, legislators might choose to stay closer and communicate more frequently with those who are similar to themselves, such as in party affiliation[1]. As spatial proximity could be related with other associated attributes – for instance, similar views – it is not straightforward to identify their independent social effects. Thus, group endogeneity problems present a challenge for the identification of peer effects within this type of network.

This paper examines how spatial proximity influences legislative decision-making by taking

[1] As Krehbiel (1993) claims: "In casting apparently partisan votes, do individual legislators vote with fellow party members in spite of their disagreement about the policy in question, or do they vote with fellow party members because of their agreement about the policy question?"

advantage of the completely randomized allocation of legislative offices to first-term male legislators in the Chamber of Deputies of Brazil. I use this office lottery to establish causal links between office proximity and legislators' behavior, as the random assignment of office locations eliminates selection bias[2]. Brazil's proportional and open list electoral system, which often leads to majority coalition formation in the legislative, presents an intriguing environment for studying peer effects dynamics. In addition, most newly allocated offices are located in the same building, which consists of eight identical floors, each with a single hallway leading to a bank of elevators. This layout fosters spontaneous communication among legislators who frequently engage with their neighbors. For instance, former Worker's Party (PT) deputy, João Grandão, once reported that he frequently ran into his officemate and political adversary, who later became president, Jair Bolsonaro. He stated, "I always talked to him, (...) and I never had a problem with him"[3], highlighting that their interactions were frequent despite any political or ideological differences. In another, more ordinary situation, deputy Gonzaga Sobrinho mentioned that he borrowed a tie from his office neighbor, Deputy Silvio Costa[4]. These examples highlight the casual interactions that often take place in office neighborhoods.

To explore the impact of office proximity on legislative behavior, I analyze voting convergence for all possible legislator pairs per roll call. The sample is restricted to pairs in which at least one member was subject to the lottery. Pairs of incumbents are excluded because both members could have chosen their office neighborhood. To create these pairs, this study uses data from the 55th and 56th legislative sessions – covering the periods from 2015 to 2022. Spatial proximity is measured by an indicator of whether a given office is directly next to, in front of, or diagonally across from any other office. I also analyze the impact of office proximity on legislative speech content using document embeddings to assess speech similarity.

The results show that, on average, office proximity increases vote convergence by 0.26 percentage points – this effect amounts to 8% of the influence of same party membership and is not statistically significant. Since most of the votes in the sample are cast on procedural issues, where there is little

[2]Other network effects, such as through friendship ties, also suffer from the same endogeneity problem. Cohen and Malloy (2014) and Battaglini et al. (2023a,b) have used alumni network data to identify the impact of friendship on legislative decision-making

[3]https://www.topmidianews.com.br/na-lata/na-lata-vizinho-de-gabinete-de-bolsonaro-deputado-do-pt-lembra/67504/.

[4]In his speech on June 3, 2015, Deputy Gonzaga Sobrinho remarked: "As they said there wouldn't be a session today and that all the Deputies had traveled, I'm now seeing that many are here and there will indeed be a session. So, I had to borrow a jacket from the guard at the entrance and got the tie from Deputy Silvio Costa, my office neighbor."

variation, as legislators usually stick to their party's position or their personal beliefs, it could be that these influences potentially mask the effects of spatial proximity on the convergence of votes within pairs of legislators. Thus, I run a heterogeneity analysis taking into account the importance of voting by emphasizing non-procedural decisions. In disputed votes where roll calls have narrow victory or defeat margins, there is strong evidence for direct influence from office-mates on voting agreement. Votes for highly contested roll calls, with a margin of less than 5 p.p., show a statistically significant estimate of 2 percentage points – roughly eight times the magnitude of "comfortable" votes. Although smaller in magnitude, the effects of spatial proximity remain robust with larger result margins (10 p.p. and 25 p.p.) and with an alternative measure of voting convergence. The results do not hold under virtual vote systems but remain robust when tested using randomization inference.

This study also explores the relationship between office proximity and expertise in disputed decisions. I show evidence that pairs with neighboring legislators that are experts present higher voting convergence in contested votes. An expert (or perceived expert) is defined as a member of the committee related to the subject of the proposal being voted. The point estimate is 4.51 percentage points for roll calls with a result margin of less than 5 p.p. – and remain significant with a result margin of less than 10 p.p. and with randomization inference. This finding suggests that when confronted with the necessity of making a decision on a disputed legislation, legislators can look to individuals in closer proximity and who are perceived to be experts. Usually these are members of the standing committee that reported the measure to the floor. Legislators coped with uncertainty about the consequences of voting one way or the other by using the voting intentions of trusted, expert colleagues to decide how to vote (Uslaner and Weber, 1977). These findings also suggest a potential mechanism of cue-taking where legislators rely on perceived experts—typically standing committee members who reported the measure – to guide their voting decisions, especially when informed decision-making is crucial, though other mechanisms, like social pressure, cannot be ignored (Lowe and Jo, 2024). A work close to mine is Fong (2018), who exploits a natural experiment wherein legislators are assigned to committees midsession because of the death, resignation, or transfer of the seat's previous occupant. He analyses co-sponsorship networks and finds that peers who are close to the legislator in the legislative network tend to vote more often with that legislator on bills from the legislator's new committee's jurisdiction after the assignment than they did before.

I also run heterogeneity analyses based on shared social characteristics, vote topics, and proposition types to explore how these factors influence the outcomes. I show that pairs sharing the

same freshman status and the same party both look statistically significant, whereas same state of origin only looks marginally distinguishable from zero. The effects of proximity are stronger on topics related to labor and social security. Furthermore, heterogeneity analysis by proposition type shows that votes on Provisional Measures – executive decrees with immediate legal force pending legislative approval – exhibit stronger agreement among closely located legislators. However, the point estimate is not statistically significant as is the other point estimate for the remaining types of proposition. It, therefore, reveals that type of proposition does not represent an important dimension of heterogeneity in the analysis of the influence of spatial proximity on co-voting patterns.

I also implement a number of robustness checks. Firstly, I use different definitions of office neighborhood and voting convergence, showing that the choices made do not impact the main results and their interpretation. Secondly, it explores alternative definitions of the importance of voting beyond their result margin. Taking in consideration the relevance of the individual vote for each legislator, it adapts the Battaglini et al. (2023a) approach, in which roll calls are classified as relevant or not based on the salience of the topic being voted on by each legislator. This approach utilizes co-sponsorship data to determine the frequency with which each legislator has supported each topic, considering the least frequent topic as not relevant. The results only offer weak evidence that office proximity is more likely to increase agreement when the vote is not relevant for either both or only one legislator in a pair. This evidence weakly indicates that cue-taking might occur through office proximity in votes that are not part of the own legislator's agenda, situations in which they tend to be less informed about the legislation being voted.

A second approach to classifying votes according to their relevance follows Saia (2018), which categorizes roll calls based on the intensity of debate. Using speech data, it calculates the average daily number of speeches to establish a threshold, classifying roll calls as non-relevant if they do not exceed this threshold. This approach finds that roll calls classified as relevant and voted on during days of higher debate intensity are more likely to align among legislators in closer office proximity. Days with heightened debate intensity imply that legislators are likely to be more present in the Chamber as well as in their offices, potentially increasing the likelihood of communication among them. Thirdly, I analyze virtual votes cast under the *Sistema de Deliberação Remota* (SDR)[5] and show, on average, a negative and not statistically significant point estimate for voting

[5]Due to social distancing measures introduced to combat the COVID-19 pandemic, the Chamber of Deputies (56th legislature) implemented the SDR, *Sistema de Deliberação Remota*, transitioning all plenary activities—including

convergence associated with sharing a similar location. For contested votes, with a result margin lower than 5 p.p., the impact of office proximity is 10 times smaller and no longer statistically significant. The effect of committee membership interacted with similar location also changes and is no longer statistically significant. Both findings support our hypothesis on the importance of physical interactions in explaining the influence of spatial proximity, also underscore the importance of keeping the standing committees operational.

This paper offers empirical evidence that adds to the existing literature on the influence of spatial proximity on legislative behavior (Alquezar-Yus and Amer-Mestre, 2024; Darmofal et al., 2023; Ferber and Pugliese, 2000; Harmon et al., 2019; Lowe and Jo, 2024; Masket, 2008; Rogowski and Sinclair, 2012; Saia, 2018), suggesting that office proximity serves as a social structure that influences non-procedural decisions and complements both ideological and non-ideological factors (Zucco and Lauderdale, 2011). By demonstrating that office proximity affects voting behavior even amid strong party and coalition influences, it advances as well our understanding of legislative dynamics. I also contribute to the informational theory of legislative committees by offering empirical evidence on their importance in the diffusion of information. In their model of strategic communication applied to the legislative context, Gilligan and Krehbiel (1987) demonstrate the informational rationale for committee power. They show that informational gains to the parent chamber are maximized when committee medians are near floor medians. Pereira and Mueller (2004) adapt this model to the Brazilian Congress and show the committees that are more representative of the floor that would have a greater effect of reducing uncertainty in equilibrium. On the other hand, if all the cue-givers in a policy domain tend to be a biased sample of the congress, the probability of unrepresentative policy decisions is very substantial. Therefore, this work contributes to the discussion by introducing a new transmission mechanism, spatial proximity, highlighting how legislative networks and expertise interact to influence policy outcomes.

Section II reviews the literature. Section III presents the institutional background of this study. Section III introduces the data and the empirical strategy. Section IV presents the peer-effects analysis on legislative behavior and its robustness checks. Section V concludes.

---

roll calls—to a remote format and temporarily suspending the functions of permanent committees. This completely remote setup lasted from March 17, 2020 until February 11, 2021, when a hybrid model was introduced, allowing legislators to return to in-person plenary sessions and resuming standing committee activities. This hybrid system remained in place until October 25, 2021, when normal operations resumed. Briefly in 2022, from the start of the legislative year in February until April 18, the Chamber reinstated the SDR. The period under the SDR allows me to perform a counterfactual exercise.

1.2 Literature Review

1.2.1 Spatial Proximity

In his hallmark work, Young (1966) explored the influence of boardinghouse groups on congressional voting over the Jeffersonian Era. He presented evidence that these intralegislative fraternal associations were influences of major significance upon the members' voting behavior, so that these social structures tended to institutionalize the difference among members. In a more recent study, Ferber and Pugliese (2000) further examined the influence of spatial proximity on the interpersonal communication patterns among legislators by analyzing seating and office shared locations. They found supportive evidence that seat proximity is positively related to the interpersonal communication patterns among members in the New York State Assembly. However, they could only weakly support the hypothesis that office proximity similarly influences these patterns. Masket (2008) explored data from three decades of roll call votes in the California Assembly and its seating rule allocation to show that legislators actively seek to their deskmates for cues, relying on this information and guidance to make voting choices that further their goals. Consequently, the influence of proximity extends to shaping how legislators interact, collaborate, and vote with one another, thereby molding the trajectory of political conflicts and legislative outcomes.

Use of randomization-based research design for the identification of spatial proximity effects has been effective in the context of homophily bias[6]. Using exogenous variation in seating within the legislature, Harmon et al. (2019) and Alquezar-Yus and Amer-Mestre (2024) identify the influence of spatial proximity on voting behavior in the European Parliament. Saia (2018) uses the random allocation of seats in the Iceland Parliament to identify the causal effect of social interaction on legislators' voting and speech behavior. He shows that the probability of failure to toe one's party line is higher, the higher the fraction of peers seated nearby casting votes that diverge from the party line, while peers' influence appears to have a sizable impact on legislator's speech behavior. Lowe and Jo (2024) and Darmofal et al. (2023) also explore the same setting to make causal claims of the effects of spatial proximity over voting behavior. Rogowski and Sinclair (2012) use the United States House of Representatives office lottery – in which newly elected members select their offices in a randomly assigned order – to obtain causal estimates of peer effects that are not biased due

[6]This methodology has been also widely used in the identification of peer effects in other different domains, such as education (Sacerdote, 2001; Yakusheva et al., 2011; Zimmerman, 2003) and workplace (Guryan et al., 2009). An alternative is to use research designs that attempt to control for homophily by incorporating additional covariates or leveraging natural experiments.

to homophily or endogeneity. Contrary to previous findings, they found no evidence that office proximity affects legislative behavior.

1.2.2 Mechanisms

Although the main specification does not allow me to perfectly distinguish a specific mechanism, my results are consistent with the mechanism of cue-taking. In their book, Matthews and Stimson (1975) point out that cue-taking refers to the mechanism of information sharing in which certain legislators serve as cue-givers, providing signals to their peers, while others act as cue-takers, receiving these signals and basing their voting decisions on them. Zelizer (2019) argues that spatial proximity among legislators leads to the formation of new cue-taking relationships. Closer spatial proximity reduces the barriers that typically hinder communication, representing a shortcut way of making reasonable decisions. For instance, sharing the same office floor reduces transaction costs associated with seeking guidance, facilitating easier access to information. Caldeira and Patterson (1987) demonstrated that spatial proximity facilitates the development of interpersonal ties among members, shaping the legislature by establishing channels of communication and enabling the connections necessary for bargaining, exchanging cues, and decision-making.

Legislators must turn to someone for cues on how to vote when there is a situation of a time constraint on the decision and the existence of an input deficiency or an input overload (Kingdon, 1989). His interview data show the informants have some claim to expertise 82 percent of the time. Legislators argue that in more complex decisions, they pick informants with some expertise on the legislation. The committee collectively and its members individually are potent cue-givers for the obvious reason of expertise. Cue-takers can pick and choose among the whole committee membership to find members whose attributes are most suitable, or with whom they have a close personal relationship. Uslaner and Weber (1977) find that for cue-seeking, legislators in state legislatures look for cues from committee leaders more often than they look for cues from party leaders and party caucuses. Expertise and committee membership are interrelated, given the opportunity that members on the committee had to attend the hearings and listen to the experts. Santos (2002) finds that patterns of committee appointments in the Brazilian Congress show a positive and significant association between deputies' previous specialization and their likelihood of being part of control committees in the Brazilian Congress.

When examining contemporaneous effects of shared location, it's challenging to distinguish cue-taking from other potential mechanisms, like social pressure and peer pressure. This difficulty arises

because both cue-taking and these other mechanisms exert influence only during social interactions and not afterward. Lowe and Jo (2024) add that legislators may take actions that conform to the neighbor's views to signal that they share an agreement or that they listen to the neighbor, perhaps to avoid stigma or conflict, and for the hedonic value of having a good relationship with neighbors. Also, in critical decisions where vote decisions are costly, legislators tend to exert peer pressure on their neighbors (Chan et al., 2019). In contested votes, individual vote becomes more important at the margin – since they have the power to change a roll call outcome. Therefore, legislators can be persuaded to deviate from their party's stance, to vote against their individual ideological position or their constituents preferences to favor a winning coalition.

1.3 Institutional Background

1.3.1 The Chamber of Deputies of Brazil

The Chamber of Deputies of Brazil[7] stands as a cornerstone of the country's democratic governance. Comprising elected representatives known as deputies, it serves as the lower house of the National Congress alongside the Federal Senate. This body embodies the essence of Brazil's representative democracy, charged with vital functions ranging from proposing and debating legislation to overseeing the use of public resources.

In total, 513 deputies are elected for a mandate of four year-term, coinciding with the legislature. The Chamber's composition reflects Brazil's regional diversity, with the number of deputies from each state determined by its population size[8]. From urban centers to remote rural areas, deputies bring the concerns and aspirations of their constituents to the legislative forefront.

Representatives are elected according to proportional electoral system. Brazil also follows an open list system. Unlike the closed list system where voters endorse a party's predetermined slate of candidates, the open list system allows to handpick individual candidates from within their chosen party or coalition. This electoral system results in a congress comprised of multiple political parties[9]. According to Ames (1995), this fragmented system prevents party leaders from exerting control over candidacies and, consequently, over party members' voting decisions within the congress. An opposite view defend that even though electoral laws may give politicians incentives

[7] In portuguese, *Câmara dos Deputados do Brasil*.

[8] Minimum of 8 representatives per State and the Federal District (e.g. Acre) and maximum of 70 representatives per State (e.g. São Paulo).

[9] In the 55th legislature, 28 parties held representation, while in the 56th legislature, this number increased to 30 parties.

to cultivate the personal vote and to defy the party line, individualistic behavior does not thrive in the milieu inside the Brazilian congress Figueiredo and Limongi (2000). They show by relying on roll call data that presidents have counted on reliable support on most of their propositions, the average level of discipline of the presidential coalition under their period of analysis is 85.6%. This evidence suggests that the legislative in Brazil functions on a coalition setting, in which presidents form governments, and the parties included in the governmental coalition provide political support for the president. More recently, Almeida (2018) revisited the president's legislative dominance over the period of 1989-2016 and proposed that its variation was caused by changes in the incentives of the legislators to delegate agenda power to the president and to participate in the legislative process, in the sense of producing laws of their own and controlling those originating from the Executive; and also that those incentives are associated with characteristics of parliamentary preferences and to the opportunity costs of these activities.

One of the Chamber's pivotal functions is its participation in the legislative process. Bills can originate in either the Chamber of Deputies or the Federal Senate, but it is within the Chamber where many legislative initiatives take shape. Proposed laws undergo meticulous scrutiny, with multiple readings, debates, and committee reviews shaping their trajectory. This rigorous process ensures that legislation reflects the diverse perspectives and interests of Brazil's populace before advancing to the Senate for further consideration. As a forum for democratic deliberation, the Chamber of Deputies fosters vibrant discourse and negotiation among its members. Across party lines, legislators engage in robust debates, championing their respective policy agendas and advocating for their constituents. Through the lens of proportional representation, the Chamber embodies the pluralism inherent in Brazilian society, facilitating dialogue and compromise to advance the common good. Moreover, the Chamber's role extends beyond legislative endeavors to encompass broader democratic functions. It serves as a check on executive power, holding the government accountable through oversight mechanisms and investigations. From scrutinizing budget allocations to examining policy implementation, legislators exercise their mandate to ensure transparency and accountability in governance.

1.3.2 The Office Lottery

At the beginning of each legislative session, a unique tradition unfolds within the Chamber – the office lottery. The procedure of the office lottery involves two urns. In the first urn, the names of the deputies who will participate in the lottery are placed. Then, one name is drawn from this urn.

Subsequently, in the second urn, the numbers corresponding to the available offices are placed. A number is drawn from this urn, and it is matched with the name drawn from the first urn. This process is repeated until all deputies have been assigned an office. This method ensures a fair and random allocation of offices among the deputies. For transparency, this procedure is broadcasted live on public television and made available on the Chamber's YouTube channel[10]. See Appendix A.1 for an illustration.

The criteria for the office lottery was established through the Act of Board No. 88, of 10/18/2006. Certain individuals are excused from the drawing process. These include former presidents of the Chamber of Deputies, persons with mobility impairments or special needs supported by a medical certificate from the House's Medical Department, individuals aged 65 or above by the beginning of the forthcoming Legislature, women, incumbents of the current legislature who have secured reelection, elected substitutes with a tenure of 365 days or more in the current Legislature, and former deputies who have served as titular members. Therefore, only first term male deputies participate in this ritual, being randomly allocated to vacant parliamentary offices. For example, in the 56th legislature, out of 262 newly elected deputies (19 of whom were former deputies), 154 actively participated in the lottery, totaling approximately 59%.

These offices serve as the operational hub for deputies and their teams, providing essential infrastructure to support their legislative endeavors. Within the main building of the Chamber, Annex I and II allocate high-ranking officials such as the current and former Speakers of the Chamber and most of the party leaders alongside with administrative offices and committee rooms. Complementing the main building are the Annexes III and IV, which offer additional office accommodations for legislators. These annexes are interconnected with the main building, ensuring seamless access for legislators and staff as they navigate their daily responsibilities. Among the deputies elected in the legislatures under analysis, all were assigned either to Annex III or Annex IV[11]. See their office plans in Appendix A.2.

Since this study examines the impact of proximity on legislative behavior, it utilizes these office plans to define a measure of shared location, *Office Neighbors*. It takes in account all surrounding offices – whether they are located adjacent to each other, either directly next to, in front of, or diagonally across from each other.

---

[10] https://www.youtube.com/live/N6P2_EsJUJo?si=TMv7Zm7TWzxgSakB

[11] In the 56th legislature, out of the deputies who took part in the lottery, 101 were randomly assigned to Annex IV, whereas 53 were randomly assigned to Annex III.

Table 1.1: Descriptive Statistics of Legislators in the 55th and 56th Legislatures

Panel A: 55th legislature (2015-2018)

	Population	Lottery Sample
Total Number	513	137 (=27%)
Number of Men	463	137
Number of Women	50	0
Mean Age	61	53

Panel B: 56th legislature (2019-2022)

	Population	Lottery Sample
Total Number	513	154 (=30%)
Number of Men	436	154
Number of Women	77	0
Mean Age	55	49

Table 1 presents the descriptive statistics of the population of legislators and of the sample of legislators that participated in the lottery in the 55th and 56th legislatures.

In the 55th legislature, there were a total of 513 legislators, consisting of 463 men and 50 women. The average age of legislators was 61 years old. The average age of the lottery sample is lower because it primarily consists of first-term legislators, who typically tend to be younger.

Moving to the 56th legislature, there were shifts in gender representation: the number of male legislators decreased to 436, while the number of female legislators increased to 77. Percentage of men decreased from 90% to 84%. The average age of legislators in this term decreased to 55 years old.

These figures indicate changes in gender composition and potentially a younger cohort of legislators in the 56th legislature compared to the 55th.

1.3.3 Voting

In the Chamber of Deputies of Brazil, two main types of voting procedures are utilized to make legislative decisions. These procedures vary depending on the nature of the proposed legislation and the level of consensus required among legislators.

Voice vote or acclamation is a voting method in which a group vote is taken on a topic or motion

by responding vocally. Standing voting is a method where legislators express their vote by raising their hands or standing up rather than verbally stating their vote individually. These methods are commonly used for routine matters or procedural votes where there is a high level of consensus among legislators, and individual voting records are not required. They allow for a quicker and more informal process compared to roll call voting. However, it lacks the transparency of roll call voting as individual legislators' votes are not recorded.

Recorded or roll call voting[12] is a method where each legislator expressly announces their vote. This method is used for more significant legislative decisions or when there is a need for individual accountability and transparency. Roll call allows constituents and the public to know how each legislator voted on a particular issue.

In general, the choice between these methods depends on the nature of the vote and the level of transparency and accountability desired by the legislative body.

A vote is always related to a legislative proposal that can be initiated by any legislator in the Chamber. Legislative proposals differ in their scope, procedural requirements, approval thresholds, and consequences. Within the same legislative proposal, there could exist more than one voting procedure (for instance, while voting a proposal for tax reform, legislators might vote for the final passage of the proposal and to amend the text of the proposal). See Appendix A.2 for the description of the main type of legislative proposals and for the the distribution of votes over them across the two legislatures under analysis.

In the Chamber, legislators can cast their votes as "yes" (*sim*), "no" (*não*), "abstain" (*abstenção*), or "obstruction" (*obstrução*). Additionally, legislators may also be marked as "absent" (*ausente*) if they are not present to cast their vote. Obstruction typically refers to a procedural tactic taken by legislators, often aligned with their party, rather than an individual voting decision. It is taken to intentionally delay or impede the legislative process. This type of action is often used strategically by political parties or blocs to achieve certain objectives or to express opposition to particular measures being considered. Thus, in this analysis, votes marked as "obstruction‘ are not considered – since the paper's objective is to analyse individual legislative behavior.

Table 2 illustrates the distribution of recorded votes across the two legislatures. Approximately 65% of recorded votes are categorized as "yes" or "no," while absences account for 31% of the total recorded votes.

---

[12]In portuguese, *votação nominal*.

Table 1.2: Distribution of recorded votes — 55th and 56th legislature — Chamber of Deputies of Brazil.

Type	Freq.	Percent
Yes	241,825	33.4
No	231,001	32
Obstruction	24,025	3.3
Abstain	2,211	0.3
Absence	224,606	31
Total	723,668	100

1.3.4 Political Debate

According to the *Regimento Interno* (en. Internal Rule), there are seven possibilities for pronouncements: 1) when presenting a proposition; 2) during the *Pequeno* and *Grande Expediente* (en. Short and the Long shift) or during the phase of *Comunicações Parlamentares* (en. Parliamentary Communications); 3) when discussing measures being voted; 4) to raise a point of order; 5) to protest; 6) to send a vote; and, 7) for self-defense in case of an accusation considered undue (at the discretion of Speaker of the House). Such possibilities focus on different moments of the legislative process. See Santos et al. (2021).

Floor debates are part of the ordinary sessions of the chamber and comprise four phases: Short Shift, Long Shift, Order of Business, and Parliamentary Communications. Each of these phases is also carefully regulated by the Internal Rule (Regimento Interno), which mainly addresses deputies' speeches.

It is clear from the presentation of regulations on the chamber's speech policies that leaders have almost absolute control over the use of speech. Hence, there are two moments where representatives can intervene in the debates more freely and systematically: a) on the floor, during the Short Shift; and b) in committees, during the Order of Business, when discussing the policies under consideration.

1.4 Empirical Strategy

1.4.1 Data

This paper uses data from both the 55th and 56th legislatures spanning from 2015 to 2022. These include the Dilma (2015-2016) and Temer (2016-2018) presidencies, as well as the Bolsonaro (2019-

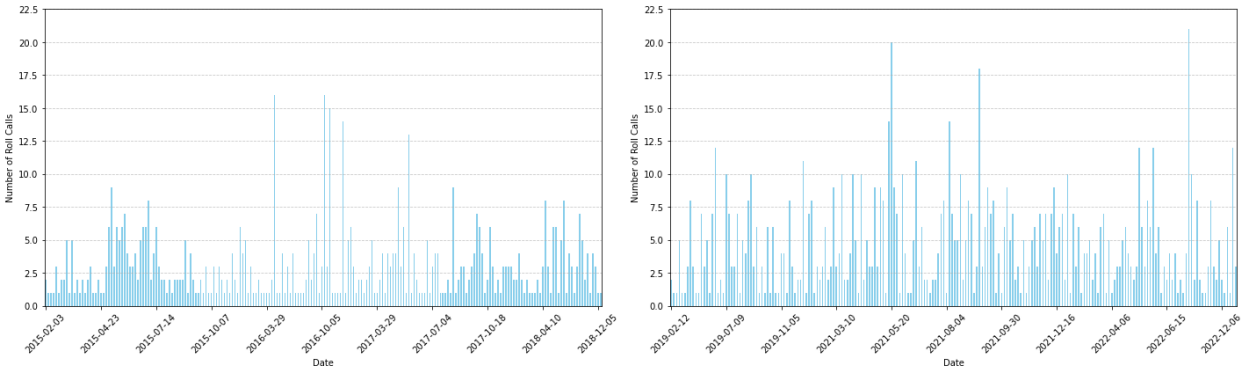
2022) presidency. The data was obtained by web scraping through the *Dados Abertos* (en. Open Data) platform[13] of the Chamber of Deputies of Brazil. Due to the empirical strategy's reliance on physical social interactions to understand peer effects, data from periods under the *Sistema de Deliberação Remota* (SDR) (en. Remote Deliberation System) were excluded[14][15].

Table 3 outlines the descriptive statistics, including the count of legislators participating in the lottery, the number of roll calls, and the number of speeches for each legislature. Figures 1 and 2 present the histogram of the daily distribution of roll calls and speeches, respectively[16].

Table 1.3: New members and activities in the 55th-56th legislatures

Legislature	No. of Legislators in the Lottery	No. of roll calls	No. of speeches
55th	137	738	17.016
56th	154	658	4.189

Figure 1.1: Number of Roll Calls per Day - 55th and 56th legislature



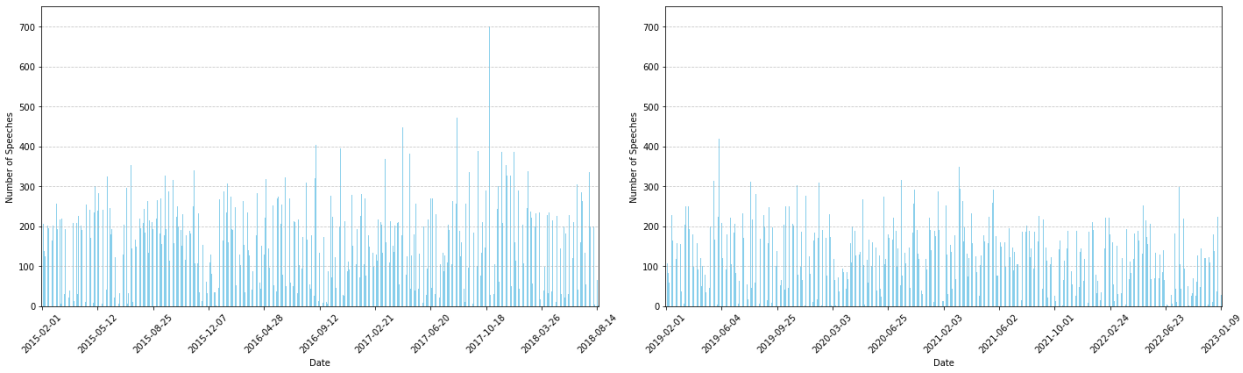
[13] https://dadosabertos.camara.leg.br/

[14] The COVID-19 pandemic caused significant disruptions to legislative operations. Beggining on March 17th, 2020, the Chamber of Deputies enacted the SDR, shifting all plenary activities, including roll calls, to a remote format and temporarily closing permanent committees. This arrangement persisted until February 11th, 2021, when a hybrid system was adopted, permitting legislators to reconvene in person in the plenary and restoring permanent committees to operation. This hybrid model continued until October 25th of the same year, when activities resumed normal operations. During a brief period in 2022, from the start of the legislative year in February until April 18th, the Chamber reinstated the SDR.

[15] This subset of data is used in section 4.1.1. for a robustness check.

[16] Data covering the periods under the SDR were included in these histograms for reference, even though they are not present in Table 3 and in the remaining part of this analysis.

Figure 1.2: Number of Speeches per Day - 55th and 56th legislature



The observed differences in the frequency of roll calls and speeches between the two legislative sessions possibly reflect variations in the political strategies, priorities, and coalition dynamics during the respective governments.

Additionally, significant events potentially affected the legislative behavior over these two legislatures. First, due to the presidential impeachment process during the 55th legislature, the Chamber's proceedings were significantly disrupted, leading to a notable impact on floor activity. The Chamber officially commenced discussions on the impeachment on December 2, 2015. Ultimately, on April 17, 2016, the legislators voted to proceed with the impeachment. As the impeachment proceedings unfolded, parliamentary activities were heavily focused on debates, committee hearings, and negotiations related to the impeachment vote. This intense focus on impeachment-related matters likely diverted attention and resources away from other legislative priorities, potentially leading to fewer roll calls on unrelated legislation during this period. Moreover, the disparity between the two legislative bodies can also be attributed to the influence of social distancing policies, including the adoption of the SDR and the closure of permanent committees, which significantly affected legislative activities during the 56th legislature.

1.4.2 Variables

Office Neighbors

Since the identification strategy relies on the random office allocation, among all pairs of legislators only those that contain at least one lottery participant are considered. There are two arguments for this restriction. The first is econometric: pairs containing two incumbents show no independence

between their average potential outcomes and the treatment. The second is more intuitive: newly elected legislators have less established relationships within Congress and are more likely to form stronger ties with their office neighbors, whether those neighbors are other freshmen or incumbents.

Then, letting ij index pairs of legislators, the sample D_t is restricted to pairs in which either the legislator i or the legislator j (or both) was subject to the office lottery. Therefore, I focus on the subset of pairs of legislators $D_t \in U_t$, where $U_t = L_t \times L_t = \{(i,j) \mid i \in L_t \text{ and } j \in L_t\}$ is the set of all possible combinations of pairs of legislators of the legislature t and L_t is the set that contains all legislators of the legislature t.

To explore the peer effects of office neighbors, I define the main independent variable $\textit{Office Neighbors}_{ijt}$ as an indicator function of whether the pair of legislators ij shares the same office neighborhood in legislature t, with i never having been taken as a neighbor of herself. Three different variations of the office neighborhood are employed. Starting from the most flexible one, $\textit{Office Neighbors}_1$, is one that defines an office neighborhood by taking into account all surrounding offices – if they are located adjacent to each other, either directly next to, in front of, or diagonally across from each other. The second definition, $\textit{Office Neighbors}_2$, considers offices adjacent or directly in front of each other. In the last and most parsimonious definition, $\textit{Office Neighbors}_3$, only offices adjacent to each other are considered neighbors. The first definition is the most comprehensive and is taken as the main measure of office proximity.

Figure 3 illustrates the first measure and also provides a general view of a neighboring location. If we select an office i (shown in dark blue), any office j located in the light blue offices is considered a neighbor, while any office j located in the red offices is not. Therefore, an office i can potentially have up to 5 neighbors j (given that $ij \in D_t$)[17].

Thus, for every pair of legislators $ij \in D_t$ with $j \neq i$ and $t \in \{55, 56\}$, I set up the main independent variable $\textit{Office Neighbors}_{ijt}$ as[18]:

$$\textbf{\textit{Office Neighbors}}_{ijt} = \begin{cases} 1 & \text{if legislators } i \text{ and } j \text{ share the same office neighborhood in legislature } t \\ 0 & \text{otherwise} \end{cases}$$

[17] If an office i is located at the end of the hallway, it will have fewer neighbors (3 instead of 5).

[18] All duplicate pairs of the form ij and ji are removed from the sample

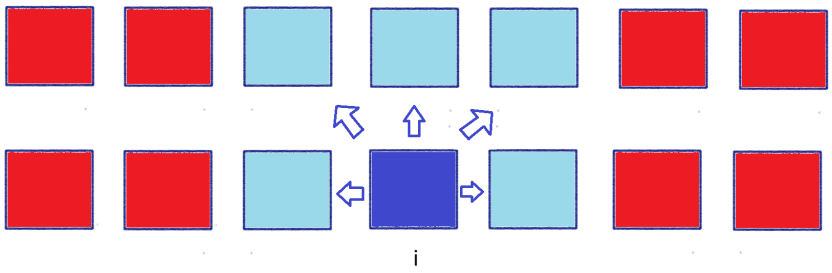


Figure 1.3: Neighboring Locations

Measures of Legislative Behavior

To explore the impact of spatial proximity on legislative behavior, I analyze all pairs of legislators $ij \in D_t$ per roll call p, in which legislators vote and are present. Thus, in the main analysis, I take into account only the "yes" and "no" vote decisions. See Appendix A.4 for an alternative analysis, which takes into account "yes", "no", "abstain" and "abstention" votes.

As a measure of legislative behavior, I define the agreement score $Convergence_{ijpt}$ as an indicator function for whether the pair of legislators ij cast the same vote on the roll call p during the legislature t.

Thus, for every pair of legislators $ij \in D_t$ voting on roll call p with $j \neq i$, $p \in R_t$ and $t \in \{55, 56\}$; where R_t is the set that contains all the roll calls of the legislature t. I set *Convergence*

Table 1.4: Pair-Level Summary Statistics

	$Convergence_1$ (only "Yes" and "No" votes)				
Variable	Mean	St.Dev.	Minimum	Maximum	N
Convergence	0.690	0.462	0	1	41,400,152
Office neighbors$_1$	0.0078	0.088	0	1	41,400,152
Office neighbors$_2$	0.0045	0.067	0	1	41,400,152
Office neighbors$_3$	0.0027	0.052	0	1	41,400,152
Same party	0.0642	0.245	0	1	41,400,152
Same coalition	0.599	0.489	0	1	41,400,152
Same state	0.0652	0.246	0	1	41,400,152
Same freshman status	0.173	0.378	0	1	41,400,152
Same gender	0.934	0.246	0	1	41,400,152
Ideological Distance	0.526	0.472	0	1.89	41,400,152
Age diff.	13,66	9.84	0	62	41,400,152

$_{ijpt}$ as:

$$\boldsymbol{Convergence}_{ijpt} = \begin{cases} 1 & \text{if legislators } i \text{ and } j \text{ cast the same vote on roll call } p \text{ during legislature } t \\ 0 & \text{otherwise} \end{cases}$$

Table 4 provides summary statistics for the main analysis and all main variables. Same party is an indicator for whether a pair of legislators is from the same party at the time of the proposal's vote. Same coalition is an indicator for whether a pair of legislators is from the same coalition at the time of the proposal's vote. Same state is an indicator for whether a pair of legislators is from the same state at the time of the proposal's vote. Same gender is an indicator for whether a pair of legislators shares the same gender. Ideological distance is defined by taking the absolute difference between the Brazilian Legislative Survey scores (Zucco, 2023) of the members of each pair, with ideological positions being assigned according to party membership. Age difference is the absolute difference between the age of the members of each pair.

1.4.3 Identification

Since the Internal Rule allows for office-swapping (after the lottery is run and the first term male legislators are randomly allocated), non-compliance cannot be ruled out. Thus, the empirical strategy follows an intent-to-treat analysis (ITT). This is called intent-to-treat analysis because it

measures the causal effect of intended treatments, rather than the treatment outcomes. Since not all legislators who are randomly assigned to offices actually remain there, we should expect the ITT effect to be strictly smaller than the average treatment effect.

With this pair-wise sample D_t and variables, the intent-to-treat (ITT) estimate, β_1, can be obtained from the following specification:

$$\text{Convergence}_{ijpt} = \beta_0 + \beta_1 \cdot \text{Office neighbors}_{ijt} + \mathbf{B_2} \cdot \mathbf{X_{ijt}} + \mu_t + \delta_p + \epsilon_{ijpt} \tag{1.1}$$

Where $\mathbf{X_{ijt}}$ is a vector of covariates (including ideological distance, same party or coalition, same state, same freshman status, same gender and age difference), μ_t represents legislature fixed effects, δ_p denotes fixed effects for legislative procedural voting and ϵ_{ijpt} is the error term. It is assumed that *Office neighbors*$_{ijt}$ is orthogonal to the error term ϵ_{ijpt} given the random office allocation.

Standard errors are two-way cluster-robust, taking into account the correlation between pairs (i, j) and (i', j') when i = i' or j = j'.

I use randomization inference to calculate Fisher's exact p-values. I simulate placebo seating assignments following the exact procedure of the Chamber for assigning offices. The advantage of randomization inference is that it provides an exact test against the sharp null hypothesis of no treatment effects without relying on asymptotic assumptions (Imbens and Rubin, 2015a).

It is important to note that the parameter of interest, β_1, measures the influence of office proximity within a pair – that is, influence could potentially go in both directions (from deputy i to deputy j and/or from deputy j to deputy i). If the influence is positive (i.e., deputies located into the same office neighborhood positively influence themselves), their behavior converge.

The covariate balance tests are shown in Table 5. I replace the left-hand variables of the main specification with other pre-determined characteristics of the legislators' pair.

The results show statistically significant effects for same gender and age difference. A potential explanation for these stronger effects is that women can self-select in specific office neighborhoods (since they do not participate in the lottery), thus forming clusters within the floors. Thus, it is more likely for the treatment to select pairs containing two men. Also, since the sample is restricted only to pairs containing at least one participant in the lottery, and those are usually younger, the likelihood of the treatment selecting pairs with smaller age differences is greater. The older incumbent also can self-select into specific office neighborhoods. By controlling for floor fixed effects, these point estimates are no longer statistically significant. The joint test for orthogonality

Table 1.5: Covariate Balance Table

	Dependent variable:					
	Same party	S. coalition	S. state	S. gender	Id. dist.	Age diff.
Office neighbors	-0.68	0.21	-0.02	1.91**	-0.0189	-0.8155***
	(0.8)	(1.6)	(0.8)	(0.9)	(0.016)	(0.310)
Observations	41,400,152	41,400,152	41,400,152	41,400,152	41,400,152	41,400,152

Notes: Two-way cluster-robust standard errors. Columns (1)-(6) report covariate balance tests using the covariates as outcomes. *Office neighbors* is an indicator of whether a pair of legislators belong to the same office neighborhood in a specific legislature. *Same party*, *Same coalition*, *Same state*, *Same gender*, and *Age difference* are defined analogously. *p<0.1; **p<0.05; ***p<0.01.

also shows no statistical significance; it suggests that these covariates are jointly orthogonal to the treatment, meaning that differences in covariates are likely due to random variation rather than selection biases. Therefore, this empirical evidence shows that the office lottery was correctly randomized.

1.5 Empirical Results

1.5.1 Estimation of Peer Effects

Table 6 displays the results of the ITT analysis. All standard errors, reported in parentheses, are two-way cluster-robust.

In columns 1, 2 and 3 the spatial proximity variable is *Office Neighbors*$_1$, which is one that defines an office neighborhood by taking in account all surrounding offices – if they are located adjacent to each other, either directly next to, in front of, or diagonally across from each other. Column 4 uses, for the spatial proximity effect, the second definition, *Office Neighbors*$_2$, in which offices adjacent or directly in front of each other are considered neighbors. Column (5) uses the most parsimonious definition, *Office Neighbors*$_3$, in which only offices adjacent to each other are considered neighbors.

In the first column (1), the main specification is estimated without including the set of covariates and the voting fixed effects (it only includes legislature fixed effects), resulting in a point estimate of 0.83 percentage points and statistically non significant. Moving to column 2, covariates and voting fixed effects are included, it similarly presents a statistically non-significant result with a smaller point estimate of 0.26 p.p. In columns 3, the covariate same party is replaced by same coalition and the point estimate remains close to the one observed in column (2). For robustness, columns

Table 1.6: Pair-Level Effects on Voting: Main Analysis (p.p.)

	Dependent variable: Convergence				
	(1)	(2)	(3)	(4)	(5)
Office neighbors	0.83	0.26	0.21	0.07	0.19
	(0.8)	(0.6)	(0.6)	(0.8)	(0.9)
Ideological distance		-28.0***	-28.0***	-28.0***	-28.0***
		(1.30)	(1.30)	(1.30)	(1.30)
Same party		3.30***		3.30***	3.30***
		(0.8)		(0.8)	(0.8)
Same coalition			1.43***		
			(0.4)		
Same state		0.15	0.22	0.15	0.15
		(0.4)	(0.4)	(0.4)	(0.4)
Same gender		0.73	0.62	0.73	0.73
		(0.9)	(0.9)	(0.9)	(0.9)
Age difference		-0.07***	-0.07***	-0.07***	-0.07***
		(0.4)	(0.4)	(0.4)	(0.4)
Voting FE	No	Yes	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes	Yes	Yes
Observations	41,400,152	41,400,152	41,400,152	41,400,152	41,400,152
Number of roll calls	1,396	1,396	1,396	1,396	1,396
Outcome Mean	0.690	0.690	0.690	0.690	0.690

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. The outcome, *Convergence*, is an indicator of whether the pair of legislators agreed in the roll call vote. Column (1)-(5) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement. Column (1) uses *Office Neighbors*$_1$ to define the spatial proximity effect, which is one that defines an office neighborhood by taking in account all surrounding offices – if they are located adjacent to each other, either directly next to, in front of, or diagonally across from each other. It does not include any covariates and only controls for congress fixed effects. Column (2) also uses *Office Neighbors*$_1$ and includes the complete set of covariates plus voting fixed effects. Column (3) keeps *Office Neighbors*$_1$ and the same specification as column (2) but the same party variable – it replaces this covariate by same coalition. Column (4) uses the same specification as column (2) but for the spatial proximity effect it considers the second definition, *Office Neighbors*$_2$, in which offices adjacent or directly in front of each other are considered neighbors. Column (5) uses the same specification as column (2) but for the spatial proximity effect it considers the most parsimonious definition, *Office Neighbors*$_3$, in which only offices adjacent to each other are considered neighbors. Ideological distance is defined by taking the absolute difference between the BLS scores of the members of each pair. *Same party* is an indicator for whether a pair of legislators is from the same party at the time of the proposal's vote. *Same coalition* is an indicator for whether a pair of legislators is from the same coalition at the time of the proposal's vote. *Same state* is an indicator for whether a pair of legislators is from the same state at the time of the proposal's vote. *Same gender* is an indicator for whether a pair of legislators shares the same gender. *Age difference* is the absolute difference between the age of the members of each pair. *p<0.1; **p<0.05; ***p<0.01

4 and 5 show results for the same specification as column 2, but use alternative definitions of office neighbors. Despite this adjustment, the ITT estimate do not present great difference and remains statistically non-significant

The results show that, on average, office proximity increases agreement by 0.26 percentage points. This effect amounts to 8 p.p. of the influence of same party membership. The point estimate is consistent with other studies analyzing the influence of spatial proximity on co-voting behavior. Rogowski and Sinclair (2012), using data from the U.S. House of Representatives, also found not statistically significant results and their point estimate lies within our confidence interval. Darmofal et al. (2023) analyze roll call voting under random seating assignment in the Iceland Parliament and also find null effects. Lowe and Jo (2024) point estimate, using a specification that emphasizes spatial proximity effects on cross-party pairs and data from the Iceland Parliament, also lies within our confidence interval range. Harmon et al. (2019) estimated, using data from the European Parliament, a 0.6 p.p. effect of spatial proximity on co-voting behavior – which is also covered by our CI.

Although I find an average treatment effect that is small and not statistically distinguishable from zero, most votes in the sample are cast on procedural matters, where there is little variation, as legislators typically toe the party line or follow their ideological position. It might be that these influences are potentially masking the effects of spatial proximity on the voting convergence within pairs of legislators. Therefore, it is important to study the potential heterogeneous effects of spatial proximity to help uncover some potential dimensions of stronger influence.

1.5.2 Heterogeneity in Peer Effects

Heterogeneity analysis is first ran taking into account the importance of voting by emphasizing non-procedural decisions. One example of non-procedural voting is disputed votes, where roll calls have narrow victory or defeat margins. Table 7 shows empirical evidence supporting the direct influence of office-mates on voting agreement in contested decisions.

Votes for highly contested roll calls, with a margin of less than 5 p.p., show a statistically significant estimate of 2 percentage points – roughly eight times the magnitude of the average effect. They are also robust to randomization inference. With 400 draws, I estimate a Fisher's exact p-value of 0.0575, see Appendix A.7.

Table 8 shows spatial proximity effects in co-voting behavior on roll calls with different result margins. It categorizes result margins into five groups: greater than 10 p.p., less than 50 p.p., less

Table 1.7: Pair-Level Effects on Voting: Result Margin Heterogeneity (p.p)

	Dependent variable: Convergence	
	$> 5\%$	$< 5\%$
Office neighbors	0.20	2.0***
	(0.6)	(1.0)
Controls	Yes	Yes
Voting type FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	40,052,049	1,308,783
Number of roll calls	1,359	36
Outcome Mean	69.1	49.7

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Convergence is an indicator of whether the pair of legislators agreed in the proposal's vote. Office neighbors is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Column (1) reports the proximity effect for roll calls with more than 5 p.p. result margin. Column (2) reports the proximity effect for roll calls with less than 5 p.p. result margin Covariates included are same party, ideological distance, same state, same gender, and age difference. These variables are self-explanatory. *p<0.1; **p<0.05; ***p<0.01

Table 1.8: Pair-Level Effects on Voting: Result Margin Heterogeneity (p.p.)

	Dependent variable: Convergence$_1$			
	$> 10\%$	$< 50\%$	$< 25\%$	$< 10\%$
Office neighbors	0.6	1.3	2.0***	1.5**
	(0.7)	(1.0)	(0.7)	(0.7)
Controls	Yes	Yes	Yes	Yes
Proposal type FE	Yes	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes	Yes
Observations	28,908,654	17,062,992	6,381,415	2,498,300
Outcome Mean	69.1	56.3	51.0	49.9

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Columns (1)-(4) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement in roll calls with different margin of victory/defeat. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office neighbors* is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables are *same party*, *same state*, *ideological distance*, *same gender*, and *age difference*. These variables are self-explanatory. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

than 25 p.p., less than 10 p.p., and less than 5 p.p.

When the result margin is greater than 10 p.p., the point estimate is 0.6 p.p., closer to the point estimate of the average effect in voting agreement. Mainly due to the predominance of comfortable and procedural votes in the sample, it suggests that legislators are less swayed by their office neighbors in uncontested voting situations.

For result margins less than 50 p.p., the impact of proximity increases the likelihood of agreement to 1.3 p.p. Though statistically insignificant, the effect size is relatively modest compared to smaller result margins. For winning margins less than 25 p.p., the probability increases to 2 p.p., suggesting a more substantial impact of office neighborhood proximity on voting behavior. This effect is statistically significant at the 1 p.p. level. For result margins less than 10 p.p., the probability decreases to 1.5 p.p., but remains statistically significant at the 5 p.p. level. Although smaller in magnitude, the effects of spatial proximity remain robust with larger result margins (10 p.p. and 25 p.p.). If we consider an agreement score that takes into account absences and abstain votes, the estimates stay consistent (see table 18 in the Appendix A.4).

This evidence suggests that office proximity, functioning as a social structure, plays an important

role in non-procedural decisions, complementing other sources of influence, whether ideological or non-ideological (Zucco and Lauderdale, 2011). Evidence of spatial proximity effects on contested decisions was identified in earlier studies. Harmon et al. (2019) estimated that the peer effects in these close votes are about twice those found for "comfortable" vote effects. Young (1966) revealed that the more evenly divided the House sentiment and the more closely contested the issue, the greater the reliance upon messmates for political cues. Masket (2008) and Lowe and Jo (2024) find stronger deskmate influence on contested votes than detected in lopsided votes.

Legislators argue that in more complex decisions, they pick informants with some expertise on the legislation (Kingdon, 1989; Matthews and Stimson, 1975). The committee collectively and its members individually are potent cue-givers for the obvious reason of expertise. Expertise and committee membership are interrelated, given the opportunity that members of the committee had to attend the hearings and listen to the experts. Santos (2002) finds that the patterns of committee appointments in the Brazilian Congress show a positive and significant association between the previous specialization of the deputies and their likelihood of being part of control committees in the Brazilian Congress.

Thus, I further explore the relationship between office proximity and committee membership to explain influence in co-voting patterns. Equation (2) presents a specification that interacts the office proximity variable, Office neighbors$_{ijt}$, with an indicator of whether a legislator within a pair is a member of the committee related to the subject of the roll call voted on, Committee$_{ijt}$:

$$\begin{aligned} \text{Convergence}_{ijpt} = \beta_0 &+ \beta_1 \cdot \text{Office neighbors}_{ijt} + \beta_2 \cdot \text{Committee}_{ijt} \\ &+ \beta_3 \cdot \text{Committee}_{ijt} \times \text{Office neighbors}_{ijt} \\ &+ \mathbf{B_4} \cdot \mathbf{X_{ijt}} + \mu_t + \delta_p + \epsilon_{ijpt} \end{aligned} \tag{1.2}$$

Where $\mathbf{X_{ijt}}$ is a vector of covariates (including ideological distance, same party or coalition, same state, same freshman status, same gender and age difference), μ_t represents legislature fixed effects, δ_p denotes fixed effects for legislative procedural voting and ϵ_{ijpt} is the error term. It is assumed that *Office neighbors*$_{ijt}$ is orthogonal to the error term ϵ_{ijpt} given the random office allocation.

Table 9 shows that the point estimate of the coefficient of the interaction term (β_3) is 4.51 percentage points for roll calls with a result margin of less than 5 p.p. – and remain robust with a result margin of less than 10 p.p. The results remain robust under randomization inference. With 400 draws, I estimate a Fisher's exact p-value of 0.0267, see see Appendix A.7.

Since committee assignments are endogenously defined, there is a potential concern that unob-

Table 1.9: Pair-Level Effects on Voting: Interaction of Spatial Proximity and Committee Membership (p.p.)

	Dependent variable: Convergence	
	$> 5\%$	$< 5\%$
Office neighbors	0.24	1.04
	(0.6)	(1.0)
Committee	-0.58	0.04
	(0.2)	(0.2)
Committee $\times$ Office neighbors	-0.17	4.51***
	(0.5)	(1.7)
Controls	Yes	Yes
Voting FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	39,556,744	1,308,783
Number of roll calls	1,347	36
Outcome Mean	69.1	49.7

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Convergence is an indicator of whether the pair of legislators agreed in the proposal's vote. Office neighbors is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Committee is an indicator of whether one legislator within a pair is member of the committee related to the subject of the roll call being voted. Column (1) reports the proximity effect for roll calls with more than 5 p.p. result margin. Column (2) reports the proximity effect for roll calls with less than 5 p.p. result margin Covariates included are same party, ideological distance, same state, same gender, and age difference. These variables are self-explanatory. *p<0.1; **p<0.05; ***p<0.01

served factors influencing committee membership could also affect legislative behavior, introducing bias into the analysis. To address this issue, I restricted the analysis to committee members selected at the beginning of each legislative session, ensuring that membership is less likely to be influenced by evolving legislative dynamics. Even without this restriction the results remain robust, supporting the reliability of the findings. Additionally, I created an indicator variable to identify pairs of legislators who both served on the same committee. The interaction between this indicator and the office neighbors variable was tested to assess whether shared committee membership among office neighbors influenced outcomes. This lack of significance indicates that the observed effects are likely driven by information flows from more informed legislators, rather than by simple co-membership in a committee.

This empirical evidence suggests that when confronted with the necessity of making a quick low-information decision on legislation, they can look to individuals who have made up their minds early and are perceived to be experts – therefore, committee membership and office neighborhood are complements on contested decisions. These experts usually are members of the standing committee that reported the measure to the floor. Legislators coped with uncertainty about the consequences of voting one way or the other by using the voting intentions of trusted, expert colleagues to decide how to vote (Uslaner and Weber, 1977). These findings demonstrate the importance of spatial proximity in the diffusion of information and emphasize the informational role of committees in Congress.

If committee membership shapes the flow of information in Congress, it raises questions about its potential impact on policy-making – especially on contested decisions, where the influence is greater and the likelihood of votes altering the roll call outcome is higher. The asymmetry of information naturally brings loss to the floor, so that committee members have greater influence over the design of the bill. This cost will be compensated for by the informational gains due to the reduction of uncertainty. Pereira and Mueller (2004) adapts Gilligan and Krehbiel (1987) model of strategic communication to the Brazilian Congress and show the committees that are more representative of the floor that would have a greater effect of reducing uncertainty in equilibrium. On the other hand, if all the cue-givers in a policy domain tend to be a biased sample of the Congress, the probability of unrepresentative policy decisions is very substantial. To illustrate this, Appendix A.5 uses the Brazilian Legislative Survey's (BLS) ideological estimates (Zucco, 2023) to compare the standing committees medians to the floor median. It shows that, in that sample, control committees (CCJC and CFT) are more representative of the floor compared to thematic committees (as exemplified by

CCULT). Therefore, this exercise highlights the potential for standing committee memberships to significantly influence policy-making, as committees that shape the flow of information may drive decisions in ways that reflect their own ideological biases, potentially leading to policy outcomes that are less representative of the broader legislative body.

Another potential source of treatment effect heterogeneity to explore is the variation across the roll calls' topics. Each roll call is categorized by the Chamber of Deputies into one of 28 possible topics[19]. Thus, I also analyze how the treatment effects varied across different roll call topics. Figure 4 shows the estimates of the Office Neighbors' coefficient using the main specification with all controls and fixed effects included. It shows that the treatment effects – the effect of belonging to the same office neighborhood on voting convergence – present a different size and direction depending on the topic being voted on. In particular, roll calls related to the topic of Social Security & Welfare show a relatively larger and statistically significant heterogeneous treatment effect of 3.4 p.p. The heterogeneous treatment effects for roll calls on the topic of Labor & Employment also present statistically significant effects of 1.52 p.p. On the other hand, the topic Science, Technology & Innovation presents an inverse relationship of our previous estimates – the same is true for other topics, such as Consumer Rights & Defense, Civil Law & Civil Procedure, and Education. All other heterogeneous treatment effects are not distinguishable from zero – or just marginally.

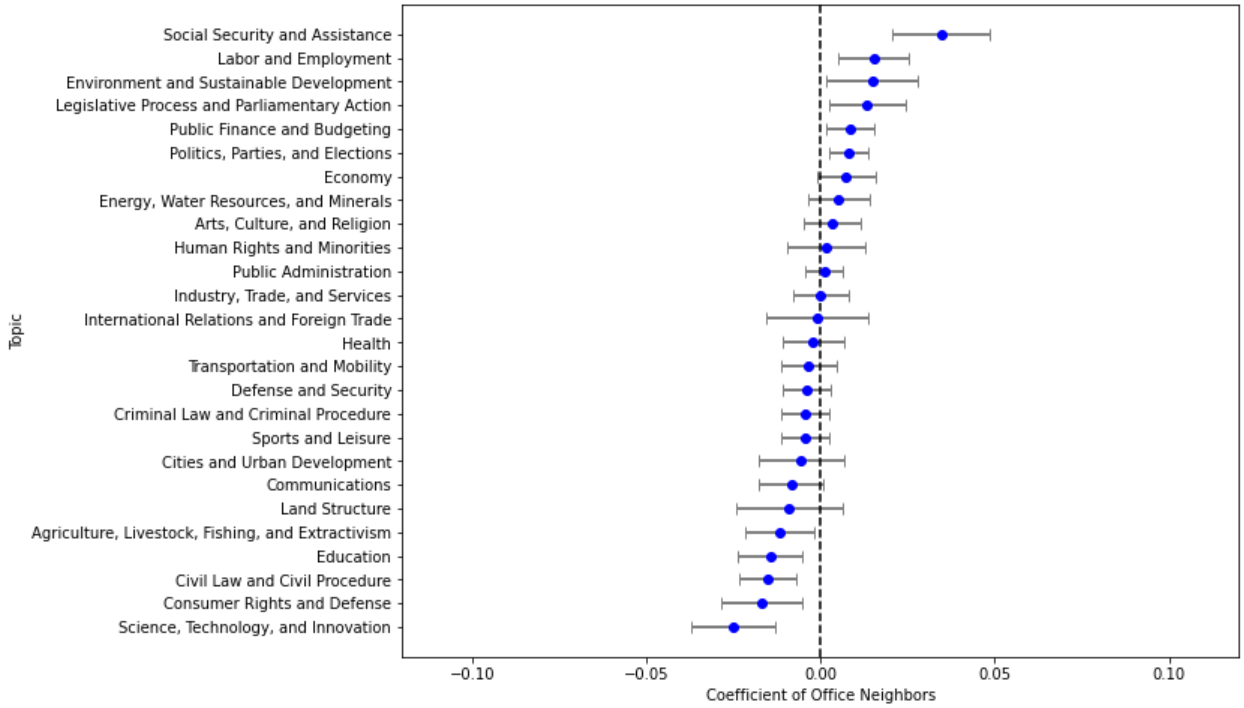
These findings show that for the two legislatures under analysis, heterogeneous treatment effects are stronger on topics related to labor and social security. The next step of the analysis is to explore other potential dimensions of influence within each topic to understand what could be driving these larger effects. Thus, in Appendix A.3, the distribution of topics by types of legislative proposal is shown. The topic that presented the largest treatment effect, Social Security & Welfare, is composed of about 90 p.p. of Provisional Measures (MPV)[20]. The second one, Labor & Employment, is composed of approximately 63 percentage points of Provisional Measures. These percentages are comparatively higher to a global average of approximately 30 p.p. of Provisional Measures per topic. The Provisional Measure is a type of executive order with the force of law immediately upon issuance, but subject to review and approval by the Chamber. Therefore,

According to Kingdon (1989), bargaining in the legislative may involve personal favor-trading,

[19]Only 26 topics are shown in Figure 4 and table 19. The topics of Law & Justice and Tributes & Commemorative Dates were removed due they presented very few observations and due to a limited number of observations – and abnormal standard errors - and very low F statistics.

[20]In particular, the set of roll calls on the topic of Social Security & Welfare involves 5 different Provisional Measures: MPV 665/2014, MPV 676/2015, MPV 780/2017, MPV 1099/2022, and MPV 1113/2022.

Figure 1.4: Heterogeneity analysis on roll calls' topics



where a vote becomes a credit which can be called in later. In multiparty presidential systems such as Brazil, an executive must exchange robustly with the legislative branch by using the influence of pork and patronage. The Brazilian executive controls the disbursement of pork to legislators through the execution of individual and collective budgetary amendments, and determines the proportionality of partisan representation within the cabinet. One significant challenge posed by this system is its multitude of potential partners that turn the office neighborhood into a potential network of influence for the executive. This social structure could help reduce the transaction costs associated with forming winning coalitions. Since Provisional Measures are the executive's primary legislative tool, they serve as the traditional means for the executive to direct policy, further emphasizing the role of these networks in shaping legislative outcomes. Thus, one can hypothesize that party and government leaders could exploit this spatial proximity to alter voting outcomes – and this effect could be perceived through the patterns of co-voting on Provisional Measures.

In table 10, I test this hypothesis and investigate how type of proposition heterogeneity influences pair-level voting behavior among legislators. It presents the analysis divided into seven

Table 1.10: Pair-Level Effects on Voting: Proposition Type Heterogeneity (p.p.)

	Dependent variable: Convergence						
	PL	PEC	PLP	MPV	PDC	PRC	PDL
Office neighbors	0.5	0.8	0.9	1.1	0.3	0.3	-0.3
	(0.6)	(1.0)	(0.9)	(1.1)	(0.9)	(2.4)	(1.1)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	19,975,715	6,665,089	3,297,855	10,106,397	1,121,658	53,656	179,782
Outcome Mean	69.2	68.9	68.6	68.7	68.6	75.7	73.4

Notes: All point estimates are presented in percentage points. Standard errors (in parentheses) are two-way cluster-robust. Columns (1)-(7) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement across different types of legislative proposals. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office neighbors* is an indicator for whether a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables include *same party*, *same state*, *ideological distance*, *same gender*, and *age difference*. *p<0.1; **p<0.05; ***p<0.01.

categories[21], respectively: PL (Proposed Law), PEC (Proposed Constitutional Amendment), MPV (Provisory Measure), PLP (Proposed Complementary Law), PDC (Proposed Decree of Congress), PRC (Resolution Project), and PDL (Proposed Decree of Legislative).

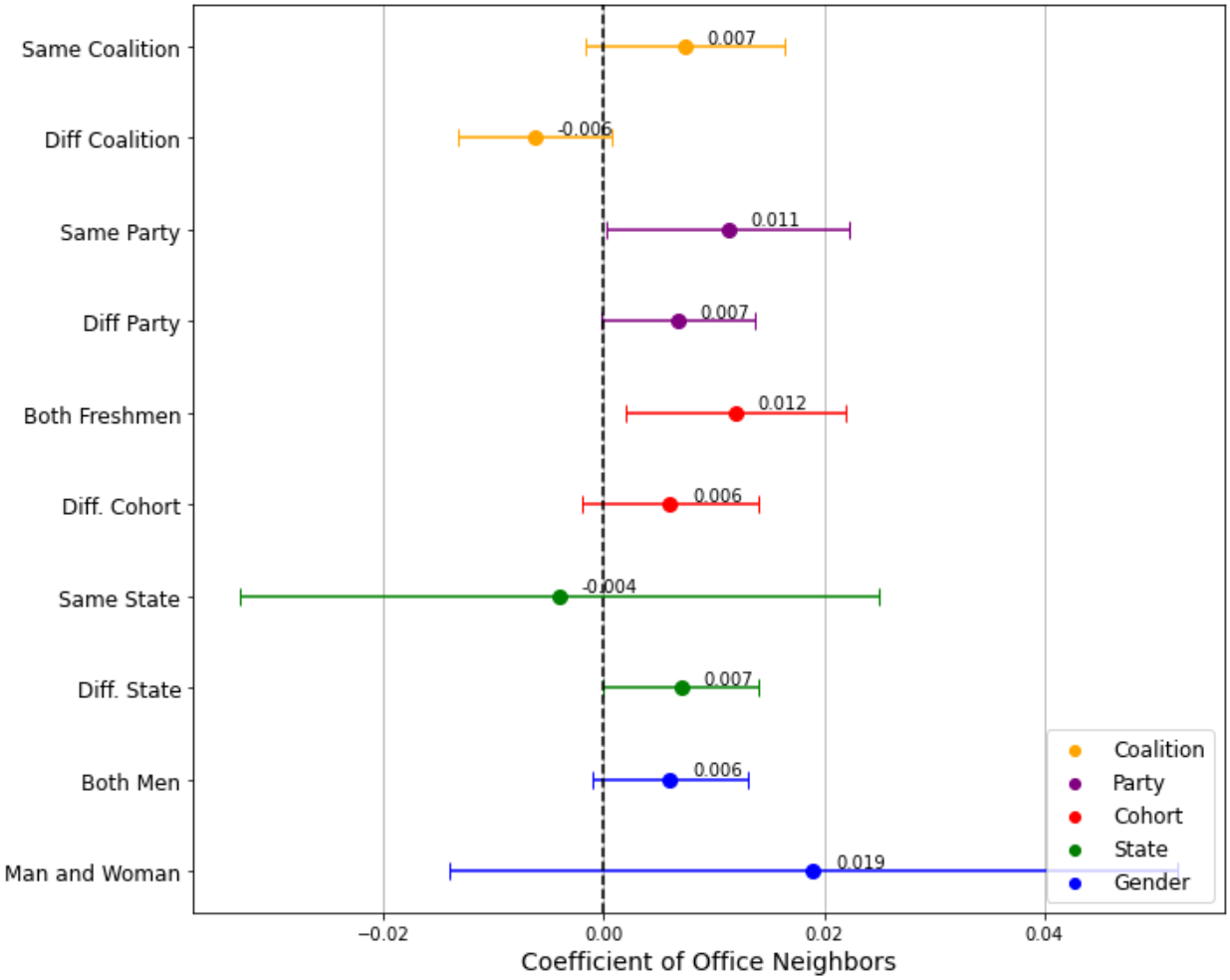
Across various types of propositions, the point estimates vary from -0.3 p.p. (PDL) to 1.1 p.p. (MPV), indicating diversity in the strength and direction of the relationship between office neighborhood proximity and voting agreement across different legislative proposals. However, the point estimates are not distinguishable from zero.

Overall, it shows that for Provisional Measures (MPV) legislators in close proximity exhibit a stronger tendency for agreement. However, this point estimate is not statistically significant, indicating that MPV alone is not a substantial influence on legislators' ability to sway the opinions of their neighbors. It, therefore, reveals this type of proposition does not represent an important dimension of heterogeneity in the analysis of the influence of spatial proximity on co-voting patterns.

Other potential dimensions of heterogeneity are their shared social characteristics. Therefore, heterogeneity analysis is also run by examining whether office neighbors that share salient social characteristics (namely gender, state of origin, freshman status, and party and coalition affiliation) influence each other more. Harmon et al. (2019) argue that shared social characteristics might

[21]See Appendix A.3 for the description of the main type of legislative proposals and for the distribution of votes over them across the two legislatures under analysis.

Figure 1.5: Heterogeneity analysis on legislators' characteristics



strengthen peer effects either because of the greater deference that individuals show toward the ideas and interests of in-group members or because social connection leads to more communication, and thus greater influence.

Figure 5 reveals that the impact of belonging to the same office neighborhood on voting agreement varies across various factors, including party and coalition affiliation, state of origin, cohort, and gender. In particular, pairs sharing the same freshman status look statistically significant, whereas the same (and different) party and the different state of origin only look marginally distinguishable from zero. Across various dimensions, the heterogeneity tends to converge to the voting agreement rate of 0.8 p.p. There is a notable difference when comparing pairs consisting of a male

and a female legislator with those formed by two male legislators[22]. In the former category, the coefficient is 1.9 p.p. with a standard error of 3.3, while in the latter one it is 0.6 p.p. with a standard error of 0.7.

Legislators from different coalitions who are in close proximity are less likely to agree in their voting decisions, with a decrease in agreement of 0.6 p.p. Additionally, when legislators from the same state are in close proximity, office proximity tends to favor disagreement in voting decisions, with a decrease in agreement of 0.4 p.p., with a standard error of 2.9.

Robustness Check: Alternative Definitions of Relevance

I also explore other definitions of voting relevance rather than the result margin of victory. An alternative is to consider the relevance of the individual vote for each legislator. Thus, this study adapts the Battaglini et al. (2023a) approach, in which roll calls are classified as relevant or not based on the salience of the topic being voted on by each legislator. This approach utilizes co-sponsorship data to determine the frequency with which each legislator has supported each topic, considering the least frequent topic as not relevant.

Table 11 reveals the highest point estimate (1.57 p.p.) for roll calls involving topics deemed not relevant for the pair of legislators. These result only offer weak evidence that office proximity is more likely to increase agreement when the vote is not relevant for either both or only one legislator in a pair. This evidence weakly indicates that cue-taking might occur through office proximity in votes that are not part to the own legislator's agenda – situations in which they tend to be less informed about the legislation being voted.

A second approach to classify votes according to their relevance follows Saia (2018), which categorizes roll calls based on the intensity of debate (see Table 12). Using speech data, it calculates the average daily number of speeches to establish a threshold, classifying roll calls as non relevant if they do not exceed this threshold. This approach offers limited evidence indicating that roll calls classified as relevant and voted on days with higher debate intensity are more likely to align among legislators in closer office proximity. Days with heightened debate intensity imply that legislators are likely to be more present in the Chamber as well as in their offices, potentially increasing the likelihood of communication among them.

[22]Since pairs must contain at least one participant of the office lottery, pairs with both female legislators are not present.

Table 1.11: Pair-Level Effects on Voting: Individual Vote Relevance Heterogeneity (p.p.)

	Dependent variable: Convergence		
	Non relevance for both	Relevant for one	Relevant for both
Office neighbors	1.57	1.53	0.18
	(1.3)	(1.0)	(0.8)
Controls	Yes	Yes	Yes
Proposal type FE	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes
Observations	863,342	4,859,778	13,254,169
Outcome Mean	70.15	67.0	69.1

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Roll calls are classified as relevant or not based on the salience of the topic being voted on by each legislator. Standard errors are clustered at the legislator individual level. This approach utilizes co-sponsorship data to determine the frequency with which each legislator has supported each topic, considering the least frequent topic as not relevant. Columns (1)-(3) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office neighbors* is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables are *same party*, *same state*, *ideological distance*, *same gender*, and *age difference*. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Robustness Check: The Remote Congress - Placebo Outcomes

Due to social distancing measures introduced to combat the COVID-19 pandemic, the Chamber of Deputies (56th legislature) implemented the SDR, *Sistema de Deliberação Remota*, transitioning all plenary activities—including roll calls—to a remote format and temporarily suspending the functions of permanent committees. This completely remote setup lasted from March 17, 2020 until February 11, 2021, when a hybrid model was introduced, allowing legislators to return to in-person plenary sessions and resuming committee activities. This hybrid system remained in place until October 25, 2021, when normal operations resumed. Briefly in 2022, from the start of the legislative year in February until April 18, the Chamber reinstated the SDR.

Objectively, while the SDR maintained the legislative workflow, it imposed tighter time frames and limited opportunities for extended debate, raising concerns about transparency and thoroughness in policy making (Santos, 2021). The congressional agenda under the SDR focused on a range of urgent pandemic-related issues, including public health measures, economic relief packages, emergency funding for healthcare systems, social assistance for vulnerable populations, and regulations to support remote work and education. The agenda also included policies to stabilize

Table 1.12: Pair-Level Effects on Voting: Roll Call Relevance Heterogeneity (p.p.)

	Dependent variable: Convergence$_1$	
	Low relevance	Relevant
Office neighbors	0.19	0.72
	(1.0)	(0.7)
Controls	Yes	Yes
Proposal type FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	2,019,460	29,118,074
Outcome Mean	66.1	67.6

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Roll calls relevance are classified based on the intensity of debate. Using speech data, it calculates the average daily number of speeches to establish a threshold, classifying roll calls as non relevant if they do not exceed this threshold. Standard errors are clustered at the legislator individual level. Columns (1)-(2) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office neighbors* is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables are *same party*, *same state*, *ideological distance*, *same gender*, and *age difference*. These variables are self-explanatory. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 1.13: Pair-Level Effects on Voting: Result Margin Heterogeneity (p.p)

	Dependent variable: Convergence	
	$> 5\%$	$< 5\%$
Office neighbors	-0.44	0.2
	(0.8)	(1.0)
Controls	Yes	Yes
Voting type FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	14,909,164	438,190
Number of roll calls	1,359	36
Outcome Mean	69.1	49.7

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Convergence is an indicator of whether the pair of legislators agreed in the proposal's vote. Office neighbors is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Column (1) reports the proximity effect for roll calls with more than 5 p.p. result margin. Column (2) reports the proximity effect for roll calls with less than 5 p.p. result margin Covariates included are same party, ideological distance, same state, same gender, and age difference. These variables are self-explanatory. *p<0.1; **p<0.05; ***p<0.01

the economy, such as financial aid for businesses, unemployment support, and amendments to labor laws, addressing the immediate needs and challenges posed by the COVID-19 crisis.

The period under the SDR allows me to perform a counterfactual exercise. To make this robustness test, I ran the same specifications (Tables 13 and 14) but using the subset of data related to the period the Congress was operating in a completely remote way. Therefore, the spatial interactions that used to happen were no longer possible, both at the Annexes and at the committees. Holding the hypothesis that spatial proximity effects would not hold in the absence of physical interactions, I expect to find smaller and statistically nonsignificant results on both models.

Analysis of votes cast under the SDR shows, on average, a negative point estimate (-0.44 p.p.) for voting convergence associated with sharing a similar location. For contested votes, with a result margin lower than 5 p.p., the impact of office proximity (0.2 p.p.) is 10 times smaller and non longer statistically significant. The effect of committee membership interacted with similar

Table 1.14: Pair-Level Effects on Voting: Interaction of Spatial Proximity and Committee Membership (p.p.)

	Dependent variable: Convergence	
	$> 5\%$	$< 5\%$
Office neighbors	-0.39	0.46
	(0.7)	(1.1)
Committee	0.71	-0.13
	(0.3)	(0.2)
Committee × Office neighbors	-0.32	-1.24
	(0.7)	(2.1)
Controls	Yes	Yes
Voting FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	14,909,164	438,190
Number of roll calls	1,347	36
Outcome Mean	69.1	49.7

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Convergence is an indicator of whether the pair of legislators agreed in the proposal's vote. Office neighbors is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Committee is an indicator of whether one legislator within a pair is member of the committee related to the subject of the roll call being voted. Column (1) reports the proximity effect for roll calls with more than 5 p.p. result margin. Column (2) reports the proximity effect for roll calls with less than 5 p.p. result margin Covariates included are same party, ideological distance, same state, same gender, and age difference. These variables are self-explanatory. *p<0.1; **p<0.05; ***p<0.01

Table 1.15: Pair-Level Effects on Speeches: Main Analysis (p.p.)

	Dependent variable: Speech Similarity		
	All	Order of Business & Small Shift	Small Shift
Office neighbors	0.08	0.21	0.14
	(0.3)	(0.5)	(0.3)
Controls	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes
Observations	435,608	76,860	69,064
Outcome Mean	15.5	16.1	17.6

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Column (1)-(3) report the impact of a pair of legislators belonging to the same office neighborhood on their speech similarity. Column (1) takes in account speeches from the Small Shift, Long Shift, Order of Business, and Parliamentary Communications. Communication. Column (2) takes in account speeches from the Small Shift and the Order of Business. Column (3) only takes in account speeches from the Small Shift. *Cosine Similarity* is taken between the speeches of the respective pair of legislators. *Office neighbors* is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables are *same party*, *same state*, *ideological distance*, *same gender* and *age difference*. *p<0.1; **p<0.05; ***p<0.01

location also changes (-1.24 p.p.) and is non longer statistically significant. Both findings support our hypothesis on the importance of physical interactions in explaining the influence of spatial proximity, also underscoring the importance of keeping the standing committees operational.

1.5.3 The Impact of Office Proximity on Speeches

This section presents an examination of the impact of proximity on speech behavior. Speech behavior is assessed using a natural language processing technique called document embedding. The similarity in speech, denoted as *Speech similarity*$_{ijt}$, is determined using word frequency vectors, also known as "bag-of-words". This approach involves tallying the frequency of words in each speech, creating a vector representation, and then computing the cosine similarity between the speeches of each pair of legislators $i - j$ in a given legislature t.

Column (1) shows that, on average, legislators in close proximity are more likely to speak similarly in 0.08 percentage points. However, this small effect is not statistically significant at conventional levels (p>0.1).

There are two moments where legislators can intervene in the debates more freely and systemat-

ically: a) on the floor, during the Short Shift; and, b) in committees, during the Order of Business, when discussing the policies under consideration. In column (2), which considers speeches from these two moments, the probability increases to 0.21 percentage points, suggesting a stronger positive effect on speech similarity when legislators can freely speak their minds. Nevertheless, similar to column (1), the effect is not statistically significant.

In column (3), focusing solely on speeches from the Small Shift, the coefficient remains positive at 0.0014 but is still not statistically significant.

Overall, while there appears to be a tendency for legislators from the same office neighborhood to have more similar speeches, the point estimates are very small and not statistically significant.

1.6 Conclusion

The office neighborhood provides an environment that facilitates communication and influence between legislators. Based on this social structure, this paper examined the peer effects of office proximity on legislative behavior, relying on the random allocation of offices in Brazil's Chamber of Deputies to estimate and identify these network effects. By applying this methodology in the analysis of peer effects, this study contributed to the literature on randomization-based research in legislative politics (Alquezar-Yus and Amer-Mestre, 2024; Darmofal et al., 2023; Harmon et al., 2019; Lowe and Jo, 2024; Rogowski and Sinclair, 2012; Saia, 2018; Zelizer, 2019). Although the results indicate that, on average, office proximity has a no statistically significant impact on legislative behavior, empirical evidence shows that this influence becomes significant and stronger in contested votes, a important type of nonprocedural decision. These findings suggest that in such situations, where each vote holds greater importance, legislators' office proximity plays an important role in voting behavior – complementing other sources of influence, whether ideological or nonideological.

This paper also emphasizes the informational role of committees by showing empirical evidence of its influence on disputed decisions. The committee system exists in order to reap the informational gains to the floor as a whole from having subgroups of its members specializing on specific topics. This specialization allows the committees to acquire information about the true consequences of a bill to be considered by the floor. I showed that when one legislator in a pair of legislators is a perceived expert (i.e., he is on the relevant standing committee for the respective roll call), sharing a same office neighborhood significantly increases the influence on vote decisions in contested votes. These findings suggest a potential mechanism of cue-taking where legislators rely on perceived experts – typically standing committee members who reported the measure – to guide

their voting decisions, especially when informed decision making is crucial. However, this paper acknowledges limitations in distinguishing between cue-taking through office proximity and other concurrent mechanisms like social pressure. Therefore, future research exploring these potentially mutual mechanisms on exogenous or endogenous legislative networks could provide further insights into this phenomenon. A good example are the two field experiments ran by Zelizer (2019) in a state legislature, in which he offers empirical evidence of cue-taking behavior through the diffusion of a randomly-assigned information treatment across an endogenous legislative network. Furthermore, this empirical evidence is limited to contested votes, as it did not show influence on other alternative measures of voting relevance (a proxy for nonprocedural decisions).

A potential endogeneity issue arises when analyzing voting convergence on subsets of roll calls with narrow result margins (contested votes), as this setting may introduce simultaneity bias. Specifically, legislators' decisions to vote in alignment with their office neighbors may be influenced by the anticipation of a close outcome. In highly contested votes, legislators may be more inclined to consult or be influenced by their peers, knowing that each vote could decisively impact the result. This simultaneity creates a feedback loop: the closeness of the vote encourages peer influence, and the influence itself may, in turn, affect the vote margin. As a result, any observed voting convergence could reflect both genuine peer effects and strategic alignment in anticipation of a close vote, rather than purely the influence of office proximity. This endogeneity complicates the causal interpretation of peer effects in contested roll calls, as the closeness of the vote is not an exogenous condition but is potentially shaped by the same forces driving voting convergence. Thus, the potential simultaneity bias in contested votes represents a limitation of this study.

Brazil's proportional and open list electoral system, which often leads to majority coalition formation in the legislative, presents an intriguing environment for studying peer effects dynamics. First, this coalition-driven legislative system centralizes influence within the formal structures of party and coalition leadership, where decisions are typically coordinated to maintain legislative support for the executive's agenda. In this context, it is generally expected that legislative behavior align primarily with party and coalition guidelines, as legislators depend on these influences for political resources and career advancement. Given these dynamics, the Chamber of Deputies provides a unique setting to test for alternative sources of influence outside party and coalition structures. If peer effects, such as those generated by office proximity, have a measurable impact on legislators' behavior, it would suggest that interpersonal dynamics play a role beyond the structured guidance of party and coalition leadership. Therefore, this paper's empirical evidences challenge the

assumption that party and coalition influence are the sole drivers of legislative behavior in Brazil's Chamber of Deputies, indicating that informal interactions can influence decisions even in a highly coordinated political system.

Other studies have explored similar coalition-based legislative systems, where informal interactions may influence formal decision-making. For instance, Saia (2018) and Lowe and Jo (2024) examined Iceland's parliamentary system, which also relies on coalition politics and centralized party influence. Their findings suggest that even within highly structured environments, informal peer effects can shape legislators' behavior in nuanced ways. Further analysis on different legislative systems, in which majority coalition formation is not the rule, could provide additional evidence to this hypothesis.

Chapter 2

FAST INTERNET AND LABOR MARKET IMPACT: EVIDENCE FROM DIFFERENTIAL BROADBAND ROLLOUT IN BRAZIL

with *Ana Almeida*

2.1 Introduction

The expansion of broadband (fast) internet infrastructure has become an important driver of economic development. Broadband enhances productivity (Czernich, 2014), improves capital and human resources (Campbell, 2024; Grimes and Townsend, 2018; Klein, 2022), and facilitates flexible work arrangements (Han, 2021). On the supply side, it accelerates innovation and efficient decision-making, boosting total factor productivity and increasing the value of capital, while also improving access to online education and enabling remote work. On the demand side, consumers benefit from significant time savings and improved access to digital services such as e-commerce.

Yet, despite these generally positive outcomes, broadband's overall impact on development is multifaceted. While it enhances education by providing access to online platforms that bolster human capital, it can also impair academic performance due to distractions from online games and social media (Cambini et al., 2024; Henriksen et al., 2022). In labor markets, broadband investment facilitates the adoption of new technologies that complement skilled workers but may substitute for unskilled workers (Akerman et al., 2015; Yang, 2023). Atasoy (2013) finds that broadband expansion tends to improve employment and wages in sectors reliant on skilled labor; however, these benefits are unevenly distributed and can widen the gap between high- and low-skilled workers. Moreover, although broadband can foster job creation and wage growth, it may simultaneously intensify competition and shift demand toward less-skilled labor, adversely affecting wage growth among highly educated workers (Dutz et al., 2017). These contrasting effects underscore the need for a comprehensive analysis of broadband's diverse impacts on economic development and, specifically, on labor markets.

This paper examines the effects of broadband infrastructure expansion on labor market outcomes – specifically, net job creation, number of hours worked, and average monthly wages – across 3,000 Brazilian municipalities. We exploit Brazil's broadband policy as a quasi-natural experiment to

identify causal impacts in the period ranging from 2011 through 2016 [1]. The policy addressed critical network bottlenecks by mandating the installation of backhaul capacity in underserved municipalities and by establishing population-based minimum speed requirements – 8 Mbps for populations up to 20,000, 16 Mbps for 20,001 to 40,000, 32 Mbps for 40,001 to 60,000, and 64 Mbps for populations above 60,001. These criteria create discontinuities in treatment assignment, which we exploit using a multicutoff regression discontinuity design that compares municipalities just above and below the respective population thresholds[2]. By isolating exogenous variation in internet speed, our analysis reveals how enhanced broadband infrastructure reshapes labor markets, with a particular emphasis on its skill-biased distributional consequences.

The broadband policy in Brazil deployed a mix of technologies – microwave radio, fiber-optics, and satellite – to meet its infrastructure investment needs. In particular, due to non-compliance in treatment assignment, actual assignments deviate from the policy's strict cutoffs, necessitating a fuzzy multicutoff regression discontinuity (FRD) framework. In the first stage, we model the probability of receiving a certain amount of internet speed as a function of instruments based on population eligibility cutoffs. In the second stage, we regress labor market outcomes on the predicted internet speed from the first stage. Our first stage analysis reveals that only the linear specification for microwave radio at the first cutoff produces robust and theoretically consistent estimates, rendering it the sole valid specification in our analysis.[3]

We first assess the aggregate effects of broadband expansion on labor market outcomes before examining heterogeneity across worker characteristics and industry sectors. Consistent with prior literature, our findings reveal nuanced trade-offs: broadband expansion stimulates net job creation among individuals with lower educational attainment and those employed in the commerce industry, yet it simultaneously reduces hours worked and wages for certain groups. In particular, highly educated workers and those in occupations characterized by non-routine tasks experience persistent declines in hours worked and temporary wage reductions that peak around 2013. These adverse effects may reflect skill mismatches or substitution effects – where automation displaces high-skilled workers in administrative or technical roles. Our results echo the skill-biased patterns observed in

[1] This policy was enacted through Decree No. 6,424, which revised Decree No. 4,769 to ensure that the infrastructure could support higher data throughput and prepare the network for future demands.

[2] Note that we compare municipalities with different internet speeds, rather than municipalities that transition from 0 to 100% internet access, as explored in Dutz et al. (2017).

[3] In contrast, no specification for fiber or the remaining cutoffs yields consistently reliable results, underscoring the instrument's limitations for those technologies.

other contexts (Akerman et al., 2015; Dutz et al., 2017), demonstrating that while broadband in theory has the potential to complement high-skilled workers in STEM fields, it in fact substitutes for roles vulnerable to technological change. Moreover, pronounced industrial and gender disparities emerge, with the most negative impacts concentrated in the service sector and among women. Overall, these findings underscore broadband's dual role as both a catalyst for economic inclusion and a driver of inequality, contingent on workers' skill profiles and industrial characteristics.

This chapter contributes to several strands of the literature. First, it extends the growing body of empirical work on the economic impact of broadband access (Akerman et al., 2015; Czernich et al., 2011; Koutroumpis, 2009), with novel evidence from a large developing economy (Dutz et al., 2017; Mendonça et al., 2021). While much of the existing literature focuses on OECD countries or urban broadband diffusion, this study leverages quasi-experimental variation to identify effects in less-developed, mid-sized municipalities—a setting where labor market frictions and digital divides are more pronounced. Second, it contributes to research on infrastructure and local development by showing that digital infrastructure, like traditional transport networks (Donaldson, 2018), can substantially reshape economic activity. Finally, by focusing on occupational and sectoral heterogeneity, the chapter provides a more granular understanding of how broadband may complement specific types of human capital, offering implications for policies aimed at fostering inclusive digital transformation.

Section II presents the institutional background of this study. Section III introduces the empirical model, describing the data and identification strategy, followed by internal validity checks. Section IV presents the main results on municipal labor market outcomes and heterogeneity analyses. Section V concludes.

2.2 Background

2.2.1 The Backhaul policy

Brazil's telecommunications sector has its basis in Law No. 9,472/1997—the General Telecommunications Law—which established the regulatory framework and defined the national government's authority over telecommunications services. The National Telecommunications Agency (Anatel) was created to oversee the privatization process and ensure that universal service obligations were met across various technologies.

Building on this strong regulatory framework, the 1990s marked a period of major reform in Brazil's telecommunications landscape. The mid-90s privatization process was not only a market-

Table 2.1: Rule for Backhaul Speed Assignment

Population	**Minimum Speed**
Up to 20,000	8 Mbps
From 20,001 to 40,000	16 Mbps
From 40,001 to 60,000	32 Mbps
From 60,001 upwards	64 Mbps

Source: Decree No. 6,424/2008, art. 13-A.

driven initiative but also accompanied by a robust governmental commitment to ensure that every citizen had access to basic communication services. This dual approach was fundamental for modernizing the nation's communication infrastructure and reducing regional disparities. In 2003, Decree No. 4,769 further advanced these efforts by establishing the Plan for the Universalization of the Public Switched Telephone Network (PSTN), with the primary objective of extending telephony services to under-served and remote areas, ensuring that even the most isolated municipalities could access the network.

As the Internet gained prominence for economic development, Brazil's policy focus expanded beyond basic telephony to encompass broadband services. In 2008, Decree No 6,424 was issued to address limitations in the existing network infrastructure – specifically, the backhaul capacity[4]. This decree not only mandated the implementation of backhaul infrastructure but also established the Broadband at School Program, aimed at bringing broadband Internet to all urban public schools across Brazil. In sum, this decree revised Decree No. 4,769 to ensure that the infrastructure could support higher data throughput, thereby reducing bottlenecks and preparing the network for future demands.

Under this decree, telecommunications companies were required to install backhaul capacity in all municipalities lacking such infrastructure. The installation schedule was designed to span three years: 40% of municipalities were to be completed by the end of 2008, 80% by the end of 2009, and full implementation was expected by the end of 2010. Moreover, the legislation set minimum speed requirements for different population size ranges, see Table 2.1. These benchmarks were largely consistent with the capacities deployed by companies, especially in regions where low demand and limited consumer purchasing power made investments less attractive.

[4]Backhaul is the support network infrastructure of the PSTN for broadband connection, linking the access networks to the operator's backbone.

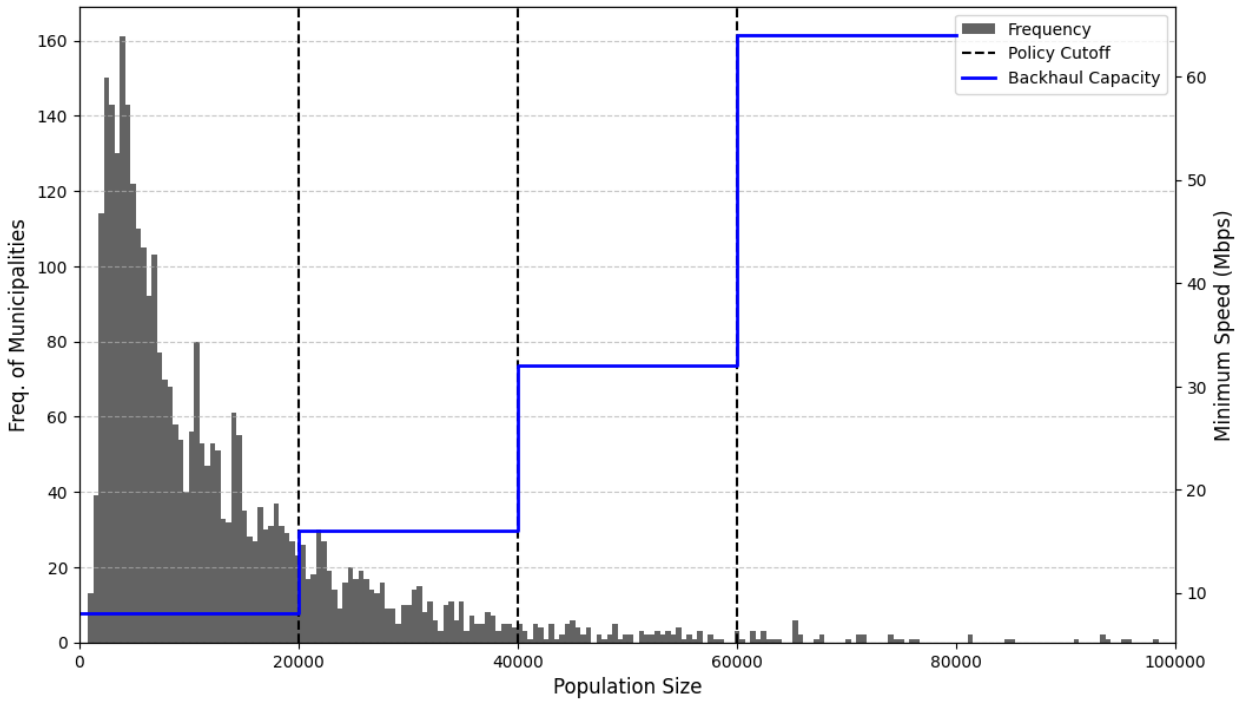


Figure 2.1: Rule of Assignment and Distribution of Municipalities

By 2008, the year Decree No. 6,424 was enacted, Brazil comprised 5,565 municipalities, of which 2,125 had already been invested with the necessary infrastructure (i.e. treated)[5]. Therefore, the policy proposed to connect 3,440 municipalities (approximately 62% of Brazilian municipalities) that did not have access to this technology.

These municipalities present varying levels of population size. If we look at their statistics, the median population size is 8,895 while the average population is notably higher at 14,283. This gap between the median and the mean indicates a skewed distribution, where most municipalities tend to be smaller and a inferior number of more populous municipalities raise the overall average. This considerable variation emphasizes the complexity Brazil faces in providing equal access to services and infrastructure across all municipalities, especially when considering remote and less populated areas. Figure 2.1 presents both the distribution of municipalities and the policy's assignment per population size. As observed, the distribution is skewed to the left, with the majority of municipalities clustered below the initial cutoff point.

Microwave radio, fiber-optics, and satellite transmission technologies were strategically adopted

[5]Municipalities are the lowest level of political division within Brazil.

Table 2.2: Backhaul technology of treated municipalities - 2008 to 2010

Year	Radio	Fiber	Satellite	**Total per Year**
2008	715	556	1	1,272 (42%)
2009	1,102	156	10	1,268 (42%)
2010	424	39	1	464 (16%)
Total per Tech.	2,241 (74.6%)	751 (25%)	12 (0.4%)	3,004 (100%)

Source: Authors' elaboration based on Anatel data.

to address the policy and delivery the necessary infrastructure. Microwave radio transmission is used in point-to-point communication systems, where it utilizes wireless signals to transmit data, making it an effective solution for less densely populated areas and regions where traditional wired networks are challenging to implement, as they are generally installed in areas with easy access for construction work and laying cable. Fiber-optics transmission, on the other hand, employs optical fibers to deliver exceptionally high-speed and reliable connections, making it ideal for urban and coastal regions with higher demand and better economic conditions. Satellite transmission provides an alternative in remote or difficult-to-access areas, such as the Amazon region, where geographical obstacles hinder the deployment of terrestrial infrastructures.

The roll-out program covered a total of 3,004 municipalities by 2010 (approximately 87% of the untreated municipalities), as detailed in Table 2.2. During the first year of implementation, 42% of municipalities were treated. By the end of the second year, an additional 42% received the infrastructure investment, bringing the total to 84%. By 2010, the remaining 16% of municipalities received the treatment. Additionally, 114 municipalities received treatment after 2010, while 271 municipalities remained untreated. About three-quarters of the municipalities in the sample received investment in radio technology infrastructure, while one quarter received investment in fiber technology infrastructure. The number of municipalities receiving investment in satellite technology was very small, less than 1%.

Figure 2.2 shows the map of Brazil illustrating the municipalities that received investment in backhaul capacity and the specific technologies adopted. A significant concentration of municipalities is found in the Northeastern region of Brazil, an area that generally experiences lower levels of development. In these less urbanized settings, particularly in the interior where population sizes tend to be smaller, microwave radio has been the preferred choice for infrastructure upgrades. Conversely, fiber optics are more commonly observed along the coast, where larger urban centers and

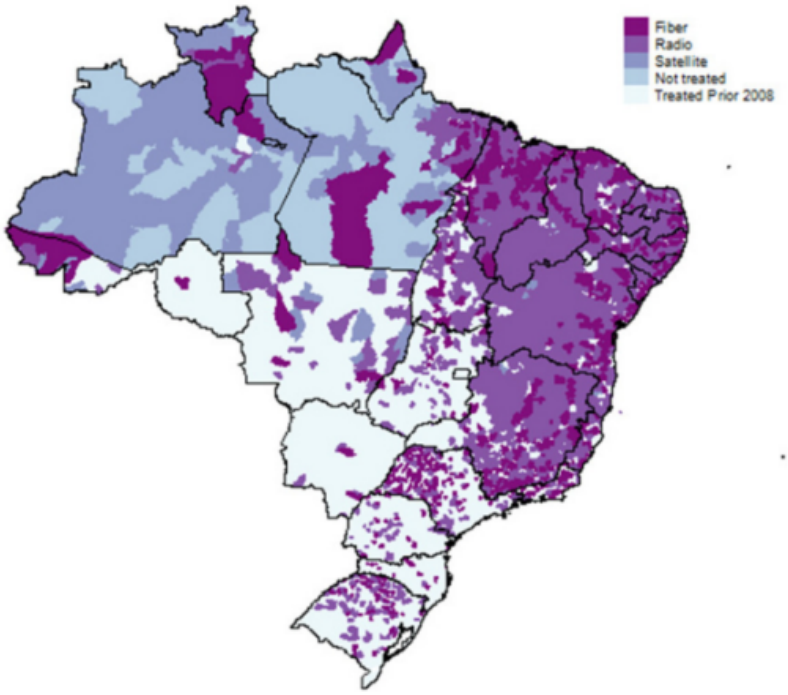


Figure 2.2: Backhaul Technologies - 2011. *Source*:Henriksen et al. (2022)

better economic conditions prevail. In the North, especially within the challenging terrain of the Amazon, satellite infrastructure has been deployed to ensure connectivity in hard-to-reach areas. Meanwhile, the patterns of technology deployment in the Midwestern and Southern regions do not display a clear trend.

Table 2.3 shows the distribution of backhaul technologies among 3,004 treated municipalities in 2010, segmented by population size.

Table 2.3: Backhaul technology by population level – treated municipalities in 2010

Population	Radio	Fiber	Satellite	**Total per Pop**
$<$ 20 K	1,854 (79.3%)	476 (20.4%)	7 (0.3%)	2,337 (100%)
20 K to 40 K	303 (60%)	199 (40%)	0 (4%)	502 (100%)
40 K to 60 K	60 (60%)	39 (39%)	1 (1%)	100 (100%)
$>$ 60 K	24 (37%)	37 (57%)	4 (6%)	65 (100%)
Total per Tech.	2,241	751	12	3,004

Source: Authors' elaboration based on Anatel data.

In municipalities with fewer than 20,000 inhabitants, microwave radio is predominant, accounting for 79.3% of installations, while fiber optics is implemented in 20.4% of cases, and satellite usage is minimal at 0.3%. For municipalities with populations between 20,000 and 40,000, 60% adopted microwave radio and 40% fiber optics, with no satellite installations reported. A similar distribution is observed for those with 40,000 to 60,000 inhabitants, where 60% received microwave radio, 39% fiber optics, and 1% satellite. In contrast, in larger municipalities with over 60,000 inhabitants, fiber optics becomes the preferred technology at 57%, with microwave radio installations at 37% and satellite at 6%.

Overall, out of the total 3,004 treated municipalities, 2,241 were equipped with microwave radio, 751 with fiber optics, and 12 with satellite, indicating that smaller municipalities predominantly rely on microwave radio while fiber optics is relatively more common in larger urban areas, with satellite playing a minor role across the board.

2.2.2 Broadband Internet and Labor Markets

Broadband expansion has been linked to both positive outcomes – such as increased employment rates and improved job-finding prospects – and negative consequences, including reduced work intensity and wage stagnation.

The adoption of broadband internet in firms complements skilled workers, particularly in high-skill-intensive industries, by enhancing productivity and wages (Akerman et al., 2015; Yang, 2023). For instance, Akerman et al. (2015) demonstrates through a Norwegian case study that broadband adoption improves labor market outcomes for skilled workers, especially college graduates in science, technology, engineering, and business fields. This is attributed to broadband's role in enabling firms to adopt skill-biased technologies, which disproportionately benefit workers performing non-routine, abstract tasks.

However, Akerman et al. (2015) also highlights broadband's substitution effect: it reduces demand for unskilled workers by lowering their marginal productivity and wages. This effect is most pronounced in routine task-intensive industries, where automation driven by broadband internet displaces workers engaged in repetitive roles.

Empirical evidence paints a nuanced picture. While broadband can boost employment by fostering firm growth and productivity – particularly in rural or geographically isolated regions (Atasoy, 2013; Bhuller et al., 2019) – it also risks exacerbating labor market inequalities. The skill-biased nature of technological change may lead to labor misallocation and wage disparities, especially

when workers' skills fail to align with evolving technological demands (Hua and and, 2024). These disparities are amplified in regions with underdeveloped digital infrastructure. Although digital infrastructure investments can mitigate misallocation by improving labor mobility, benefits often remain unevenly distributed across worker groups (Hua and and, 2024). Dutz et al. (2017), using a quasi-experimental design, found that municipalities transitioning to full internet access experienced wage declines in higher-skilled workers and higher-skill service-industry occupations, even after controlling for confounding factors.

These findings underscore the dual-edged impact of broadband expansion. To harness its benefits while addressing adverse effects, targeted policy measures are fundamental. Investments in digital literacy programs (Yang, 2023), equitable regional infrastructure development (Hua and and, 2024; Jin et al., 2025), and firm-level incentives (Akerman et al., 2015) could help maximize the gains of expanded broadband connection.

2.3 Empirical Model

This section outlines the empirical model employed in our analysis. We begin with a comprehensive overview of the data, summarizing key sources and variables. Next, we explain our sample selection criteria, detailing the rationale behind the inclusion of observations. In sequence, we elaborate on our identification strategy and the empirical estimation techniques used to derive our results. Finally, we present two internal validity checks.

2.3.1 Data

We use data from multiple sources to construct our analysis. The primary dataset is sourced from the Annual Social Information Report (RAIS), an administrative record derived from a mandatory yearly survey conducted by the Brazilian Ministry of Labor and Employment. The survey covers all formal employers – those possessing a signed work card – which entitles them to the complete range of benefits and labor protections provided under the legal employment system[6]. This survey provides comprehensive details on all formally employed workers in Brazil – approximately 60 million per year – and the establishments where they work. Firms are required to file, and face

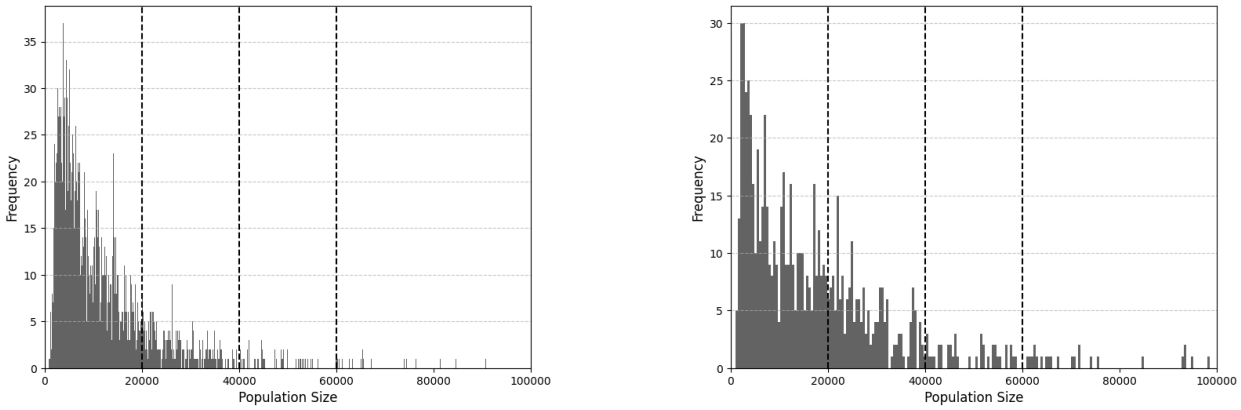
[6]According to Dix-Carneiro and Kovak (2015), workers have an incentive to ensure that their employer is filling the required information, as it is the main tool used by the government to enable the payment of the *abono salarial* to eligible workers. This is a government program that pays one additional minimum wage at the end of the year to workers whose average monthly wage was not greater than two times the minimum wage, and whose job information was correctly declared in RAIS, among other minor requirements.

fines until they do so. It omits those without signed work cards, including interns, the self-employed, elected officials, domestic workers, and other minor employment categories. At the worker level, we extract information such as occupation codes (CBO, Brazilian Classification of Occupations), industry codes (CNAE, National Classification of Economic Activities), wages, hours worked, dates of admission and separation, salary ranges calculated relative to the minimum wage, educational attainment, gender, and age. Additionally, we obtain firm-level data on both firm size and firm entry/exit, and we enhance these datasets with exporting status information from the Brazilian Secretariat of Foreign Trade (SECEX).

We incorporate data from Anatel, which provides precise information on the backhaul capacity installed in each municipality as well as data on fixed broadband internet use. We also utilize data from the Brazilian Institute of Geography and Statistics (IBGE) to obtain information on population sizes and the occupation codes.

We then use RAIS to construct municipality-level outcomes, leveraging its panel structure for longitudinal analysis. Our primary outcomes include net job creation, hours worked, and wages. In addition to aggregating worker-level data to the municipality level, we also develop outcomes that capture specific worker characteristics. First, workers are organized into three educational groups based on their highest attained degree. Workers are considered high-educated if they hold a college degree (bachelor's, master's, or PhD), medium-educated if their highest qualification is a high school diploma, and low-educated if they have not completed high school.

Next, to classify occupations, we translate the Brazilian Classification of Occupations (CBO) codes into skill levels using a suite of natural language processing techniques. Our first approach employs an unsupervised learning method that begins with leveraging a transformers-based model, specifically SentenceTransformer (Reimers and Gurevych, 2019), to extract BERT embeddings from the textual descriptions of occupation names and associated tasks. These embeddings transform the textual information into high-dimensional vector representations that capture the nuanced semantic meanings of each occupation. Once these representations are obtained, we apply K-Means clustering with two centroids to automatically partition the occupation codes into two distinct groups—one corresponding to skilled occupations and the other to unskilled occupations. This method allows us to uncover underlying patterns in the data without relying on pre-labeled training sets.



(a) Radio

(b) Fiber

Figure 2.3: Histogram of Population Size

2.3.2 Sample Definition and Summary Statistics

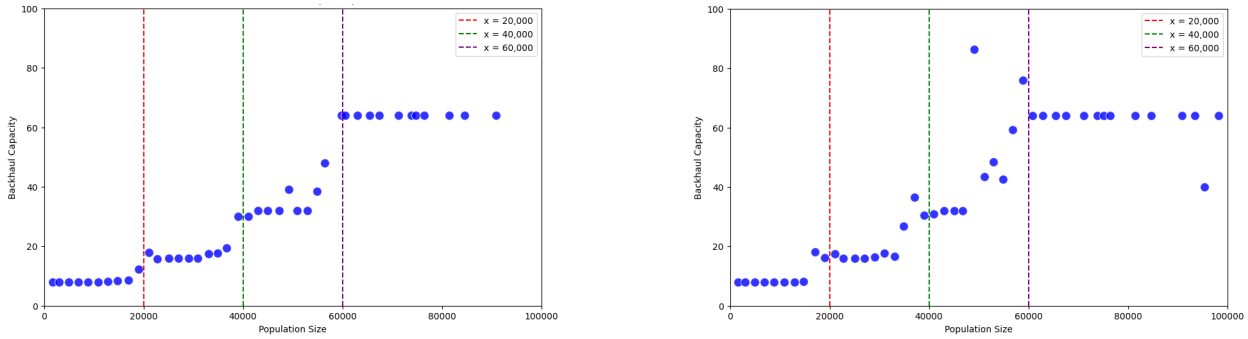
As noted in Section 2.1, 99.6% of municipalities received investments in either microwave radio or fiber optics. Consequently, our analysis focuses solely on these two technologies, excluding satellite-treated municipalities – which account for only 0.4% of the population – and we run the analysis separately for each technology.

Figure 2.3 illustrates the distribution of municipalities by population size. As observed in Table 2.3, the overall distribution is left-skewed. Specifically, municipalities equipped with radio technology tend to fall below the 20,000-resident threshold, while fiber technology is relatively more prevalent in municipalities with larger populations.

Figure 2.4 illustrates the actual backhaul capacities installed in our targeted municipalities.

According to the policy, municipalities under 20,000 inhabitants should receive 8 Mbps, those with 20,000 to 40,000 should receive 16 Mbps, from 40,000 to 60,000 should receive 32 Mbps, and those above 60,000 should receive 64 Mbps. Ideally, this creates a step function with clear capacity jumps at the 20k, 40k, and 60k thresholds. In practice, however, many municipalities do not strictly adhere to these cutoffs (i.e. non-compliers). Notably, some municipalities near the lower end of a given population range have been assigned capacities intended for the next bracket, potentially leading to an underestimation of treatment effects. This issue is most prominent in the fiber sample, though it also appears around the second cutoff in the radio sample.

To understand the underlying reasons for this lack of compliance, we interviewed Marcelo Alves



(a) Radio

(b) Fiber

Figure 2.4: Backhaul Capacity vs Population Size

da Silva, a specialist in telecommunications regulation at Anatel—Brazil's national telecommunications regulator [7]. He noted that municipalities served by fiber—a comparatively inexpensive technology to deploy in peri-urban settings—often lie close to large metropolitan areas and cluster just below the population thresholds. Providers therefore have an incentive to install higher-than-mandated capacities in these towns: the incremental cost of laying additional fiber is low, and pre-provisioning bandwidth allows them to accommodate future demand surges while avoiding repeated civil-works permits. Fiber links also benefit from straightforward capacity upgrades via equipment swaps rather than physical reconstruction, so oversizing backhaul ex-ante is commercially sensible.

By contrast, point-to-point radio backhaul is costlier to install, entails higher tower and spectrum expenses, and is typically reserved for remote municipalities that are hard to reach by fiber and usually have smaller populations. In such cases providers face tighter budget constraints and little strategic value in supplying capacities beyond the legal minimum. As a result, compliance is much stricter at the first radio cut-off (20,000 inhabitants), where large population reclassifications are unlikely and the reputational or regulatory gains from exceeding the requirement are minimal. Observing better compliance in the radio series and systematic "over-provisioning" in the fiber series is therefore consistent with the economic incentives highlighted by industry practitioners.

Tables B.1 and B.2 in the Appendix B.1 present the summary statistics of the main outcomes—log of wages, hours worked, and net job creation—per different technology organized by population threshold, treatment status, and subgroup.

[7] Authors' interview with Marcelo Alves da Silva (Anatel), May 10th, 2025.

2.3.3 Identification and Empirical Strategy

Our empirical model exploits a quasi-natural experiment induced by the broadband policy to assess the causal impact of improved internet speeds on labor market outcomes at the municipal level in Brazil. We employ local polynomial regression discontinuity (RD) point estimators with robust bias-corrected confidence intervals following Calonico et al. (2023, 2014) and Cattaneo et al. (2016). This non-parametric method leverages the predetermined population thresholds set by the policy, which create sharp discontinuities in treatment assignment. In essence, in the absence of manipulation, the probability of receiving treatment jumps discontinuously at these cutoffs, yielding variation that is plausibly exogenous to potential confounders. By comparing municipalities that lie just below and just above each threshold, our multicutoff design approximates a randomized controlled trial, thereby allowing for credible causal inference. For estimation, we first select a data-driven, mean squared error (MSE)-optimal bandwidth to define the effective sample around each cutoff, and we weight observations using a kernel[8].

The multicutoff approach enables us to assess treatment heterogeneity across different population levels. Ideally, municipalities with population sizes below 20,000 should receive 8 Mbps, those between 20,000 and 40,000 should get 16 Mbps, between 40,000 and 60,000 should receive 32 Mbps, and those above 60,000 should have 64 Mbps. However, due to non-compliance in treatment (see Figure 2.4), the actual assignments deviate from these strict cutoffs, making a fuzzy multicutoff regression discontinuity (FRD) framework appropriate for our analysis[9].

Additionally, it is essential to ensure that our running variable – population size, which serves as our first-stage instrument – affects labor market outcomes only through its impact on internet speed assignment. To maintain the validity of the instrument's exclusion restriction, we must control for any confounding factors, such as other policies based on population size that might also influence labor market outcomes. One notable example is the Municipal Participation Fund (FPM), a federal program that provides block grants to municipal governments. Under the FPM

[8] A narrower bandwidth includes observations very close to the cutoff, which minimizes bias but can lead to high variance due to a smaller sample size. Conversely, a wider bandwidth increases the number of observations, thereby reducing variance, but may introduce bias by incorporating data points that are less comparable to those near the cutoff. The optimal bandwidth is determined by minimizing the mean squared error (MSE) of the estimator, ensuring that the bias-variance trade-off is well managed. We employ two different types of kernel: uniform and triangular. The uniform kernel assigns equal weight to all observations within a chosen bandwidth around the cutoff, treating each observation equally regardless of its distance from the threshold. In contrast, the triangular kernel places more weight on observations closer to the cutoff and less on those further away, reflecting the intuition that observations nearer the threshold are typically more informative about the discontinuity.

[9] We also present a normalizing-and-pooling approach, in which we normalize the three scores and pool all observations into a single dataset for analysis.

allocation mechanism, municipalities are categorized into population brackets, with each bracket receiving a coefficient that determines the share of state resources allocated to the FPM – smaller populations correspond to lower coefficients (see Litschig and Morrison (2012) and Brollo et al. (2013))[10].

Our estimation, therefore, follows a two-stage procedure. In the first stage, we model the probability of receiving treatment dosage (i.e. internet speed), S_i, as a function of instruments based on the population eligibility cutoffs. For each municipality i and for each cutoff n (with $n = 1, 2, 3$ corresponding to the three population thresholds), we define an indicator variable C_{in} that equals 1 if municipality i meets the criteria for cutoff n, and 0 otherwise. In order to control for the confounding factor, we also include dummies variables, FPM_{ik}, for each population bracket k for every municipality i. Additionally, we let $f(\text{p}_i)$ be a flexible polynomial function of population size p_i and ϕ_i represents state fixed effects.

The first-stage specification is:

$$S_i = \sum_{n=1}^{3} \delta_n C_{in} + \gamma f(p_i) + \sum_{k=1}^{K} \theta_k FPM_{ik} + \phi_i + \eta_i,$$

where η_i is the error term.

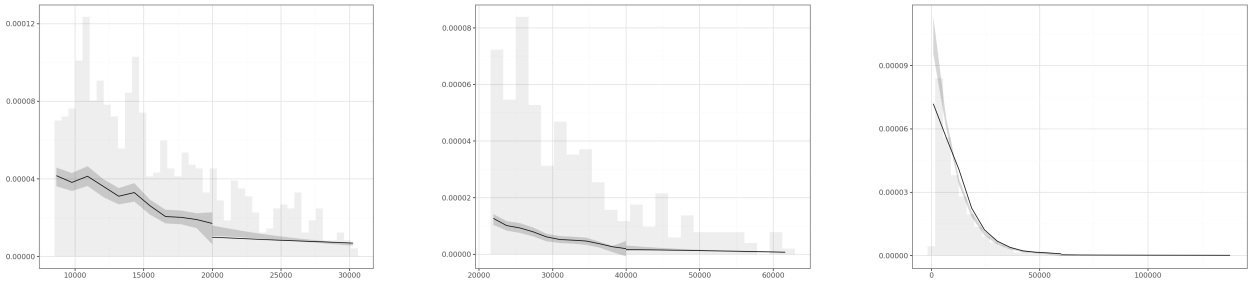
In the second stage, we regress the outcome Y_i on the predicted treatment status $\hat{S}_i$ obtained from the first stage. We use the same bandwidth calculated in the first stage, therefore taking into account the same effective sample:

$$Y_i = \beta S(C_i) + \rho f(p_i) + \sum_{k=1}^{K} \phi_k FPM_{ik} + \phi_i + \epsilon_i,$$

where ϵ_i is the error term. This two-stage approach, adapted for multiple cutoffs, allows us to estimate the local average treatment effects (LATE) of the policy-induced improvements in internet speed.

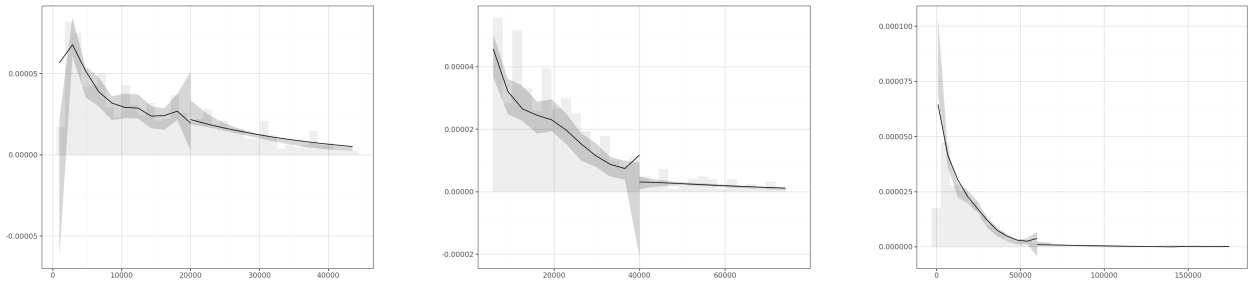
This targeted investment in backhaul capacity is expected to influence labor outcomes by enhancing connectivity and reducing information frictions within municipalities. Improved internet speeds can lower transaction costs, facilitate more efficient communication and access to digital platforms, and promote the adoption of modern technologies in the workplace. These channels

[10]This fund consists of automatic federal transfers established by the Federal Constitution of Brazil (Art. 159 lb). FPM transfers amount to 75 percent of all federal transfers and, according to the rules that regulate the allocation of these funds, municipal governments must spend 15 percent of them on education and 15 percent on health care, while the remainder is unrestricted.



(a) 20k (b) 40k (c) 60k

Figure 2.5: Density Plots of Manipulation Tests: Radio



(a) 20k (b) 40k (c) 60k

Figure 2.6: Density Plots of Manipulation Tests: Fiber

may contribute to changes in productivity, job creation, and ultimately, wages.

2.3.4 Internal Validity Tests

In our setting, we acknowledge that manipulation of the running variable could potentially compromise the validity of our estimates. This would violate the continuity assumption of our estimator, meaning that treated and untreated units may differ systematically beyond just their assignment to the policy. To address this concern, we conduct manipulation tests following Cattaneo et al. (2018). This approach allows us to detect any significant discontinuities in the density of the running variable at the cutoff, which would indicate potential manipulation and thus undermine the credibility of our design, by introducing bias in the estimated treatment effects.

Figures 2.5 and 2.6 display the density plots for the manipulation tests at each cutoff for the different technologies, while Table 2.4 presents the corresponding test statistics.

Both the visual evidence and the statistical results indicate signs of manipulation at the second

Table 2.4: Manipulation Tests

Technology	Cutoff	Bw.	N. Obs.	Test Statistic $T_p(h)$	p-value
Radio	20 k	3,786	265	-1.1813	0.2375
	40 k	6,011	68	0.6569	0.5112
	60 k	30,819	175	1.6592	0.0971*
Fiber	20 k	7,034	236	-0.5148	0.6067
	40 k	11,410	96	-2.1608	0.0307***
	60 k	38,175	223	-0.5955	0.5515

Notes: Tests are conducted at the cutoffs of 20,000; 40,000; and 60,000 inhabitants.

and third cutoffs, 40k and 60k, respectively. The manipulation issues may stem partly from the small sample sizes observed around these cutoffs, which can lead to less stable density estimates and increase the likelihood of observing apparent discontinuities. Additionally, Figure 2.3 shows evidence of non-continuous support, which suggests that the distribution of the running variable may not be smooth around these thresholds.

Brollo et al. (2013) and Litschig and Morrison (2012) observe that while local elites in Brazil had incentives and some ability to manipulate population counts, their control over the process was likely imperfect, making it implausible that manipulation systematically violated the continuity assumption. Moreover, the population thresholds were adjusted based on national population growth, making it difficult for municipalities to precisely anticipate their positions relative to the cutoffs. Since population estimates were derived from the most recent Census (2000) and the policy was introduced only in 2008, local governments had limited ability to manipulate their classification in advance. Additionally, the equidistant nature of the thresholds makes it unlikely that central government officials set them to favor specific political groups. Empirical evidence from these studies also shows no systematic over-representation of politically aligned municipalities near the cutoffs, further supporting the validity of the design.

However, issues observed at the second and third cutoffs may also stem from the discrete nature of the running variable, which can lead to irregularities in its support around these thresholds. Unlike a continuous running variable where units are evenly distributed near the cutoff, a discrete running variable with gaps can cause density-based manipulation tests to produce misleading results, as apparent discontinuities may emerge simply due to the limited number of observations at specific points. This lack of smooth support may also impact estimation, as standard nonparamet-

Table 2.5: Covariance Balance Tests

Poly. Order		Covariates (pre-treatment)											
		Cutoff 20k				Cutoff 40k				Cutoff 60k			
		log(FPM Transfers)	Fertility Rate	HDI-M	log(GDP)	log(FPM Transfers)	Fertility Rate	HDI-M	log(GDP)	log(FPM Transfers)	Fertility Rate	HDI-M	log(GDP)
Panel A: Radio													
1	Est.	-0.097	-0.068	0.029	0.048	0.187	6.722	0.049	-8.416	-0.068	0.689	-0.038	0.157
	p-value	0.843	0.700	0.163	0.973	0.372	0.454	0.782	0.416	0.188	0.075	0.019	0.562
	Bw.	2352	2091	4337	1475	1585	773	1424	836	7911	8155	8016	9968
	N.	171	149	296	99	12	5	12	6	27	27	27	29
Panel B: Fiber													
2	Est.	-0.184	0.397	-0.059	-0.658	0.395	0.250	-0.010	0.256	-0.169	0.613	-0.054	-0.341
	p-value	0.306	0.091	0.076	0.009	0.630	0.916	0.999	0.387	0.184	0.897	0.773	0.429
	Bw.	1357.8	2708.5	2504.5	2246.0	1591.0	1875.3	2522.4	2135.1	11699.1	13274.2	13928.0	14642.5
	N.	43	96	90	81	12	14	22	17	32	34	34	35

Notes: Tests are conducted at three different cutoffs (20k, 40k, 60k) using triangular kernel and linear polynomial. "Est." indicates the point estimate, "B.w." denotes the bandwidth, and "N." the number of observations. Robust bias-corrected standard errors.

ric techniques rely on a sufficient density of observations close to the cutoff. When the sample size is limited, these methods may require extrapolation beyond the immediate neighborhood of the discontinuity, weakening internal validity (Lee and Card, 2008). Consequently, observed irregularities in density estimates at these cutoffs should be interpreted with caution, as they may reflect structural limitations of the data rather than true manipulation or violations of the identifying assumptions.

We present in Table 2.5 the covariance balance tests. I replace the left-hand variables of the main specification with other pre-treatment variables, including the logarithm of FPM transfers, fertility rates, HDI-M (Human Development Index at the municipality level), and the logarithm of GDP (Gross Domestic Product).

For radio, the first and second cutoffs do not exhibit statistically significant differences in the covariates between the treatment and control groups (i.e., municipalities just to the right and left of the discontinuity). However, at the third cutoff, fertility rates and HDI-M are unevenly distributed between the groups. For fiber, the log of GDP shows an unequal distribution at the first cutoff, while the second and third cutoffs do not indicate significant imbalances in covariates.

The observed imbalances in pre-treatment covariates at certain cutoffs also likely stem from specification error rather than true discontinuities. When a second (and third) order polynomial is used in the estimation, none of the previously detected imbalances remain statistically significant. This aligns with findings in Litschig and Morrison (2012). Additionally, joint F-tests do not reject the null hypothesis of no systematic pre-treatment differences in local development or overall public resources, providing further evidence that treatment and control municipalities were comparable prior to the intervention. Nonetheless, the first cutoff for radio remained robust across all tests, even when varying specifications.

Table 2.6: First Stage Estimates: Backhaul Capacity - Radio

	Pooled		20k		40k		60k	
	Tri.	Uni.	Tri.	Uni.	Tri.	Uni.	Tri.	Uni.
1. Poly. Order: Linear								
Est.	6.84**	6.91**	5.14*	4.75*	-0.463	-3.98	0.966	-0.14
S.E.	(3.13)	(3.06)	(2.58)	(2.60)	(6.66)	(4.92)	(17.43)	(19.23)
Bw.	5812	4327	5412	3761	1221	1986	10462	5978
N.	513	344	386	269	9	15	29	19
2. Poly. Order: Quadratic								
Est.	6.90*	7.46**	4.17	3.96	-4.62	-19.82	-2.70	-3.09
S.E.	(3.66)	(3.93)	(3.63)	(3.27)	(11.51)	(12.56)	(21.58)	(23.00)
Bw.	6953	5729	5246	4495	2178	1889	17233	8393
N.	637	501	364	307	19	14	34	28
3. Poly. Order: Cubic								
Est.	7.39*	8.42**	4.83	3.96	11.11	20.82	-1.42	-1.82
S.E.	(4.59)	(4.59)	(5.39)	(4.38)	(17.33)	(17.62)	(22.61)	(26.04)
Bw.	8401	6643	5539	5702	3290	2683	20532	12536
N.	796	610	403	417	24	22	34	30

Notes: This table reports first-stage estimates for Backhaul Capacity for Radio technology using two different kernels (Triangular and Uniform) across different population size cutoffs 20k, 40k, and 60k) and a pooled model with normalized score. "Est." indicates the point estimate, "S.E." represents the robust bias-corrected standard error, "B.w." denotes the bandwidth, and "N." the number of observations.

2.4 Results

This section presents the results of the empirical model. We show the first and second stage estimates, followed by an investigation on heterogeneous treatment effects. We also document the effects of fast internet on internet use.

2.4.1 First Stage

Tables 2.6 and 2.7 present the first stage estimates of backhaul capacity for municipalities receiving microwave radio and fiber-optic technologies, respectively, using three different polynomial orders (linear, quadratic, and cubic) and two kernel weighting schemes (triangular and uniform) across varying policy population size cutoffs (20,000, 40,000, and 60,000; a normalizing-and-pooling approach is also shown (Cattaneo et al., 2016).

Within municipalities receiving microwave radio infrastructure (Table 2.6), the linear models

Table 2.7: First Stage Estimates: Backhaul Capacity - Fiber

	Pooled		20k		40k		60k	
	Tri.	Uni.	Tri.	Uni.	Tri.	Uni.	Tri.	Uni.
1. Poly. Order: Linear								
Est.	1.54	2.09	4.88	5.44	-3.26	-19.09	-11.45	-12.28
S.E.	(4.93)	(7.20)	(11.16)	(8.42)	(2.88)	(7.20)	(24.24)	(23.92)
Bw.	3096	2551	4944	5396	1798	943	14987	10935
N.	155	122	166	180	13	7	35	31
2. Poly. Order: Quadratic								
Est.	7.18	11.75	3.00	8.26	-11.53	2.24	-25.86**	-19.36**
S.E.	(9.12)	(15.84)	(4.94)	(8.07)	(8.92)	(9.67)	(43.75)	(42.27)
Bw.	6683	5956	6223	3211	2782	1887	16527	11981
N.	294	265	203	119	26	15	36	34
3. Poly. Order: Cubic								
Est.	-0.06	5.81	-0.28	15.13*	-21.88*	0.97	-121.49	-127.83
S.E.	(14.21)	(7.54)	(14.4)	(9.47)	(12.62)	(39.03)	(25.72)	(89.91)
Bw.	5864	6823	5211	3489	3473	2776	19390	13707
N.	263	297	174	124	34	26	36	34

Notes: This table reports first-stage estimates for Backhaul Capacity for Fiber technology using two different kernels (Triangular and Uniform) across different population size cutoffs (20k, 40k, and 60k) and a pooled model with normalized score. "Est." indicates the point estimate, "S.E." represents the robust bias-corrected standard error, "Bw." denotes the bandwidth, and "N." the number of observations.

indicate positive and statistically significant effects at the first cutoff (20k) for both kernels, with treatment effects of approximately 5%, whereas the point estimates become statistically insignificant at the second (40k) and third (60k) cutoffs. The quadratic and cubic specifications show statistically insignificant point estimates and larger standard errors, suggesting sensitivity to model specification and bandwidth choice, which is particularly pronounced when working with small sample sizes.

In contrast, within municipalities receiving investment in fiber-optics infrastructure (Table 2.7) we reveal a different pattern: the linear specification largely produces insignificant effects, while the quadratic models show substantial negative and statistically significant effects at the 60k cutoff. The cubic specification for fiber also produces variable results with some large negative estimates, particularly at the second and third cutoffs, albeit with considerable imprecision.

As the distribution of municipalities in Brazil across both technologies is fairly concentrated around the first cutoff—about 90% of the effective sample—this concentration leads to a lack of

precision in the point estimation for the second and third cutoffs, as well as abnormally large standard errors indicating great uncertainty. These findings underscore that the estimated impact of the policy instrument on backhaul capacity is highly contingent on the technology under consideration, as well as on the chosen polynomial order.

Consequently, only the linear specification for microwave radio at the first cutoff (20k) produces robust and theoretically consistent estimates, making it the sole valid first-stage specification. In contrast, no specification for fiber yields consistently reliable results, underscoring the instrument's limitations for this technology.

Therefore, we adopt the linear specification with a triangular kernel for microwave radio at the 20k cutoff as the primary model for my second-stage analysis. For completeness, we present second-stage results for fiber in the Appendix B.1.

2.4.2 Fast Internet and Internet Use

Figure 2.7 presents estimates assessing the policy impact on fixed broadband access in Brazil, measured by internet access density per 100 households with data sourced from Anatel. The effect on internet use appears to rise beginning in 2012 and persists thereafter, suggesting a roughly two-year lag before the policy's impact fully materialized at the household level.

Furthermore, as the bandwidth sizes decrease over time – resulting in smaller sample sizes – the standard errors increase, indicating greater uncertainty in the estimates within these narrower windows, likely due to the heightened sensitivity to local variations.

This delayed effect could reflect the time needed for internet service providers to finalize last-mile connections, roll out commercial plans, and market their services effectively. Households, in turn, may have taken additional time to adopt and integrate these new services, which would explain why the observed impact does not manifest immediately but becomes more pronounced a couple of years after the policy roll-out ends.

Figure 2.8, in contrast, consistently shows null effects over the same period, indicating that the policy's influence on fiber-based connections may have been weaker or altogether absent. One possibility is that the infrastructure investments in fiber did not translate into a comparable increase in household-level adoption—perhaps due to higher deployment costs, limited local service offerings, or other market barriers. Consequently, even though fiber technology was deployed, it may not have reached a critical mass of subscribers needed to register a discernible effect on internet access density.

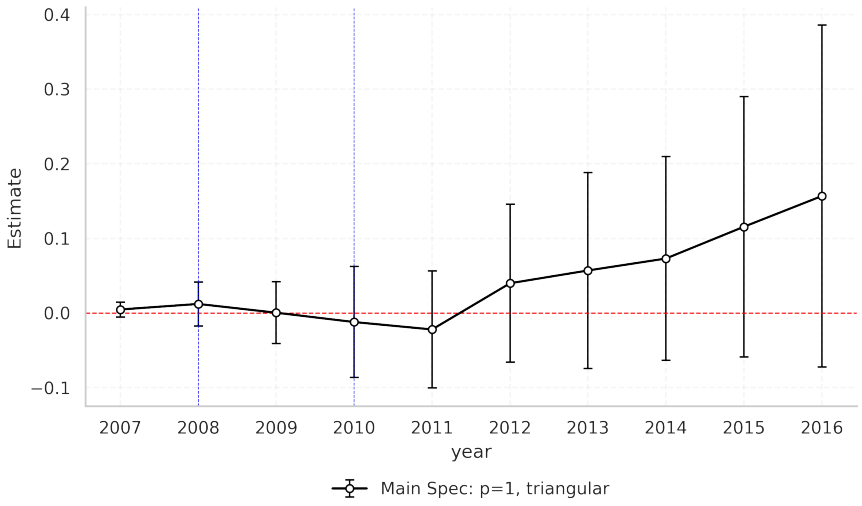


Figure 2.7: Effects on Fast Internet on Internet Use - Radio

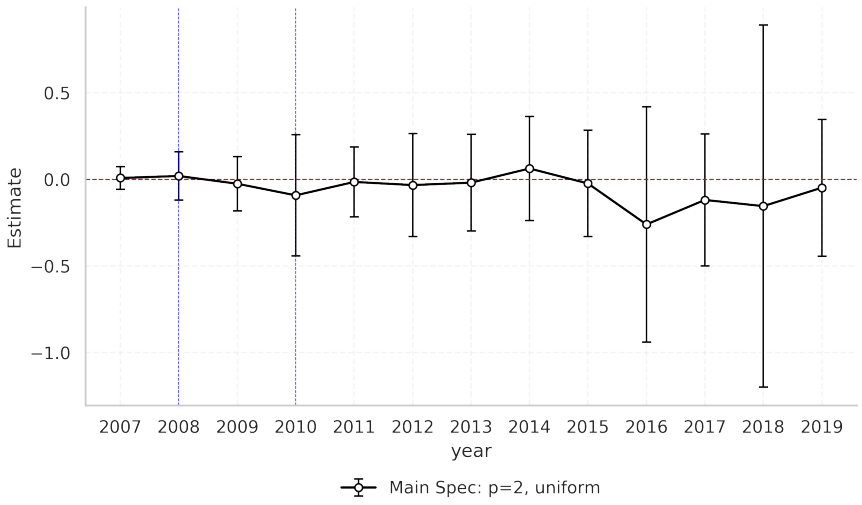


Figure 2.8: Effects on Fast Internet on Internet Use - Fiber

2.4.3 Second Stage

This section examines the causal impact of Brazil's broadband expansion policy – specifically, investments in radio-based backhaul infrastructure in municipalities located near the 20,000-population size threshold – on three municipal-level labor market outcomes: net job creation, hours worked, and wages.

Table 2.8 reports second-stage estimates of the policy's effects on these outcomes between 2006 and 2016, leveraging a fuzzy regression discontinuity design. The analysis employs a specification with a linear polynomial and triangular kernel. This specification was selected due to its superior first-stage performance in addressing endogeneity concerns.

Table 2.8: Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) - Radio

	Pre-treatment		Treatment Roll-out			Post-treatment					
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation											
Est.	-44.23	-5.50	21.31	-19.09	15.48	10.42	-46.74	33.94	22.96	2.49	14.72
S.E.	(68.28)	(8.58)	(18.09)	(12.72)	(14.00)	(23.54)	(34.00)	(24.18)	(17.24)	(12.31)	(12.84)
Bw.	5161.10	5246.53	3814.22	4398.79	6127.06	5597.15	5036.82	3910.92	5352.78	8131.61	6121.56
N.	2161	2161	2159	2161	2161	2161	2161	2161	2161	2161	2161
Panel B: Hours Worked											
Est.	0.1161	0.0602	-0.1233	-0.0516	-0.0847	-0.1592	-0.1979	-0.1165	-0.1775	-0.2942*	-0.2811*
S.E.	(0.2974)	(0.2018)	(0.2083)	(0.2063)	(0.2192)	(0.1833)	(0.1399)	(0.1506)	(0.1530)	(0.1648)	(0.1691)
Bw.	5260.24	7729.10	6741.53	6545.18	5757.49	6416.95	7671.73	8204.74	7290.26	6696.84	6047.46
N.	2161	2161	2158	2161	2161	2161	2161	2161	2161	2160	2161
Panel C: Log(Wages)											
Est.	-0.01084	-0.01738	-0.01198	-0.01490	-0.00880	-0.01227	-0.01188	-0.03406*	-0.02488*	-0.01097	-0.00248
S.E.	(0.01174)	(0.01232)	(0.01389)	(0.01106)	(0.01244)	(0.00837)	(0.00740)	(0.01898)	(0.01482)	(0.00681)	(0.00742)
Bw.	5708.62	5489.80	5167.31	5788.80	5669.10	7970.29	7929.26	5281.60	5412.03	7815.44	6197.57
N.	2161	2161	2159	2161	2161	2161	2161	2161	2161	2161	2161

Notes: This table presents second-stage estimates for three labor outcomes – net job creation, hours worked, and log(wages) – across years 2006–2016. Years 2006–2007 serve as falsification tests, 2008–2010 represent the policy roll-out period, and 2011–2016 reflect the policy impact period. "Est." indicates the point estimate, "S.E." represents the robust bias-corrected standard error, "Bw." denotes the bandwidth, and "N." the number of observations.

The pre-treatment period (2006–2007) serves as a falsification test, showing no significant effects, which supports the validity of the identification strategy. During the treatment roll-out period (2008–2010), estimates for net job creation fluctuate without a clear pattern, suggesting no immediate large-scale employment effects. Similarly, hours worked and wages exhibit no significant changes in this period, indicating that labor market adjustments may take time. In the post-treatment period (2011–2016), hours worked show a declining trend, with statistically significant reductions in 2015 and 2016 (-0.29 and -0.28, respectively), suggesting that internet expansion may

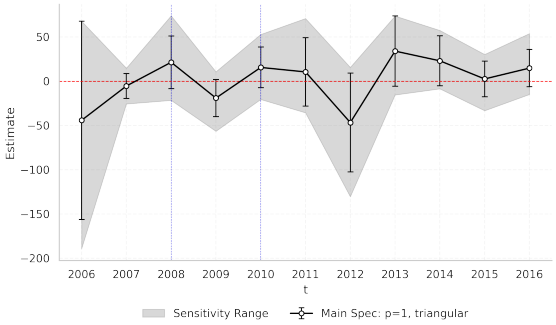
have led to reduced labor demand or productivity changes. Wages exhibit a small but significant decline in 2013 and 2014 (3.4% and 2.5%, respectively), although the effect does not persist in later years. Overall, the findings suggest that while internet expansion via microwave radio technology influenced labor market dynamics, its effects were more evident in work intensity (hours worked) and earnings rather than net job creation.

Figure 2.9 displays the point estimates and corresponding standard errors from the main specification, along with sensitivity bands for alternative specifications – quadratic and cubic polynomials, as well as uniform kernel. The figure highlights how the estimated effects vary depending on the polynomial degree and kernel choice. In particular, for the outcome of hours worked, the main specification – which employs a linear polynomial – tends to underestimate the negative impact of the broadband policy. This mechanical limitation arises because a linear specification imposes a constant slope over the range of the running variable, thereby failing to capture potential nonlinear relationships that may be present. As a result, the linear model smooths over more pronounced declines in hours worked that could be better detected with higher-order polynomials, which allow for curvature in the estimated relationship.

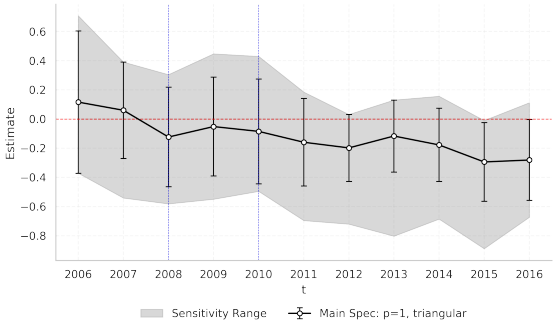
Overall, the regression-discontinuity evidence indicates that enlarging municipal backhaul capacity, by itself, has not triggered a sweeping transformation of local labour markets. The yearly estimates reveal a gradual fall in work intensity—average hours worked edge downward after the policy's rollout and reach statistically significant reductions only in 2015 and 2016—yet the wide confidence bands around those coefficients caution against treating the effect as either large or persistent.

The pooled specification reported in Table B.3 (Appendix B.1) reinforces this mixed picture. Aggregating all post-treatment years (2011–2016) yields a clear spike of roughly eleven additional jobs per municipality at the 20,000-population threshold, but the employment surge is offset by a 0.53-hour contraction in weekly hours. Average log wages sit effectively unchanged—the point estimate is near zero and far from statistical significance. Additionally, we report in Appendix B.1 how the availability of faster internet impacts firms by looking at net firm entry and exports. We show that faster Radio-based connectivity neither catalyzed sustained entry of new firms nor boosted export activity in the affected municipalities. On the other hand, by looking at value added to different industries, we observe small but statistically significant post-treatment effects in services.

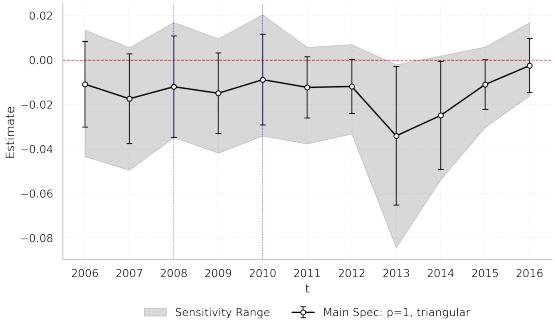
Taken together, these findings suggest that the broadband expansion reallocated labor rather



(a) Net Job Creation



(b) Hours Worked



(c) Log of Wages

Figure 2.9: Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) - Radio

than boosting overall productivity or pay in the short run. However, it is important to emphasize that these are pooled, aggregate results by year. Given the heterogeneous structure of Brazil's labor market – including differences in skill levels and industrial composition – local average treatment effects may mask non-constant impacts across different subpopulations.

Therefore, the next section explores heterogeneous treatment effects to investigate whether specific groups of workers or industries might experience different impacts from the expansion of broadband infrastructure.

2.4.4 Fast Internet and Employment in Skilled and Unskilled Jobs

Educational Level

Conducting a heterogeneous analysis by education level is particularly salient in evaluating a fast internet infrastructure expansion policy, as the benefits of such investments are inherently tied to individuals' ability to leverage digital tools. Education serves as a critical proxy for skill-based disparities in accessing and utilizing high-speed internet, shaping both economic opportunities and labor market outcomes. For instance, high-education workers – often employed in knowledge-intensive functions – are more likely to capitalize on improved connectivity through remote work, digital innovation, or participation in global markets, potentially driving wage growth and job creation in these fields. Conversely, low-education workers, who are overrepresented in manual or routine sectors, may experience muted benefits or even displacement if automation accelerates with digital infrastructure upgrades. By disaggregating effects, this analysis reveals whether the policy exacerbates the digital divide or fosters inclusive growth.

We group employees into three educational brackets based on educational achievement. We define an individual as being high-educated if the highest degree attained is a college degree (bachelor, master, or PhD), as medium-educated if it is a high school degree, and as low-educated if a high school degree has not been attained.

Table 2.9 presents second-stage estimates of a policy's impact on municipal labor markets, stratified by education level (high, medium, low) across pre-treatment (2006–2007), treatment roll-out (2008–2010), and post-treatment (2011–2016) periods.

For high-education workers (tertiary education), post-treatment effects reveal statistically significant declines in hours worked and wages. Hours worked (Panel B) fell sharply, with point estimates of -0.67 ($p<0.10$) in 2011, -0.52 ($p<0.01$) in 2012, and -0.37 ($p<0.05$) in 2013, suggest-

Table 2.9: Second-stage Estimates on Municipal Labor Market Outcomes by Education Level and Year (2006–2016)

		Pre-treatment		Treatment Roll-out			Post-treatment					
Education		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation												
High	Est.	1.45	-4.06	5.93	3.63	-3.09	-5.08	1.32	1.32	-0.47	-1.48	-8.08
	S.E.	(5.17)	(2.43)	(4.93)	(8.01)	(3.26)	(3.84)	(5.41)	(5.41)	(4.64)	(3.33)	(6.77)
	Bw.	7149.59	7632.60	4333.63	4010.66	5078.13	5737.19	5466.47	5466.47	6040.27	9138.76	4243.26
	N.	545	594	294	280	349	422	396	396	469	734	290
Medium	Est.	6.52	3.14	2.49	15.43	-8.07	-8.74	10.06	-18.14	-1.05	0.13	13.65
	S.E.	(8.29)	(6.41)	(7.88)	(11.89)	(8.76)	(9.49)	(8.46)	(11.67)	(6.93)	(3.06)	(10.88)
	Bw.	5740.30	6455.63	5287.65	4723.05	3659.40	5488.68	6738.96	5031.13	4221.58	7136.84	7505.94
	N.	422	500	371	325	263	398	516	346	289	547	582
Low	Est.	-92.759	-1.455	14.107	3.484	8.074	8.506	-17.958	3.241	19.579*	0.638	14.528
	S.E.	(72.650)	(6.122)	(12.433)	(6.709)	(7.722)	(12.852)	(14.574)	(10.144)	(10.190)	(8.562)	(9.438)
	Bw.	5047.239	5813.031	4483.794	8033.196	7155.544	10302.969	9559.282	6199.350	6595.330	8638.801	6371.946
	N.	348	432	306	640	548	860	798	479	509	688	495
Panel B: Hours Worked												
High	Est.	-0.00	-0.12	-0.41	-0.48	-0.43*	-0.67*	-0.52***	-0.37**	-0.20	-0.32	-0.31
	S.E.	(0.28)	(0.29)	(0.26)	(0.26)	(0.25)	(0.26)	(0.26)	(0.24)	(0.21)	(0.23)	(0.23)
	Bw.	6904.26	7626.17	6777.57	6552.40	6982.59	7130.60	7758.55	7956.56	8990.30	7945.66	7945.66
	N.	528	591	519	503	535	545	610	627	722	626	626
Medium	Est.	0.02	-0.01	-0.14	-0.07	-0.05	-0.11	-0.08	-0.08	-0.21	-0.24	-0.23
	S.E.	(0.31)	(0.20)	(0.23)	(0.21)	(0.23)	(0.20)	(0.15)	(0.14)	(0.15)	(0.16)	(0.15)
	Bw.	5282.31	9122.60	7176.17	7400.99	6289.85	7039.30	8202.86	9159.19	7641.41	7251.72	6794.47
	N.	369	734	549	573	492	542	650	737	597	553	523
Low	Est.	0.30	0.16	0.10	0.07	-0.01	-0.03	-0.04	0.02	-0.07	-0.14	-0.13
	S.E.	(0.31)	(0.21)	(0.20)	(0.21)	(0.19)	(0.15)	(0.12)	(0.14)	(0.14)	(0.15)	(0.14)
	Bw.	6119.83	7666.68	7837.12	6572.77	6043.51	6944.57	7992.54	8413.85	7176.42	6658.97	6902.25
	N.	474	598	614	507	469	535	631	672	549	513	531
Panel C: Log(Wages)												
High	Est.	-0.001	-0.001	0.001	-0.003	-0.006	-0.014	-0.015*	-0.023*	-0.032**	-0.028	-0.023
	S.E.	(0.011)	(0.020)	(0.023)	(0.013)	(0.020)	(0.018)	(0.011)	(0.014)	(0.016)	(0.013)	(0.013)
	Bw.	8680.197	7170.285	6594.193	7144.243	5460.849	6088.751	8192.881	7334.054	6360.464	7454.007	7334.054
	N.	687	545	504	547	394	470	648	564	495	578	564
Medium	Est.	-0.002	-0.013	-0.005	-0.005	0.001	-0.009	-0.010	-0.016	-0.011	-0.002	0.000
	S.E.	(0.008)	(0.010)	(0.011)	(0.010)	(0.010)	(0.007)	(0.007)	(0.010)	(0.011)	(0.007)	(0.008)
	Bw.	7211.154	5863.071	6394.680	5991.735	6272.941	8629.712	7974.591	7006.148	6820.465	7122.458	6338.367
	N.	549	445	496	462	489	686	630	538	526	544	492
Low	Est.	-0.012	-0.018	-0.014	-0.022	-0.018	-0.010	-0.009	-0.010	-0.005	-0.008	0.010
	S.E.	(0.015)	(0.013)	(0.014)	(0.015)	(0.016)	(0.009)	(0.008)	(0.009)	(0.006)	(0.010)	(0.008)
	Bw.	7058.589	7047.321	6940.127	5580.334	5718.411	6647.706	7681.055	8210.234	8267.065	5922.447	6940.436
	N.	541	542	534	407	419	512	602	651	654	455	535

Notes: This table presents second-stage estimates for three labor outcomes — net job creation, hours worked, and log(wages) — stratified by education level (high/medium/low) across years 2006–2016. Education categories correspond to workers with tertiary education (high), secondary education (medium), and primary/no formal education (low). Pre-treatment (2006–2007), treatment roll-out (2008–2010), and post-treatment (2011–2016) periods follow the policy timeline. Standard errors (S.E.) are robust and bias-corrected; bandwidth (Bw.) and observation counts (N.) are reported.

ing sustained reductions. Similarly, wages (Panel C) exhibited a gradual decline, reaching -3.2% ($p<0.05$) by 2014. Net job creation (Panel A) for this group showed no consistent trends, with volatile and non-significant estimates across years.

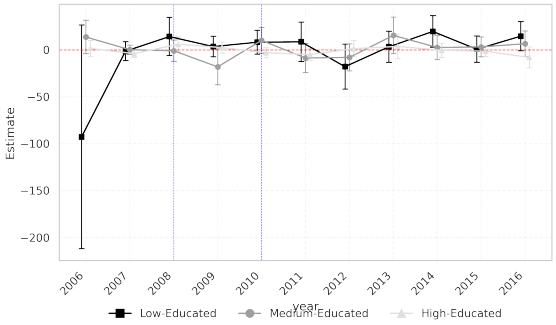
Medium-education workers (secondary education) experienced pronounced volatility in net job creation, including a sharp pre-treatment increase (approximately 6 in 2006) and a post-treatment drop (approximately 18 in 2013), though estimates lacked statistical significance. Hours worked and wages for this group remained largely stable, with minor non-significant fluctuations.

Low-education workers (primary/no formal education) saw a significant positive shift in net job creation in 2014 of approximately 20 jobs ($p<0.10$), contrasting with pre-treatment instability. Hours worked and wages for this group showed negligible or non-significant changes, except for a modest post-treatment wage increase in 2016 of 1%.

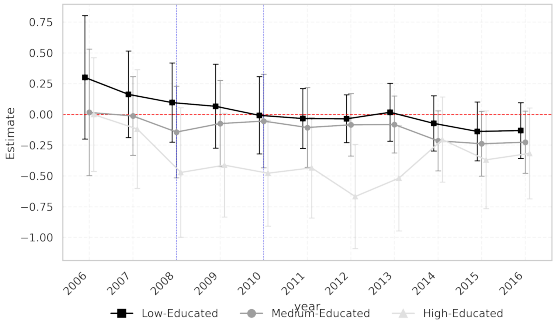
Notably, precision varied across education levels and years, as reflected in bandwidths (ranging from 4,010 to 9,139 for high education) and sample sizes (ranging from 306 to 860 for low education). Pre-treatment estimates were largely non-significant, aligning with our identification assumptions, while post-treatment effects highlight divergent outcomes by skill level. These results suggest that the policy disproportionately affected high-skill workers, potentially due to labor market substitution or reduced demand for tertiary-educated labor near the cutoff.

Figure 2.10 displays the point estimates and corresponding standard errors of Table 2.9.

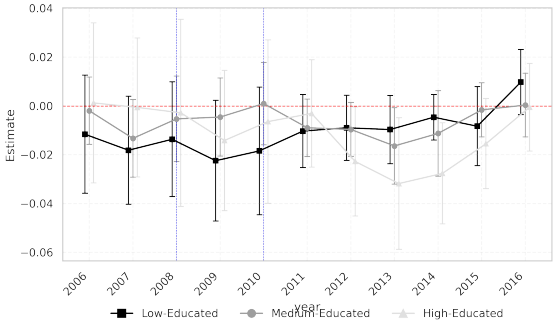
The findings in Table 2.9 align closely with the hypothesized skill-specific dynamics of a fast internet infrastructure rollout. The significant post-treatment wage and hours reductions for high-education workers may reflect automation or outsourcing pressures amplified by improved connectivity, as firms near the rollout cutoff could adopt digital tools that substitute high-skill roles (e.g., remote management). Conversely, the delayed 2014 surge in low-education net job creation suggests that infrastructure expansion eventually stimulated demand for manual or platform-based roles (e.g., e-commerce logistics), albeit without corresponding wage gains. This delayed impact is potentially related to the delayed adoption of new digital tools through enhanced internet connection, highlighting important adjustment costs for firms. Therefore, these divergences underscore how fast internet connection interacts with preexisting skill hierarchies: high-education workers faced displacement risks despite their adaptability, while low-education groups saw job gains concentrated in low-wage sectors (see next section).



(a) Net Job Creation



(b) Hours Worked



(c) Log of Wages

Figure 2.10: Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) per Educational Level - Radio

Occupational Skill Level

We propose an alternative approach that derives occupational skill levels directly from the nature of work by leveraging textual descriptions from the Brazilian Classification of Occupations (CBO). This unsupervised, task-based method, combined with educational stratification, yields a more comprehensive understanding of skill differentiation in the labor market. By analyzing the semantic content of occupation titles and their associated tasks, this approach captures latent patterns in task complexity, managerial responsibilities, and specialized knowledge. It offers a granular perspective that aligns with the CBO's emphasis on knowledge, skills, and personal attributes. Combining this task-based classification with educational stratification allows for a more holistic understanding of skill differentiation in the labor market.

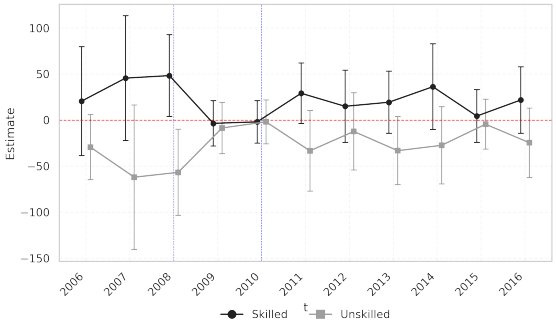
To classify occupations as routine or non-routine (or unskilled and skilled), we first preprocess the CBO descriptions by concatenating occupation titles and task lists into a single input – for instance, merging titles like "Maintenance Operator" with tasks such as "Assemble Refrigeration Piping" or "Maintenance Director" with tasks like "Manage People" and "Administer Production". These inputs are then encoded into dense vector representations using SentenceTransformer (Reimers and Gurevych, 2019), which capture linguistic patterns, task complexity, and contextual relationships within the CBO texts. The resulting embeddings reflect nuances such as managerial verbs (e.g., manage, define) indicative of non-routine tasks and operational verbs (e.g., assemble, apply) linked to routine tasks. Finally, we apply K-Means clustering to partition the occupations into two groups based on the Euclidean distance between their embeddings, grouping those with semantically similar tasks into clusters that represent distinct skill levels. Table 2.10 presents the main point estimates.

The empirical evidence in Table 2.10 suggests that effects on net job creation (Panel A) is statistically insignificant and unstable across years, precluding robust conclusions about employment effects. However, the results for hours worked (Panel B) and wages (Panel C) reveal meaningful skill-based disparities tied to the policy. For skilled workers, post-treatment years (2011–2016) show statistically significant reductions in both hours worked and wages. Hours worked declined sharply, with coefficients ranging from 0.57 ($p<0.05$) in 2012 to 0.65 ($p<0.05$) in 2016, while wages fell by approximately 3% ($p<0.10$) in 2013 and 2.5% ($p<0.10$) in 2014. These patterns suggest that skilled labor markets near the policy cutoff experienced sustained negative pressures on labor demand, potentially due to automation or reduced demand for non-routine tasks as firms adopted

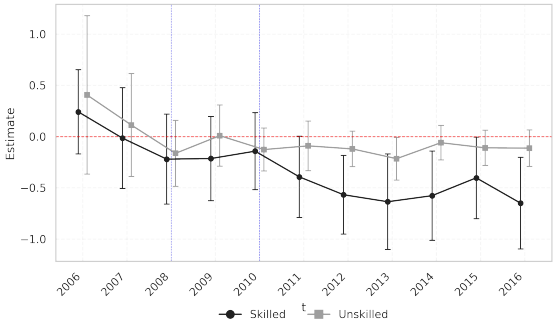
Table 2.10: Second-stage Estimates on Municipal Labor Market Outcomes by Occupation Skill Level and Year (2006–2016)

		Pre-treatment		Treatment Roll-out			Post-treatment					
Skill Level		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation												
Skilled	Est.	20.64	45.62	48.34*	-3.51	-1.85	29.19	15.01	19.43	36.28	4.41	21.83
	S.E.	(36.03)	(41.24)	(27.03)	(14.97)	(14.08)	(20.08)	(23.99)	(20.51)	(28.35)	(17.49)	(21.97)
	Bw.	4274.35	4451.82	6130.93	5873.95	7775.14	6100.92	5616.43	4627.67	3581.84	5347.08	5477.10
	N.	291	304	475	447	611	473	408	318	259	380	397
Unskilled	Est.	-29.31	-61.96	-56.68*	-8.51	-1.85	-33.31	-12.28	-33.18	-27.32	-4.42	-24.54
	S.E.	(21.50)	(47.85)	(28.48)	(17.00)	(14.56)	(26.75)	(25.62)	(22.55)	(25.53)	(16.52)	(22.94)
	Bw.	8075.18	4290.24	5973.15	4836.90	6417.34	4221.09	5314.52	4917.23	3889.47	5664.16	5443.92
	N.	644	294	460	330	496	289	375	338	275	413	393
Panel B: Hours Worked												
Skilled	Est.	0.24	-0.02	-0.22	-0.21	-0.14	-0.39	-0.57**	-0.64**	-0.58**	-0.40*	-0.65**
	S.E.	(0.25)	(0.30)	(0.27)	(0.25)	(0.23)	(0.24)	(0.23)	(0.28)	(0.27)	(0.24)	(0.27)
	Bw.	8577.71	4729.77	5630.70	5029.88	6103.35	4835.58	5087.12	4683.91	4590.54	5108.65	4235.45
	N.	683	325	409	346	473	330	350	321	316	353	290
Unskilled	Est.	0.41	0.11	-0.16	0.01	-0.13	-0.09	-0.12	-0.22*	-0.06	-0.11	-0.11
	S.E.	(0.47)	(0.31)	(0.20)	(0.18)	(0.13)	(0.15)	(0.11)	(0.13)	(0.10)	(0.10)	(0.11)
	Bw.	2654.91	3237.41	3942.58	4730.78	5104.98	3591.19	4782.26	3869.96	5084.21	4308.95	4847.03
	N.	195	230	277	325	351	260	329	275	350	295	331
Panel C: Log(Wages)												
Skilled	Est.	-0.006	-0.012	-0.011	-0.016	-0.005	-0.012	-0.009	-0.028*	-0.025*	-0.014	-0.006
	S.E.	(0.011)	(0.012)	(0.013)	(0.014)	(0.014)	(0.012)	(0.009)	(0.015)	(0.014)	(0.010)	(0.008)
	Bw.	7059.886	5807.467	6965.883	6524.734	7364.954	7378.432	8617.888	6460.150	6830.305	6854.285	8557.891
	N.	543	432	535	504	566	568	686	500	527	527	682
Unskilled	Est.	0.000	-0.015	-0.015	-0.015	-0.001	-0.007	-0.005	-0.021	-0.007	-0.004	0.005
	S.E.	(0.010)	(0.011)	(0.013)	(0.010)	(0.007)	(0.006)	(0.006)	(0.013)	(0.007)	(0.006)	(0.006)
	Bw.	6588.122	5362.323	7057.145	5317.948	5833.783	7479.604	7213.244	4711.686	5636.147	7443.305	6326.080
	N.	508	381	541	376	434	579	549	323	409	577	492

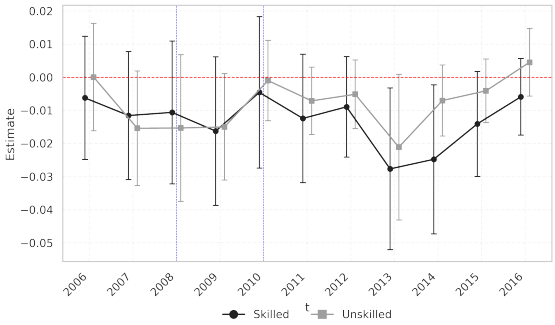
Notes: This table presents second-stage estimates for three labor outcomes — net job creation, hours worked, and log(wages) — stratified by skill/occupational level (skilled/unskilled) across years 2006–2016. Skill categories correspond to workers with non-routine tasks (skilled), and routine tasks (unskilled). Pre-treatment (2006–2007), treatment roll-out (2008–2010), and post-treatment (2011–2016) periods follow the policy timeline. Standard errors (S.E.) are robust and bias-corrected; bandwidth (Bw.) and observation counts (N.) are reported.



(a) Net Job Creation



(b) Hours Worked



(c) Log of Wages

Figure 2.11: Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) per Skill Level - Radio

digital tools enabled by faster internet infrastructure.

Unskilled workers, in contrast, saw smaller and less consistent effects. Hours worked declined modestly post-treatment (e.g., 0.22, $p<0.10$ in 2013), but wages remained stable, with no significant coefficients. This asymmetry implies that unskilled workers, while less affected by wage suppression, still faced reduced labor demand in routine sectors, possibly due to indirect competition or productivity shifts.

Figure 2.11 displays the point estimates and corresponding confidence intervals in a timeline. The diverging outcomes underscore how digital infrastructure expansion may exacerbate skill-based inequalities: skilled workers bear the brunt of wage and hours reductions, while unskilled workers experience muted or non-significant effects.

These findings align with theories of skill-biased technological change, where connectivity upgrades disproportionately disrupt high-skilled workers. While broadband internet has been shown to complement skilled workers in many cases, it can also lead to job displacement and wage stagnation

for certain groups of highly educated workers. For example, in the United States, the computerization risk of occupations has been linked to a high likelihood of job switching or non-employment, particularly for older workers and those with high levels of formal education (Fossen and Sorgner, 2022). This suggests that even highly educated workers may face negative labor market outcomes if their skills are not aligned with the demands of the digital economy.

2.4.5 Fast Internet and Employment Across Industries

This section examines how digital infrastructure's impact varies according to industry-specific characteristics. I stratify the data into four distinct industries (primary, manufacturing, commerce, and services) based on 2-digit Classification of Economic Activities (CNAE) codes. Table 2.11 displays the mapping of each 2-digit code to its corresponding industry classification, following the classification in Barbosa et al. (2021).

Industry heterogeneity analysis is important as digital infrastructure's labor market impacts depend on industry-specific exposure to automation, adaptability to digital tools, and task. We expect that industries with routine tasks (e.g., manufacturing) face higher displacement risks, while tech-complementary sectors (e.g., service) may gain from productivity or e-commerce growth. Testing these predictions validates the model's assumptions about skill- versus routine-biased technological change and informs targeted policies to address industry-specific vulnerabilities in the digital transition.

Table 2.12 presents the second-stage estimates on municipal labor market outcomes by industry and year.

Table 2.12 highlights asymmetric effects of high-speed internet upgrades (8mb to 16mb), with the service sector experiencing significant post-treatment declines in labor market outcomes. Between 2011 and 2016, service-sector hours worked fell sharply, culminating in a reduction of approximately one hour per week ($p<0.01$) by 2016, while wages dropped by 5.2% ($p<0.10$) in 2013. These results suggest that higher internet speeds, while enabling productivity-enhancing technologies (e.g., automation, remote work platforms), disproportionately disrupted service roles –potentially displacing in-person jobs (e.g., finance, real estate, administrative support) or compressing wages due to increased competition from remote labor markets.

In contrast, the commerce sector exhibited a transient job creation effect in 2015 (29 jobs, $p<0.10$), potentially linked to e-commerce or digital payment adoption facilitated by faster speeds. However, the absence of corresponding wage or hours gains suggests that these jobs were low-quality

Table 2.11: Industry Categorization

Sector	**Code**	**Description**
Primary	1-9	Agriculture, Livestock, and Related Services; Forestry Production; Fishing and Aquaculture; Coal Mining; Oil and Natural Gas Extraction; Metallic Mineral Extraction; Non-Metallic Mineral Extraction; Support Activities for Mineral Extraction.
Manufacturing	15-19, 20-37, 40-41, 45	Production of food and drinks; Tobacco industry; Textile industry (production of textiles and clothes); Leather processing and production of leather products; Water, gas, and electricity; Construction; Basic metallurgy; Production of metal products excluding machinery and equipment; Production of machinery and equipment; Production of office machinery and computer equipment; Production of electrical machinery and electric materials; Production of electronic material and communication equipment; Production of medical and precision machinery; Production of precision and optical instruments, automation machinery and clocks; Automotive industry; Production of other transportation equipment; Wood processing; Paper and cellulose production; Editing, printing, and reproduction of recordings; Production of coke, oil refining, production of nuclear fuels, and alcohol; Chemical industry, Rubber and plastic industry; Production of non-metallic minerals (glass, cement, etc.); Production of furniture and diverse industries; Recycling.
Commerce	50-52	Trade and repair of automobiles and motorcycles and fuel trade; Retail trade and repair of personal and domestic objects.
Services	55, 60-67, 70-74, 80, 85, 92	Mail and telecommunications; Informatics and related services; Research and development; Transportation-related activities and travel agencies; Business-to-business services; Cinematographic and audiovisual works, news agencies; Real estate; Hotels and food-related services; Financial intermediation; Insurance and pensions; Auxiliary activities related to insurance and pensions; Education; Health and social services.

Note.— This table shows the grouping of industries into "primary","manufacturing," "commerce," and "services" based on the 2-digit Classification of Economic Activities (CNAE) contained in the RAIS database.

Table 2.12: Second-stage Estimates on Municipal Labor Market Outcomes by Industry and Year (2006–2016)

		Pre-treatment		Treatment Roll-out			Post-treatment					
Education		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation												
Primary	Est.	24.54	-1.70	-10.17	-13.08	-353.31	1288.81	-156.64	87.26	261.03	76.32	-0.04
	S.E.	(28.80)	(19.33)	(30.95)	(38.24)	(446.49)	(2180.78)	(167.70)	(94.28)	(295.99)	(69.69)	(15.11)
	Bw.	6066.40	5649.35	5393.32	5574.98	4762.29	3728.18	4288.23	4389.88	4191.60	4042.35	6259.92
	N.	304	270	254	277	221	191	203	223	210	202	338
Manufacturing	Est.	16.72	23.94	-7.12	-7.70	35.92	2.50	-76.47	-30.64	-20.47	5.54	0.11
	S.E.	(17.10)	(20.95)	(18.74)	(23.04)	(35.71)	(41.77)	(67.55)	(27.78)	(31.37)	(23.85)	(26.07)
	Bw.	6595.28	4564.86	6264.92	4027.83	4076.72	4760.05	4669.89	4805.38	4616.41	4460.00	5177.18
	N.	364	230	365	223	225	266	260	277	267	267	310
Commerce	Est.	15.77	20.74	14.35	18.30	23.12	30.40	15.47	12.67	28.83	28.99*	-3.50
	S.E.	(12.41)	(13.62)	(9.78)	(9.36)	(15.59)	(21.63)	(13.07)	(12.08)	(19.29)	(16.66)	(9.01)
	Bw.	5101.05	4489.93	5816.24	6557.71	4972.72	4260.78	5067.29	6941.40	4128.12	4388.31	7042.91
	N.	343	304	429	503	342	290	348	532	285	299	542
Service	Est.	-31.90	-19.56	-16.20	-8.20	-3.93	-46.11	2.78	2.14	-30.94	-38.37	10.34
	S.E.	(19.44)	(18.19)	(18.37)	(14.70)	(17.65)	(28.16)	(15.51)	(18.63)	(21.24)	(23.88)	(17.26)
	Bw.	5265.32	4538.18	5990.80	6800.80	5149.47	6006.85	6232.57	5774.45	5284.45	4541.89	5551.56
	N.	369	309	462	524	358	465	484	424	369	310	404
Panel B: Hours Worked												
Primary	Est.	2.44	0.34	0.02	-0.25	-0.17	1.90	-5.89	0.02	39.05	3.31	-1.22
	S.E.	(6.77)	(0.66)	(0.64)	(0.60)	(0.70)	(6.48)	(12.88)	(0.36)	(80.20)	(3.99)	(1.50)
	Bw.	2645.26	3020.95	3198.33	2759.84	2534.85	2863.26	2987.40	5646.73	3111.57	2926.53	4151.70
	N.	125	141	149	133	125	142	149	276	158	151	196
Manufacturing	Est.	0.20	-0.05	0.98	0.22	0.00	0.12	0.39	0.26	0.24	-0.07	-0.25
	S.E.	(0.42)	(0.12)	(0.71)	(0.26)	(0.21)	(0.57)	(0.45)	(0.36)	(0.39)	(0.29)	(0.21)
	Bw.	3445.85	2677.72	4354.09	4504.71	2270.85	3787.26	3721.69	3336.41	3516.59	1746.90	3919.54
	N.	174	143	220	235	127	213	212	198	209	105	235
Commerce	Est.	0.02	-0.03	0.07	0.13	0.06	0.04	0.00	-0.10	0.05	-0.00	-0.02
	S.E.	(0.15)	(0.14)	(0.14)	(0.23)	(0.10)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
	Bw.	2876.87	5649.22	2597.18	2520.30	3876.22	5112.58	3620.05	2162.33	3686.43	2705.85	2729.74
	N.	202	406	191	187	275	354	261	154	267	198	198
Service	Est.	0.23	-0.03	-0.28	-0.24	-0.21	-0.38	-0.60**	-0.92**	-0.79**	-0.52*	-0.97***
	S.E.	(0.39)	(0.29)	(0.30)	(0.29)	(0.25)	(0.26)	(0.26)	(0.38)	(0.33)	(0.29)	(0.36)
	Bw.	5421.19	5606.82	5374.17	4746.04	6626.21	5405.93	5319.62	4359.82	4388.63	5114.24	3891.37
	N.	387	407	382	326	512	384	376	297	299	355	275
Panel C: Log(Wages)												
Primary	Est.	-0.31	-0.15	-0.20	-0.01	0.08	0.18	2.13	2.69	9.57	0.56	0.05
	S.E.	(0.87)	(0.24)	(0.43)	(0.19)	(0.61)	(1.81)	(6.44)	(3.29)	(25.16)	(1.73)	(0.10)
	Bw.	2856.25	4071.70	3470.00	3063.51	2746.69	2619.39	2802.89	3539.01	2895.64	2321.61	4261.34
	N.	133	182	164	147	132	133	141	183	149	122	198
Manufacturing	Est.	-0.014	-0.022	-0.030	-0.021	0.049	-0.042	-0.006	-0.006	-0.025	-0.054	-0.055
	S.E.	(0.034)	(0.034)	(0.036)	(0.039)	(0.073)	(0.060)	(0.030)	(0.030)	(0.046)	(0.096)	(0.120)
	Bw.	2787.543	2732.056	4217.847	4387.325	2352.943	4219.781	2604.420	3973.751	4134.458	2745.254	2308.787
	N.	144	143	215	228	134	227	153	229	236	170	146
Commerce	Est.	0.016	-0.005	0.003	0.003	0.007	0.013	-0.001	0.004	-0.002	-0.001	-0.008
	S.E.	(0.012)	(0.010)	(0.006)	(0.008)	(0.009)	(0.011)	(0.009)	(0.005)	(0.007)	(0.005)	(0.011)
	Bw.	2867.437	4381.245	4299.532	2308.925	2770.514	1930.360	3538.117	1138.209	2929.331	5155.767	4261.177
	N.	201	296	292	168	203	138	253	79	210	358	290
Service	Est.	-0.012	-0.018	-0.023	-0.037	-0.018	-0.032	-0.013	-0.052*	-0.029	-0.032	-0.026
	S.E.	(0.017)	(0.015)	(0.017)	(0.034)	(0.020)	(0.021)	(0.015)	(0.031)	(0.031)	(0.027)	(0.021)
	Bw.	2643.714	4196.069	4444.780	2359.307	4249.694	3006.944	4576.753	3691.358	2533.988	2576.915	3087.950
	N.	195	288	304	173	290	216	314	267	188	191	220

Notes: This table presents second-stage estimates for three labor outcomes — net job creation, hours worked, and log(wages) — stratified by sector (Primary, Manufacturing, Commerce and Service) across years 2006–2016. Pre-treatment (2006–2007), treatment roll-out (2008–2010), and post-treatment (2011–2016) periods follow the policy timeline. Standard errors (S.E.) are robust and bias-corrected; bandwidth (Bw.) and observation counts (N.) are reported.

or short-term, reflecting gig economy expansion rather than stable employment growth. Meanwhile, manufacturing and primary sectors showed no robust trends, with volatile estimates underscoring limited evidence, consistent with their reliance on physical capital or low-tech processes less sensitive to marginal speed improvements.

These findings align with the hypothesis that returns to internet speed upgrades are contingent on industries' digital absorptive capacity: service sectors, reliant on real-time data and connectivity, face dual pressures of automation and labor market restructuring, while commerce reaps fleeting gains from digitized transactions. Shi and Li (2023) find that in the services industry, the increased competition facilitated by broadband internet can lead to wage stagnation, particularly in sub-sectors where labor is more abundant. For example, in China, the expansion of broadband internet has been linked to a slight decrease in average wages, despite the overall increase in employment rates.

The services industry has been one of the primary beneficiaries of broadband internet expansion. In Germany, for example, broadband availability has been linked to robust employment growth in the service sector, particularly in knowledge- and computer-intensive industries (Stockinger, 2019). This growth is driven by the complementary nature of broadband internet in the production processes of service firms, which often rely on high-speed internet for operations.

However, the wage effects of broadband internet in the services industry are more mixed. While some studies suggest that broadband expansion can lead to higher starting wages and more stable employment relationships, others find that the average wages in the services sector may slightly decrease due to the increased supply of labor facilitated by broadband internet (Bhuller et al., 2019; Bu, 2023).

2.4.6 Fast Internet and Gender Inequality

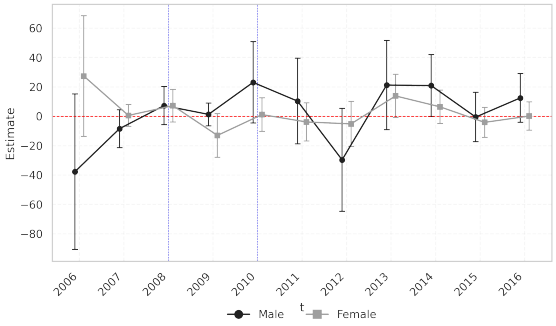
The availability and use of high-speed internet can influence labor market outcomes differently for men and women, often exacerbating existing gender disparities. For instance, skill-biased technological changes may disproportionately benefit men, and women may face barriers in accessing and utilizing digital technologies effectively. Therefore, this section investigates the gender heterogeneous effects of higher internet infrastructure investment. Table 2.13 presents the second-stage estimates on labor market outcomes by gender and year.

The gender-stratified analysis in Table 2.13 reveals asymmetric labor market effects, with female workers experiencing more pronounced post-treatment declines in hours worked and wages

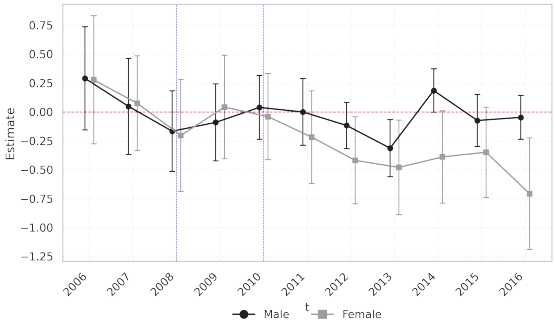
Table 2.13: Second-stage Estimates on Municipal Labor Market Outcomes by Gender and Year (2006–2016)

		Pre-treatment		Treatment Roll-out			Post-treatment					
Education		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation												
Male	Est.	-37.73	-8.44	7.27	1.31	23.11	10.33	-29.65	21.28	20.93	-0.51	12.50
	S.E.	(32.29)	(7.87)	(7.94)	(4.72)	(16.90)	(17.79)	(21.41)	(18.52)	(12.86)	(10.27)	(10.12)
	Bw.	6709.48	3921.21	4469.74	6543.60	4464.91	6274.89	5748.70	3485.80	4896.83	5135.35	4928.67
	N.	514	277	305	504	305	489	422	251	336	357	340
Female	Est.	27.39	0.59	7.24	-13.02	1.16	-3.87	-5.17	13.88	6.45	-4.13	0.16
	S.E.	(25.02)	(4.54)	(6.79)	(9.05)	(7.00)	(7.92)	(9.39)	(8.96)	(6.91)	(6.22)	(5.85)
	Bw.	4487.63	4626.27	4576.83	2951.08	5516.06	5015.48	3861.18	4600.56	5269.24	4698.28	5562.58
	N.	306	318	314	214	401	345	273	317	369	323	407
Panel B: Hours Worked												
Male	Est.	0.29	0.05	-0.17	-0.09	0.04	0.00	-0.12	-0.31*	0.19	-0.07	-0.05
	S.E.	(0.27)	(0.25)	(0.21)	(0.20)	(0.17)	(0.18)	(0.12)	(0.15)	(0.12)	(0.14)	(0.12)
	Bw.	5293.83	3836.02	4060.42	3868.60	6372.84	5057.73	6040.21	4617.01	8811.47	5337.58	6722.26
	N.	371	273	283	274	495	349	469	318	700	378	516
Female	Est.	0.28	0.08	-0.20	0.04	-0.04	-0.22	-0.42*	-0.48*	-0.39	-0.35	-0.71**
	S.E.	(0.34)	(0.25)	(0.30)	(0.27)	(0.23)	(0.24)	(0.23)	(0.25)	(0.24)	(0.24)	(0.29)
	Bw.	5744.41	6533.21	4767.53	5523.12	6841.07	5235.18	5104.16	5253.08	4976.95	5195.52	3762.28
	N.	422	504	327	403	527	364	351	365	343	362	269
Panel C: Log(Wages)												
Male	Est.	-0.009	-0.016	0.001	0.001	0.005	-0.015	-0.012	-0.007	-0.018	-0.010	-0.004
	S.E.	(0.010)	(0.014)	(0.012)	(0.012)	(0.014)	(0.012)	(0.017)	(0.021)	(0.016)	(0.014)	(0.010)
	Bw.	4011.394	3268.322	3247.249	2321.198	2257.175	4045.319	2187.706	1959.802	3380.054	2516.054	4168.870
	N.	280	234	231	171	165	283	159	140	243	188	287
Female	Est.	-0.014	-0.026	-0.021	-0.031	-0.017	-0.032	-0.022	-0.037*	-0.034	-0.032	-0.024
	S.E.	(0.014)	(0.018)	(0.015)	(0.026)	(0.020)	(0.021)	(0.021)	(0.021)	(0.023)	(0.024)	(0.023)
	Bw.	3099.644	2632.575	4495.094	2377.693	2738.524	2815.793	2222.354	4223.781	3503.508	2779.049	2456.750
	N.	222	194	307	177	198	204	161	289	253	203	184

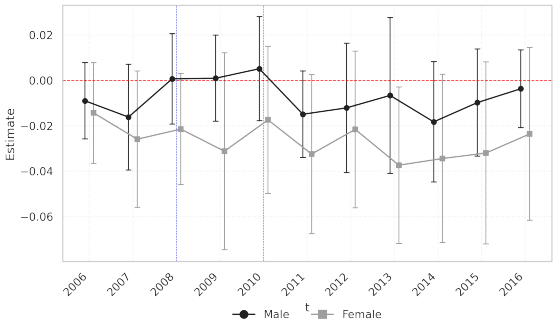
Notes: This table presents second-stage estimates for three labor outcomes — net job creation, hours worked, and log(wages) — stratified by gender (male/female) across years 2006–2016. Pre-treatment (2006–2007), treatment roll-out (2008–2010), and post-treatment (2011–2016) periods follow the policy timeline. Standard errors (S.E.) are robust and bias-corrected; bandwidth (Bw.) and observation counts (N.) are reported.



(a) Net Job Creation



(b) Hours Worked



(c) Log of Wages

Figure 2.12: Effects of Broadband Expansion on Municipal Labor Market Outcomes - Radio Technology, 2011–2015

compared to males. While net job creation (Panel A) lacks statistical significance for both genders, hours worked (Panel B) and wages (Panel C) exhibit gendered disparities. For female workers, hours worked declined significantly post-treatment, intensifying to -0.71 ($p<0.05$) in 2016, while wages fell by 3.7 % ($p<0.10$) in 2013. Male workers, in contrast, saw smaller and less consistent effects. Hours worked declined modestly in 2013 (-0.31, $p<0.10$), but wages remained stable.

These effects suggest that female labor markets near the policy cutoff faced sustained negative pressures, potentially due to over-representation in service roles vulnerable to automation, or caregiving constraints limiting adaptability to digital shifts. This asymmetry implies that digital infrastructure expansion exacerbated potential preexisting gender inequalities, disproportionately displacing or reducing labor demand for female workers.

Kusumawardhani et al. (2023) find that wider internet availability has provided women in Indonesia with an opportunity to work longer hours. However, they compare the arrival of internet to locations that did not have early internet access.

The results align with theories of gendered occupational sorting and digital disruption: women, often concentrated in routine or care-based sectors, may face heightened automation risks or reduced flexibility to transition into tech-driven roles. Policymakers must address these disparities through gender-targeted re-skilling programs, childcare support, and incentives for female participation in high-growth digital sectors to mitigate regressive labor market outcomes.

2.5 Conclusion

In conclusion, this study provides new insights into the multifaceted impacts of broadband expansion on local labor markets in Brazil. Leveraging a quasi-natural experiment and a robust regression discontinuity framework across more than 2,000 municipalities, our analysis reveals that the broadband policy – primarily through investments in microwave radio technology around a population size of 20,000 – has yielded mixed outcomes. While enhanced internet infrastructure has spurred net job creation among lower-educated and commerce sector workers, suggesting potential benefits in traditionally underserved segments, our findings do not reveal large aggregate effects on job creation. Instead, any overall gains or losses appear to be small and short-lived, potentially masked by regional heterogeneity or emerging only over a longer time horizon than captured in the initial years following implementation. To harness its benefits while addressing adverse effects, targeted policy measures are fundamental. Investments in digital literacy programs (Yang, 2023), equitable regional infrastructure development (Hua and and, 2024; Jin et al., 2025), and firm-level incentives (Akerman et al., 2015) could help maximize the gains of expanded broadband connectivity. Moreover, a heterogeneity analysis by region could further elucidate these dynamics and guide more tailored interventions.

These findings highlight the inherent trade-offs of digital infrastructure policies, wherein the same technological improvements that foster inclusiveness and job growth can also precipitate labor market disruptions, possibly through automation and substitution effects. Overall, this work not only advances our understanding of the distributional consequences of broadband expansion but also underscores the need for complementary policy measures to mitigate adverse effects while harnessing the full potential of digital transformation.

Chapter 3

DOUBLE MACHINE LEARNING FOR PRICE ELASTICITY ESTIMATION: LEVERAGING UNSTRUCTURED DATA FROM THE STEAM DIGITAL STORE

3.1 Introduction

Understanding the price elasticity of demand – how quantity demanded responds to price changes – is elemental to economic analysis. Decision-makers rely on accurate elasticity estimates to optimize pricing strategies, forecast revenue streams, design effective marketing campaigns, and navigate market competition. Yet despite its fundamental importance, obtaining reliable estimates remains one of economics' most persistent methodological challenges.

When pioneering studies (Schultz, 1933; Stigler, 1939; Working, 1943; Wright, 1928) first attempted to estimate demand curves from observational market data, they immediately encountered an elegant but formidable problem: the simultaneous determination of prices and quantities. Unlike controlled experiments, real markets rarely feature random price variations. Instead, prices emerge from the dynamic interplay between supply and demand forces, influenced by many factors that researchers often do not observe.

This identification challenge has only grown more complex in modern markets, where products have so many different features that affect both what consumers buy and how firms price their goods. The resulting high-dimensional confounding problem stretches traditional econometric approaches to their limits, necessitating innovative methodological approaches. To illustrate this, I analyze *Steam*, Valve Corporation's digital storefront that has dominated PC gaming since its 2003 launch. Steam lists more than games: it also distributes creative suites, development tools, and other software. With millions of users worldwide, Steam provides extensive data on game pricing, performance metrics, and user engagement [1]. This rich, multimodal environment represents both the promise and the complexity of modern data: it supplies granular signals about product quality and engagement, yet its very richness magnifies the confounding problem that motivates my

[1] AS Valve does not release unit-sales figures, we use review counts as demand proxies. Prior work documents a remarkably stable ratio—popularized as the *Boxleiter method*—where roughly 1–2% of purchasers leave a review; multiplying observed reviews by the reciprocal of that rate yields sales estimates that align with occasional publisher disclosures and third-party audits. See discussions in vginsights.com and gamalytic.com.

approach.

I propose an empirical strategy that integrates recent advances in causal inference and artificial intelligence (AI) for price elasticity of demand estimation. This approach utilizes the Double Machine Learning (DML) framework (Chernozhukov et al., 2018), which is specifically designed to estimate low-dimensional causal parameters (e.g. price elasticity of demand) in the presence of high-dimensional confounding variables. It achieves this by employing flexible machine learning (ML) algorithms to model the complex relationships between the confounders and both the outcome and the treatment (in my setting, demand and price, respectively), which allow for valid statistical inference [2].

To handle the rich, multimodal product data, I combine DML with modern AI-driven feature generation techniques. Specifically, I employ multimodal embeddings to represent the complex product characteristics captured in text and images (He et al., 2015; Liu et al., 2019). These embeddings transform the unstructured and high-dimensional raw data into dense numerical vectors that capture semantic and visual nuances. After encoding textual and visual data into embeddings, I fine-tune these representations specifically for price and demand prediction—critical components for my downstream task of estimating price elasticity of demand. I then evaluate these fine-tuned embeddings through both qualitative and quantitative analyses, demonstrating not only how effectively they represent the products but also quantifying their incremental predictive power compared to a baseline linear model using only tabular (structured) data. These fine-tuned embedding vectors then serve as the high-dimensional confounder set within the DML framework. This approach, leveraging AI-generated features within a robust causal inference structure, mirrors the architecture successfully developed by Klaassen et al. (2024) and applied by Bach et al. (2024) in their demand analysis of toy cars, demonstrating its potential for empirical economic analysis [3]. This combination offers a pathway to utilize the full richness of digital product data for causal analysis, separating the complex prediction task (handled flexibly by ML and embeddings) from the causal estimation task (handled rigorously by DML).

I find a positive and inelastic relationship between the quantity demanded and prices, showing that on average, a 10% increase in price would be associated with approximately a 1% increase

[2]Importantly, DML incorporates mechanisms like Neyman orthogonality and cross-fitting to ensure that biases inherent in the ML estimation of these nuisance functions do not contaminate the final estimate of the causal parameter of interest, thereby preserving desirable statistical properties like $\sqrt{N}$-consistency and asymptotic normality.

[3]See also Chapter 10 in Chernozhukov et al. (2024), Bajari et al. (2023), and Compiani et al. (2025).

in quantity demanded. Consequently, raising prices would likely increase overall revenue (since the percentage increase in price outweighs the percentage increase in quantity), while lowering prices would counterintuitively reduce both quantity demanded and revenue—a finding with significant implications for pricing strategy on digital platforms where quality signals are fundamental. This aggregate estimate, however, masks substantial heterogeneity across different categories of products. Entertainment genres demonstrate the most consistent positive elasticity patterns. Adventure (approximately 0.30), Action (0.27), and Sports (0.28) all show relatively stronger positive elasticities, indicating consumers in these categories may use price as a quality signal when making purchasing decisions. This suggests strong brand loyalty, differentiated experiences, or price insensitivity among core gamers who prioritize content over cost. Whereas professional software categories reveal more diverse patterns: Software Training exhibits the highest elasticity (0.53), in contrast Animation & Modeling is the only category showing negative elasticity (-0.08). However, because the professional software categories have wider confidence intervals than the entertainment ones, these point estimates should be interpreted with appropriate caution.

This chapter makes several contributions. Empirically, it applies the DML-embedding methodology to estimate the price elasticity of demand for a large dataset comprising approximately 14,000 digital products available on the Steam digital store, published between 2000 and 2019. This provides a contemporary estimate for a significant digital market, accounting for a richer set of product characteristics than typically feasible with traditional methods. Methodologically, it serves as a case study demonstrating the application, interpretation, and potential pitfalls of using AI-generated features derived from unstructured data within a formal causal inference framework.

This research is situated within the burgeoning field of Causal Machine Learning (Chernozhukov et al., 2024; Kaddour et al., 2022; Li and Chu, 2023), which signifies a paradigm shift towards using sophisticated ML tools not merely for prediction, but for the more challenging task of estimating causal effects from complex observational data. The ability to integrate unstructured data into causal models is particularly relevant, as demonstrated by emerging applications in diverse domains. For instance, in medicine, researchers are using text from electronic health records combined with DML to estimate treatment effects more reliably from real-world data (Masukawa et al., 2022). Similarly, in epidemiological studies, there is potential to use DML to evaluate the health effects of multiple mismeasured pollutants (Xu et al., 2024). This chapter contributes to this broader methodological movement by providing a detailed application in the context of economic demand analysis, highlighting both the power and the necessary caution required when bridging AI and

causal inference.

The remainder of this paper is structured as follows: Section 2 provides a background on demand estimation challenges, causal inference principles, the DML framework, and multimodal embeddings. Section 3 describes the Steam dataset, variable construction, and data preprocessing steps. Section 4 details the specific methodology employed, including the model specification, identification assumptions, the DML estimation procedure, and the incorporation of embeddings. Section 5 presents the empirical results, including the estimated price elasticity and its heterogeneity analysis. Section 6 acknowledges the limitations of the study. Finally, Section 7 concludes with a summary of findings and directions for future research.

3.2 Background and Literature Review

3.2.1 Demand Estimation

Estimating the price elasticity of demand is a fundamental task in applied economics. However, it is complicated by the problem of endogeneity, as prices are rarely exogenous. Prices may respond to unobserved demand shocks or firm-level strategic decisions, leading to biased and inconsistent estimates of demand when naive methods are used. This motivates the use of specialized econometric techniques that can recover causal estimates of demand parameters under various data structures and economic environments.

The most basic econometric method is Ordinary Least Squares (OLS), which estimates a linear demand equation of the form:

$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \epsilon_i \tag{3.1}$$

In this specification, Q_i denotes quantity demanded, P_i is price, X_i is a vector of observed demand covariates (such as income), and ϵ_i captures unobserved influences on demand. The validity of OLS relies crucially on the assumption that price is exogenous, meaning that $\text{Cov}(P_i, \epsilon_i) = 0$. However, in most market settings, price is determined endogenously through the interaction of supply and demand, and may also be influenced by unobservable product characteristics or market-level shocks. Consequently, when $\text{Cov}(P_i, \epsilon_i) \neq 0$, the OLS estimator becomes biased and inconsistent, typically underestimating the true price responsiveness. Thus, while OLS is easy to implement and interpret, it is generally inappropriate in settings where price endogeneity is suspected.

Instrumental Variables (IV)

Instrumental Variables (IV) estimation provides a solution to the endogeneity problem by leveraging an instrument I_i that shifts price but is not directly related to the unobserved determinants of demand. A valid instrument must satisfy two core conditions: it must be relevant, meaning that it is correlated with the endogenous regressor ($\text{Cov}(I_i, P_i) \neq 0$), and it must be exogenous, implying that it is uncorrelated with the error term ($\text{Cov}(I_i, \epsilon_i) = 0$). The most common implementation of IV is the Two-Stage Least Squares (2SLS) procedure. In the first stage, price is regressed on the instrument and other exogenous variables:

$$P_i = \pi_0 + \pi_1 I_i + \pi_2 X_i + v_i \tag{3.2}$$

This produces fitted values $\hat{P}_i$ that reflect exogenous variation in price. In the second stage, the demand equation is estimated using $\hat{P}_i$ in place of the endogenous price:

$$Q_i = \beta_0 + \beta_1 \hat{P}_i + \beta_2 X_i + u_i \tag{3.3}$$

This approach yields consistent estimates of demand parameters, assuming the instrument is valid. Nonetheless, IV methods face practical challenges. Finding strong and credible instruments is often difficult, and weak instruments can result in biased estimates that converge toward OLS. Furthermore, over-identification tests such as the Hansen J-test may be required when multiple instruments are used. Foundational contributions to the IV literature include Wright (1928), Angrist et al. (1995), and Stock and Yogo (2002).

Simultaneous Equation Models (SEMs)

Simultaneous Equation Models (SEMs) offer a structural framework for jointly modeling supply and demand, thereby addressing the simultaneity problem explicitly. In this approach, both the demand and supply equations are specified:

$$\text{Demand:} \quad Q_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \epsilon_i \tag{3.4}$$

$$\text{Supply:} \quad Q_i = \gamma_0 + \gamma_1 P_i + \gamma_2 W_i + \eta_i \tag{3.5}$$

Here, X_i denotes variables that shift demand but not supply (e.g., income), while W_i represents

supply-side shifters such as input costs. A crucial issue in SEMs is identification, which refers to the ability to uniquely estimate the parameters of the demand equation. Identification is generally ensured by satisfying the order and rank conditions, which require that each equation exclude at least one exogenous variable that appears in the other. Once identified, SEMs can be estimated using methods such as 2SLS, Three-Stage Least Squares (3SLS), or Limited Information Maximum Likelihood (LIML). These models are particularly useful when the goal is to recover structural elasticities or when both supply and demand curves are of interest, as discussed in Greene (2012) and Wooldridge (2010).

Time Series Models

When demand data are collected over time, it is essential to account for temporal dependencies and dynamic behavior. Autoregressive Integrated Moving Average (ARIMA) models provide a flexible framework for univariate time series forecasting. The ARIMA model can be written as:

$$\phi(L)(1-L)^d Q_t = \theta(L)\epsilon_t \tag{3.6}$$

where $\phi(L)$ and $\theta(L)$ are lag polynomials, d is the order of integration, and ϵ_t is a white noise error term. ARIMA models are well-suited to capturing trends, cycles, and seasonality in demand data.

In contrast, Vector Autoregression (VAR) models generalize ARIMA to the multivariate case, allowing for the joint modeling of variables such as price, quantity, and advertising:

$$Y_t = A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \epsilon_t \tag{3.7}$$

where Y_t is a vector of endogenous variables. VAR models are particularly useful for analyzing dynamic interactions and conducting impulse response analysis. Both ARIMA and VAR require stationarity; otherwise, unit root tests and differencing may be necessary. See Box et al. (2015) and Sims (1980).

Panel Data Models

Panel data models are designed for data that follow multiple entities over time, such as firms or consumers. This structure allows researchers to control for unobserved heterogeneity across units. One popular approach is the Fixed Effects (FE) model, which includes unit-specific intercepts:

$$Q_{it} = \beta_0 + \beta_1 P_{it} + \alpha_i + \epsilon_{it} \tag{3.8}$$

Here, α_i captures all time-invariant unobserved characteristics of unit i. By focusing on within-entity variation, the FE estimator effectively eliminates bias from unobserved time-invariant confounders. The Random Effects (RE) model also accounts for unit-specific heterogeneity, but assumes that α_i is uncorrelated with the regressors:

$$\text{Cov}(X_{it}, \alpha_i) = 0 \tag{3.9}$$

This assumption allows for more efficient estimation under correct specification. The Hausman test is commonly used to compare FE and RE models and determine whether the stricter RE assumptions are justified. See Hsiao (2014) and Wooldridge (2010).

Discrete Choice Models

Discrete choice models are used when individuals select one option from a finite set of alternatives, such as products or brands. These models are rooted in Random Utility Maximization (RUM) theory, where the utility from choice j for individual i at time t is given by:

$$U_{ijt} = V_{ijt} + \epsilon_{ijt}, \quad V_{ijt} = \beta X_{ijt} \tag{3.10}$$

The basic logit model specifies the probability of choosing option j as:

$$P_{ijt} = \frac{e^{V_{ijt}}}{\sum_k e^{V_{ikt}}} \tag{3.11}$$

However, the logit model imposes the Independence of Irrelevant Alternatives (IIA) assumption, which restricts substitution patterns in unrealistic ways. The probit model relaxes this by allowing for normally distributed, potentially correlated error terms. Mixed logit models offer even greater flexibility by allowing parameters to vary randomly across individuals, thereby capturing heterogeneity in preferences. The Berry-Levinsohn-Pakes (BLP) (Berry et al., 1995) model extends the mixed logit to the context of aggregate market data. It accounts for unobserved product characteristics and the endogeneity of price through a structural error term:

$$s_{jt} = \delta_{jt} + \mu_{jt} + \xi_{jt} \tag{3.12}$$

In BLP, the term ξ_{jt} captures unobserved demand shocks and is instrumented to address endogeneity. This framework has become the workhorse model for empirical industrial organization. Other references include McFadden (1974) and Train (2009).

3.2.2 Causal Inference: Foundations

The goal of causal inference is to move beyond mere correlation and estimate the effect of a specific action, intervention, or "treatment" on an outcome of interest. The Potential Outcomes Framework (Holland, 1986; Imbens and Rubin, 2015b; Rubin, 1974) provides a rigorous foundation for defining and estimating causal effects. For a given unit i (e.g., a Steam game), let P_i be the treatment status (e.g., the price level). I conceptualize potential outcomes: $Q_i(P_i = p)$ represents the outcome (e.g., demand) unit i *would* exhibit if it received the treatment (e.g., a high price), and $Q_i(P_i = p')$ represents the outcome the *same* unit i *would* exhibit if it received a different treatment (e.g., a low price).

The individual treatment effect (ITE) for unit i is defined as the difference between its potential outcomes: $ITE_i = Q_i(P_i = p) - Q_i(P_i = p')$. However, the "fundamental problem of causal inference" is that we can only observe one of these potential outcomes for any given unit at a specific point in time – the outcome corresponding to the treatment actually received. We cannot simultaneously observe what would have happened under the alternative treatment scenario (the counterfactual).

Consequently, causal inference typically focuses on estimating *average* treatment effects across a population or subpopulation. Common estimands include the Average Treatment Effect (ATE), $E[Q_i(P_i = p) - Q_i(P_i = p')]$, which is the average effect across the entire population, and the Average Treatment Effect on the Treated (ATT), $E[Q_i(P_i = p) - Q_i(P_i = p')|P_i = p]$, which is the average effect specifically for those units that actually received the treatment level p.

In observational studies, where treatment assignment is not controlled by the researcher, estimating these average effects is complicated by confounding bias. Confounding occurs when variables exist that influence both the treatment assignment P_i and the potential outcomes $Q_i(P_i = p)$. These variables are known as confounders, denoted by X_i. If confounders are not adequately accounted for, any observed association between P_i and Q_i may be spurious, reflecting the influence of X_i rather than a true causal effect of P_i on Q_i. Naive estimators, such as simple correlations or regressions that fail to control appropriately for X_i, will generally yield biased estimates of the causal effect. While randomized controlled trials (RCTs) are considered the gold standard for establishing

causality because randomization, on average, breaks the link between confounders and treatment assignment, they are often infeasible or unethical in many economic and social contexts. Therefore, causal inference from observational data relies heavily on identifying and adjusting for confounders, which requires specific assumptions about the data generating process.

3.2.3 Machine Learning for Causal Inference: The Rise of DML

Traditional econometric methods, while well-established for causal inference under specific assumptions, often face limitations when confronted with modern datasets characterized by high dimensionality and complex, non-linear relationships (Chernozhukov et al., 2018, 2024). Estimating causal effects typically requires conditioning on confounders, but when the number of potential confounders (p) is large relative to the sample size (n), or when the functional forms relating confounders to treatment and outcome are unknown, standard methods like Ordinary Least Squares (OLS) can suffer from the "curse of dimensionality" and are prone to bias from model misspecification. Relying on parametric assumptions (e.g., linearity) in such settings often lacks strong theoretical justification and risks producing misleading results.

This challenge has motivated the integration of machine learning techniques into the causal inference toolkit (Chernozhukov et al., 2024; Kaddour et al., 2022; Li and Chu, 2023). ML algorithms excel at prediction in high-dimensional spaces and can automatically learn complex, non-linear patterns from data without requiring pre-specified functional forms. The core idea behind methods like Double Machine Learning (DML) is to leverage these predictive capabilities of ML for the "nuisance" parts of a causal model—specifically, estimating the conditional expectations of the outcome and the treatment given the high-dimensional confounders—while employing econometric principles to ensure that the final estimate of the target causal parameter remains statistically valid for causal inference.

DML (Chernozhukov et al., 2018) provides a general framework for combining machine learning with causal inference through three interconnected principles. At its core, Neyman Orthogonality ensures that the estimation relies on a moment condition that shields the target causal parameter from first-order bias when small errors occur in the machine learning estimation of nuisance functions—a critical safeguard against the regularization bias that ML estimators often introduce. This orthogonality works in tandem with the second principle: employing high-quality ML algorithms that can estimate these nuisance functions at sufficiently fast convergence rates, though not necessarily at the $\sqrt{N}$ rate desired for the causal parameter itself. Various supervised learning methods

serve this purpose effectively, including random forests, gradient boosting, neural networks, and lasso. To further protect against overfitting bias, DML implements its third principle, cross-fitting, a systematic form of sample splitting where the data is divided into multiple folds. For each fold, nuisance functions are estimated using only data from the other folds, thereby ensuring that the causal parameter estimation for any observation draws on nuisance function estimates trained independently of that observation—effectively breaking the dependency that could otherwise introduce bias.

In the context of demand estimation, DML offers an attractive alternative to traditional solutions to endogeneity, such as instrumental variables (IV), which require strong assumptions and valid exclusion restrictions (Ackerberg et al., 2007; Berry et al., 1995). Rather than rely on exogenous instruments, DML handles endogeneity by flexibly adjusting for high-dimensional confounders—what the causal inference literature refers to as "nuisance functions"—using modern machine learning algorithms (Chernozhukov et al., 2018). In my application, these confounders include multimodal embeddings derived from unstructured text and image data, which serve as proxies for unobserved product quality, a key source of omitted variable bias in price elasticity estimation. While DML cannot fully resolve bias from unobserved confounding without valid instruments, it improves upon classical approaches by conditioning on richer, behaviorally-relevant features in a flexible, non-parametric way. This makes the conditional independence assumption more plausible and reduces reliance on restrictive functional form assumptions. In doing so, this chapter contributes to the literature by illustrating how combining DML with unstructured data sources provides a scalable and robust strategy for estimating demand elasticities in digital marketplaces where traditional identification strategies are limited or infeasible.

3.2.4 Representing Unstructured Data: Multimodal Embeddings

A significant challenge in applying methods like DML to datasets common in digital markets and other domains is the prevalence of unstructured data, such as text and images. Traditional quantitative models are typically designed for structured, tabular data. Incorporating the rich information contained in free-form text (product descriptions, user reviews, tags) or visual elements (product images, logos, banners) requires methods to convert this unstructured data into a format suitable for statistical modeling.

Deep learning has provided powerful tools for this task in the form of embeddings (Bengio et al., 2003). Embeddings are dense, relatively low-dimensional vector representations of high-

dimensional, sparse, or unstructured data. Models like Word2Vec, GloVe, ResNet (He et al., 2015; Mikolov et al., 2013; Pennington et al., 2014) and particularly transformer-based architectures (e.g., BERT, RoBERTa for text; BEiT for images) learn mappings from raw inputs (words, sentences, pixels) to vectors such that items with similar semantic or visual meaning are located close to each other in the embedding space (Bao et al., 2022; Devlin et al., 2019; Liu et al., 2019; Vaswani et al., 2017). These embeddings capture complex patterns and relationships within the data in a format amenable to downstream ML tasks.

Furthermore, multimodal embedding techniques aim to integrate information from different data types (e.g., text, image, audio, tabular data) into a single, unified vector representation (Klaassen et al., 2024). This allows models to leverage complementary information from various sources. For instance, a multimodal embedding for a Steam game could jointly represent its textual description, its store banner image, and its structured genre information. Bach et al. (2024) utilized transformer-based multimodal embeddings to represent toy car products based on text descriptions, images, and tabular covariates, finding that these AI-driven representations significantly improved demand prediction and yielded more credible price elasticity estimates when used within a causal inference framework. Bajari et al. (2023) generate abstract product attributes from text descriptions and images using BERT and ResNet, and then use these attributes to estimate an adjusted hedonic price function.

In the context of this study, multimodal embeddings serve as the mechanism to transform the rich text and image data X_i characterizing each Steam game into embeddings E_i. These embeddings E_i, capturing nuanced product attributes, then function as the set of control variables (along with the tabular data) within the DML framework, allowing the estimation of price elasticity while conditioning on these detailed, AI-generated features. See Appendix C.1 for more details.

3.3 Data

This section describes the data source, sample selection, variable definitions, and preprocessing steps undertaken for the empirical application of estimating price elasticity on the Steam platform using DML with multimodal embeddings.

3.3.1 Data Source and Sample

The sample consists of 13,905 unique video games (including their downloadable contents) from the Steam digital store [4]. To ensure relevance and a sufficient history for most games (circumventing the computational complexity as well), I limited the sample to games published between the years 2000 and 2019, inclusive. This timeframe covers a significant period of Steam's growth and evolution, allowing for robust analysis of pricing patterns and their effects on demand across different market conditions and platform development stages. Game data including prices, descriptions, user reviews, and metadata were obtained from Kaggle [5]. I complemented the data set by extracting images using Python-based web scraping tools (Richardson, 2007) from the Steam digital store and from SteamDB [6].

3.3.2 Variable Definitions

The primary outcome variable Q_i is the natural logarithm of the number of user reviews received by a game i accumulated up to the point of data collection.

In digital markets where direct sales data is often unavailable, researchers frequently use online product reviews as proxies for demand, with metrics like review volume shown to correlate with sales across various products (Dellarocas et al., 2007; Zhu and Zhang, 2010). However, using review metrics as outcome variables when estimating price elasticity presents some limitations: the relationship between reviews and sales involves complex bidirectional causality, it remains unclear whether reviews primarily predict sales or actively influence purchasing decisions; reviewers represent a self-selected, potentially biased sample of consumers; and, review volume have different dynamics depending on product type and lifecycle stage. While this study uses log total reviews as the outcome variable due to data constraints, the resulting elasticity estimate specifically reflects how price changes associate with changes in online discussion volume rather than true market demand—a limitation requiring careful interpretation.

The treatment variable P_i is the natural logarithm of the game i's base price listed on the Steam store in US dollars. Taking the logarithm of both quantity and price allows the estimated

---

[4] https://store.steampowered.com/. Downloadable Content (DLC) refers to additional content, such as expansions, new characters, cosmetic items, or game modes, that can be purchased and downloaded separately from the base game.

[5] https://www.kaggle.com/datasets/trolukovich/steam-games-complete-dataset.

[6] https://steamdb.info/.

coefficient to be interpreted as an elasticity.

The control variables (confounders) X_i combine features from tabular, text, and image data, designed to capture characteristics that might confound the price-demand relationship. The tabular data includes genres (converted to dummy variables indicating the presence of specific game genres such as Action, RPG, Strategy), features/categories (dummy variables for key game attributes listed on the store page such as Single-player, Multi-player, Co-op, Steam Achievements, Controller Support, VR Support), release date (the date the game was initially released on Steam as indicated on the store page), and baseline review metrics (to capture pre-existing quality or popularity signals without inducing direct simultaneity with the outcome).

Text embeddings were generated using a pre-trained transformer model, RoBERTa (Liu et al., 2019), applied to the concatenation or combination of the game title, developer name(s), publisher name(s), game description, and user-defined tags (e.g., "Horror", "Medieval", "Survival"). Image embeddings were similarly generated using features extracted from a pre-trained convolutional neural network, ResNet-50 (He et al., 2015) applied to the main store "capsule" image (the small banner representing the game in lists). The final input Z_i for each game i is constructed by concatenating the tabular feature vector to the text and image embedding vectors. Thus, for each game i,

$$E_i = \begin{bmatrix} \text{RoBERTa}(\text{text}_i) \\ \text{ResNet50}(\text{image}_i) \end{bmatrix},$$

and the final (input) feature vector is

$$Z_i = \begin{bmatrix} E_i \\ T_i \end{bmatrix}$$

where T_i stands for the tabular feature vector.

I train the final input Z_i via a two-stage approach targeting both price and quantity prediction. First, I perform linear probing: I freeze all backbone weights and train only a linear head on the dual objectives of predicting price and quantity. Second, I unfreeze the full network and fine-tune on the same objectives. This blending seems to produce substantial gains in generalization ability and accuracy of the resulting predictive model (Kumar et al., 2022). The resulting optimized embedding $\hat{E}_i$ then form the final covariate set for our Double Machine Learning (DML) pipeline:

$$Z_i = \left[T_i,\ \hat{E}_i\right].$$

3.3.3 Multimodal Embeddings

Building on the approach of Bach et al. (2024), I proceed as follows. First, for each game i I take the combined vector of its text and image embeddings (including the tabular data), Z_i. I then apply a Johnson–Lindenstrauss mapping to project Z_i down to a 256-dimensional vector $\tilde{Z}_i$ (Johnson and Lindenstrauss, 1984), which approximately preserves pairwise distances. Next, I center these projected embeddings by subtracting their overall mean, and rescale them to unit length so that they lie on the surface of a hypersphere:

$$X_i^e = \frac{\tilde{Z}_i - \frac{1}{n}\sum_{j=1}^{n} \tilde{Z}_j}{\left\| \tilde{Z}_i - \frac{1}{n}\sum_{j=1}^{n} \tilde{Z}_j \right\|}.$$

These normalized vectors form the basis of all subsequent analysis.

I then evaluate the embeddings in two complementary ways:

1. **Qualitative inspection.** Visualize clusters and nearest-neighbor relationships on the hypersphere to check whether games with similar characteristics are grouped together.

2. **Quantitative performance.** Include the new features as covariates in predictive models of price and quantity, testing whether they materially improve out-of-sample forecasts—an important prerequisite for their use in downstream causal and hedonic-price analyses.

Together, these steps ensure that the learned representations both reflect meaningful product similarities and contribute useful information to demand-and-price modeling.

Qualitative Inspection

For the qualitative inspection, I firstly employ a clustering analysis using k-means to distribute games into five distinct groups based on their embedding representations. To assess how visual information impacts clustering quality, I conduct parallel analyses using multimodal (text+image) embeddings and text-only embeddings. The resulting clusters are presented in Tables 3.1 and 3.2, which demonstrate that multimodal embeddings produce clusters with greater visual coherence and internal consistency — highlighting the value of incorporating multiple data modalities. Appendix A.2 presents tables showing the 5 most representative games within each cluster along with their full text data.

Table 3.1: Text-only Model

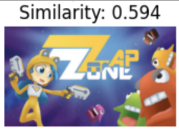
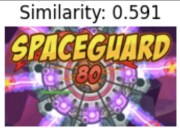
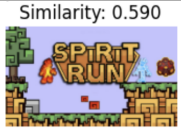
Cluster	**Representative Games**
Cluster 1	Similarity: 0.594  Similarity: 0.591  Similarity: 0.590  Similarity: 0.590  Similarity: 0.587 
Cluster 2	Similarity: 0.666  Similarity: 0.666  Similarity: 0.663  Similarity: 0.662  Similarity: 0.659 
Cluster 3	Similarity: 0.671  Similarity: 0.669  Similarity: 0.668  Similarity: 0.663  Similarity: 0.661 
Cluster 4	Similarity: 0.603  Similarity: 0.602  Similarity: 0.602  Similarity: 0.600  Similarity: 0.599 
Cluster 5	Similarity: 0.677  Similarity: 0.676  Similarity: 0.676  Similarity: 0.676  Similarity: 0.674 

Table 3.2: Multimodal (Text + Image) Model

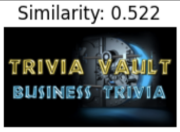
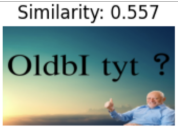
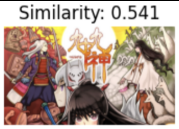
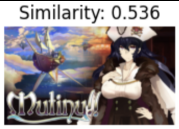
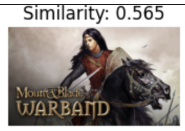
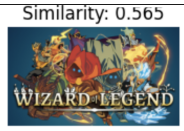
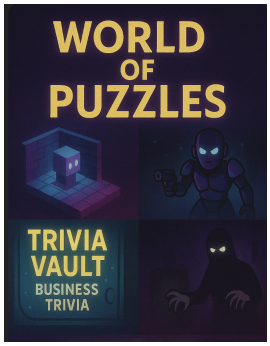
Cluster	Representative Games
Cluster 1	Similarity: 0.527  Similarity: 0.526  Similarity: 0.523  Similarity: 0.522  Similarity: 0.522 
Cluster 2	Similarity: 0.557  Similarity: 0.557  Similarity: 0.556  Similarity: 0.553  Similarity: 0.551 
Cluster 3	Similarity: 0.541  Similarity: 0.539  Similarity: 0.538  Similarity: 0.537  Similarity: 0.536 
Cluster 4	Similarity: 0.551  Similarity: 0.550  Similarity: 0.549  Similarity: 0.549  Similarity: 0.548 
Cluster 5	Similarity: 0.569  Similarity: 0.568  Similarity: 0.565  Similarity: 0.565  Similarity: 0.565 

Table 3.3: "Average" Game by Cluster (Multimodal Model)

Cluster	Representative Image	Description
Cluster 1		**World of Puzzles** is a mind-bending indie adventure where physics and logic collide in surreal, shifting environments. Set in a mysterious digital realm where gravity is optional and perception is your greatest tool, players must solve intricate puzzles to restore harmony to a fragmented world.

Cluster 2		**Quest of Quirks** is a delightfully oddball indie game that mashes together pixelated action, lighthearted storytelling, and bite-sized puzzle mechanics. Jump between wildly different game styles—from rescuing princesses in trap-filled castles to merging mysterious tiles in a magical 4x4 grid.
Cluster 3		**Magical Enigma** is a fantasy visual novel with anime-style characters, magical puzzles, and branching storylines. Join a young mage and companions as they uncover secrets across a magical realm.
Cluster 4		**Modern Flight Simulator** combines civil aviation, military helicopters, and iconic train routes in one highly realistic experience. Fly aircraft like the Airbus A320 and Embraer 195, pilot tactical Mi-8 helicopters, or operate trains through scenic routes like Marias Pass.
Cluster 5		**Ancient Realms** features era-spanning combat between knights, vikings, mages, and archers in mythic battle arenas. Choose your class and engage in tactical PvP and co-op action in a fractured fantasy world.

Next, to further describe these cluster characteristics, I utilize generative AI (OpenAI, 2025) to create an "average" game of each cluster, including a descriptive summary and an image (Table 3.3). These generated insights align closely with human interpretations of the clusters, confirming the effectiveness of combining textual and visual embeddings for more comprehensive product

categorization.

Quantitative Performance

Building on the qualitative model inspection, I now present a quantitative performance evaluation of the embeddings' predictive power. I examine how effectively the embeddings predict price and quantity. Table 3.4 compares performance across modeling approaches and data modalities, revealing the impact of incorporating progressively richer information sources on the model's predictive capabilities.

Table 3.4: Test R^2 scores for predicting quantity and price

Method	Q_i	P_i
Linear Reg [tabular only]	52.72%	24.48%
Boosted Trees [tabular only]	58.09%	29.83%
Boosted Trees [text + tabular]	75.53%	56.41%
XGBoost [text + tabular]	76.50%	57.22%
Boosted Trees [text + image + tabular]	76.53%	55.86%
XGBoost [text + image + tabular]	77.29%	56.54%

The results demonstrate a clear pattern of improvement when incorporating multimodal data into predictive models. Starting with tabular-only approaches, we observe relatively modest predictive performance (52-58% R^2 for quantity prediction and 24-30% for price prediction). The introduction of text embeddings yields a substantial performance boost across all model types, increasing R^2 scores by approximately 17-18 percentage points for quantity prediction and by a remarkable 26-27 percentage points for price prediction.

The addition of image embeddings to create fully multimodal models (text+image+tabular) provides a further incremental improvement, particularly for tree-based methods. XGBoost consistently delivers the strongest performance within each data modality group, with the multimodal XGBoost model achieving the highest overall R^2 scores (77.29% for quantity and 56.54% for price prediction). Fuhr et al. (2024) observe that XGBoost performs very well across a broad range of settings in their analyses, and they recommend it as a baseline or default method within DML.

These findings highlight the significant value of incorporating textual product information into demand forecasting models, while also suggesting that visual product representations provide additional predictive signal beyond what text alone can capture. The substantially larger gains observed for price prediction compared to quantity prediction when moving from tabular-only to

text-enhanced models indicate that textual features may be particularly valuable for understanding price elasticity dynamics.

3.3.4 Descriptive Statistics

To provide context on the dataset, Table 3.5 presents summary statistics for key variables.

Table 3.5: Summary Statistics for Steam Digital Store Dataset (N=13,905)

Variable	Mean	Std. Dev.	Min	Max	Notes
Price (USD)	11.24	11.46	0.50	199.99	Launch price; includes downloadable content
Log Price (P)	1.99	0.97	-0.69	5.29	Used as treatment variable
Total Reviews	1,187	10,851	10	836,608	Highly skewed distribution
Log Total Reviews (Q)	4.54	1.83	2.30	13.63	Used as outcome variable
Release Date	1,428	1,143	11	7,304	Days since release date
Avg. Genres per Game	2.35	1.43	0	11	Based on publisher-assigned genre tags
Avg. Features per Game	4.27	2.55	0	18	Based on publisher-assigned feature tags
% Positive Reviews	74%	18%	0%	100%	Based on total number of reviews
Text Embedding Dim.	768	-	-	-	Encoder: RoBERTa-base
Image Embedding Dim.	2,048	-	-	-	Encoder: ResNet50
Tabular Dim.	61	-	-	-	
Total Dim. of Z_i	2,877	-	-	-	Combined tabular, text, image features

This table reveals the wide range in prices and especially review counts, indicating significant heterogeneity in the market. The high dimensionality of the final control vector Z_i, driven primarily by the embeddings, underscores the necessity of using methods like DML that are designed for such settings. The construction of this comprehensive Z_i vector is pivotal, as the credibility of the causal estimates relies heavily on its ability to capture the relevant confounding factors influencing both price and the demand proxy (review volume).

3.4 Empirical Model: DML with Multimodal Embeddings

This section details the methodological framework employed to estimate the price elasticity of demand for Steam games, combining the Partially Linear Regression (PLR) model structure with the Double Machine Learning (DML) estimation strategy and utilizing multimodal embeddings as high-dimensional control variables (Chernozhukov et al., 2018; Klaassen et al., 2024).

3.4.1 Model Specification: Partially Linear Regression (PLR)

I adopt a partially linear regression model, a common structure used within the DML framework, particularly suitable for estimating the effect of a low-dimensional treatment variable P_i (price) while controlling for high-dimensional confounders (Z_i). The model assumes that the outcome variable Q_i (quantity) is linearly related to the treatment variable, conditional on the confounders, but allows the relationship between the confounders and the outcome, as well as the relationship between the confounders and the treatment, to be arbitrarily complex and non-linear.

The structural equations are specified as follows:

$$Q_i = \theta \times P_i + g_0(Z_i) + \zeta_i \tag{3.13}$$

$$P_i = m_0(Z_i) + V_i \tag{3.14}$$

Where Q_i represents the logarithm of game i's number of user reviews (my proxy for demand), P_i denotes the logarithm of the game i's launch price, and Z_i encompasses a high-dimensional vector of control variables combining multimodal embeddings from text and image, including also tabular data. Our target parameter θ represents the constant price elasticity of demand, while $g_0(Z_i) = E[Q_i|Z_i]$ and $m_0(Z_i) = E[P_i|Z_i]$ function as potentially complex nuisance functions capturing conditional expectations. The stochastic error terms ζ_i and V_i are defined such that $E[V_i|Z_i] = 0$, with identification requiring $E[\zeta_i|Z_i, P_i] = 0$—implying that after conditioning on Z_i, any remaining variation in price (represented by V_i) is uncorrelated with remaining variation in demand (represented by ζ_i). This specification assumes a constant and additive causal effect of log price on log demand across all games.

This analysis not only focuses on estimating the average effect but also extends the DML framework to explore heterogeneous treatment effects (HTE), fully leveraging DML's advantages. While the partially linear regression model provides a solid foundation for estimating constant effects, I also implement extensions that allow θ to vary across different game characteristics, capturing how price elasticity may differ across different attributes. This approach maintains DML's ability to flexibly estimate high-dimensional nuisance functions using machine learning without imposing restrictive parametric assumptions, while simultaneously revealing important variations in consumer price sensitivity throughout the marketplace.

3.4.2 Identification Assumptions

For the estimated parameter θ from the PLR model to be interpreted as the causal price elasticity of demand, several key identification assumptions must hold.

The most crucial is the Conditional Independence Assumption (CIA), which requires that, conditional on observed covariates Z_i, treatment assignment P_i is independent of potential outcomes $Q_i(P_i = p)$, where p represents a specific price level that the random variable P_i can take. Formally: $Q_i(P_i = p) \perp P_i | Z_i$ for all relevant price levels p. This means that once we account for all game characteristics captured in vector Z_i (including rich information from multimodal embeddings), a game i's price is assigned as if randomly with respect to any remaining factors influencing potential demand. While our comprehensive embeddings make this assumption more plausible than using only basic tabular controls, CIA remains fundamentally untestable since potential outcomes are never fully observed.

The second key assumption is Positivity, requiring that for any characteristics $Z_i = x$ in the population, there's a non-zero probability of observing games at different price levels being compared: $0 < P(P_i = p | Z_i = x) < 1$ for all supported z and relevant price levels p, where $P(\cdot)$ denotes probability. This ensures that for any game type defined by features z, we can find comparable games with different prices.

Finally, the Stable Unit Treatment Value Assumption (SUTVA) comprises two parts: No Interference (one game's potential outcomes are unaffected by prices assigned to other games) and Consistency (the observed outcome equals the potential outcome under the actual assigned price: $Q_i = Q_i(P_i = p)$ when $P_i = p$). If these assumptions hold, the parameter θ identified by the DML procedure corresponds to the average causal effect of log price on log demand.

3.4.3 DML Estimation Procedure

The DML estimation procedure for the PLR model (Equations 1 and 2) aims to obtain an estimate of θ that is robust to the biases introduced by using machine learning to estimate the high-dimensional nuisance functions $g_0(Z_i)$ and $m_0(Z_i)$. It achieves this through orthogonalization and cross-fitting.

Following Bach et al. (2024), I implement a simplified version of the cross-fitting procedure by splitting the dataset into just two groups rather than using K-fold cross-validation. The core algorithm begins with this two-fold cross-fitting setup, where I randomly partition the dataset

of 13,905 games into two mutually exclusive subsets (80% for training data, and the remaining 20% for test data). It's worth noting that during the linear probing and fine-tuning stages of the neural network, I also create a separate validation set to prevent overfitting. Next, I perform nuisance function estimation. I train two machine learning models : a model $\hat{g}(Z_i)$ to predict the outcome Q_i from the features Z_i, estimating $g_0(Z_i) = E[Q_i|Z_i]$; and a model $\hat{m}(Z_i)$ to predict the treatment P_i from the features Z_i, estimating $m_0(X_i) = E[P_i|Z_i]$. The model architecture leverages XGBoost (Chen and Guestrin, 2016) capable of successfully handling high-dimensional multimodal embeddings (Fuhr et al., 2024).

The algorithm continues with residual calculation using out-of-sample predictions. For the test set, I use the models trained on the training set to compute residuals: outcome residual $\tilde{Q}_i = Q_i - \hat{g}(Z_i)$ and treatment residual $\tilde{P}_i = P_i - \hat{m}(Z_i)$. This step effectively partials out the influence of the confounders Z_i from both the outcome and the treatment, analogous to the Frisch-Waugh-Lovell theorem but using flexible neural networks instead of linear regression. The use of out-of-sample predictions is the essence of cross-fitting and prevents overfitting bias. Finally, after computing residuals for all observations across both subsets, I estimate the target parameter θ using a regression of the outcome residuals on the treatment residuals, pooling all observations.

The specific structure of the estimation procedure, particularly the use of residuals in the final stage, corresponds to using a Neyman-orthogonal score function (or moment condition) for θ. For the PLR model, the orthogonal score function can be written as $\psi(W_i; \theta, \eta) = (Q_i - \theta P_i - g_0(Z_i))(P_i - m_0(Z_i))$, where $W_i = (Q_i, P_i, Z_i)$ and $\eta = (g_0, m_0)$ represents the nuisance functions. The DML estimate $\hat{\theta}_{DML}$ solves the empirical analogue of the moment condition $E[\psi(W_i; \theta_0, \eta_0)] = 0$. The crucial property of Neyman orthogonality is that the derivative of the expected score function with respect to the nuisance functions η, evaluated at the true values θ_0, η_0, is zero: $\partial_\eta E[\psi(W_i; \theta_0, \eta)]|_{\eta=\eta_0} = 0$. This mathematical property ensures that first-order errors in the ML estimates $\hat{g}$ and $\hat{m}$ do not introduce first-order bias into the estimate $\hat{\theta}_{DML}$. It effectively immunizes the causal parameter estimate against the "regularization bias" that is often inherent in the neural network estimators used for the nuisance functions.

Cross-fitting, as mentioned, mitigates overfitting bias. If the same data were used both to train the ML models $\hat{g}$ and $\hat{m}$ and to compute the residuals for the final regression, any overfitting by the models to the specific noise in the training data would induce a spurious correlation between the residuals $\tilde{Q}$ and $\tilde{P}$, biasing the estimate of θ. Cross-fitting breaks this dependence by ensuring that the residuals for each observation are calculated using models trained on independent

data. Together, Neyman orthogonality and cross-fitting allow DML to provide $\sqrt{N}$-consistent and asymptotically normal estimates for θ, enabling valid statistical inference, under relatively weak assumptions on the convergence rates of the ML nuisance estimators. This elegant separation of the complex prediction task (for nuisance functions) from the simpler, low-dimensional causal estimation task (for θ) allows leveraging the power of modern ML within a framework that maintains econometric rigor.

3.5 Results and Interpretation

This section presents the results from applying the Double Machine Learning methodology with multimodal embeddings to estimate the price elasticity of demand (proxied by log review volume) for the sample of 13,905 Steam games.

3.5.1 Homogeneous Elasticity Model

Table 3.6 presents the primary estimation results for the price elasticity parameter θ. For comparison purposes, results from a naive Ordinary Least Squares (OLS) regression are also included. The OLS model regresses our proxy for demand (Q), the log review volume, directly on the log price (P) and only basic tabular controls (e.g., release year, genre dummies, game features), without incorporating the high-dimensional embeddings or the DML procedure.

Table 3.6: Price Elasticity Estimates (θ) for Steam Games

Model Specification	Coefficient ($\hat{\theta}$)	Std. Error	p-value	95% Conf. Interval
1. Naive OLS	0.31	0.012	<0.001	[0.291, 0.339]
2. DML	0.11	0.022	<0.001	[0.071, 0.158]

Notes: Model 1 (Naive OLS) only controls for tabular data as confounders. Model 2 (Double Machine Learning) controls for both tabular data and multimodal embeddings (text and image) as confounders.

The positive elasticity estimates (0.31 for Naive OLS and 0.11 for DML) indicate that video games on Steam display characteristics of Giffen goods or status goods, where higher prices are associated with higher demand rather than lower demand. This contradicts the traditional Law of Demand but makes sense in the context of experience goods, such as videogames.

Consequently, raising prices would likely increase overall revenue (since the percentage increase in price outweighs the percentage increase in quantity), while lowering prices would counterintu-

itively reduce both demand and revenue—a finding with significant implications for pricing strategy.

When I account for more confounding factors using DML with multimodal embeddings, the elasticity estimate decreases from 0.31 to 0.11, suggesting that some of the positive price-quantity relationship may be explained by unobserved quality factors captured in the text and image embeddings.

This improvement in causal identification stems from DML's ability to leverage machine learning models that can detect complex non-linearities between confounders and outcomes. Unlike traditional linear approaches, the machine learning algorithms employed in the DML framework can capture nuanced relationships in the data, providing more robust causal estimates by accounting for interactions and threshold effects that linear models would miss. The multimodal embeddings prove particularly valuable in this context, as they encode rich information about consumer perceptions of the games. Text embeddings capture nuanced descriptions, feature sets, and narrative elements, while image embeddings encode visual quality, art style, and graphical fidelity—all factors that significantly influence consumer expectations and purchasing decisions.

By controlling for these perception-based factors, I obtain a clearer picture of the actual price-demand relationship. This aligns with the experience goods framework, where consumers use price as a signal of quality in the absence of perfect information. Video games are classic experience goods since players cannot fully evaluate quality before purchasing. Higher prices may signal higher production values, more content, or greater expected enjoyment. The reduced coefficient in the DML model suggests that once I control for the actual quality features (through text descriptions and images), the "price as a quality signal" effect diminishes but remains positive.

However, this aggregate estimate likely masks substantial variation across different market segments. Given the diverse nature of video games on Steam—spanning various genres, price points, and quality levels—I now turn my attention to exploring the heterogeneous price elasticities across distinct game clusters and genres to uncover more nuanced demand patterns in this complex digital market.

3.5.2 Heterogeneous Elasticity Model

Figure 3.1 shows the estimated price elasticities of demand by game genre using the DML framework and controlling for multimodal embeddings. It reveals remarkable heterogeneity in price elasticity across different Steam genres, validating our market segmentation hypothesis.

Professional software categories reveal more diverse patterns. Software Training exhibits the

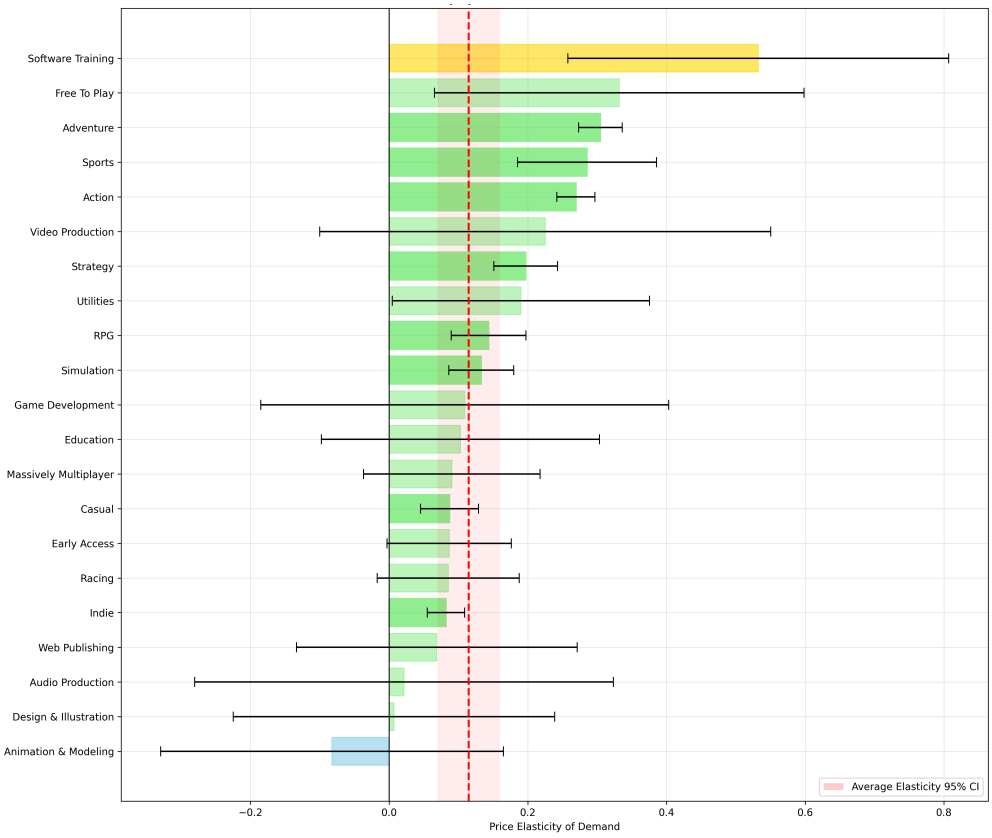


Figure 3.1: Price Elasticity by Game Genre

highest elasticity (0.53), suggesting quality signaling is particularly strong in this category. In contrast, Animation & Modeling is the only category showing negative elasticity (-0.08), following traditional economic principles where price increases reduce demand. Most other professional categories (Audio Production, Design & Illustration) display elasticities close to zero. The confidence intervals provide important context—most professional software genres have intervals crossing zero, indicating we cannot conclude with statistical precision whether their true elasticities are positive or negative.

In contrast, Entertainment genres demonstrate the most consistent positive elasticity patterns. Adventure (approximately 0.30), Action (0.27), Free To Play (0.33), and Sports (0.28) all show relatively stronger positive elasticities, indicating consumers in these categories may use price as a quality signal when making purchasing decisions. This suggests strong brand loyalty, differentiated experiences, or price insensitivity among core gamers who prioritize content over cost.

Figures 3.2 and 3.3 further explore the relationship between price elasticity, market size, and price points, revealing other interesting patterns within the Steam digital store.

Examining the relationship between price elasticity and average price (Figure 3.2) reveals intriguing price-point dynamics. Entertainment genres with the strongest positive elasticities (Adventure, Action, Free To Play) cluster in the affordable $7-10 price range, suggesting this may be an optimal price zone where quality signaling effects are maximized. In contrast, higher-priced professional tools ($15-25) show divergent patterns—Software Training benefits substantially from quality signaling despite its premium price point (approximately $19), while similarly priced categories like Game Development and Education trend toward more traditional economic responses. This non-linear relationship between price and elasticity challenges simplistic pricing models and indicates category-specific pricing norms may be more influential than absolute price levels.

Figure 3.3 adds another dimension to our understanding by revealing that elasticity patterns do not strongly correlate with market size. Indie represents the largest category (approximately 8,000 titles) yet shows relatively modest elasticity (0.08), suggesting size alone doesn't determine price sensitivity. Meanwhile, mid-sized categories like Adventure and Action (4,000-5,000 games) demonstrate some of the strongest positive elasticities, indicating these established entertainment markets may have developed robust quality-signaling mechanisms. The smallest categories (<1,000 games) show the most varied elasticity responses, from Software Training's strongly positive elasticity to Animation & Modeling's negative elasticity, highlighting how specialized markets develop distinct pricing dynamics.

Figure 3.2: Price Elasticity vs Average Price by Game Genre

Figure 3.3: Price Elasticity vs Sample Size by Game Genre

Together, this dimension of heterogeneity demonstrates that Steam's marketplace consists of distinct economic ecosystems with varying relationships between price, market size, and consumer behavior. These findings reinforce the importance of targeted pricing strategies that consider not just genre norms but also relative market positioning and the varying role of price as a quality signal across different segments of the digital marketplace.

3.6 Limitations

Despite the methodological advancements, this study is subject to several limitations. The use of log review volume as the outcome variable Q_i represents a significant constraint, as repeatedly emphasized. The estimated elasticity reflects the responsiveness of this specific proxy to price changes, conditional on controls. While likely correlated with actual sales or player engagement, it is not a direct measure of demand. Factors influencing review-writing behavior itself, beyond purchasing behavior, could affect the results.

Additionally, the validity of the causal interpretation rests on the untestable Conditional Independence Assumption. While the rich multimodal embeddings used for Z_i make CIA more plausible by controlling for many observable characteristics, there might still be unobserved confounders. Examples could include uncaptured aspects of game quality or novelty, external marketing campaigns not reflected in store data, platform-specific promotional events, or major changes in competitor offerings. Similarly, the Overlap assumption might be violated for highly unique games, and SUTVA could be challenged by strong network effects or competitive spillovers. Sensitivity analyses exploring the potential impact of unobserved confounding would be a valuable addition.

The DML estimates can potentially be sensitive to the choice of the machine learning algorithms used to estimate the nuisance functions (g_0 and m_0) and the tuning of their hyperparameters. While DML is designed to be robust to small estimation errors in the nuisance functions due to Neyman orthogonality, substantial misspecification or poor performance of the ML models could still impact the final estimate of θ.

This analysis also treats price and game characteristics largely as static snapshots. In reality, prices change over time (sales, permanent reductions), and the perceived characteristics or popularity of a game can evolve. A dynamic panel data approach could potentially capture these richer dynamics but would require more complex data structures and estimation techniques. Finally, the findings are specific to the sample of games analyzed (published 2000-2019 on Steam) and the chosen demand proxy. Generalizing the specific elasticity estimate to other platforms, time periods,

or different measures of demand should be done with caution.

3.7 Conclusion

This study leverages Double Machine Learning methods enhanced with multimodal embeddings to robustly estimate the price elasticity of demand in the Steam gaming market. Findings reveal an unconventional modestly positive price elasticity overall, reflecting consumer perceptions of price as a quality indicator within digital markets for experience goods. Detailed analysis further uncovers substantial elasticity variation across distinct gaming genres, emphasizing the market's heterogeneity. This highlights the critical importance of accounting for nuanced consumer perceptions and differentiated product characteristics in pricing decisions.

Methodologically, the paper demonstrates the viability and value of incorporating rich, AI-generated features derived from unstructured textual and visual data into causal inference models. Despite its innovations, the study acknowledges limitations such as reliance on review count proxies and assumptions underpinning causal identification.

Future research should further refine these methodological frameworks, potentially incorporating dynamic panel data methods or richer outcome measures. Overall, the integration of AI-driven multimodal embeddings with rigorous causal inference frameworks marks an important advancement in empirical economic analysis, providing both methodological insights and practical strategic guidance for digital market actors.

BIBLIOGRAPHY

Ackerberg, D., Benkard, C. L., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics*, 6:4171–4276.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The skill complementarity of broadband internet *. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Almeida, A. (2018). *Governo presidencial condicionado: delegação e participação legislativa na Câmara dos Deputados*. PhD thesis, Universidade do Estado do Rio de Janeiro.

Alquezar-Yus, M. and Amer-Mestre, J. (2024). Reverse revolving doors: The influence of interest groups on legislative voting. Working paper.

Ames, B. (1995). Electoral strategy under open-list proportional representation. *American Journal of Political Science*, 39(2):406–433.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Atasoy, H. (2013). The effects of broadband internet expansion on labor market outcomes. *ILR Review*, 66(2):315–345.

Athey, S. and Imbens, G. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Bach, P., Chernozhukov, V., Klaassen, S., Spindler, M., Teichert-Kluge, J., and Vijaykumar, S. (2024). Adventures in demand analysis using ai.

Bajari, P., Cen, Z., Chernozhukov, V., Manukonda, M., Vijaykumar, S., Wang, J., Huerta, R., Li, J., Leng, L., Monokroussos, G., and Wan, S. (2023). Hedonic prices and quality adjusted price indices powered by ai.

Bao, H., Dong, L., Piao, S., and Wei, F. (2022). Beit: Bert pre-training of image transformers.

Barbosa, A., Casagrande, D., Maier, P., and Trevisan, G. (2021). Changing the pyramids: The impact of broadband internet on firm employment structures. *Working Paper*.

Battaglini, M., Sciabolazza, V., and Patacchini, E. (2023a). Abstentions and social networks in congress. Working paper.

Battaglini, M., Sciabolazza, V., and Patacchini, E. (2023b). Logrolling in congress. Working paper.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Benhabib, J., Bisin, A., and Jackson, M. O., editors (2011). *Handbook of Social Economics*. Elsevier Science, Amsterdam.

Berry, S. T., Levinsohn, J., and Pakes, A. (1995). Dynamic models of oligopolistic competition. *Journal of Econometrics*, 117:126–163.

Bhuller, M., Kostol, A. R., and Vigtel, T. C. (2019). How Broadband Internet Affects Labor Market Matching. Technical report.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th edition.

Briglauer, W., Krämer, J., and Palan, N. (2024). Socioeconomic benefits of high-speed broadband availability and service adoption: A survey. *Telecommunications Policy*, 48(7):102808.

Brollo, F., Nannicini, T., Perotti, R., and Tabellini, G. (2013). The political resource curse. *The American Economic Review*, 103(5):1759–1796.

Bu, W., . T. Y. (2023). The effects of broadband internet on employment and wages. *Journal of Global Information Management*, 31(6).

Caldeira, G. A. and Patterson, S. C. (1987). Political friendship in the legislature. *Journal of Politics*, 49:953–957.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2023). *rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs*. R package version 2.2.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cambini, C., Sabatino, L., and Zaccagni, S. (2024). The faster the better? advanced internet access and student performance. *Telecommunications Policy*, 48(8):102815.

Campbell, R. C. (2024). Need for speed: Fiber and student achievement. *Telecommunications Policy*, 48(6):102767.

Case, A. C. and Katz, L. F. (1991). The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths. NBER Working Papers 3705, National Bureau of Economic Research, Inc.

Cattaneo, M. D., Jansson, M., and Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, 18(1):234–261.

Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248.

Chan, J., Gupta, S., Li, F., and Wang, Y. (2019). Pivotal persuasion. *Journal of Economic Theory*, 180:178–202.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ml and ai.

Cingano, F. and Rosolia, A. (2012). People i know: Job search and social networks. *Journal of Labor Economics*, 30(2):291 – 332.

Cohen, L. and Malloy, C. (2014). Friends in high places. *American Economic Journal: Economic Policy*, 6(3):63–91.

Compiani, G., Morozov, I., and Seiler, S. (2025). Demand estimation with text and image data.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Cox, G. and McCubbins, M. (2007). *Legislative Leviathan: Party Government in the House*. Cambridge University Press, 2 edition.

Czernich, N. (2014). Does broadband internet reduce the unemployment rate? evidence for germany. *Information Economics and Policy*, 29:32–45.

Czernich, N., Falck, O., Kretschmer, T., and Woessmann, L. (2011). Broadband infrastructure and economic growth. *The Economic Journal*, 121(552):505–532.

Darmofal, D., Finocchiaro, C. J., and Indridason, I. H. (2023). Roll-call voting under random seating assignment. *Political Science Research and Methods*, pages 1–20.

Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21(4):23–45.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dix-Carneiro, R. and Kovak, B. K. (2015). Trade reform and regional dynamics: Evidence from 25 years of brazilian matched employer-employee data. *Policy Research Working Paper, World Bank Group, Washington, DC*, (7205).

Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Duflo, E. and Saez, E. (2003). The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *The Quarterly Journal of Economics*, 118(3):815–842.

Dutz, M. A., Mation, L. F., O'Connell, S. D., and and, R. D. W. (2017). Economy-wide and sectoral impacts on workers of brazil's internet rollout. *Forum for Social Economics*, 46(2):160–177.

Evans, W. N., Oates, W. E., and Schwab, R. M. (1992). Measuring Peer Group Effects: A Study of Teenage Behavior. *Journal of Political Economy*, 100(5):966–991.

Ferber, P. and Pugliese, R. (2000). Partisans, proximates, and poker players: The impact of homophily and proximity on communication patterns of state legislators. *Polity*, 32(3):401–414.

Figueiredo, A. and Limongi, F. (2000). Presidential power, legislative organization, and party behavior in brazil. *Comparative Politics*, 32(2):151–170.

Fong, C. (2018). Expertise, networks, and interpersonal influence in congress. *The Journal of Politics*.

Fossen, F. M. and Sorgner, A. (2022). New digital technologies and heterogeneous wage and employment dynamics in the united states: Evidence from individual-level data. *Technological Forecasting and Social Change*, 175:121381.

Fowler, J. H. (2006). Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487.

Fuhr, J., Berens, P., and Papies, D. (2024). Estimating causal effects with double machine learning – a method evaluation.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Gilligan, T. and Krehbiel, K. (1987). Collective decision-making and standing committees: An informational rationale for restrictive amendment procedures. *Journal of Law, Economics, and Organization*, 3:287–335.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548.

Greene, W. H. (2012). *Econometric Analysis*. Pearson Education, 7th edition.

Grimes, A. and Townsend, W. (2018). Effects of (ultra-fast) fibre broadband on student achievement. *Information Economics and Policy*, 44:8–15.

Guryan, J., Kroft, K., and Notowidigdo, M. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4):34–68.

Han, L. (2021). Broadband adoption and self-employment — evidence from the american community survey. *Available at SSRN: https://ssrn.com/abstract=3936667*.

Harmon, N., Fisman, R., and Kamenica, E. (2019). Peer effects in legislative voting. *American Economic Journal: Applied Economics*, 11(4):156–180.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Henriksen, A. L., Zoghbi, A. C., Tannuri-Pianto, M., and Terra, R. (2022). Education outcomes of broadband expansion in brazilian municipalities. *Information Economics and Policy*, 60:100983.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hsiao, C. (2014). *Analysis of Panel Data*. Cambridge University Press, 3rd edition.

Hua, Y. and and, H. Z. (2024). Labour misallocation and digital infrastructure: evidence from a quasi-natural experiment of 'broadband china strategy'. *Applied Economics Letters*, 0(0):1–7.

Imbens, G. W. and Rubin, D. B. (2015a). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. and Rubin, D. B. (2015b). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jin, G., Wang, G., Hu, C., Sheng, H., Li, W., and Zhang, X. (2025). Spatial spillover of digital infrastructure construction and employment of highly skilled urban workers: Empirical evidence from technological and emotional service industries in china. *Available at SSRN: https://ssrn.com/abstract=5091856*.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. (2022). Causal machine learning: A survey and open problems.

Kawaguchi, D. (2004). Peer effects on substance use among american teenagers. *Journal of Population Economics*, 17:351–367.

Kingdon, J. W. (1989). *Congressmen's Voting Decisions*. University of Michigan Press, 3 edition.

Klaassen, S., Teichert-Kluge, J., Bach, P., Chernozhukov, V., Spindler, M., and Vijaykumar, S. (2024). Doublemldeep: Estimation of causal effects with multimodal data.

Klein, G. J. (2022). Fiber-broadband-internet and its regional impact—an empirical investigation. *Telecommunications Policy*, 46(5):102331.

Koutroumpis, P. (2009). The economic impact of broadband on growth: A simultaneous approach. *Telecommunications Policy*, 33(9):471–485.

Krehbiel, K. (1993). Where's the party? *British Journal of Political Science*, 23:235–266.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution.

Kusumawardhani, N., Pramana, R., Saputri, N. S., and Suryadarma, D. (2023). Heterogeneous im-

pact of internet availability on female labor market outcomes in an emerging economy: Evidence from indonesia. *World Development*, 164:106182.

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674. The regression discontinuity design: Theory and applications.

Li, S. and Chu, Z. (2023). Machine learning for causal inference.

Litschig, S. and Morrison, K. (2012). Government spending and re-election: Quasi-experimental evidence from brazilian municipalities. *Economics Working Papers 1233, Department of Economics and Business, Universitat Pompeu Fabra*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Lowe, M. and Jo, D. (2024). Legislature integration and bipartisanship: A natural experiment in iceland. Working paper.

Masket, S. (2008). Where you sit is where you stand: The impact of seating proximity on legislative cue-taking. *Quarterly Journal of Political Science*, 3(3):301–311.

Masukawa, K., Aoyama, M., Yokota, S., Nakamura, J., Ishida, R., Nakayama, M., and Miyashita, M. (2022). Machine learning models to detect social distress, spiritual pain, and severe physical psychological symptoms in terminally ill patients with cancer from unstructured text data in electronic medical records. *Palliative Medicine*, 36(8):1207–1216. Publisher Copyright: © The Author(s) 2022.

Matthews, D. R. and Stimson, J. A. (1975). *Yeas and Nays: Normal Decision-Making in the U. S. House of Representatives*. John Wiley & Sons, New York.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Mendonça, M. J., Loureiro, P. R. A., Nascimento, A., and Ellery, R. (2021). Assessment of the effect of broadband expansion on the economy reviewed. *Review of Development Economics*, 25(4):2414–2432.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

OpenAI (2025). Chatgpt (may 1 version). https://chat.openai.com. Large language model (GPT-4.0) by OpenAI.

Patacchini, E. and Venanzoni, G. (2014). Peer effects in the demand for housing quality. *Journal of Urban Economics*, 83(C):6–17.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pereira, C. and Mueller, B. (2004). A theory of executive dominance of congressional politics: The committee system in the brazilian chamber of deputies. *The Journal of Legislative Studies*, 10(1):9–49.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Rogowski, J. and Sinclair, B. (2012). Estimating the causal effects of social interaction with endogenous networks. *Political Analysis*, 20(3):316–328.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *Quarterly Journal of Economics*.

Saia, A. (2018). Random interactions in the chamber: Legislators' behavior and political distance. *Journal of Public Economics*.

Santos, F. (2002). Parties and committees in the coalition presidential system. *Dados - Revista de Ciências Sociais*, 45(2):237–264.

Santos, F. (2021). *Congresso remoto: a experiência legislativa brasileira em tempos de pandemia*. Sociedade e Política Collection, Rio de Janeiro. 141 p.

Santos, F., Guarnieri, F., and Salles, N. (2021). 175brazil: Legislative debate under coalition presidentialism model. In *The Politics of Legislative Debates*. Oxford University Press.

Schultz, H. (1933). A comparison of elasticities of demand obtained by different methods. *Econometrica*, 1(3):274–308.

Shi, C. M. and Li, D. (2023). The impact of broadband internet on public media: Evidence from china. *Information Economics and Policy*, 65:101058.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., and Goldstein, T. (2021). Saint: Improved neural networks for tabular data via row attention and contrastive pre-training.

Stigler, G. J. (1939). The limitations of statistical demand curves. *Journal of the American Statistical Association*, 34(207):469–481.

Stock, J. H. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Stockinger, B. (2019). Broadband internet availability and establishments' employment growth in germany: evidence from instrumental variables estimations. *J Labour Market Res*, 53(7).

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition.

Uslaner, E. and Weber, R. (1977). *Patterns of Decision Making in State Legislatures*. Praeger Publishers.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wahlke, J. C. (1962). *Legislative System: Explorations in Legislative Behavior*.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edition.

Working, H. (1943). Statistical laws of family expenditure. *Journal of the American Statistical Association*, 38(221):43–56.

Wright, P. G. (1928). The tariff on animal and vegetable oils. *New York: The Macmillan Company*, page 286–319.

Xu, G., Zhou, X., Wang, M., Zhang, B., Jiang, W., Laden, F., Suh, H. H., Szpiro, A. A., Spiegelman,

D., and Wang, Z. (2024). Causal inference with double/debiased machine learning for evaluating the health effects of multiple mismeasured pollutants.

Yakusheva, O., Kapinos, K., and Weiss, M. (2011). Peer effects and the freshman 15: Evidence from a natural experiment. *College of Nursing Faculty Research and Publications*, (92).

Yang, G., Y. S. . D. X. (2023). Digital economy and wage gap between high- and low-skilled workers. *DESD*, 1(7).

Young, J. P. (1966). *The Washington Community: 1800-1828*. Columbia University Press, New York.

Zelizer, A. (2019). Is position-taking contagious? evidence of cue-taking from two field experiments in a state legislature. *American Political Science Review*, 113(2):340–352.

Zenou, Y. (2008). Social interactions and labor market outcomes in cities. IFN Working Paper No. 755.

Zhu, F. and Zhang, X. M. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2):133–148.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *The Review of Economics and Statistics*, 85(1):9–23.

Zucco, C. (2023). Brazilian legislative surveys (waves 1-9, 1990-2021). https://doi.org/10.7910/DVN/WM9IZ8. Harvard Dataverse, V1, UNF:6:h2YTat2Lb1dDZFgDj2u/Qg==.

Zucco, C. and Lauderdale, B. (2011). Distinguishing between influences on brazilian legislative behavior. *Legislative Studies Quarterly*, 36(3):363–396.

Appendix A

THE IMPACT OF OFFICE PROXIMITY IN LEGISLATIVE DECISION-MAKING: EVIDENCE FROM BRAZIL

A.1 Office Lottery

Figure A.1: Drawing for one of the offices

Figure A.2: Office Plan - Annex III

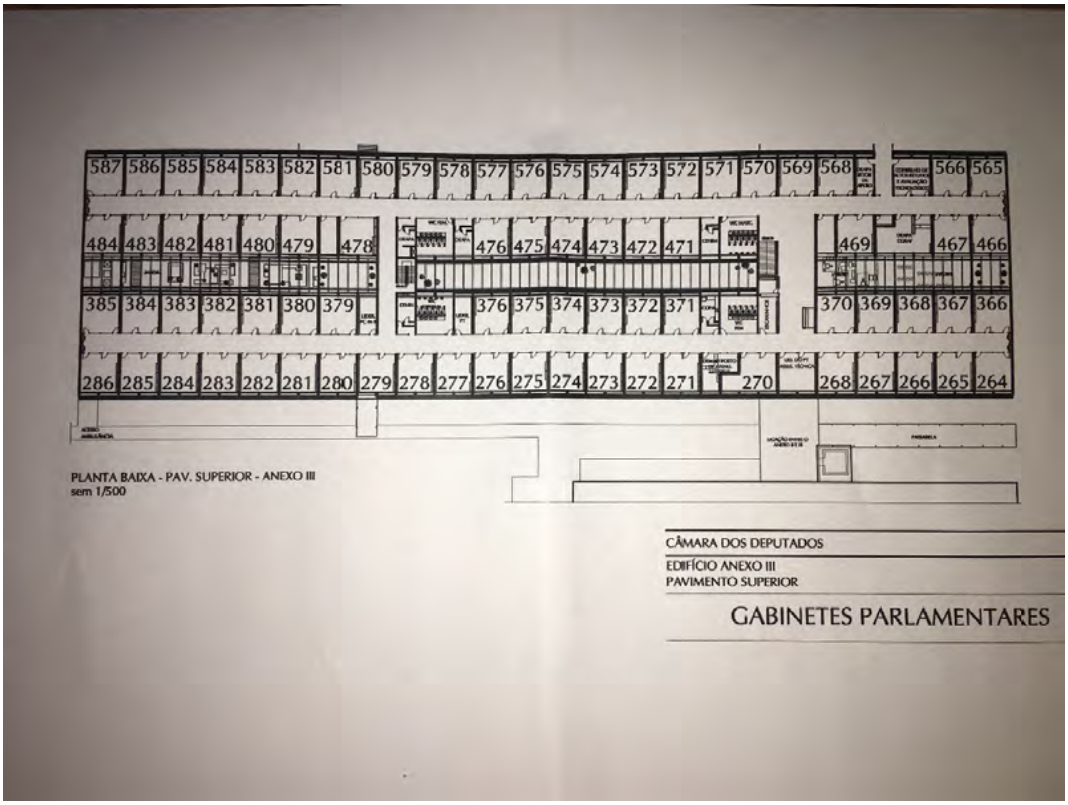
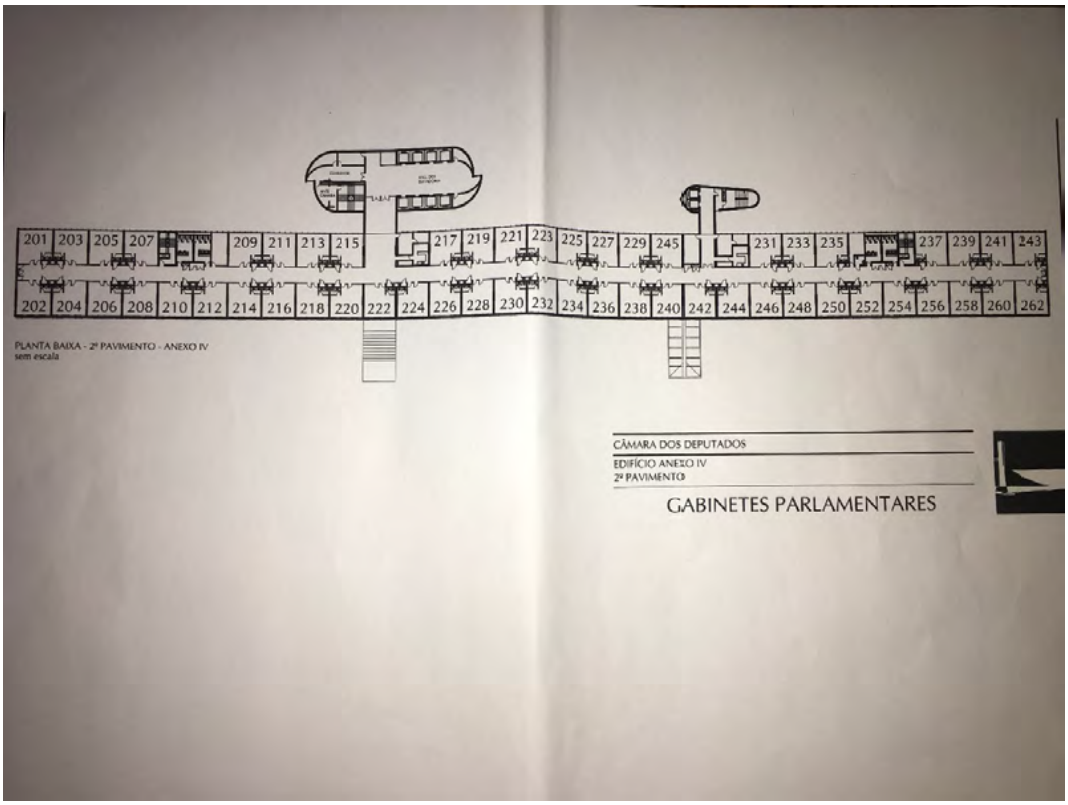


Figure A.3: Office Plan - Annex IV



A.3 Types of Legislative Proposals

Projeto de Lei or PL (en. Proposed Law) is a legislative proposal introduced by members of the Chamber of Deputies to create new laws or amend existing ones. PLs cover a wide range of topics, including social policies, economic regulations, and public administration. They undergo a comprehensive legislative process, starting with submission by a legislator, followed by review by relevant committees, debate, and voting in the Chamber of Deputies. Once approved, PLs are forwarded to the Federal Senate for further consideration. Approval of a PL requires a simple majority vote in both chambers of Congress.

Projeto de Decreto Legislativo or PDC (en. Proposed Decree of Congress) are legislative proposals introduced by members of the Chamber of Deputies to regulate matters within the legislative branch's authority. PDCs typically address issues such as ratifying international agreements, approving government acts, or revoking presidential decrees. The process for passing a PDC involves submission, review by relevant committees, debate, and voting in the Chamber of Deputies. Like PLs, PDCs require a simple majority vote in both chambers for approval.

Proposta de Emenda à Constituição or PEC (en. Proposed Constitutional Amendment) aim to modify the Constitution of Brazil, which is the highest legal document in the country. PECs address fundamental principles, rights, and the organization of government institutions. They can be initiated by the President of Brazil, one-third of the members of the Chamber of Deputies, or one-third of the members of the Federal Senate. PECs undergo a rigorous legislative process, including approval by special committees in both chambers, debate, and voting. Approval of a PEC typically requires a supermajority vote (two-thirds of the total number of legislators) in each chamber.

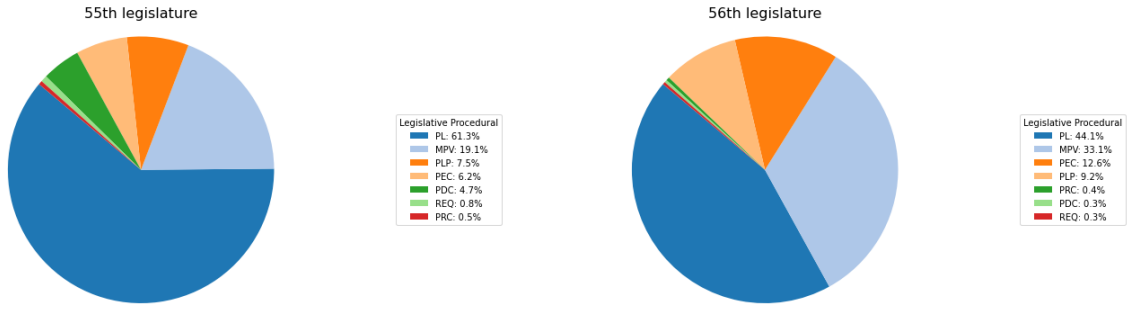
Projeto de Lei Complementar or PLP (en. Proposed Complementary Law) complement existing laws, particularly those established by the Constitution or other primary legislation. PLPs often address specific areas of law that require specialized regulation or provisions. They can be introduced by any member of the Chamber of Deputies. The process for passing a PLP is similar to that of PLs, involving submission, committee review, debate, and voting in both chambers. PLPs may require a higher threshold for approval, such as a qualified majority vote.

Medida Provisória or MPV (en. Provisional Measure) are legislative decrees issued by the President of Brazil in urgent or exceptional situations. MPVs have the force of law immediately upon issuance but require approval by Congress within a specified period to remain in effect permanently. MPVs are often used to address pressing issues such as economic crises, public health emergencies, or national security concerns. The legislative process for MPVs involves review and approval by both chambers of Congress.

Requerimento or REQ (en. Legislative Inquiry) a proposal, either verbal or written, used to make a request to the President, the Board, or the Plenary. There are various types of important requests presented by Deputies. Some examples include: request for roll call voting; urgency requests; request for the summoning of a Minister of State; request for the inclusion of out-of-agenda items; request for the creation of a parliamentary inquiry commission (CPI). [FIX]

Projeto de Resolução or PRC (en. Resolution Project) is used to regulate matters within the exclusive competence of the Chamber of Deputies and also for the Chamber to address issues within its jurisdiction, such as the loss of a legislator's mandate. The Resolution Project is deliberated only in the Chamber and is not subject to the President of the Republic's sanction. An approved Resolution Project becomes a Chamber Resolution.

Figure A.4: Distribution of Votes - 55th and 56th legislatures



A.4 Convergence$_2$

Table A.1: Summary Statistics

	Convergence$_2$ (Yes, No, Abstain and Absences)				
Variable	Mean	St.Dev.	Minimum	Maximum	N
Convergence$_2$	0.433	0.495	0	1	99,410,565
Office neighbors$_1$	0.0077	0.087	0	1	99,410,565
Same party	0.0607	0.238	0	1	99,410,565
Same coalition	0.599	0.489	0	1	99,410,565
Same state	0.0649	0.246	0	1	99,410,565
Same gender	0.928	0.257	0	1	99,410,565
Age diff.	13,65	9.85	0	62	99,410,565

Table A.2: Pair-Level Effects on Voting: Main Analysis (p.p.)

	Dependent variable: Convergence$_2$	
	(1)	(2)
Office neighbors$_1$	0.27	0.02
	(0.5)	(0.5)
Controls	No	Yes
Proposal type FE	Yes	Yes
Legislature FE	Yes	Yes
Observations	99,410,565	99,410,565
Outcome Mean	43.3	43.3

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Columns (1)-(2) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office Neighbors*$_1$ is an indicator of whether or not a pair of legislators belong to the same office neighborhood defined by taking in account all surrounding offices – if they are located adjacent to each other, either directly next to, in front of, or diagonally across from each other. Control variables are *same party*, *same state*, *same gender* and *age difference*. *p<0.1; **p<0.05; ***p<0.01

Table A.3: Pair-Level Effects on Voting: Result Margin Heterogeneity (p.p.)

	Dependent variable: Convergence$_2$				
	$> 10\%$	$< 50\%$	$< 25\%$	$< 10\%$	$< 5\%$
Office neighbors	-0.2	0.4	1.0*	1.2*	1.4*
	(0.5)	(0.6)	(0.5)	(0.6)	(0.8)
Controls	Yes	Yes	Yes	Yes	Yes
Proposal type FE	Yes	Yes	Yes	Yes	Yes
Legislature FE	Yes	Yes	Yes	Yes	Yes
Observations	35,048,688	24,042,371	7,476,884	2,712,067	1,478,644
Outcome Mean	43.6	40.0	36.5	35.4	35.5

Notes: All point estimates are presented in percentage points. Standard errors in parenthesis are two-way cluster-robust. Columns (1)-(5) report the impact of a pair of legislators belonging to the same office neighborhood on their voting agreement in roll calls with different margin of victory/defeat. *Convergence* is an indicator of whether the pair of legislators agreed in the proposal's vote. *Office neighbors* is an indicator of whether or not a pair of legislators belong to the same office neighborhood in a specific legislature. Control variables are *same party*, *ideological distance*, *same state*, *same gender*, and *age difference*. These variables are self-explanatory. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A.5 Committee Median Preferences versus Floor Preferences

Figure A.5: CCJC

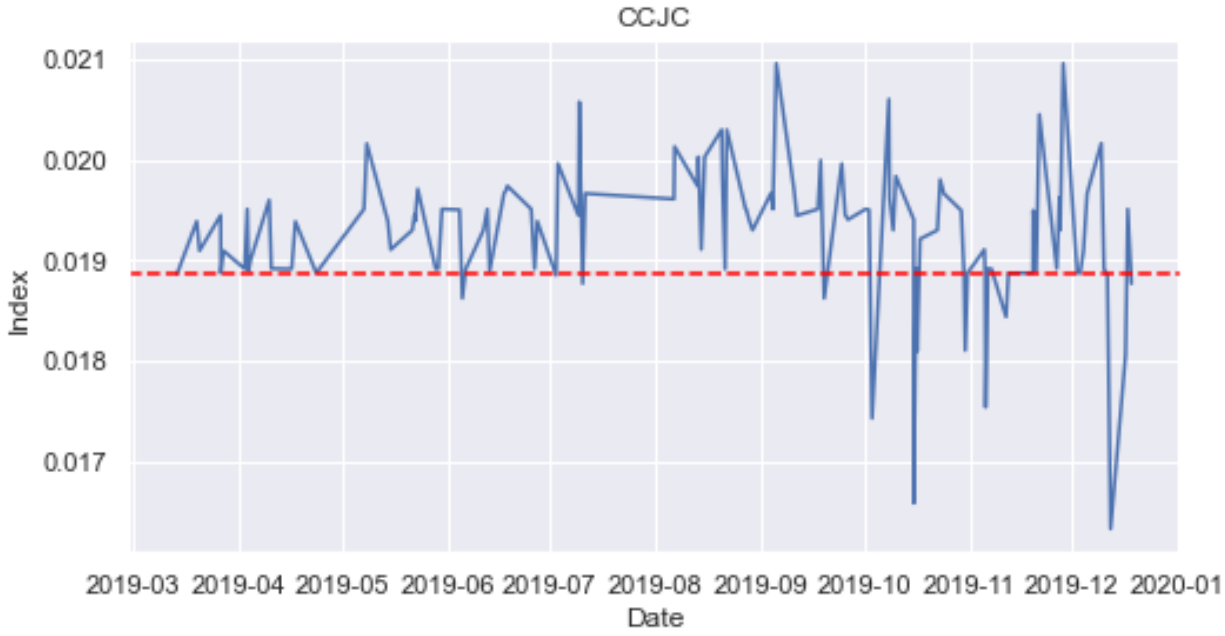


Figure A.6: CFT

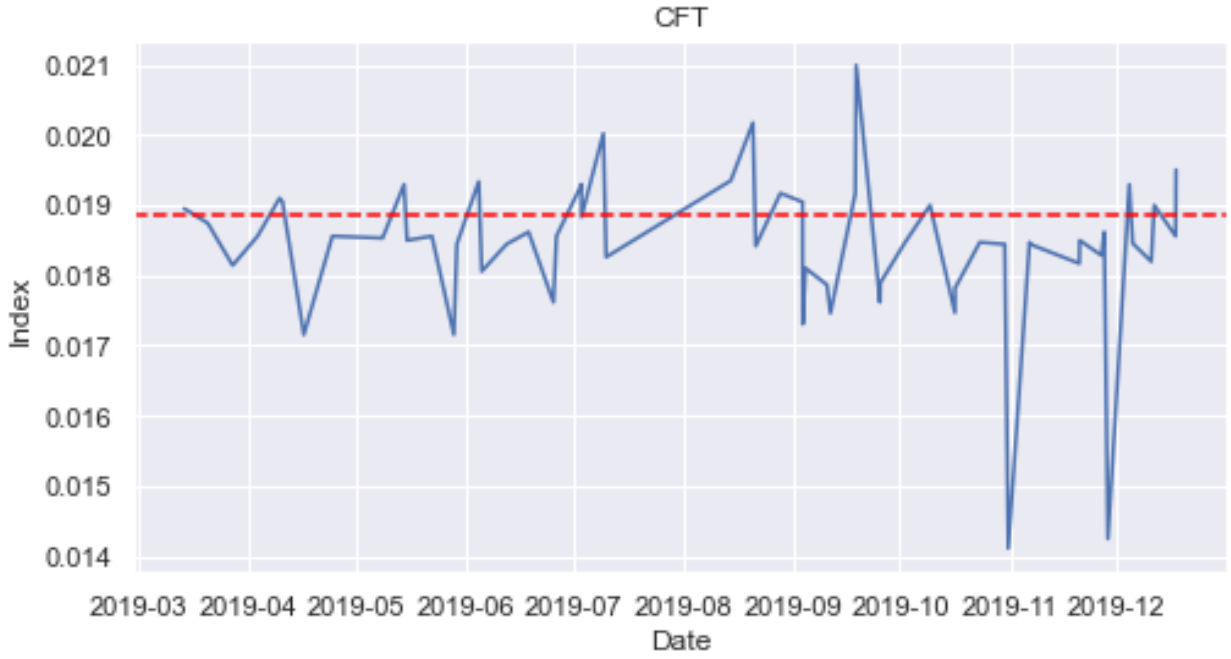
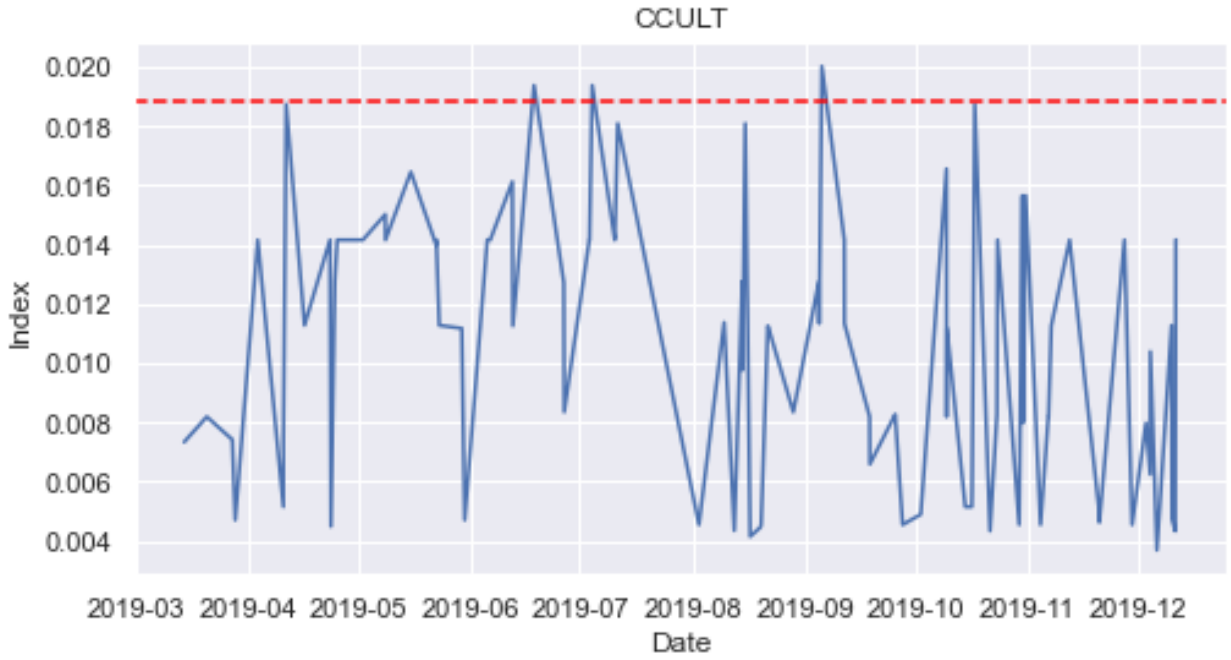


Figure A.7: CCULT



A.6 Distribution of Types across the Topics

Topics	MPV	PDC	PDL	PEC	PL	PLP	PRC	**Total**
Public Administration	237	3	4	189	252	121	4	810
Agriculture, Livestock, Fishing, and Extractivism			2			31		68
Art, Culture, and Religion	9			3	18	3	1	35
Cities and Urban Development	35				19			54
Science, Technology, and Innovation	18	1	3	2	14	16		54
Communications	28				10	1		39
Defense and Security	1	4	3	3	55			66
Civil and Procedural Law	1			3	26	4		34
Consumer Rights and Protection	39		1		15			55
Criminal and Procedural Law					38	2		40
Human Rights and Minorities	32	4	1	7	47	15	15	122
Economy	157			23	72	27		279
Education	16	2		25	50	1		94
Energy, Water Resources, and Minerals	47		1	1	17	59	23	148
Sports and Leisure	3				25			28
Land Structure	23				3			26
Public Finance and Budget	60		2	87	44	89		282
Industry, Commerce, and Services					9	4		13
Environment and Sustainable Development	13				34			47
Politics, Parties, and Elections				41	28			69
Social Security and Welfare	39				2			41
Legislative Process and Parliamentary Actions				14			9	24
International Relations and Foreign Trade	9	1	7		3			20
Health	4	2		16	28			50
Labor and Employment	57	1		3	29			90
Transportation and Mobility	40				42			82

A.7 Randomization Inference

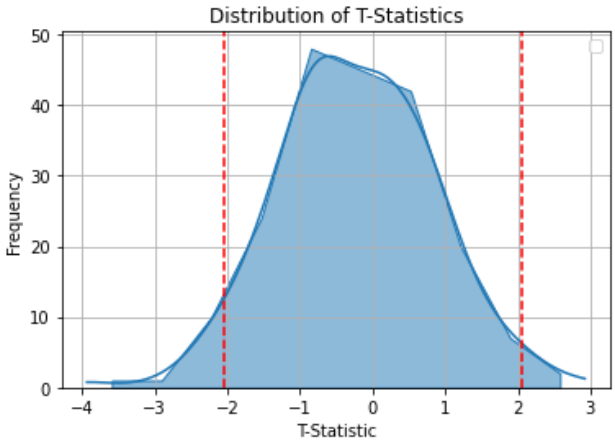


Figure A.8: Randomization Inference for causal estimate (sample: 5 p.p. result margin)

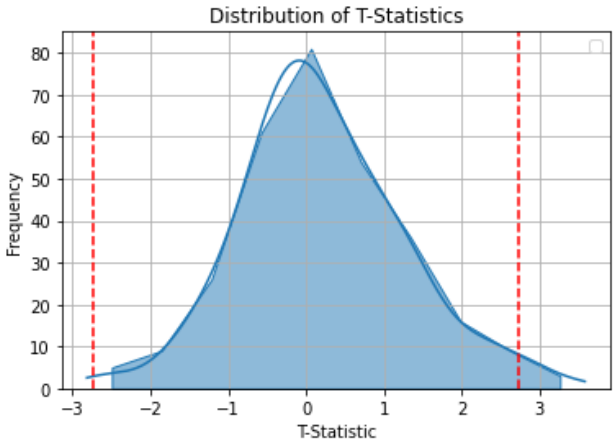


Figure A.9: Randomization Inference for Interaction of Office Neighbors and Committee.

Appendix B

FAST INTERNET AND LABOR MARKET IMPACT: EVIDENCE FROM DIFFERENTIAL BROADBAND ROLLOUT IN BRAZIL

Fast Internet and Labor Market Impact: Evidence from Differential Broadband Rollout in Brazil

B.1 Summary Statistics

B.2 Second-stage Estimates

Table B.3: Pooled RDD Estimates on Municipal Labor-Market Outcomes, 2011–2016 (Radio)

Post-treatment (pooled 2011–2016)			
Est.	S.E.	Bw. (N)	
Panel A: Net Job Creation			
10.96^{***}	(1.86)	1642.72	(660)
Panel B: Hours Worked			
-0.530^{***}	(0.039)	1244.23	(492)
Panel C: Log(Wages)			
−0.00072	(0.00052)	388.07	(186)

Notes: Estimates are obtained from a pooled fuzzy RDD covering the post-treatment period 2011–2016. "Est." is the point estimate; "S.E." is the robust bias-corrected standard error. "Bw." is the MSE-optimal bandwidth chosen by `rdrobust`; the value in parentheses is the effective number of observations inside that bandwidth (N).

	20k						40k						60k					
	Control			Treatment			Control			Treatment			Control			Treatment		
	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs
Education																		
High	6.78	34.71	31.78	6.83	35.00	94.26	6.83	35.00	94.26	6.90	33.83	175.06	6.90	33.83	175.06	6.96	36.52	256.52
	(0.20)	(44.82)	(2.17K)	(0.15)	(33.25)	(16.91K)	(0.15)	(33.25)	(16.91K)	(0.15)	(39.90)	(32.36K)	(0.15)	(39.90)	(32.36K)	(0.07)	(13.59)	(97.38K)
Mid	6.23	37.71	95.96	6.21	38.21	343.58	6.21	38.21	343.58	6.23	38.11	701.32	6.23	38.11	701.32	6.27	40.73	1384.33
	(0.06)	(29.30)	(22.88K)	(0.04)	(23.61)	(171.42K)	(0.04)	(23.61)	(171.42K)	(0.03)	(24.89)	(584.79K)	(0.03)	(24.89)	(584.79K)	(0.03)	(3.33)	(2.66M)
Low	6.05	40.01	156.20	6.00	40.68	473.67	6.00	40.68	473.67	6.02	40.71	966.73	6.02	40.71	966.73	6.10	41.91	1768.10
	(0.07)	(23.97)	(116.4K)	(0.07)	(15.61)	(670.8K)	(0.07)	(15.61)	(670.8K)	(0.04)	(19.64)	(3.04M)	(0.04)	(19.64)	(3.04M)	(0.02)	(2.35)	(6.32M)
Occupation																		
High-skill	6.10	41.51	-22.42	6.06	42.09	-108.31	6.06	42.09	-108.31	6.06	42.47	-93.62	6.06	42.47	-93.62	6.13	43.17	-268.63
	(0.06)	(13.93)	(42.60K)	(0.04)	(9.45)	(193.63K)	(0.04)	(9.45)	(193.63K)	(0.05)	(6.33)	(1.43M)	(0.05)	(6.33)	(1.43M)	(0.03)	(0.56)	(1.11M)
Low-skill	6.28	36.54	10.91	6.26	37.17	56.22	6.26	37.17	56.22	6.31	36.06	32.98	6.31	36.06	32.98	6.35	38.96	294.52
	(0.07)	(31.86)	(42.82K)	(0.05)	(25.62)	(167.98K)	(0.05)	(25.62)	(167.98K)	(0.03)	(33.45)	(1.38M)	(0.03)	(33.45)	(1.38M)	(0.02)	(8.11)	(725.10K)
Gender																		
Female	6.15	36.76	129.68	6.11	37.49	407.32	6.11	37.49	407.32	6.16	36.34	760.70	6.16	36.34	760.70	6.18	39.45	1341.46
	(0.06)	(32.13)	(33.33K)	(0.06)	(26.54)	(210.10K)	(0.06)	(26.54)	(210.10K)	(0.02)	(34.36)	(594.07K)	(0.02)	(34.36)	(594.07K)	(0.01)	(7.78)	(1.84M)
Male	6.21	40.04	153.44	6.16	40.56	504.19	6.16	40.56	504.19	6.19	40.60	1082.42	6.19	40.60	1082.42	6.24	42.13	2067.50
	(0.05)	(19.05)	(107.40K)	(0.04)	(11.97)	(681.91K)	(0.04)	(11.97)	(681.91K)	(0.03)	(12.18)	(3.21M)	(0.03)	(12.18)	(3.21M)	(0.02)	(0.91)	(8.46M)
Industry																		
Primary	5.96	42.82	-25.73	5.98	43.20	-36.06	5.98	43.20	-36.06	6.01	43.41	-24.36	6.01	43.41	-24.36	5.99	43.17	-172.04
	(0.21)	(14.09)	(7.10K)	(0.18)	(13.28)	(37.10K)	(0.18)	(13.28)	(37.10K)	(0.06)	(2.41)	(47.25K)	(0.06)	(2.41)	(47.25K)	(0.86)	(3.91)	(531.85K)
Manufacturing	6.09	43.04	-31.85	6.11	43.15	-105.93	6.11	43.15	-105.93	6.16	43.43	-236.50	6.16	43.43	-236.50	6.14	43.63	-597.92
	(0.20)	(14.24)	(16.81K)	(0.19)	(5.82)	(63.91K)	(0.19)	(5.82)	(63.91K)	(0.10)	(1.80)	(218.38K)	(0.10)	(1.80)	(218.38K)	(0.06)	(0.17)	(941.09K)
Commerce	6.05	42.81	-33.76	6.01	43.17	-23.93	6.01	43.17	-23.93	6.00	43.32	52.99	6.00	43.32	52.99	6.03	43.62	266.48
	(0.25)	(3.45)	(105.01)	(0.14)	(2.24)	(240.45)	(0.14)	(2.24)	(240.45)	(0.09)	(0.83)	(393.65)	(0.09)	(0.83)	(393.65)	(0.09)	(0.42)	(505.51)
Services	6.19	35.86	83.15	6.16	36.39	161.50	6.16	36.39	161.50	6.21	34.53	223.27	6.21	34.53	223.27	6.26	37.81	506.79
	(0.27)	(6.02)	(148.19)	(0.26)	(5.74)	(326.44)	(0.26)	(5.74)	(326.44)	(0.17)	(6.85)	(569.04)	(0.17)	(6.85)	(569.04)	(0.14)	(3.86)	(1111.76)

Table B.1: Summary statistics (mean and standard deviation) for log wages, hours worked, and net job creation, by population threshold, treatment status, and subgroup. Municipalities treated by microwave radio backhaul.

	20k						40k						60k					
	Control			Treatment			Control			Treatment			Control			Treatment		
	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs	Wage	Hours	NetJobs
Education																		
High	6.83	34.05	47.09	6.86	35.41	121.24	6.86	35.41	121.24	6.88	36.11	216.85	6.88	36.11	216.85	6.87	35.68	522.16
	(0.14)	(41.15)	(6.9K)	(0.15)	(32.38)	(321.0K)	(0.15)	(32.38)	(321.0K)	(0.13)	(34.46)	(52.8K)	(0.13)	(34.46)	(52.8K)	(0.14)	(20.39)	(568.5K)
Mid	6.28	38.55	146.51	6.24	39.11	432.83	6.24	39.11	432.83	6.25	40.03	948.31	6.25	40.03	948.31	6.28	40.38	2083.13
	(0.06)	(24.87)	(63.5K)	(0.05)	(18.56)	(298.9K)	(0.05)	(18.56)	(298.9K)	(0.04)	(22.27)	(1.09M)	(0.04)	(22.27)	(1.09M)	(0.02)	(6.36)	(10.37M)
Low	6.08	41.04	230.98	6.01	41.02	669.61	6.01	41.02	669.61	6.07	42.50	1325.73	6.07	42.50	1325.73	6.10	42.08	2236.10
	(0.15)	(17.20)	(203.2K)	(0.10)	(13.03)	(1.26M)	(0.10)	(13.03)	(1.26M)	(0.03)	(4.61)	(3.12M)	(0.03)	(4.61)	(3.12M)	(0.02)	(2.99)	(14.88M)
Occupation																		
High-skill	6.13	42.17	-0.34	6.06	42.25	-33.40	6.06	42.25	-33.40	6.14	43.30	-59.04	6.14	43.30	-59.04	6.15	42.88	-679.98
	(0.17)	(10.53)	(61.72K)	(0.14)	(7.71)	(599.86K)	(0.14)	(7.71)	(599.86K)	(0.06)	(0.60)	(674.23K)	(0.06)	(0.60)	(674.23K)	(0.02)	(1.26)	(3.79M)
Low-skill	6.37	36.83	-9.92	6.32	37.65	-29.37	6.32	37.65	-29.37	6.31	38.77	-100.12	6.31	38.77	-100.12	6.37	38.69	386.76
	(0.09)	(27.58)	(62.48K)	(0.06)	(23.20)	(475.64K)	(0.06)	(23.20)	(475.64K)	(0.03)	(27.29)	(655.34K)	(0.03)	(27.29)	(655.34K)	(0.05)	(9.66)	(2.96M)
Gender																		
Female	6.19	37.27	178.55	6.13	37.92	496.96	6.13	37.92	496.96	6.16	39.02	897.44	6.16	39.02	897.44	6.21	38.83	1871.11
	(0.08)	(28.68)	(68.53K)	(0.05)	(24.43)	(453.30K)	(0.05)	(24.43)	(453.30K)	(0.02)	(27.35)	(855.27K)	(0.02)	(27.35)	(855.27K)	(0.03)	(10.24)	(7.79M)
Male	6.25	39.50	-5.13	6.19	39.95	-31.39	6.19	39.95	-31.39	6.23	41.03	-79.58	6.23	41.03	-79.58	6.26	40.79	-146.61
	(0.15)	(26.17)	(62.09K)	(0.11)	(20.74)	(537.03K)	(0.11)	(20.74)	(537.03K)	(0.05)	(19.01)	(660.92K)	(0.05)	(19.01)	(660.92K)	(0.04)	(9.86)	(3.64M)
Industry																		
Primary	5.98	42.59	-22.82	5.99	43.34	-53.69	5.99	43.34	-53.69	6.10	43.02	-20.40	6.10	43.02	-20.40	6.13	43.26	-302.43
	(0.40)	(3.67)	(100.72)	(0.34)	(2.74)	(203.74)	(0.34)	(2.74)	(203.74)	(0.45)	(4.73)	(216.39)	(0.45)	(4.73)	(216.39)	(0.93)	(2.90)	(1051.60)
Manufacturing	6.11	43.15	-41.07	6.13	43.11	-172.43	6.13	43.11	-172.43	6.05	43.36	-276.22	6.05	43.36	-276.22	6.24	43.40	-574.71
	(0.44)	(2.39)	(158.00)	(0.37)	(2.99)	(443.03)	(0.37)	(2.99)	(443.03)	(0.69)	(2.78)	(577.63)	(0.69)	(2.78)	(577.63)	(0.26)	(0.83)	(1260.17)
Commerce	6.06	42.81	-15.86	6.01	43.27	30.75	6.01	43.27	30.75	6.04	43.43	77.86	6.04	43.43	77.86	6.04	43.57	423.08
	(0.18)	(2.82)	(108.52)	(0.15)	(1.52)	(221.45)	(0.15)	(1.52)	(221.45)	(0.11)	(0.90)	(382.97)	(0.11)	(0.90)	(382.97)	(0.10)	(0.39)	(970.60)
Services	6.27	35.85	79.24	6.19	36.58	192.87	6.19	36.58	192.87	6.22	37.74	238.64	6.22	37.74	238.64	6.28	37.07	463.40
	(0.33)	(5.71)	(174.78)	(0.27)	(5.40)	(474.13)	(0.27)	(5.40)	(474.13)	(0.19)	(6.23)	(719.61)	(0.19)	(6.23)	(719.61)	(0.22)	(4.15)	(1286.25)

Table B.2: Summary statistics (mean and standard deviation) for log wages, hours worked, and net job creation, by population threshold, treatment status, and subgroup. Municipalities treated by fiber backhaul.

Table B.4: Second-stage Estimates on Municipal Labor Market Outcomes by Year (2006–2016) - Fiber

	Pre-treatment		Treatment Roll-out			Post-treatment					
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Net Job Creation											
Est.	-50.12	-7.89	18.45	-22.67	12.34	8.91	-50.23	30.87	20.45	1.98	12.35
S.E.	(70.45)	(9.23)	(19.32)	(13.45)	(15.23)	(25.12)	(35.76)	(26.45)	(18.56)	(13.12)	(13.78)
Bw.	5200.12	5300.45	3850.78	4450.23	6180.34	5650.89	5100.23	3950.12	5400.45	8200.56	6200.78
N.	2161	2161	2159	2161	2161	2161	2161	2161	2161	2161	2161
Panel B: Hours Worked											
Est.	0.1023	0.0541	-0.1304	-0.0456	-0.0912	-0.1650	-0.2034	-0.1208	-0.1823	-0.3012	-0.2879
S.E.	(0.3102)	(0.2103)	(0.2156)	(0.2134)	(0.2250)	(0.1901)	(0.1456)	(0.1556)	(0.1589)	(0.1703)	(0.1754)
Bw.	5300.56	7800.12	6800.78	6600.34	5800.45	6500.23	7700.89	8250.45	7350.23	6750.89	6100.12
N.	2161	2161	2158	2161	2161	2161	2161	2161	2161	2160	2161
Panel C: Log(Wages)											
Est.	-0.01234	-0.01845	-0.01356	-0.01678	-0.00987	-0.01345	-0.01267	-0.03512	-0.02645	-0.01234	-0.00321
S.E.	(0.01245)	(0.01312)	(0.01478)	(0.01234)	(0.01345)	(0.00923)	(0.00812)	(0.01987)	(0.01567)	(0.00745)	(0.00812)
Bw.	5800.23	5550.12	5200.78	5850.34	5700.23	8050.89	8000.12	5300.45	5450.89	7900.78	6250.45
N.	2161	2161	2159	2161	2161	2161	2161	2161	2161	2161	2161

Notes: This table presents second-stage estimates for three labor outcomes – net job creation, hours worked, and log(wages) – across years 2006–2016. Years 2006–2007 serve as falsification tests, 2008–2010 represent the policy roll-out period, and 2011–2016 reflect the policy impact period. "Est." indicates the point estimate, "S.E." represents the robust bias-corrected standard error, "Bw." denotes the bandwidth, and "N." the number of observations.

B.3 Fast Internet and Firms Impact

We investigate in this section how the availability of faster internet impacts firms by looking at net firm entry and exports. Table B.5 shows the results.

Table B.5: Second-stage Estimates on Municipal Outcomes, Radio Technology (2006–2016)

		Pre-treatment		Treatment Roll-out			Post-treatment					
Outcome		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Panel A: Firm Entry												
	Est.	–	0.20	1.46	0.94	2.43	-4.86	1.62	3.76*	-2.44	-5.19*	3.00
	S.E.	–	(2.02)	(2.85)	(2.04)	(2.62)	(3.99)	(2.99)	(2.13)	(3.02)	(2.76)	(2.42)
	Bw.	–	5,852	5,675	4,924	6,140	5,546	5,036	5,365	6,274	5,906	5,499
	N.	–	2161	2161	2161	2161	2161	2161	2161	2161	2161	2161
Panel B: Log(Exports)												
	Est.	-0.42	0.54	1.35	5.20	0.85	1.43	0.07	0.29	0.52	0.31	2.24
	S.E.	(1.14)	(1.49)	(1.99)	(11.09)	(0.70)	(1.56)	(0.61)	(0.51)	(0.65)	(1.94)	(2.32)
	Bw.	3,983	2,982	4,328	5,321	4,338	5,161	6,983	5,108	4,894	4,763	4,423
	N.	238	256	262	246	242	249	256	273	271	280	298

Notes: Estimates are obtained via local-polynomial regressions with a triangular kernel and optimal bandwidth (Bw.). Standard errors (S.E.) are robust and bias-corrected. *N* reports the number of municipal observations used in each yearly regression. Pre-treatment (2006–07), roll-out (2008–10), and post-treatment (2011–16) periods follow the policy timeline.

Table B.5 suggests that the roll-out of higher-speed Radio backhaul produced no systematic

Figure B.1: Net Firm Entry

improvement in local business dynamism. During the official implementation phase (2008–2010) the point estimates for net firm entry are positive but highly imprecise, and exports display a similarly erratic pattern (e.g., a large but noisily estimated spike in 2009). In the post-treatment period (2011–2016) the results continue to lack a coherent trend: net firm entry jumps by about 3.8 additional firms in 2013 but then falls by roughly 5 firms in 2015, with both effects significant at the 10 percent level yet opposite in sign; every other year's estimate is indistinguishable from zero. Log-exports remain essentially flat throughout, never achieving statistical significance—even the largest estimate, in 2016, is less than one standard error away from zero. Taken together, these findings imply that faster Radio-based connectivity neither catalysed sustained entry of new firms nor boosted export activity in the affected municipalities. The year-to-year fluctuations, coupled with wide confidence intervals, also point to limited statistical power and emphasise that any positive effects of Radio broadband are at best transitory and, more plausibly, economically negligible.

Appendix C

DOUBLE MACHINE LEARNING FOR PRICE ELASTICITY ESTIMATION: LEVERAGING UNSTRUCTURED DATA FROM THE STEAM DIGITAL STORE

C.1 Incorporating Multimodal Embeddings as Confounders

The Transformer architecture, introduced in Vaswani et al. (2017), marked a significant shift in deep learning for sequence modeling, especially within Natural Language Processing (NLP). By replacing sequential processing with attention mechanisms, Transformers enable parallel computation and excel at capturing long-range dependencies, leading to state-of-the-art results in machine translation, text summarization, and other tasks. A key innovation lies in transformer embeddings: dynamic, context-sensitive vector representations of input tokens.

The efficacy of these representations stems from several key AI principles integrated into the models (Bach et al., 2024). First, Self-Supervised Learning tackles the scarcity of labeled data by generating learning tasks directly from unlabeled data. For instance, a model might be trained to predict masked or corrupted portions of its input (e.g., reconstructing "Well made diecast model truck with metal body" from a masked version like "Well made [m] model truck with [m] body"), effectively transforming each input sample into a self-labeled instance and enabling the model to capture complex syntactic and semantic relationships inherent in the data without requiring explicit annotations. The resulting internal representations (embeddings) encapsulate learned features, and this approach is generalizable across data modalities, including images. Second, the Attention Mechanism, a core component of transformer models, permits the model to dynamically and selectively weigh the importance of different components within the input data when forming representations. This adaptive weighting produces contextually rich embeddings that capture nuanced relationships, offering an advantage over earlier context-free models like Word2Vec or GloVe (Mikolov et al., 2013; Pennington et al., 2014) and contributing to state-of-the-art performance across various tasks.

While pre-trained embeddings capture general features, their utility for specific downstream tasks can be significantly enhanced through fine-tuning. In our context, the primary goal is causal inference—specifically, estimating the price elasticity of demand. Therefore, we implement Causal

Fine-Tuning, adapting the learned embeddings by further training the model to optimize the prediction of relevant target variables: the quantity signal (Q_{it}) and the price signal (P_{it}). This aligns the fine-tuning objective precisely with the requirements for orthogonal estimation of causal effects, a technique discussed previously in Section 2.3 and further in Section 4. During fine-tuning, the embeddings serve as inputs to specialized prediction layers, and the resulting prediction errors are used to update the parameters throughout the model, including the embedding layers, via gradient descent and back-propagation. A diagram illustrating this process (Bach et al., 2024):

$$
\begin{array}{ccccc}
 & & A_i^{tx} & & \\
 & & \uparrow & & \\
X_i^{in} = \begin{bmatrix} \text{Text}_i \\ \text{Image}_i \end{bmatrix} & \overset{e}{\longmapsto} & E_i := \begin{bmatrix} T_i \\ I_i \end{bmatrix} & \overset{m}{\longmapsto} & \{\hat{Q}_i, \hat{P}_i\}_{i=1}^N. \\
 & & \downarrow & & \\
 & & A_i^{im} & &
\end{array}
$$

In this diagram, X_i^{in} represents the raw multimodal inputs (textual content Text_i and visual components Image_i). The embedding function e transforms these unprocessed inputs into their corresponding vector representations E_i. The notations A_i^{im} and A_i^{tx} indicate the auxiliary masked targets utilized within the self-supervised learning approach. The model develops effective representations by learning to reconstruct these masked elements using only the visible portions of the input data. The function m denotes a subsequent prediction layer that employs the generated embeddings E_i to forecast variables of economic relevance, particularly $\hat{Q}_{it}$ (quantity metrics) and $\hat{P}_{it}$ (price metrics).

Through continuous improvement of its ability to reconstruct these auxiliary targets, the model progressively enhances the quality and informativeness of its internal vector representations. Fine-tuning adjusts both e and m to ensure that the embeddings and downstream predictions align with the target predictive and causal inference questions.

C.2 Qualitative Assessment

C.2.1 Only Text Clusters

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	Zap Zone	Mighty Mob Games — Casual, Action, Indie — Attention! The Zugs have escaped the detention zone! Zap Zone is a maze shooter that borrows inspiration from classics like Pac-Man. Here you're put on the scene as Zoey, a space vigilant who finds herself in the middle of a re...
2	Spaceguard 80	Draftline Games — Action, Indie, Casual, Adventure — The intergalactic union is collapsing dragging the galaxy into chaos. No safe place anywhere anymore! The rescue team led by Spacecat receives a distress call. He has to make a difficult decision - whether to help th...
3	Trash Squad	Enitvare — Action, Indie, Shooter — Trash Squad is a dynamic shooter with RPG elements. Stand up for the fight against hordes of monsters in over a dozen randomly generated levels. Every one of the many available characters has its own unique skills and free choice while distr...
4	Spirit Run - Fire vs. Ice	Libredia — Lunagames — Indie, Action — Get ready for the run of a lifetime. In this endless platformer you play an elemental spirit that has the ability to change form. Spirit Run: Fire vs Ice is inspired by games such as the platformer Canabalt and the classic shoot 'em...
5	Glow	Impetus Games — Action, Indie, Casual — Glow is a fast-paced action game in the spirit of the good old 80's & 90's 2D top-down shooters, presented in modern 3D graphics. You control a lone but brave Firefly, to help her fight evil spiders and critters that try to hunt you down. ...

Table C.1: Cluster #1 - Casual Indie Action Games with Retro Influences

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	IMAZE.EXE	AFBIK Studio — AFBIK Studio — Casual, Indie, Adventure, Puzzle, 2D — IMAZE.EXE - game where you need to found the way from maze. Features: - 40 levels. - Pleasant music. - More than 2 hours of gameplay. - Simple controls. - Steam Achievements. Credits: Programming: AFBIK Studio Music and eff...
2	KNACK!	Ardi Studio — Ardi Studio — Indie, Casual, Simulation, Action, Adventure, Singleplayer, Platformer, Relaxing, Minimalist, Colorful — Minimalistic arcade platformer. You are able to move only in one direction, can you get through it? 35+ Level 150+ Achievements Trading Cards Emoticons ...
3	Countryballs: Over The World	Divertic — Garage Games — Indie, Adventure, Memes, Casual — Hello there, player! Countryballs: Over The World is an adventure game obviously based on Countryballs jokes. You are playing as Polandball, who just ran out of his ointment. Will you succeed and get it? List of ...
4	DEADLY WHEELS	meokigame — Zotdinex — Indie, Casual, Simulation, Racing, Sports, RPG — "Deadly wheels" - simulator of the race for survival. You are expected to: two fascinating locations; a wide range of cars with unique characteristics; weapons for every taste; a bright and memorable low-poly design;...
5	TheLooppy	Flatingo — Atriagames — Adventure, Indie, Platformer — TheLooppy is a hardcore 2D platformer with puzzle elements, in which you help the hero Looppy to go through the magical valley to the crystal of desires, so that his dream of becoming big can finally be realized. Do not leave this cuti...

Table C.2: Cluster #2 - Small-Scale Indie Casual Games

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	Moonstrider	Triple Lasers — Triple Lasers — Early Access, Action, Indie, Adventure, Early Access, Rogue-like — Moonstrider is an action-packed space rogue-like where you sail through space with rocket boots, fight deadly and cunning aliens, and explore the galaxy. Each playthrough features uniquely gen...
2	Last Encounter	Exordium Games — Exordium Games — Action, Indie, Local Co-Op, Rogue-lite, Space, Top-Down Shooter, Co-op, Family Friendly, 4 Player Local, Procedural Generation, Great Soundtrack, Colorful, Sci-fi, Rogue-like, Shooter, Top-Down, Shoot 'Em Up, Dungeon Crawler, Local Multiplayer, Twin Stick Shooter — La...
3	Achtung! Cthulhu Tactics	Auroch Digital — Ripstone — Strategy, Action, Violent, Indie, RPG, Lovecraftian, Turn-Based Tactics — Challenge the horrifying reign of Nazi terror and battle an immortal evil in Achtung! Cthulhu Tactics; a turn-based tactical strategy game set in the award-winning Achtung! Cthu...
4	Warden: Melody of the Undergrowth	Cardboard Keep — Cardboard Keep — Action, Adventure, RPG, Indie, 3D Platformer, Platformer, Stylized, Colorful, Swordplay, Puzzle-Platformer, Multiple Endings, Lore-Rich, Inventory Management, 1990's, Character Action Game — Trapped in an ancient forest, a young prince searches...
5	Brut@l	Stormcloud Games — Rising Star Games — Action, Gore, Violent, Dungeon Crawler, Rogue-like, Co-op, Indie, Rogue-lite, Local Co-Op — Brut@l is a modern reimagining of the classic ASCII dungeon crawler, fusing old-school gaming with a stunning 3D visual style to create an adventure that's t...

Table C.3: Cluster #3 - Indie Action Games with Roguelike Elements

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	FSX Steam Edition: Airbus Series Vol. 4 Add-On	FeelThere — Dovetail Games - Flight — Simulation — About This Content FeelThere's Airbus Series Vol. 4 is a detailed depiction of the fascinating world of Fly-By-Wire technology - as well as the Airbus' high-end systems. These powerful aircraft ...
2	Universal Combat CE	3000AD — Strategy, Simulation, Action, Space, Indie, Sci-fi, Space Sim — THE MOST ADVANCED SPACE COMBAT CAPITAL SHIP SIMULATOR. EVER. SERIOUSLY. In Feb 2015, as part of the Battlecruiser twenty-five year anniversary celebration, a refresh of Universal Combat v2.0 was release...
3	FSX Steam Edition: Airbus Series Vol. 3 Add-On	FeelThere — Dovetail Games - Flight — Simulation — About This Content FeelThere's Airbus Series Vol. 3 is a detailed depiction of the fascinating world of Fly-By-Wire technology as well as the Airbus' high-end systems. These powerful aircraft of...
4	FSX: Steam Edition - Discover Australia and New Zealand Add-On	First Class Simulations — Dovetail Games - Flight — Simulation — About This Content Explore some of the most dramatic scenery on Earth with the latest action-packed mission add-on from First Class Simulations! Australia and New Zealan...
5	FSX Steam Edition: Falcon 7X Add-On	Wilco — Dovetail Games - Flight — Simulation — About This Content The Falcon 7X is the flagship offering of Dassault's business jet line. This fully fly-by-wire long-range trijet is the first bizjet to use fighter jet technology in conjunction with a large ex...

Table C.4: Cluster #4 - Flight simulation games

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	RAGE 2	id Software, Avalanche Studios — Bethesda Softworks — Action, FPS, Open World, Post-apocalyptic, Gore, Singleplayer, Shooter, Violent, First-Person, Blood, Multiplayer, Funny, Exploration, Sci-fi, Adventure, Racing, Cyberpunk, Sandbox, Co-op, Nudity — Dive headfirst into a dystopian world devoid of society, la...
2	Bound By Flame	Spiders — Focus Home Interactive — RPG, Action, Fantasy, Singleplayer, Adventure, Hack and Slash, Third Person, Action RPG, Story Rich, Dark Fantasy, Demons, Atmospheric, Great Soundtrack, Difficult, Masterpiece, Character Customization, Female Protagonist, Open World, Souls-like, Choices Matter — You...
3	Mirror's Edge	DICE — Electronic Arts — Parkour, First-Person, Action, Female Protagonist, Singleplayer, Great Soundtrack, Adventure, Platformer, Atmospheric, Dystopian, Stylized, Futuristic, Sci-fi, FPS, Cyberpunk, Puzzle, Time Attack, Colorful, Masterpiece, Classic — In a city where information is heavily mo...
4	The Evil Within	Tango Gameworks — Bethesda Softworks — Horror, Survival Horror, Psychological Horror, Gore, Atmospheric, Action, Singleplayer, Third Person, Survival, Zombies, Dark, Stealth, Adventure, Third-Person Shooter, Difficult, Cinematic, Shooter, Story Rich, Masterpiece, Walking Simulator — Developed by Shin...
5	Mass Effect	BioWare — Electronic Arts — RPG, Sci-fi, Story Rich, Action, Third-Person Shooter, Space, Singleplayer, Great Soundtrack, Third Person, Choices Matter, Shooter, Female Protagonist, Character Customization, Adventure, Masterpiece, Open World, Atmospheric, Romance, Real-Time with Pause, Action RPG — As Com...

Table C.5: Cluster #5 - AAA Action Games with Strong Narrative Elements

C.2.2 Text+Image Clusters

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	Anomaly 1729	Anvil Drop, LLC — Black Shell Media — Adventure, Indie, Casual, Puzzle — Come explore the world of Phiohm, where up can be down, left can be right, and the way forward is never what it seems. Guide newly cognizant Ano on a journey of self-discovery by learning how to manipulate this e...
2	Grim Nights	Edym Pixels — Edym Pixels — Strategy, Indie, Tower Defense, Pixel Graphics, Survival — A side scrolling, pixel art, survival strategy indie-game. Gather resources, expand your village, explore the underground for riches and train soldiers to defend against hordes of the undead. Features: -...
3	Random Access Murder	Team Murder — Team Murder — Indie, Action, FPS — You're a robot. Murder other robots with your data-destroying weapons and create a stack of dismantled enemies. Be the last heap standing and DO NOT RUN OUT OF MEMORY! RAM is a 1st person arena shooter with a unique look and fast...
4	Trivia Vault: Business Trivia	Ripknot Systems — Ripknot Systems — Casual, Simulation, Action, Family Friendly, 2D, Singleplayer, Indie, Text-Based, Time Management, Choices Matter, Replay Value, Point & Click, 1980s, Word Game, Logic, 1990's, Management, Relaxing, Strategy, Story Rich — Welcome to the Trivia Vault c...
5	Paralysis	Justin Jackson — Justin Jackson — Early Access, Action, Indie, Violent, Early Access, Gore, Casual, Horror, Multiplayer — Paralysis is a first person multiplayer horror game. In which survivors must use teamwork to evade a player controlled creature(s). Will the survivors make it out alive? Or w...

Table C.6: Cluster #1 - Eclectic Indie Games with Varied Mechanics

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	Knight Adventure	Syberstudio — Syberstudio — Action, Adventure, Indie — You are a valiant knight and you must save a beautiful Princess! She is imprisoned by the Evil Sorcerer in The Castle. You'll have to overcome many obstacles and traps to get to her. Yet there's another and more dangerous threat...
2	oldbI tyt ?	Easy game — Game for people — Indie, Action, Strategy — oldbI tyt- clone popular game called 2048. Is simple and surprisingly fun game where you have this 4 by 4 field and numbered tiles on it. Each turn a new tile appears. When two tiles collide they merge into a bigger tile. Player mus...
3	Plandzz	Brzezinski — SIFAKA DIGITAL — Casual, Indie, Puzzle, 2D, Minimalist, Singleplayer, Cute, Great Soundtrack, Atmospheric, Relaxing — Plandzz is a game where you need to solve both simple and challenging puzzles, arrange all the pieces to get a whole block. Interesting puzzles. Pleasant music ...
4	Woody Blox	Frolov Pavel — Frolov Pavel — Indie, Casual — A classic and challenging puzzle. Your goal is to move the blocks so that the main green block you can move to the arrows. Move the green block to the far right and close the arrows. Each new level will be harder and harder. Move the blocks...
5	The Chronicles of Quiver Dick	Crankage Games — Crankage Games — Adventure, Indie, RPG, Memes, Funny, Comedy, Dark Comedy, RPG-Maker — From the creative minds that brought you Metal as Phuk, based on one of the most hilarious characters ever written into an RPG, The Chronicles of Quiver Dick takes you on ...

Table C.7: Cluster #2 - Mixed Small-Scale Indie Games

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	Tsukumogami	TORaIKI — Fruitbat Factory — Indie, RPG, Adventure, JRPG, Anime — Tsukumogami is the original Japanese version of 99 Spirits. 99 Spirits is an RPG/Puzzle game that revolves around the popular Japanese folklore of Tsukumogami, everyday objects coming alive on their 100th birthday. Hanabusa...
2	Princess Evangile W Happiness - Steam Edition	MOONSTONE — MangaGamer — Sexual Content, Adventure, Casual, Visual Novel, Anime, Nudity — It's a new Princess Evangile, with nine new love stories! This time around, you can finally date all nine heroines: the original four heroines in their "epilogues," a...
3	ChronoClock	Purple Software — Sekai Project — Visual Novel, Nudity, Casual, Adventure, Anime, Time Travel, Sexual Content, Story Rich, Cute, Time Manipulation, Multiple Endings, Romance, Choices Matter — Preface Next in line to manage a multi-million dollar corporation is our protagonist, Rei Sawatari. He inh...
4	Umineko: Golden Fantasia	07th Expansion — MangaGamer — Action, Fighting, 2D Fighter, Anime, Indie, Great Soundtrack, Multiplayer, Memes, Psychological Horror — The tag-team fighting game based on the original "Umineko When They Cry" series, finally released on Steam, in English and with improved rollback ...
5	Mutiny!!	Lupiesoft — MangaGamer, Lupiesoft — Sexual Content, Nudity, Visual Novel, Indie, Female Protagonist, Pirates, Anime, Puzzle — Mutiny!! is set in a world of fantasy, and adventure. A realm where airships can sail among the clouds as well as the sea, entering the flogiston to journey to other plan...

Table C.8: Cluster #3 - Japanese Anime-Styled Visual Novels and Games

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	FSX Steam Edition: Airbus Series Vol. 1 Add-On	FeelThere — Dovetail Games - Flight — Simulation — About This Content FeelThere's Airbus Series Vol. 1 is a detailed depiction of the fascinating world of Fly-By-Wire technology as well as the Airbus' high-end systems. Developed by highly qualif...
2	Train Simulator: Marias Pass Route Add-On	Dovetail Games — Dovetail Games - Trains — Simulation, Trains — About This Content This DLC is only available to purchase in the USA. Spectacular scenery and large sweeping curves through the Rocky Mountains are just some of the features of Marias Pass, av...
3	DCS: Mi-8 MTV2 Magnificent Eight	Belsimtek — The Fighter Collection, Eagle Dynamics SA — Simulation, Flight, Free to Play — About This Content DCS: Mi-8MTV2 Magnificent Eight is a highly realistic PC simulation of the Mi-8MTV2, a combat transport and fire support helicopter and an upgraded variant ...
4	FSX: Steam Edition - Embraer E-Jets 175 & 195 Add-On	FeelThere — Dovetail Games - Flight — Simulation — About This Content The Embraer 175/195 aircraft are domestic range airliners with state-of-the-art avionics, fly-by-wire technology, superior cabin comfort and uncompromising performance. These...
5	FSX Steam Edition: Airbus Series Vol. 3 Add-On	FeelThere — Dovetail Games - Flight — Simulation — About This Content FeelThere's Airbus Series Vol. 3 is a detailed depiction of the fascinating world of Fly-By-Wire technology as well as the Airbus' high-end systems. These powerful aircraft of...

Table C.9: Cluster #4 - High-Fidelity Vehicle Simulation Add-Ons

Order	**Game Name**	**Text** (Developer/Publisher/User-tags/Description)
1	FOR HONOR	Ubisoft Montreal, Ubisoft Quebec, Ubisoft Toronto, Blue Byte — Ubisoft — Medieval, Action, Swordplay, Multiplayer, PvP, Third Person, Fighting, War, Co-op, Gore, Singleplayer, Online Co-Op, Hack and Slash, Realistic, Strategy, Atmospheric, RPG, Fantasy, MOBA, Story Rich — Enter the chaos of war as a bo...
2	ARK: Survival Evolved	Studio Wildcard, Instinct Games, Efecto Studios, Virtual Basement LLC — Studio Wildcard — Survival, Open World, Dinosaurs, Multiplayer, Crafting, Building, Adventure, Base Building, Co-op, Action, First-Person, Sandbox, Early Access, Massively Multiplayer, Singleplayer, Dragons, RPG, Sci-fi, MMOR...
3	Rising Storm 2: Vietnam	Antimatter Games, Tripwire Interactive — Tripwire Interactive — FPS, War, Realistic, Multiplayer, Military, Shooter, Tactical, Action, First-Person, Historical, Team-Based, Simulation, Atmospheric, Gore, Cold War, Violent, Massively Multiplayer, Strategy, Singleplayer, Indie — Rising Storm 2: ...
4	Mount & Blade: Warband	TaleWorlds Entertainment — TaleWorlds Entertainment — Medieval, RPG, Open World, Strategy, Sandbox, Action, Multiplayer, Moddable, Military, Adventure, Horses, Realistic, Singleplayer, First-Person, Historical, Third Person, Hack and Slash, Simulation, Fantasy, Indie — In a land torn asunder b...
5	Wizard of Legend	Contingent99 — Contingent99 — Rogue-like, Pixel Graphics, Action, Dungeon Crawler, Adventure, Magic, Local Co-Op, Indie, Rogue-lite, Multiplayer, Hack and Slash, Difficult, 2D, Local Multiplayer, Co-op, Procedural Generation, Singleplayer, RPG, Fast-Paced, Great Soundtrack — Wizard of Legend is a fa...

Table C.10: Cluster #5 - Popular Multiplayer Action Games