UAB
Universitat Autònoma de Barcelona

Essays on Urban Economics

Diego Buitrago-Mora

UAB
Universitat Autònoma de Barcelona

Essays on Urban Economics

by

Diego Buitrago-Mora

Ph.D. Candidate

Doctoral Dissertation

Miquel-Àngel Garcia-López and Anna Matas Prat

Supervisors

September 2023

Acknowledgments

To Myriam, Carol, and Oscar, my family.

I want to express my heartfelt thanks to my supervisors, Anna Matas and Miquel Àngel Garcia-López, for their dedicated support in guiding my development as an economist and researcher. I also want to show my gratitude to Rosella Nicolini, Emilio Padilla, and Javier Asencio, who lead the Ph.D. Program in applied economics and the Department of Applied Economics. Their contributions have been incredibly valuable on my academic journey.

I'd also like to express my gratitude to the entire community of the Department of Applied Economics at the Universitat Autonoma de Barcelona, including colleagues, professors, and the academic and administrative board. Special thanks go to Yadira Gómez-Hernández, David Castells-Quintana, Ronald Miranda, and Imma Ruano for their exceptional support.

In the realm of international collaborations, I had the privilege of working with Professor Jos Van Ommeren from Vrije Universiteit in Amsterdam, whose guidance was fundamental to my third essay, and Professor Araceli Ortega Diaz from Tecnológico de Monterrey. I am equally thankful to Raul Ramos, a colleague and expert whose expertise was instrumental in my third essay.

I would like to extend my appreciation to the jury members, whose expertise and insightful comments played a crucial role in enhancing my research: Hans Koster, Jordi Jofre, Maria Borjesson, Hugo Silva, and Mariona Segú.

My sincere thanks go out to all the participants in the various congresses and workshops where I had the privilege to present my research. I also wish to express my gratitude to the experts who assisted at different points with data, expertise, software, and more.

I owe a debt of gratitude to my non-academic support network, whose unwavering support has been pivotal throughout this journey. Special thanks to Sergio Martinez, Jhonathan Báez, Or Tzur, Sandra Mora, Cristobal Mora, Maria Mora, Guy Whitworth, Ali Al Sakkaf, and Andrea Murillo.

I am also deeply thankful for the financial support received from the Ministerio de Ciencia e Innovación of Spain (research projects RTI2018-097434-B-100), the Ministerio de Ciencia, Tecnología e Innovación of Colombia (convocatoria doctorado en el exterior 860-2019), and COLFUTURO of Colombia (programa credito-beca 2017).

In closing, I'd like to share a quote by Erich Fromm: *Creativity requires having the courage to let go of certainties.*

Contents

Part I

General Introduction

Urbanization, or the rapid concentration of people in urban areas, has undergone significant changes over time, changing the physical, economic, and social makeup of cities worldwide. This phenomenon has lead to numerous scholarly investigations into the complex relationships that shape urban landscapes and their effects on societies. The interaction of employment decentralization, transportation infrastructure, and land use regulations within urban areas can have a significant impact on development patterns, employment opportunities, and quality of life. These topics are thoroughly addressed in three chapters of this research, each of which focuses on a different aspect of urbanization's effects.

This doctoral dissertation presents three empirical research studies that cross-reference transportation economics with the study of urban and regional economies. As each chapter progresses, the shared effort to understand the complex interactions between policies, infrastructure, and spatial configurations in the context of urban and regional dynamics becomes clearer. The research projects extend the fields of urban, regional, and transportation economics, each with its own narrative, to produce a comprehensive investigation of the many different factors that influence cities and regions. Despite the fact that each investigation has a distinct focus, when taken together, they contribute to a better understanding of urban transformation and its numerous consequences.

Like other fields, urban economics is divided into theory and practice. The theoretical study of urban economics deals with understanding the outcomes that result from the interaction of the various agents that inhabit cities and regions, and it began in the 1960s as an independent science of theoretical analysis of urban resource allocation with what is commonly known as the Alonso-Mills-Muth model.

The applied perspective, on the other hand, employs theoretical concepts in a specific context to produce conclusions that can be extrapolated or compared. This branch of urban studies has grown exponentially since the 1970s, thanks to advances in data processing software and increased availability of specialized spatial data.

In this context, the goal of this study is to contribute to the applied study of urban economics while maintaining the discipline's theoretical rigor. For this, I propose three case studies that examine urban economics broadly and how it interacts with other disciplines. The primary goal of this research is to contribute empirically to several fundamental topics in the economic study of cities: urban form, urban land regulation, transportation system expansion, cities in developing countries, real estate prices, environmental sustainability of cities, income and residential segregation, and internal migration.

To address such a broad range of topics, it is necessary to investigate various subjects of study, which necessitates an analysis of various urban contexts. This allows me to demonstrate that the richness of theoretical robustness is precisely its applicability in diverse urban and regional settings. As a result, initially, I considered it critical to include a study close to my initial knowledge, a study subject that would allow me to connect firsthand with the research, as Mills (2000) calls home Bias. It is for this reason that I chose to study my hometown of Bogotá. This research began as a master's thesis, and an advanced version is presented here[1]. Bogotá is an intriguing example of the dynamics of a capital city in a middle-income country, with enough data to conduct a quality study.

Following that, this study focuses on two settings in developed countries, with subjects and settings that enrich the research findings. The first is an examination of urban areas in the United States. This analysis allows me to investigate the quintessential urban context that has inspired many of the most significant outcomes in urban economics, both theoretically and practically. The richness and detail of data in over 270 urban areas makes the United States a one-of-a-kind research opportunity.

Finally, the study concentrated on Europe, specifically one of its most important urban agglomerations, Madrid. This portion of the research allows me to conduct an impact study (as in the case of Bogotá), but in the context of a high-income country where the main challenges are less related to financing or technical capacity. Furthermore, this section focuses on public transportation, particularly trains, which opens up new avenues for search, especially on a continent where they are widely used.

No less important is that most of this investigative process took place in Barcelona, which, in addition to concentrating an important research dynamic in the study of cities, is an inspiring, vibrant city that in many ways captures the reason why research in urban economics is essential: understanding and planning better cities that lead to a better quality of life for their residents. Lastly, it is important to note that this research was nourished by doctoral stays in Amsterdam (The Netherlands) and Monterrey (Mexico), both of which contributed to the international perspective that I intend to provide.

Although urban environments are inspiring and play an important role in this research, I hope to make contributions that go beyond contexts. As a result, I reach conclusions that contribute significantly to fundamental relationships in the study of urban economics. First, I examine the relationship between urban land regulation and real estate prices in developing countries. The reasons for the urban forms are then discussed, as well as their potential consequences, which I examine from the perspectives of environmental, socioeconomic, and economic outcomes. Finally, I investigate whether there is a link between public transportation expansion and population redistribution patterns.

In addition, I use a variety of data processing techniques that are customized to the various study environments. Some of the tools that are used are difference in difference indicators, locally weighted regressions, decentralization indicators, Poisson regressions, radial density of employment, instrumental variables (geological, historical, and Bartik-type), fixed effects, population aerial interpolation, a based-opportunity accessibility indicator, and propensity score matching.

I use a combination of freely accessible administrative data (censuses, road networks, trains, highways, and public transportation stations, location of pollution stations, employment, population characteristics, historical data on weather conditions, historical variables, and so on) and

[1] In July 2023, a peer-reviewed version was published in the journal Regional Science and Urban Economics (RSUE).

exclusive data (microdata on real estate prices; continuous data on soil characteristics, among others) for the databases, which are the main input of this research.

Because the majority of the information used is georeferenced spatial data, I use the most advanced data treatment techniques with the assistance of specialized software. Similarly, the level of analysis varies greatly, ranging from the block level to neighborhoods, districts, census tracts, metropolitan regions, and even average results at the national level.

Overall, the intersection of land use regulations, transportation networks, and urban form has broad implications for urban development, equity, and sustainability. The empirical evidence presented in these chapters emphasizes the complexities of these relationships, which frequently produce heterogeneous effects across different contexts. While the dynamics of developed and developing countries differ, shared lessons emerge. Regulatory frameworks must strike a balance between encouraging growth and maintaining livability. Transportation investments must take into account the long-term impact on urban spatial patterns and quality of life. This research informs policy discussions and emphasizes the need for comprehensive approaches to urban planning and development in a constantly evolving global landscape by comprehensively exploring these dynamics.

As a result, this investigation is divided into three chapters and finishes with general conclusions. In the first I present the study *Real estate prices and urban land regulations: evidence from the Law of Heights in Bogotá*. The second chapter, *Causes and consequences of urban form: evidence from US cities*, delves into the examination of urban areas in the United States. Finally, the third chapter, *Railroad network expansion, opportunity-based accessibility, and population redistribution*, concentrates on Europe, specifically in Madrid. These chapters collectively contribute to a comprehensive understanding of the multifaceted dynamics that shape urban areas and underscore the importance of thoughtful urban planning and development strategies in an ever-changing global landscape.

Part II

Real estate prices and land use regulations: Evidence from the Law of Heights in Bogotá.

Real estate prices and land use regulations: Evidence from the Law of Heights in Bogotá

Diego Buitrago-Mora

Ph.D. Candidate

Miquel-Àngel Garcia-López

Supervisor

Julio 2023

Abstract: Between 2015 and 2017, the Law of Heights (Policy-562) regulated areas of urban renewal in specific locations of Bogotá (Colombia). Using a novel dataset based on detailed information at the block level between 2008 and 2017, we study whether this policy affected real estate prices. Our empirical strategy compares the price per square meter before and after Policy-562 in treated blocks and in control blocks with similar pre-treatment traits. Results show that prices increased more in treated blocks than in the rest of the city. We also provide evidence that results are heterogeneous from a temporal, land use and strata point of view.

Key words: Real estate prices, land regulations.
JEL: R14, R31, R58

1. Introduction

The relationship between land use regulations and real estate prices is well documented in developed countries (Quigley and Rosenthal, 2005, Turner, Haughwout, and van der Klaauw, 2014, Freemark, 2020, Greenaway-McGrevy, Pacheco, and Kade, 2021). In general, empirical evidence centered on housing markets finds that a greater degree of regulation not only increases housing prices (Ihlanfeldt, 2007), but also accelerates their reduction in an economic recession (Huang and Tang, 2012), and the effects vary considerably at the intra-city level (Kok, Monkkonen, and Quigley, 2014).

On the other hand, little is known about this relationship in developing countries. Mayo and Sheppard (1996) compare housing supply regulations in Malaysia, Thailand, and South Korea. Brueckner and Sridhar (2012) find that building height limits caused spatial expansion of Indian cities. Monkkonen (2013) focuses on Indonesia, a country with an important informal housing market, with particularly stringent rules on urban land use, but with a low level of enforcement, and finds that the impact of a greater degree of regulation on formal market prices is unclear. Monkkonen and Ronconi (2013) finds a negative relationship between regulation and land prices in the three major Argentinian metropolitan areas with higher levels of regulation and lower levels of compliance. For the case of Beijing (China), Ling, Dao-lin, and Ke-lin (2013) find that land control policies accelerated housing prices when they were implemented. Finally, Brueckner, Fu, Gu, and Zhang (2017) find that building height restrictions in terms of floor area ratio increases land prices in Chinese cities.

This paper aims to contribute to this literature by studying the impact of a particular regulation, the so-called Law of Heights (Policy-562), on real estate prices in a city of a developing country, Bogotá (Colombia), between 2008 and 2017. Using annual data for 837,505 registered lots[2] grouped in 42,993 blocks, we rely on an empirical strategy based on Difference-in-Differences (DiD) techniques to compare real estate prices before and after the implementation of Policy-562 in treated blocks and in control blocks with similar pre-treatment traits. Besides the average effects, we also explore the heterogeneity of the effects by year of the treatment subperiod, the main land uses of the blocks, and the strata where they are located[3]

There are various reasons why Bogotá and its Law of Heights (Policy-562) provide an excellent testing ground of the relationship between regulation and real estate prices. First, the new policy aimed to regulate the conditions for urban renewal not in the whole city, but only in some specific areas. As a result, it is possible to identify treated blocks. Second, the treatment period of this policy is also easy to identify: It was in force between 2015 and February 2016, but new projects were still approved and executed between March 2016 and December 2017. Third, the Law of Heights increased the degree of land use regulation in Bogotá because, despite relaxing the height limits for the new buildings (which required a monetary compensation), the new set of

[2]In this research, we interpret a *lot* as an area of land with one or more owners that may contain one or several *properties*.

[3]The strata system in Colombia is a system of subsidizing public services by regulating their prices. Every block in Colombia's urban areas, including Bogotá, are assigned to an specific strata level, from 1 to 6, depending on its physical characteristics and surrounding conditions. Strata 1 to 3 receive a subsidy from the higher strata 4 to 6. See Appendix A. for a further definition of the strata system in Bogotá and its relationship with Policy-562.

rules clearly increased construction costs. Finally, detailed data at the lot level is available for the 2008 to 2017 period.

In general, this paper furthers our understanding of the effects of land use regulations. The related empirical literature shows that they limit city size (Hannah, Kim, and Mills, 1993), increase real estate prices (Quigley and Rosenthal, 2005, Ihlanfeldt, 2007, Huang and Tang, 2012), follow the market (Wallace, 1988, Garcia-López, Solé-Ollé, and Viladecans-Marsal, 2015), and, in general, affect many other aspects of development (Cheshire and Sheppard, 2004). Furthermore, regulations seems to negatively affect welfare (Turner *et al.*, 2014). As above mentioned, most of the literature has focused on developed countries, and only few recent works has analyzed other countries with inconclusive and, sometimes, opposite results. This paper contributes to this literature by providing empirical evidence for a particular regulation in a city of a middle-income developing country.

Our results show that, on average, Policy-562 positively affected real estate prices. In particular, our pure DiD approach reports an estimated effect of 33.5% in treated blocks. This result holds when we consider more balanced samples of treated and untreated blocks in terms of observables by combining DiD with Propensity Score and Nearest Neighbor Matching techniques. When we follow Brueckner *et al.* (2017) matched-pair approach to consider balanced samples in terms of unobservables, we estimate a Policy-562 effect of 16.4%. Finally, in Appendix D. we show that the effect of Policy-562 is heterogeneous in three dimensions. By year, the effect decreased during the treatment subperiod. By main land uses of blocks, Policy-562 only affected Residential and Services prices. By strata, while Policy-562 increased prices in low strata 1 and 3 and high strata 6 treated blocks, it decreased prices in high strata 4 and 5.

The rest of the paper is structured as follows. In Section 2, we briefly describe land use regulation in Colombia and in Bogotá, with an especial attention to the Law of Heights. In Section 3, we present the city of Bogotá, the dataset to study real estate prices at the block level, and the procedure to identify the blocks (un)affected by Policy-562. The empirical strategy based on Difference-in-Differences techniques is discussed in Section 4. Section 5 presents the main results and robustness checks, and Section 6 concludes.

2. The Law of Heights (Policy-562)

Colombia has a national land use regulatory framework that can be considered strong in the Latin American region[4]. Law 388 of 1997 exemplifies this. This Law enshrines how to use urban land and grants cities with more than 100,000 inhabitants the freedom to draft their master zoning plan or Plan de Ordenamiento Territorial (POT). According to Cámara de Comercio de Bogotá (2018), a POT comprises a set of goals, guidelines, policies, strategies, programs, actions, and norms aimed at directing and managing the physical development and land use in the territory.

[4]According to Cabeza (2006), Latin American countries can be classified according to their level of land use regulation. First, countries with specific (centralized) national laws on land use planning (Uruguay, Colombia, Salvador, Honduras, and Cuba). Second, countries with several (decentralized and non-coordinated) regional laws (Argentina, Bolivia, Ecuador, Venezuela, and Mexico). Finally, coungries without land use regulation laws (Chile, Paraguay, Brazil, Panama, Costa Rica, and the Dominican Republic).

Thus, the zoning plan constitutes a road map for the long-term (12-year) development of urban and rural areas to consolidate a 'coherent' city model.

The first POT of Bogotá was approved in 2004 and it classified the territory according to three structures: (i) a main ecological structure, (ii) a functional and services structure, and (iii) a socioeconomic-spatial structure. More specifically, the 2004 POT regulated height limits, floor area ratio and developer payments that affected all areas of the city indistinctly. The norm remained in force until December 30th 2021, when a new POT (Law 555 of 2021) was approved for the 2022-2035 period.

In December 2014, Bogotá implemented a new policy (562 of 2014) regulating the conditions for urban renewal in defined city areas. The policy aimed to promote the improvement, beautification, development and, in particular, densification of some specific parts of the city with public and private interventions. Unfortunately, there is no technical document justifying the selection of the areas (see Figure 2a). It seems that they were close to public transportation (Transmilenio) and main roads, to metropolitan and zonal parks, to facilities (public safety, defense and justice, food supply and consumption, hospitals, fairgrounds, cemeteries and public administration services), and they were not protected (not developable land). However, it is also true that other areas satisfied the above mentioned characteristics and were not selected (for example, areas in the south of the city with many illegal settlements).

To achieve these goals, Policy-562 first removed height limits on new buildings conditional on some payments from the developers. In general, these payments in Colombia refer to the amount of area (A) that developers must give to the city. This land comes from the lots to be developed and it is used to satisfy the 'needs' of the surrounding area in terms of public space, road infrastructure, parking lots, front gardens, or public services, among others. It is calculated as follow:

$$A = P \times K$$

where P is the total lot area, and K is the payment factor.

Secondly, Policy-562 modified developer payments (A) by updating the value of K. Under the 2004 POT, K had a unique value of 0.20. Under Policy-562, the value of K depended on the floor area ratio (FAR, the ratio between a building's total floor area and the total lot area).

Table 1 reports K values for different floor area ratio intervals: The higher the FAR, the higher the K. It also shows that developer payments were lower under Policy-562 when the floor area ratio was bellow 4. On the contrary, Policy-562 payments were higher than 2004 POT ones for higher floor area ratios.

Using an example discussed by Ruiz and Moncada (2017), in Table 2 we compute developer payments under the 2004 POT and Policy-562 for a residential project with 100 m^2 apartments in a lot of 8,694 m^2 (138 m $\times$ 63 m). First, to build 100 apartments (Columns 1 and 2) a developer would have to give to the city 1,739 m^2 of the lot area under the 2004 POT, but only 52 m^2 under Policy-562. Second, developer payments would be roughly the same with the two policies when building 310 apartments (Columns 3 and 4). Third, to build 433 apartments[5] (Columns 5 and 6), developer payments under Policy-562 would be 120% higher (3,817 m^2 vs. 1739 m^2).

[5]Because of the 2004 POT height limits (10-story buildings), computations in Column 5 are hypothetical.

Finally, if we consider the maximum number of floors that could be built according the 2004 POT (10) and the maximum number of apartments per floor[6] according to each policy (31 and 43), developer payments by apartment would increase by 58% (from 5.61 (=1739/310) (Column 3) to 8.88 (=3817/430) m^2 per apartment (Column 6)).

It is important to clarify that, under this policy, developer payments could be also monetary. That is, if the amount of land (A) that was to be given to the city was not available in the area (or was less than 2,000 m^2), the developer could make a monetary payment (based on cadastral values) that the city would use for infrastructures and urban amenities in other areas.

Table 1: Policy-562 K values to compute developer payments in Bogotá

Floor Area Ratio	K
$2.0 < FAR \leq 2.4$	0.006
$2.4 < FAR \leq 2.8$	0.035
$2.8 < FAR \leq 3.3$	0.092
$3.3 < FAR \leq 4.0$	0.197
$4.0 < FAR \leq 4.4$	0.322
$4.4 < FAR \leq 5.0$	0.439
$5.0 < FAR \leq 6.5$	0.553
$6.5 < FAR \leq 9.0$	0.655
$9.0 < FAR \leq 14$	0.757
$FAR > 14$	0.833

Table 2: Developer payments in a residential project: 2004 POT vs. Policy-562

	200 Apartments		310 Apartments		433 Apartments	
	2004 POT [1]	Policy-562 [2]	2004 POT [3]	Policy-562 [4]	2004 POT [5]	Policy-562 [6]
Total lot area (m^2)	8,694	8,694	8,694	8,694	8,694	8,694
Number of floors	6	5	10	7	14	10
Total floor area (P) (m^2)		20,000		31,000		43,000
Floor area ratio (FAR)		2.30		3.57		4.98
Payment factor (K)	0.20	0.006	0.20	0.197	0.20	0.439
Developer payment (A) (m^2)	1738.80	52.16	1738.80	1712.72	1738.80	3816.67

In February 2016, Policy-562 was repealed, among other reasons, because its approval was considered illegal. By that date, 901 projects were approved, and 2,362 applications had been filled while the new policy was in force. Between March 2016 and December 2017, most applications were approved and executed. The 2016 Resolution 079 revoked Policy-562. The cancellation of the decree meant that Policy-562 had no effect on newly issued construction licenses as of February 22, 2016. However, any license requested prior to February 21, 2016, if authorized, was governed

[6] This number depends on other requirements of the policies (e.g., the land use index) and explains why the number of floors is different for the two policies in the three studied scenarios in Table 2.

by Policy-562. Similarly, all projects approved and under construction with Policy-562 continued to adhere to this policy even after the repeal declaration and until project completion.[7].

Policy-562 was also important for the city budget. According Secretaria de Hacienda de Bogotá (2015), 200,000 million COP (US$ 50 million) in developer payments were raised in 2015, representing 20% and 2% of non-tax revenues and total revenues, respectively. Compared to 2004 POT payments between 2005 and 2014, Policy-562 raised 50% of them in just 15 months (Cámara de Comercio, 2015).

Finally, it is important to mention that Colombia and, in particular, Bogotá have an active law enforcement system with a low percentage of informality and a reasonable time to approve building permits. On average, 12,000 building permits are issued every year in Bogotá. Each permit is issued in an average of 50 calendar days. Secretaría Distrital de Gobierno (SDG) is responsible of the related law enforcement according to article 135 of the National Police Code. On average, 900 stop-workers orders are issued every year: 62% of them for not having any type of building permit, 30% due to breach obligations related to the construction process itself, 7% for allocating a property to a use other than that authorized in the building permit, and 1% to protect properties of cultural, historical and architectural interest. This scenario differs from other developing countries like Indonesia, with restrictive land registration and building permits (160 days), and inefficient law enforcement (Monkkonen, 2013).

3. Data

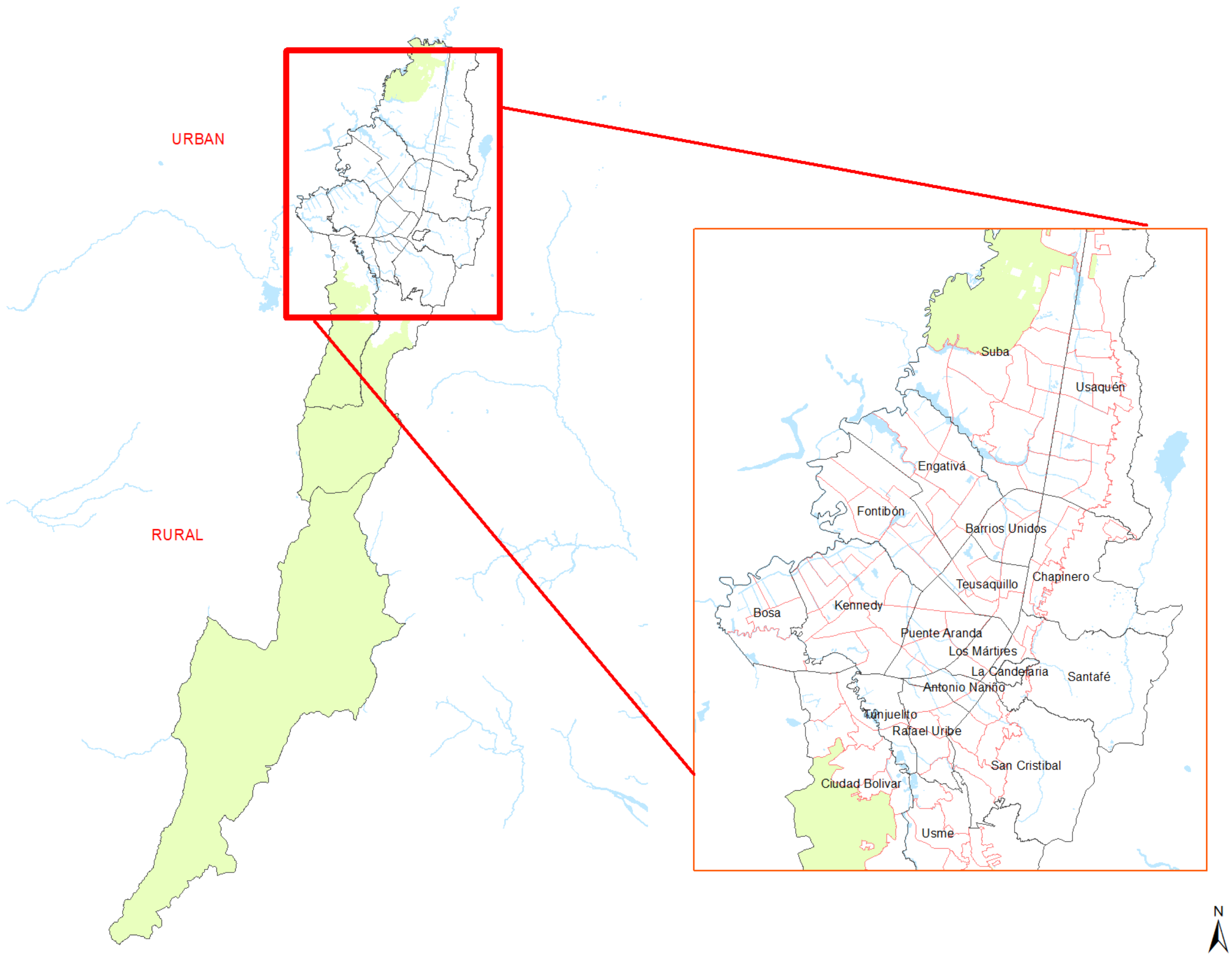
3.1 Bogotá (Colombia)

We study the metropolitan area of Bogotá, with 10,121,956 inhabitants in 2021 according to Departamento Administrativo Nacional de Estadística (DANE) living in 4,000 km^2, that is, with roughly 2,530 inhabitants per km^2.

Figure 1 shows the urban and rural areas of metropolitan Bogotá. As can be noticed, two-thirds of the city is rural (in green). We focus the analysis on the urban areas, which includes 19 municipalities (black lines). After the city, the municipality is the largest level of zoning. For planning purposes, the city is also divided into 108 zonal planning units (ZPU) (red lines) and their 1090 neighborhoods.

[7]The repeal decree literally says:"... If, during the term that elapses between the application for a license or its modification and the issuance of the administrative act that grants the license or authorizes the modification, there is a change in the urban regulations that affect the project submitted ... the applicant will have the right to have the license or modification granted based on the urban planning regulations in force at the time of the filing of the application, provided that it has been submitted legally and duly ..."

Figure 1: Urban and rural areas in Bogotá



3.2 *Real estate prices*

To measure real estate prices, we use the dataset developed by Secretaría Distrital de Planeamiento (SDP). It is based on annual studies of the real estate market monitoring the trends in the commercial value of properties. Opposed to the traditional cadastral values, these SDP values contain real estate market elements such as sales offers, leases and financial transactions, and appraisals[8]. SPD prices represent the commercial reference values (per m^2) and reflect the dynamics of the real estate market[9].

The SPD dataset also includes information about the floor area (m^2) and the predominant land use of the lots (residential, manufacturing and services). Unfortunately, no other property

[8]The annual appraisal process is carried out by the cadastral unit (Unidad Administrativa Especial de Catastro Distrital, UAECD), an autonomous entity belonging to the Bogotá finance office and independent from SPD. Appraisals are processes that reflect the characteristics of homogeneous geographical and economic zones to determine the current value of properties. New projects and development plans only affect these values once the properties are physically changed. In other words, SDP prices do not respond to regulatory changes via appraisals that happened at the same time that the norm changed. On the contrary, SPD prices adjust in the medium and long term.

[9]As a robustness check, we compared the SPD dataset with the best available alternative dataset (Coordenada Urbana developed by Cámara Colombiana de Construcción CAMACOL), which includes average transaction prices at the neighbourhood level. Both datasets are highly correlated and a simple test for difference of the means shows that they are not statistically different. Unfortunately, we did not have access to individual transaction prices.

characteristic (e.g., height) is included in the dataset.

Our initial sample includes data for 837,505 registered lots in 2017. They represent 88% of registered lots[10]. To avoid inconsistencies due to missing values in previous years[11], we fix these 2017 lots for the whole studied period. By doing so, we avoid inconsistencies due to missing values in the previous years. Then, we group lot data into blocks and we end up with 42,993 blocks. The real estate price at the block level is then computed as the average price (per m^2) of the lots that make up each block.

3.3 *Areas (un)affected by Policy-562*

As we explain in more detail in the next sections, we study the impact of Policy-562 on real estate prices with a before–after analysis that compares the evolution of prices in treated areas (affected by Policy-562) and untreated areas (unaffected by Policy-562).

The identification of the affected areas of the city is challenging because, first, this information is not at the same spatial level of aggregation as that of real estate prices (block level), and, second, we do not have a map of the blocks (we only know their municipalities, ZPUs and neighborhoods). In fact, all we can resort to are documents and paper maps of the city in which the areas affected by Policy–562 are presented schematically and without precise geographic detail. For example, Figure 2a is a paper map published by the planning authority identifying the ZPUs of the city affected by Policy-562 (in yellow), non-affected (in white), and under special protection (in red). It is important to notice that not all blocks that make up each ZPU were affected by Policy-562.

To identify whether or not each of the 42,993 defined blocks are affected by the Policy-562, we follow a top-down analysis, i.e. from the largest level of aggregation to the smallest one, in order to obtain a dummy variable that takes a value of 1 for areas included under Policy–562 and 0 otherwise.

We begin by identifying with zero the blocks located in ZPUs of municipalities without areas designated under Policy–562. Then, we use a lower level of aggregation, the ZPUs, and assign a value of 1 to blocks located in ZPUs with more than 75% of their total area affected by Policy–562. For ZPUs with less than 75% of affected area, we use an smaller spatial unit, the neighborhood, and repeated the exercise: We assign a value of 1 to blocks located in neighborhoods with more than 75% of their total area affected by Policy–562.

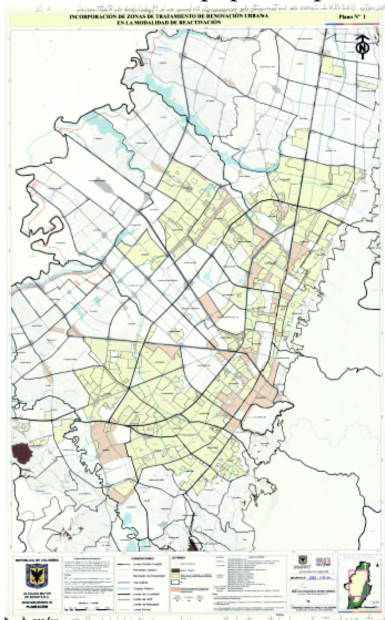
At the end of this procedure, we identify 7,700 blocks affected by Policy-562 (18% of blocks) (the blue areas in Figure 2b) and 35,293 unaffected blocks (the yellow areas in Figure 2c). The former are our (initial) treatment group and the latter our (initial) control group.

[10] According to the 2017 cadastral census, there were 2,543,290 properties in 951,749 registered lots.

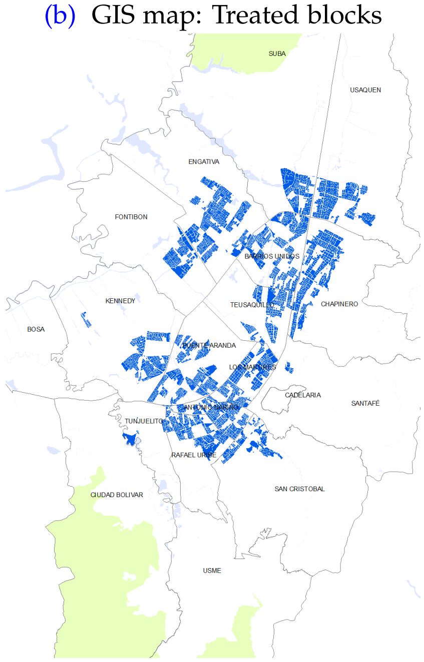
[11] For example, when new lots are added to the city boundaries, or when lots are excluded because they are merged due to the construction of new buildings.

Figure 2: From a paper map to GIS maps of (un)affected blocks by Policy-562

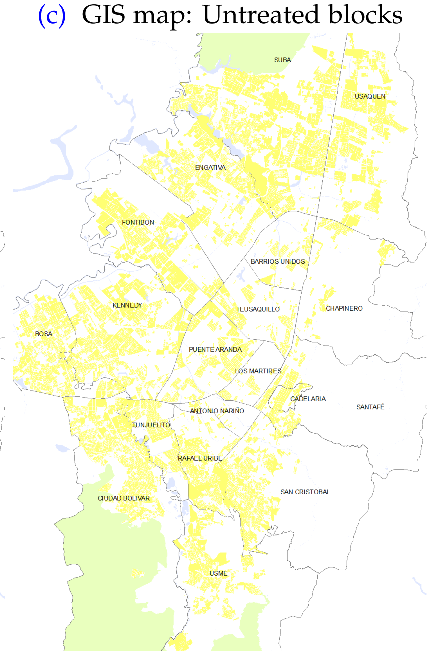
(a) Official paper map



(b) GIS map: Treated blocks



(c) GIS map: Untreated blocks



Notes: In Figure 2a, yellow and white zones are ZPUs affected and unaffected by Policy-562, respectively. Red zones are ZPUs under special protection. In Figures 2b and 2c, blue and yellow zones are blocks in areas affected and unaffected by Policy-562, respectively. In both figures, gray lines are municipality boundaries.

4. Empirical strategy

4.1 *Timing of the analysis*

Using the SPD dataset, we have information on real estate prices from 2008 to 2017. We split this period into two subperiods. First, the treatment subperiod (2015–2017) considers the years in which Policy-562 was in effect (2015 and February 2016) and the years in which the last projects approved by Policy-562 were developed (March 2016 and 2017). Second, the subperiod 2008–2014 is the period before treatment.

4.2 Estimated equation

We estimate the effect of the Law of Heights (Policy-562) on real estate prices using a Difference-in-Differences (DiD) strategy. In particular, with our 10 year dataset, we estimate the following equation:

$$\begin{aligned}\ln(\text{Price}_{it}) = \beta_0 + \beta_1 \times \text{Policy-562}_i \times \text{After-562}_t \\ + \beta_2 \times \text{Time-variant controls}_{it} + \beta_3 \times \text{Time-invariant controls}_i + v_t + \epsilon_{it}\end{aligned} \tag{1}$$

where $\ln(\text{Price}_{it})$ is the log of the average property price in block i in year t.

Policy-562_i is a dummy equal to one if block i is affected by the new policy, and zero otherwise. After-562_t is a dummy equal to one if year t corresponds to the period of implementation of the Law of Heights (2015–2017), and zero otherwise. We are interested on the DiD estimator, that is, on the estimated value of β_1, the coefficient of the interaction between Policy-562_i and After-562_t. It measures the effect of the new policy in treated vs. untreated (control) areas.

$\text{Time-variant controls}_{it}$ is a vector of time-variant block and ZPU characteristics. First, we control for the log of the average floor area (m^2) in the block. Second, to control for socioeconomic characteristics, we add the log of the number of inhabitants per hectare (density) and the log of population per household. Summary statistics are reported in Table 7 of Appendix B..

$\text{Time-invariant controls}_i$ is a vector of time-invariant ZPU characteristics. First, we control for time-invariant socioeconomic characteristics with dummy variables for each of the five strata. Second, we add controls for the accessibility to the city's main services such as the log of km^2 of metropolitan parks, the log of km^2 of zonal parks, the number of health–related private institutions (small and medium), and the number of facilities (public safety, defense and justice, food supply and consumption, hospitals, fairgrounds, cemeteries and public administration services). These variables are from 2017. In the same group, we added the number of Transmilenio stations, the system of Bus Rapid Transit (BRT) responsible for the majority of public transport trips in the city[12].

Finally, v_t are year fixed-effects, and ϵ_{it} is an error term with the usual properties.

In our preferred specification we replace the time-invariant controls with block fixed-effects (α_i) that fully control for all time invariant differences between blocks:

$$\ln(\text{Price}_{it}) = \beta_1 \times \text{Policy-562}_i \times \text{After-562}_t + \beta_2 \times \text{Time-variant controls}_{it} + v_t + \alpha_i + u_{it} \tag{2}$$

4.3 On the parallel trends assumption

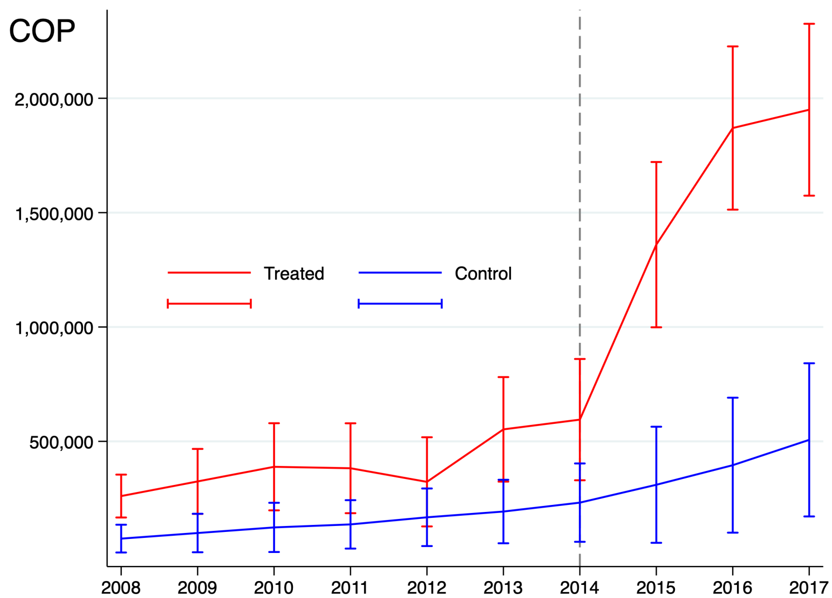
To use the DiD strategy, we assume parallel trends, which implies that the time effects (v_t) take account of any time trend in the data that is common to both the treatment and control groups

[12] We include this variable as time-invariant using most recent values because there was no new construction of lines or stations between 2013 and 2020. The last one before such a pause was the enlargement to connect Soacha (the neighbouring municipality in the south of Bogotá) in 2013.

(Jones, 2009). The presence of this common trend prior to the implementation of Policy–562 means that the behavior of the two groups should be homogeneous and independent of the future impact that will affect the treated group. Several authors stress the importance of studying this assumption by comparing the observable characteristics of the treated and control groups (Zhang, 2017, Givord, Quantin, and Trevien, 2018) which, in this case, means verifying if there is a systematic difference in the behavior of the real estate prices prior to the introduction of Policy–562.

Figure 3 shows the evolution of the average prices in treated and control groups between 2008 and 2017. It shows that, before the Law of Heights (2008-2014), real estate prices of the two groups evolved in a similar way and, in fact, they were not statistically different. These parallel pre-trends are suggestive evidence in support of the parallel trends assumption. On the other hand, it is clear that the average prices of the two groups followed different trends when Policy-562 was in place (2015-2017).

Figure 3: Evolution of real estate prices in treated and control groups: Mean and S.D.



Notes: 7,700 treated blocks and 35,293 untreated (control) blocks as described in Section 3.3.

5. Results

5.1 *Main results*

Table 3 reports DiD results when we regress the log of price on the interacted Policy-562 variable. In Column 1, we follow a pooled strategy and estimate Equation (1) without control variables. Then, we gradually add time-variant (Column 2) and time-invariant (Column 3) controls. Column 4 shows results when we follow a block fixed-effects panel strategy and estimate Equation (2). Since our dependent variable is based on the average price of the lots that make up each block, we weight block-year observations by the number of lots-year.

The estimated coefficient of interest is positive and statistically significant in all columns and decreases when we add control variables and, in particular, when we control for block fixed-effects. Our preferred result is in Column 4, it reports an estimated coefficient of 0.289 indicating that blocks affected by Policy-562 experienced an increase in real estate prices around 33.5% higher than untreated blocks.

Table 6 in Appendix B. shows that average prices of treated blocks increased from 592,000 to 1,942,000 COP/sq.m. between 2014 and 2017, which represents a total growth of 228.1% in the treatment period. As a result, the Law of Heights explains roughly 15% of this growth. Similarly, if we consider that average prices of untreated blocks increased by 120.7% (from 241,000 to 533,000 COP/sq.m.), Policy-562 would explain a third of the difference in growth rates between treated and untreated blocks.

Table 3: The effect of Policy-562 (Law of Heights) on real estate prices: DiD main results

	[1]	[2]	[3]	[4]
Policy-562×After-562	1.130[a]	1.168[a]	0.850[a]	0.289[a]
	(0.032)	(0.033)	(0.032)	(0.037)
Time-variant controls		✓	✓	✓
Time-invariant controls			✓	
Block fixed-effects				✓
Time fixed-effects	✓	✓	✓	✓
Adjusted R^2	0.100	0.133	0.139	0.217

Notes: 429,930 observations (= 42,993 blocks × 10 years) in each regression. Regressions are weighted by the number of lots that make up each block. Robust standard errors are clustered by ZPU and are in parenthesis. The coefficient of interest remains significant when clustering at the neighborhood and block levels. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively.

Since Bogotá's real estate market is not perfectly segmented by block, in Appendix C. we consider potential spillover effects when prices in one block are affected by prices (or their determinants) in nearby blocks. In particular, in Table 8 we add controls for the log of average price per square meter in neighbouring blocks located at different distances and ZPUs. The estimated coefficient of interest remains positive and statistically significant in all specifications. Furthermore, these results are not statistically different from our preferred specification in Column 4 of Table 3.

In Appendix D. we investigate the heterogeneity of the above results. First, we study whether the effect of Policy-562 changed over time during the treatment period. Results in Column 1 of Table 9 shows that the positive effect of this policy on prices decreased every year (from 2015 to 2017). We relate this decreasing effect with the political context of Bogotá during these years and, in particular, the announcement and effective repeal of the Law.

Second, we also explore heterogeneous effects related to the main land use of the blocks. Results in Column 2 indicates that the Law of Heights only affected Residential and Services treated blocks. On the contrary, Manufacturing prices were not significantly affected.

Finally, we consider the strata where blocks are located. Results in Column 3 confirm heterogeneous effects of Policy-562 at the strata level. While prices in low strata 1 and 3 and high strata 6 treated blocks were positively affected, prices in high strata 4 and 5 zones were negatively affected by the Law.

5.2 *Robustness checks*

Despite the parallel pre-trends reported in Figure 3, we fear that treated and control groups might be different in terms of observables. To alleviate this concern, we consider three alternative methods that aim to redefine our treated and control groups. First, we apply a Propensity Score Matching (PSM) to select treated and controls that are similar in terms of explanatory variables[13]. We end up with 34,449 blocks (80% of the initial sample). The treated and controls groups are made up of 6,186 and 28,263 blocks, respectively. Alternatively, we consider a Nearest Neighbor Matching (NNM) using the 100-nearest neighbors on all explanatory variables[14]. With this method, we select a total of 6,177 blocks, 3,818 treated and 2,359 untreated. Finally, we follow Brueckner *et al.* (2017) matched-pair approach and consider what we name the Geographical Approach (GA): We focus on the control group to select those untreated blocks that are adjacent to treated blocks. The idea is that, at this spatial level, adjacent blocks may only differ on the treatment. In this case, we end up with a total of 13,546 blocks, that is, the original 7,700 treated blocks and 5,846 untreated blocks (16.7% of the initial untreated sample).

Table 4 reports results when we combine the DiD approach with the PSM (Column 1), the NNM (Column 2) and the GA (Column 3). As previously, the estimated coefficient of interest is positive and statistically significant in all three alternative approaches.

Regarding the magnitude of the estimated coefficients, the PSM and NNM ones (0.296 and 0.324) are statistically similar to their pure DiD counterpart (0.289) in Column 4 of Table 3. They show that Policy-562 increased prices by 34.5% and 38.3%, respectively.

On the other hand, the GA estimated coefficient (0.152) is statistically smaller and differs by a factor of 2 with the pure DiD estimated coefficient (0.289) in Column 4 of Table 3. This GA result indicates that Policy-562 (only) caused a 16.4% growth in real estate prices in treated blocks[15].

Table 4: The effect of Policy-562 (Law of Heights) on real estate prices: Alternative methods

	PSM + DiD	NNM + DiD	GA + DiD
	[1]	[2]	[3]
Policy-562×After-562	0.296^a (0.040)	0.324^a (0.083)	0.152^a (0.050)
Adjusted R^2	0.214	0.207	0.203
Observations	344,490	61,770	135,460

Notes: Regressions include time-variant controls, block fixed-effects, and year fixed-effects. They are also weighted by the number of lots that make up each block. Robust standard errors are clustered by ZPU and are in parenthesis. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

[13] In Appendix E. we provide further details on the method and its implementation.

[14] Unfortunately, smaller 'neighborhoods' do not provide enough number of observations. On the contrary, bigger 'neighborhoods' do not significantly change the number of observations and results hold.

[15] In some additional robustness checks that are available upon request, we apply the geographical approach (GA) to the PSM and the NNM samples. In both cases, results hold with significant and smaller estimated coefficients.

We may also fear that the cutoff used in the definition of blocks affected by Policy-562 is somehow arbitrary. As we explain in detail in Section 3, treated blocks are those located in ZPUs with more than 75% of their total area affected by Policy-562. For ZPUs with less than 75% of affected area, we apply this threshold to each of their neighborhoods.

In Table 5 we explore the sensitivity of the results to the chosen cutoff. First, we consider an smaller cutoff of 25% in Column 1 and a more demanding cutoff of 100% in Column 2. Using these alternative thresholds, the number of treated blocks increases from 7,700 to 10,488 (25% threshold) and decreases to 3,075 (100% threshold). The results of estimating Equation (2) confirm the positive and significant effect of Policy-562 for the two thresholds. Furthermore, when comparing with the result counterpart in Table 3 Column 4 (75% threshold), it is clear that the estimated positive effect increases the higher the threshold.

Second, in Column 3 we consider a multilevel treatment by simultaneously using different threshold intervals: Blocks with 25% to less than 75% of affected area, blocks with 75% to less than 100% of affected area, and blocks with 100% of affected area. The omitted category refers to blocks with less than 25% of affected area. The estimated coefficients confirm the positive effect of Policy-562, which is more important for the most affected blocks (100%).

Finally, in Column 4 we consider a continuous treatment variable by directly using the percentage of affected area (instead of a dummy). The significant and positive estimated coefficient indicates that each additional 1 p.p. in the percentage of affected area, increased real estate prices by 0.52%. In other words, blocks with a 100% of affected area experienced a 52% increase in their real estate prices.

Overall, these alternative threshold results confirm results when using the 75% threshold.

Table 5: The effect of Policy-562 (Law of Heights) on real estate prices: Alternative measures

	Thresholds		Intervals	Continuous
	25%≥ [1]	100% [2]	[3]	[4]
Policy-562×After-562	0.234[a] (0.033)	0.538[a] (0.050)		
25-75% Policy-562×After-562			0.060[a] (0.010)	
75-100% Policy-562×After-562			0.119[a] (0.008)	
100% Policy-562×After-562			0.560[a] (0.009)	
Continuous Policy-562×After-562				0.520[a] (0.044)
Adjusted R^2	0.217	0.217	0.217	0.217

Notes: 429,930 observations (= 42,993 blocks × 10 years) in each regression. Regressions include time-variant controls, block fixed-effects, and year fixed-effects. They are also weighted by the number of lots that make up each block. Robust standard errors are clustered by ZPU and are in parenthesis. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively.

In summary, results in Tables 3, 4 and, 5 confirm that Policy-562 affected real estate prices in Bogotá. In particular, while the pure DiD specification in Column 4 of Table 3 shows that prices increased 33.5% more in treated blocks, the GA specification in Column 3 of Table 4 reports an effect of 16.4%.

A qualifier is important here. There are some identification issues that might affect the magnitude of the estimated coefficient. In this sense, our research faces an endogeneity problem. First, we are worried that some unobserved variable determines both real estate prices and Policy-562. The DiD, the PSM, and the NNM approaches are elaborate ways of comparing blocks that are similar on observable quantities. By comparing near neighbors, the GA approach may do better at controlling for unobservables. Second, as shown in Table 5, we also face a measurement error in our measure of Policy-562. Therefore, the magnitude of the positive effect estimated in the DiD, the PSM, the NNM and the GA specifications should be read with caution.

6. Conclusions

In this paper, we investigate the effect of the Law of Heights (Policy-562) on real estate prices in Bogotá between 2008 and 2017. Our results show that treated blocks experienced an increase in real estate prices. On average, the effect of Policy-562 ranges between 16.4% (GA approach) and 33.5% (pure DiD approach). This effect is also heterogeneous from a temporal, land use and strata point of view: It decreases in time, it is only related to Residential and Services land uses, and it is positive in low strata 1 and 3 and high strata 6 and negative in high strata 4 and 5.

We think that the contributions made by this paper are relevant. First, it provides empirical evidence for a city (Bogotá) in a middle-income developing country (Colombia) and shows that, similar to developed countries, a greater degree of regulation increases real estate prices. Second, while most papers focus on the average effects of the regulation, this research also provides empirical evidence on its heterogeneous effects. In particular, the paper furthers our understanding of how regulation affects different land uses and income groups.

A qualifier is important here. As we previously acknowledge, our research faces an unsolved endogeneity problem related to unobserved variables and potential measurement errors. As a result, the magnitudes of the estimated coefficients in our preferred specification should be read with caution.

Appendix A.

A. *Strata system in Bogotá*

Definition

The strata system in Colombia regulates prices of utility services such as water, sewer, electricity, gas, and telephone. According to the 1991 Constitution, these services are basic. At the national level, the strata system is regulated by law 142 of 1994.

In Bogotá, the system divides the city into six strata. Residents in strata 5 and 6 pay utility services according to their consumption plus an additional 20%, which is used to partially pay consumption bills by residents in lower strata. Specifically, strata 1, 2 and 3 residents receive a discount of 50%, 40% and 15%, respectively, on the utility prices. Finally, residents in strata 4 pay utility prices without additional charges or subsidies.

The classification in the different strata depends on the structural differences between the areas of the city, mainly in the housing characteristics/conditions and urban amenities. As a result, the strata system is not directly related to income, but only indirectly through these characteristics.

Policy-562 and strata system

Changes in the strata classification of a block are related to two main facts. The first is the change in the physical characteristics of the houses in a block, and the second is related to the improvements of urban amenities (such as new public transportation, the reconstruction of sidewalks and public places, and improvements in the sewage system, among others) that directly improve conditions around the block.

Regarding the first point, it is important to note that being part of the affected areas of Policy-562 did not directly affect block strata. It could be the case, however, that in the period when Policy-562 was in force (less than two years) plus the period of execution of the projects, some projects generated sufficient conditions to change the strata of a block. Nevertheless, it was neither a massive nor an automatic process. Thus, changes were slower than the window time we are analyzing in the present study.

Regarding the second reason that generated the change of strata in Bogotá (improvements of urban amenities around a block), Policy-562 did not regulate the construction of large-scale public infrastructure works. In addition, it did allow the developer to make a monetary payment instead of the construction of urban amenities in the area of direct influence of the new building (see table 1). Furthermore, since the objective was to regulate the constructions in urban renewal areas, most of the new buildings developed were built in consolidated areas, so developers often opted for the payment of the monetary compensation to the city (see section 2), which means that there was no improvement of urban amenities in the surrounding areas.

Finally, the process of changing strata is constant since there is a continuous committee evaluating every case. However, the legal formalization of the strata is made official by decree law. Since the strata system exists, Bogotá generated decree laws in 1997, 2002, 2004, 2008, 2009, 2013, 2017 and 2019. When analyzing the strata change by blocks between 2008 and 2017, we observe that only 4% (1,692 blocks) presented changes. Of these changes, 3.7% were from low strata (1, 2 and 3), and 0.3% were from high strata (4, 5 and 6). Of the blocks included in the areas of influence of Policy-562, less than 2% changed strata between 2008 and 2017.

B. Summary statistics

Table 6: Average (and standard deviation) of real estate prices ('000 COP / sq.m.)

	Full sample			Treated			Control		
	2008	2014	2017	2008	2014	2017	2008	2014	2017
All blocks	172	430	1,078	286	592	1,942	147	241	533
	(143)	(584)	(810)	(184)	(278)	(355)	(118)	(246)	(395)
	42,993 blocks			7,700 blocks			35,293 blocks		
Residential	115	401	833	275	427	1,862	193	517	840
	(123)	(275)	(599)	(171)	(149)	(233)	(148)	(425)	(697)
	38,780 blocks			6,420 blocks			32,360 blocks		
Manufacturing	81	323	625	189	175	1,419	141	208	469
	(105)	(287)	(555)	(142)	(125)	(241)	(105)	(180)	(303)
	1,773 blocks			370 blocks			1,403 blocks		
Services	263	1,139	1,773	405	1,079	2,715	120	113	338
	(206)	(1,119)	(1,509)	(226)	(470)	(734)	(91)	(100)	(189)
	2,440 blocks			910 blocks			1,530 blocks		
Strata 1	70	290	487	143	703	993	304	1,140	1,834
	(91)	(576)	(668)	(170)	(518)	(844)	(177)	(716)	(927)
	11,619 blocks			845 blocks			10,774 blocks		
Strata 2	129	478	879	125	524	983	294	1,120	1,779
	(63)	(208)	(378)	(64)	(235)	(461)	(184)	(733)	(957)
	15,241 blocks			362 blocks			14,879 blocks		
Strata 3	232	864	1,466	234	881	1,506	360	1,389	2,074
	(77)	(316)	(465)	(73)	(277)	(392)	(254)	(1,009)	(1,338)
	11,888 blocks			4,502 blocks			7,386 blocks		
Strata 4	352	1,193	1,886	356	1,249	2,017	276	1,068	1,701
	(105)	(376)	(521)	(97)	(410)	(536)	(191)	(762)	(996)
	2,389 blocks			1,003 blocks			1,386 blocks		
Strata 5	474	1,643	2,431	548	1,974	2,907	268	1,031	1,661
	(173)	(887)	(1,214)	(156)	(952)	(1,285)	(171)	(669)	(872)
	1,016 blocks			499 blocks			517 blocks		
Strata 6	635	2,462	3,432	725	2,907	4,019	256	969	1,588
	(246)	(1,023)	(1,375)	(181)	(635)	(809)	(141)	(548)	(744)
	840 blocks			489 blocks			351 blocks		

Table 7: Average (and standard deviation) of control variables

	Full sample			Treated			Control		
Time-variant controls	2008	2014	2017	2008	2014	2017	2008	2014	2017
Block Floor area (m^2)	733	913	940	825	950	989	713	905	929
	(3,368)	(7,368)	(4,237)	(2,996)	(3,703)	(3,770)	(3,443)	(7,947)	(4,332)
ZPU Density (inh/ha)	101	108	111	124	126	128	97	104	108
	(133)	(143)	(150)	(112)	(115)	(117)	(137)	(149)	(156)
ZPU Population per household	1.5	1.4	1.3	2.2	2.0	1.8	1.4	1.3	1.2
	(1.7)	(1.6)	(1.5)	(1.6)	(1.5)	(1.3)	(1.7)	(1.6)	(1.5)
Time-invariant controls									
ZPU Metropolitan parks (km^2)		3.5			1.7			3.9	
		(51.6)			(6.8)			(56.8)	
ZPU Zonal parks (km^2)		1.5			2.2			1.4	
		(1.3)			(1.5)			(1.2)	
ZPU Num. health-related inst.		46			163			20.20	
		(171)			(329)			(92)	
ZPU Number of facilities		58			88			52	
		(55)			(33)			(57)	
ZPU Num. Transmilenio stations		0.7			2.1			0.4	
		(1.5)			(2.1)			(1.1)	

C. Spillover effects

Since Bogotá's real estate market is not perfectly segmented by block, we now consider potential spillover effects when prices in one block are affected by prices (or their determinants) in nearby blocks.

In Table 8 we control for these spillover effects by including the log of average price per square meter in neighbouring blocks located at 50 m (Columns 1 and 4), 100 m (Columns 2 and 5) and 500 m (Columns 3 and 6) and belonging to any ZPU (Columns 1, 2, and 3) or only to the same ZPU (Columns 4, 5 and 6).

The estimated coefficient of interest is positive and statistically significant in all six specifications, confirming that blocks affected by Policy-562 experienced an increase in real estate prices. Furthermore, these results are not statistically different from our preferred specification in Column 4 of Table 3, indicating that spillover effects from neighbouring blocks do not bias our preferred estimates.

Table 8: The effect of Policy-562 (Law of Heights) on real estate prices: Controlling spillovers

	[1]	[2]	[3]	[4]	[5]	[6]
Neighbouring blocks	50 m	100 m	500 m	50 m	100 m	500 m
Belonging to	Any ZPU	Any ZPU	Any ZPU	Same ZPU	Same ZPU	Same ZPU
Policy-562×After-562	0.275^a	0.277^a	0.327^a	0.276^a	0.307^a	0.317^a
	(0.036)	(0.037)	(0.037)	(0.036)	(0.037)	(0.037)
Adjusted R^2	0.262	0.267	0.268	0.262	0.265	0.268

Notes: 429,930 observations (= 42,993 blocks × 10 years) in each regression. Regressions include time-variant controls, block fixed-effects, year fixed-effects, and the log of the average price per square meter of neighbouring blocks. Regressions are weighted by the number of lots that make up each block.. Robust standard errors are clustered by ZPU and are in parenthesis. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

D. Heterogeneous results

We now turn our attention to studying the heterogeneity of our results by year of the treatment subperiod, for each of the three main land uses, and for each of the six strata.

Treatment years

To study whether the effect of the Law of Heights changed over time, we estimate Equation (A 1) allowing for different effects in the years belonging to the treatment subperiod.

$$\begin{aligned} \ln(\text{Price}_{it}) = {} & \beta_1 \times \text{Policy-562}_i \times \text{Year 2015} \\ & + \beta_2 \times \text{Policy-562}_i \times \text{Year 2016} \\ & + \beta_3 \times \text{Policy-562}_i \times \text{Year 2017} \\ & + \beta_4 \times \text{Time-variant controls}_{it} + v_t + \alpha_i + u_{it} \end{aligned} \tag{A 1}$$

Column 1 of Table 9 reports the main results: The estimated coefficients for the three years are positive and statistically significant (0.542, 0.207 and 0.139), and show that the effects of Policy-562 decreased over time. In particular, treated blocks experienced an increase in real estate prices around 72% in 2015, 23% in 2016, and 15% 2017.

To understand the decreasing effect of the Law of Heights, it is necessary to look at the political context of Bogotá during these years. In 2015, the campaign for the mayor of Bogotá was advanced. The two candidates with more options to win opposed the current administration that promoted Policy-562, and in both cases, they proposed to repeal the law. A new mayor was elected in October 2015 and, although his term came into effect in January 2016, one of his first announcements was the repeal of the Law of Heights (effective February 2016). As explained in Section 4.1, between March 2016 and December 2017, the last projects approved by Policy-562 were developed.

Land use

We also explore the heterogeneous effects related to the main land use of the block. To do so, we estimate Equation (A 2):

$$\begin{aligned} \ln(\text{Price}_{it}) = {} & \sum_{j}^{3} \left(\beta_{1j} \times \text{Policy-562}_i \times \text{After-562}_t \times \text{Land Use}_j \right) \\ & + \beta_2 \times \text{Time-variant controls}_{it} + v_t + \alpha_i + u_{it} \end{aligned} \tag{A 2}$$

where Land Use_j are the three main land uses available in the SPD dataset (Residential, Manufacturing, and Services).

Column 2 of Table 9 reports results showing that Residential and Services blocks were significantly affected by the Law of Heights with estimated coefficients (0.268 and 0.281) similar to their average counterpart in Column 4 of Table 3 (0.289). These estimated coefficients mean that the related blocks experienced increases in their prices around 30.8% (Residential) and 32.4%

(Services). On the other hand, the estimated coefficient for Manufacturing is lower (0.151) and non-significant.

Table 6 in Appendix B. shows that, between 2014 and 2017, Residential and Services treated blocks experienced huge increases in their prices. For the case of Residential blocks, average prices grew from 427,000 to 1,862,000 COP/sq.m., which represents a growth of 336.7% in four years. Policy-562 explains 9% of this growth. Similarly, prices in Services blocks increased 151.8% (from 1,079,000 to 2,715,000 COP/sq.m.), and the Law of Heights explains 21% of the Services growth.

Table 9: The effect of Policy-562 (Law of Heights) on real estate prices: Heterogeneity

	Year		Land use		Strata
	[1]		[2]		[3]
Policy-562×Year 2015	0.542^{a}	Policy-562×After-562	0.268^{a}	Policy-562×After-562	0.932^{a}
	(0.057)	×Residential	(0.038)	×Stratum 1 (low)	(0.198)
Policy-562×Year 2016	0.207^{a}	Policy-562×After-562	0.151	Policy-562×After-562	-0.211
	(0.047)	×Manufacturing	(0.360)	×Stratum 2 (low)	(0.141)
Policy-562×Year 2017	0.139^{a}	Policy-562×After-562	0.281^{b}	Policy-562×After-562	0.465^{a}
	(0.025)	×Services	(0.120)	×Stratum 3 (low)	(0.039)
				Policy-562×After-562	-0.193^{b}
				×Stratum 4 (high)	(0.091)
				Policy-562×After-562	-1.652^{a}
				×Stratum 5 (high)	(0.217)
				Policy-562×After-562	1.720^{a}
				×Stratum 6 (high)	(0.173)
Adjusted R^2	0.220	Adjusted R^2	0.220	Adjusted R^2	0.218

Notes: 429,930 observations (= 42,993 blocks × 10 years) in each regression. Regressions include time-variant controls, block fixed-effects, and year fixed-effects. They are also weighted by the number of lots that make up each block. Robust standard errors are clustered by ZPU and are in parenthesis. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

Strata

Finally, we study the effects of Policy-562 according to the strata where blocks are located by estimating Equation (A 3):

$$\begin{aligned}\ln(\text{Price}_{it}) = &\sum_{j}^{6} \left(\beta_{1j} \times \text{Policy-562}_i \times \text{After-562}_t \times \text{Strata}_j\right) \\ &+ \beta_2 \times \text{Time-variant controls}_{it} + v_t + \alpha_i + u_{it}\end{aligned} \tag{A 3}$$

where Strata_j are the strata of the city (1, 2 and 3 low strata and 4, 5 and 6 high strata).

Column 3 of Table 9 reports main results using de 2017 strata definition[16]. They indicate that Policy-562 positively affected prices in low strata 1 (0.932) and 3 (0.465) zones. In particular, their treated blocks increased real estate prices by 154% (strata 1) and 59% (strata 3), respectively. On the contrary, prices in low strata 2 were negatively but not significantly affected. For the high

[16] As commented in Appendix A., the definition of strata changed in some specific years of the studied period. However, only 4% of the blocks were affected by this change of classification. As a robustness check, we re-estimated Equation (A 3) using the 2008 strata definition. Results hold and are available upon request.

strata zones, all estimated coefficients are significant but differ in their sign. While strata 4 and 5 were negatively affected (-0.193 and -1.652), the effect was positive in strata 6 (1.720). These results mean that the Law of Heights decreased real estate prices in strata 4 and 5 by 18% and 81%, and increased them in strata 6 by 458%.

It is important to note that most of these Policy-562 effects at the strata level are either of a different sign (strata 4 and 5) or are larger (strata 1 and 6) than the final price growth experienced by the treated blocks during the treatment period (which can be computed using the average values reported in Table 6 in Appendix B.). This means that there were other confounding factors with opposite and compensating effects.

E. Propensity score matching

As a robustness check, in Columns 1 and 2 of Table 4 we use the 'propensity score matching' technique to adjust imbalances in the explanatory variables between treated and controls to ensure that the post-treatment effects. We first estimate a logit using the dummy Policy-562$_i$ as the dependent variable. As explanatory variables, we use the set of control variables of Eq. (1) measured in 2007. We then compute the 'propensity score' and control blocks are matched to their treated counterparts based on a similar 'propensity score'. To do so, we use the 'nearest neighbor matching with replacement' method, whereby a given control unit can be matched to more than one treatment unit, which increases the average quality of matching and reduces the bias.

Panel A of Table 10 reports the results of the logit. They show that areas with more population density, inhabitants per household, metropolitan parks, private institutions and new buildings approved were less likely to be included under the influence of Policy–562. Recall that one of the objectives of the Law of Heights was to reactivate areas suffering some degree of abandonment. On the other hand, the group of variables that seeks to measure access to the city's facilities, and the general advantages related to the location of each lot within the city, presents a positive and significant effect on the probability of being included in the treated areas, where the existence of Transmilenio stations and zonal parks are highly significant factors that increase this probability by 50 and 61%, respectively.

In Panel B of Table 10 we perform several tests to determine whether or not we achieve a good matching. First, we perform a comparison of means between treated (column 3) and controls (column 4) in the unmatched and matched samples (see Rosenbaum and Rubin, 1985). In the unmatched sample, the treated group are disproportionately located in ZPUs with low population density, inhabitants per household and km^2 of metropolitan parks, and with a large number of private institutions, facilities, Transmilenio stations, km^2 of approved projects, new buildings approved and km^2 of zonal parks. In the matched sample, most of the differences in means between the treated and the control group are statistically non-significant (columns 5 and 6). Second, we also examine standardized bias both before and after matching (column 7); before matching, many variables presented a bias greater than 20%, which is the level above which some authors suggest linear regression coefficients risk being highly biased (Imbens and Wooldridge, 2009). After matching, all the variables present a bias below this level and with a substantial reduction in % bias (column 8) (between 79 and 99%). Third, we also re-estimate the propensity score on the matched sample and compare the pseudo-R^2 before and after matching, which are actually 0.422 and 0.003, respectively. LR tests of joint significance of the regressors before and after the matching present values of 17061 and 5.41, respectively. All these tests suggest that matching is successful in balancing the sample.

At the end of the matching process, the new sample includes 34,449 blocks (80% of the initial sample). The treated and controls groups are made up of 6,186 and 28,263 blocks, respectively. This indicates that the original sample (with 35,293 and 7,700 blocks, respectively) considers blocks that are structurally different and that do not have a counterpart.

Table 10: Propensity score matching: Selecting treated and control groups

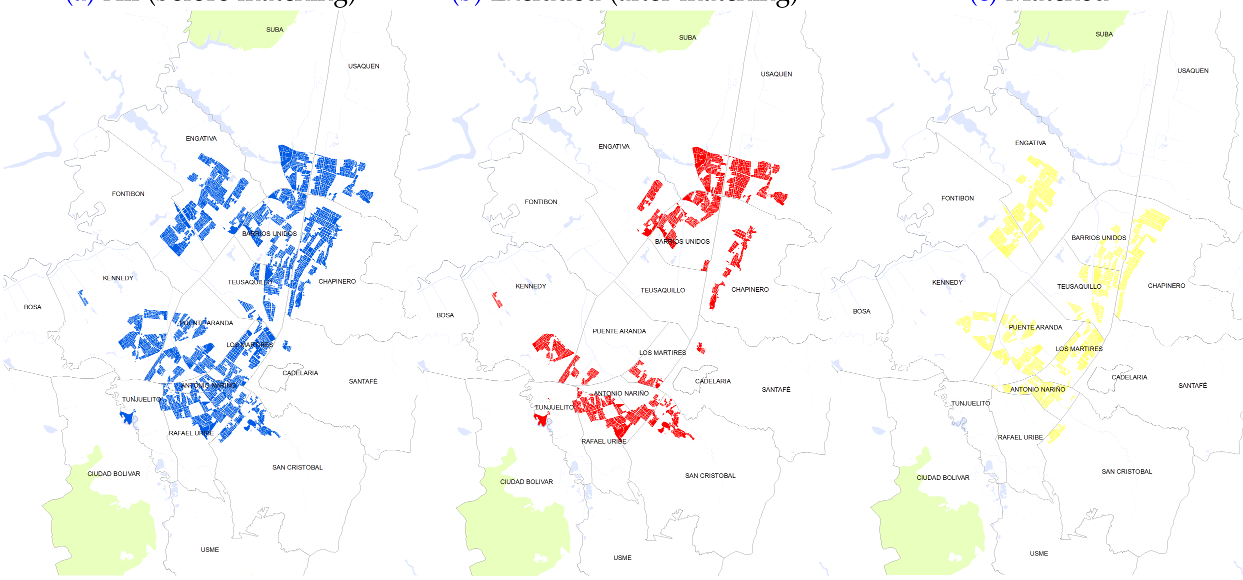
	Panel A Logit	Panel B Matched vs. unmatched samples						
			Mean		t-test			
		Sample	Treated	Control	t	$p-value$	% $bias$	%∇ $\|bias\|$
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
ln(Population per hectare)	-1.056^{a}	Unmatched	4.785	5.145	-22.24	0.00^{a}	-26.2	
	(0.038)	Matched	4.845	4.850	-0.20	0.85	-0.4	98.6
ln(Population per household)	-2.015^{a}	Unmatched	1.510	1.528	-15.12	0.00^{a}	-21.1	
	(0.161)	Matched	1.512	1.512	-0.80	0.43	-0.8	96.3
ln(km^2 of metropolitan parks)	-0.280^{a}	Unmatched	0.138	0.230	-11.88	0.00^{a}	-16.0	
	(0.034)	Matched	0.163	0.143	2.02	0.04^{b}	3.4	79.0
ln(km^2 of zonal parks)	0.609^{a}	Unmatched	0.566	0.200	69.42	0.00^{a}	74.1	
	(0.037)	Matched	0.564	0.589	-2.34	0.02^{b}	-5.1	93.1
Number of private institutions	-0.005^{a}	Unmatched	163.0	20.20	70.00	0.00^{a}	59.1	
	(0.001)	Matched	169.0	163.5	0.85	0.39	2.3	96.2
Number of facilities	0.004^{a}	Unmatched	88.4	51.5	55.2	0.00^{a}	79.6	
	(0.001)	Matched	91.9	91.9	-0.14	0.89	-0.2	99.8
Number of Transmilenio stations	0.501^{a}	Unmatched	2.1	0.5	100.53	0.00^{a}	101.2	
	(0.011)	Matched	1.9	1.9	-0.21	0.84	-0.5	99.5
ln(km^2 of approved projects)	0.294^{a}	Unmatched	9.579	5.491	72.31	0.00^{a}	107.8	
	(0.007)	Matched	9.449	9.427	0.47	0.64	0.6	99.5
Number of approved new buildings	-0.005^{a}	Unmatched	142.5	95.6	21.29	0.00^{a}	29.7	
	(0.001)	Matched	152.3	149.2	1.40	0.16	2.0	93.4
Strata dummies	✓							
Pseudo-R^2 0.439 LR-χ^2 17759		Unmatched	Pseudo-R^2 0.422	LR-χ^2 17061	Median bias 37.9			
		Matched	Pseudo-R^2 0.003	LR-χ^2 5.41	Median bias 1.7			

Notes: 42,993 observations. Standard errors in column 1 are in parenthesis. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

Figure 4 shows the stages followed by the matching process for the case of the treated group. Figure 4a shows the location of all blocks affected by the Law of Heights. Figure 4b shows the blocks that are excluded after the matching process. Although there are excluded blocks in several locations of the city, they appear to be concentrated mainly in two with different strata. On the one hand, blocks are excluded in the municipalities of Tunjuelito and Rafael Uribe with a predominance of low-strata properties (1 and 2). On the other hand, blocks are excluded in the municipality of Chapinero with a predominance of high-strata properties (4 and 6). Finally, Figure 4c shows the selected treated block after the matching process.

Figure 4: Stages of the matching process for treated blocks

(a) All (before matching) (b) Excluded (after matching) (c) Matched



Part III

Causes and consequences of urban form: Evidence from US cities.

Causes and consequences of urban form: Evidence from US cities

Diego Buitrago-Mora

Ph.D. Candidate

Miquel-Àngel Garcia-López

Supervisor

Abstract: This paper investigates the causes and consequences of urban form in US metropolitan areas. Using data at the tract level for the years 1990, 2000 and 2010, we first investigate the spatial pattern of decentralized jobs by identifying employment subcenters using McMillen (2001)'s nonparametric method and characterizing and cataloging cities according to their urban spatial structure (monocentric or polycentric). Then we study why subcenters form and emerge focusing on the role of agglomeration forces (population) and dispersion forces (congestion). Finally, we analyze the effects of density on economic, social and environmental city outcomes and explore the role of urban form.

Key words: Urban form, polycentrism, income segregation, residential segregation, pollution, per capita income.

JEL: R11, R12

2. Introduction

Since the mid-twentieth century, cities in the United States have experienced a decentralization of employment (Glaeser and Kahn, 2001, 2004). The typical urban area in the United States at the turn of the century was dense and compact, but has since shifted to a more dispersed area. Residents are able to work in the Central Business District (CBD) and reside in the suburbs. Glaeser and Kahn (2001) mention the emergence and accelerated massification of the automobile as a determining factor (in 1910, the ratio of cars per household was 0.02; in 1990, it was 1.43[17]). Employment centers within a given urban region form an interdependent system with a similar size distribution and pattern of specialization to that of cities within a broader regional or national economy Anas, Arnott, and Small (1998).

This paper employs a number of previously proposed techniques and proposes a number of new ones to contribute to the discussion about the causes and effects of polycentrism in the United States.

The study of the phenomenon of decentralization of employment in the United States has focused on identifying new centers of activity (subcenters) in Metropolitan Statistical Areas (MSAs[18]) as an indicator of decentralization throughout the nation. McDonald and McMillen (1990) work is an up-to-date review of studies conducted until 1990. This study also indicates that the Chicago metropolitan statistical area has multiple subcenters (polycentric city). They conduct an empirical comparison of the pattern of employment decentralization between 1956 and 1970. Meanwhile, Giuliano and A. Small (1999) describe the emergence of 32 subcenters in the Los Angeles metropolitan area between 1970 and 1980, emphasizing the role that various industries played in the subcenters' development and expansion. Small and Song (1994) examines Los Angeles and estimates the employment and population density functions. Glaeser and Kahn (2001) examines the phenomenon at the national level. They estimate traditional density functions and display results by ZIP codes, MSAs, and Census Bureau regions (South, West, Northeast, and Midwest). Thus, we first examine the structure and urban form of MSAs in the United States by identifying subcenters.

Regarding the methodologies to identify them, some authors use employment and population density thresholds (Giuliano and Small, 1991, Shunfeng Song, 1994, Cervero and Wu, 1997, McMillen and McDonald, 1998, William T. Bogart and William C. Ferry, 1999, Anderson and Bogart, 2001); Another methodology is identifying peaks, in which they regress employment (or population) density to the distance to the CBD as the main explanatory (McDonald, 1987, McDonald and McMillen, 1990). Furthermore, the behavior of mobility patterns (i.e., commuting) (Gordon and Richardson, 1996) and, finally, the study of the residuals of the density functions from the monocentric model. This last approach was introduced by McDonald and Prather (1994) with Ordinary Least Square (OLS); McMillen (2001), in turn, improves it by introducing a two-dimensional density function (log of employment density on the distance to CBD) (Garcia-López, Hémet, and Viladecans-Marsal, 2017). We follow McMillen (2001)'s approach.

[17] See the highway statistics summary of 1995

[18] We use cities and MSA indistinctly in this paper. We also include Consolidated Metropolitan areas (CMSA). The explanation of its specific characteristics is presented in section 3.1

The author suggests using a non-parametric estimator to smooth the employment density, followed by a weighted least squares (WLS) and then running a regression (parametric or semiparametric) to see if potential subcenters have a significant effect on employment density. The proposed strategy has the advantage of being suitable to any urban environment. That is why we will use national census data from 1990, 2000, and 2010. When we finish this section, we will identify the subcenters within each MSA and classify them as monocentric or polycentric. In the same section, we use density functions to assess the degree of job decentralization at various levels of geographic aggregation and propose an analysis with an intra- and inter- MSAs characterization of the identified subcenters.

In the next section, we study the causes of polycentrism in USA. We restored Fujita and Ogawa (1982)'s hypothesis to determine the primary drivers of employment spread. We investigate whether commuting costs and population density play a significant role in determining why employment relocates outside the CBD using the classification of MSAs based on their urban form obtained in the previous section. Prior research has attempted to identify the primary cause of the emergence of subcenters in the United States. The construction of roads and transportation infrastructure is a hypothesis that Baum-Snow (2007*a*), Duranton and Turner (2011*a*), Garcia-López *et al.* (2017) have all investigated. As well as, population and commuting costs (Fujita and Ogawa, 1982, McMillen and Smith, 2003)

Once we provide evidence of the causes, we move to the consequences, of polycentrism in the United States. The starting point once more, is the classification of MSAs from the first section. Moreover, we defined employment density within three kilometers of each subcenter as our main measure of urban form. With this kind of measurement, we can get a full picture of MSA physical structures, their association with employment, and how they relate to economic, socioeconomic, and environmental outcomes. Thus, we introduce a set of changing dependent variables to assess such potential consequences of urban form.

Regarding the first group of outcomes, existing research acknowledges that the dispersion and subsequent re-clustering of employment in subcenters other than the original one (polycentrism) relieves the pressure on the traditional center and mitigates the drawbacks of over-accumulation of employment. However, the literature demonstrates mixed results. Meijers and Burger (2010) suggests that the economic benefits of polycentricity are greater in smaller cities in the United States; Cheshire and Carbonaro (1996) compares European regions according to their main productive activities (agricultural, resource-based industries, and services) and reports a high correlation with income growth; Wu, Shen, and Sun (2016) examines the morphology of urban form in Chinese cities and concludes that the dependence of the urban economy on spatial structure depends on city size. We assessed the economic performance of MSAs by examining per capita income.

For the second group of consequences (socioeconomic), we dived into income segregation and residential segregation. With respect to the first group, we use the rank-order information theory index (H), first suggested by Reardon (2011). We explore three arguments for employing this index as a measure of income segregation among MSAs in the United States. The first is that it distinguishes between changes in income disparity and income segregation (Reardon, 2011). The second is that, despite the fact that its magnitude lacks an intuitive interpretation, comparing it across MSAs and time horizons reveals where and when segregation is highest and lowest (Reardon, 2011), making it suitable for the present investigation. And third, it is not affected by the level of income inequality in a metropolitan area, so it measures the extent to which families of different incomes are segregated among neighborhoods more accurately than a measure that calculates the ratio of census tract median family income to MSA median income. As complementary measures, we also use the segregation of poverty index (PovH-10) and the segregation of affluence index (AffH-10) as indicators of disparities between the extreme groups of the income distribution.

As for residential segregation, there are numerous indexes that measure it and can be associated, in the context of this research, with urban form. Massey and Denton (1988) classify them into 5 groups: measures of evenness, exposure, concentration, centralization, and clustering. In such a theoretical context, the purpose of this study is to determine the degree of equality between demographic groupings within each MSA, in the same way that we seek to obtain results at the national level. Consequently, we employ a dissimilarity index by racial group (between Whites, African-Americans, Hispanics, and Asians). The Census Bureau defines the dissimilarity index as the proportion of a population group that would have to relocate to a different neighborhood (census tract) in order for each census tract to have the same proportion of that population group as the MSA overall. The addition of this index to the income segregation study will provide a larger view of the intra-metropolitan dynamics of American families, as well as a national perspective.

The last part of the study of the consequences aims to distinguish the environmental impact of cities and their relationship with urban form. The above by measuring the quantity of particles ($PM10$) and polluting gases (Co, SO_2, $Ozone$, and NO_2). There are two ways to figure out the amount of particles that cause pollution in urban areas of the United States. First, satellite pictures, whose primary source is typically NASA. Carozzi and Roth (2023) studies the air quality of American cities using this tool. The second source is the ground stations (sites). van Donkelaar, Martin, Brauer, and Boys (2015) demonstrates that the results from the two information sources might be fairly similar; Borck and Schrauth (2021) uses the German stations but acknowledges that aerial photographs are also a legitimate alternative.

Regarding the first option, despite its enormous potential, photos have been available in the United States since 1998; thus, it does not correspond to our period of interest. On the other hand, in some cases, the availability of information from sites dates back to the 1950s. In addition, spatial availability is widespread throughout the whole country, particularly in urban contexts. Hence, we utilize daily data from pollutant gases and particle monitoring stations (sites).

Following the introduction, the sections of this investigation are as follows: A second section that investigates urban form and produces the results needed for the following sections. In the third section, we tested the major hypotheses regarding the causes of polycentrism in the United States, and in the final section, we introduced a set of variables that measure the economic, socioeconomic, and environmental consequences of polycentrism in the United States between 1990 and 2010. The final sections contain conclusions and appendices.

3. Urban spatial structure of US cities

In this section, we define spatial trends within and between MSAs. By classifying them based on the number of employment sub-centers, we can determine whether they are polycentric or monocentric. Next, we examine their degree of centrality, regional trends, and then intra-metropolitan trends (characterization of sub-centers).

We defined a Metropolitan Statistical Area (MSA), following the United States Office of Management and Budget (OMB) in 1999, as a unit of counties that must include at least one city, or an Urbanized Area (UA), as defined by the Census Bureau, with at least 50,000 inhabitants. Aside from that, it must reach a metropolitan population of at least 100,000 people. Furthermore, a Consolidated Metropolitan Statistical Area (CMSA) is made up of more than one MSA and must have a population of at least one million people[19]. In addition to the 1999 concept, we consider the Economic Census delimitation of a CBD in 1982, which defines them as areas of concentration of economic activity (retail and services), characterized by high traffic flow, and defined in terms of existing census tract boundaries (Bureau, 1987). We complete our database by matching 1982 and 1999 definitions, which includes 205 MSAs spread across 48 states in the continental United States[20].

Furthermore, we employ the Census Bureau's four geographic regions established in 1950. The union of states is the largest grouped statistical entity used for several US censuses. Their formation is the result of historical factors (beginning with the first census in 1790) and geographical factors (which naturally determine the formation of some state boundaries)[21]. The Northeast, Midwest, South, and West are the four regions. They are defined by the following states:

Northeast: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York,and Pennsylvania. **Midwest**: Indiana, Illinois, Michigan, Ohio, Wisconsin, Iowa, Nebraska, Kansas, North, Dakota, Minnesota, South Dakota and Missouri. **South**: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, and Texas. **West**: Arizona, Colorado, Idaho, New Mexico, Montana, Utah, Nevada, Wyoming, Alaska, California, Hawaii, Oregon, and Washington.

Finally, the employment data comes from the United States Decennial Census, specifically the section on employment status. We used the censuses from 1990, 2000, and 2010 and the Longitudinal Tract Data Base (LTDB) created by Logan, Xu, and Stults (2014) to reduce geographic differences between decennial censuses.

3.1 Monocentric and polycentric MSA

In this research, we want to differentiate between polycentric and monocentric MSAs in the United States using Mcmillen's Methodology. Such a process is essential since it establishes

[19]We refer to MSA without specifying whether it is consolidated or not. More information can be found at https://www.census.gov

[20]We exclude Alaska and Hawaii because the states must be physically continuous in order to apply the subcenter identification methodology (McMillen, 2001).

[21]See chapters 2 and 6 of the geographic-areas-reference-manual at https://www.census.gov

the group of MSAs to study in subsequent steps. However, this section of the research itself is a novelty since it uses Census tracts from 2010.

To do this, we will use the logarithm of the employment density (y), defined as the number of jobs in each census tract over the kilometers of land that compose it.

The first step is a locally weighted regression (LWR) to smooth the density over space. The primary function of the LWR estimator is:

$$\sum_{i=1}^{n}(y_i - m - \beta(x^*)(x_i - x))^2 k_i \tag{4}$$

In Equation 4, k_i determines the weight given to each point i to project at x. Then, separate functions can be calculated for each point: $\hat{y}_j$=$\hat{m}(x_j)$ for $j = 1,...,n$.. In our case, we resorted to Nadayara-Watson kernel estimator, which is a simpler version that equals to zero β^* (McMillen, 2001). So that, we use distances to the CBD (north-south and east-west) from each census tract.

$$y_i = m + \beta_1(x^*)(x_{1i} - x_1) + \beta_2(x^*)(x_{2i} - x_2), \tag{5}$$

where y_i is the employment density in tract i, and x_{1i} and x_{2i} the distance to that point. x_1 and x_2 are the target points and x^* is the one for expansion.

Each regression is run with the observations around each census tract and assigned a greater weight (k) to the closest observations. In this regard, we must define ¿how close? That being said, a window of the data that receives some weight must be chosen, and since we need a monocentric benchmark (with no employment density picks), this is a critical decision, as pointed out by McMillen (2001).

We must choose a window that is not too small, or there will be no difference between the smooth function and the density peaks (perfect fit), making it difficult to identify the subcenters. Not very large, either, because if the initial smooth deviates significantly from the estimate prediction, small density peaks may be identified as subcenters. Therefore, we use different window sizes: 10%, 30%, 50%, 70%, and 90%. Afterwards, we select two critical thresholds (significance levels): 1.96 (5%) and 1.64 (10%). This methodology allows us to generate different benchmarks, which we compare with the regressions of each census tract. If the residuals generated are significant and continuous[22] we identify a subcenter.

Table 11 shows the results for the subcenter identification technique for the 1990, 2000, and 2010 decennial censuses. The first two rows of each panel present the initial groups of 272 MSAs divided between polycentric and monocentric. From row 3 to 10 (of each panel), the polycentric cities are grouped by number of subcenters. Overall, as the window grows, more polycentric (less monocentric) cities are found. In panel A, we can see that the majority of polycentric cities are concentrated in the first three groups (from 1 to 15 subcenters). Panel B presents the results with a critical threshold of 1.64. The change between windows generates greater changes in the distribution between monocentric and polycentric cities: 62 cities on average with 1 to 5 subcenters with a window of 10%, while the average is 191 in the window of 90%.

[22] As in Garcia-López *et al.* (2017), we use a "queen" criterion for continuity, which means that two census tract are continuous if they share at least one side

Table 11: Polycentric and monocentric MSAs by number of subcenters.USA, 1990, 2000 and 2010

Window		10%			30%			50%			70%			90%		
Year		90	00	10	90	00	10	90	00	10	90	00	10	90	00	10
Panel A = 0.05																
	M	168	174	173	114	121	109	88	94	89	74	67	73	66	63	66
	P	104	98	99	158	151	163	184	178	183	198	205	199	206	209	206
Subcenters	1 -5	65	59	60	112	107	116	139	134	137	152	160	156	167	171	165
	6-10	19	20	16	24	26	26	23	24	25	25	28	24	22	21	26
	11-15	7	7	10	9	6	9	8	7	7	7	3	7	7	7	6
	16-20	2	3	5	4	3	3	5	6	5	5	6	2	6	4	3
	21-30	5	4	3	5	5	6	6	3	7	6	5	8	2	3	4
	31-40	4	3	3	2	2	2	2	2	1	2	1	1	2	1	1
	41-60	2	2	1	2	2	0	1	2	1	1	2	1	0	2	1
	>60	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Panel B= 0.10																
	M	151	156	152	78	75	79	41	44	41	20	25	22	15	24	23
	P	121	116	120	194	197	193	231	228	231	252	247	250	257	248	249
Subcenters	1 -5	63	62	63	124	130	125	160	161	162	184	181	186	201	187	187
	6-10	25	19	24	33	31	32	31	30	33	31	28	29	27	30	35
	11-15	13	13	10	16	15	11	22	16	12	15	17	11	10	11	10
	16-20	4	4	4	6	6	9	3	7	10	6	7	10	7	7	5
	21-30	5	8	7	6	5	5	5	4	3	7	6	3	6	7	7
	31-40	6	5	6	4	5	6	5	5	7	4	4	7	3	2	2
	41-60	3	3	4	3	3	3	3	3	2	3	2	2	1	2	1
	>60	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Notes: M: Monocentric MSAs. P: Polycentric MSAs. The table displays the number of monocentric and polycentric MSAs in the country and the number of subcenters found, when considering the different thresholds (0.05 and 0.10). Panel A has a critical threshold of 1.96 (5%). Panel B critical threshold of 1.64 (10%).

3.2 *Metropolitan and regional pattern*

We now extend our analysis of the trends of job decentralization by showing stylized facts of the previous findings of numbers of subcenters. That is why, our starting point is the results of the section 3.1, particularly the results of the window of 50% and a significance level of 5%. This choice is arbitrary based on McMillen (2001) and Garcia-López *et al.* (2017). In any case, both suggest not to choose an extreme scenario. The former uses a 50% window, and the latter windows from 10 to 90%. Nevertheless, we present some robustness tests in the apendix A. In addition, we use the MSA area as a fixed effect to account for higher employment densities (Glaeser and Kahn, 2001). Then, we run the log of employment density to each census tract's distance to the CBD, as show in Equation 6.

$$log\left(\frac{Employment}{Squarekilometer}\right)_{i,t} = \alpha_i + \beta \times DistancefromCBD + \epsilon_{i,t} \qquad (6)$$

where i is a MSA or Region and t is 1990, 2000 or 2010. The β is our indicator of decentralization with respect to the CBD; that is, the more negative, less decentralized, and correspondingly, the more positive (or near to zero) more decentralized is the MSA[23].

The allocation of MSAs to each region was straightforward in all but 9 cases. Such MSAs are at the borders between regions and states. We allocated them according to the highest concentration of employment in 2010[24].

Table 12: Employment decentralization indicator by MSA and region.

Year	1990	2000	2010		1990	2000	2010
Panel A: top 5 MSAs by number of subcenters.				**Panel B: Regions**			
New York	-0.0292[c]	-0.0289[c]	-0.0310[b]	Northeast	-0.0320[a]	-0.0352[a]	-0.0321[a]
	(0.0088)	(0.0080)	(0.0081)		(0.0021)	(0.0020)	(0.0019)
Los Angeles	-0.0562[a]	-0.0432[a]	-0.0428[a]	Midwest	-0.0364[a]	-0.0344[a]	-0.0342[a]
	(0.0030)	(0.0031)	(0.0025)		(0.0025)	(0.0021)	(0.0021)
Chicago	-0.0544[a]	-0.0475[a]	-0.0486[a]	South	-0.0274[a]	-0.0244[a]	-0.0247[a]
	(0.0030)	(0.0032)	(0.0033)		(0.0022)	(0.0020)	(0.0019)
Houston	-0.0211[a]	-0.00627	-0.0056	West	-0.0132[a]	-0.0151[a]	-0.0169[a]
	(0.0055)	(0.0063)	(0.0054)		(0.0017)	(0.0016)	(0.0015)
Whashington	-0.0529[a]	-0.0404[a]	-0.0362[a]				
	(0.0032)	(0.0029)	(0.0030)				

Notes: [a], [b], and [c] indicate significance at 1, 5, and 10 percent level, respectively. MSAs are made up by several continues cities, respectively: *New York*, Northern New Jersey, and Long Island. *Los Angeles*, Riverside, and Orange County. *Chicago*, Gary, and Kenosha. *Houston*, Galveston, and Brazoria. *Washington*, and Baltimore.

Table 12 shows the decentralization patterns from 1990 to 2010 by city and region with the decentralization indicator (β). In panel A, we show the five cities with the most subcenters in 2010. All but New York, present a decentralization process.[25]. Washington, in turn, presents the most accelerated decentralization process, going from -0.0529 in 1990 to -0.0362 in 2010, followed by Houston, Los Angeles, and Chicago in that order. In 2010, the most centralized city was Chicago and the least Houston. The information from these MSAs gives us clues about regional behavior. Indeed, each region has at least one of the five MSAs: Northeast, New York. West, Los Angeles. South, Huston, and Washington. Midwest, Chicago.

In panel B, we can see results by region. The West region is the most decentralized in the three decades. In the upper map of Figure 5, we present the five cities that concentrated the most employment in 2010. Of those, Las Vegas (-0.021), Seattle (-0.035), San Francisco (-0.023), and Denver (-0.023) show a smaller degree of decentralization (the β of the individuals regressions) than the regional average (-0.017). Los Angeles, in contrast, is more decentralized (-0.0140), and being the largest urban agglomeration in this part of the country, makes the regional average

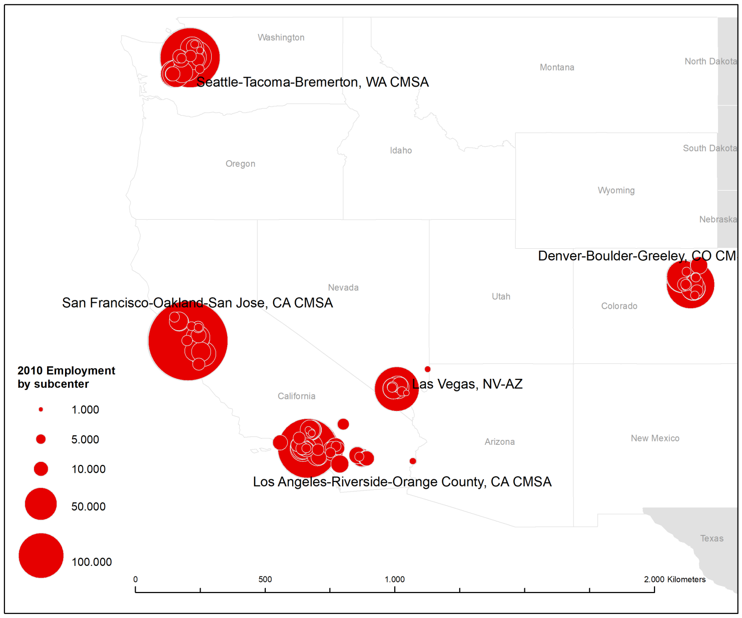
[23]These regressions are standard in Urban Economics and are based on the monocentric model Glaeser and Kahn (2001))

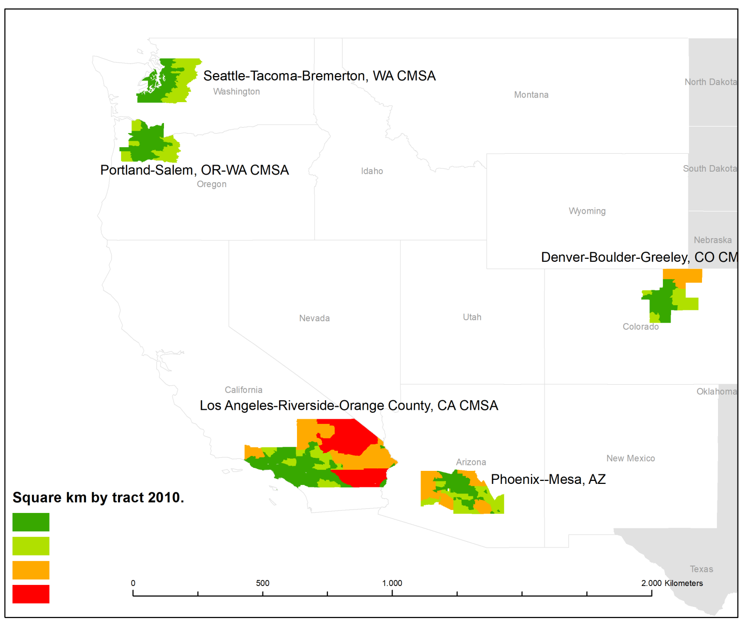
[24]Cincinnati-Hamilton, allocated in the Midwest. Evansville–Henderson, allocated in the Midwest. Huntington–Ashland, WV–KY–OH, allocated in the South. Las Cruces, allocated in the West. Louisville, allocated in the South. Parkersburg–Marietta, allocated in the South. Philadelphia-Wilmington-Atlantic City, allocated in the Northeast. Steubenville–Weirton, allocated in the Midwest. Wheeling, allocated in the South.

[25]This could be attributed to the re-consolidation of the CBD and particularly the subcenters around it in Manhattan. This is happening even though, we found increasing number of subcenters between 1990 and 2010 (see Figure 7)

drop. The lower decentralization indicator in San Francisco, for its part, supports the point made by Glaeser and Kahn (2001) that the "ideas–intensive" industries tend to be closer to CBD.

Figure 5: Largest MSAs in the West Region by total employment and land consume.





In the South, Miami and Atlanta are the most decentralized MSAs (-0.064 and -0.064). Miami has proportionally consolidated the most subcenters, going from 5 in 1990 to 12 in 2010. The Midwest region is the least decentralized. In this case, out of the five MSAs with the highest number of jobs in 2010, only Detroit is more decentralized than the regional average (-0.026). Minneapolis is the least decentralized (-0.074).

Finally, throughout the MSAs in the Northeast, the degree of decentralization is similar to the regional average except for Hartford that presents a β of (-0.071), being the least decentralized in this part of the country[26].

We finally complement this analysis with a mention of the land-occupancy patterns of subcenters by region. This provides us with elements for spatially dimensioning the subcenters in the United States. To do so, we use GIS data and present in the bottom panel of Figure 5 the results for the West region where it shows that it is the most decentralized region, mainly due to the composition of the Los Angeles CMSA.

In the Midwest, Cleveland and Detroit are not part of the top five on land occupation, but they are in employment. Similarly, after adding the subcenters identified in 2010, Minneapolis and Chicago have a similar land extension (171 and 179 km2 respectively). However, the total employment of the former represents 35% of employment of the latter. At the Northeast Region we can see a clear land-use pattern, which is lower near the CBD and increases as it recedes. Except for Buffalo, the cities that use the most land are also the ones that generate the most jobs. For its part, the South region presents the most significant difference between land consumption and employment concentration. If we compare with the upper map, Washington and Miami are not in the metropolitan areas that occupy the most land. Also, four out of the five largest cities are in the state of Texas (Houston, Dallas, Austin, and San Antonio), where consumption of land increases as it moves away from CBD (red dots), contrary to Atlanta, which remains relatively constant in the land–consumption by subcenters (does not show areas with red).

3.3 *Characterization of the subcenters*

We present an intra-metropolitan assessment to round out our analysis of the dynamics of job decentralization in the United States. In other words, we want to create a subcenter profile. Furthermore, we chose three cases that helped us understand the dynamics of employment within MSAs during the analysis period due to their decentralization patterns: Atlanta, New York, and Miami. McMillen (2001)'s methodology aids in the identification of subcenters via density peaks; however, as the authors point out, conclusions about the size of the subcenters cannot be drawn unless critical values for employment density are included, as in Giuliano and Small (1991)'s method. As a result of combining the two techniques, we obtain results with an intra-metropolitan scope.

[26] In Appendix B. are the South, Midwest and Northeast maps

Table 13: Subcenters composition. Main statistics.

	Regions	Census tracts[a] total	Census tracts[a] mean	Census tracts[a] s.d.	Land[b] total	Land[b] mean	Land[b] s.d.	Employment[c] total	Employment[c] mean	Employment[c] s.d.
Panel A= 1990	USA	1565	6	14	9235	50	102	13 796 517	74 981	175 544
	Northeast	343	10	26	583	25	37	3 490 079	151 743	370 307
	Midwest	335	5	11	1616	34	54	2 980 685	62 098	115 892
	South	548	5	10	5517	67	135	4 823 576	58 824	116 372
	West	339	7	15	1519	49	81	2 502 177	80 715	138 551
Panel B= 2000	USA	1494	5	14	4629	26	45	13 481 807	76 614	181 203
	Northeast	351	10	28	477	22	31	3 272 279	149 627	378 551
	Midwest	327	5	11	1079	23	42	2 884 916	62 474	125 469
	South	521	4	10	2311	30	52	4 705 800	62 668	122 328
	West	295	6	13	761	23	35	2 618 812	80 192	142 796
Panel C= 2010	USA	1592	6	15	5455	30	52	14 892 155	82 510	208 164
	Northeast	395	12	32	533	23	36	3 916 645	171 385	460 379
	Midwest	341	5	10	1429	29	53	2 923 760	60 402	122 396
	South	566	5	10	2718	34	58	5 192 445	65 546	133 235
	West	290	6	12	774	26	44	2 859 305	96 285	147 012

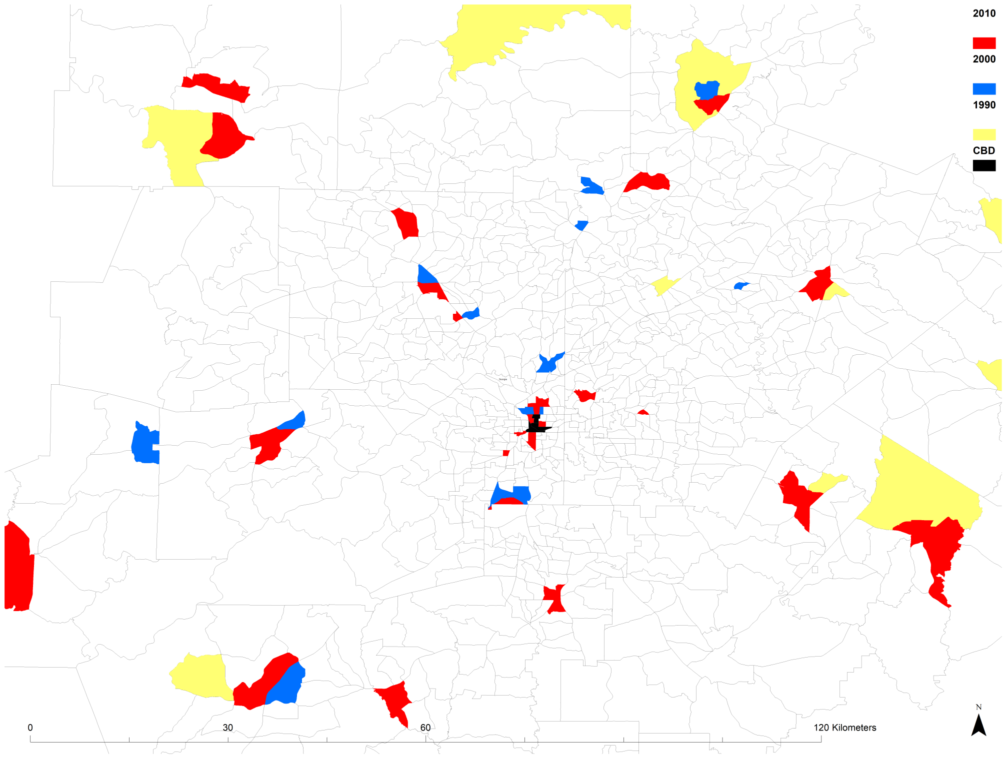
Notes:[a] Census tracts identified as part of subcenters (CBD included).[b] Sum of square kilometers of tract's extension .[c] Number of jobs within subcenters identified (CBD included).

Table 13 shows the main statistics of subcenters[27]. Columns 1 to 3 include statistics by census tract. The average number of census tracts by subcenter remains at six for the entire country between 1990 and 2010. When looking by region, the tract´s total number increases in the Northeast, Midwest, and South, which accounts for the increase in the average size of the individual subcenters and, therefore, decentralization of employment.

It is possible that the census tract size change between decennial censuses and as we mentioned, We do handle this difference with the use of LTDB (Logan *et al.*, 2014). As an example, we can see the case of Atlanta. In Figure 6 we can see the MSA divided by 2010 census tracts. If the tract do not change in size during the period 1990 to 2010, we only see red. But if we see smaller red census tracts over yellow or blue ones, it means it has changed. It is evident in some subcenters that their average size has decreased.

[27]Identified with the 50% window and the 5% significance level.

Figure 6: Subcenters in Atlanta (2010 census tracts)

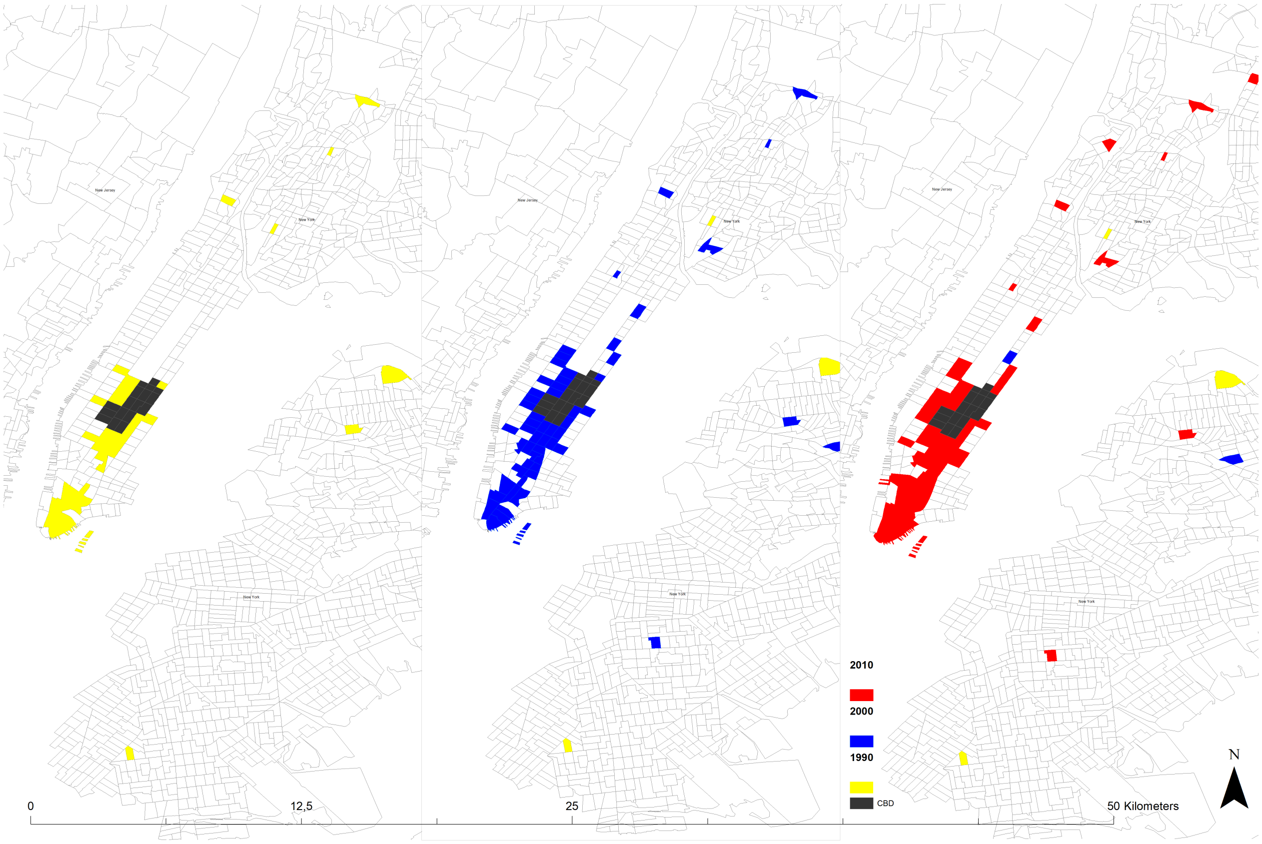


Columns 4 to 6 give information on the extension of land by subcenter. In 1990 the average subcenter included 50 km2 of land, while in 2010 30 km2. The most remarkable changes are in the South (from 67 to 34km2) and in the West (from 49 to 26km2). Recall that the lower maps by region (appendix B.) show the five MSAs with the highest concentration of land supporting these results.

Further, columns 7 to 8 of Table 13 present the sum of the jobs by census tract that are part of the subcenters. Between 1990 and 2010, the average number of jobs increased across the United States. Furthermore, the average number of jobs per subcenter in the United States is 82,510, with a clear difference in the Northeast (171,385 jobs per subcenter).

So, if we want to draw conclusions about the subcenters' intra-metropolitan characterization, we can say that they are more land-consuming in the South; however, they account for twice as many census tracts in the Northeast as the rest of the country. Furthermore, the Midwest subcenters had fewer jobs than the other regions.

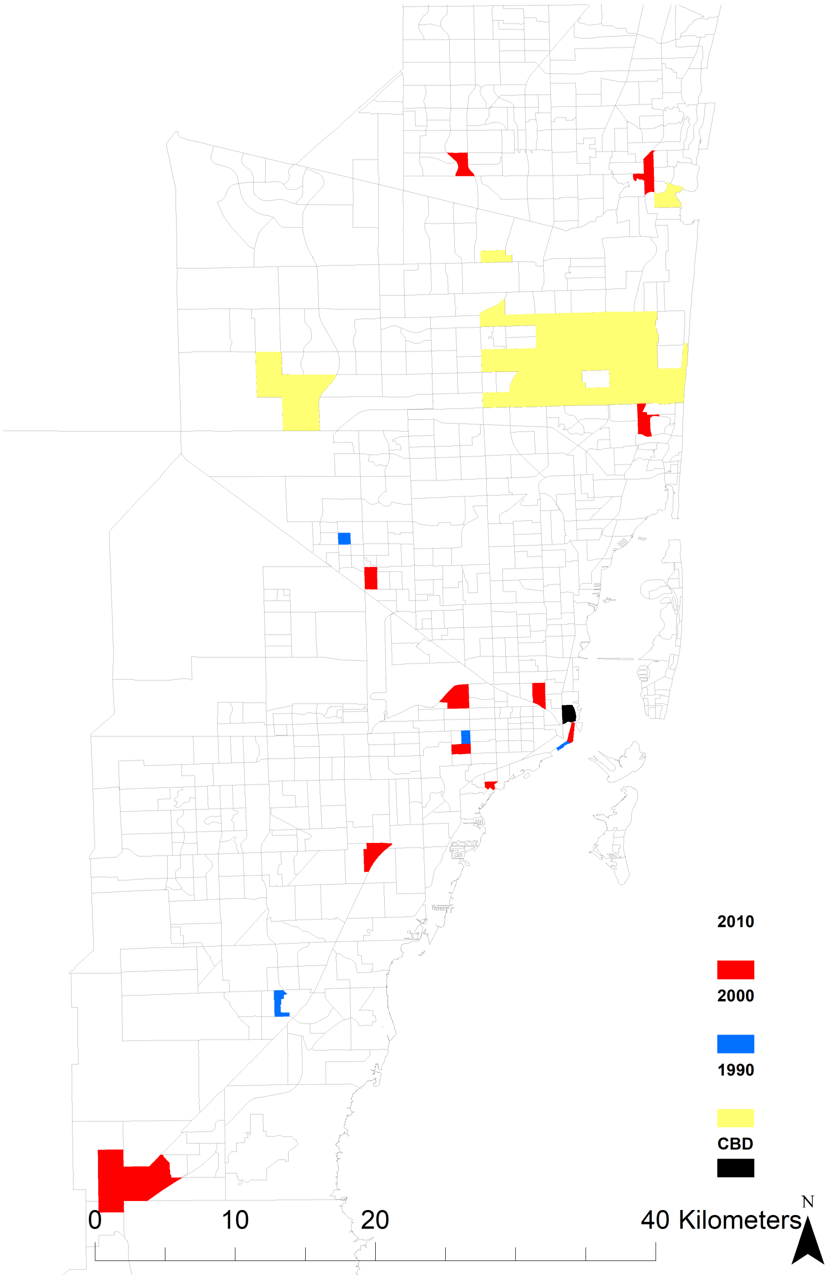
Figure 7: Subcenters in New York-Northern New Jersey-Long Island CMSA (2010 census tracts).



We arbitrarily select two more cases to illustrate intra-metropolitan patterns. In the Figure (7) from left to right are the subcenters identify for 1990, 2000, and 2010 (the CBD in black) for New york CMSA. The subcenter´s consolidation around the CBD is progressive in each decade, their expansion towards the other areas of the MSA is evident. Unlike the case of Atlanta (Figure 6), in New York, the size of the census tracts is more stable, in such a manner that 2010's map in Figure 7 allows us to visualize another important phenomenon, the origin, and disappearance of subcenters, both in 1990 (yellow) and in 2000 (blue).

Miami is another interesting case. The CMSA (which includes Fort Lauderdale) proportionally increases the number of subcenters like no other in the country, demonstrating an accelerated expansion process in the last two decades. Furthermore, the variation in the tract´s size makes the extent of land occupied by the subcenters substantially different between 1990 and 2010 (see Figure 8).

Figure 8: Subcenters in Miami-Fort Lauderdale CMSA by 2010 (census tracts).



4. Causes of urban form in the United States

In this section, we will look at the primary drivers of job decentralization in the United States. Based on previous research, particularly the approach of Fujita and Ogawa (1982). Our starting point is the classification of MSAs proposed in the previous section, which is based on the number of subcenters and their spatial distribution in 1990, 2000, and 2010. Our outcome variable is total employment in the subcenters. Furthermore, we expand the study by including a second outcome variable: the number of subcenters. We present a Poisson model for this purpose.

First, we present the data sources for the major explanatory variables and controls; second, we present our econometric strategy, which is divided into OLS and IV results for total employment and the Poisson model for the number of subcenters. We treat for the possibility of endogeneity between employment (total and number of subcenters) and population density or commuting costs (main explanatory variables) using IVs.

4.1 Data

We use transport improvement variables, which are related to the change in commuting costs, and population density as the main explanatory variables. The population data is derived from the 1990, 2000, and 2010 censuses at the census tract level. We use the Logan *et al.* (2014) database once more, which allows us to reduce inconsistencies caused by differences in the sizes of administrative units between censuses.

Regarding the measurement of commuting costs, our main variable is the highway-lane length by MSA in kilometers. Additionally, we use the number of Vehicle Traveled Kilometers (VKT) and the highway length, both by MSA. This data comes from the annual average daily traffic (AADT) and a description of the road network from the US Highway Performance and Monitoring System (HPMS) for 1983, 1993, and 2003. For the three variables, we utilize the data from Duranton and Turner (2011*a*). They use a county identifier to match every segment of interstate highway to an MSA and then calculate lane kilometers, highway kilometers, and, VKT. The advantage of using these data is that most of the highways-related infrastructure results come from federal programs in the 50's and 60's in USA, which means they are exogenous or predetermined to the period analyzed here. However, in the scenario of persistence of potential endogeneity, we will treat it with historic variables.

As controls, we include four variables that geographically characterize MSAs. Burchfield, Overman, Puga, and Turner (2006) and Duranton and Turner (2011*a*) utilized these variables. Specifically, we use *Elevation range*, which is the difference between each MSAs highest and lowest points. *Ruggeness*, which is computed by imposing a regular 90-meter grid on each MSA and calculating the mean elevational difference between each cell and adjacent cells. *Heating Degree Days (HDD)* are the number of degrees that the average temperature in a day is below 65 °F (18 °C), which is the temperature below which a building needs heating. And, we use *Cooling Degrees Days (CDD)*, which employs the same calculation but for temperatures above 65 °F. In the United

States, these measures are widely used to determine energy demand [28].

We use the industrial mix which is the distribution of employment among a city's principal productive activities. This controls for the occupational profile and partly for the socioeconomic characteristics; moreover, it is the best predictor of the degree of decentralization in the context of studying urban form (Glaeser and Kahn, 2001). We specifically use two variables: the manufacturing and the services employment shares. This information was compiled by Gilles Duranton, BeHy Wang, and Hongyu Xiao using the County Business Patterns (CBP) from the United States Census Bureau.

Furthermore, we take the average household income, the proportion of poor people, and the proportion of college-educated workers. Following the methodology of Duranton and Turner (2011*a*), who derived these variables from the decennial censuses of 1980, 1990, and 2000, we extended the analysis through the 2010 census.

Finally, to account for the long-term growth of MSAs, we control for the decadal population between 1920 and 1980, as suggested by Duranton and Turner (2011*a*). This group of variables may be associated with all of the unobservable conditions of each MSA; therefore, including them in our regressions is a good robustness test.

Table 14: Variables: main statistics by MSA (panel-data).

Variable	Obs.	Mean	Std. Dev.	Min	Max
Population	615	987,612	2,124,307	97,478	20,042,616
Highway-lane length (km)	615	1,349	1,978	1	17,187
Highway length (km)	615	267	331	0	2,622
VKT[a]	615	16,066	32,251	3.6	331,928
Elevation range (meters)	615	625	924	4	4,367
Ruggedness	615	9	11	0	78
Heating DD[b]	615	4,475	2,238	243	9,892
Cooling DD[b]	615	1,413	927	108	3,973
Share services	615	0.76	0.09	0.38	0.94
Share manufacturing	615	0.17	0.09	0.01	0.59
Mean annual income[c]	615	48,648	13,162	15,749	99,358
Share poor	615	0.14	0.04	0.05	0.38
Share college-educated	615	0.51	0.09	0.30	0.75
Share of clay[d]	615	29	12	3	73
Share aquifers[d]	615	33	38	0	100

Notes:[a] Millions of kilometers. [b] Sum of Annual Daily Degrees. [c] Average annual household income in USD. [d] As a percentage of the total subsoil of each MSA, this two variables are included in the next section as controls.

[28] Additional information about these variables can be found in Burchfield *et al.* (2006) and at diegopuga.org/data/sprawl/

4.2 OLS results

Following the existing literature (Fujita and Ogawa, 1982, McMillen, 2003), we study the role played by highway improvements and population (density) on the spatial pattern of jobs. We index MSAs by i and years by t:

$$\begin{aligned} ln(Employment_{it}) = & \beta_0 + \beta_1 \times ln(Transport_{it}) + \beta_2 \times ln(PopulationDensity_{it}) + \\ & \beta_3 \times Geography_i + \beta_4 \times Socioeconomic_{it} + \\ & \beta_5 \times Economic.Sectors + \epsilon_{it} \end{aligned} \tag{7}$$

In Table 15, we present the OLS results after maximizing Equation 7 to verify the causal relationship between the logs of total employment in the subcenters and transport infrastructure improvement (as a proxy for commuting costs) and population density employing OLS. Recall that, we initially identified 272 MSAs, but we only kept the ones that had no missing values in the variables we just mentioned and had more than 100,000 people living in them in 1990. The result is 205 MSAs per year (615 in the panel data).

Within each MSA, population density is measured as a density per km2 (we use log values). Commuting costs are measured using three distinct variables: highway-lane length (columns 1 to 6), interstate highway length (columns 7 to 9) and vehicle kilometers traveled (VKT) (columns 10 to 12). We intend to demonstrate that the results are consistent with various metrics of commuting costs and also to study two econometric approaches (a pool structure and an FE structure).

Table 15: Total employment in subcenters: OLS results

	Highway-lane length						**Highway length**			**VKT**		
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
ln (highway-lane)	1.926^a	1.761^a	1.781^a	1.551^a	1.574^a	1.061^a						
	(0.176)	(0.229)	(0.279)	(0.304)	(0.327)	(0.400)						
ln (I. highway length)							1.631^a	1.584^a	1.114^b			
							(0.347)	(0.325)	(0.493)			
ln (VKT)										1.309	1.314^a	0.929^a
										(0.265)	(0.246)	(0.335)
ln(Population density)		0.452	0.553	0.634	-0.603	-1.58	-0.101	0.932	-1.416	-0.478	0.405	-1.853
		(0.332)	(0.368)	(0.404)	(1.449)	(2.023)	(1.296)	(0.445)	(2.031)	(1.296)	(0.474)	(2.003)
Time Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Census division			Y	Y	Y		Y	Y		Y	Y	
Geography			Y	Y	Y		Y	Y		Y	Y	
Socieconomic				Y	Y	Y	Y	Y	Y	Y	Y	Y
Industrial Mix				Y	Y	Y	Y	Y	Y	Y	Y	Y
Past population					Y		Y			Y		
MSA Fixed Effects						Y			Y			Y
R2	0.168	0.171	0.185	0.202	0.208		0.229	0.2236		0.228	0.223	
N° of observations	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. VKT= Vehicle Kilometers Travel.

Regarding the log of highway-lane length results, Columns 1 through 5 use a pooled database, whereas Column 6 adds MSA fixed effects and keeps time fixed effects and time-variant variables. In Column 1, we include only the fixed effects per year, and the number of controls gradually increases until Column 5. Throughout the regressions, the coefficient lowers while remaining significant, and $R^(2)$ also increases, indicating that we are controlling correctly. Regarding the log of population density, no statistical significance is observed, and the coefficient becomes negative when past populations are added; this could be as result of the biased results due to the unsolved endogeneity between population density and total employment.

In the second section of the Table (columns 7–12), we offer two alternative variables to measure commuting costs, while maintaining the same technique. (where Columns 6 and 12 are fixed effect specifications). The results are statistically significant in every scenario. Although the log of highway length has larger coefficients, we choose to measure commuting costs using the log of highway-lane length since it provides information about the actual increase in highway capacity and not just their coverage. Our preferred specification is shown in Column 5, which indicates a 1.5 elasticity between the log of highway lane length and the log of total employment.

The log of total employment in subcenters is positively affected by the comparison between the three variables of commuting costs and the two techniques. In addition, we prefer the especification of highway-lanes length since it would account for two effects: the enhancement of accessibility levels for the initial network and the extension of the network to areas that previously lacked highways.

4.3 *Poisson Model*

We find that with a significant level of 5% and a window of 50%, the number of subcenters goes from 0 to 56 across the 272 MSAs initially studied (see Table 11). The next step is to identify the main drivers of such phenomenon. Since our response variable, the number of subcenter, is a count variable, the natural stochastic model is a Poisson point process for the event's occurrence[29] Cameron and Trivedi (2007). In fact, we present the model following their approach (as in McMillen and Smith (2003)). The Poisson distribution to account for the number of occurrences of the event has a density:

$$Pr[Y = y] = \frac{e^{-\mu} - \mu^y}{y!}, y = 0, 1, 2..., \tag{8}$$

Where μ is the intensity parameter. The first two moments of the Poisson distribution ($P[\mu]$) are $E[Y] = \mu = V[\mu]$, and typically it is assumed to follow the following exponential function for its mean:

$$\mu = exp(x_i'\beta), i = 1,...,n. \tag{9}$$

Given 8 and 9 and the fact that this model implies independence of observations, we expect a quasi-maximum likelihood estimator with function

[29] Appendix C. provides supplementary technical justification for employing this model.

$$\ln L(\beta) = \sum_{i=1}^{n} \{y_i x_i' \beta - exp(x_i'\beta) - ln y_i!\}. \tag{10}$$

The assumption of the quasi-maximum estimator relaxes the rule of the conditional variance and conditional mean being equal, as well as permitting the inclusion of randomness with the addition of the asymptotically normal error term. (Cameron and Trivedi, 2007).

In the first part of the paper, we estimate the number of subcenters for the decennial censuses of 1990, 2000, and 2010. So, we use a panel data within the Poisson framework. In a short panel with a T small and $n \to \infty$, both fixed effects and random effects models are possible (Cameron and Trivedi, 2007)

$$y_{it}|x_{it}, \beta, \alpha_i \sim \text{Poiss}[\alpha_i \exp(x_{it}'\beta)] \sim Poiss[exp(ln\alpha_i + x_{it}'\beta)] \tag{11}$$

If α_i is unobserved but is not correlated, we use random effects, and if it is unobserved and is possibly correlated with x_{it} if we use fixed effects.

We select a Poisson model to account for the number of subcenters.

Poisson model for the number of subcenters:

$$\begin{aligned} NumberSubcenters_{it} =& \beta_0 + \beta_1 \times ln(Transport_{it}) + \beta_2 \times ln(PopulationDensity_{it}) + \\ & \beta_3 \times Geography_i + \beta_4 \times Socioeconomic_{it} + \\ & \beta_5 \times Economic.Sectorsx_{it} + \epsilon_{it} \end{aligned} \tag{12}$$

Table 16: Number of subcenters: Poisson model results

	Highway-lane length					Highway length		VKT	
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
ln (Highway-lane)	1.037^a	0.914^a	0.818^a	0.770^a	0.789^a				
	(0.041)	(0.076)	(0.059)	(0.074)	(0.070)				
ln (I. Highway length)						0.801^a	0.774^a		
						(0.078)	(0.082)		
ln (VKT)								0.700^a	0.690^a
								(0.063)	(0.062)
ln(Population density)		0.255^a	0.384^a	0.319^a	0.467^b	0.587^b	0.416^a	0.216	0.131^b
		(0.084)	(0.078)	(0.081)	(0.235)	(0.256)	(0.082)	(0.249)	(0.086)
Year Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Census division			Y	Y	Y	Y	Y	Y	Y
Geography			Y	Y	Y	Y	Y	Y	Y
Socieconomic				Y	Y	Y	Y	Y	Y
Industrial Mix				Y	Y	Y	Y	Y	Y
Past population					Y	Y		Y	
N° of observations	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*615*	*612*

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. VKT= Vehicle Kilometers Travel.

Table 16, displays our findings with the Poisson model, which explores the link between the number of subcenters and the factors that, according to our hypothesis and the available literature, are the most important contributors to their occurrence. For this, we employ the same structure as Table 15: the log of population (density) and the log of commuting cots serve as the primary variables on the right-hand side.

The first difference compared to the total employment figures is that the log population (density) is significant and continues to be significant after controlling for past populations (the only exception is with VKT). This suggests a causal relationship between population density and the number of subcenters in the United States.

As for the log of highway-lane length coefficients, they range from 1.037 to 0.77. Our preferable specification is in Column 5, where we have a coefficient of 0.47 for the population density and an elasticity of 0.79 for the commuting costs variable.

4.4 Two-stage least square

The OLS regressions are most likely biased due to the risk of endogeneity. That is, an increase in highway capacity or population (density) could increase total employment or the number of subcenters. As a result, IVs will be used to search for external sources of exogeneity variation. Thus, in this section, we present the model that combines the findings from the OLS and Poisson estimates, using our two primary right-hand variables: highway-lane length and population density.

4.4.1 Instrument for highways–lane lenght

The growth in the capacity or coverage of the highways partially explains the spatial location of employment. The connection between locations is easier, generating employment by finding new opportunities for locations outside the traditional centers. However, increasing roadway capacity may also be a response to rising demand brought on by the quick increase in employment. Therefore, we must look for an exogenous source of variation in order to include highway-related variables in an equation that describes total employment in sub-centers (and the number of subcenters).

Thus, we used two instruments proposed by Duranton and Turner (2012*a*)[30] to control potential endogeneity via reverse causality. They converted the paper maps of the highway plan of 1947 and the railroad system of 1988 into digital ones and assigned a number of kilometers to each MSA. Since the planning and the majority of the implementation of these historical instruments were not done in response to the employment dynamics that we investigate here, we could assume that they are exogenous.

First, we take the *Highway Plan of 1947*. This is the most ambitious highway expansion plan in US history. Because of the historical setting in which it was planned, exogeneity is assumed. It was first proposed in 1921 and began construction in 1937 as ”a strategic highway network suggested by the War Department, the location of military establishments, inter-regional traffic

[30]for more detail, see Duranton and Turner (2012*a*)

demand, and the distribution of population and economic activity at that time". It was established in 1956 after being formally declared in 1947. It was finished in the 1980s (Duranton and Turner, 2011*a*).

Our second instrument is the *1,898 USA Railroad System*, which we believe is relevant because building railroad tracks and vehicle roads necessitates leveling and grading a roadbed. The initial assumption is that modern highways were constructed alongside train routes. To satisfy the exogeneity criteria, we can conclude that the path of the railroad lines forecasts the current highway system. It is worth noting that the railroad system connected the United States during a period when agricultural employment predominated. Furthermore, the rail network was built by private companies hoping to profit from railroad operations in the near future (Duranton and Turner, 2011*a*).

Table 17: Pooled TSLS, instruments first stage results.

Dependent variable	ln(lane kilometers)			ln(Population density)		
	[1]	[2]	[3]	[4]	[5]	[6]
ln (Highway 1947)	0.301^{a}	0.159^{a}	0.148^{a}			
	(0.039)	(0.035)	(0.031)			
ln (Rail 1898)	0.189^{b}	0.251^{a}	0.288^{a}			
	(0.094)	(0.071)	(0.078)			
ln (Expected density population)				0.974^{a}	0.969^{a}	0.906^{a}
				(0.009)	0.009	(0.039)
Time Fixed Effects	Y	Y	Y	Y	Y	Y
Census division		Y	Y		Y	Y
Geography		Y	Y		Y	Y
Socieconomic		Y	Y		Y	Y
Past population			Y			Y

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. ** We are not assuming i.i.d so we refer Kleibergen-Paap rk Wald F statistic.

4.4.2 *Instrument for total employment in subcenters.*

Finally, population density may be a result of new employment, similar to how the number of highway-lane kilometers can be an endogenous factor when predicting total employment (or the number of subcenters). As a result, we propose a Bartik-type instrument[31] that meets our need for exclusion constraint. These instruments, known colloquially as "shift-share," seek to predict growth rates by interacting lag (initial) local factors with national growth rates (such as industry composition, labor shares, local population, or employment). The exogeneity of the instrument is determined by the local composition, which is from a previous period. As a result, the Bartik instrument we use to calculate the Expected Population (EP) in 1990, 2000, and 2010 takes the following form:

[31] Although the origin of shift-share instruments is earlier, it was initially proposed by Bartik (1991) and popularized by Blanchard and Katz (1992). This type of instrument is used in various lines of economic research, such as labor economics and migration economics.

$$\mathrm{EP}_{i,t} = \sum_{k=1}^{k} \mathrm{Z}_{i,k,t-10}\mathrm{N}_{k,t}^{US} \tag{13}$$

Where t is 1990, 2000, or 2010, and i is one of the 205 MSAs being considered. Our *shift* measure is the population from the prior decade; therefore Z is the lagged population and $t-10$ is 1980, 1990, and 2000, respectively. And the *share* component is the projected increase in employment in the USA, which is stated as a decadal growth rate between 1977 and 1988, 1988 and 1996, or 1996 and 2006. With the weighting coming from the initial mix of economic activity in each MSA, this means that the expected population of an MSA in a given year t is a weighted average of how much US employment is rising. To determine the expected population density, we divide the result by the area of each MSA.

In Table 17, we report the results of the first stage. Using these findings, we demonstrate how historical instruments explain for the number of kilometers of highway-lane length after adding various controls (columns 1–3). There is also a correlation of 0.63 between the log of the 1947 highway plan and the log of highway-lane kilometers by MSA, as well as a correlation of 0.51 between the latter and the log of the 1898 network rail. We also present the results of the expected density of the population.

We present the TSLS results in Table 34. In panel A, we provide findings for total employment in subcenters, and in panel B, we present results for the number of subcenters (Poisson model)[32]. The two panels have the same layout: from Column 1 to Column 5, we only added the highway-lane length instruments (historical instruments) and progressively added the controls and fixed effects. We add the instrument to the population (density) in Column 6. In Column 7, we utilize a different set of past population variables (from 1920 to 1940) . In Column 8, the complete specification was executed without the historical population variables.

Regarding total employment in subcenters (panel A), the elasticity with highway-lane length is rather stable across all specifications. The value decreases from 1.67 to 1.36. Regarding population density, the results are higher than those presented in Table 6, and the majority of specifications are significant, indicating the relevance of the instrument. However, the significance is lost when the historical population of the MSAs is included, probably due to the correlation with total employment. In column 7, we use the oldest decades (1920–1940), and although the coefficient is still not statistically significant, it becomes positive. Our favorite specification is in Column 8, where total employment in US subcenters between 1990 and 2010 exhibited an elasticity of 1.5 with highway-lane length and 0.8 with population density. Also, the combined use of the instruments passed the weak identification test (using Stock and Yogo (2005) critical values). The number of subcenters in panel B, is explained by log kilometers of highway-lane length and population density, after testing numerous instrument specifications and combinations.

The Poisson regression coefficient can be interpreted as follows: for a one-unit change in the log of highway lane or population density, the difference in the logs of the expected number of subcenters is expected to change by 0.64 and 0.37, respectively (in our preferred specification in column 8).

[32]In the appendix A., we present robusstness tests for other windows and significance levles.

Table 18: Total employment and number of subcenter: TSLS results

Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Panel A: Total Employment (TSLS)							
ln (highway-lane)	1.667[a]	1.449[a]	1.457[a]	1.399[b]	1.366[b]	1.502[b]	1.359[a]	1.497[a]
	(0.307)	(0.367)	(0.477)	(0.519)	(0.578)	(0.545)	(0.551)	(0.501)
ln(pop. density)		0.654[c]	1.052[b]	0.861[c]	-0.229	-0.747	0.59	0.803[c]
		(0.361)	(0.491)	(0.484)	(1.261)	(1.639)	(0.718)	(0.492)
IV: Highway-lane	✓	✓	✓	✓	✓	✓	✓	✓
IV: Population						✓	✓	✓
Time Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Census division			Y	Y	Y	Y	Y	Y
Geography			Y	Y	Y	Y	Y	Y
Industrial Mix				Y	Y	Y	Y	Y
Socieconomic				Y	Y	Y	Y	Y
Past population					Y	Y[d]	Y	
Obs.	615	615	615	615	615	615	615	615
F-S. F-Statistics	58.19	43.99	33.59	36.67	34.16	22.00	21.14	23.84
	Panel B: Number of subcenters (Poisson TSLS)							
ln (highway-lane)	0.800[a]	0.697[a]	0.625[a]	0.620[a]	0.665[a]	0.686[a]	0.672[a]	0.636[a]
	(0.041)	(0.053)	(0.069)	(0.076)	(0.093)	(0.097)	(0.096)	(0.078)
ln(pop. density)		0.306[a]	0.436[a]	0.382[a]	0.41	0.22	0.428[a]	0.368[a]
		(0.07)	(0.08)	(0.08)	(0.29)	(0.39)	(0.12)	(0.09)
IV: Highway-lane	✓	✓	✓	✓	✓	✓	✓	✓
IV: Population						✓	✓	✓
Time Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Census division			Y	Y	Y	Y	Y	Y
Geography			Y	Y	Y	Y	Y	Y
Industrial Mix				Y	Y	Y	Y	Y
Socieconomic				Y	Y	Y	Y	Y
Past population					Y	Y[d]	Y	
Obs.	615	615	615	615	615	615	615	615
F-S. F-Statistics	58.19	43.99	33.59	36.67	34.16	22.00	21.14	23.84

Notes:[a], [b], and [c] indicate significance at 1, 5, and 10 percent level, respectively. First Stage F- Statistics. [d] In this specification,Past population is decennial from from 1920 to 1940 in this specification.

After controlling for a comprehensive set of variables and employing instruments to address the possibility of endogeneity due to reverse causation, we can suggest that polycentricity in the United States, measured as the number of alternative centers (subcenters) to the traditional center and as the total employment concentrated in these subcenters, can be partially explained by reductions in commuting costs between subcenters, which are the result of highway-lane expansion. Similarly, population (density) in the United States between 1990 and 2010 was a precursor to polycentricity, meaning that larger cities tend to be more polycentric.

5. Consequences of urban form in the United States

Beyond the causes of urban form, there is a need to comprehend its economic, socioeconomic, and environmental implications. As the primary concentrations of human activity, MSAs are also a significant sources of challenges that can be managed, at least in part, by proper planning.

In this section, we weigh the benefits and disadvantages of job decentralization in order to better understand whether urban form influences the various outcomes in contemporary cities. Even though compact cities are easier to manage, employment outside of the traditional center occurs naturally. Such employment spatial redesign may have a positive impact on land price pressures, commuting costs, worker productivity, and company productivity, among other factors. But what if decentralization fails to generate new centers where people can benefit from agglomeration economies, and all it creates is sprawl? In that case, there could be an increase in dependence on traditional centers, an increase in commuting costs, or a decrease in productivity. Furthermore, decentralization of employment may only attract a portion of the population, resulting in urban segregation; it may also accumulate factories with more labor-intensive activities, lower wages, and higher pollution (to name a few examples).

To that end, our econometric approach in this section characterizes cities from a complex perspective by incorporating a number of dependent variables such as per capita income, measures of income inequality, residential segregation indexes, and the amount of polluting particles and gases that contaminate the air in cities. Furthermore, we distinguish between monocentric and polycentric cities using the city categorization provided in Section 3.1.

3-km employment density

The density of employment (or population) is accepted in the literature as a metric for understanding the spatial pattern of economic activity in a city or metropolitan area. For instance, Anas *et al.* (1998) employs density gradients to investigate urban spatial structure. Such a "traditional" use of density (number of jobs per unit of land area) allows us to capture spatial concentration, but tells us nothing about the intra-metropolitan distribution of employment (Duranton and Turner, 2018). Consequently, the density may over- or underestimate the spatial concentration of employment. As an illustration for the United States, Duranton and Puga (2020*a*) state that metropolitan areas are defined on the basis of counties. However, if a metropolitan area includes counties with substantial rural portions, such a calculation will underestimate the density experienced by the majority of economic actors.

In light of the high quality of the data, we choose a more specialized measurement. Roca and Puga (2016) suggest *experienced density* as an alternative method for reducing these disparities. The objective is to determine the employment density within a specified radius. García-López and Moreno-Monroy (2016), use it in Brazilian cities to calculate the employment density in a 1-km radius around each administrative unit and then produce a mean by Urban Area (UA). We adhere to the same proposal, but modify the density calculation for the United States with a 3-km radius measurement. This makes sense as we are examining a very heterogeneous group of metropolitan areas, the 3-km radius measurement allows us to capture the density within each census unit (the mean length of census tracts is 4.7 square kilometers, while the standard deviation exceeds 6 square kilometers); therefore, when calculating the average per MSA, we consider the densities surrounding each tract regardless of the administrative division's size.

$$\text{3-km ED} = \frac{1}{n}\sum_{i}\left(\frac{\text{MSA jobs in the surrounding 3-km radius}}{\text{3.1416 sq km}}\right) \tag{14}$$

Where n is the total number of census tracts in every MSA in USA

As demonstrated by the formula 14, we calculated the employment density surrounding each of the 55k census tracts (2010 limits) for 1990, 2000, and 2010 and then averaged it for each of the defined metropolitan areas (in Section 3.1). In Table 19, we display the 3-km density and the number of subcenters (others than the CBD) with population over one million for 45 cities.

Therefore, we define the employment density within a 3 km radius of each census tract unit as our primary explanatory variable for urban form[33].

Expected Employment

We use *Expected Employment* as an instrument, to control for potential endogeneity between employment density within 3-km of each census tract and each of the variables that we will analyze as urban-form outcomes. Specific justification is given in each of the following sections. We apply a Bartik-type instrument once more, but this time to determine the Expected Employment (EE). Thus, we modify the given Equation 13:

$$\mathrm{EE}_{i,t} = \sum_{k=1}^{k} \mathrm{M}_{i,k,t-10}\mathrm{N}_{k,t}^{US} \tag{15}$$

Where, t is 1990, 2000, or 2010, and i is one of the 205 MSAs being considered. As our "shift" measure, we take employment data from a prior decade, making M equal to the lagged employment, or $t - 10 = 1980, 1990$, and 2000, respectively. And the "share" portion represents the anticipated growth in employment in the United States over the next decade, thus N represents the anticipated growth over the next decade between 1977 and 1988, 1988 and 1996, or 1996 and 2006. With the weighting being set by the baseline composition of economic activity in each MSA, this means that the predicted employment in an MSA for a given year t is a weighted average of how much employment in the United States was growing. To determine the anticipated employment density, we divide the result by the total area of each MSA.

5.1 Economic consequences

Does the physical structure of cities affect their economic performance? Does the spatial pattern of employment location produce comparative economic advantages amongst metropolitan statistical areas (MSAs) within a country? This section aims to contribute this discussion. Specifically, the purpose of this section is to determine if the spatial urban shape, as assessed by the 3-km

[33] Nonetheless, we calculate the *experienced density* at 1, 5, 7, and 10 km and present the results in the Appendix H.

Table 19: Urban spatial structure in 45 MSAs with population over a million, 1990, 2000, and 2010

MSA	1990		2000		2010	
	Sub.	**3-km**	**Sub.**	**3-km**	**Sub.**	**3-km**
Atlanta	16	2622	15	3013	21	3334
Austin-San Marcos	2	1978	2	2781	4	3223
Buffalo-Niagara Falls	0	1947	0	1724	0	1756
Charlotte-Gastonia-Rock Hill	3	1868	5	2073	7	2797
Chicago-Gary-Kenosha	12	1568	13	1493	15	1695
Cincinnati-Hamilton	4	2445	5	2337	4	2352
Cleveland-Akron	4	1375	5	1228	6	1333
Columbus	6	2107	4	2158	4	2307
Dallas-Fort Worth	13	1719	14	1970	13	2335
Denver-Boulder-Greeley	5	1673	6	1969	5	2231
Detroit-Ann Arbor-Flint	14	1389	13	1290	12	1276
Grand Rapids-Muskegon-Holland	2	2091	2	2029	1	2318
Greensboro-Winston-Salem-High P.	6	2140	4	2091	3	2595
Houston-Galveston-Brazoria	11	2121	13	2282	16	2762
Indianapolis	4	1978	2	1965	1	2084
Jacksonville	4	2139	3	2289	2	2514
Kansas City	3	1830	4	1805	5	1977
Las Vegas	3	906	2	1390	4	1885
Los Angeles-Riverside-Orange C.	29	1712	31	1517	35	1817
Memphis	2	1940	1	1850	1	2019
Miami-Fort Lauderdale	2	489	6	1843	8	2227
Milwaukee-Racine	2	1479	1	1281	2	1395
Minneapolis-St. Paul	6	2243	7	2371	6	2502
Nashville	2	2514	5	2760	7	3001
New Orleans	4	1698	2	1644	4	1406
New York-Northern New Jersey-Lo. Is.	25	1449	18	1397	26	1686
Norfolk-Virginia Beach-Newport News	5	1645	6	1667	8	1876
Oklahoma City	3	1481	3	1438	3	1764
Orlando	9	1646	7	1881	8	2486
Philadelphia-Wilmington-Atlantic City	8	2098	5	1843	5	2082
Phoenix-Mesa	7	1315	9	1685	11	1946
Pittsburgh	8	1948	5	1807	3	1914
Portland-Salem	5	1950	3	2312	3	2625
Raleigh-Durham-Chapel Hill	4	1732	4	2038	6	2763
Richmond-Petersburg	2	2470	4	2353	3	2547
Sacramento-Yolo	3	1706	4	1805	4	2191
Salt Lake City-Ogden	3	1489	3	1875	3	2317
San Antonio	5	1728	7	1899	5	2282
San Diego	6	1708	7	1677	8	1998
San Francisco-Oakland-San Jose	4	1991	6	2025	8	2615
Seattle-Tacoma-Bremerton	12	2407	12	2746	14	3107
St. Louis	7	2113	4	1998	4	2094
Tampa-St. Petersburg-Clearwater	6	1472	7	1606	8	1831
Washington-Baltimore	25	2282	24	2162	28	2641
West Palm Beach-Boca Raton	1	1200	0	1412	2	1698

Notes: Cities with population over one million in 2010. Sub.= number of subcenters other than the CBD. Identified in Section 3. 3-km= employment density in a 3 kilometers radius.

density of employment, has an independent effect on the determination of per capita income in US MSAs from 1990 to 2010. In order to accomplish this, we use panel-data setting with year and sub-regions fixed effects in the United States:

$$\ln(\text{Percapita income}_{i,t}) = \beta_0 + \beta_1 \times \ln(\text{3-km Employment density})_{i,t} \\ + \sum_g(\beta_{2,g} \times \text{Geography}_{i,g}) \\ + \sum_p(\beta_{3,p} \times \text{Past Population}_{i,p,t=1920...1950(\Delta 10)}) \\ + \sum_m(\beta_{4,m} \times \text{Industrial mix}_{i,s,t}) \\ + \sum_s(\beta_{5,s} \times \text{Socioeconomic}_{i,m}) \tag{16}$$

We obtain the per capita income data from the Census Bureau using LTDB data (Section 3). Details of the control variables are provided in the section 4.1; however, it is important to note that we exclude average household income from the socioeconomic variables (so we control for share of poor and percentage of college-educated workers).

Table 20: Urban form and economic outcomes: Per capita income

City structure	All					Mono		Poly	
Method Variables	OLS [1]	TSLS [2]	TSLS [3]	TSLS [4]	TSLS [5]	TSLS [6]	TSLS [7]	TSLS [8]	TSLS [9]
ln(3-km empl. den.)	0.194^{a} (0.062)	5.168 (4.952)	3.834 (2.446)	1.533^{a} (0.410)	1.039^{b} (0.518)	1.443^{a} (0.490)	0.617^{b} (0.343)	1.443^{a} (0.426)	1.159^{c} (0.72)
IV: Expected empl.		✓	✓	✓	✓	✓	✓	✓	✓
Years FE		Y	Y	Y	Y	Y	Y	Y	Y
Division FE				Y	Y	Y	Y	Y	Y
Geography	Y		Y	Y	Y	Y	Y	Y	Y
Past population	Y			Y	Y	Y	Y	Y	Y
Indutrial mix	Y		Y	Y	Y	Y	Y	Y	Y
Socioeconomic	Y				Y		Y		Y
N° of observations	612	612	612	612	612	163	163	446	446
F-S. F-Statistics		1.15	2.96	15.7	6.25	9.63	5.02	12.91	3.83

Notes: NOTES. ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. 3-km employment density. Expected employment. FE= fixed effects. First Stage F statistics. Mono= monocentric. Poly= polycentric.

We are concerned about endogeneity due to reverse causality in the association between urban-spatial structure and per capita income, particularly with our urban-shape variable, the 3-km employment density. For instance, regions with a greater per capita income may attract new firms and employees. In addition, a wealthier city may invest more in urban amenities that could spur the development of new employment hubs. We apply the instrumental variables (IV) technique for our primary right-hand variable. Our instrument is the Expected Employment explained in the previous section.

From Column 1 to Column 5 of Table 20, we displayed the results for all MSAs. In Column 1, the OLS results are significant and positive. The elasticity of the per capita income becomes significant from Column 2 to Column 5 and stabilizes at 1.5 in Column 4. The result with populations from previous decades (from 1920 to 1950) is displayed in Column 5, which, despite reducing the coefficient to 1.03, remains significant. The results of the 163 monocentric MSAs with and without previous populations are displayed in columns 6 and 7. And columns 8 and 9 do the same for MSAs with multiple centers. Unexpectedly, the coefficients of 6 and 8 are identical, which means that before controlling for socioeconomic characteristics there is not difference between the economic performance of a city, however, after adding that set of characteristics, polycentric cities present a coefficient 1.9 times larger.

We found that, under the conditions and controls used in this setup, employment density around employment concentration centers has a positive effect on the economic success of MSAs in the United States. Polycentric MSAs present a higher result (Column 9). All in all, urban form matters when it comes to establishing the economic performance of a city.

5.2 *Socioeconomic consequences*

Urbanization exacerbates the emergence of socioeconomic disparities; in fact, according to Monkkonen (2011), urbanization is likely the most fundamental cause of socioeconomic segregation. There are incentives for urban residents to locate in areas with characteristics similar to their own (such as income and race), resulting in the appearance of vast, homogeneous urban parts.

Understanding this dynamic is critical due to the risks associated with spatial-urban separation. For example, if there is a high-income neighborhood, it may increase interest in all types of services, such as health care (a hospital has monetary incentives to be close because patients may have private insurance, the hospital may charge higher prices, and it does not rely on government transfers). In areas with a higher concentration of low-income families, the situation is reversed (the same hospital faces several disincentives to locate there). This trend has an impact on the quality and quantity of amenities in these areas, resulting in disparities in living standards between groups living in different parts of the same MSA. Adding factors like family preferences, average age, or racial self-identification can amplify the segregation. According to Mills and Hamilton (1994), as cities grow larger, commuting distances and land value disparities grow, resulting in more differentiated neighborhoods. Furthermore, it is commonly believed that racial differences play the most important role in the study of segregation as a result of polycentrism in the United States (Monkkonen, 2011).

We intend to contribute to this line of research by studying segregation as a result of city spatial composition using our existing model of urban form and consequences. That is why we look at income segregation first, followed by residential segregation. We investigate whether income and residential segregation in the United States induce disparities in opportunities amongst families residing in the same MSA and we also produce national average results. In order to comprehend the average segregation profile of MSAs, we employ a variety of indexes.

The expression for the income and residential segregation indexes is generalized in Equation 17.

$$
\begin{aligned}
\text{Segregation Index (income and residential)}_{i,t} = & \beta_0 + \beta_1 \times \ln(\text{3-km Employment density})_{i,t} \\
& + \sum_g (\beta_{2,g} \times \text{Geography}_{i,g}) \\
& + \sum_p (\beta_{3,p} \times \text{Past Population}_{i,p,t=1920...1950(\Delta 10)}) \\
& + \sum_m (\beta_{4,m} \times \text{Industrial mix}_{i,s,t})
\end{aligned}
\tag{17}
$$

After maximizing Equation 17, we obtain the income segregation index (or residential segregation index) for the MSA i at year t. In comparison with Equation 16, we are not considering the socioeconomic controls because potential correlation with the socioeconomic indexes (dependent variables).

We are aware of the risk of endogeneity due to reverse causality between our residential and income segregation indexes and the 3-km employment density, and we employ the Bartik-type expected employment instrument once more.

5.2.1 *Income segregation and urban form*

According to Bayer, McMillan, and Rueben (2004), the distance between different socioeconomic groups increases based on their ability to pay for land and housing, their preferences for land in relation to commuting costs, and the accessibility of amenities in each location. In such a way that a family's income is a determining factor in the sorting and decision-making process for housing in one part of an MSA vs. another. Since people with comparable incomes tend to locate in close proximity, this makes rental and home purchase costs a vehicle for spatial inequality.

For this, we utilize three measures of income segregation calculated for each MSA by Logan *et al.* (2014) in the LTDB ; this will allow us to compensate for census discrepancies caused by differences in administrative limits between censuses. Hence, variables are computed at the census tract level using data from the decennial census and the American Household Survey (AHS).

The first indication of income segregation that we employ is the rank-order information theory index (H). This metric compares the difference in family income between census tracts to that of the metropolitan statistical area. It can range between a theoretical minimum of zero to a theoretical maximum of one (no segregation and complete segregation, respectively).

The second and third measures, provide information on the top and bottom of the income distribution. The segregation of poverty index (PovH-10) indicates the distance between the 10% poorest families in a metropolitan statistical area (MSA) and the remaining 90%. Homologously, the segregation of affluence (AffH-10) indicates the degree to which the 10% richest families are separated from the remaining 90%. The calculation of the indexes is based on the rank-order information theory index (H) and is described in the appendix D., along with a technical explanation based on Logan *et al.* (2014)'s study.

Table 21: Urban form and socioeconomic outcomes: Index of income segregation (H), index of the 10% poorer (PovH) and, index of the 10% more affluent (AffH).

City structure	All						Mono			Poly		
Method	OLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS
Index	H	H	H	H	PovH	AffH	H	PovH	AffH	H	PovH	AffH
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
ln(3-km empl. density)	0.002	0.852	0.702^{b}	0.121^{a}	0.082^{b}	0.188^{b}	0.751	0.262	0.128^{c}	0.122^{a}	0.089^{a}	0.183^{a}
	(0.007)	(0.888)	(0.455)	(0.041)	(0.360)	(0.360)	(0.058)	(0.577)	(0.069)	(0.049)	(0.034)	(0.061)
IV: Expected emp.		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Years FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Division FE	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y
Geography	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Past population	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y
Industrial Mix	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N° of obsservations	*615*	*615*	*615*	*615*	*615*	*615*	*163*	*163*	*163*	*452*	*452*	*452*
F-S F-Statistics		1.01	2.56	14.53	14.53	14.53	8.01	8.01	8.01	12.55	12.55	12.55

Notes: OLS results in Column 1. ln(3-km Employment Density). Instrumental Variable: Expected Employment. FE= fixed effects. First Stage F-Statistics. ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. Mono= monocentric. Poly= polycentric .

In Table 21, we present the most significant outcomes obtained by, maximizing Equation 17 for the three measures of income segregation (H, PovH-10, and AffH-10). Columns 1 through 6 contain the results of the three indexes for all of the MSAs. The H index (with all the controls) is displayed in Column 4, and the results indicate that the density of employment partially explains the income segregation in USA cities. This segregation is more pronounced in cities with the highest median income (Column 6), but it is also positive and significant in the lowest 10% of the income distribution (Column 7). In addition, there is evidence, under the proposed conditions and constraints, that as employment density rises in the monocentric cities (columns 7–9), the income top 10% of the population becomes more segregated.

Lastly, polycentric cities (columns 10–12) present significant and positive results in the three indexes; in fact, the coefficients are very similar to those of all cities, indicating that they have the most impact on the overall result. Again, the highest-income group exhibits the most segregation due to the higher employment density; the coefficient is more than twice that of the poorest 10% of the population.

In general, there is evidence of a relationship between urban form and urban income segregation. Additionally, the wealthiest groups are more segregated by income in all of the city structures we examine, which brings more services and amenities to the areas where they reside. As for the segregation of low-income individuals, there is evidence that it occurs primarily in cities with multiple centers. Nonetheless, the coefficient is twice as small as the opposite (10% wealthier), indicating that they are less concentrated and more dispersed across urban agglomerations.

5.2.2 *Residential segregation and urban form*

We use the dissimilarity index from LTDB treated by (Logan *et al.*, 2014). While describing the index, Massey and Denton (1988) stated that it calculates the weighted mean absolute deviation of each census tract's minority proportion from the MSA's minority proportion and expresses this number as a fraction of its theoretical maximum.

The dissimilarity index, ranges from one (complete segregation) to zero (fully integrated). In this instance, the minority groups will be non-Hispanic black (b), Hispanic (h), and non-Hispanic Asian (a), and the "majority" reference group will be non-Hispanic whites (w). So, the three pairs of whites and other ethnic groups are the dissimilarity indexes we use. Appendix E. contains the technical information that follows the explanation of Massey and Denton (1988).

Table 22: Urban form and socioeconomic outcomes: Dissimilarity index.

City structure	**All**						**Mono**		**Poly**	
Method	OLS	OLS	OLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS	TSLS
Demographic group	b	h	a	b	h	a	b	h	b	h
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
ln(3-km empl. density)	-2.639	-1.524	4.967^a	26.354^b	29.067^b	-1.284	34.239	39.015^b	22.416^c	27.066^c
	(3.236)	(3.009)	(2.426)	(12.659)	(14.839)	(6.225)	(25.580)	(23.156)	(12.018)	(15.940)
IV: Expected emp.				✓	✓	✓	✓	✓	✓	✓
Years FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Division FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geography	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Past population	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Indutrial mix	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
N° of observations	*615*	*615*	*615*	*615*	*615*	*615*	*163*	*163*	*452*	*452*
F-S F-Statistics				15.7	15.7	15.7	9.63	9.63	12.91	12.91

Notes: h= Hispanic, b= non-Hispanic black, and a= non-Hispanic Asians. Non-Hispanic white is the reference group. ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively.

In Table 22, we present the primary results after maximizing equation 17 with the dissimilarity index as the dependent variable. From Column 1 to 6, we present the results for all MSAs for the three demographic groups. After adding the instrument, we found that the residential inequality between the majority group and the minority groups is partially explained by the concentration of employment around the CBD and the subcenters of the MSAs in the USA, when making the comparison between whites and African Americans and whites and Hispanics. We do not find the same evidence when comparing whites and Asians. For that reason, we omitted that group from the next part of the table, where we introduced different city structures.

Columns 7 and 8 show results for monocentric cities, while columns 9 and 10 show results for polycentric cities. In general, we can say that Hispanics are highly segregated in both types of cities; however, in small cities with a single center of activity (monocentric), segregation is more pronounced. As for the African American population, we find clear evidence that in polycentric cities there is residential segregation for this population group.

Under the controls and conditions described in this study, polycentric cities are the site of residential segregation based on racial origins, particularly for Hispanic and African-American residents.

5.3 *Environmental consequences*

Now we'll look at the effects of urban form on environmental outcomes, asking whether urban form influences environmental outcomes in metropolitan statistical areas in the United States. Can the amount and concentration of polluting particles and gases be linked to the shape of a city? We do not attempt to provide definitive answers to these questions, as we do in other sections of this section. Instead, we want to contribute to the debate that is gaining traction among academics and policymakers because it has the potential to affect the lives of millions of city dwellers. Using the same identification strategy We want to generate aggregated results for the United States as well as in monocentric and polycentric cities.

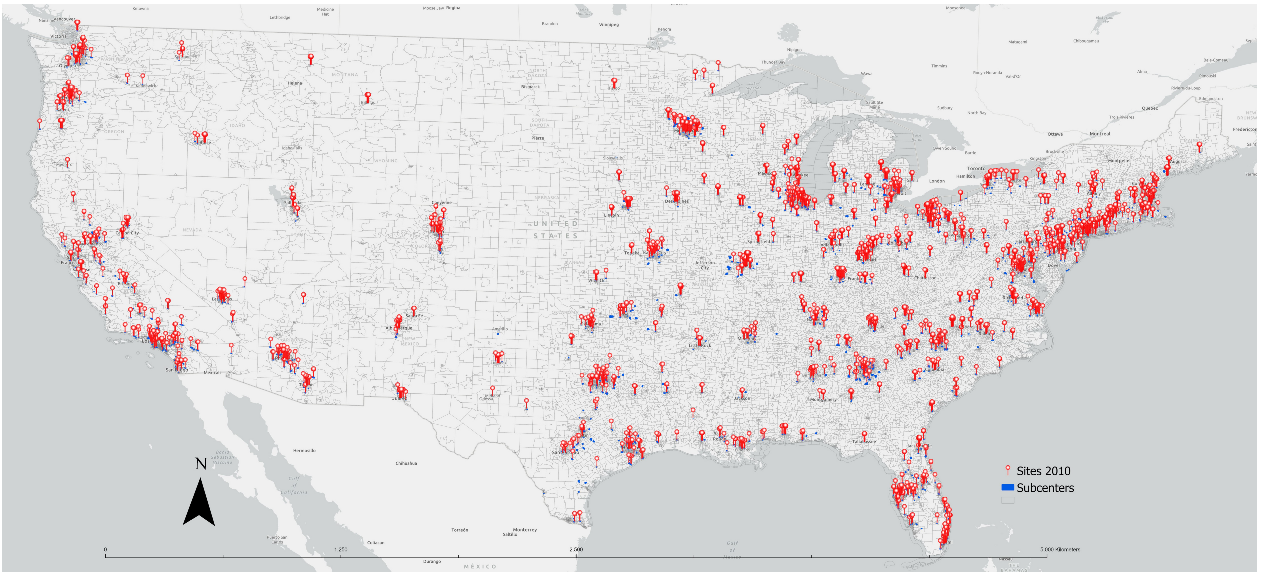
Using data from EPA (Environmental Protection Agency)[34], we initially added a time filter in such a way that we only examined sites active in 1990 and new sites (opened between 1990 and 2010). The initial number of stations is 20,779, while after the time filter there are 9,445 So that, after applying this filter, we are left with the stations that have recordings throughout the period of our interest. Next, we filter by the continental United States, eliminating Puerto Rico (78 stations), Hawaii (74 stations), and Alaska (117 stations) since, per Section 3.1, we are interested in sites near the subcenters described in Section 3.1. We maintain 42 stations in Mexico and four locations in Canada to facilitate border-MSAs measurements. Our data base contains 9,181 stations. In Figure 9 we show the distribution of the stations in 2010.

As is typical for this type of database, many of the stations had no record of the pollutant particles we sought to analyze (PM10, SO4, CO4, Ozone, and NO2), so we eliminated 4,871 sites. There are 4,310 stations that we maintain. Considering that we are working with daily values and that February 2000 was a leap year, we end up with 4,723,760 observations in this period. Lastly, we eliminate all daily records with values of zero for each of the five particles and gases of interest. Hence, our unbalanced 1990, 2000, and 2010 panel contain a total of 1,734,886 daily values.

Matching of Sites

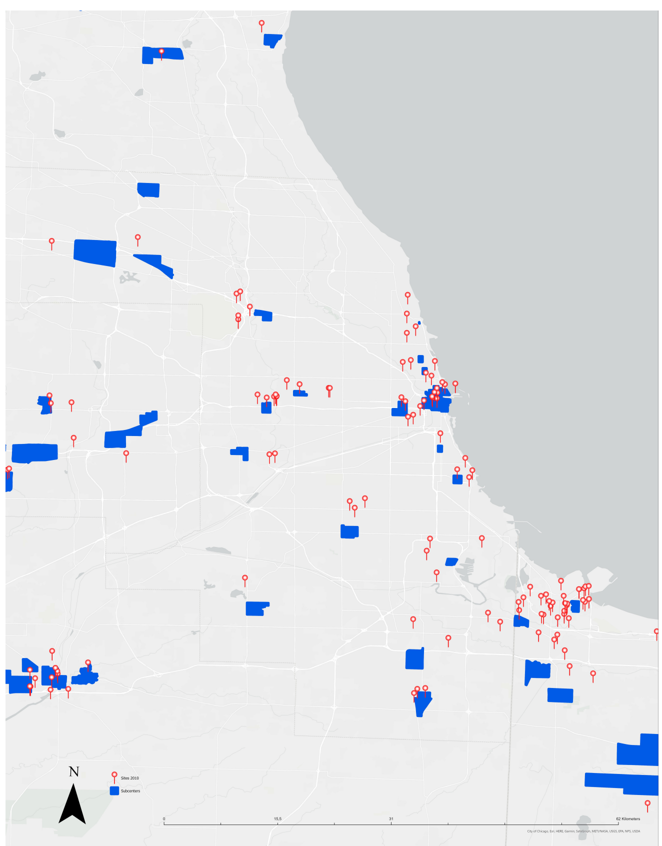
As our objective is to measure the degree of air pollution surrounding each of the subcenters specified in Section 3.1, we construct a 3-kilometer buffer to assign stations to each subcenter (thus controlling for the sites that were in urban areas). In addition to being consistent with the calculation of our urban-form measure (employment density in 3-kilometer buffers), and after testing with 1-km, 5-km, and 10-km, the selected measure did not cause considerable overlapping[35], and we excluded the sites with overlap. Also, each subcenter had at least one

Figure 9: Stations of measurement of particles on site and subcenters census tracts 2010.



Notes: Census tracts identified as part of subcenters with the methodology proposed in section 3.1. Sites: monitoring stations across continental USA in 2010.

Figure 10: Chicago-Gary-Kenosh CMSA.



Notes: Census tracts identified as part of subcenters with the methodology proposed in section 3.1. Monitoring stations withing a 3-km buffer around subcenters.

Table 23: Main statistics and sources of emission.

Pollutant	Unit of measure	Obs.	Mean	Std. Dev.	Min	Max
Carbon Monoxide (CO)	parts per million	1,734,886	0.20	0.49	0.00	12.40
Sulfur Dioxide (SO_2)	parts per billion	1,734,886	1.77	4.83	0.00	399.26
Nitrogen Dioxide (NO_2)	parts per billion	1,734,886	3.20	8.08	0.00	231.17
Ozone	parts per million	1,734,886	0.02	0.02	0.00	0.15
Particle Matter PM10	μ/m3	1,734,886	5.31	13.83	0.00	497.00

Notes:μ/m3= milligrams per cubic meter.

allocated station within 3 kilometers.

In Table 23 we present the main statistics of the final data-set. Regarding the main sources of emission we follow the World Health Organization (WHO) and Air Quality Index agency (AQI) concept. CO comes from vehicle exhaust, which accounts for approximately 75 percent of all emissions countrywide and up to 95 percent in cities. The highest levels of SO_2 are typically found near large industrial complexes and power plants, refineries, and industrial boilers are major suppliers of it. NO_2 emissions are mostly produced by automobiles, trucks, buses, power plants, diesel-powered heavy construction equipment, other mobile engines, and industrial boilers.

When primary sources of pollutants (cars, power cars, refineries, industrial plants, etc.) react chemically with sunlight, *badozone* (or simply ozone) is produced, which increases during warm months. Finally, $PM10$ are emitted by direct sources of pollution or as a reaction of various pollutants emitted to the atmosphere.

The concentration of particles and dangerous gases is closely correlated with human activity; contaminating particles are present in areas where people dwell and are directly linked to cardiovascular and pulmonary disorders (WHO). Moreover, according to AQI, these effects have been linked to both short-term (typically over 24 hours, but possibly as short as one hour) and long-term (generally over many months) exposures (years). In appendix F., we expand on the notions and selection method for EPA database contaminants.

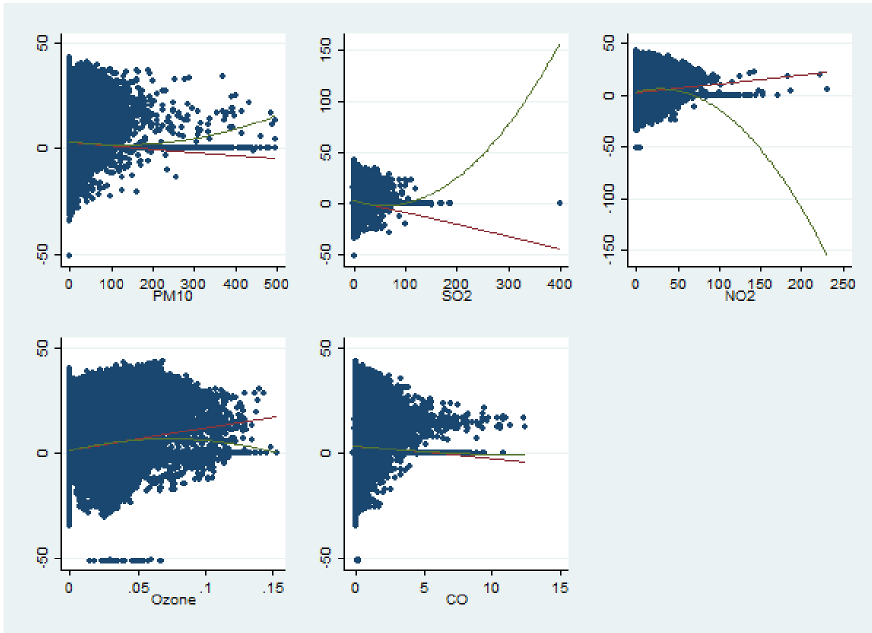
Following the analysis of urban density and air quality in Germany by Borck and Schrauth (2021), we employ a two-step methodology that aids the treatment of harmful gases and particles in the present study.

$$\begin{aligned}\text{Pollutant}_{i,t} =& \beta_0 + \beta_1 \times (\text{Barometric pressure})_{i,t} + \beta_2 \times (\text{Relative Humidity}_{i,t}) + \\ & \beta_3 \times (\text{Dew Point}_{i,t}) + \beta_4 \times (\text{Temperature}_{i,t}) + \beta_5 \times (\text{Wind-Speed}_{i,t}) + \\ & \beta_6 \times (\text{Temperature}_{i,t})^2 + \beta_7 \times (\text{Temperature}_{i,t})^3 + \beta_8 \times (\text{Wind-Speed}_{i,t})^2 + \\ & \beta_9 \times (\text{Temperature} \times \text{Wind}_{i,t}) + \mu_{month} + \lambda_{day} + \epsilon_{i,t}\end{aligned} \tag{18}$$

[34]Official entity in the United States responsible for aggregating the results of the sites.

[35]There is overlap between MSAs if a site is within three kilometers of the subcenters of two different MSAs. We identified four overlaps in 1990 and three in 2000.

Figure 11: Pollutants interaction with Temperature



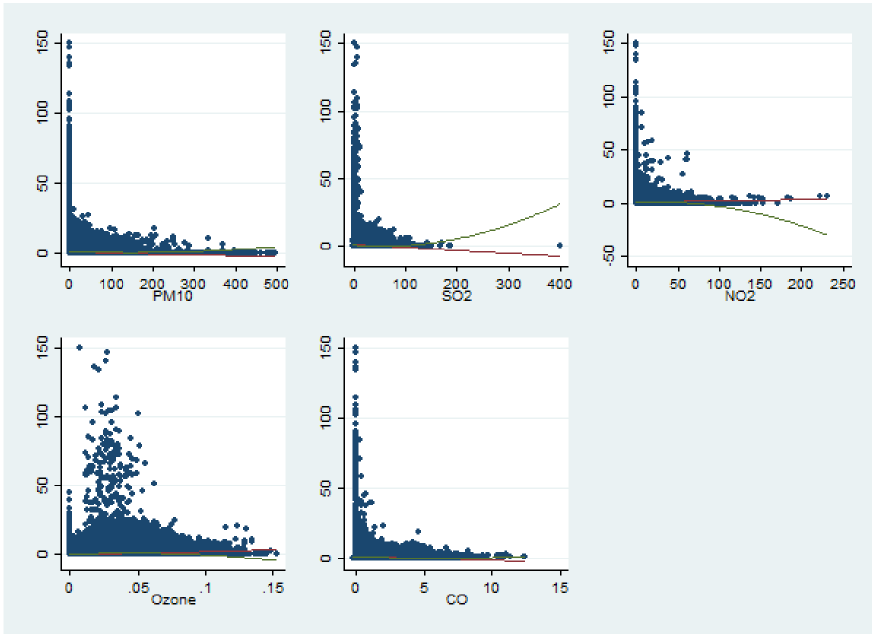
Notes: The red line represents a linear relationship. The green line represents a quadratic relationship.

In the first phase, we will discount from the panel daily data of contaminants, meteorological conditions, and time controls. We performed a spatial join with each of the sites for the years 1990, 2000, and 2010 using EPA data. Thus, we maximized Equation 18 where pollutant can be CO, SO_2, NO_2, $ozone$ or $PM10$; i represents each monitoring station (site); and t represents each day in 1990, 2000, and 2010; μ_{month} and λ_{day} are months and days of the week fixed effects; Finally, ϵ equals the daily residuals, which will be utilized in the next step.

As shown in Equation 18, we add daily values of barometric pressure, relative humidity, dew point, wind speed, and temperature, as well as the squared values of temperature and wind speed and the cubed value of temperature to account for nonlinear interactions (see graphs 11 and 12). Furthermore, Stone, Auffhammer, Carey, Hansen, Huggel, Cramer, Lobell, Molau, Solow, Tibig, and Yohe (2013) and Borck and Schrauth (2021) account for the relationship between wind and temperature; they all emphasize the importance of incorporating all climatic variables when calculating the concentration of contaminants in the atmosphere. In fact, because weather variables are intrinsically correlated over time and space, Stone *et al.* (2013) suggests that all possible weather variables be included in the regression. On warm days, for example, ozone is more likely to condense, whereas particles may be washed out of the atmosphere on rainy days. Appendix G. contains a more detailed description of the variables. Finally, because pollutants are closely linked to human activity cycles in the short and medium term, we include fixed effects of time by weekday and month.

With the residuals arising from the first part of the technique, we will apply our consequences chapter setup. Hence, we begin by grouping the sites by yearly average to produce an annual value for each MSA.

Figure 12: Pollutants interaction with Wind-speed



Notes: The red line represents a linear relationship. The green line represents a quadratic relationship.

$$
\begin{aligned}
\text{Pollutant Residuals}(\epsilon_{i,t}) = & \beta_0 + \beta_1 \times \ln(\text{3-km Employment density})_{i,t} \\
& + \sum_g (\beta_{2,g} \times \text{Soil Characteristics}_{i,g}) \\
& + \sum_p (\beta_{3,p} \times \text{Past Population}_{i,p,t=1920...1980(\Delta 10)}) \\
& + \sum_m (\beta_{4,m} \times \text{Industrial mix}_{i,s,t})
\end{aligned} \tag{19}
$$

ϵ are the residual of each of the five pollutants we are studying; i represents each MSA, and t represents 1990, 2000, or 2010. The controls are outlined in Section 4.1. We also include two variables for soil characteristics: the percentage of clay concentration and the percentage of each MSA's territory that is on subsurface aquifers since we consider that geological characteristics could play an important role in an environmental setting by MSA.

We are concerned about the endogeneity caused by the reverse causality between the employment density at 3 kilometers and the amount of polluting particles and gases. The high concentration of pollutants in the dense and monocentric cities of the early 20th century in the United States may have prompted the population to relocate beyond the typical urban core, away from the contamination of the CBD. Ultimately, the next cluster reached levels of CBD contamination, and the cycle repeated. Therefore, we utilize expected employment (explained in the first part of this section) as an exogenous variable for employment density.

The main findings regarding the relationship between urban form and environmental outcomes are presented in Table 24. From Column 1 to Column 3, all MSA results are listed. The higher concentration of employment near employment centers (including CBDs) results in greater exposure to total pollutant particles and polluting emissions in American cities.

Table 24: Urban form and environmental outcomes: polluting gases and particles.

City structure	All			Mono			Poly		
Pollutant Variables	PM10 [1]	SO2 [2]	Ozone [3]	PM10 [4]	SO2 [5]	Ozone [6]	PM10 [7]	SO2 [8]	Ozone [9]
ln(3-km empl. density)	2.049^c (1.061)	3.02^b (1.230)	1.862^a (0.708)	4.851 4.693	6.572 (8.364)	4.093^c (2.422)	1.919^c (1.102)	2.659^b (1.282)	1.909^b (0.761)
IV: Expected emp.	✓	✓	✓	✓	✓	✓	✓	✓	✓
Subsoil	Y	Y	Y	Y	Y	Y	Y	Y	Y
Years FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Division FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geology	Y	Y	Y	Y	Y	Y	Y	Y	Y
Past population	Y	Y	Y	Y	Y	Y	Y	Y	Y
Indutrial mix	Y	Y	Y	Y	Y	Y	Y	Y	Y
N° of observations	*609*	*570*	*598*	*163*	*158*	*161*	*446*	*412*	*437*
F-Stage F-Statistics	50.3	45.85	49.53	5.22	4.19	5.97	36.4	33.22	35.77

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. Monocentric and polycentric cities classified according with the subcenters identification procedure in section 3.1. CO_2 and NO_2 do not generate significant values in any scenario.

The results from columns 4 to 6 are for monocentric cities, while the results from columns 7 to 9 are for polycentric cities. We find a positive correlation between $PM10$ particles and employment density in polycentric cities. Even though this is less than the national average, the outcome is consistent, and the instrument appears to be functioning properly. Keep in mind that present particles can have multiple sources, so this indicator includes smaller particles ($PM2.5$) as well as the accumulation and reaction of other pollutants that are dispersed in the air and reach these levels.

The results for SO_2 are similar; the outcome is positive and significant in polycentric cities, but slightly lower than the average for all cities. Given that industries are the largest source of SO_2 emissions, our evidence suggests that in dispersed urban areas, exposure is not reduced but rather redistributed to sub-employment centers.

The concentration of *ozone* in monocentric cities is more than twice as high as in polycentric cities. As a result, polycentric cities generate less ozone exposure[36]; however, the levels remain statistically significant in relation to employment density. Remember that *ozone* is related to various sources of pollution, such as automobiles, buses, and industries, and that while it increases during the warmer months, the result is still positive after accounting for time-fixed effects and climatological conditions (including temperature). Polycentric cities, overall, are associated with lower concentration of particles and gases, which is in line with Castells-Quintana, Dienesch, and Krause (2021).

[36] Since we are defining urban form as the concentration of employment around a subcenter, we (arguably) define exposure as the amount of pollutant gases and particles that people in such a 3-km buffer is vulnerable to.

6. Conclusions

In conclusion, our study explores the causes and consequences of employment decentralization in the United States from 1990 to 2010. We identify non-traditional CBD subcenters in metropolitan statistical areas, with 89 out of 272 initial areas not having alternative employment centers in 2010. Utilizing McMillen (2001)'s nonparametric methodology and traditional employment density functions, we derive regional results, revealing the West as the most decentralized region, followed by the Midwest.

We validate the significance of population density and commuting costs in explaining the number of subcenters and total employment in these centers. Our findings indicate an elasticity of 1.5 between total employment and highway lane length, and 0.80 with population density. Subcenters emerge due to agglomeration forces (population density) and dispersion forces (congestion and commuting costs), aligning with previous research.

Our study generates comprehensive insights into the potential consequences of employment dispersion in the United States, offering valuable guidance for policymakers in city planning. The methodologies employed are applicable to diverse urban and regional scenarios. We analyze economic, socioeconomic, and environmental outcomes to provide a comprehensive understanding.

In terms of economic outcomes, we find that higher employment density yields better results for cities across the sample. Additionally, evidence suggests that polycentric cities tend to be more successful than monocentric ones, particularly after controlling for socioeconomic factors.

Regarding socioeconomic implications, we observe income segregation in polycentric cities at both ends of the income distribution, with the wealthiest 10% being segregated regardless of city structure. Hispanic and African-American populations experience racial segregation, but polycentrism reduces segregation within the Hispanic group.

Finally, we investigate the environmental effects, finding a weaker correlation between particle concentration (PM10) and employment density in polycentric cities than the national average. Moreover, ozone concentration is twice as high in monocentric cities compared to polycentric ones, indicating that employment decentralization reduces exposure to polluting emissions.

In light of these findings, policymakers should carefully consider the outcomes presented here while planning cities. The results offer valuable insights into the economic, social, and environmental impacts of employment decentralization, providing a basis for informed policy decisions aimed at creating sustainable, prosperous, and inclusive urban environments.

Appendix A.

A. Windows and significance level, robustness tests.

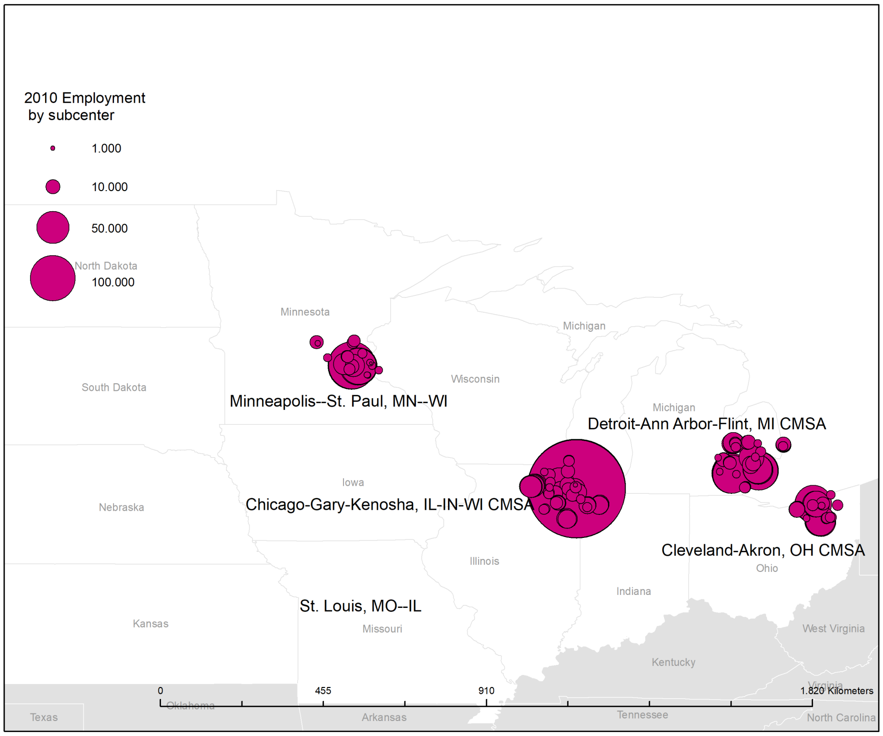
Table 25: Total employment and number of subcenters (TSLS) results: robustness tests with different windows and critical thresholds (significance levels).

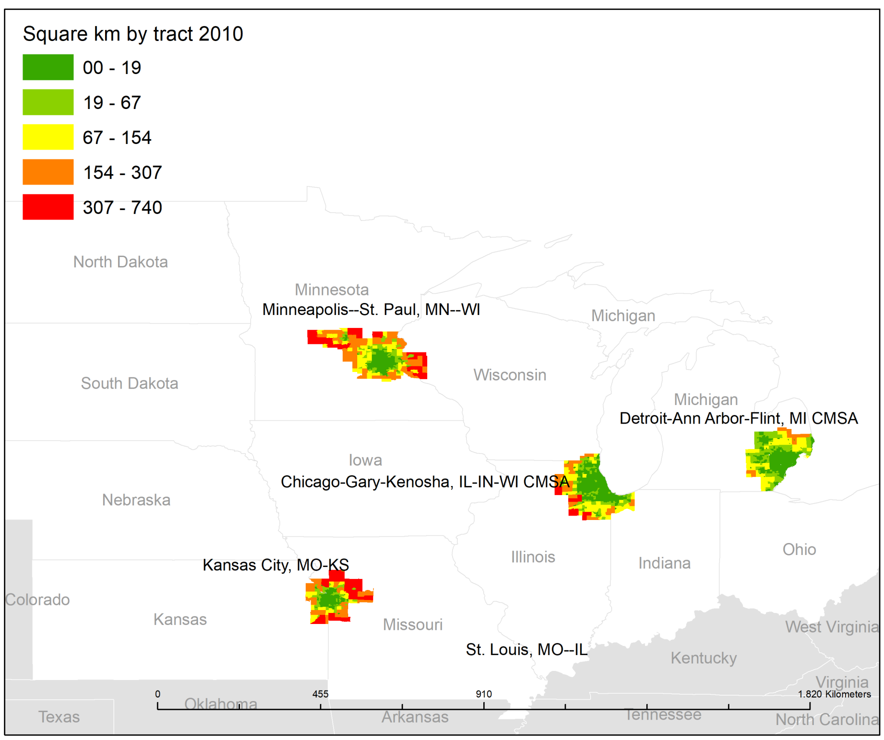
	Panel A: Total Employment (TSLS)					Panel B: Number of subcenters (Poisson TSLS)				
Window (%)	10	30	70	90	50	10	30	70	90	50
Sig. level (%)	10	10	10	10	5	10	10	10	10	5
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
ln (highway-lane)	3.285^a	1.705^a	1.178^a	0.823^a	2.146^a	2.319^a	0.991^a	0.774^a	0.683^a	0.918^a
	(0.36)	(0.40)	(0.20)	(0.24)	(0.45)	(0.32)	(0.95)	(0.71)	(0.08)	(0.10)
ln (pop. density)	2.076^a	1.585^a	1.043^a	1.48^a	1.493^a	1.149^a	0.463^a	0.412^a	0.491^a	0.474^a
	(0.39)	(0.42)	(0.20)	(0.30)	(0.43)	(0.26)	(0.13)	(0.97)	(0.10)	(0.11)
IV: Highway-lane	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IV: Population	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Census division	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geography	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industrial Mix	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Socioeconomic	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	615	615	615	615	615	615	615	615	615	615

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. IV= instrumental variable. Sig. level= critical thresholds (significance level): 1.96 (5%) and 1.64 (10%).

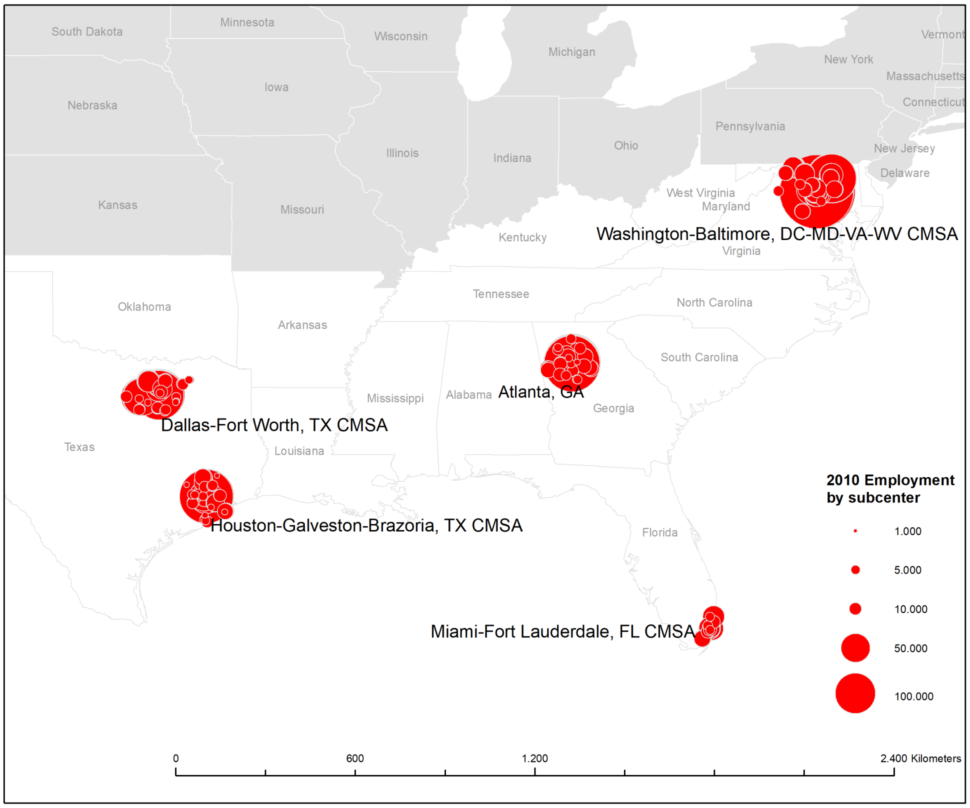
B. *Regions USA: top five MSAs employment centers and top five land-consuming sub centers*

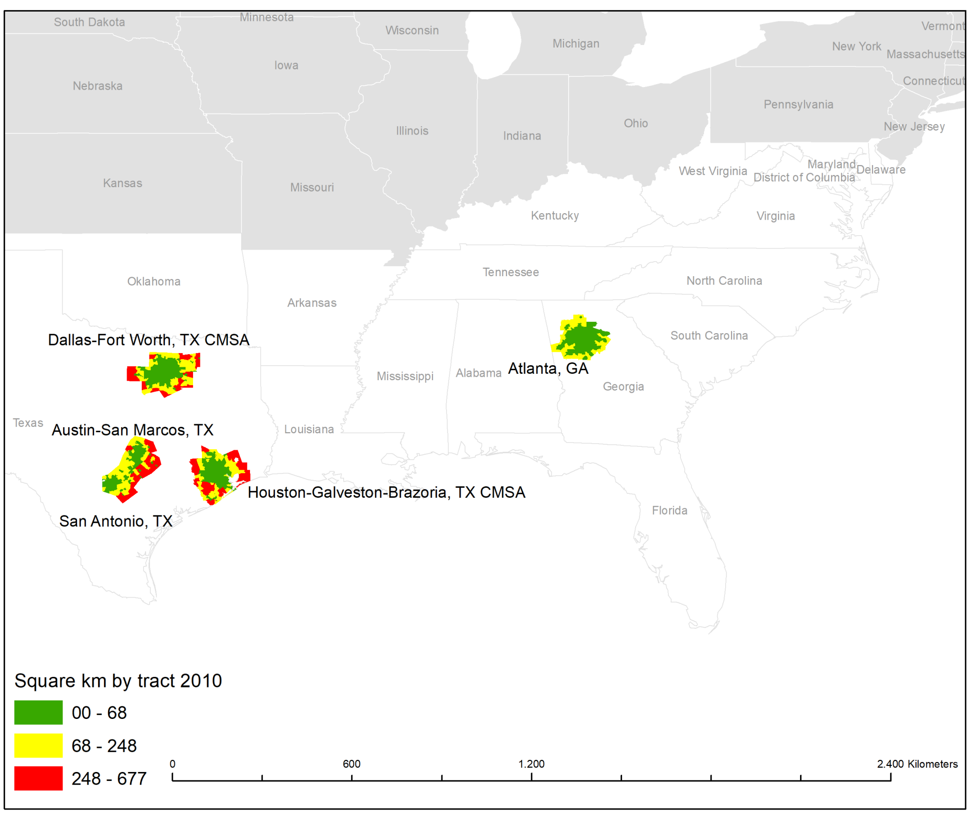
Midwest

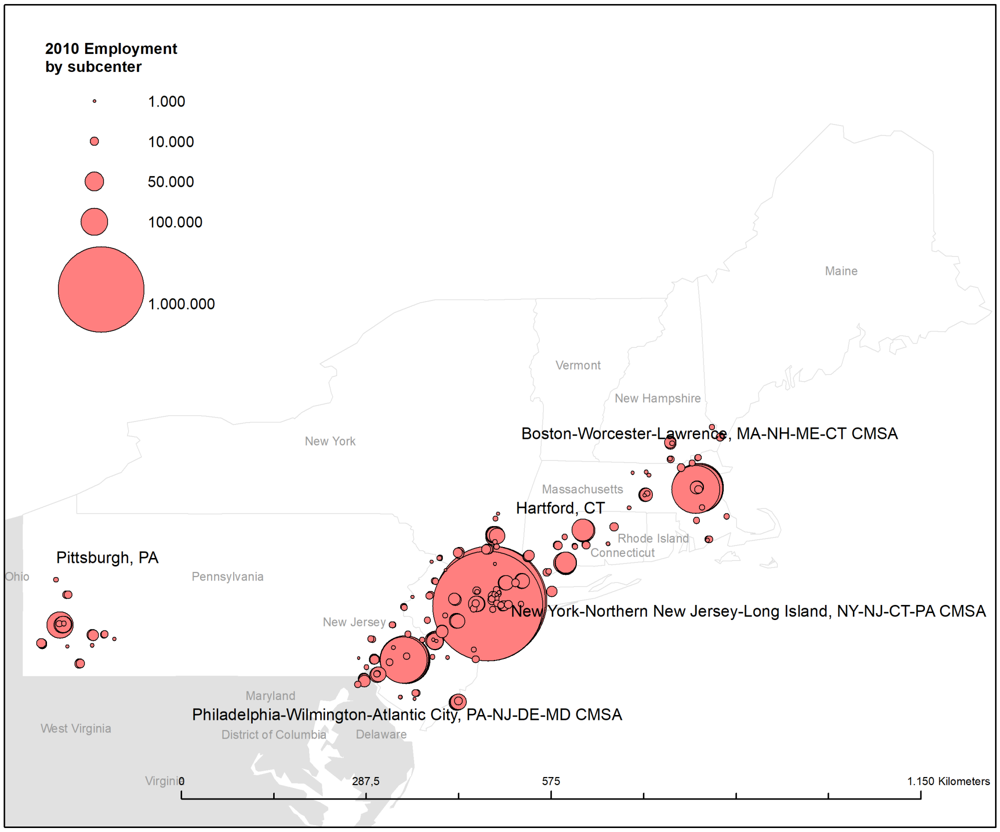




South


South Dakota
Minnesota
Wisconsin
Michigan
Vermont
New York
Massachusetts
Connecticut
Nebraska
Iowa
Illinois
Indiana
Ohio
Pennsylvania
New Jersey
Delaware
Kansas
Missouri
West Virginia
Maryland
Kentucky
Washington-Baltimore, DC-MD-VA-WV CMSA
Virginia
Oklahoma
Arkansas
Tennessee
North Carolina
South Carolina
Mississippi
Alabama
Atlanta, GA
Georgia
Dallas-Fort Worth, TX CMSA
Texas
Louisiana
Houston-Galveston-Brazoria, TX CMSA
Florida
Miami-Fort Lauderdale, FL CMSA
2010 Employment
by subcenter
1.000
5.000
10.000
50.000
100.000
0
600
1.200
2.400 Kilometers


South Dakota
Minnesota
Wisconsin
Michigan
Vermont
New York
Massachusetts
Connecticut
Nebraska
Iowa
Illinois
Indiana
Ohio
Pennsylvania
New Jersey
Kansas
Missouri
West Virginia
Maryland
District of Columbia
Delaware
Kentucky
Virginia
Oklahoma
Tennessee
North Carolina
Arkansas
South Carolina
Dallas-Fort Worth, TX CMSA
Atlanta, GA
Mississippi
Alabama
Georgia
Texas
Austin-San Marcos, TX
Louisiana
Houston-Galveston-Brazoria, TX CMSA
San Antonio, TX
Florida
Square km by tract 2010
00 - 68
68 - 248
248 - 677
0
600
1.200
2.400 Kilometers


2010 Employment
by subcenter
1.000
10.000
50.000
100.000
1.000.000
Maine
Vermont
New Hampshire
Boston-Worcester-Lawrence, MA-NH-ME-CT CMSA
New York
Massachusetts
Hartford, CT
Rhode Island
Connecticut
Pittsburgh, PA
Ohio
Pennsylvania
New Jersey
New York-Northern New Jersey-Long Island, NY-NJ-CT-PA CMSA
Maryland
Philadelphia-Wilmington-Atlantic City, PA-NJ-DE-MD CMSA
West Virginia
District of Columbia
Delaware
Virginia
0
287,5
575
1.150 Kilometers

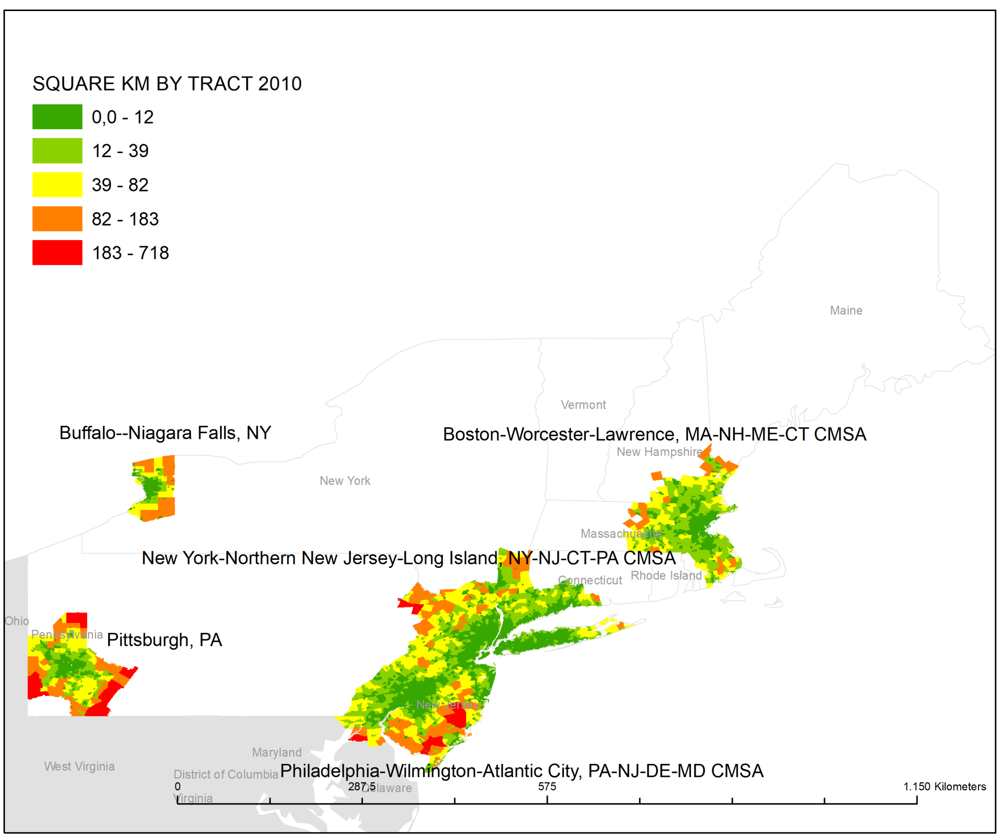
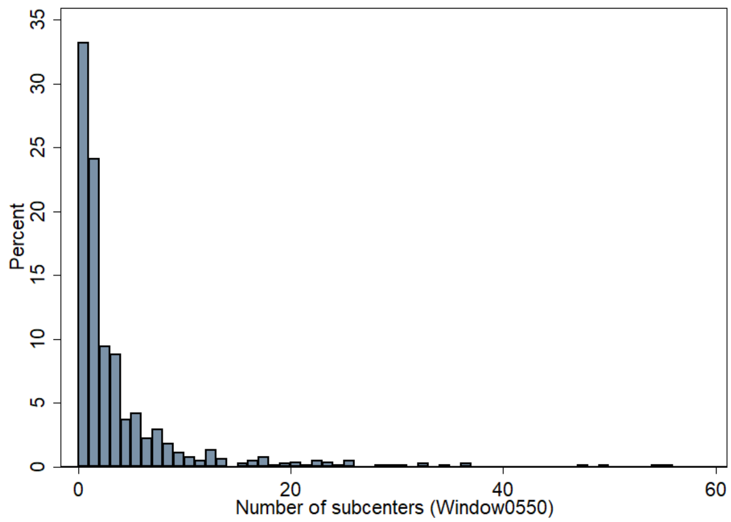

SQUARE KM BY TRACT 2010
0,0 - 12
12 - 39
39 - 82
82 - 183
183 - 718
Maine
Vermont
Buffalo--Niagara Falls, NY
Boston-Worcester-Lawrence, MA-NH-ME-CT CMSA
New Hampshire
New York
Massachusetts
New York-Northern New Jersey-Long Island, NY-NJ-CT-PA CMSA
Connecticut
Rhode Island
Ohio
Pennsylvania
Pittsburgh, PA
West Virginia
Maryland
District of Columbia
Philadelphia-Wilmington-Atlantic City, PA-NJ-DE-MD CMSA
0
287,5
575
1.150 Kilometers

Figure 13: Number of subcenters distribution



C. *Poisson model*

Working with count data brings several challenges. The most common ones are possible overdispersion and the excess of zeros. In both cases, the model's selection could help us reduce the potential loss on the prediction power of our proposal.

Fig 13 shows the distribution of the number of subcenters with a considerable zeros concentration in the left tail, accounting for 33% of the total. That is the share of monocentric cities we find. Literature (Silva, Tenreyro, and Windmeijer, 2015, Desmarais and Harden, 2013, Cameron and Trivedi, 2007) suggest a threshold of around 40% to consider a database zero-inflated. However, it is also essential to understand if the zero and the positive count numbers come from the same generator process or respond to different dynamics.

Thus, we run a Zero-inflated Poisson regression with corrected Vuong tests [37]. The Akaike test rejects the *Ho* of the Zero Inflated Model over a Poisson model; besides, the process that leads to the emergence of subcenters is the same (the lack of it) that keeps some cities as monocentric.

Overdispersion has similar consequences to the failure of the assumption of homoskedasticity. With heteroskedasticity, the variance of the error term is not constant, as with overdispersion. We approach the potential overdispertion with a negative binomial model with mean= $\mu(X)$ and Variance= $(1+\alpha)\mu(x)$. Let's recall that the Poisson model's first two moments are equal $E[Y] = \mu = V[\mu]$; $\mu = exp(x_i'\beta)$ and if $\alpha \sim \infty$ in 8 the negative binomial model becomes an identical Poisson. For the above, with the inclusion of α we are able to control for overdispersion. In addition, we use robust standard errors to control the incorrect stochastic process generated by the overdispersion. King and Roberts (2015) show that if the process generator of data is far from Poisson ($E[Y] = \mu = V[\mu]$), robust standard errors are more efficient which is our case since the mean of the number of subcenters is 3.32, and the standard deviation 6.26.

[37]Desmarais and Harden (2013) present a correction of the traditional Vuong test (Vuong, 1989) to select models. They argue it is biased towards zero-inflated models and propose a correction based on the Akaike and Bayesian (Schwarz) information criteria. We run such correction.

D. Income segregation index (H)

For any given value of p, we can dichotomize the income distribution at p and compute the residential (pairwise) segregation between those with income ranks less than p and those with income ranks greater than or equal to p. Let $H(p)$ denote the value of the traditional information theory index (H) of segregation computed between the two groups so defined. Likewise, let $E(p)$ denote the entropy of the population when divided into these two groups (Theil and Finizza, 1971). That is,

$$E(p) = plog_2\frac{1}{p} + (1-p)log_2\frac{1}{(1-p)} \tag{A 1}$$

and

$$H(p) = 1 - \sum_j \frac{t_j E_j(p)}{TE(p)}, \tag{A 2}$$

Where T is the population of the metropolitan area and t_j is the population of neighborhood j. Then the rank-order information theory index (H^R) can be written as

$$H^R = 2ln(2)\int_0^1 E(p)H(p)dp \tag{A 3}$$

We obtain the rank-order information theory index by calculating the segregation between families with incomes above and below each point in the income distribution, averaging these segregation values, and giving the greatest weight to the segregation between families with incomes above and below the median. The rank-order information theory index ranges from a minimum of 0, obtained in the case of no income segregation (when the income distribution in each local environment (e.g. census tract) mirrors that of the region as a whole) to a maximum of 1, obtained in the case of complete income segregation (when there is no income variation in any local environment). Estimates of income segregation at points in the income distribution can be obtained by estimating the function $H(p)$ to provide a measure of segregation at any threshold. To calculate the level of income segregation between families above and below the 90th percentile of the income distribution (H90), for instance, we calculate H(0.9) using the estimated parameters of the function H. (p). Similarly, to calculate the level of income segregation between families above and below the 10th percentile of the income distribution (H10), we compute H 0.1 using the estimated parameters of the function H. (p).

E. Dissimilarity index

Following Massey and Denton (1988), we adapt the dissimilarity formula to our econometric setting,

$$D = \sum_{i=1}^{n} \left[\frac{t_i |p_i - P|}{2TP(1-P)} \right] \tag{A 4}$$

Where t_i and p_i represent the total population and minority percentage of census tract I T and P represent the population and minority percentage of MSA I, and T and P represent the population size and minority percentage (non-Hispanic blacks, non-Hispanic Asians, and Hispanics) of the entire MSA, which is subdivided into census tracts.

This dissimilarity index is produced from the Lorenz curve, which curves the cumulative proportion of minority group X versus the cumulative proportion of majority group Y across census tracts, which are ordered from lowest to highest proportion. D represents the maximum vertical distances between this curve and the diagonal line of evenness. The dissimilarity index is strongly affected by random departures from evenness when the number of minority members is small relative to the number of census tracts (Massey and Denton, 1988), and it is highly sensitive to the redistribution of minority members across aerial units with minority proportions above or below the city minority proportion (James and Taeuber, 1985). Only transfers of minority members from areas in which they are over represented (above the MSA´s minority proportion) to areas in which they are underrepresented (below the minority proportion) have an effect on segregation as measured by the dissimilarity index.

F. Polluting particles and gases.

In this appendix, we provide a more in-depth explanation of the conceptual definition of polluting particles and gases in accordance with the Air Quality Index (AQI) and the World Health Organization (WHO). We also provide specifics regarding the processing of EPA data.

Particulate matter that contributes to air pollution.

As a result of human and natural activities, the atmosphere contains masses of a mixture of solids and liquid droplets that are polluting. According to the World Health Organization, human sources include combustion engines (both diesel and gasoline), solid-fuel (coal, lignite, heavy oil and biomass) combustion for energy production in households and industry, and other industrial activities (building, mining, manufacture of cement, ceramic and bricks, and smelting), as well as pavement erosion and tire wear. Agriculture is the primary ammonium source. The natural sources are related to soil and dust re-suspension, which is a significant source of PM, especially in arid regions or during episodes of long-distance dust transport, such as from the Sahara to southern Europe (WHO).

PM refers to the mass concentration of particles with a diameter of less than 10 micrometers (PM10) and less than 2.5 micrometers (PM2.5). PM2.5, often known as fine PM, consists of ultra fine particles with a diameter of less than 0.1 m.

The database contains measurements of PM 2.5 and Pm10 particles, but the monitoring of PM 2.5 is relatively new. In 1990, we identified only 47 stations that specifically calculated them. We are thus left with PM10. We eliminate negative values (40 daily results) and outliers above 500 μ/m3 (milligrams per cubic meters). These are particulate matter concentrations which only occur when there is a huge fire or another atypical cause of high pollution (such as New Year's Eve fireworks) (Borck and Schrauth, 2021).

Gases

Carbon monoxide (CO)

The gas carbon monoxide is odorless and colorless. It is produced when the carbon in fuels is incompletely burned. Automobile exhaust accounts for around 75% of all carbon monoxide emissions in the United States, and up to 95% in urban areas. Additional sources include industrial fuel combustion and natural causes such as wildfires. Typically, carbon monoxide levels are higher during cold weather because cold temperatures reduce the efficiency of combustion and generate inversions that trap pollutants near to the ground. In this study, we used an 8-hour sample to find the average ending time of a run. The value is the hourly arithmetic mean. The unit of measurement is parts per million for carbon monoxide.

SO2 (sulfur dioxide)

When sulfur-containing fuels such as coal and oil are burned, sulfur dioxide, a colorless, reactive gas, is released. The largest concentrations of sulfur dioxide are typically found near large industrial complexes. Power plants, refineries, and industrial boilers are major suppliers. In this investigation, the average sample duration was three hours. The value is the hourly arithmetic mean. The unit of measurement is parts per billion of sulfur dioxide.

NO2 (nitrogen dioxide)

Nitrogen dioxide is a member of the group of gases known as nitrogen oxides or nitrogen oxides (NO). These gases are extremely reactive. The indicator for the wider category of nitrogen oxides is NO2. The primary source of NO2 in the atmosphere is the combustion of fuels. NO2 is produced by automobiles, trucks, buses, power plants, and off-road equipment. In this study, we employed a duration sample of one hour. Measurement units, parts per billion

Ozone

Ozone is a gas present in the atmosphere. Ozone can be beneficial or harmful depending on its location. Natural ozone is present approximately 6 to 30 miles above the Earth's surface in the high atmosphere. This ozone layer protects humans from the sun's UV rays. When pollutants (emitted by sources such as automobiles, power plants, industrial boilers, refineries, and chemical factories) react chemically in sunlight, they produce ground-level ozone. During the warmer months, ozone pollution is more likely to emerge. This is when the weather conditions typically required for the formation of ground-level ozone—ample sunlight—occur. In this investigation, the sample duration was based on an eight-hour run average beginning at the hour. The measurement unit is parts per million.

G. Metereological variables.

The atmospheric pressure, commonly referred to as barometric pressure, is the force exerted by the air on the earth's atmosphere. At heights above sea level, air pressure is expected to drop. The measurement unit is millibars.[38]

Dew Point is the ambient temperature at which water vapor condenses from the air. Degree is the unit of measurement (in Centigrade or Fahrenheit). It provides information about the amount of moisture in the compressed air.

Humidity is the amount of water evaporating from the air. High temperatures provide greater water vapor absorption than cold air. It is expressed as a relative value between 0 and 100.

Outside Temperature is recorded on two distinct scales: Celsius and Fahrenheit. Graphs 11 depict the link with pollutants.

The flow of air from low to high pressure creates **wind speed**. It is measured using an anemometer. It can also be measured in kilometers per hour.

[38] The millibar (mbar) is a unit of pressure equivalent to one thousandth of a bar; one bar is equal to 1000 (thousand) millibars. It is one of the standard scientific units for measuring the weight of the atmosphere (or pressure produced by gravitational pull) on the surface of the earth.

H. Employment density robustness tests

Table 26: Economic and socio economic outcomes. Robustness tests: alternative buffers of employment density.

	Economic			Socioeconomic											
Population group									b	h	a	b	h	b	h
Variables	I.P.	I.P.	I.P.	H	Pov	Aff	H	H	Diss	Diss	Diss	Diss	Diss	Diss	Diss
City Structure	All	Mono	Poly	All	All	All	Mono	Poly	All	All	All	Mono	Mono	Poly	Poly
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]
ln(1-km E.D.)	1.50	0.49^{b}	2.50	0.16^{c}	0.11^{b}	0.11^{b}	0.06	0.19^{b}	35.07^{b}	38.68^{c}	-1.78	28.83	32.85	36.40	43.96
	(1.02)	(0.28)	(2.78)	(0.64)	(0.55)	(0.05)	(0.05)	(0.08)	(18.84)	(20.07)	(8.31)	(23.88)	(20.33)	(23.04)	(27.13)
ln(5-km E.D.)	0.96^{b}	0.59^{b}	1.07^{c}	0.12^{a}	0.08^{b}	0.08^{a}	0.07	0.12^{a}	25.73^{b}	28.37^{b}	-1.25	31.95	36.41	22.37^{b}	27.01^{b}
	(0.42)	(0.34)	(0.56)	(0.37)	(0.03)	(0.03)	(0.05)	(0.03)	(11.78)	(13.82)	(6.09)	(24.36)	(22.33)	(11.25)	(14.85)
ln(7-km E.D.)	1.04^{b}	0.65^{c}	1.15^{b}	0.11^{a}	0.08^{b}	0.08^{b}	0.07	0.12^{a}	25.61^{b}	28.2^{a}	-1.25	31.34	35.70	22.27^{a}	26.89^{b}
	(0.43)	(0.39)	(0.55)	(0.03)	(0.32)	(0.00)	(0.05)	(0.03)	(11.21)	(13.01)	(6.08)	(23.85)	(21.83)	(10.60)	(13.87)
ln(10-km E.D.)	1.10^{a}	0.72	1.18^{b}	0.11^{a}	0.08^{a}	0.08^{a}	0.07	0.12^{a}	25.35^{b}	27.96^{a}	-1.23	31.50	35.89	22.03^{a}	26.60^{a}
	(0.42)	(0.46)	(0.50)	(0.31)	(0.03)	(0.03)	(0.05)	(0.03)	(10.63)	(12.29)	(6.03)	(24.50)	(22.18)	(9.89)	(12.81)
IV: Expected emp.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Subsoil	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Years FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Division FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geology	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Past pop.	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Indutrial mix	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	*615*	*163*	*452*	*615*	*615*	*615*	*163*	*452*	*615*	*615*	*615*	*163*	*163*	*452*	*452*

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. b= black, h= Hispanic, a= Asian.I.P.= Income Percapita. H.= index of segregation. PovH= index of segregation of the 10% poorer. AffH= index of segregation pf the 10% more affluent. Diss= Dissimilarity index. Mono= monocentric. Poly= polycentric. E.D.= employment density. Iv= instrumental variable.

Table 27: Environmental outcomes robustness tests: alternative buffers of employment density.

Pollutant City Structure	Pm10 All [1]	So2 All [2]	ozone All [3]	Pm10 Mono [4]	So2 Mono [5]	ozone Mono [6]	Pm10 Poly [7]	So2 Poly [8]	ozone Poly [9]
ln(1-km empl. density)	3.755^b (2.170)	5.544^b (2.694)	3.454^b (1.609)	5.555 (7.072)	7.268 (10.921)	4.737 (3.170)	4.463 (2.913)	6.272^c (3.747)	4.43^c (2.372)
ln(5-km empl. density)	2.062^b (1.017)	3.049^b (1.224)	1.871^a (0.710)	4.219 (3.780)	5.643 (6.703)	3.580 (2.219)	1.921b (1.051)	2.668^b (1.289)	1.909^b (0.760)
ln(7-km empl. density)	2.11^b (0.990)	3.148^b (1.264)	1.912^a (0.713)	3.897 (3.876)	5.132 (5.921)	3.331 (2.045)	1.972^b (1.025)	2.773^b (1.356)	1.960^a (0.764)
ln(10-km empl. density)	0.251^b (1.012)	3.384^b (1.378)	2.039^a (0.767)	3.640 (3.114)	4.776 (5.417)	3.137 (1.980)	2.137^b (1.058)	3.034^b (1.517)	2.123^b (0.834)
IV: Expected emp.	✓	✓	✓	✓	✓	✓	✓	✓	✓
Subsoil	Y	Y	Y	Y	Y	Y	Y	Y	Y
Years FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Division FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Geology	Y	Y	Y	Y	Y	Y	Y	Y	Y
Past population	Y	Y	Y	Y	Y	Y	Y	Y	Y
Indutrial mix	Y	Y	Y	Y	Y	Y	Y	Y	Y
N° of observations	*615*	*615*	*615*	*163*	*163*	*163*	*452*	*452*	*452*

Notes: ^{a}, ^{b}, and ^{c} indicate significance at 1, 5, and 10 percent level, respectively. Mono= monocentric. Poly= polycentric. Iv= instrumental variable.

Part IV

Railroad network expansion, opportunity–base accessibility and population redistribution.

Railroad network expansion, opportunity–base accessibility and population redistribution

Diego Buitrago-Mora

Ph.D. Candidate

Anna Matas

Supervisor

Abstract: In this study, we look at the relationship between the expansion of the rail road system in the Madrid metropolitan area and population redistribution patterns between 1998 and 2020. We focus on two Metro expansion plans that represent roughly 40% of the total network; we also introduce a measure of accessibility that captures the opportunities gained (in terms of jobs) as a result of the transportation infrastructure improvements. We propose using geological variables as instruments (IV) to treat potential bias due to reversal causality. We generate results in terms of total population, population density, age groups, and nationality.

Key words: Accessibility, population patterns, market potential, GTFS
JEL: R23, R42

2. Introduction

In this study, we investigate the relationship between the fast growth of the railroad network and population redistribution patterns in the Madrid Metropolitan Area (MMA)[39]. We look at two expansion plans for the Metro of Madrid and light rail, which account for roughly 40% of the total network, as well as the suburban train network variation. We measure such growth by observing the variation in accessibility caused by the opening of new stations and whether that variation is a determining factor in redistributing the population of the region, which increased by more than 1.7 million people between 1998 and 2020. We use census tract-level data (3,908 per year).

This study adds to a large body of literature that investigates the relationship between transportation infrastructure expansion and population (or employment) relocation. The study of this relationship has important implications for policymakers. First, when planning the construction of highways or urban trains, it is critical to understand the medium- and long-term impact of such projects, as well as their potential to generate decision patterns in housing sorting among the population. Second, to investigate the changes in transportation and commuting costs, which are regarded as one of the main causes of population decentralization (Fujita and Ogawa, 1982), as these changes are an important driver of the increase in agglomeration benefits. Third, in order to plan and build more sustainable cities with a higher quality of life, it is necessary to have adequate urban transportation planning. Furthermore, understanding where the population will (potentially) be located and which areas will be more appealing than others is critical to addressing the issues related to housing provision. Finally, understanding the characteristics of the population groups that are relocating, such as age groups, whether they are immigrants or not, is a cross-cutting goal in this field of study.

There has been a substantial amount of research carried out to investigate the relationship between highways and sub-urbanization. Baum-Snow (2007*b*) investigates the expansion of the interstate highway system in this regard and discovers that it explains one-third of the population in metropolitan areas in the United States between 1950 and 1990. Duranton and Turner (2012*b*) also shows that the stock of highways in the United States contributed to population growth within cities between 1980 and 2020. Garcia-López *et al.* (2015) investigate highway expansion in Spain and its association with a decline in CBD population between 1960 and 2011. In this sense, even though the primary focus of this study is on trains, our findings consider the expansion of the main highway network to be a fundamental control.

Regarding investment in rails and its relation to urban population spatial patterns, Mayer and Trevien (2017) demonstrates that the extension of the Paris regional train system (RER) resulted

[39]Unlike other European metro areas (including Barcelona), the Madrid Metropolitan Area (MMA) has no official boundaries. The most recent official definition dates from 1964. As a result, in this study, we will define MMA using the definition proposed in the Comunidad de Madrid (2002), which includes twenty-eight municipalities. When referring to the internal division of Madrid city, we will use the term *districts*. We could refer, as well, to the *Region of Madrid* as the autonomous community that includes the MMA and the city of Madrid. The municipalities that make MMA are: Madrid, Alcobendas, San Sebastián de los Reyes, Colmenar Viejo, Tres Cantos, Pozuelo de Alarcón, Majadahonda, Las Rosas, Boadilla del Monte, Villaviciosa de Odon, Villanueva de la Cañada, Villanueva del Pardillo, Brunete, Alarcón, Leganés, Getafe, Móstoles, Fuenlabrada, Parla, Pinto, Coslada, San Fernando de Henares, Torrejón de Ardoz, Alcalá de Henares, Paracuellos de Jarama, Mejorada del Campo, Velilla de San Antonio, and Rivas-Vaciamadrid.

in a nine percent increase in employment without affecting population growth. Garcia-López (2012) demonstrates that improvements to highways and railroads stimulate suburban population growth. Redding and Turner (2015) investigate, among other things, the effects of transportation infrastructure improvements on the spatial distribution of the population, they present evidence at various spatial scales, both within and between cities. Gonzalez-Navarro and Turner (2018) examine the connection between the city's subway system, population, and spatial configuration. They use data on nighttime lighting as centralization indicators for 138 subway systems worldwide.

The majority of cited studies measure the difference in connectivity to transportation networks (whether highways or trains) as the difference in distance to the nearest access (ramps or train stations). We propose the use of an opportunity-based accessibility measure that takes into account, beyond the local effect, the variation in accessibility caused by the change in the entire system. In this regard, prior research has analyzed the impact of variation in accessibility to employment (Matas, Raymond, and Roig, 2015, Graham, 2007) or education (Tiznado-Aitken, Muñoz, and Hurtubia, 2021) in order to determine the effect of transport infrastructure on suburbanization. We will follow the accessibility measure proposed by Hansen (1959); Specifically, we measure the attractiveness of each census as the number of jobs reachable in Madrid (the central municipality that concentrates more than 60% of the total employment) with the rail road system, so our primary right-hand variable is the difference in accessibility between 1998 and 2020, a first-difference form of specification.

Furthermore, despite the significance of the Madrid metropolitan area for Spain and Europe, there are few existing studies; in fact, no study was found on the total impact of the rail system's expansion despite of its magnitude. Calvo, de Oña, and Arán (2013) analyzed the expansion of lines 1 and 10 with a focus on land use and described the population density surrounding the new stations. Our results take into account the expansion of all Metro, light rail, and suburban train lines, as well as precise data on connections between the three rail modes, as a result of the introduction of the accessibility measure and the use of GTFS files.

Due to the probability that the variation in accessibility is endogenous to population growth, we recommend the use of instrumental variables. Our proposal is to select geological variables that determine the possibility of building tunnels, because tunnel construction was a critical component of the system's expansion.

Following this introduction, the remainder of this study is organized as follows: First, we have a data section where we explain the main sources as well as the treatments and techniques we used before incorporating it into our main specification. Starting with the population, we move on to the rail-road system (including expansion plans and the accessibility measure), and finally to the instruments. The empirical strategy is presented in the following section. Then comes a section of results, followed by a section of conclusions.

3. Data

This section provides a general overview of the various data sets that we used in the empirical strategy section. The key statistics and general descriptions will be covered in the first section, and the processes and techniques will be covered in the subsequent subsections. since the construction of tunnels was a fundamental part of the expansion of the system.

Our variable of interest is the population of the Metropolitan Area of Madrid (MMA) at the census tract level. We obtained the data from the Institute of Statistics of the Madrid annual register (padrón) for each of the 3,908 census tracts between 1998 and 2020. Furthermore, based on the cadastral register, we identified census tracts in nonresidential areas, so we excluded them[40]. We had 3,693 census tracts at the end of the process. There are 2,360 that correspond to the city of Madrid and 1,333 that correspond to the MMA without the city of Madrid. We obtain information by nationality and age group from the same database. Finally, in accordance with Duranton and Turner (2011*a*) and Garcia-López (2012), we consider the municipal population in the years 1900, 1920, 1940, and 1960 to account for population evolution over time.

The Madrid transport consortium provided us with georeferenced information on Metro, light rail, and suburban train entrances for the year 2020. We built the network in 1998 based on this information.

Table 28: Statistics of the Main variables.

	Metro Area			Metro area (no Mdr.-city)		
Variable	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.
98-20 Δ ln(tot. Pop.)	3,693	0.91	2.46	1,333	1.46	3.03
98-20 Δ (tot. Pop.)	3,693	213.13	781.52	1,333	301	989.38
98-20 Δ (spa. Pop.)	3,693	176.26	135.66	1,333	185	144.53
98-20 Δ (fore. Pop.)	3,693	36.86	763.13	1,333	115	977.08
98-20 Δ (accessibility)	3,693	93,885	64,088	1,333	43,247	44,984
98-20 Δ ln(accessibility)	3,693	0.64	0.68	1,333	0.80	1.01
Distance to CBD ^{a}	3,693	9,292	6,884	1,333	16,899	5,542
90-17 Δ distance (hwy. ramp) ^{a}	3,693	-1,222	2,590	1,333	-1,921	3,869
Area census tract ^{b}	3,693	0.24	0.89	1,333	0.47	1.29
Population 1,900	3,693	368,894	275,182	1,333	2,801	3,168
Population 1,920	3,693	527,489	394,205	1,333	3,045	3,225
Population 1,940	3,693	846,671	633,665	1,333	3,650	4,080
Population 1,960	3,693	1,393,747	1,042,493	1,333	6,825	6,724
Population 1,998 (initial)	3,693	1,217	676	1,333	1,304	851
Closets fault (meters)^a	3,693	4,480	6,707	1,333	12,412	5,116
Closets highly permeable (meters)^a	3,693	2,179	1,755	1,333	1,539	1,272

Notes: Population of 1990, 1920, 1940 and 1960 at the municipality level. ^{a}, variable in meters. ^{b}, variable in square kilometers, Distance hwy. Ramp= distance to the closets highway ramp access. Tot.pop.= Total population. Spa. Pop.= Spanish population. Fore.pop.= Foreign-born population.

[40]This data was obtained from the European Land Cover Project (CORINE) and enabled us to identify the use at the lot level. We excluded Census tracts that had more than 90% nonresidential use.

On the other hand, we will include the impact of the expansion of the road network that connects the Madrid municipalities (major roads). The Madrid Statistics Institute publishes georeferenced data, which in 2017 totaled 661.8 kilometers. To account for this effect, we compute

$$2017 - 1990\Delta\text{distance}(\text{highway ramp access}) = distance_{2017} - distance_{1990} \tag{5}$$

which is the variation in connectivity (distance in meters) from each census unit (centroid). We built the state of the network in 1990 based on information from 2017, excluding the main highways built as part of the Spanish main expansion plan during the period (The National Road Plan of 1983).

The distance to the CBD[41] of the metropolitan region and the area of each census tract are the result of self-calculation analysis using georeferenced information from census tracts and with the support of GIS packages. Finally, we receive vector data for the set of instrumental variables from the Geological and Mining Institute of Spain (IGME).

In the following sections, we will go over the procedure we use to treat our main variables: population, railroad-related variables (including accessibility measurement), and geological instrumental variables. The statistics for the main variables used in this study are shown in Table 28. The data is presented for the Madrid metropolitan area, including and excluding Madrid city.

3.1 Population patterns in Madrid

From 1998 to 2020, the total population of Spain increased by 17%, while the population of the region of Madrid (Autonomous Community[42]) increased by 35% during the same time, reaching 6,779,888 people. The Madrid Metropolitan Area (MMA) is the second-largest metropolitan agglomeration in the European Union, behind Paris but ahead of Berlin, Barcelona, and Rome.

As shown in figure 14, foreign and local populations have driven the growth. However, the variability has been more significant among foreigners, which in 2020 represented around 14% of the total. Madrid's most important foreign population groups are from South America and the European Union.

Figure 15 depicts the change in population (logs) between 1998 and 2020 at the census level (the darker the census tract, the greater the variation). When comparing the first and last years of our study, the raw data suggests that the population has moved to the periphery of the metropolitan area, the limits of the city of Madrid, and outside of it. Some areas in the south, north, and northeast were empty in 1998 and have since been transformed into areas that will accommodate a large proportion of the total population in 2020.

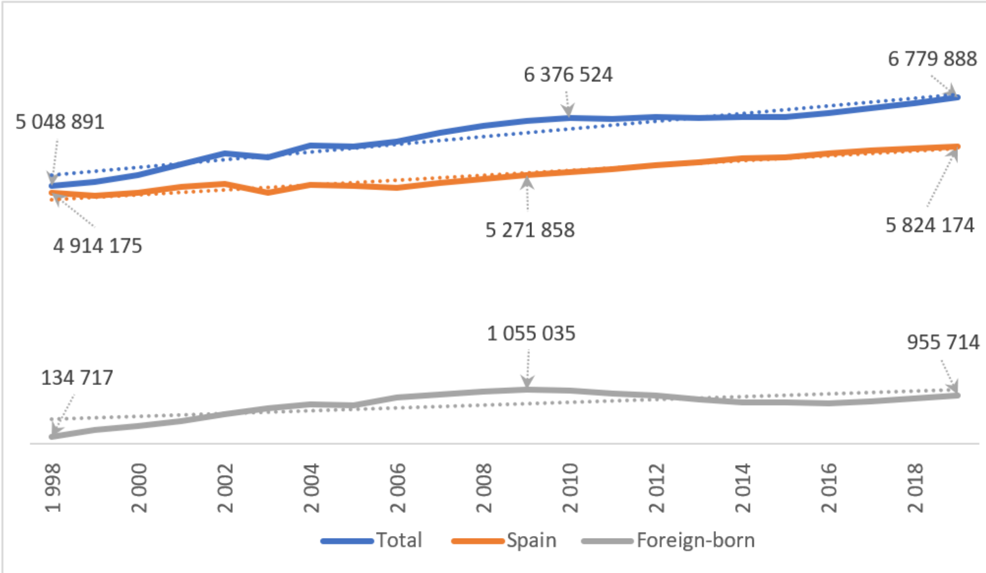
3.1.1 Aerial Interpolation

When working with population data, and especially when studying a dynamic effect, it is necessary to deal with the disparities that emerge due to changes in census units size. As a

[41] We set it in the Madrid Plaza de Sol, which also serves as the zero point of the radial highway system.

[42] Spain is divided into seventeen autonomous communities. In the case of Madrid, the name is identical to the capital city of Spain.

Figure 14: Madrid population trends between 1998 and 2020.



Notes: Notes: Total population based on the annual official register (padrón). Own development using data of the Spanish Institute of Statistics INE.

result, they must be homogenized. In such a sense, despite being a common situation faced by researchers who use spatial information (Logan *et al.*, 2014), the treatment depends on each case. In our case, we must homologate the census units to the limits of 2020. In Madrid, there is an annual update. As seen in Table 29, the number of new census tracts (compared to 2020) was 20% less in 1998, and those that changed size were 3%, although the tendency as the years goes by is to increase the share of matches; in 2009 they reach 90%.

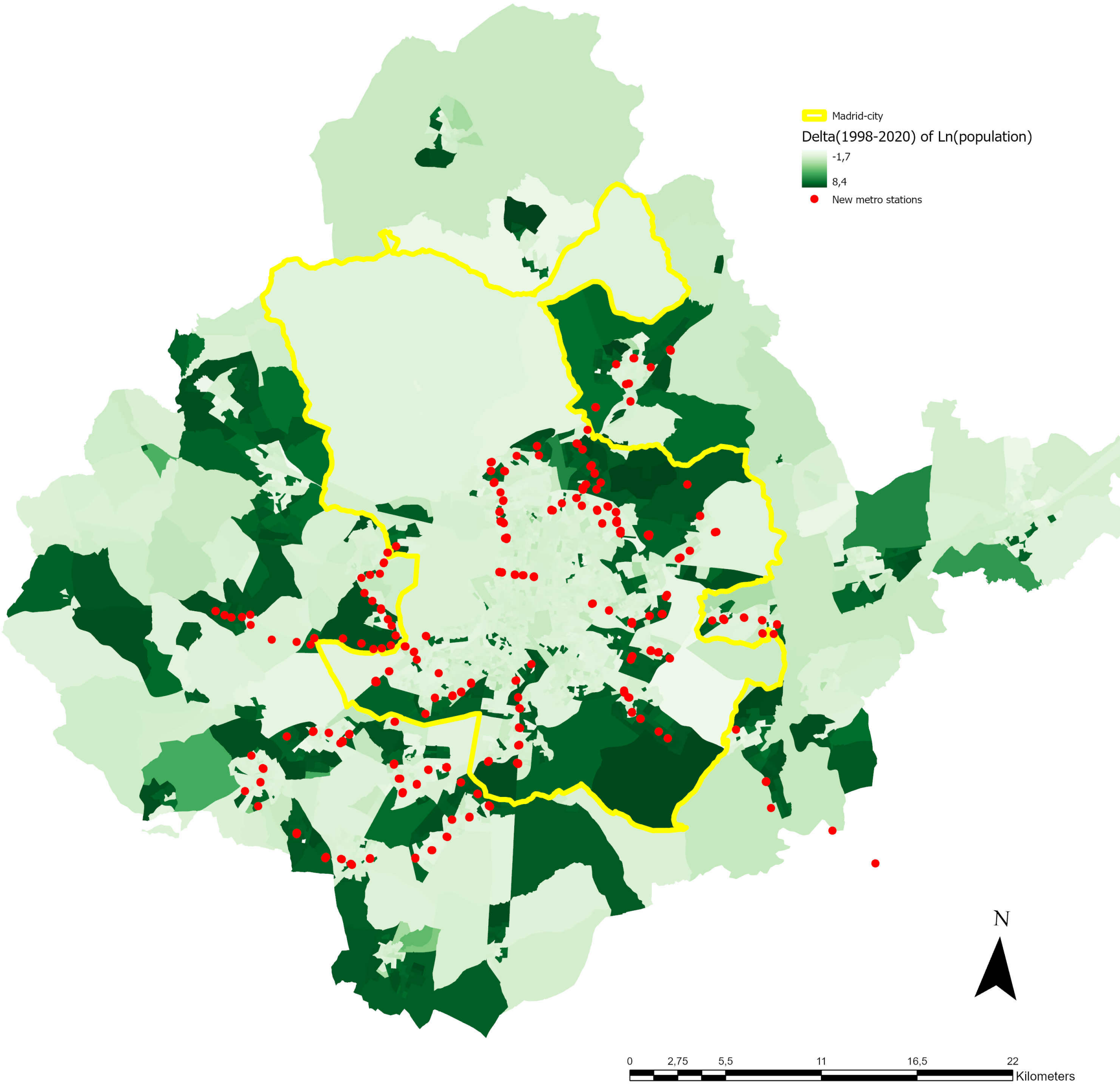
Therefore, two main problems arose. The first is related to the accuracy of the maps; advances in GIS estimations have changed over the years, and the most recent versions of the census tracts are more precise with their limits, for example, of platforms, and vehicular roundabouts, among others (see figure 16 where the black lines correspond to 2020 limits and the yellow ones to 1998). The second is the change in the size of the census tracts, either by consolidation or division (see figure 17 again, where the black lines correspond to 2020 limits and the yellow ones to 1998).

Concerning the first of our potential issues, as illustrated in Figure 16, these discrepancies can result in differences when comparing the same section over time. Logan *et al.* (2014) and Martin, Dorling, and Mitchell (2002) provide suggestions for our approach. First, we compare each year to the 2020 limits and apply an initial area filter. If the difference is greater than 2%, the census unit is interpolated. Otherwise, we assume that it is unchanged and that no interpolation is required. We use the 95% and 90% cut-offs as a robustness test because this value appears arbitrary. The number of tracts has remained constant.

Regarding the change in census tract size, we use the same interpolation with area weights as Logan *et al.* (2014)[43]. The inputs are the two georeferenced census tract data sets that we intend to interpolate. In our case, each year from 1998 to 2019, with 2020 limits.

[43]They interpolate the census tracts of the decennial census of the United States from 1970 to the limits of 2010 using a combination of aerial and population interpolation.

Figure 15: Madrid Metropolitan area and Madrid City: Δ(2020-1998) ln (total population) by census tract



Notes: Notes: Own development with data from INE.

As a result, we assigned the official population register (padrón) of that specific year if the census tract coincided with the area in 2020 by at least 98%. We used the remaining census tracts for interpolation (either because their size increased or decreased, or because they did not exist in the year compared to 2020), and then we assigned the population as follows:

$$InterpolatedPopulation_{it} = \sum_{i=1}^{n} \left(\frac{area_\gamma}{area_i} \times Pop_{it}\right) \tag{6}$$

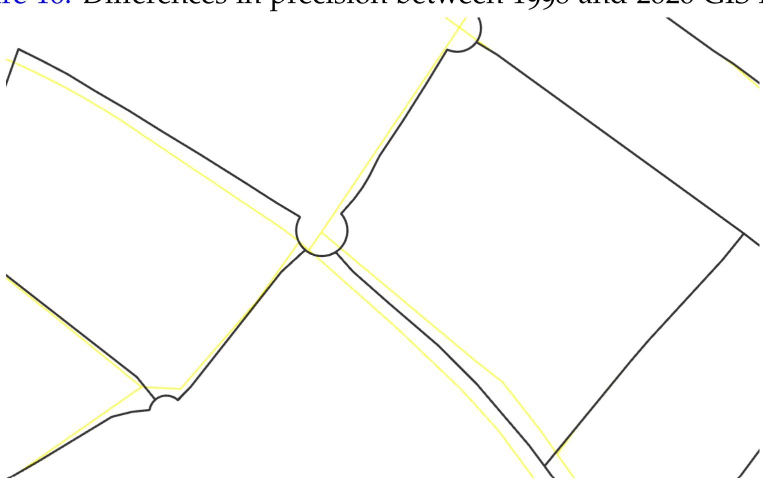
Where i represents the 2020 census tract limits, t represents a year between 1998 and 2019, and γ represents a tract in year t that intercepts i. The population of the ith section with 2020 limits is the sum of the parts in year t that comprise that section. We know that interpolation does not

Table 29: Region of Madrid: aerial Interpolation. Census tracts between 1998 and 2020.

Year	Number of census units			%		
	Change size	Match	New	Change size	Match	New
1,998	149	3,406	862	0.03	0.77	0.20
1,999	150	3,503	764	0.03	0.79	0.17
2,000	150	3,503	764	0.03	0.79	0.17
2,001	149	3,565	703	0.03	0.81	0.16
2,002	135	3,578	704	0.03	0.81	0.16
2,003	135	3,578	704	0.03	0.81	0.16
2,004	126	3,646	645	0.03	0.83	0.15
2,005	125	3,685	607	0.03	0.83	0.14
2,006	114	3,806	497	0.03	0.86	0.11
2,007	91	3,889	437	0.02	0.88	0.10
2,008	87	3,937	393	0.02	0.89	0.09
2,009	80	3,992	345	0.02	0.90	0.08
2,010	71	4,094	252	0.02	0.93	0.06
2,011	64	4,159	194	0.01	0.94	0.04
2,012	64	4,163	190	0.01	0.94	0.04
2,013	54	4,221	142	0.01	0.96	0.03
2,014	41	4,261	115	0.01	0.96	0.03
2,015	20	4,305	92	0.00	0.97	0.02
2,016	16	4,309	92	0.00	0.98	0.02
2,017	15	4,313	89	0.00	0.98	0.02
2,018	4	4,409	4	0.00	1.00	0.00
2,019	1	4,416		0.00	1.00	0.00

Notes: Change in size: if the census tract's area differed by more than 2% from the previous year. We use 1997, which is why we have 1998 results. New: If there was no register of that census tract the previous year, usually due to a subdivision.

Figure 16: Differences in precision between 1998 and 2020 GIS Maps



take into account the internal distribution of the population in each census unit. However, we are using the smallest level of aggregation.

Figure 17: Differences in the internal distribution of the same census tract

3.2 *Expansion of the railroad system in Madrid between 1998 and 2020.*

Madrid's railway network includes 370 kilometers of suburban trains and 295 kilometers of Metro, including 30 kilometers of light rail[44]. The most ambitious expansion plans in the history of the Madrid railway system were carried out between 1998 and 2020, with the main emphasis on the expansion of the Metro network. Table 30 shows how, in 1998, the metro only served the central municipality; by 2020, it had expanded to 11 municipalities, and the light metro had also emerged. Although the suburban train expanded, it did not do so at the same rate, and the majority of the new stations were in municipalities that already had a connection to the system. The train system is fully integrated, which means that if a family lives near one Metro, light rail, or suburban train station, they live near the entire system. In this section, we will go over the two Metro expansion plans (including light rail) in detail, as well as mention the suburban train expansion. This analysis is critical in the context of this research because the accessibility measure we will propose will account for these changes.

3.2.1 Metro and light metro of Madrid

The Madrid metro was inaugurated in 1919 with a length of 4 km. Since then, the network has grown to become one of the largest in the world (ranked eighth), with more stations (ranked fifth) and escalators (ranked fourth). The metro is a vast network of mostly underground tunnels through which nearly 2,400 trains pass daily; it is the transportation center of Madrid. The

[44]In addition, the municipality of Parla has an approximately 8-kilometer-long light rail. It is not included in this study because it only provides local service.

Table 30: Rail mode by municipality in 2020. Metropolitan Area of Madrid

Railroad mode	Metro		Ligth rail		Suburban	
Municipality	1998	2020	1998	2020	1998	2020
Alcalá de Henares					✓	✓
Alcobendas		✓				✓
Alcorcón		✓			✓	✓
Boadilla del Monte				✓		
Brunete						
Colmenar Viejo						✓
Coslada		✓			✓	✓
Fuenlabrada		✓			✓	✓
Getafe		✓			✓	✓
Leganés		✓			✓	✓
Madrid	✓	✓		✓	✓	✓
Majadahonda					✓	✓
Mejorada del Campo						
Móstoles		✓			✓	✓
Paracuellos de Jarama						
Parla				✓*	✓	✓
Pinto					✓	✓
Pozuelo de Alarcón				✓		
Rivas-Vaciamadrid		✓				
Rozas de Madrid (Las)					✓	✓
San Fernando de Henares		✓				
San Sebastián de los Reyes		✓				✓
Torrejón de Ardoz					✓	✓
Tres Cantos					✓	✓
Velilla de San Antonio						
Villanueva de la Cañada						
Villanueva del Pardillo						
Villaviciosa de Odón						

Notes: * The light rail of Parla was excluded. However, the municipality is integrated into the system through the suburban train. Several new suburban train stations were opened as alternatives in municipalities where another station already existed; in this table, we compare municipalities with and without a station in 1998 and 2020.

primary focus of this research will be on two expansion plans, one from 1995 to 2003 and one from 2003 to 2007.

Table 31: Expansion plans of the Metro of Madrid, since 1985

Period	Nº lines	New km.	Length(km.)	2020 (%)
1985-1990	10		112.5	38
1990-1995	11	8.3	120.8	41
1995-1998	12	14.5	135.3	46
1998-2003	12	110.5	245.8	84
2003-2007	13	32.1	277.9	98
2007-2019	13	16.1	298	100

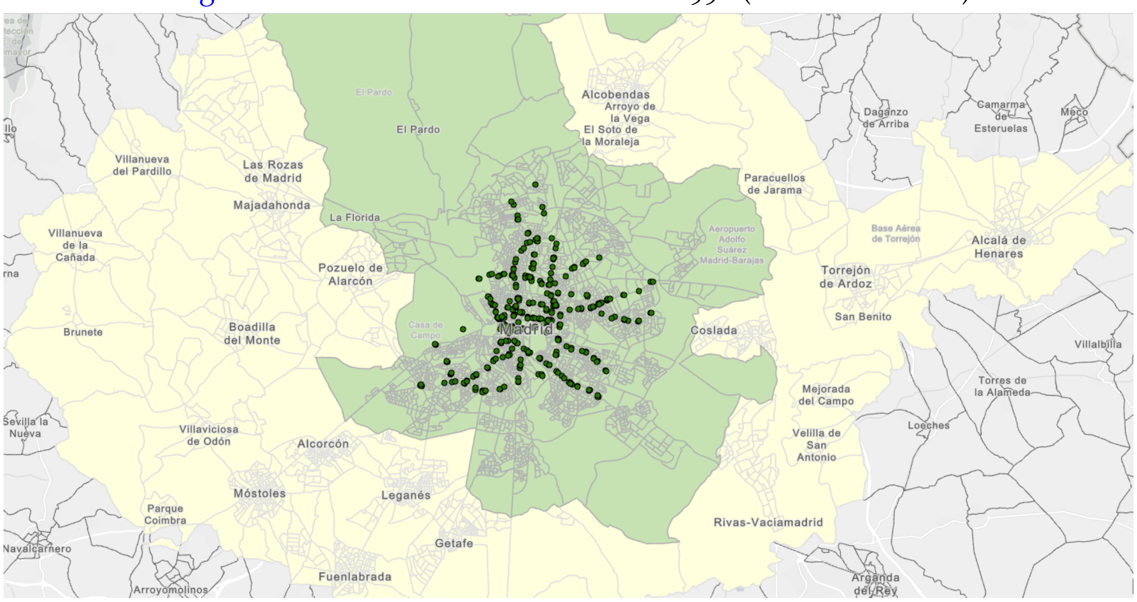
Notes: It includes the light rail.

Table 31 shows that the network of 1995 represented 46% of the 2020 system, implying that more than half corresponds to constructions that occurred over the last three decades.

Madrid's underground network expanded from 120 kilometers to 294 kilometers and from 164 to 302 stations between 1995 and 2020. Two entirely new lines, as well as three new light rails, were built. This expansion occurred over two consecutive investment plans, with the following goals in mind:

1. To connect densely populated suburbs of Madrid to the underground.
2. To provide access to strategic areas (the existing airport, the Olympic Village related to the failed bid for the Olympic Games, or a new airport never constructed).
3. Construct a circular line with 27 stations connecting major suburban towns south of Madrid.
4. The extension to municipalities outside Madrid without a train connection.
5. The extension of the Madrid underground to peripheral suburbs and relatively small towns near Madrid (with a population between 40,000 and 113,000 inhabitants).
6. The connection of new urban residential developments to the network and the construction of three intermediate stations on new lines.
7. The connection of the new residential developments located in the north and southeast of Madrid city.

Figure 18: Entrances Metro of Madrid 1998 (own elaboration)

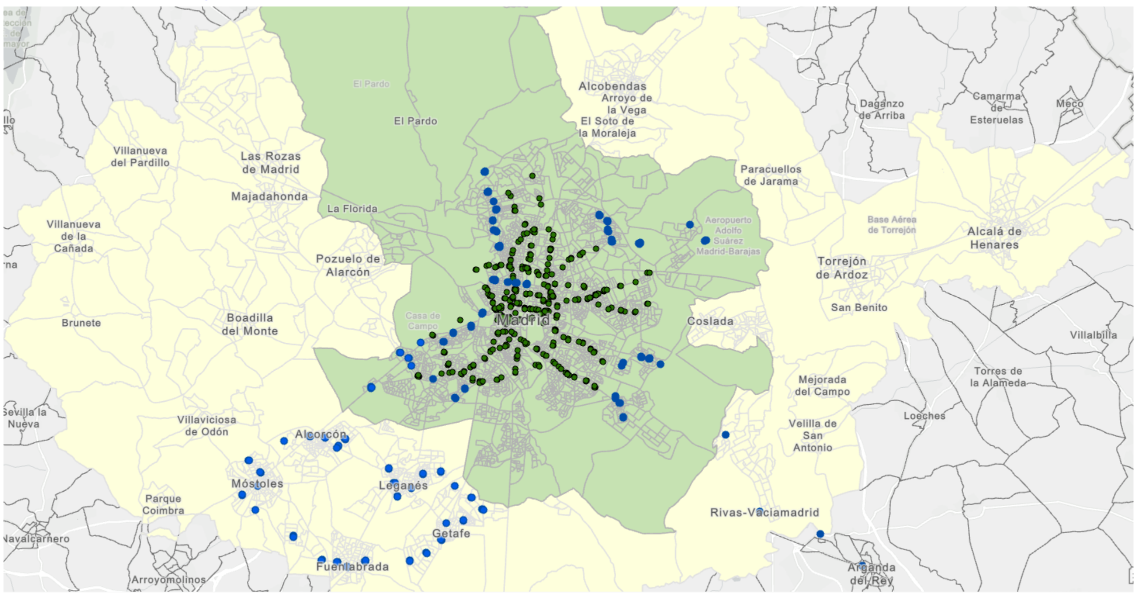


Although the expansion plan has been in place since 1995, we will evaluate the impact on the population since 1998 due to the availability of data at the census unit level. In figure 18, the green dots represent each entrance to the metro system in 1998, prior to the expansion plans. The metro network did not extend beyond the Madrid city limits (the green zone). Buses and the suburban train system connected the MMA (yellow area).

First Plan 1995-2003

Between 1995 and 2003, the first extension plan was implemented in two sub-periods. The first, which took place between 1995 and 1999, included 56 new kilometers, 38 new stations, and 223 new rolling stock (figure 19, blue dots). The second period was from 1999 to 2003, when 54.6 new kilometers and 36 new stations (419 rolling stock) were built. The objectives of these two successive plans were as follows: the extension of Madrid underground to densely populated suburbs (extensions of lines 1, 4, 7, 9, and the new line 11); the improvement of underground structure and connections (central portions of lines 7 and 10); the improvement of accessibility to strategic areas (line 8 to the airport and trade fair venue); and Metrosur.

Figure 19: Entrances Metro of Madrid 2003 (own elaboration)



The timeline of the construction between 1999 and 2003 was:

- 1994 Line 1: Puente de Vallecas – Miguel Hernández, 5 new stations.
- 1995 Line 6: closes the circle, extension from Laguna to Ciudad Universitaria, six new stations, May.
- 1996 Line 10: Plaza de España – Príncipe Pío, one station.
- 1998 Line 8: Mar de Cristal – Campo de las Naciones, June. Line 9: Pavones – Vicálvaro, December. Line 10: Alonso Martínez – Nuevos Ministerios (two stations); it continues as far as Fuencarral taking advantage of the previous L8 (changes its name to L10). Line 11: Plaza Elíptica – Pan Bendito, new line, 3 stations, November. Line 4: Canillas – Mar de Cristal (one station), April. Line 4: Mar de Cristal – Parque de Santa María (2 stations), December. Line 7: Avenida de América – Canal, 3 new stations, March and October.
- 1999 Line 8: Campo de las Naciones – Aeropuerto T1, T2 and T3, June. Line 8: Aeropuerto – Barajas, September. Line 9: Vicálvaro – Arganda del Rey, September, 18 kms. Line 1:

Miguel Hernández – Congosto (Vallecas), 3 new stations, 2.7 kms, March. Line 5: New station, Eugenia de Montijo, built between Aluche and Carabanchel. Line 7: Extension from Canal to Pitis, ten stations.

- 2002 Line 8: Nuevos Ministerios – Colombia- Mar de Cristal. Line 10: extended from Casa de Campo to Colonia Jardin.

Second Plan 2003-2007

The metro added 59 kilometers of new lines, 28 kilometers of light rail, and 81 new stations (see red dots in figure 20).

The timeline of the construction between 1999 and 2003 was:

- 2003 Line 12: Circle line around municipalities south of Madrid, April. Line 10: extension from Colonia Jardín - Cuatro Vientos – Puerta del Sur (connected to L12). Line 5: extension from Aluche to Casa de Campo (3 stations), October.

- 2006 Line10, new station Aviación española, built between Colonia Jardín and Cuatro Vientos. Line11: extensión from Pan Bendito to La Peseta, 3 new stations (Carabanchel), 2.7 kms. Line 5: extension from Canillejas to Alameda de Osuna, two stations, November, 2.4 km.

- 2007 Line 8: Nuevos Ministerios – Aeropuerto T4; last extension in 2007. Line 8: new station Pinar del Rey, built between two existing stations. Line 8: extended from Barajas to Airport T4, one station, 2.4 km. Line 10: From Fuencarral until Hospital Infanta Sofia, 15.8 km, incorporating 11 new stations. Line 3: Extension from Legazpi to Villaverde Alto, seven new stations, 8.4 km. Line 2: Extension from Ventas to Elipa (one station), 1.5 km. Line 1: Extension from Congosto to Valdecarros (three stations), South, 5.3 km. Line 1: Extension from Plaza Castilla to Pinar de Chamartín (three stations), North. Line 4: Extension from Parque de Santa María to Pinar de Chamartin (three stations). Line 6: New station, Arganzuela-Planetario, built between Mendez-Alvaro and Legazpi. Line 7: Extension from Las Musas to Henares, seven new stations. Line 1: Light Rail: nine stations, 5.4 kms, Pinar de Chamartín-Las Tablas. Line 2: Light Rail: tertheen stations, 8.4 km, Colonia Jardín – Estación de Aravaca. Line 3: Light Rail, 16 stations, 13.5 kms, Colonia Jardín – Puerta de Boadilla.

- 2008 Line 9, new station Rivas Futura, built between Rivas – urbanizaciones and Rivas Vaciamadrid. Line 7: extension from Henares to Hospital de Henares, one station, February.

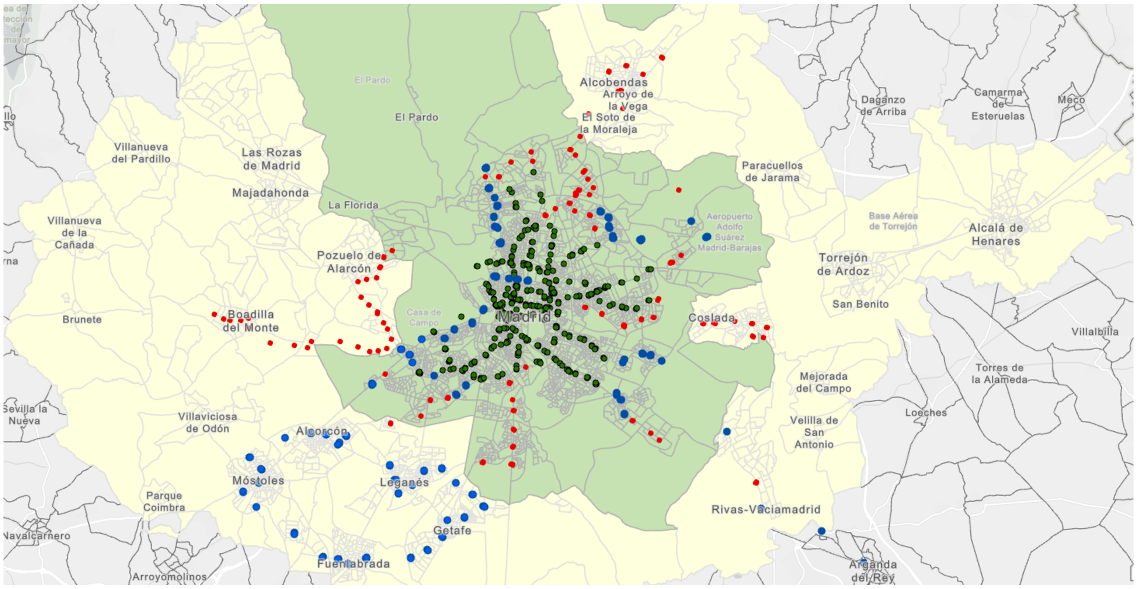
- 2010 Line 11: La Peseta – La Fortuna (Leganés), one station (last extensión), October.

- 2011 Line 9: extension from Herrera Oria to Mirasierra, one station. Line 2: extension from Elipa to Las Rosas, four stations.

- 2015 Line 9: extension from Mirasierra to Paco de Lucia (one station).

• 2019 Line 7: extension from Lacoma to Pitis (one station).

Figure 20: Entrances Metro of Madrid 2019 (own elaboration)



3.2.2 *Suburban train in Madrid*

In 1851, the first line of the Madrid suburban train network opened between Atocha (Madrid's downtown and current most important transportation hub) and the municipality of Aranjuez. There are currently 10 lines with 89 stations, 17 of which were opened between 1998 and 2020, but coverage did not change much by municipality during our study period, and some stations were even closed. In table 30, we see that the suburban train serves 17 of the MMA's 28 municipalities.

Though its expansion was less ambitious than that of the Metro (in part because it was already the most extensive rail mode in the MMA), we included it in this study because we use an accessibility metric (described in the following section), and thus we consider the overall system's network effect. This means that municipalities where there was no variation of suburban train (i.e. they already had a station and no new one was opened) will in any case experience a change in accessibility if new stations are open in another municipality.

3.2.3 *Opportunity-base accessibility measure*

We aim to measure the variation in accessibility between the census tracts of the metropolitan area of Madrid and the districts of the city of Madrid, as a result of the expansion of the rail road system in the period 1998–2020. In this section, we define how to measure said accessibility as well as the data we use for it. Incorporating an accessibility measure has an advantage over traditional methods such as Euclidean distances or binary variables, because it includes the network effect beyond the local impact of the new stations (Zheng and Ramos, 2022).

The accessibility measure we propose is a gravity-based measure that follows Hansen (1959) and Geurs and van Wee (2004). It seeks to assign a weight or importance to each area before and after the transportation innovation in order to measure its attractiveness in comparison with the other ones. Such a weight comes from the number of opportunities reachable, which is related to the number of jobs to which the inhabitants living in a census tract will have access. Regarding the employment data, is not available at the census tract level but at the district level for the year 2001 and comes from the decennial census[45].

$$A_{it} = \sum_j E_j exp(-\beta d_{ij,t}) \tag{7}$$

For the above, we define the accessibility A as the sum of jobs E that can be reached from the *ith* census tract (centroid) in the metropolitan area of Madrid, to the *jth* district in Madrid city. d_{ij} is the origin-destination matrix that includes the three existing train modes in Madrid (Metro, light rail, and suburban train) as well as walking times to the system[46].

Equation 7, is an opportunity-based accessibility (Miller, 2020). The advantage of using this type of accessibility is that we can define a distance decay effect with β. That is, the jobs closest to the *ith* centroid have a higher weight, which decreases as we move away from the *ith* centroid. Based on O'Kelly and Horner (2003), we only need to determine the β parameter, that is why from equation 7 we could say,

$$exp(-\beta d_{ij,max}) = Q_{ij,max} \tag{8}$$

Where $d_{ij,max}$ is the maximum amount of time spent commuting, and $Q_{ij,max}$ is the weight assigned to $d_{ij,max}$. Then, taking the natural logarithm, rearranging terms, and explicitly solving for β results in,

$$\beta = -\frac{ln(Q_{ij,max})}{d_{ij,max}} \tag{9}$$

As a result, we can obtain β from $d_{ij,max}$ and $Q_{ij,max}$. In order to define them, we use real data on commute times generated by the Spanish Institute of Statistics (INE) based on a decennial survey[47]. We filter by municipalities in the Madrid metropolitan region (excluding Madrid city) and by those who have to commute to work to a different municipality (we exclude people working from home or in the same municipality). The sample represents around 660,000 workers. Of those, 11% reported spending more than 2 hours per day commuting (round trip), or more than 1 hour each way. As a result, we define $Q_{ij,max}$ = 0.11 and $d_{ij,max}$ = 60 (minutes). Thus, we found a β of -0.0367 from equation 9[48]. In map 21, we show the findings of the differences in accessibility (logs) between 1998 and 2020. The darkest areas represent census tracts with the greatest increase in accessibility.

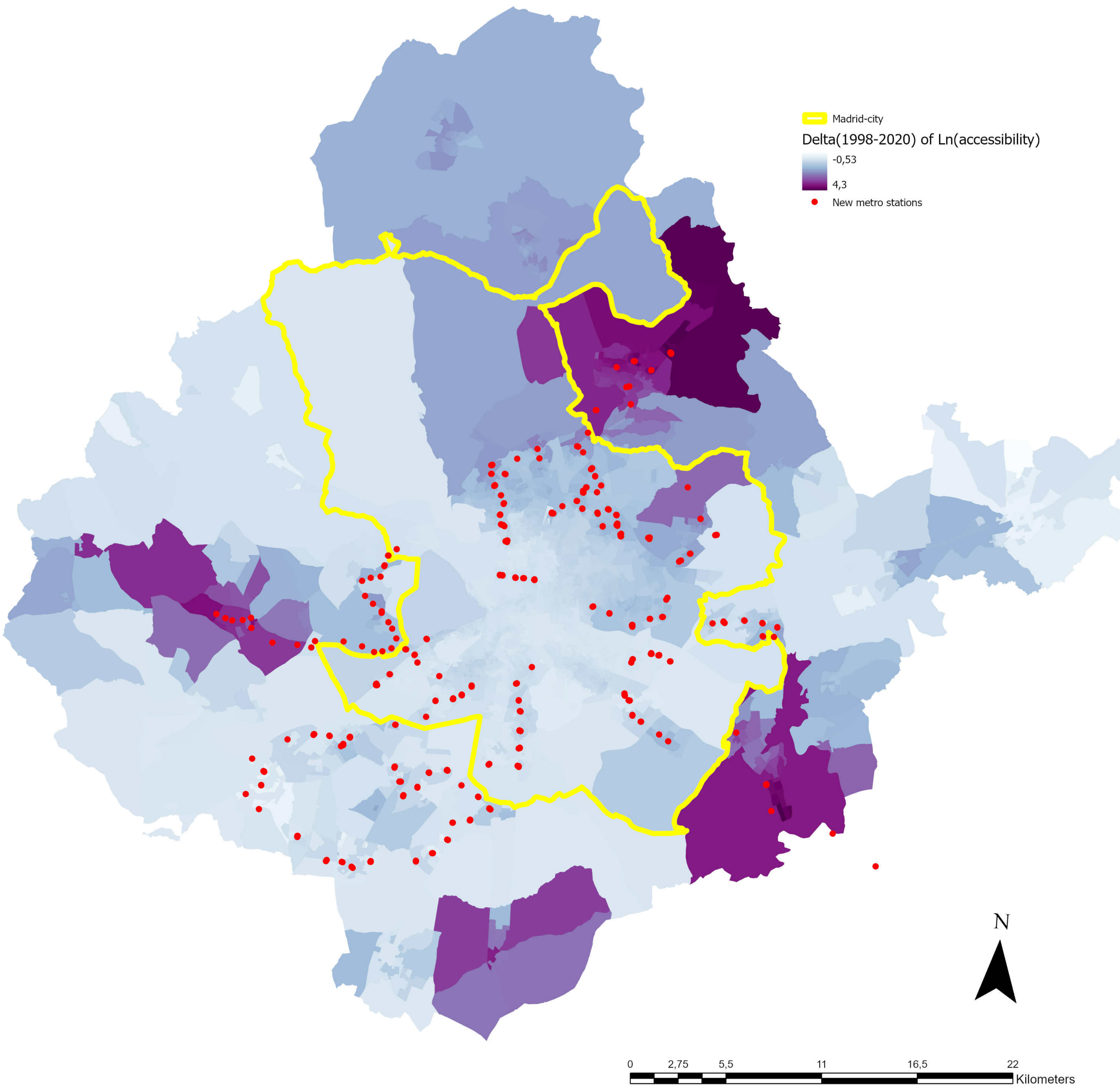
[45]Madrid City is divided into 21 districts. We did not use time-variant employment to avoid inconsistencies due to variations in the job composition between censuses. We fixed 2001 employment

[46]We expand on the method to compute d_{ij} in the next section

[47]We use the Survey of Essential Characteristics of Population and Housing 2021 (*Encuesta de Características Esenciales de la Población y las Viviendas)* which complements the 2021 Population and Housing Census.

[48]In addition, for the definition of β, we performed a likelihood ratio analysis and a comparison of the squared residuals with different values for $Q_{ij,max}$ (0.05, 0.07, 0.11, 0.15, and 0.25). We find that the best-fitting $Q_{ij,max}$ is between 0.07 and 0.11, which lines up with our choice of Q based on real commuting times.

Figure 21: Madrid Metropolitan area and Madrid City: Δ(2020-1998) ln (Accessibility) by census tract.



Notes: Own development with data from INE.

Origin-destination matrix (GTFS files)

To determine accessibility using equation 7, the origin-destination matrix ($d_{ij,t}$) between each census tract (centroid) and the 21 districts of Madrid must be generated. The matrix will contain the time frames (in minutes) that a passenger takes in the rail road system to travel between all feasible census tract-district combinations. We utilize General Transit Feed Specification Files (GTFS) for the computation. In addition to actual information on the network of streets from which to walk to the railway station.

GTFS files are created by public transportation operators in order to include georeferenced and homogenous data structures into spatial analysis using a standard vocabulary. The GTFS includes the structure of schedules, frequencies, fares, routes, stops, calendars, and trips for each mode of transport. In our case, the files for the three modes are published by the Madrid Regional Public Transport Consortium (CRTM).

We include the network of streets where people may move using a GIS management application. In other words, the trip simulation includes the time it takes a person to walk from a centroid to the nearest railway station. The streets chosen are just those that can be walked on (highways and avenues with large capacity are excluded). Finally, we ran simulations of the network in 1998 and 2020, using the same hours (a non-holiday Wednesday at 9 a.m.).

3.3 *Instruments*

Transportation improvements, as Baum-Snow (2007*b*) and Garcia-López (2012) point out, are predicted to be endogenous to population growth. As a result, the assumption that population growth is not correlated with transportation improvements may be biased. That is, causality could result from transportation improvements leading to population growth or vice versa. Policymakers may choose to serve regions that have rising projected population growth or, conversely, those with low prospects. Reverse causation would be at work in both circumstances. To address this issue, we present two instruments that meet the relevance (to be correlated with the expansion of the railroad system) and exogeneity criteria (cannot be correlated with the error term).

Geological variables and tunnels construction

Modern tunnel construction is a complicated engineering operation that must take into consideration technical features of the subsoil such as rock strength, water inflow, and ground support needs, in addition to other variables such as archaeology, risk management, and so on.

Tunneling development in Spain, for example, has been inextricably linked to the expansion of the country's rail network. Initially, with the extension of freight trains from Madrid to the ports; later, in the early twentieth century, with the establishment and quick expansion of urban rail networks Ministerio de Fomento (2009), The Madrid Metro is a prime illustration of this connection. Tunnels account for 78% of the system's total length, or approximately 230 kilometers[49]. In New York, it is 59%, while in London, it is 45%. Several underground parts of the Madrid Metro system are among the longest of it kind in Europe (Lines 12 and 7, with 42 and 32 kilometers of tunnels, respectively).

We propose employing geological characteristics to capture the likelihood of tunnel building for the reasons stated above. We provide parameters linked to water input and rock strength using the soil permeability variable. Furthermore, we must account for ground stability; therefore, we include the geological faults variable. The Geological and Mining Institute of Spain (IGME) supplies vector data for this, which sets us apart from other studies that use raster data (for example, Combes, Duranton, Gobillon, and Roux (2010) uses a 1km raster for France). Specifically, we were provided access to continuous lithostratigraphic and permeability maps. As far as we know, this information has never been used in this context.

[49]There are 41 kilometers of surface light rail, 1,9 kilometers between Eugenia Montijo and Empalme, 20 kilometers between Puerta de Arganda and Arganda del Rey, and 3 kilometers between Lago and Casa de Campo

3.3.1 *Rocks permeability*

Our first instrument incorporates rock permeability. The permeability of a rock indicates how quickly a fluid can penetrate it. If the permeability is high, rainwater will easily permeate the pores. If the permeability is low, however, rainwater will tend to accumulate on the surface or move along it if the terrain is sloped. The IGME lithological classification divides rock types based on their permeability, from highly permeable to extremely impermeable (see table 32). The more permeable the rock, the more challenging it is to construct a tunnel due to the increased risk of infiltration. In other words, the closer a census tract is to a highly permeable rock system, the less likely it is that a subway will be constructed due to stability, cost, work duration, etc.

Table 32: Lithology and subsoil permeability. IGME classification.

Lithology	Permeability				
	Highly permeable	Permeable	Medium	Impermeable	Extremely Impermeable
Carbonated	C-HP	C-P	C-M	C-I	C-EI
Detrital (Quaternary)	Q-HP	Q-P	Q-M	Q-I	Q-EI
Detrital	D-HP	D-P	D-M	D-I	D-EI
Volcanic (pyroclastic and lava)	V-HP	V-P	V-M	V-I	V-EI
Meta-Detrital	M-HP	M-P	M-M	M-I	M-EI
Igmeas	I-HP	I-P	I-M	I-I	I-EI
Evaporitic	E-HP	E-P	E-M	E-I	E-EI

On map 22, the highly permeable soils of the Madrid region are depicted. Since the IGME has provided information on the lithology of highly permeable areas, we can conclude that gravel, sand, and sediment (alluvial deposits, valley bottoms, and low terraces) predominate there.

Our first instrument is the distance between each census tract centroid and the nearest highly permeable rock system; we anticipate a negative relationship as proximity makes tunneling challenging. We therefore assume that the greater the distance, the greater the likelihood that a tunnel will be constructed in that section (and in its region of influence). When analyzing census sections, the instrument captures an individual effect that, when added, tells us where a tunnel could not be constructed due to rock permeability conditions.

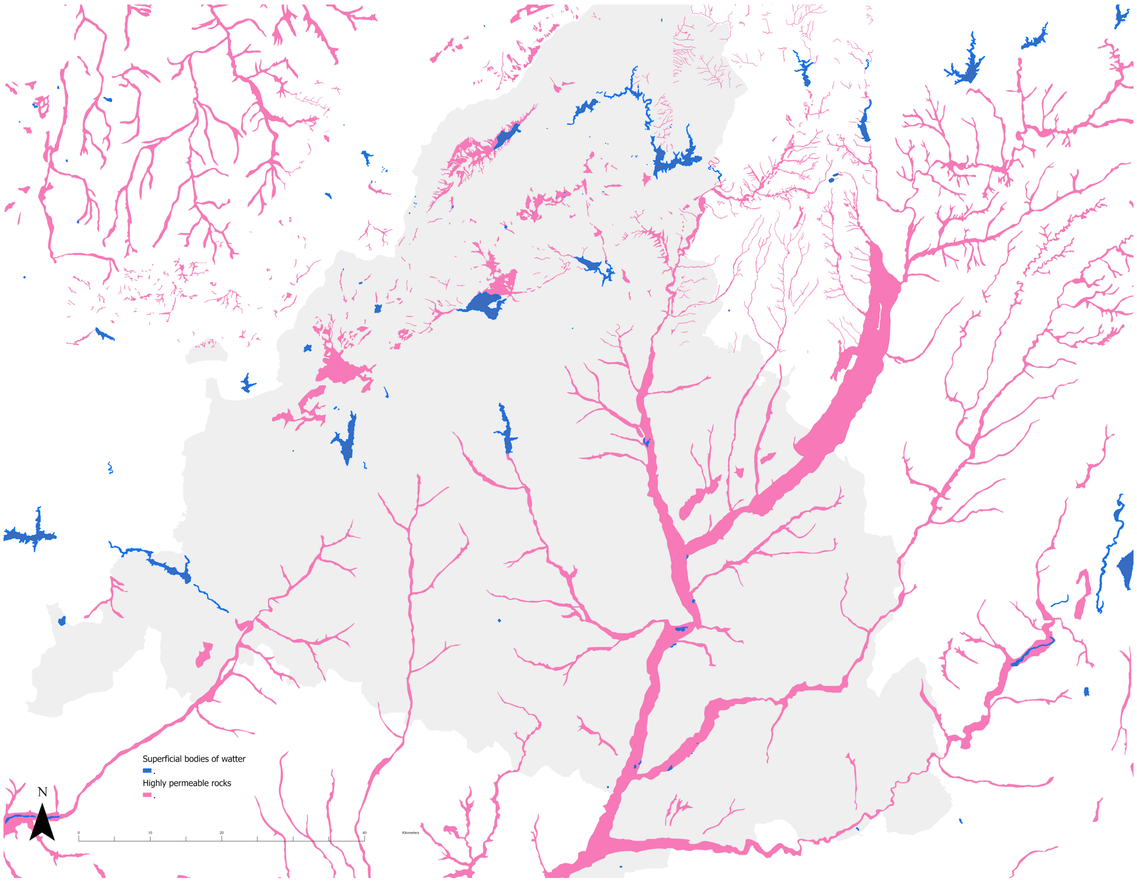
3.3.2 *Geological faults*

The second geological variable used as an instrument is faults. Again, we received detailed vector data information from IGME, including a continuous map of the contacts system[50].

Drilling in continuous rock systems is preferred for tunneling, but this is not always the case, so if there is a fault close to the route of a tunnel (normal faults or inverted faults (thrust)), it may influence the decision to build a tunnel or its route due to the necessitate of a greater engineering or financial effort. As a result, our second instrument is the distance between each centroid

[50] Contacts are an umbrella term for the boundaries between rocky bodies, and a fault is the contact between tectonic structures.

Figure 22: Madrid region, highly permeable rocks



Notes: Notes: Own development with continues vector data from IGME.

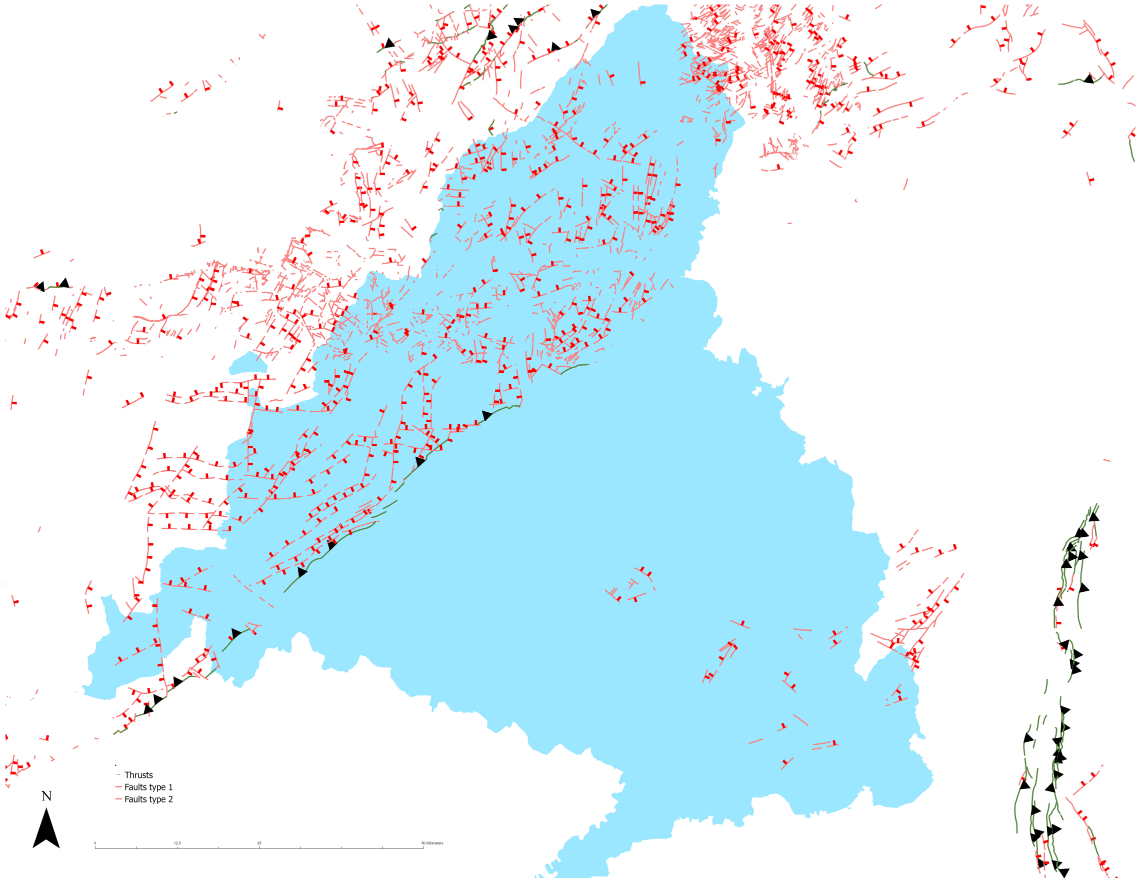
and the nearest geological fault. We believe that the distance to faults is inversely related to the potential of tunnel construction and, by extension, Metro system expansion.

The Madrid fault system is illustrated on the map 23. It is important to note that proximity to the fault has regional rather than local implications (i.e., if you are close to one part of the fault, you are close to the entire fault), making it a useful instrument for predicting tunnels construction, particularly in this case where we are studying the expansion of the train system that connects Madrid city with the municipalities in the region.

Validity of the instruments

To maintain the validity of our geological instruments, we must ensure that the distance between each census tract and the highly permeable zones and the nearest fault explains tunnel construction. Given that geological instruments have already been employed to quantify agglomeration economics (Combes *et al.*, 2010, Duranton and Turner, 2011*b*), it is vital to elaborate on what sets them apart. Remember that in this scenario, we are focusing on geological variables that directly explain tunnel building, as these structures are the axis of the railway system expansion that we are analyzing. So, in our study, we utilize geological variables for different reasons than we would if we were explaining housing, total employment, or the location of firms. We specifically consider the presented instruments to be valid because:

Figure 23: Madrid region: geological faults and thrust faults



Notes: Notes: Own development with continues vector data from IGME.

1. We did not include bodies of water, which have been utilized to explain human density or as a control variable in agglomeration economics functions and urban form (Duranton and Turner (2012*b*), for example, use under groundwater). That is, only highly permeable terrain is considered, not bodies of water. This is crucial because highly permeable soil can be linked to water sources, but not all highly permeable soils are water sources. In other words, while the population may locate near local water sources, our instrument assesses the *system* of highly permeable rocks and specifically excludes water bodies (see map 22).

2. The relevance of the instruments, which we split into two categories: scale relevance and historical relevance. The former refers to the fact that soil geological factors such as average soil richness, water presence, average stability indicators, and so on are frequently utilized as aggregated values. We are using continuous data in our scenario, and each census tract has a unique value [51] This is important since, according to Graham, Melo, Jiwattanakulpaisarn, and Noland (2010), aggregate instruments (including geological instruments), may be less useful for disaggregate work evaluating urbanization at smaller geographical scales. Combes *et al.* (2010), for example, utilize changes in mineral content to forecast population density since mineral-rich soil enhances food production, which is critical for human life.

[51] the average length of a census tract in the MMA is 322 meters.

Although the assumption is important for measuring aggregate population density (e.g., municipal), it is ineffective for our geographical scale of research.

Second, in terms of historical relevance. The historical context suggests that lithography and geological faults were much less understood at the time the MMA municipalities were established[52]. Due to the aforementioned, the expansion of the population between 1998 and 2020 in the towns surrounding Madrid has already internalized the "established" conditions. In contrast, it is highly likely that geological faults and highly permeable zones were accounted for in the planning and final placement of the tunnels (planned in the 1990s).

3. Local versus regional. We may deduce from the figure 23 and 22 that soil features have a regional influence that extends beyond the local dimension. That is, if geological faults or soils containing highly permeable rocks cross a census tract, it has no direct effect on population growth because (ceteris paribus) you can always decide to build in another part of the same town, because there is an incentive to build close to an already established city. However, the closeness to a fault or highly permeable soil system does influence the choice to build a tunnel or not, given the regional implications it has.

[52]the majority of municipalities in the Madrid region have evidence of human settlement in the modern, middle, and even ancient periods. The only exceptions are two completely contemporaneous municipalities. First the 1980-founded municipality of Tres Cantos, which is not served by the Metro or light rail, and second the municipality of Rivas Vaciamadrid, which is connected to the Metro via external trains at ground level (i.e., without tunnels)

4. Empirical strategy

In this paper, we want to add to the literature on the effects of transportation infrastructure expansion on population redistribution by using data from Madrid. We pay special attention to population variation in the metropolitan area outside the central municipality (the city of Madrid), among other things, because the metro system did not extend beyond the city limits at the start of the study period. The idea is to capture the effect of train expansion, with a focus on two Metro expansion plans, though the accessibility measure also takes into account all new light rail and suburban train station openings (and closures).

In order for our empirical strategy to evaluate the relationship between the variation of population (logs) at the census unit level in the Madrid metropolitan area between 1998 and 2020 and the variation of accessibility (logs) over the same 23-year period, we only included areas considered of residential use, and excluded the areas with no population at the beginning of the period (population of 1998 population = 0) since we consider population change in such areas to be clearly endogenous in our study (i.e., they have been the object of housing promotion and therefore have attracted population)[53]. Thus, the total population variation specification takes the form,

$$\Delta ln(Pop_{it}) = \beta_0 + \beta_1 \Delta(A_{it}) + \beta_2 \Delta(ramp_{it}) + X'\beta + \eta_{it} \tag{10}$$

Where,

- Δ ln(Pop_{it}) = $ln\ Pop_{i,2020} - ln\ Pop_{i,1998}$
- $\Delta(A_{it})$ = $ln\ accessibility_{i,2020} - ln\ accessibility_{i,1998}$
- $\Delta(ramp_{it})$ = $ramp_{i,2017} - Ramp_{i,1990}$ (distance to the closest ramp entrance)
- X' = Initial population, census tract area, history (past populations) and, geography (distance to the CBD).

To address the identification problem that arises considering railroad improvement is expected to be endogenous to total population growth, we employ a two-step methodology to answer our main question. First, we calculate the value of the variation in accessibility (logs) with equation 11.

$$\Delta(A_{it}) = \theta_0 + \theta_1 Z_{1i} + \theta_2 Z_{2i} + \theta_3 \Delta(ramp_{it}) + \mu_{it} \tag{11}$$

Where Z_{1i} and Z_{2i} are the two geological variables proposed as instruments. Then, using equation 12, we compute the variation of the total population (logs) with the estimated accessibility $\widehat{\Delta(A_{it})}$.

[53]We also omitted the areas with zero population in 1998 because we are using the variation of the log of population and the log of zero does not exist. To demonstrate that the results are not influenced by this exclusion, we run the average result from our main regression (Table 34, column 7) as a semilogarithmic (population variation in levels and the rest in logarithmic form). Then we compared the predicted variation in population levels between the two models, so we ran a test of comparison of the means of the OLS results with and without the excluded areas, and the results show that they are not statistically different.

$$\Delta ln(Pop_{it}) = \beta_0 + \beta_1 \widehat{\Delta(A_{it})} + \beta_2 \Delta(ramp_{it}) + X'\beta + \eta_{it} \quad (12)$$

5. Results

5.1 OLS results

Table 33 displays the main outcomes of maximizing equation 10. As previously stated, OLS results serve as a reference point because reverse causality introduces a risk of bias in our main results. Columns one to three present the results in three different regions of the Madrid metropolitan area: the total metropolitan area, the city of Madrid, and the first minus the second. The preliminary OLS findings indicate that the variation in accessibility (logs) caused the increase in total population (logs) in the entire metro area, with the outcome in areas outside of Madrid being particularly important. We found no evidence of the same result for the city of Madrid.

Table 33: Madrid, 2020-1998 Δ ln(total population). OLS results.

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
2020-1998 $\Delta ln(accessibility)$	0.025^a	-0.028	0.066^a	0.030^a	0.071^a	0.058^a	0.068^a
	(0.008)	(0.021)	(0.011)	(0.011)	(0.011)	(0.010)	(0.012)
Initial population	x	x	x		x	x	x
Distance to CBD	x	x	x		x	x	x
Area census tract	x	x	x			x	x
1990-2017 Δ(hwy ramp)	x	x	x				x
History							x
Metro area	x		x	x	x	x	x
Madrid city	x	x					
R2	0.50	0.248	0.289	0.490	0.310	0.289	0.288
Nº of observations	*3246*	*2176*	*1070*	*1070*	*1070*	*1070*	*1070*

Notes: Accessibility as defined in section 3.2.3. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Metro area: metropolitan area of Madrid. History: population of the municipality in 1900, 1920, 1940, and 1960. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively. Appendix 40 contains more detailed results.

Column four makes use of the same census tracts as column three but is unconditional (no controls). From columns five through seven, we added the logarithmic form of the initial population, distance to the CBD of the metropolitan area, variation in distance to the closest highway access ramp, census tract area, and past population of the municipality every twenty years from 1900 to 1960. Following Duranton and Turner (2011*a*), Garcia-López (2012), we included the last group of population variables because they allow us to control for long-run municipal population growth, which is very likely correlated with unobserved attributes that influence family housing sorting, making them a useful robustness check.

Throughout the specifications, the coefficients remain positive and statistically significant. When we compare column seven to the third column (with and without past population), we see that the coefficients are very similar, indicating that we are successfully controlling for the main characteristics that drive population change before adding the past population variables.

5.2 *TSLS results*

Table 34 shows the main results of maximizing equation 12 using a Two State Least Square methodology. To facilitate comparison, we use the same structure as in Table 33. We propose using geological soil characteristics as Instrumental Variables. As explained in Section 3.3, our IVs aim to predict the likelihood of constructing a tunnel.

From columns one to three, we examine whether accessibility variation (logs) has an effect on population variation (logs) of different geographical levels of Madrid. As in the OLS results, the metro area without Madrid is positive and significant. Furthermore, although not statistically significant, the results of Madrid City are negative in both specifications, which could indicate a dynamic of land competition. This is not the scope of this research, but it does provide a new avenue for investigation. Both the unconditional (column four) and conditional (columns five to seven) specifications produce a positive and significant result, and the addition of past populations has no significant effect on the coefficient. Furthermore, in our preferred specification, the coefficient is 2.1 times larger than the OLS results in column 3. This result is in line with what was found in earlier studies, which showed that when the source of exogenous variation (IVs) was added to control the potential bias, the coefficient went up compared to the OLS result. That is the case with Garcia-López (2012), who observed that the difference between OLS and TSLS is between 2.5 and 3.7 times larger on railroad and highway coefficients in non-central areas when explaining population growth in Barcelona between 1991 and 2001. Furthermore, Baum-Snow (2007*b*) demonstrates that the coefficients are between 1.7 and 2.8 times larger when studying the variation of central city population growth due to highway-ray variation in the United States between 1950 and 1990.

Table 34: Madrid, 2020-1998 Δ ln (total population). TSLS results with geological IVs.

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
2020-1998 $\Delta ln(accessibility)$	0.001	-0.077	0.143^a	0.347^a	0.257^a	0.114^a	0.141^a
	(0.251)	(0.069)	(0.036)	(0.096)	(0.035)	(0.030)	(0.038)
IV: Faults	✓	✓	✓	✓	✓	✓	✓
IV: Permeability	✓	✓	✓	✓	✓	✓	✓
Initial population	x	x	x		x	x	x
Distance to CBD	x	x	x		x	x	x
Area census tract	x	x	x			x	x
1990-2017 Δ(hwy ramp)	x	x	x				x
History							x
Metro Area	x		x	x	x	x	x
Madrid city	x	x					
F-Stage F-Statistics	85.4	133.0	67.3	46.5	77.7	95.32	42.5
N° of observations	*3246*	*2176*	*1070*	*1070*	*1070*	*1070*	*1070*
Overidentification	*0.36*	*0.00*	*0.029*	*0.005*	*0.221*	*0.016*	*0.067*

Notes: Accessibility as defined in section 3.2.3. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. Metro area: metropolitan area of Madrid. History: population of the municipality in 1900, 1920, 1940, and 1960. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively. Appendix 41 contains more detailed results.

We also present Stock and Yogo (2002)'s weak identification test (F-Stage F-Statistics), which suggests a critical value greater than 19.93 when using two instruments. We additionally provide the p-value (with distribution χ^2) of Sargan-Hansen's over-identification test, where the null hypothesis is that the instruments are valid, i.e., uncorrelated with the error term. Though our instruments appear to fit the model, we also run first-stage regressions, and the main results are shown in Table 35. Both of our instruments have the expected sign and significance level. We ran the specification with the historical population once more, and the results are consistent (column four).

Table 35: Madrid*, 2020-1998 Δ ln(total population). TSLS results with geological IV. First Stage results.

Dependent	**2020-1998 ln(Accessibility)**			
	[1]	[2]	[3]	[4]
ln(Closets fault)	-0.182[a]	-0.171[a]	-0.164[a]	-0.152[a]
	(0.019)	(0.017)	(0.017)	(0.025)
ln(Permeability)	-0.022[b]	-0.097[a]	-0.086[a]	-0.091[a]
	(0.019)	(0.010)	(0.010)	(0.010)
Initial population		x	x	x
Distance to CBD		x	x	x
1990-2017 Δ(hwy ramp)			x	x
Area census tract			x	x
History				x
Weak identification F-Statistics	46.47	77.71	67.30	42.53
Overidentification test	0.01	0.22	0.03	0.07
N° of observations	*1070*	*1070*	*1070*	*1070*

Notes: *Madrid metropolitan area without Madrid city. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. History: municipality population of the years 1900, 1920, 1940 and, 1960. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively.

5.2.1 *Results by nationality and age group*

We have previously presented total population results by census tract. Preliminary findings indicate that the improved accessibility has an effect on the decision of families to relocate to areas where the train system, particularly the Metro, has arrived. However, our conclusions do not offer information about the characteristics of the population, which can provide information about the mechanisms at work when it comes to understanding the relationship between transportation infrastructure improvement and population redistribution. Although we do not attempt to provide a complete answer, we do hope to supply additional information that will aid in our understanding of the phenomenon. As a result, we propose defining the data by age and nationality. The first set of variables applies to people between the ages of 20 and 39, 40 and 59, and over 60. The nationalities under consideration are the local (Spanish) and no local (Foreign-born). We had access to detailed information at the census tract level in both cases, as explained in Section 3.

Table 36 summarizes our main findings from our study of the heterogeneous effects of accessibility on population redistribution. We found that the population between 40 and 59 years old drives the average results from Table 34, followed by the population between 20 and 40 years old. We found no evidence that the variation in accessibility (logs) causes people over the age of 60 to relocate within the metropolitan area.

In terms of nationalities, we found evidence that the local population grows the most when accessibility changes (logs). However, despite the fact that the proportion of foreign-born people increased the most during the study period, we did not find that the total variation of this group responds to increased accessibility. One reason could be that the residence of immigrants is more volatile.

Table 36: Madrid*, 2020-1998 Δ ln (total population). TSLS results with geological IV. Heterogeneous results by age group and nationality.

Age-group / nationality	**20-39**	**40-59**	**>60**	**Spain**	**Foreign**
	[1]	[2]	[3]	[4]	[5]
2020-1998 $\Delta ln(accessibility)$	0.140^{a}	0.392^{a}	-0.178^{b}	0.126^{a}	-0.028
	(0.044)	(0.055)	(0.072)	(0.038)	(0.083)
IV: Faults	✓	✓	✓	✓	✓
IV: Permeability	✓	✓	✓	✓	✓
Initial population	x	x	x	x	x
Distance to CBD	x	x	x	x	x
1990-2017 Δ(hwy ramp)	x	x	x	x	x
Area census tract	x	x	x	x	x
F-Stage F-Statistics	67.29	67.29	67.29	67.29	67.29
N° of observations	*1070*	*1070*	*1070*	*1070*	*1070*

Notes: *Madrid metropolitan area without Madrid city. Accessibility as defined in section 3.2.3. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. Metro area: metropolitan area of Madrid. History: municipality population of the years 1900, 1920, 1940 and, 1960. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

5.2.2 *Robustness*

Density of population

To support our findings, we calculated the effect of accessibility variation (logs) on population density variation (logs). It is expected that as the total population grows, so will the density in zones with fixed areas (census tracts). However, density variation can also be an indicator of productivity, innovation, access to goods and services, and, in general, an increase in the benefits of agglomeration economies, in addition to complementing our result [54].

The main findings are presented in Table 37. In terms of age groups, the coefficient for the $40-59$ age group is 1.4 higher than the total population, implying that one of the possible reasons why this segment of the population has increased as a result of the new train stations is that they value the benefits of urban agglomeration. The sign and level of significance are the expected

[54] This study does not intend to investigate the effects of transportation infrastructure expansion on productivity or agglomeration economies, but it does open the door for future research in the case of Madrid and the expansion of the train system.

for the population with ages between 20 to 39; however, there is no difference with the total population (logs) result from table 34.

The Spanish population, on the other hand, increases its density (logs), with the coefficient being 2.3 times greater than the total population (logs). The result for the foreign-born population indicates a negative relationship. As mentioned by Duranton and Puga (2020*b*), anything that makes a city more appealing (such as increased accessibility) draws people from other places, putting upward pressure on house prices, which translates into higher land prices that many foreign-born families cannot afford.

Table 37: Madrid*, 2020-1998 Δ ln (density of population). TSLS results with geological IV. Heterogeneous results by age group and nationality.

Age-group / nationality	**20-39**	**40-59**	**>60**	**Total**	**Spain**	**Foreign**
	[1]	[2]	[3]	[4]	[5]	[6]
2020-1998 $\Delta ln(accessibility)$	0.140^{a}	0.533^{a}	-0.074	0.277^{a}	0.290^{a}	-0.179^{c}
	(0.044)	(0.057)	(0.056)	(0.036)	(0.038)	(0.076)
IV: Faults	✓	✓	✓	✓	✓	✓
IV: Permeability	✓	✓	✓	✓	✓	✓
Initial population	x	x	x	x	x	x
Distance to CBD	x	x	x	x	x	x
1990-2017 Δ(hwy ramp)	x	x	x	x	x	x
Area census tract	x	x	x	x	x	x
F-Stage F-Statistics	71.77	71.77	71.77	71.77	71.77	71.77
N^{o} of observations	*1070*	*1070*	*1070*	*1070*	*1070*	*1070*

Notes: *Madrid metropolitan area without Madrid city. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

Connectivity v.s accessibility: euclidean distance

The use of the variation of the Euclidean distance to measure the change in connectivity is preferred in studies of the relationship between transportation infrastructure improvements and population or employment redistribution (Baum-Snow, 2007*b*, Garcia-López, 2012). As previously shown, we propose an accessibility measure; however, in this subsection, we also present the results of replacing our main right-hand variable with a change in the Euclidean distance. As a result, in Table 39, we present the results of the impact that a change in connectivity (changes in the distance to the closest station) has on the logs of the total population, the Spanish population, and the foreign population.

The results are statistically significant and have the expected sign. The weak identification test for the instruments, on the other hand, provides less robust results: it passes the 15% maximum IV size but not the 10% (as it did in the counterpart results in tables 34, and 37). Which makes sense given that the regional nature of our instruments captures the network effect rather than the local effect captured when measuring connectivity change (Euclidean distance variation). Overall, the results support our main specification: whether measured by Euclidean distance or accessibility,

Table 39: Madrid*, 2020-1998 Δ ln (total population). Euclidean distance, TSLS results with geological IV. Heterogeneous results by nationality.

	Total	Spain	Foreign
	[1]	[2]	[3]
2020-1998 $\Delta ln(EuclideanDistance)$	-0.301^a	-0.274^a	-0.056
	(0.102)	(0.103)	(0.177)
IV: Faults	✓	✓	✓
IV: Permeability	✓	✓	✓
Initial population	x	x	x
Distance to CBD	x	x	x
1990-2017 Δ(hwy ramp)	x	x	x
Area census tract	x	x	x
F-Stage F-Statistics	12.24	12.24	12.24
N° of observations	*1070*	*1070*	*1070*

Notes: *Madrid metropolitan area without Madrid city. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. ^{a}, ^{b}, and ^{c} indicates significant at 1, 5, and 10 percent level, respectively.

improvements in railroad infrastructure in the Madrid Metropolitan area resulted in an increase in total population (logs) and population density (logs).

6. Conclusions

In conclusion, our study has demonstrated that the expansion of the train system in the Madrid metropolitan area has had a significant impact on population dynamics. The methodology employed, which utilizes accessibility measurements based on GTFS files and real-time travel times, allowed us to identify a positive relationship between accessibility increase and population growth. Specifically, a 1% increase in accessibility (logs) in a census tract resulted in a notable 0.14 (Given a maximum time of 60 minutes and a decay parameter β of -0.0367,) increase in total population (logs) over a 23-year period, with consistent results when considering population density (logs) or using Euclidean distance (logs) as alternative measure.

However, our analysis also revealed heterogeneities across age and nationality groups. Notably, the Spanish population aged 40 to 59 displayed the highest sensitivity to accessibility changes, highlighting their preference for the benefits of urban agglomeration facilitated by improved transportation infrastructure.

From a policy perspective, these findings hold crucial implications for urban planning and transportation infrastructure development. Understanding the impact of accessibility improvements on population growth can guide decision-making regarding train system expansions, housing planning, and city development. Policymakers should consider the specific needs and preferences of different demographic groups when formulating strategies to promote urban growth and address potential disparities.

Furthermore, our innovative use of unique data to examine characteristics related to tunnel construction as instrumental variables, offers valuable insights into factors influencing transportation-infrastructure development. This approach can inform more effective and targeted policy interventions in future infrastructure projects.

Finally, our study opens up two promising research avenues. Firstly, exploring population redistribution within Madrid city is essential, given the identified dynamic of land competition. Understanding how accessibility improvements affect population movements within the city can lead to informed strategies for managing urban growth and land use. Secondly, investigating the interaction between the Madrid Metropolitan Area's transportation infrastructure and productivity and firm location can shed light on the broader economic impact of transportation developments, guiding efforts to enhance regional economic growth and competitiveness.

In conclusion, our study provides essential empirical evidence on the effects of the train system expansion in the Madrid metropolitan area and offers valuable insights for policymakers to shape urban development, transportation planning, and economic growth strategies.

Appendix A.

A. OLS results

Table 40: Madrid, 2020-1998 Δ ln(total population). OLS results. All coefficients

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
2020-1998 $\Delta ln(access.)$	0.025[a]	-0.028	0.066[a]	0.030[a]	0.071[a]	0.058[a]	0.068[a]
	(0.008)	(0.021)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)
ln(Initial population)	-0.633[a]	-0.519[a]	-0.733[a]		-0.744[a]	-0.729[a]	-0.738[a]
	(0.043)	(0.074)	(0.049)		(0.056)	(0.049)	(0.049)
ln(distance to CBD)	0.000[a]	-0.000[b]	0.000[a]		0.000[a]	0.000[a]	0.000[a]
	(0.001)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
2017-1990 $\Delta ln(hwyramp)$	-0.006[a]	-0.010[a]	0.005[c]				0.006[b]
	(0.001)	(0.001)	(0.002)				(0.002)
ln(area census tract)	0.293[a]	0.253[a]	0.263[a]			0.276[a]	0.255[a]
	(0.035)	(0.097)	(0.031)			(0.029)	(0.031)
ln(pop. 1900)							-0.092
							(0.078)
ln(pop. 1920)							0.046
							(0.079)
ln(pop. 1940)							0.017
							(0.046)
ln(pop. 1960)							0.01
							(0.030)
M.Area (no Madrid-City)	x		x	x	x	x	x
Madrid city	x	x					
R2	0.50	0.248	0.289	0.490	0.310	0.289	0.288
N° of observations	*3246*	*2176*	*1070*	*1070*	*1070*	*1070*	*1070*

Notes: Accessibility as defined in section 3.2.3. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Metro area: metropolitan area of Madrid. History: population of the municipality in 1900, 1920, 1940, and 1960. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively.

B. TSLS results

Table 41: Madrid, 2020-1998 Δ ln (total population). TSLS results with geological IVs. All coefficients

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
2020-1998 $\Delta ln(access.)$	0.001	-0.077	0.143[a]	0.347[a]	0.257[a]	0.114[a]	0.141[a]
	(0.251)	(0.069)	(0.036)	(0.096)	(0.035)	(0.036)	(0.038)
ln(Initial population)	-0.630[a]	-0.515[a]	-0.742[a]		-0.753[a]	-0.733[a]	-0.745[a]
	(0.045)	(0.076)	(0.048)		(0.053)	(0.048)	(0.048)
ln(distance to CBD)	0.000[b]	-0.000[b]	0.000[a]		0.000[a]	0.000[a]	0.000[a]
	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
2017-1990 $\Delta ln(hwyramp)$	-0.007[a]	-0.009[a]	0.011[a]				0.012[a]
	(0.001)	(0.001)	(0.004)				(0.004)
ln(area census tract)	0.301[a]	0.265[a]	0.233[a]			0.265[a]	0.227[a]
	(0.036)	(0.095)	(0.033)			(0.029)	(0.032)
ln(pop. 1900)							-0.088
							(0.087)
ln(pop. 1920)							0.047
							(0.088)
ln(pop. 1940)							-0.059
							(0.061)
ln(pop. 1960)							0.086
							(0.049)
IV: Faults	✓	✓	✓	✓	✓	✓	✓
IV: Permeability	✓	✓	✓	✓	✓	✓	✓
Metro Area	x		x	x	x	x	x
Madrid city	x	x					
F-Stage F-Statistics	85.4	133.0	67.3	46.5	77.7	67.3	42.5
N° of observations	*3246*	*2176*	*1070*	*1070*	*1070*	*1070*	*1070*
Overidentification	*0.36*	*0.00*	*0.029*	*0.005*	*0.221*	*0.029*	*0.067*

Notes: Accessibility as defined in section 3.2.3. 1990-2017 Δ (hwy ramp): variation of the distance to the closest highway access ramp. Faults: distance in meters to the closets fault. Permeability: closets distance in meters to highly permeable soil. Metro area: metropolitan area of Madrid. History: population of the municipality in 1900, 1920, 1940, and 1960. First Stage F-statistics: Weak identification test, Stock and Yogo (2002) critical values. [a], [b], and [c] indicates significant at 1, 5, and 10 percent level, respectively.

Part V

General conclusions

This compilation of three empirical research studies in the field of urban and regional economies, with a focus on transportation economics, offers a thorough examination of the multifaceted dynamics that shape urban areas. Globally, urbanization, or the concentration of people in cities, has transformed societies and economies. The complex interplay of factors such as land use regulation, employment decentralization, and transportation infrastructure significantly impact development, opportunities, and quality of life within urban regions. This study connects the theoretical-practical divide in urban economics by conducting three distinct case studies, each of which sheds light on different aspects of urbanization and its interactions with other disciplines.

The first chapter looks at how urban land regulation, specifically the Law of Heights, affects real estate prices in Bogotá, Colombia. The findings show that increased regulation led to higher real estate prices, with significant differences depending on land use and income strata. However, acknowledging potential endogeneity issues highlights the importance of careful interpretation.

The second chapter investigates the causes and consequences of employment decentralization in the United States, focusing on non-traditional CBD subcenters and their effects on economic, socioeconomic, and environmental outcomes. It emphasizes the benefits of polycentric cities in terms of economic success, reduced segregation, and lower pollution levels.

The third chapter focuses on the expansion of the Madrid metropolitan area's train system and its impact on population redistribution patterns. The study found a link between accessibility improvements and population growth, emphasizing the importance of transportation infrastructure in urban development. Heterogeneity across age and nationality groups emphasizes the importance of nuanced policy considerations.

Throughout these studies, advanced data processing techniques and specialized software are used, allowing for a detailed investigation of urban contexts at various levels of analysis. This research adds to our understanding of the complex connections and dynamics that shape urban economies and societies by addressing diverse topics in various urban settings.

Overall, this research compilation adds significantly to the applied study of urban economics while maintaining a solid theoretical foundation. The findings provide valuable insights for policymakers, urban planners, and researchers working to create cities that improve the quality of life for their residents. It emphasizes that comprehensive approaches to urban development, equity, and sustainability are required, taking into account the intersection of land use regulations, transportation networks, and urban form.

The study of the relationship between urban land regulation and real estate prices reveals that, similar to developed countries, greater regulation in developing countries such as Colombia can raise real estate prices. Furthermore, the study provides empirical evidence of regulation's heterogeneous effects on different land uses and income groups, emphasizing the need for tailored policy solutions.

The study of the causes and consequences of job decentralization in the United States provides a more nuanced understanding of urban economic outcomes. It emphasizes the significance of population density and commuting costs in explaining the formation of subcenters and their effects on economic success and socioeconomic factors. The environmental findings of the study highlight the potential benefits of employment decentralization in reducing pollution exposure, reinforcing the case for polycentric city planning.

The study of the expansion of the train system in the Madrid metropolitan area demonstrates the positive impact of accessibility improvements on population growth. This information is critical for policymakers who are designing transportation infrastructure, housing, and urban development. The study's unique use of tunnel construction data as instrumental variables provides valuable insights into factors influencing transportation infrastructure development, informing future policy interventions.

These chapters highlight the complexities of urban dynamics and the importance of well-informed, targeted policies. They also propose promising research directions, such as investigating population redistribution within cities and the broader economic impact of transportation developments. As cities continue to shape our collective future, this research provides us with critical information to inform strategies for sustainable, inclusive, and prosperous urban development.

Finally, this three-chapter investigation provides a comprehensive understanding of the multifaceted dynamics that shape urban areas. It emphasizes the significance of careful urban planning and development strategies in our constantly shifting global landscape. With its blend of theory and practice, urban economics plays a critical role in shaping the cities of the future.

References

Anas, Alex, Richard Arnott, and Kenneth Small. 1998. Urban spatial structure. *Journal of Economic Literature* 36(3): 1426–1464. URL https://EconPapers.repec.org/RePEc:aea:jeclit:v:36:y:1998:i:3:p:1426-1464.

Anderson, Nathan B. and William T. Bogart. 2001. The Structure of Sprawl: Identifying and Characterizing Employment Centers in Polycentric Metropolitan Areas. *American Journal of Economics and Sociology* 60(1): 147–169. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/1536-7150.00058. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1536-7150.00058.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research.

Baum-Snow, Nathaniel. 2007*a*. Did Highways Cause Suburbanization? *The Quarterly Journal of Economics* 122(2): 775–805. URL https://academic.oup.com/qje/article/122/2/775/1942140.

Baum-Snow, Nathaniel. 2007*b*. Suburbanization and transportation in the monocentric model. *Journal of Urban Economics* 62(3): 405–423.

Bayer, Patrick, Robert McMillan, and Kim S. Rueben. 2004. What drives racial segregation? New evidence using Census microdata. *Journal of Urban Economics* 56(3): 514–535.

Blanchard, Olivier Jean and Lawrence F. Katz. 1992. Regional evolutions. *Brookings Papers on Economic Activity* 1992: 1–75.

Borck, Rainald and Philipp Schrauth. 2021. Population density and urban air quality. *Regional Science and Urban Economics* 86: 103596. URL https://www.sciencedirect.com/science/article/pii/S0166046220302817.

Brueckner, Jan K., Shihe Fu, Yizhen Gu, and Junfu Zhang. 2017. Measuring the Stringency of Land Use Regulation: The Case of China's Building Height Limits. *The Review of Economics and Statistics* 99(4): 663–677.

Brueckner, Jan K. and Kala Seetharam Sridhar. 2012. Measuring welfare gains from relaxation of land-use restrictions: The case of India's building-height limits. *Regional Science and Urban Economics* 42(6): 1061–1067.

Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew A. Turner. 2006. Causes of sprawl: A portrait from space. *The Quarterly Journal of Economics* 121(2): 587–633.

Bureau, US Census. 1987. *History of the Economic Censuses*. U.S. Department of Commerce, Bureau of the Census.

Cabeza, Ángel Massiris. 2006. Políticas latinoamericanas de ordenamiento territorial. *Universidad Nacional de Colombia, revista maestría en ordenamiento urbano-regional* : 19.

Calvo, Francisco, Juan de Oña, and Fernando Arán. 2013. Impact of the madrid subway on population settlement and land use. *Land Use Policy* 31: 627–639. URL https://www.sciencedirect.com/science/article/pii/S0264837712001792. Themed Issue 1-Guest Editor Romy Greiner-Themed Issue 2- Guest Editor Davide Viaggi.

Cameron, A. and Pravin Trivedi. 2007. *Essentials of Count Data Regression*. 331–348.

Carozzi, Felipe and Sefi Roth. 2023. Dirty density: Air quality and the density of american cities. *Journal of Environmental Economics and Management* 118: 102767. URL https://www.sciencedirect.com/science/article/pii/S0095069622001206.

Castells-Quintana, David, Elisa Dienesch, and Melanie Krause. 2021. Air pollution in an urban world: A global view on density, cities and emissions. *Ecological Economics* 189: 107153.

Cervero, R. and K.-L. Wu. 1997. Polycentrism, commuting, and residential location in the San Francisco Bay area. *Environment and Planning A* 29(5): 865–886. Publisher: Pion Limited 865.

Cheshire, Paul and G. Carbonaro. 1996. Urban economic growth in europe: Testing theory and policy prescriptions. *Urban Studies* 33(7): 1111–1128.

Cheshire, Paul and Stephen Sheppard. 2004. Land markets and land market regulation: Progress towards understanding. *Regional Science and Urban Economics* 34(6): 619 – 637.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, and Sébastien Roux. 2010. Estimating agglomeration economies with history, geology, and worker effects. In *Agglomeration economics*. University of Chicago Press, 15–66.

Comunidad de Madrid. 2002. *Atlas de la comunidad de Madrid en el umbral del siglo XXI. Imagen socioeconómica de una región receptora de inmigrantes*. 1st edition. Editorial Complutense, S.A.

Cámara de Comercio, Bogotá. 2015. Observatorio de la gestión urbana en bogotá, 4. URL https://bibliotecadigital.ccb.org.co/handle/11520/14094.

Cámara de Comercio de Bogotá. 2018. Plan de ordenamiento territorial. URL http://www.ccb.org.co.

Desmarais, Bruce A. and Jeffrey J. Harden. 2013. Testing for zero inflation in count models: Bias correction for the vuong test. *The Stata Journal* 13(4): 810–835.

Duranton, Gilles and Diego Puga. 2020*a*. The Economics of Urban Density. NBER Working Papers 27215, National Bureau of Economic Research, Inc. URL https://ideas.repec.org/p/nbr/nberwo/27215.html.

Duranton, Gilles and Diego Puga. 2020*b*. The economics of urban density. *Journal of economic perspectives* 34(3): 3–26.

Duranton, Gilles and Matthew A Turner. 2011*a*. The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review* 101(6): 2616–2652. URL http://pubs.aeaweb.org/doi/10.1257/aer.101.6.2616.

Duranton, Gilles and Matthew A. Turner. 2011*b*. The fundamental law of road congestion: Evidence from us cities. *American Economic Review* 101(6): 2616–2652.

Duranton, Gilles and Matthew A. Turner. 2012*a*. Urban growth and transportation. *Review of Economic Studies* 79(4): 1407–1440.

Duranton, Gilles and Matthew A. Turner. 2012*b*. Urban growth and transportation. *Review of Economic Studies* 79(4): 1407–1440.

Duranton, Gilles and Matthew A. Turner. 2018. Urban form and driving: Evidence from US cities. *Journal of Urban Economics* 108(C): 170–191.

Freemark, Yonah. 2020. Upzoning chicago: Impacts of a zoning reform on property values and housing construction. *Urban Affairs Review* 56(3): 758–789. URL https://doi.org/10.1177/1078087418824672.

Fujita, Masahisa and Hideaki Ogawa. 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics* 12(2): 161–196. URL https://www.sciencedirect.com/science/article/pii/016604628290031X.

Garcia-López, Miquel-Àngel. 2012. Urban spatial structure, suburbanization and transportation in Barcelona. *Journal of Urban Economics* 72(2–3): 176–190.

Garcia-López, Miquel-Àngel, Camille Hémet, and Elisabet Viladecans-Marsal. 2017. Next train to the polycentric city: The effect of railroads on subcenter formation. *Regional Science and Urban Economics* 67: 50 – 63. URL http://www.sciencedirect.com/science/article/pii/S0166046216303271.

Garcia-López, Miquel-Àngel, Albert Solé-Ollé, and Elisabet Viladecans-Marsal. 2015. Does zoning follow highways? *Regional Science and Urban Economics* 53: 148–155.

García-López, Miguel Ángel and Ana I. Moreno-Monroy. 2016. Income Segregation and Urban Spatial Structure: Evidence from Brazil. workingPaper, CAF. URL http://scioteca.caf.com/handle/123456789/957.

Geurs, Karst T. and Bert van Wee. 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography* 12(2): 127–140. URL https://www.sciencedirect.com/science/article/pii/S0966692303000607.

Giuliano, Genevieve and Kenneth A. Small. 1999. The determinants of growth of employment subcenters. *Journal of Transport Geography* 7(3): 189–201. URL http://www.sciencedirect.com/science/article/pii/S096669239800043X.

Giuliano, Genevieve and Kenneth A. Small. 1991. Subcenters in the Los Angeles region. *Regional Science and Urban Economics* 21(2): 163–182. URL http://www.sciencedirect.com/science/article/pii/016604629190032I.

Givord, Pauline, Simon Quantin, and Corentin Trevien. 2018. A long-term evaluation of the first generation of French urban enterprise zones. *Journal of Urban Economics* 105: 149–161.

Glaeser, Edward L and Matthew E Kahn. 2001. Decentralized Employment and the Transformation of the American City. Working Paper, National Bureau of Economic Research. URL http://www.nber.org/papers/w8117.

Glaeser, Edward L. and Matthew E. Kahn. 2004. Sprawl and urban growth. In J. V. Henderson and J. F. Thisse (eds.) *Handbook of Regional and Urban Economics Vol. 4: Cities and Geography*, *Handbook of Regional and Urban Economics*, volume 4. Amsterdam: Elsevier Ltd., 2481–2527.

Gonzalez-Navarro, Marco and Matthew A. Turner. 2018. Subways and urban growth: Evidence from earth. *Journal of Urban Economics* 108: 85–106. URL https://www.sciencedirect.com/science/article/pii/S009411901830072X.

Gordon, P. and H.w. Richardson. 1996. Employment decentralization in US metropolitan areas: Is Los Angeles an outlier or the norm? *Environment and Planning A* 28(10): 1727–1743. Publisher: Pion Limited 1727.

Graham, Daniel J. 2007. Agglomeration, productivity and transport investment. *Journal of Transport Economics and Policy* 41(3): 317–343. URL http://www.jstor.org/stable/20054024.

Graham, Daniel J, Patricia S Melo, Piyapong Jiwattanakulpaisarn, and Robert B Noland. 2010. Testing for causality between productivity and agglomeration economies. *Journal of regional Science* 50(5): 935–951.

Greenaway-McGrevy, Ryan, Gail Pacheco, and Sorensen Kade. 2021. The effect of upzoning on house prices and redevelopment premiums in auckland, new zealand. *Urban Studies* 58(5): 959–976. URL https://doi.org/10.1177/0042098020940602.

Hannah, Lawrence, Kyung-Hwan Kim, and Edwin S. Mills. 1993. Land use controls and housing prices in korea. *Urban Studies* 30(1): 147–156.

Hansen, Walter G. 1959. How accessibility shapes land use. *Journal of the American Institute of Planners* 25(2): 73–76. URL https://doi.org/10.1080/01944365908978307.

Huang, Haifang and Yao Tang. 2012. Residential land use regulation and the US housing price cycle between 2000 and 2009. *Journal of Urban Economics* 71(1): 93–99.

Ihlanfeldt, Keith R. 2007. The effect of land use regulation on housing and land prices. *Journal of Urban Economics* 61(3): 420–435.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 47(1): 5–86.

James, David R and Karl E Taeuber. 1985. Measures of segregation: An overview. *Demography* 22(3): 315–325.

Jones, Andrew M. 2009. Panel Data Methods and Applications to Health Economics. In Terence C. Mills and Kerry Patterson (eds.) *Palgrave Handbook of Econometrics*. London: Palgrave Macmillan UK, 557–631.

King, Gary and Margaret E. Roberts. 2015. How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis* 23(2): 159–179. URL http://pan.oxfordjournals.org/content/23/2/159.

Kok, Nils, Paavo Monkkonen, and John M. Quigley. 2014. Land use regulations and the value of land and housing: An intra-metropolitan analysis. *Journal of Urban Economics* 81: 136–148.

Ling, L., Z. Dao-lin, and H. Ke-lin. 2013. The effect of real estate regulation policy on house price in Beijing city. In *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*. 2078–2086.

Logan, John R., Zengwang Xu, and Brian Stults. 2014. Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longtitudinal Tract Database. *The Professional geographer : the journal of the Association of American Geographers* 66(3): 412–420. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4134912/.

Martin, David, Danny Dorling, and Richard Mitchell. 2002. Linking censuses through time: Problems and solutions. *Area* 34(1): 82–91. URL http://www.jstor.org/stable/20004208.

Massey, Douglas S. and Nancy A. Denton. 1988. The dimensions of residential segregation. *Social Forces* 67(2): 281–315.

Matas, Anna, Josep-Lluis Raymond, and Josep-Lluis Roig. 2015. Wages and accessibility: The impact of transport infrastructure. *Regional Studies* 49: 1236–1254. URL http://ejournals.ebsco.com.are.uab.cat/direct.asp?ArticleID=47B9A445E32C4A0D9CBC. Themed Issue 1- Guest Editor Romy GreinerThemed Issue 2- Guest Editor Davide Viaggi.

Mayer, Thierry and Corentin Trevien. 2017. The impact of urban public transportation evidence from the paris region. *Journal of Urban Economics* 102: 1–21. URL https://www.sciencedirect.com/science/article/pii/S0094119017300608.

Mayo, S. K. and Stephen Sheppard. 1996. Housing supply under rapid economic growth and varying regulatory stringency: An international comparison. *Journal of Housing Economics* 5: 274–289.

McDonald, J. F. and D. P. McMillen. 1990. Employment Subcenters and Land Values in a Polycentric Urban Area: The Case of Chicago. *Environment and Planning A: Economy and Space* 22(12): 1561–1574. URL https://doi.org/10.1068/a221561.

McDonald, John F. 1987. The identification of urban employment subcenters. *Journal of Urban Economics* 21(2): 242–258. URL http://www.sciencedirect.com/science/article/pii/0094119087900179.

McDonald, John F. and Paul J. Prather. 1994. Suburban employment centers: The case of chicago. *Urban Studies* 31(2): 201–218.

McMillen, Daniel P. 2001. Nonparametric employment subcenter identification. *Journal of Urban Economics* 50(3): 448–473.

McMillen, Daniel P. 2003. Identifying sub-centres using continguity matrices. *Urban Studies* 40(1): 57–69.

McMillen, Daniel P. and John F. McDonald. 1998. Suburban Subcenters and Employment Density in Metropolitan Chicago. *Journal of Urban Economics* 43(2): 157–180. URL http://www.sciencedirect.com/science/article/pii/S0094119097920386.

McMillen, Daniel P. and Stefani C. Smith. 2003. The number of subcenters in large urban areas. *Journal of Urban Economics* 53(3): 321–338. URL https://www.sciencedirect.com/science/article/pii/S0094119003000263.

Meijers, Evert J and Martijn J Burger. 2010. Spatial structure and productivity in us metropolitan areas. *Environment and Planning A: Economy and Space* 42(6): 1383–1402.

Miller, Eric J. 2020. Measuring accessibility: Methods and issues. International Transport Forum Discussion Paper.

Mills, Edwin S and Bruce W Hamilton. 1994. Urban economics and real estate markets. *Journal of Economic Perspectives* 8(1): 137–148.

Ministerio de Fomento . 2009. TÚneles de espaÑa. de la ingeniería medieval a los modernos túneles del siglo xxi. 586(6). URL https://cvp.mitma.gob.es/revista-del-ministerio-de-fomento-n%3F-586--extra-julio-agosto-2009---tuneles-de-espana--de-la-ingenieria-medieval-a-los-modern

Monkkonen, Paavo. 2011. Housing finance reform and increasing socioeconomic segregation in mexico. *Housing Policy Debate* 21(3): 367–393.

Monkkonen, Paavo. 2013. Urban land-use regulations and housing markets in developing countries: Evidence from Indonesia on the importance of enforcement. *Land Use Policy* 34: 255–264.

Monkkonen, Paavo and Lucas Ronconi. 2013. Land Use Regulations, Compliance and Land Markets in Argentina. *Urban Studies* 50(10): 1951–1969.

O'Kelly, Morton E and Mark W Horner. 2003. Aggregate accessibility to population at the county level: Us 1940-2000. *Journal of Geographical Systems* 5(1): 5.

Quigley, John M. and Larry A. Rosenthal. 2005. The Effects of Land Use Regulation on the Price of Housing: What Do We Know? What Can We Learn? *Cityscape* 8(1): 69–137.

Reardon, Sean F. 2011. Measures of income segregation. Technical report, Stanford University.

Redding, Stephen J. and Matthew A. Turner. 2015. Chapter 20 - transportation costs and the spatial organization of economic activity. In Gilles Duranton, J. Vernon Henderson, and William C. Strange (eds.) *Handbook of Regional and Urban Economics*, *Handbook of Regional and Urban Economics*, volume 5. Elsevier, 1339–1398. URL https://www.sciencedirect.com/science/article/pii/B978044459531700020X.

Roca, Jorge De La and Diego Puga. 2016. Learning by Working in Big Cities. *The Review of Economic Studies* 84(1): 106–142. URL https://doi.org/10.1093/restud/rdw031.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician* 39(1): 33–38.

Ruiz, Geraldine and Ivan Moncada. 2017. Análisis multitemporal de la variación del valor del suelo en bogotá durante el periodo 2013-2016, teniendo en cuenta los cambios de normativa urbana en el plan de ordenamiento territorial de la ciudad (decretos 190 de 2004, 364 de 2013, 562 de 2014, 575 de 2015 y 079 de 2016). URL https://repository.udistrital.edu.co/bitstream/handle/11349/5151/RuizPalomoGeraldineJohanna2017.pdf?sequence=1.

Secretaria de Hacienda de Bogotá. 2015. Estados financieros. URL https://www.shd.gov.co/shd/estados-contables-basicos.

Shunfeng Song. 1994. Modelling Worker Residence Distribution in the Los Angeles Region. *Urban Studies* 31(9): 1533. Publisher: Carfax Publishing Company.

Silva, João Santos, Silvana Tenreyro, and Frank Windmeijer. 2015. *Journal of Econometric Methods* 4(1): 29–46.

Small, Kenneth A. and Shunfeng Song. 1994. Population and Employment Densities: Structure and Change. *Journal of Urban Economics* 36(3): 292 – 313. URL http://www.sciencedirect.com/science/article/pii/S0094119084710370.

Stock, Jame H. and Motohiro Yogo. 2005. Testing for weak instruments in linear iv regression. In Donald W.K. Andrews and James H. Stock (eds.) *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press, 80–108.

Stock, James H and Motohiro Yogo. 2002. Testing for weak instruments in linear iv regression.

Stone, Dáithí, Maximilian Auffhammer, Mark Carey, Gerrit Hansen, Christian Huggel, Wolfgang Cramer, David Lobell, Ulf Molau, Andrew Solow, Lourdes Tibig, and Gary Yohe. 2013. The challenge to detect and attribute effects of climate change on human and natural systems. *Climatic Change* 121(2): 381–395.

Theil, Henri and Ana Finizza. 1971. A note on the measurement of racial integration of schools by means of information theory. *Journal of Mathematical Sociology* 1(2): 187–194.

Tiznado-Aitken, Ignacio, Juan Carlos Muñoz, and Ricardo Hurtubia. 2021. Public transport accessibility accounting for level of service and competition for urban opportunities: An equity analysis for education in santiago de chile. *Journal of Transport Geography* 90: 102919. URL https://www.sciencedirect.com/science/article/pii/S0966692320309960.

Turner, Matthew A., Andrew Haughwout, and Wilbert van der Klaauw. 2014. Land use regulation and welfare. *Econometrica* 82(4): 1341–1403.

van Donkelaar, Aaron, Randall V Martin, Michael Brauer, and Brian L Boys. 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental Health Perspectives* 123(2): 135–143.

Vuong, Quang. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–33.

Wallace, Nancy E. 1988. The market effects of zoning undeveloped land: Does zoning follow the market? *Journal of Urban Economics* 23(3): 307 – 326.

William T. Bogart and William C. Ferry. 1999. Employment Centres in Greater Cleveland: Evidence of Evolution in a Formerly Monocentric City. *Urban Studies* 36(12): 2099. URL https://login.are.uab.cat/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.43084408&site=eds-live. Publisher: Carfax Publishing, Taylor & Francis Ltd.

Wu, Jie, Jianfa Shen, and Bindong Sun. 2016. Effects of polycentricity on economic performance and its dependence on city size: The case of china. *Habitat International* 58: 69–77.

Zhang, Junsen. 2017. The Evolution of China's One-Child Policy and Its Effects on Family Outcomes. *Journal of Economic Perspectives* 31(1): 141–160.

Zheng, Siqi and Camila Ramos. 2022. Subway expansion, job accessibility improvements, and home value appreciation in four global cities: Considering both local and network effects .