Essays on Urban and Development Economics

Cong Peng

June 2019

A thesis submitted to the London School of Economics for the degree of Doctor of Philosophy, London, UK

Declaration

I certify that the thesis I have presented for examination for the Mphil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 55,000 words.

London, June 2019

Cong Peng

Statement of conjoint work

I certify that Chapter 3 of this thesis is co-authored with Stephen Gibbons and Cheng Keat Tang. I contributed 50% of the work. Chapter 4 of this thesis is co-authored with J. Vernon Henderson and Neeraj G. Baruah. I contributed 50% of the work.

Contents

Acknowledgements

I wish to express deep appreciation to my supervisors, Professor Steve Gibbons and Professor Olmo Silva, for their guidance during my research, without which the completion of my thesis would not have been possible. I am particularly grateful for their continued support and encouragement in exploring research ideas, their generosity with their time and patience, and their academic rigor and enthusiasm.

I would like to offer sincere gratitude to Professor Vernon Henderson. I had the honor to work closely with him on a number of research projects during my doctorate journey; he is my coauthor in one of the chapters. Our weekly discussions became the best way for me learn how to do research in practice. His continuous feedback and support have been invaluable for growing my knowledge and ability in solving problems as an economist.

I am also grateful to Neeraj Baruah, from whom I have benefited enormously in GIS and remote sensing expertise. I would also like to thank other experts both at LSE and at other institutions, that I met through conferences, for their feedback on my research. I benefited greatly from discussions with Daniel Sturm, Jos van Ommeren, Adam Storeygard, Alex Moradi, Lindsay Relihan, Gabriel Kreindler, Steve Pischke, Yizhen Gu, Christian Hilber, Henry Overman, Gabriel Ahlfeldt, Felipe Carozzi, and Giacomo Ponzetto.

As I learned over time, most research ideas would ultimately fail. Being able to handle failure and staying optimistic is important for any research to come to fruition. I was lucky to have many great friends at LSE to share frustrated moments and brighten me up. Special mentions go to Cheng Keat Tang, Wenfan Chen, Ying Chen, Hayoung Kim, Pascal Jaupart, Tanner Regan, Dzhamilya Nigmatulina, Ulrich Eberle, Matthew Sharp, Andreas Diemer, Yue Yuan, Bhargavi Sakthivel and Vivian Liu.

I gratefully acknowledge the generous financial support from the China Scholarship Council (CSC) and the LSE.

Lastly and most importantly, I would like to thank my wife, Xiaoyi Lyu, and my parents for their support and understanding.

Abstract

This thesis consists of four independent chapters on urban and development economics. The first chapter estimates the causal effect of access to new subway stations on employment. I focus on a narrow band on either side of an equidistant line to existing and new subway lines in Shanghai. Comparing employment outcomes between firms on the side of the line nearest to new stations with firms on the side of the line closest to existing stations, identifies the causal effect of the new subway service. I find that every 1km decrease in distance to the subway service improves employment growth of manufacturer firms by 55-58% over five years, while it surprisingly reduces the employment growth of consumer service firms.

The second chapter exploits the exogenous shock of an influential online shopping retail discount event in China (similar to Cyber Monday), to investigate how rapid growth of e-commerce affects urban traffic congestion. In the week after the event, intracity traffic congestion dropped by 1.7% during peak hours and 1% during off-peak hours. Using Baidu Index (similar to Google Trends) as a proxy for the changes in online shopping, I find that a 10% increase in online shopping causes a 1.4% reduction in traffic congestion. A welfare analysis conducted for Beijing suggests that the congestion relief effect has a monetary value of around 239 million US dollars a year.

The third chapter (co-authored) studies how differential institutions persisting from colonial rule affect the spatial structure of African cities, especially how new patches of development are scattered and spatially disconnected from existing developments. The paper finds that Francophone cities are more compact than Anglophone cities. Geocoded Demographic Health Survey (DHS) data further show that compact cities have better connection to public services.

The last chapter (co-authored) evaluates the environmental value of canals in England using a hedonic property price approach. Results reveal that proximity to canals increases house prices and the effect is highly localized. Houses within 100 meters of a canal have a price premium of around 5%, relative to those beyond 1km. A difference-in-differences analysis suggests that re-opening of the Droitwich canals caused price increases of around 10% within 100m of the restored canals, which is consistent with results in the national sample.

Introduction

This thesis explores four pressing themes in the realms of urban and development economics: infrastructure, digital economy, institutions, and environment.

Chapter 1 investigates the impact of a new subway line on employment, based on locations and firms, in Pudong New Area, Shanghai. On a location basis, I provide suggestive evidence supporting a positive impact of subways on employment. By aggregating firm-level employment to 500 by 500 meters grids, I show that although locations within 2 km to subway stations experience little growth in overall employment relative to locations further off, these locations experience greater increases in firm entrants, which hints at a shift in the mix of economic activity. In addition, I explore firm sorting based on the gains from subway access by tracking locations of individual firms over years. I show that firms that move into the locations that gained improved access to subways experienced higher employment growth. On a firm basis, I develop a novel identification strategy to estimate the causal effect for a subsample of firms. First, I focus on a narrow band on either side of an equidistant line to existing and new subway lines. Second, I focus on survived non-mover firms that can be observed both before and after the new line opened and did not move locations. Comparing employment outcomes between firms on the side of the line nearest to new stations, with firms on the side of the line closest to existing stations, identifies the causal effect of the new subway service. I find that every 1km decrease in distance to a subway service improves employment growth of manufacturer firms by 55-58% over five years, while it surprisingly reduces employment growth of consumer service firms.

Chapter 2 explores the impact of e-commerce on traffic congestion. Traditional retail generates vehicular traffic both from warehouses to stores and from consumers to the stores. E-commerce reduces intermediate traffic by delivering goods directly from the warehouses to the consumers. Although evidence has shown that vans servicing e-commerce are a growing contributor to traffic and congestion, by shopping online, consumers are also making fewer shopping trips using vehicles. This poses the question of whether e-commerce can reduce overall traffic congestion. The paper exploits the exogenous shock of a large-scale online shopping retail discount event in China (similar to Cyber Monday), to investigate how the

rapid growth of e-commerce affects urban traffic congestion. Portraying e-commerce as trade across cities, I specify a CES demand system with heterogeneous consumers to model consumption, vehicle demand and traffic congestion. I track hourly traffic congestion data in 94 Chinese cities in one week before and two weeks after the event. In the week after the event, intracity traffic congestion dropped by 1.7% during peak hours and 1% during off-peak hours. Using Baidu Index (similar to Google Trends) as a proxy for online shopping, I find online shopping increasing by about 1.6 times during the event. Based on the model, I find evidence for a 10% increase in online shopping causing a 1.4% reduction in traffic congestion, with the effect most salient from 9 am to 11 am and from 7 pm to midnight. A welfare analysis conducted for Beijing suggests that the congestion relief effect has a monetary value of around 239 million US dollars a year.

Chapter 3, completed with J. Vernon Henderson and Neeraj G. Baruah, studies how differential institutions imposed during colonial rule continue to affect the spatial structure and urban interactions in African cities. Based on a sample of 318 cities across 28 countries using satellite data on built-cover over time, Anglophone cities exhibit more sprawl compared to Francophone ones. Anglophone cities show less intensive land use and more irregularity in the layout of the older colonial portions of cities, and more "leapfrog development" at the extensive margin. Results are impervious to a border experiment, many robustness tests, different measures of sprawl, and different subsamples. Why would colonial origins matter? The British operated under indirect rule and a dual mandate within cities, allowing colonial and native sections to develop separately, without an overall coordinated plan. In contrast, integrated city planning and land allocation mechanisms were a feature of French colonial rule, which was inclined towards direct rule. The results also bear public policy implications. From the Demographic and Health Survey, similar households that are located in areas of the city with more leapfrog development have poorer connections to piped water, electricity, and landlines, presumably because of higher costs of providing infrastructure with urban sprawl.

Chapter 4, completed with Steve Gibbons and Cheng Keat Tang, studies the environmental value of canals in the United Kingdom. The canal and waterway network provides a potentially valuable recreational and environmental amenity. We value this amenity using a revealed preference framework by estimating how much households are willing to pay for properties closer to canals. To deal with potential omitted confounding factors in our house price regressions we adopt two strategies. First we conduct a cross-sectional analysis, but control for local area fixed effects so we estimate from marginal differences in distance from homes to canals within small geographical neighborhoods. Secondly, we apply a difference-in-differences method to analyze the effect of the restoration of the Droitwich canals in the later 2000s. Both methods yield similar conclusions. There is a price premium for living close to a canal, but this is very localized – around 3.4% in 2016 within 100 meters and zero beyond that. The implication is that the effect is driven predominantly by canal-side

properties and others with a direct outlook on the canals or immediate access. The premium fell substantially from the pre-recession to post recession periods. We also find evidence that canal-side locations are attractive for developers, with a much higher proportion of new-build sales within 100 meters of canals relative to elsewhere - a 5.9% increase on a 7.8% baseline. Some back of the envelope calculations indicate the land value uplift from the canal network was around 0.8-0.9 billion British Pounds in 2016.

Chapter 1

Does Subway Improve Employment?

1.1 Introduction

Many Chinese cities experienced a construction boom in urban rail transit systems over the past decade. From 2009 to 2015, 87 fast transit rail lines, with a total length of 3100 km, have been built in 25 cities at the cost of around 150 billion US dollars. By the end of 2015, 39 more cities have been approved by the central government to undertake their subway construction plan, with a total budget at around 685 billion US dollars, which is about 6% of China's GDP in 2015, underlining the importance of assessing the return to the investments. This trend is particularly strong in cities that serve as regional leaders. As the economic and financial center of the country, Shanghai aims to expand its subway network from 14 lines to 25 lines over the next decade, to facilitate business and industry clusters, employment generation and to achieve other "wider" economic gains. This paper investigates whether the investment in subways increases employment in target locations (on a location basis), and particularly aims to isolate the effects for existing firms (on a firm basis), using firm-level employment data surrounding a new subway line opening in Pudong New Area, Shanghai.

How does a subway line affect employment on a location and firm basis? I lay the theoretical foundation based on the model developed by Ahlfeldt et al. (2015) (Henceforth, "ARSW" model). In this model, a city consists of a set of discrete locations. Locations differ in terms of their productivity, residential amenities, supply of floor space, and access to the transport network which determines travel times between any two locations in the city. The utility that an individual derives from a pair of residence and employment locations depends on the fundamentals in both locations adjusted by an idiosyncratic utility shock, which follows the Fréchet distribution (Extreme Type-II distribution). The probabilities for an individual to commute to a specific location to work can be obtained using properties of the Fréchet distribution, similar to McFadden (1973) and Eaton & Kortum (2002). The demand for

employment in a location is a decreasing function of the travel cost from all other locations to this specific location, and an increasing function of wages. The demand to work in a location with high productivity puts downward pressure on wages. As a result, firms hire more labor (intensive margin) and more firms enter this location (extensive margin). In equilibrium, commuting technology facilitates a spatial separation of the workplace and residence, i.e., enabling individuals to work in relatively high productivity locations, while living in high amenity locations. Therefore, new subway access can facilitate the formation of firm and employment clusters. It is worth noting that because firms in this model are homogeneous, how commuting technology innovation affects the intensive margin and extensive margin remains an open question.

Growing empirical evidence supports the predictions from the model; however, few studies focus on subways (Gonzalez-Navarro & Turner, 2018), especially the impact of subways on employment within cities. The lack of attention on subways primarily reflects the challenges in studying this subject. First, rapid transit in urban areas is usually built in areas where other types of transport network have been densely placed. The add-on connectivity provided by subway is not easy to detect (Gibbons et al., 2019). Second, given the relative rarity of subways in the world, collecting enough observations for statistical analysis is another hurdle in studying the effects of subways quantitatively. The third challenge is the endogenous placement of subway stations. Subway systems and stations are not constructed at random times and places. For example, policymakers may locate subway lines in places with higher economic growth potential to enhance its development or place the network in poor neighborhoods to counter poverty. Naive comparison of employment outcomes in the neighborhoods close to subways with those further away leads to misleading results. Finally, and most importantly, the time-consuming process of subway construction allows firms to relocate to the area close to new subways, which may be an organic part of the benefit of subways on a location basis, but could cause a selection bias in the estimation of the effects of subways on a firm basis.

The unique context and research design in this paper allow for addressing these issues empirically. First, I focus on a new subway line that connects the city center with peripheries where preexisting transportation links are insufficient, which provides large shocks on accessibility in surrounding areas. Second, the study uses firm-level panel data that allow for a microscopic analysis with sufficient observations, even though I only look at one subway line. Third, drawing insights from Gibbons & Wu (2017), I adopt a novel identification strategy that can overcome the issue of endogenous placement of subways. I exploit the random variation in the discontinuity in the change of the nearest distances to subway stations induced by the new subway line. An additional subway line only changes some locations' nearest distances to subways, leaving the locations that were relatively closer to existing subway lines unchanged. Consequently, there exists a set of points in the middle of two subway lines from

which the nearest distances to stations are unchanged on one side but changed on the other side, which I term "equidistance line". As the equidistance line is not directly targeted by the policymakers, it is arguably as good as randomly assigned. Thus, comparing employment outcomes between firms on the side of the line nearest to new stations, with firms on the side of the line closest to existing stations, identifies the causal effect of the new subway service. In addition, I focus on a narrow band on either side of the equidistance line to further reduce estimation biases due to unobservables, which is an idea similar to regression discontinuity design. Lastly, I track the changes in the locations of firms to examine the spatial sorting of firms, and address the selection bias due to firm sorting on the gains from access to subways.

More importantly, this paper advances our understanding of the effects of transportation policy by distinguishing the changes in employment on a location basis from that on a firm basis. Employment change in a location includes the change on the intensive margin, which includes change in employment of existing firms, and the changes on the extensive margin, which includes change in employment due to firm entries, exits and sorting. I define firm sorting as existing firms relocate to maximize profits after transportation accessibility across locations changes due to newly added links in the network. Understanding the effect of transportation policy on employment on a location basis is important for policymakers, particularly for project appraisal and place-based policymaking. However, it does not provide much insight into the underlying economic mechanisms, as it is a composite effect. Particularly, an increase in employment in a location does not imply an increase in employment in a specific firm located in the location. As to be shown in the paper, firms in different sectors are found to respond in opposite directions to the treatment. Thus, estimating the causal effect for firms in different sectors, and ideally for firms with different characteristics, is important for firms to value subways services.

Based on the shocks on transportation accessibility induced by the new subway line, the paper has three main findings. The first two findings focus on the locations that are directly targeted by the new subway line, from which, I infer the association between subway access and employment growth on a location basis. The third finding is based on a subsample of firms that are incidentally treated and can be interpreted causally[1]. First, in order to understand the effects of subways for locations, I present a naive comparison of the change of employment between locations within a 2 km proximity of the new line and locations further off. Employment in a location is measured by aggregate firm-level employment in 500 $\times$ 500 meters grids. The treatment group consists of grids that are within 2 km of new subway stations, and the control group consists of grids that are out of the 2 km but within the 15 km radius of ostations[2]. This analysis follows a difference-in-differences framework, despite

[1]Note that results estimated from different samples may not necessarily be comparable given the potential heterogeneous effects of subways.

[2]I exclude firms that are out of the 15 km radius of stations because those areas are semi-rural.

recognizing the potential bias in the estimates arising from the nonrandom allocation of the treatment. I find that the locations in the treatment group attract new establishments of larger size (in terms of employment) relative to the control group, while the new line appears not to improve the overall employment in the treatment group. This suggests a structural change in the composition of firms and employment in the treatment group. Restricting the sample to locations that fall into the Special Economic Zones (SEZs), the new subway services seem to empower the SEZs locations in the treatment group, in which employment grow much faster than that in the SEZs further away. Second, I investigate firm sorting by dividing firms into five groups based on their initial locations, end locations, and whether a firm moved locations. I find that firms that move into the 2 km proximity of new subway stations experience higher employment growth relative to firms that move to other locations. This provides strong evidence on the existence of firm sorting for the gains from the treatment, which could exaggerate the effects of subways on a firm basis due to the selection bias.

Finally, I delve into the causal estimation of the effect of subways on employment on a firm basis, with an eye on the relative size of the effect due to sorting. Specifically, I inspect the employment growth along a 1km narrow band of the equidistance line, and within 10km of the terminal station of Line 2 original part. The area on the side of the new subway line is defined as the treatment area as the nearest distances to subways changed; the other side is defined as the control area as the nearest distances remain unchanged. I focus on the firms that are observed in both 2008 and 2013, divide them into treatment and control groups accordingly based on the area it fell into in 2013. Some of these firms move locations and could cause selection bias. For this reason, I further restrict the sample to firms that did not move locations. Eventually, I examine two samples: one sample consists of firms that survived from 2008 to 2013 (survived firms); the other sample excludes firms that moved locations from the survived firms sample (survived non-mover firms). Results for new firms that are only observed in 2013 are also reported; however, I cannot rule out selection bias as I do for the survived non-mover firms. In response to a clear pattern of firms heterogeneity across sectors found in the data, I always split the sample by three major sectors[3]: Manufacturers, producer service firms (PSFs), and consumers service firms (CSFs). The ARSW model abstracts firms from different sectors. Thus, the empirical results presented here may be helpful for enriching the theory in terms of understanding sectoral heterogeneity. Manufacturers have very balanced observed firm characteristics between the two groups before the subway opening. This strengthens the belief that the treatment is "randomly assigned", at least for the manufacturers. The samples for PSFs and CSFs are not as balanced as manufacturers; however, the growth of revenue and employment in the firms observed prior to the treatment (from 2004 to 2008) is balanced between the treatment and control areas. This suggests that the parallel trends assumption is likely to hold for a causal interpretation of a difference-in-differences estimate. The results

[3]Using a combined sample with sectors dummies yields similar results.

are strong and present substantial sectoral heterogeneity, as follows.

(i) Manufacturers. The results reveal that a 1 km reduction in the nearest distances to subways increases employment by about 61-67% for survived firms. These effects reduce to 55-58% after excluding firms that move into this area. This indicates that the effect of firm sorting on employment is small but non-negligible. Firm entries increase by about 12% for a 1 km reduction in the nearest distances to subways.

(ii) PSFs. The new subway line increases employment of survived PSFs, but the effects are not statistically significant at conventional levels after excluding movers.

(iii) CSFs. I find that employment of CSFs appears to be at a decline, surprisingly. The effects are found for both the survived firms sample and the survived non-mover firms sample.

These results suggest that the sectoral composition of employment in this area undergoes substantial changes. In addition, the results above are robust to alternative choices of distance bandwidths and are stronger for firms that are closer to the district center. Notably, the third finding echoes the first finding estimated from the locations in the proximity of the new subway stations. The decrease of employment in CSFs may offset the increase of employment in manufactures and PSFs. As a result, the overall employment shows little growth.

Although a broad body of literature studies the effect of transportation infrastructure on urban development (Gibbons et al., 2019; Donaldson, 2018; Baum-Snow et al., 2014; Michaels, 2008), this paper is among the few that investigates the impact of subways. Existing studies show that subways promote decentralization of cities (Gonzalez-Navarro & Turner, 2018), reduce air pollution (Gendron-Carrier et al., 2018), increase housing prices in its proximities (Gibbons & Machin, 2005). Studies on the impact of subways on employment are rare. The closest study to this paper is Pogonyi et al. (2018), who use firm-level panel data to study the impact of the 1999 London Jubilee Line Extension on employment and firms. They find that areas within walking distance to stations experience a significant positive effect, whereas areas further off but still within 2000 meters experience a significant negative impact, which suggests that the line mostly shifted economic activity closer to the stations while appear not stimulating economic growth. Another closely related paper is Mayer & Trevien (2017), who study the impact of Regional Express Rail (RER) on firm location, employment and population growth based on data of 101 municipalities in the Paris region. They find that the RER opening caused an 8.8% rise in employment in the municipalities connected to the network between 1975 and 1990, using municipality-level data. It is also worth noting that Tsivanidis (2019) develops the ARSW model by incorporating multiple types of workers, firms and transit modes, and provides both reduced-form and structural estimates on the

impact of the Bus Rapid Transit (BRT) system in Bogotá on economic activities. The paper finds positive effects of BRT on employment and the number of establishments. However, these studies either aggregate firm-level data to blocks or use employment data reported at the administrative boundary level, and thus mainly focus on the effects of employment on a location basis, while I measure the reduction of distances to subways for each firm and estimate the causal effect of subways on a firm basis. Moreover, none of these papers exploit the change in locations of firms to study firm sorting and selection bias.

The remainder of the paper is organized as the follows. Section 1.2 describes the data used in this study. Section 1.3 shows the aggregate effects for locations. Section 1.4 investigates firm sorting. Section 1.5 estimates the causal effects for firms. Section 1.6 concludes.

1.2 Context and Data

The paper identifies the causal impact of the Line 2 Extension (Hence, Line 2E) in Pudong New Area, Shanghai on employment in locations and firms. The Pudong New Area, to the east of the Huangpu River, is an administrative district almost as large as Singapore. Figure A.1 shows the geography of the district. Its population is about 5 million, almost a quarter of the overall population of Shanghai. Historically, the area was mainly farmland and warehouses and wharves near the shore administered by the districts of on the west bank of the Huangpu River. Since 1993, the area has become home to many Special Economic Zones (SEZs), and the economy in this area soon caught up with the economic development in the west bank. The GDP in Pudong New Area was almost a third of the GDP of Shanghai in the year 2015. The most recent population census data show that population in this area grow by about 58% from 2000 to 2010. Along with the massive population and industry growth, the coverage of subways improved substantially. The subway Line 2 original part started service in the year 2000 and connected the central business area of the west side of the district with the traditional city center in the west bank. Line 6 started service in the year 2007, linking locations along the Huangpu river in the west side of Pudong, where industries concentrate. Line 2E started service in the year 2010 and expanded the original subway Line 2 substantially. This extension in Pudong added twelve new stations and linked the district center and city center with Pudong International Airport located in the far east side of the district. This is a substantial shock to the local transportation network given that the east side of the district is much less developed and existing local transportation links to the west bank were very limited.

The estimates of the effect of transportation are based on the economic census in Pudong in the years 2004, 2008 and 2013, and the digitalized maps of subway Line 2E, Line 2 original part, and Line 6. The timing of the census and the opening of the subway lines allow observing employment of firms before and after each subway line were opened to the

public. There are seventeen SEZs located in this area, which provide tax reduction, priority financial support, and other policy convenience to attract firms to agglomerate. To study the heterogeneous effects of SEZs, I separate the firm sample by SEZs area and non-SEZs area in some specifications. The basic geography setting of this study is shown in Figure 1.1. The brown symbols mark the Line 2 original part; the pink symbols mark Line 6; the blue symbols mark Line 2E. The connecting buffers around Line 2E stations are the area within 2 km of the subway stations. The SEZs areas are digitalized and shown in the polygons marked by shades. As the area in the south of the district is semi-rural, I restrict the overall study area to 2 km away from existing subway lines and within 15 km of Line 2E.

The economic census in China started in the year 2004. The methods and questionnaires change significantly in the following two economic censuses, which limits the comparability of variables across years. All firms were asked to answer a short questionnaire, and some shortlisted large firms are asked to answer a long questionnaire. I obtained a limited number of variables from the short questionnaire, but the data cover almost all economic entities in this area. These variables include employment, revenue, broad industry classification (manufacturer, producer service firms, and consumer service firms), year established, and a plant dummy indicating whether a unit is a headquarter or a local business unit or a plant. A firm has at least one local business unit, but it can also have several ones (supermarket chains, logistics, etc.). The data record the information at a local unit level. An advantage of the dataset is that it also provides the addresses at a local unit level. For simplicity, I use the term "firm" for both headquarters or local business units or plants. In regressions, I include a dummy indicating whether a firm is a local business unit or a plant. To map these firms, I cleaned the addresses of the firms and geocoded these addresses using the geocoding service provided by map APIs. As an overview of the data, the spatial distribution of the employment growth in 2004, 2008, and 2013 can be found in Figure A.2 in Appendix A. The color of the grids shows the level of employment. Visual inspection suggests that the area went through an employment boom and employment has been clustering around the subway stations.

1.3 The Effect of Subway Access on Employment on a Location Basis

The section examines the effect of the new subway Line 2E (opened in 2010) on employment in the direct target locations using two periods of data in 2008 and 2013. This effect comes from a few sources, including changes in employment in existing firms, new firms entering, firms exiting, and the spatial sorting of firms. I call the first part the intensive margin and the rest the extensive margin. Even though the aggregate level estimates are a composite effect in a location and the estimates are likely to be biased due to a lack of adequate counterfactuals, I present these estimates for two reasons. First, the aggregate effect in the target location

is of interest from the perspective of policymaking. It provides a benchmark for assessing the impact of the subway line. Second, as the economic landscape of Pudong New Area is highly dynamic, firms enter, exit and migrate very frequently. For example, half of the firms are established between the years 2008 and 2013. The overall effects may not be a simple sum of the effects from the various sources mentioned above, due to the complex interactions among existing firms and new firms. Estimating the effect at the aggregate level captures these interactions. To examine the overall effect for locations, I aggregate firm-level outcomes into grids of 500 × 500 meters, with each grid representing a location. Figure A.2 presents the spatial distribution of aggregate employment in each year. Grids that are within a 2 km radius of Line 2E stations are defined as the treatment group, and the others within the study area (from 2 km to 15 km) are defined as the control group. As mentioned earlier, the area beyond 15 km in the south of the district is semi-rural, so I exclude them from the sample. There are no firms from the agriculture sector in this sample.

Table 1.1 shows the average (mean) log employment and log revenue of all firms and new firms across locations in the treatment and control groups respectively, before and after the opening of the new subway line. Specifically, I first take the sum of firm-level employment of all firms or only new firms in each grid, then take the average of log employment across grids by the two groups[4]. In panel (a), I compare these variables of interest without distinguishing SEZs and non-SEZs. Columns 1 and 2 show that employment increased by about 27% ($= exp(3.02 - 2.78) - 1$) from 2008 to 2013 in the overall study area. The comparison between Columns 3 and 5 and between Columns 4 and 6 show that the treatment and control groups are *unbalanced*: The treatment group has higher values of the outcome variables both before and after the subway opening. This observation indicates that subway stations were placed in areas of economic advantage. Comparing Column 3 with 4 and comparing Column 5 with 6 reveal that employment and revenue are growing in both the treatment and control groups. Column 7 shows the difference in these differences, which can be interpreted as the effect of subway access on these outcomes, under the assumption that in the absence of the new subway line, the increase in these outcomes would not have been systematically different in the treatment and control groups. This assumption cannot be taken for granted as the economic trends between the two groups could vary systematically. The results in Column 7 show that the overall employment and revenue of the locations in the treatment group are not statistically different from those in the control group. However, the growth of employment and revenue of new firms in the treated locations are higher than those in the locations in the control group. Nevertheless, as shown in Column 8, the difference-in-differences estimate is only weakly significant for new employment.

The following panels split the sample by SEZs and non-SEZs areas, where I compare the

[4]The value of log employment in grids with zero employment is replaced with zero to avoid missing values. This is equivalent to increasing employment from zero to one for those grids.

difference in the outcomes between groups within SEZs and non-SEZs areas, respectively. The pattern of outcomes presented in Panel (b) for the non-SEZs area is very similar to that for all areas. The non-SEZs areas comprise about 80% of the overall study area. In contrast, subways appear to have a substantial effect on employment in the rest 20% of the study area that falls into SEZs. Relative to Panel (b), Columns 3-6 in Panel (c) show that both the level and the change of employment in the SEZs locations in the treatment group are much larger than those in the non-SEZs locations. In Column 7, the difference-in-differences estimates on employment and revenue in the treatment group are much larger than those in the control group within the SEZs locations. The effect for overall employment is 45% ($= exp(0.37) - 1$) over five years and is statically significant at the 5% level. Panels (b) and (c) combined suggest an interaction effect between subway access and SEZs. This is consistent with the findings by Briant et al. (2015), who have shown that better access to transportation is associated with a larger positive effect of economic zones in attracting firms. Neumark & Simpson (2015) provide a comprehensive review on the this topic.

1.4 Spatial Sorting

The context of this study provides an opportunity to explore firm sorting. Redding & Turner (2015) show that firm sorting leads to an overestimation of the aggregate effect in the treated area (also see Fujita et al. (2001)). Their illustration divides the space to three categories: "treated" is a region which is directed affected by a transportation scheme, "untreated" is a region which is adjacent to the treated region but not subject to the scheme, and the rest of the space is "residual". Assuming that the treated area is better off by a unit economic output, the scheme may attract firms in the untreated region to relocate to the treated region and displace the d unit of economic output from the untreated region to the treated region. The treatment effect estimated from comparing the treated region with the untreated regions is $a + 2d$ unit. Thus the impact of the scheme is overestimated due to sorting[5]. This is essentially the problem of selection bias due to "sorting on the gain" as described in Heckman et al. (2006). This highlights the importance of distinguishing firms that moved locations from firms that did not move in mitigating the selection bias problem. The context of the research and the data are insufficient to causally estimate how sorting contributes to employment growth in locations, because sorting decisions are based on firm characteristics that are largely unobserved, such as local business network or distance to supply chains. This section provides a descriptive analysis of how firms gain from sorting, while aiming to mitigate the omitted variable bias problem by carefully choosing reference groups that are arguably comparable.

I divide firms into four groups: Firms that stayed in the area within 2 km of Line 2E stations

[5]Note although the illustration is simple, it is difficult to find a clear cut-off to define the three categories of areas in practice.

(stay-in); firms that were within 2 km radius of Line 2E stations but moved out of this area during 2008 to 2013 (move-out); firms moved in during this period (move-in); firms that stay out of this area during this period (stay-out). Besides these mutually exclusive and collectively exhaustive groups, I also focus on the firms that move locations further than 500 meters but have been at least 2 km away from subway stations in both years (stay-out-mover). These stay-out-mover firms may be a better alternative counterfactual for the move-in group relative to the stay-out firms because firms that move are likely to be different from firms that do not move. For example, the management of the former could be more motivated for growth.

Table 1.2 reports basic summary statistics for the outcomes of interest and other characteristics in the five groups defined above. From the year 2008 to 2013, firms in all groups experience a slight drop in employment, as shown in the first row, except the move-in group. Move-in firms expand their labor force, on average, by 8% over the five years. In addition, these firms are characterized by slightly higher revenue in 2008 and fewer years of establishment. In contrast, move-out firms experience about 15% decrease in employment despite having a higher revenue in 2008 than stay-in firms. These results based on observed characteristics suggest that move-in and move-out firms are probably also different from the stay-in firms in many unobserved dimensions. One way to limit unobservables is to conduct pairwise regressions controlling for observed firm characteristics listed from the second row to the last row in Table 1.2. The groups of firms included in the four sets of regressions are move-out versus stay-in; move-in versus stay-out; move-in versus stay-out-mover; move-in versus stay-in. The first four columns in Table 1.3 show the results for the four sets in the period of 2008-2013. The difference in employment change between the move-out and stay-in firms is much less salient after adding these controls (the difference in the mean is about 0.072 in Table 1.2). Column 2 shows that move-in firms experience a much higher increase in employment than stay-out firms. It is possible that stay-out firms are very different from move-in firms. Column 3 uses stay-out-mover firms as the reference group, which potentially narrows the differences in unobservables with the move-in firms as both moved locations. As expected, the coefficient drops from 0.176 in Column 2 to about 0.126 in Column 3. Column 4 shows that the move-in firms also have much higher growth in employment relative to the stay-in firms. More importantly, the relocation of firms to this area is very likely driven by the new subway access. To show this, Columns 5-8 replicate the regressions in Columns 1-4 using firms observed from 2004 to 2008 in the data. The estimates are all very close to zero and not statically significant. Firms that moved into or out of the treatment area prior to the new subway opened did not benefit or lose. This rules out other advantages in the treatment area other than the new subway service. Taken together, it appears very likely that firms sort spatially for gains of subway access and benefit from it.

1.5 The Effect of Subway Access on Employment on a Firm Basis

The section aims to identify the causal effects of subways on existing firms. There are two challenges in identifying the causal effect. First is the endogenous placement of subway lines. The assumption that the treated and control groups follow parallel trends in section 1.3 is unlikely to hold, and therefore the difference-in-differences estimates presented are likely to be biased. Even if the economic trends were the same prior to the treatment, there could be mean reversion (Duflo, 2001). Second, the existence of spatial sorting of firms as demonstrated in section 1.4 implies selection bias in estimating the causal effect of subway access improvement on employment. In this section, I develop a novel identification strategy to overcome these two challenges to identify the causal effects.

1.5.1 Identification Strategy

My identification strategy exploits random variations in the discontinuity in the access to nearest subway stations induced by the new subway line. When there is only one subway line in the space, the nearest distance to subway service is continuous. Adding a new subway line breaks such continuity. The nearest distance to subway services becomes the minimum of distances to either subway lines, which is a discontinuous function of locations. Thus, there exists a set of points from which the nearest distance is the same to both subway lines, which I term "equidistance line". Importantly, on the side of the existing subway line, the nearest distance does not change after the new subway line is added, while on the side of the newly added subway line, the nearest distance changes. The key element of this design is that I compare employment outcomes between firms on the side of the line nearest to new stations, which are defined as the treatment group, with firms on the side of the line closest to existing stations, which are defined as the control group. Specifically, I focus on a narrow band on either side of the equidistance line to further address the effects of unobservables. As shown in Figure 1.2, the equidistance line divides firms into two groups, with the firms in the control group marked by green dots and those in the treatment group marked by red dots. I set the bandwidth to 1 km in the baseline sample, and later explore alternative bandwidths for robustness checks.

To address the spatial sorting of firms, I further restrict the sample to firms that have not moved between 2008 and 2013, given that the data allow for tracking locations of firms. To enter the sample, firms need to satisfy two conditions. First, their business survived from 2008 to 2013; Second, these firms did not move locations[6]. For simplicity, I call these firms "survived non-mover firms". These firms are thus the recipients of the treatment of

[6]Given the potential geocoding errors, firms that moved within 50 meters are classified to as "not moved".

subway access improvement and did not select their locations based on the treatment. In the regressions, I first show the effects estimated from survived firms, and then show the effects estimated from the survived non-mover firms. The difference between the two effects indicates the selection bias.

As we move away from the district center, the impact of the reduction in the nearest distances to subways may fade away, because individuals may not take the subways at all when the nearest distances are above a certain threshold. I explore restricting samples to firms that are within different distance ranges to the district center. In Figure 1.2, the two buffers with a radius of 6km and 10km centered on the original end of Line 2 mark the spatial boundary of the confined samples. Note, as this design restricts the sample to the narrow band along the equidistance line, the effect identified from these firms may not be generalizable to the overall area. Also, as there are only a few firms located in SEZs in this sample, this section does not study the SEZs and non-SEZs areas separately.

1.5.2 Econometric Framework

I start from the basic regression equation:

$$lnL_{it} = \beta D_{it} + \{\iota_i + \tau_t + \varepsilon_{it}\} \tag{1.1}$$

where i indexes firms, t indexes years. L_{it} represents employment in firm i in the year t. D_{it} is firm i's nearest distance to subway stations. D_{it} changes with time due to the new subway access. Unobservable factors include firm-specific time-invariant components ι_i, year-specific firm-invariant components τ_t, and year-by-firm components ε_{it}. Given that I only observe two periods of data, this is equivalent to the change specification below:

$$\Delta lnL_i = \beta \Delta D_i' + \Delta\tau + \Delta\varepsilon_i \tag{1.2}$$

where firm-specific time-invariant components ι_i are eliminated and year-specific firm-invariant components τ_t reduce to the constant of the regression model. In order to ease interpretation, let $\Delta D_i' = -(D_{i1} - D_{i0})$, which measures the absolute change in the nearest distances. β estimates the causal effect of subway access improvement on employment of firms. Note that this is essentially a difference-in-differences specification if I replace the continuous measure on the change of the nearest distance $\Delta D_i'$ with a dummy variable indicating treatment group.

$$\Delta lnL_i = \beta T_i + \Delta\tau + \Delta\varepsilon_i \tag{1.3}$$

where T_i is the dummy variable indicating the treatment group. Equation 1.2 offers more information relative to the standard difference-in-differences specification in equation 1.3 as it provides estimates on changes in employment with respect to changes in distance. I use

equation 1.2 for the main results while providing results for equation 1.3 in the Appendix.

Figure 1.3 illustrates how the equidistance line works, where the distance from the equidistance line to both subway lines are set to 2.5 km for an example. Consider firm i located on the boundary of the 1 km band on the side closest to Line 2E. The distance from the firm to the equidistance line is 0.5 km. The nearest distance to subway services D_{i0} is 3 km before Line 2E opening, and reduces to $D_{i1} = 2$ km after Line 2E opening. Therefore, $\Delta D_i' = 1$ km. Note that the coverage of the treatment group is 0.5 km as marked in Figure 1.3.

This research design addresses the concern that the changes in the nearest distances ΔD_i are correlated with the time-varying unobservables $\Delta\varepsilon_i$ for two reasons. First, the equidistance line is not a direct target in the decision making of the placement of the subway line. Therefore, I expect that the unobserved characteristics of firms along the equidistance line are independent of the placement of the subway. Second, focusing on a narrow band along the equidistance line further amplifies the random variation rooted in the discontinuity in the nearest distances to subway services when crossing the equidistance line, while minimizing the endogenous elements in the location of the subway lines. This is essentially an "Inconsequential Units Approach" (Chandra & Thompson, 2000; Michaels, 2008). In the next section, I examine whether the observed characteristics are balanced across the equidistance line in the sample and then show the regression results.

1.5.3 Empirical Results

Table 1.4 summaries the differences in the level of firms characteristics in the year 2008 (first five columns) and the pretreatment trends in two key outcomes (last two columns) from the year 2004 to 2008 between the treatment area and the control area. I restrict the sample to survived non-mover firms to eliminate firm sorting. I show the results for three broad sectoral classifications in three panels. For manufacturer firms, the observed characteristics of firms in the treatment area and control area are balanced in both levels and pretreatment trends. This indicates that the treatment is "randomly assigned" for the manufacturers located along the band. However, the second and third panels show that service firms in the treatment group are at a disadvantage. Columns 1 and 2 suggest that revenue and employment of service firms in the treatment group are less than half of those in the control group. Nevertheless, the pretreatment growth of revenue and employment in Column 6 and 7 suggests that the two sides of the equidistance line follow a parallel economic trend despite the difference in the levels of economic outcomes. This supports the key identification assumption $Cov(\Delta D_{it}, \Delta\varepsilon_i) = 0$ as discussed in section 1.5.2. Therefore, I proceed with a difference-in-differences framework to estimate the causal effects.

Table 1.5 shows the coefficients from the regressions of the changes in log employment on the

changes in the nearest distances to subway stations. Results in Columns 1-3 use all survived firms that are observed in both year 2008 and 2013. Columns 4-6 focus on the survived non-mover firms. Columns 1 and 3 do not include any control variables. Columns 2 and 4 control for log revenue in the year 2008. Columns 3 and 6 include other control variables including firm age, SEZs dummy, plant dummy, and distances to the terminal station of Line 2 original part. Columns 7 and 8 show the results for the number of new establishments and log employment conditional on entry with all controls except for lag revenue. In the first panel, Columns 1-3 show that a 1 km reduction in the nearest distances to subways increases employment by about 61% ($= exp(0.477) - 1$) to 67% ($= exp(0.512) - 1$). The point estimates are large and highly statistically significant. It is not a surprise to find adding control variables barely change the coefficients, because the sample is very balanced between the groups, which provides strong evidence supporting that the treatment is "as good as random assigned". Columns 4-6 exclude movers to eliminate the potential bias due to sorting. The remaining effects should be interpreted as the causal effect of subways on employment in firms. These effects are slightly smaller, falling to about 55% ($= exp(0.438) - 1$) to 58% ($= exp(0.460) - 1$). This suggests that firm sorting contributes to the employment growth in the treatment group; however, the effect from sorting is about 10% to 13% of the effect estimated from survived firms, which is relatively small, but not negligible. Column 7 shows that the reduction in the nearest distances also increases the number of new establishment (about 12% per 1 km). However, conditional on firm entering, the size of new firms is not statistically different, although the estimate is imprecise due to small sample size (only 56 observations).

The second panel presents a similar pattern for producer service firms. The effects from survived firms are slightly larger than the effects found in manufacturers. After restricting the sample to the survived non-mover firms, the point estimates reduce but remain large; however, the estimates are imprecise due to small sample size. Also, I do not find clear evidence on the impact of the subway on new producer service firm entering. The last panel shows that consumer service firms in the treatment group, surprisingly, experience a decline after the new subway opening. A 1 km reduction of the nearest distance to subway decreases employment in survived firms by 32%($= exp(-0.382) - 1$), and employment in survived non-mover firms by about 22%($= exp(-0.246) - 1$). Additionally, Columns 7 and 8 show that despite no difference in firm entrants, consumer service firms in the treatment group hire fewer employees. One possible explanation is that manufacturers and producer service firms are more productive and can afford higher rent than consumer service firms and pushed the latter out of the area. Table A.1 presents the results using the standard difference-in-differences specification in equation 1.3. The results are consistent with the specification using the reduction in the nearest distances to subway stations as the treatment variable.

Table 1.6 presents what happens when I restrict the sample to different distances to the

terminal station of Line 2 original part, which approximates the importance of subways for an individual to commute to a firm located in the narrow band. The reason is that the geometry of the positions of Line 6, Line 2E and the equidistance line necessitates that the distance from the firm to its nearest subway station increases with its distance to the terminal station of Line 2 original part, as shown in Figure 1.2. Commuters may consider taking subways when stations are within walking distance or bus-connecting-trip distance, and are less likely to take subways when the nearest subway stations are beyond a certain threshold (Gibbons & Machin, 2005). I consider two alternative samples in Table 1.6, including firms that are within 6 km from the terminal station of Line 2 original part, and firms that are more than 10 km away. The comparison between Columns 1 and 2, Columns 4 and 5 shows that the effect of subway access improvement is stronger in the area within 6 km to the center relative to the baseline sample which uses 10 km distance threshold. In Columns 3 and 6, the effects of subways are mostly not significant, with point estimates decreasing sharply. This implies that the new subway access is somewhat irrelevant in the areas that are beyond the 10 km distance threshold to the new subway line. Notably, the coefficients for consumer service firms remain significantly negative. This may echo the results in Tsivanidis (2019), who finds that the effect on commute distances is mildly greater for low-skill workers after transportation accessibility improves. The longer commute distance for low-skill workers may explain why subway access can still influence employment in the relative further area, given that the consumer service sector is likely to employ a larger proportion of low-skill workers relative to the other sectors. The reason is that manufactures in Pudong New Area are mostly in advanced manufacturing industry, for example, microelectronics, and therefore the skill levels of labor in manufacture firms are likely to be higher than those in consumer service firms. However, it requires firm-level data on skill levels of employees to pin down the underlying mechanism, which I do not have access to.

Table 1.7 shows that the results presented are robust to changing the width of the narrow band along the equidistance line. Columns 2 and 5 show the effects in the baseline sample. Columns 1 and 4 reduce the bandwidth to 800 meters. The effects appear to be stronger for manufacturer firms and producer service firms relative to the baseline sample. Columns 3 and 6 expand the bandwidth to 1200 meters. The point estimates decrease slightly for manufacturer firms but increase modestly for producer service firms. Now the effect for producer service firms from non-movers in Column 6 in the second panel is marginally statistically significant. In the bottom panel, the estimates for consumer service firms remain negative when the bandwidth changes, but are noisier. In sum, the results earlier are robust to choosing alternative bandwidths. However, expanding the bandwidth is at the cost of losing the balance between the treatment and control groups. The upper panel in Figure 1.4 presents the effect under a broader range of distance bands by the three categories of sectors. The effects diminish gradually as the bandwidths increase, reaching almost zero at the 2 km

distance threshold. The data provide an opportunity to conduct a placebo test by regressing the change of log employment from 2004 to 2008 when the treatment had not occurred on the changes in the nearest distances to subways. The placebo estimates are shown in the lower panel in Figure 1.4. They are mostly very close to zero and not significant, which reassures that the results in the upper channel are not false positives.

Taken together, these results estimated from firms along the equidistance line provide extra insights to the earlier results estimated on a location basis. The overall employment may experience little growth; however, the new subway access caused the sectoral mix in this area to change substantially.

1.6 Conclusion

The new subway service appears not to improve the overall employment in its surrounding area, but is likely to improve new firm entrants, and empower locations in the Special Economic Zones (SEZs) to attract workers. However, due to the endogenous placement of subway lines, simply comparing the locations next to subway stations with locations further off yields biased estimates. By tracking the locations of firms before and after the subway opening, I show evidence on firms sorting based on the gains from access to subways, which leads to selection bias in the estimates. This paper develops a novel identification strategy to overcome these issues. I find robust causal evidence supporting that subways increase employment in manufacturer firms, while decreasing employment in consumer service firms, surprisingly. The point estimates for producer service firms are large and positive, but less precise. Therefore, the new subway service causes a substantial change in the sectoral mix in this area.

Admittedly, these empirical results come with limitations. First, firms located in the narrow band along the equidistance line are mostly far from the subway stations, which indicates that they may be less competitive relative to firms next the subway locations. Thus, the causal effect estimated from these firms may not be generalizable to the whole population. Second, I cannot distinguish growth effects from reorganization effects at either the location level or firm level. The increase in employment in a location or firm that is close to the subway may be at the cost of locations or firms that in other locations in the district. Thus, the effects presented so far may not be informative about the general equilibrium effects. Third, the mechanism underlying the heterogeneous effects across sectors remains a puzzle. Possible mechanisms include high productivity firms crowding out low productivity firms, and skillful labors crowding out low skill labors. The mechanisms can be better studied using data on firm-level productivity, skill levels of employees and land values, which is unavailable in this study.

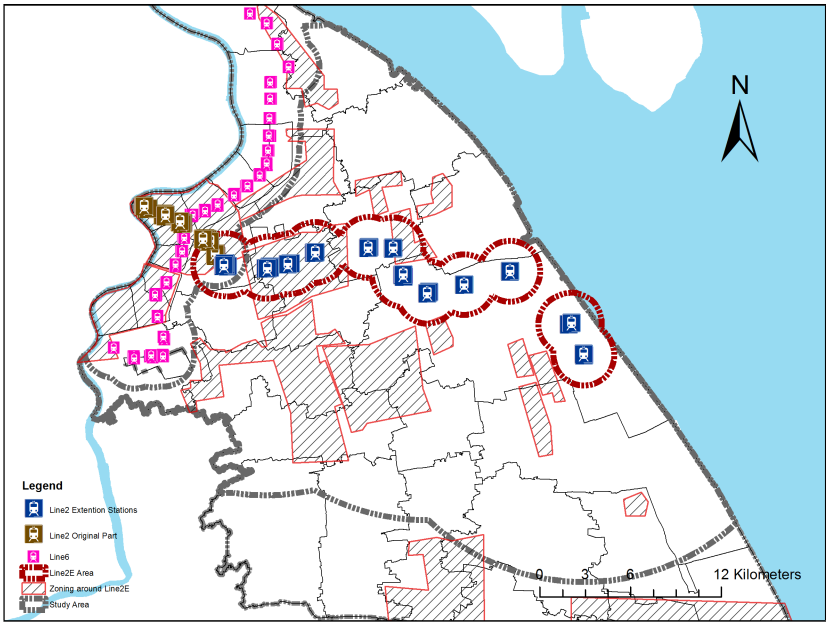


Figure 1.1: Study area ,treatment area and control area

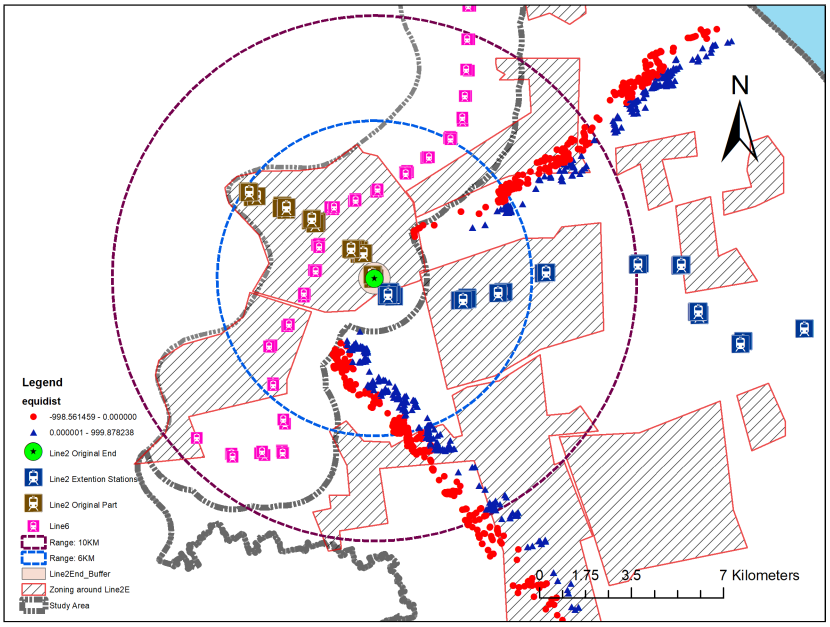


Figure 1.2: Equidistance line to Line 2E and Line 6

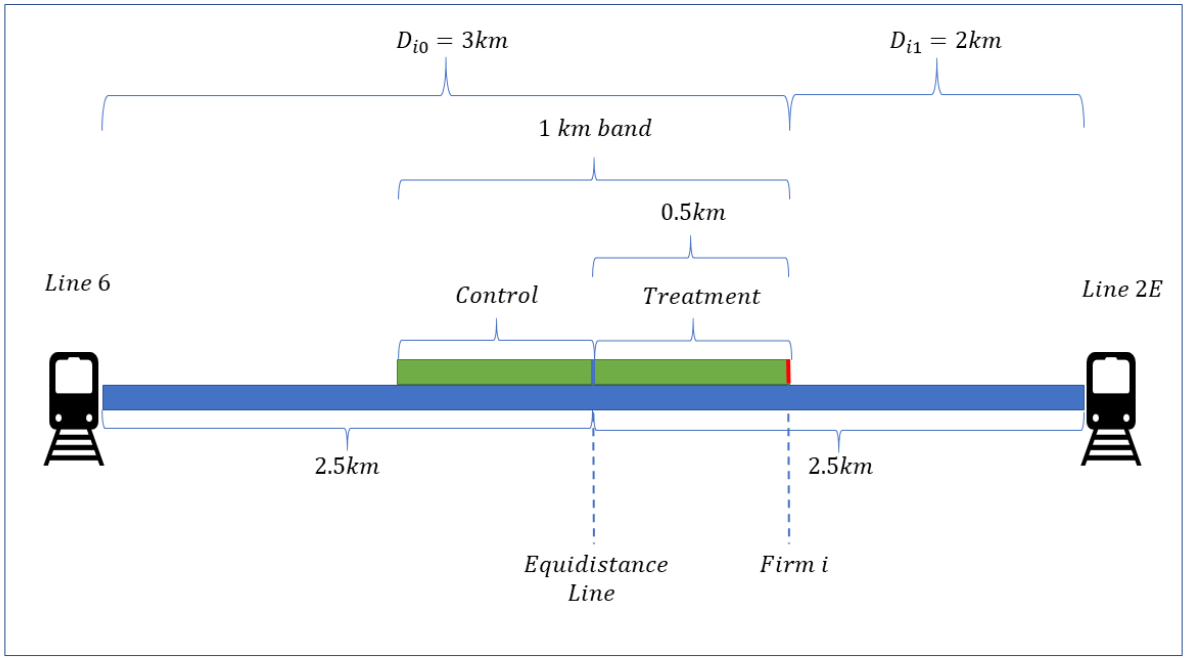


Figure 1.3: Illustration on how the equidistance line works

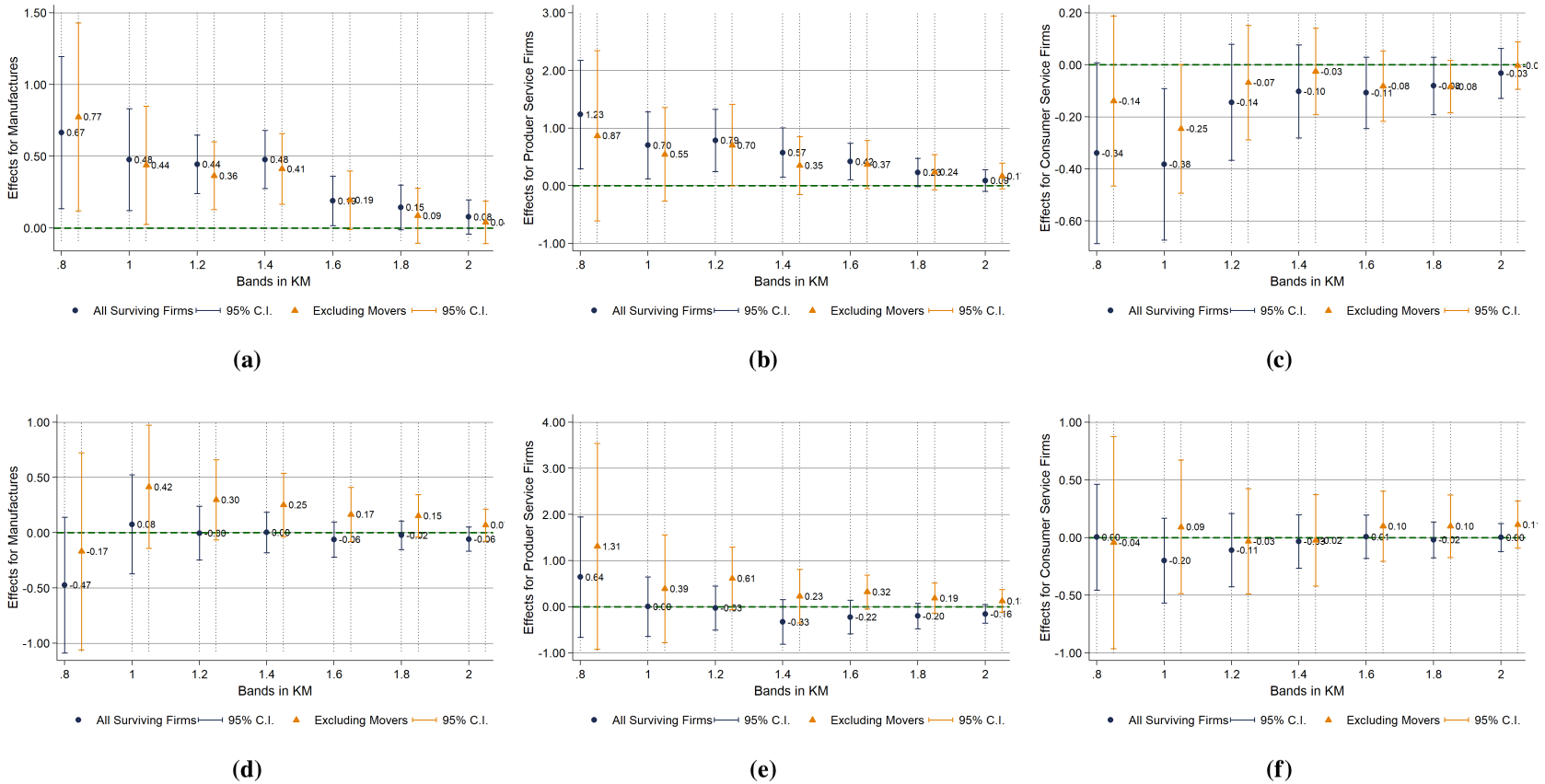


Figure 1.4: The effect of subway access on employment in difference distance bands and years

Note: The figures show bandwidths measured in kilometers on the x-axis. Each band corresponds to a sample of firms located from the equidistance line up to the distance of the band on each side. The y-axis represents the estimates from equation 1.2. Each dot is a coefficient and its value is shown alongside. The capped bars indicate 95% confidence intervals. The upper panel shows the results using data from 2008 to 2013. The lower panel shows the results using data from 2004 to 2008. By column, the graphs show results for manufacturers (a,d), producer service firms (b,e), and consumer service firms (c,f), respectively. The samples only include firms located up to 10 km from the terminal station of the Line 2 original part. Robust standard errors are applied for estimating the confidence intervals.

Table 1.1: Difference-in-difference estimates

	All		Treatments		Controls		Estimates	
	Before	After	Before	After	Before	After	Dif-in-dif	T-stat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel (a): All areas								
Ln employment	2.78	3.02	3.6	3.87	2.64	2.88	0.03	0.38
Ln revenue	5.75	6.49	7.31	7.98	5.49	6.25	-0.09	-0.45
Ln new employment	1.58	1.71	2.06	2.33	1.5	1.61	0.16	1.7
Ln new revenue	3.64	4.2	4.69	5.36	3.47	4.01	0.13	0.65
Number of observations	2921		408		2513			
Panel (b): non-SEZs areas								
Ln employment	2.54	2.76	2.94	3.09	2.47	2.71	-0.09	-0.92
Ln revenue	5.28	6.03	6.14	6.63	5.15	5.94	-0.3	-1.39
Ln new employment	1.41	1.55	1.49	1.74	1.4	1.53	0.12	1.35
Ln new revenue	3.33	3.87	3.69	4.26	3.27	3.82	0.02	0.1
Number of observations	2307		297		2010			
Panel (c): SEZs areas								
Ln employment	3.69	3.98	5.36	5.96	3.32	3.55	0.37	1.99
Ln revenue	7.49	8.23	10.42	11.58	6.85	7.49	0.52	1.41
Ln new employment	2.18	2.28	3.58	3.9	1.87	1.92	0.27	1.11
Ln new revenue	4.82	5.44	7.35	8.32	4.26	4.81	0.42	0.88
Number of observations	614		111		503			

Notes: Unequal variances are applied in the mean-comparison tests. The outcomes are the average of log employment in a 500×500 meters grid. The number of grids are the same in both periods.

Table 1.2: Summary statistics by firm migration groups

	Stay-in	Move-out	Move-in	Stay-out	Stay-out-mover
	(1)	(2)	(3)	(4)	(5)
Change of ln employment	-0.075	-0.147	0.077	-0.092	-0.127
	(0.861)	(1.012)	(0.910)	(0.946)	(0.856)
Lag ln revenue	7.725	7.866	8.099	7.912	7.882
	(2.736)	(2.381)	(2.369)	(2.236)	(2.415)
Dummy manufacture	0.250	0.544	0.360	0.496	0.406
	(0.433)	(0.499)	(0.481)	(0.500)	(0.491)
Dummy PSFs	0.288	0.202	0.245	0.171	0.159
	(0.453)	(0.402)	(0.431)	(0.376)	(0.366)
Dummy CSFs	0.462	0.254	0.395	0.334	0.434
	(0.499)	(0.436)	(0.490)	(0.472)	(0.496)
Age	10.507	10.746	10.067	10.447	10.710
	(4.685)	(4.389)	(4.037)	(4.536)	(4.700)
SEZ	0.387	0.147	0.301	0.129	0.112
	(0.487)	(0.354)	(0.459)	(0.335)	(0.316)
Plant dummy	0.176	0.094	0.094	0.083	0.132
	(0.381)	(0.292)	(0.293)	(0.276)	(0.338)
N	2304	531	286	4563	19601

Note: Stay-in group includes firms that stayed in the area within 2 km of Line 2E stations; move-out group include firms that previously existed within a 2 km radius of Line 2E stations, but moved out of this area in the between the years 2008 to 2013; move-in group includes firms that were outside of this area but moved in during this period; stay-out group includes firms that stay out of this area during this period; and stay-out-mover group include firms that move their locations further than 500 meters but have been at least 2 km away from subway stations in both 2008 and 2013. For each variable, means are shown in the first row and standard deviations are shown in the second row in parentheses.

Table 1.3: Spatial sorting of firms

	2008-2013				2004-2008			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Move-out	Move-in	Move-in	Move-in	Move-out	Move-in	Move-in	Move-in
Movers	-0.013	0.176***	0.126**	0.174***	0.010	0.028	-0.013	0.075
	(0.049)	(0.053)	(0.055)	(0.055)	(0.043)	(0.058)	(0.059)	(0.063)
R^2	0.019	0.020	0.033	0.032	0.024	0.013	0.017	0.026
N	2835	19887	4849	2590	2045	14369	4484	1854

Note: These regressions have different samples. Columns 1 and 5 include move-out and stay-in firms; Columns 2 and 6 include move-in and stay-out firms; Columns 3 and 7 include move-in and stay-out-mover firms; Columns 4 and 8 include move-in and stay-in firms. All regressions control for lag log revenue, manufacture dummy, PSFs dummy, CSFs dummy, age, SEZ dummy, and plant dummy. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 1.4: Balancing regressions: Differences in levels in 2008 and trends from 2004 to 2008 between treatment and control areas, by sectors, within 1km band

	Levels in 2008					Trends from 2004 to 2008	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Revenue	Employment	Age	SEZ	Plant	Revenue Growth	Employment Growth
Manufacture firms							
Treated	-0.210	-0.095	0.360	-0.050	-0.021	0.034	0.087
	(0.589)	(0.384)	(1.025)	(0.104)	(0.043)	(0.196)	(0.166)
N	103	103	103	103	103	121	121
Producer service firms							
Treated	-1.780***	-0.893***	0.946	-0.099	-0.110*	-0.027	0.207
	(0.527)	(0.299)	(0.812)	(0.081)	(0.066)	(0.656)	(0.453)
N	80	80	80	80	80	33	33
Consumer service firms							
Treated	-0.847**	-0.019	2.159***	-0.006	0.026	-0.416	-0.014
	(0.344)	(0.174)	(0.620)	(0.041)	(0.054)	(0.377)	(0.193)
N	311	311	311	311	311	76	76

Note: This table summarizes the differences in five firm characteristics in 2008 and two key outcomes from 2004 to 2008 between the treatment and control areas. The samples only include survived non-movers firms and only include firms located up to 10 km from the district center and within 1km to the equidistance line on both sides. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 1.5: The effect of reduction in nearest distances to subway on employment

	All survived Firms			Excluding Movers			New firms	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Manufacture firms								
Changes in nearest distance	0.512***	0.520***	0.477***	0.460**	0.457**	0.438**	0.124**	-0.392
	(0.171)	(0.170)	(0.179)	(0.198)	(0.198)	(0.207)	(0.060)	(0.447)
Adjusted R^2	0.03	0.03	0.04	0.03	0.02	0.02	0.47	-0.01
N	162	162	162	103	103	103	264	56
Producer service firms								
Changes in nearest distance	0.638**	0.625**	0.701**	0.561	0.484	0.545	0.052	-0.031
	(0.319)	(0.300)	(0.294)	(0.378)	(0.351)	(0.406)	(0.059)	(0.308)
Adjusted R^2	0.02	0.01	0.03	0.01	0.00	-0.02	0.61	0.02
N	119	119	119	80	80	80	357	148
Consumer service firms								
Changes in nearest distance	-0.418***	-0.449***	-0.382**	-0.293**	-0.324**	-0.246*	0.016	-0.383**
	(0.142)	(0.147)	(0.148)	(0.129)	(0.129)	(0.126)	(0.037)	(0.187)
Adjusted R^2	0.02	0.02	0.04	0.01	0.01	0.03	0.57	0.03
N	414	414	414	311	311	311	1401	705

Note: The dependent variable is changes in log employment. The key regressor of interest is changes in the nearest distance to subway stations. The sample in Columns 1-3 includes all survived firms that are observed in both year 2008 and 2013. The sample in Column 4-6 excludes movers. Column 1 and 4 do not include any control variables. Column 2 and 5 control for log revenue in the year 2008. Column 3 and 6 include other control variables including age, SEZ dummy, plant dummy, and distances to the terminal station of Line 2 original part. Column 7 and 8 include all control variables except for lag revenue. All samples include firms located up to 10 km from the terminal station of Line 2 original part and within 1km to the equidistance line on both sides. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 1.6: Robustness regressions: Different distances to the district centre

	All survived Firms			Excluding Movers		
	(1)	(2)	(3)	(4)	(5)	(6)
	<6 km	<10 km	>10 km	<6 km	<10 km	>10 km
Manufacture firms						
Changes in nearest distance	0.977***	0.477***	-0.042	1.194***	0.438**	0.204
	(0.297)	(0.179)	(0.151)	(0.375)	(0.207)	(0.181)
Adjusted R^2	0.05	0.04	0.03	0.04	0.02	0.03
N	59	162	334	38	103	177
Producer service firms						
Changes in nearest distance	0.853**	0.701**	-0.613	0.499	0.545	-1.327
	(0.364)	(0.294)	(0.539)	(0.481)	(0.406)	(0.806)
Adjusted R^2	0.02	0.03	-0.00	-0.01	-0.02	-0.15
N	86	119	47	63	80	21
Consumer service firms						
Changes in nearest distance	-0.468***	-0.382**	-0.711**	-0.339**	-0.246*	-1.185***
	(0.155)	(0.148)	(0.280)	(0.133)	(0.126)	(0.430)
Adjusted R^2	0.06	0.04	0.05	0.04	0.03	0.11
N	331	414	152	260	311	92

Note: The dependent variable is changes in log employment. The key regressor of interest is changes in the nearest distance to subway stations. The samples include firms located within 1km to the equidistance line on both sides. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 1.7: Robustness regressions: Different distance bandwidths

	All survived Firms			Excluding Movers		
	(1)	(2)	(3)	(4)	(5)	(6)
	800 m	1000 m	1200 m	800 m	1000 m	1200 m
Manufacture firms						
Changes in nearest distance	0.666**	0.477***	0.444***	0.775**	0.438**	0.365***
	(0.267)	(0.179)	(0.103)	(0.328)	(0.207)	(0.120)
Adjusted R^2	0.03	0.04	0.07	0.08	0.02	0.03
N	121	162	221	79	103	143
Producer service firms						
Changes in nearest distance	1.235**	0.701**	0.786***	0.867	0.545	0.702*
	(0.472)	(0.294)	(0.272)	(0.734)	(0.406)	(0.356)
Adjusted R^2	0.02	0.03	0.03	-0.06	-0.02	-0.01
N	84	119	143	55	80	98
Consumer service firms						
Changes in nearest distance	-0.339*	-0.382**	-0.144	-0.139	-0.246*	-0.068
	(0.176)	(0.148)	(0.113)	(0.166)	(0.126)	(0.112)
Adjusted R^2	0.04	0.04	0.03	0.01	0.03	0.03
N	295	414	482	222	311	359

Note: The dependent variable is changes in log employment. The key regressor of interest is changes in the nearest distance to subway stations. The samples include firms located up to 10 km from the terminal station of the Line 2 original part. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Chapter 2

Does E-commerce Reduce Traffic Congestion?

2.1 Introduction

As the digital economy takes shape, e-commerce continues to grow around the world. The past decade has seen an explosive growth of the Chinese e-commerce market, with the Gross Merchant Value increased by more than ten times[1]. Today, more than 40% of the world's e-commerce transactions take place in China, a dramatic increase from only 1% about a decade ago. In the year 2016, there were about 460 million online consumers and the share of online consumption is about 12.6% of total consumption. Traditional retail involves traffic both from warehouses to stores and from consumers to stores. E-commerce cuts intermediate traffic by delivering goods directly from the warehouses to the consumers and improve the efficiency of logistics of goods. This poses the question of whether e-commerce is more traffic efficient than traditional retail. This paper investigates the possible reduction of traffic channeled through the rapid expansion of e-commerce and its new logistics.

Although plenty of evidence has shown that vans that are servicing the e-commerce are a growing contributor to traffic and congestion, consumers are also found making less shopping trips using vehicles (Braithwaite, 2017). The trade-off between the two effects is crucial to assess the overall effect of e-commerce on traffic. Punakivi (2003) simulated the replacement of traditional retailing by electronic retailing and found that this potentially leads to 54-95% reduction in traffic depending on delivery methods. In a similar vein, Cairns (2005) estimates that a direct substitution of car trips by van trips could reduce vehicle-km by at least 70%. A recent comprehensive report by Braithwaite (2017) gathered suggestive evidence indicating

[1]According to the Chinese E-commerce Association. http://www.100ec.cn/zt/upload_data/2018dzswfzbg.pdf

that online shopping is likely to reduce overall shopping traffic in probably modest scale in the real world. Measuring congestion reduction is difficult because of latent travel demand suppressed by traffic congestion itself. When traffic congestion was reduced, individuals who chose not to travel may decide to travel after observing less traffic. In fact, the reduction of traffic congestion might never be observed in the long term due to the fundamental law of road traffic congestion (Duranton & Turner, 2011).

In this paper, I provide the first available estimates of the effect of e-commerce on traffic and congestion. The analysis utilizes temporary price shocks caused by a nationwide online shopping event in 2016 as the foundation of my identification strategy. The Singles' Day shopping event on the day of 11 November each year is the largest online sales event in China, equivalently as popular as the shopping event of Cyber Monday in the United States[2]. In the year 2016, the largest Chinese online shopping platform, Alibaba, reached a new sales record at 17.6 billion US dollars during a 24-hours promotion period. To put this figure into context, the annual online consumption is 587 billion US dollars. The average price in the online channel on the event day is about 80% of the average price in the month before the event[3].

A reduced price in the online channel during the sale event encourages consumers to switch from offline channel to online channel. I measure the change in traffic congestion in each hour one week before and two weeks after the event. Traffic congestion is measured by an index that is the ratio of actual passing time and free-flow passing time of vehicles in a road segment. The road-level information is first collected by Global Positioning System (GPS) trackers based on millions of users of a navigation service company, then aggregated to city-hour level, and finally, released for public use. Change in online shopping is measured by Baidu Index, which is similar to Google Trends but can track search Internet Protocol (IP) addresses to cities. In the week after the event, intracity traffic congestion dropped by 1.7% during peak hours and 1% during off-peak hours, while online shopping increased by about 1.6 times. Cities with a higher increase in online shopping experienced a greater reduction in traffic congestion.

How does online shopping affect traffic congestion? To answer this question, I derive the relationship between online consumption quantity, vehicle demand, and a traffic congestion index. Based on the speed-density relationship proposed in Adler et al. (2017), the logarithm of the traffic congestion index can be expressed as the ratio of vehicle density to the free-flow

[2]The concept of Singles' Day was initiated by young college students to buy presents to celebrate singlehood and resist social pressure to get married. 11 November was chosen as the date because each of its lone digits (11.11) represents a "bare stick" as the symbol of uncoupled individuals. The day was then promoted by Alibaba as a cultural event of online shopping, bearing similarity with Cyber Monday in the United States.

[3]The online shopping platform did not publish relevant price change information. An independent credit rating company, CCX Credit Technology, monitored the price of a sample of online products surrounding the event and published the findings online (see http://www.01caijing.com/article/12269.htm)

vehicle density. The change in the logarithm of the traffic congestion index is decided by the change in traffic density, which in turn is decided by the change in vehicle demand. If the vehicle demand per unit goods of online shopping is only a fraction of that of offline shopping, which I call a "vehicle-saving ratio", then the total vehicle demand for shopping changes with the online-offline substitution of consumption induced by the price shock. With a one unit increase in online consumption, the net reduction in traffic is the difference between the following two parts: the reduction in offline shopping vehicle subject to the online-offline substitution of consumption and the share of shopping made through private vehicles, and the increase in vehicle demand arising from online shopping. I call this net reduction a "traffic-saving factor". In addition, I measure the online-offline substitution using the ratio of the increase in online consumption over the reduction in offline consumption under a price shock, which I call a "online-offline substitution ratio". With simple accounting of traffic, the model further reveals that the elasticity of traffic congestion index to online consumption quantity is the traffic-saving factor adjusted by the current congestion level and a function of three shares: the share of online shopping, the share of shopping made through private vehicles, and the share of vehicles used due to shopping. Particularly, the condition for online shopping to reduce traffic is the traffic-saving factor being negative. To be specific, the condition requires that the vehicle-saving ratio is sufficiently low, or the amount of offline shopping that consumers are willing to substitute with online shopping is sufficiently large. The first part can be estimated from an operations management perspective using data relatively accessible. For example, the vehicle-saving ratio can be estimated using data on the number of online goods delivered in a van in an hour, and the number of offline goods purchased in a private vehicle[4]. However, the second part requires data that are more difficult to obtain[5]. For this reason, I developed a demand model to predict the online-offline substitution ratio.

The model is characterized by two key assumptions. First, the utility that a consumer obtains from a product depends on the matching quality between the consumer and the product. The matching quality is assumed to be a random variable that follows the Fréchet distribution. The quantity of consumption adjusted by matching quality is aggregated across products by a CES functional form to derive a consumer's utility. Second, I assume a mechanism of how consumers choose shopping channels for a product. The matching quality is assumed to be the maximum of two underlying channel-specific draws of matching quality. Specifically, given a product, a consumer draws an online matching quality from an online Fréchet distribution, and then draw an offline matching quality from an offline Fréchet distribution. The consumer

[4]I provide a crude estimate in Section 2.2.3 based on this simple logic. The vehicle-saving ratio is about 0.067.

[5]The data required to measure the online-offline substitution ratio empirically are unavailable. It requires weekly data on the online and offline consumption surrounding the event day. Even if the online shopping platform would like to share the online consumption surrounding the event, obtaining offline shopping consumption on a weekly basis remains challenging. Potentially, I could conduct a consumer survey, but the large data requirements are outside the scope of this study.

then chooses the maximum of the two draws, and thus self-selects into a type of either online consumer or offline consumer based on the channel that yields the maximum draw[6]. I assume that the two channel-specific Fréchet distributions have the same shape (which decides variability) but different scales (a higher scale means that a high-value draw is more likely). The relative value of the channel-specific scale parameters, along with the shape parameter, determine the probability under which consumers choose each channel. Intuitively, the channel with a larger scale parameter attracts a higher proportion of consumers.

As the maximum of two Fréchet random variables also follows a Fréchet distribution[7], I can derive the overall demand (sum of demand from the two channels) for a specific product. The expectation of the overall demand for the product is used by monopolist firms to set the equilibrium price, which is the marginal cost of the product multiplied by a constant mark up – a well-known result under the assumptions of CES utility and monopolistic firms. Given the fixed price of a specific product, consumers divide into online or offline consumers according to the probability under which consumers choose channels. Conditional on being online (or offline) consumers, the quantity consumed within these online (offline) consumers turns out to follow the same distribution as the overall quantity consumed by all consumers. In other words, the demand distribution in each channel is independent of the channel. This is similar to a key finding in Eaton & Kortum (2002)[8]. I find that this independence property has two important implications[9]:

(i) The expected values of the quantity consumed in each channel are equal.

(ii) The expected channel-specific quantity consumed equals the share of consumers that choose that channel multiplied by the expected overall demand from both channels.

Importantly, these properties hold when prices change. As a result, the change in channel-specific demand due to price shock can be approximated by a derivative formula. This allows for calculating the ratio of the increase in online shopping quantity over the reduction in offline shopping quantity, which gives the online-offline substitution ratio. Further, the ratio can be expressed as a concise function of the elasticity of online shopping quantity to the relative price of online to offline channel, and the elasticity of substitution between varieties. The former can be estimated from the data, and the latter is a well-studied parameter in the trade literature.

Armed with the formula of the online-offline substitution ratio, I conduct a quantitative analysis to explore whether the condition for online shopping to reduce traffic congestion holds

[6]The same consumer can be an online consumer for one product and an offline consumer for another product.

[7]The parameters of the former can be derived from the parameters of the latter.

[8]One of the key findings in Eaton & Kortum (2002) is that the price distribution of the varieties that any given origin actually sends to any given destination is independent of its origin regions.

[9]Note these properties rely heavily on the assumptions of the Fréchet distribution.

and the quantitative importance of the elasticity of traffic congestion to online consumption quantity, with reasonable guesses on the parameters and sample statistics of variables in the model. Particularly, the online-offline substitution ratio is estimated to be around -1.9, which indicates that the reduction in offline consumption due to one unit increase in online shopping is about a half unit. Several well-educated guesses on the parameters in the quantitative analysis show that the condition is likely to hold. However, the elasticity of traffic congestion to online consumption quantity appears to have a wide range of values due to heterogeneity. For example, the estimates vary from -0.06 to -0.25 for different cities. Therefore, in order to identify an average magnitude of the elasticity, I turn to empirical estimates.

In the empirical section, Ordinary Least Square (OLS) regression models estimate that a 10% increase in online shopping reduces traffic congestion by about 0.13%, an elasticity of -0.013. The effect mostly comes from peak hours, that is from 9 am to 11 am and from 7 pm to midnight. To address potential endogeneity problems arising from omitted variable bias, I instrument the change of online shopping in the event with the reduction in postage fee. Postage fee was waived on the day of the shopping event and introduces exogenous incentives for consumers to switch to online shopping conditional on market access. Using the waived online delivery postage fee as the instrumental variable (IV), the IV estimation reveals much stronger effects than the OLS estimates. A 10% increase in online shopping reduces traffic congestion by about 1.4%, an elasticity of -0.14, which is consistent with the range from the quantitative analysis. The weights in the IV estimates appear to be assigned towards cities that experience a higher increase in online shopping. Using the IV estimates to conduct a welfare analysis for Beijing, the congestion relief effect of 10% increase in online shopping can be converted to monetary terms of 239 million US dollars a year for peak hour commuters, which is equivalent to about a third of the average effect of providing access to an additional subway line, according to the welfare gains from the congestion relief effect of subways estimated by Gu et al. (2019).

This paper contributes to a growing literature in the spatial economic impact of the digital economy, as reviewed in Goldfarb & Tucker (2019). As a digital purchasing technology, e-commerce reduces transportation cost. On the consumption side, consumers choose online shopping to reduce travel cost despite certain disutility in online shopping (Forman et al., 2009). E-commerce also contributes to overcoming the logistical barrier in rural areas and leads to sizable gains in real household income. Consumers in villages benefit from greater product variety and lower prices driven by a significant reduction in travel costs (Goolsbee & Klenow, 2018; Couture et al., 2018). Consequently, e-commerce reduces spatial inequality of consumption between large cities and small cities, and increases access to varieties (Dolfen et al., 2019; Fan et al., 2018). On the supply side, online shopping causes structural changes in offline shopping economies. It shifts the market from high-cost producers to low-cost producers (Goldmanis et al., 2010). In the retail industry, offline retailers that directly compete

with online retailers are negatively impacted, while indirect competitors that can adapt the change in e-commerce revolution and take advantage of the online-offline complementarities can be winners in the competition (Relihan, 2017). Dolfen et al. (2019) concludes that consumer gains are about 1.1% of all consumption using credit card data in the US, which is tantamount to 1,150 US Dollars per household in the year of 2017. This paper advances our understanding of the impact of e-commerce on urban traffic congestion.

This paper is also linked to a long literature in traffic congestion relief policies. Many policy options have been extensively studied in the past, for example, congestion charge (Yang et al., 2018b; Tikoudis et al., 2015) and quantity-based restriction (Yang et al., 2014) on the demand side; transport infrastructure expansion and subsidies (Parry & Small, 2009; Anderson, 2014; Yang et al., 2018a; Gu et al., 2019) on the supply side. These measures are either politically controversial or expensive (Adler et al., 2017). The promotion of e-commerce is considered a "soft" policy that seeks to encourage people to reduce their car usage through enhancing the awareness and attractiveness of alternative options (Cairns et al., 2004).

The model in this paper draws insights from three strands of models. The first strand is a long list of trade models with constant elasticity demand (Armington, 1969; Krugman, 1980; Eaton & Kortum, 2002; Helpman et al., 2004; Antràs et al., 2017). The elements in these models are helpful in understanding trade flows across locations. Particularly, Fan et al. (2018) and Dolfen et al. (2019) have applied this type of trade model to analyze e-commerce. The second strand of models is in the literature of marketing economics and industrial organization. One key insight is that firms can employ random sales to compete over consumers with lower willingness to pay and reach higher profit, relative to only serving a fraction of consumers with higher willingness to pay (Varian, 1980; Seim & Sinkinson, 2016). However, their study does not allow for consumers with continuously distributed taste. Drawing insights from random coefficients demand models such as Coşar et al. (2018) and Berry et al. (1995), I introduced heterogeneous consumers into the demand side but specify their tastes for channels following Fréchet distribution [10]. Different to their study, I focus on the online retail event and analyze the substitution between online and offline products when the relative prices change during a sale. The third strand includes models developed by transportation engineers that study the relationships between speed, density and flow. Particularly, I adopt the speed-density relationship proposed as in Adler et al. (2017), and derive an exact functional form that links the change of online consumption, vehicle demand and traffic congestion index.

This paper is structured as follows. Section 2.2 sets out a model to connect the change in online shopping in the event with the change in traffic congestion. Section 2.3 outlines data and shows a descriptive analysis. Section 2.4 delves into the econometric framework.

[10] Redding (2016) specified worker's taste for amenity to follow Fréchet distribution.

Section 2.5 presents initial evidence on the link between the increase in online shopping and the reduction in traffic congestion. Section 2.6 presents the causal estimates of the effect of online shopping on traffic congestion. Section 2.7 analyzes the welfare impact of congestion relief effect of e-commerce. Section 2.8 discusses the long-term effect. Section 2.9 concludes.

2.2 Theoretical Framework

This section provides a theoretical framework to link the price change in online shopping and traffic congestion. The first part of the section lays out a demand model to predict the online-offline substitution. The analysis of the quantity demanded in both channels draws from Eaton & Kortum (2002)'s approach in analyzing the price distribution across origin countries. Although I use a multiple regions trade model framework, channel choices and online-offline substitution are the main focus. The key insight of the model remains even if reducing the multiple regions in the model to a single city. The second part of the section shows the derivation of the exact functional form for the relationship between traffic congestion index and online shopping quantity, and the condition for online shopping to reduce traffic congestion. The last part of the section explores the quantitative importance of the congestion relief effect of online shopping based on the model. In contrast to the introduction section, I start with the micro-foundation of the model and then move on to the aggregate consumption by channels in cities and its relationship with traffic congestion, which is a macro-phenomenon.

2.2.1 Demand Model and Online-Offline Substitution of Consumption

I develop a demand model to quantify the online-offline substitution of consumption (quantity). On the demand side, I allow for heterogeneous consumers. For a given variety[11], Fréchet distributed consumer-variety matching quality separates consumers into two types: online and offline consumers. On the supply side, I assume that each firm produces a unique variety and acts as a monopolist. It prices the variety uniformly in the two market channels, based on expected demand. In order to analyze a series of changes in the economy during the sale event, I introduce exogenous price shocks in both channels so that firms can price discriminate temporarily[12] across the two channels. When online prices decrease, online consumers will consume more, and some offline consumers will switch their types to online consumers. As a result, the online shopping event alters the share of online shopping and offline shopping and the overall amount of online shopping at the aggregate level. The ratio of the increase in online shopping quantity over the reduction in offline shopping quantity gives the online-offline substitution of consumption.

[11] I use the word variety and product interchangeably in the paper.

[12] There was no parallel offline sales during the event in 2016. Competing offline sales started in 2017.

2.2.1.1 The Set-Up

There are I cities, with each city indexed by i or j depending on whether the region in question is the origin, i, or the destination, j, of a trade. There is a continuum of consumers indexed by μ in city j, with the set of consumers U_j and the mass of consumers L_i. Each city is endowed with L_i units of workers (the same as the the mass of consumers) where each worker supplies one unit of labor inelastically and receives wage w_i. Suppose that labor is the only factor of production. Consumers buy (or firms sell) from online and offline channels, indexed by $m \in \{o, f\}$. Each variety is indexed by ω. Suppose that every firm in the world produces a distinct variety. The concept of a city is thus a cluster of firms with the same productivity but distinct products, and consumers with the same income but different tastes for online shopping.

2.2.1.2 Consumers

Assume individuals obtain utility $z(\mu, \omega)q_j(\mu, \omega)$ from consuming $q_j(\mu, \omega)$ units of variety ω, where $z(\mu, \omega)$ is the matching quality for each pair of consumer μ and variety ω. Consumers maximize a CES objective:

$$u_j(\mu) = \left(\int_{\Omega_j} (z(\mu, \omega)q_j(\mu, \omega))^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} \tag{2.1}$$

I assume that the unobserved matching quality is a random component and follows the Fréchet (Type-II Extreme Value) distribution. This idea is fundamentally similar to Redding & Weinstein (2016) where their CES preference parameters are assumed to vary across both product type and consumer type[13][14][15].

I assume $z(\mu, \omega)$ is generated following the process below. Given variety ω, each consumer receives two draws $z_o(\mu, \omega)$ and $z_f(\mu, \omega)$ from two Fréchet distributions $F(\theta, s_o(\omega))$ and $F(\theta, s_f(\omega))$ for online and offline channels, respectively. The Fréchet distribution for channel m is:

$$F_m(z) \equiv Pr\{z_m(\mu, \omega) \leq z\} = exp\{-\left(\frac{z}{s_m(\omega)}\right)^{-\theta}\} \tag{2.2}$$

where θ is the shape parameter. It governs the amount of variation within the distribution. $s_m(\omega)$ captures general preference for a specific channel and is assumed to be the same for

[13]Their preference parameters are not random variables, but the authors introduce a Fréchet distributed shock as the multiplier to preference parameters, so the composite is equivalent to z here.

[14]As summarized in Redding & Weinstein (2016), heterogeneous random utility models have been studied in demand system estimation literature such as Coşar et al. (2018), Berry et al. (1995), McFadden (1973), etc.

[15]McFadden (1973) argues that consumers select the products that maximize their utility from a set of alternatives. Kortum (1997) shows a model where research leads to draws from a Pareto distribution, causing the technological frontier to be distributed Fréchet. As a conjecture, if consumers search for the best products in suiting their preferences, the search process may give rise to the Fréchet distribution of the matching quality.

all consumers but different across varieties (thus not index by μ). A bigger $s_m(\omega)$ implies a higher draw of matching quality $z_m(\mu, \omega)$ for any variety ω in channel m is more likely.

$$q_{mj}(\mu, \omega) = Y_j {P_j}^{\sigma-1} (\frac{p_j(\omega)}{k_m})^{-\sigma} z_m(\mu, \omega)^{\sigma-1} \tag{2.3}$$

where $p_j(\omega)$ is the price of variety ω in city j. It is the normal price listed by the retailers without considering any sales events. As shown later in this model, the price is marginal cost adjusted by a mark-up. k_m represents a channel-specific price shock[16], which captures any temporary changes of prices due to sales on the supply side. P_j is price index in city j[17],

$$P_j = (\int_{\Omega_j} (\frac{p_j(\omega)}{z_{max}(\mu, \omega)})^{1-\sigma})^{\frac{1}{1-\sigma}} \tag{2.4}$$

In a nutshell, consumers choose the shopping channel which gives higher $q_{mj}(\mu, \omega)$. This means that consumers compare $k_o^\sigma z_o(\mu, \omega)^{\sigma-1}$ with $k_f^\sigma z_f(\mu, \omega)^{\sigma-1}$ and pick the channel that yields higher value. The quantity consumed is the max of $q_{oj}(\mu, \omega)$ and $q_{fj}(\mu, \omega)$. The assumed process above provides the mechanism to decide the type of the consumer (online versus offline) and the probability of making such choices. Consumer μ in city j maximizes utility under the constraint of income Y_j. Here, consumers in the same city are assumed to have the same income, for model simplicity. The demand function of an individual consumer μ is,

$$q_j(\mu, \omega) = max\{q_{oj}(\mu, \omega), q_{fj}(\mu, \omega)\} \tag{2.5}$$

Given the distribution of channel specific matching quality, $z_m(\mu, \omega)$, the distribution of $q_j(\mu)$ in consumers for variety ω in channel m is

$$\begin{aligned} G_{mj}(q) &\equiv Pr\{q(z_m(\mu, \omega)) \leq q)\} \\ &= exp\{-(\frac{q}{p_j(\omega)^{-\sigma} Y_j {P_j}^{\sigma-1} (k_m^{\frac{\sigma}{\sigma-1}} s_m(\omega))^{\sigma-1}})^{-\frac{\theta}{\sigma-1}}\} \end{aligned} \tag{2.6}$$

Consumers choose the channel that *potentially* gives the higher quantity, so the distribution

[16] k_m is defined in the way that k_o increases from 1 to about 1.25, when there is 20% online discount.

[17] I omit the k_m from the price index for three reasons: First, k_m is always one when there is not a sales event. Second, I assume that the price shock will not be large enough to affect the overall price index. Third, consumers perception of price index is unlikely to adjust in a week. In other words, consumers do not feel becoming relative richer because of the sales event.

of $q_j(\mu)$ that consumers *actually* buys from either channel for variety ω in city j is:

$$G_j(q) \equiv Pr\{max\{q_o(\mu,\omega), q_f(\mu,\omega)\} \leq q\} \tag{2.7}$$
$$= exp\{-(\frac{q}{p_j(\omega)^{-\sigma} Y_j {P_j}^{\sigma-1}((k_o^{\frac{\sigma}{\sigma-1}} s_o(\omega))^\theta + (k_f^{\frac{\sigma}{\sigma-1}} s_f(\omega))^\theta)^{\frac{\sigma-1}{\theta}}})^{-\frac{\theta}{\sigma-1}}\}$$

The calculation of the aggregated demand for variety ω in city j requires integrating individual demand function on the probability distribution of $z_m(\mu,\omega)$, which is equivalent to calculating the expectation of $q_j(\mu,\omega)$. Using the mean function of the Fréchet distribution, the overall demand for variety ω in city j is:

$$Q_j(\omega) = \int_{\mu\in U_j} q_j(\mu,\omega) d\mu \tag{2.8}$$
$$= L_j p_j(\omega)^{-\sigma} Y_j {P_j}^{\sigma-1} \Gamma(1 - \frac{\sigma-1}{\theta})((k_o^{\frac{\sigma}{\sigma-1}} s_o(\omega))^\theta + (k_f^{\frac{\sigma}{\sigma-1}} s_f(\omega))^\theta)^{\frac{\sigma-1}{\theta}}$$

where Γ is the Gamma function. Note that the shape parameter that governs the dispersion of matching quality θ is the same for both channel-specific distribution of $z_m(\mu,\omega)$, and the distribution of $z(\mu,\omega)$, which is the matching quality in the chosen channel. This again relies on the properties of the Fréchet distribution. Mathematically, the expectation exists only when $\theta > \sigma - 1$[18]. This means that the dispersion of $z(\mu,\omega)$ should be large enough so that the dispersion of matching quality is larger than the elasticity of substitution between varieties[19].

2.2.1.3 Firm Decisions

Without considering any sales event, firms price two channels uniformly to avoid arbitrage[20]. A firm is characterized by its productivity parameter of ϕ_i. Firms are assumed to have dual channels in the sense that all firms in city i can sell through both online channel and offline channel to consumers in city j[21]. Therefore, firms optimize price based on the aggregate demand in each city, assuming that firms have such information.

$$W_{ij}(\omega) = (p_{ij}(\omega) - \frac{w_i \tau_{ij}}{\phi_i}) Q_{ij}(\omega) \tag{2.9}$$

[18]The expectation of a random variable with Fréchet distribution exists only if the scale parameter of the distribution is larger than 1.

[19]Given that the dispersion of matching quality measures the variability of alternative choices (in my context, shopping channels), I interpret this condition as that the variability of alternative choices for a variety is larger than the substitutability between the chosen varieties.

[20]Cavallo (2017) shows that online and offline prices are identical about 72% of the time of study period from December 2014 to March 2016.

[21]I only focus on dual channel firms because the event only allows small shops (with both online and offline distribution channels) to participate. In a similar vein to Fan et al. (2018) and Helpman et al. (2004), it is possible to derive the fraction of online-only firm and dual-channel firms based on the trade-off between additional revenue from satisfying a greater variety of consumers in taste for channels and the additional cost in setting up physical stores.

where $W_{ij}(\omega)$ is the profit, w_i is the wage in city i, ϕ_i is the productivity in city i, and $Q_{ij}(\omega)$ is the expected demand for ω in city j. τ_{ij} is the iceberg transport cost to ship the goods from city i to j. For simplicity, intercity transportation costs are assumed equal across the two channels. This is plausible as moving goods across cities in both channels use the same technology, which is mostly rail freight or highway freight transport. Another reason is that I do not have data on the freight cost. Given the assumption that firms price based on the expected aggregate demand in a city, so that the uncertainty of individual consumption induced by $z(\mu, \omega)$ does not affect the well-known, constant mark-up under CES utility function and monopolistic competition[22]. The optimal price is,

$$p_{ij}(\omega) = \frac{\sigma}{\sigma - 1} \frac{w_i \tau_{ij}}{\phi_i} \tag{2.10}$$

Channel-specific price shock k_m is not included here because the long-term price does not include discounts. Discount in the sales event is treated as an external shock in the model[23]. The model does not include the dynamics that consumers may anticipate the event and shift their budget for three reasons. First, given that this is a yearly event, and the exact rules and the products on sale change every year, there are many unanticipated elements. Second, the goal of the model is to analyze online-offline substitution corresponding to price change. Including those dynamics may overly complicate the model without adding much insight. Third, I address the anticipation effects separately in the empirical section.

2.2.1.4 Equilibrium Quantity and Share of Consumers by Channels

The expected equilibrium consumption of good from city i in city j is,

$$\begin{aligned} Q_{ij}(\omega) &= L_j E(q(\mu, \omega)) \\ &= L_j \left(\frac{\sigma}{\sigma - 1} \frac{w_i \tau_{ij}}{\phi_i}\right)^{-\sigma} Y_j {P_j}^{\sigma-1} \Gamma\left(1 - \frac{\sigma - 1}{\theta}\right) \left(k_f^{\frac{\sigma}{\sigma-1}} s_f(\omega)\right)^{\sigma-1} \left(1 + \left(k^{\frac{\sigma}{\sigma-1}} s(\omega)\right)^{\theta}\right)^{\frac{\sigma-1}{\theta}} \equiv C_{ij} A \end{aligned} \tag{2.11}$$

where $k \equiv \frac{k_o}{k_f}$, $s(\omega) \equiv \frac{s_o(\omega)}{s_f(\omega)}$. Here, I normalize the price shocks and average preference for both channels by the values in the offline channel to reduce the number of unknown variables. For simplicity, let $A \equiv \left(1 + \left(k^{\frac{\sigma}{\sigma-1}} s(\omega)\right)^{\theta}\right)^{\frac{\sigma-1}{\theta}}$, and C_{ij} denotes a collection of items not related to k and $s(\omega)$.

Given the probability of $z_m(\mu, \omega)$, we can calculate the probability that a consumer buys through channel m. Given the law of large numbers, the probability gives the fraction of

[22] Gabaix et al. (2016) presents an interesting discussion on how noise in prices may affect mark-up.

[23] Discount can be modeled endogenously as a part of the pricing strategy as in Seim & Sinkinson (2016). However, since the goal of my model is to provide a framework to evaluate the ratio of the increase of sales in the online channel to the decrease of sales in the offline channel, endogenous discounts are beyond the scope of this study.

consumers that choose channel m,

$$\pi_{oj}(\omega) = Pr\{k_o^{\frac{\sigma}{\sigma-1}} z_o(\mu,\omega) \geq k_f^{\frac{\sigma}{\sigma-1}} z_f(\mu,\omega)\} = \frac{(k^{\frac{\sigma}{\sigma-1}} s(\omega))^\theta}{1 + (k^{\frac{\sigma}{\sigma-1}} s(\omega))^\theta} \tag{2.12}$$

$$\pi_{fj}(\omega) = 1 - \pi_{oj}(\omega) = \frac{1}{1 + (k^{\frac{\sigma}{\sigma-1}} s(\omega))^\theta} \tag{2.13}$$

The share of online consumers is decided by the relative channel preference $s(\omega)$ and the shape parameter θ[24]. Similar to Redding & Weinstein (2016), the expenditure share for each variety ω and each consumer μ is:

$$f_j(\mu,\omega) = (\frac{p(\mu,\omega)/z(\mu,\omega)}{P_j})^{1-\sigma} \tag{2.14}$$

The online expenditure share of consumer μ across varieties can be expressed as $\int_{\Omega_j} f_j(\mu,\omega)\pi_{oj}(\omega)d\omega$, and the online expenditure share in a city is,

$$\pi_{oj} = \int_{\Omega_j}\int_{U_j} f_j(\mu,\omega)\pi_{oj}(\omega)d\mu d\omega \tag{2.15}$$

Note that because the expenditure share for ω for consumer μ depends on the price index P_j of the city where consumer μ resides, and P_j depends on shipping cost τ_{ij}, wage w_j and productivity ϕ_j, the shares of online shopping π_{oj} are not the same across cities. In the absence of the data on city characteristics, and given the intention to focus on the analysis of switching shopping channels, I assume that consumers have the same relative preference for all products, $s(\omega) = s$ for any ω. I also assume that all firms can serve all cities in both channels, that is, $\Omega_j = \Omega_i$ for any $i, j \in I$. As a result of these assumptions on symmetry, the model predicts that the shares of online shopping across cities are the same[25]. I discuss the data limitation in allowing for a heterogenous s in Section 2.2.3.

2.2.1.5 Price Shock and Channel Substitution

Due to the symmetry assumptions for s across varieties, I focus on a specific variety ω produced in city i and sold in j. I suppress index ω for simplicity. One of the key findings in Eaton & Kortum (2002) (EK model) is that the price distribution of the varieties that any given origin actually sends to any given destination is independent of its origin regions. In my model, firms do not compete in price given the monopolistic competition assumption, and the listing price is not a random variable as the listing price is based on the expectation of

[24]These fractions are similar to those in Dolfen et al. (2019). In their model, the relative preference for online shopping s is conceptualized as two factors: relative quality of online merchants and ease of access.

[25]However, the model allows firms in different cities to be different in productivity, wage, shipping cost, which implies that consumers in different cities allocate their income differently. Classic predictions from many trade model hold here. For example, remote cities have higher price index due to higher shipping cost assuming all else being equal.

aggregate demand. Instead, the quantity demanded is a random variable as matching quality is a random variable. Similar to the EK model, the distribution of quantity demanded across consumers in city j through a channel $m \in \{o, f\}$ is independent of channels[26],

$$Pr\{q_o(\mu) \leq \tilde{q} | q_o(\mu) \geq q_f(\mu)\} = Pr\{q_f(\mu) \leq \tilde{q} | q_f(\mu) \geq q_o(\mu)\} = G_j(\tilde{q}) \tag{2.16}$$

Intuitively, what is happening is that the channel with better consumer appeal (higher s_m) can serve a greater number of consumers exactly up to the point where the distribution of quantities for what it sells through channel m is the same as m's overall quantity distribution. For example, if offline grocery shopping in Waitrose[27] is more convenient than online grocery shopping on Amazon, then Waitrose has a larger consumer base than Amazon, to the point at which the quantity served by Waitrose will have the same distribution as the quantity that consumers shopped on Amazon[28]. Because the distributions of quantities are the same in the two channels, the expected quantity will be the same in the two channels.

$$E(q_o | q_o(\mu) \geq q_f(\mu)) = E(q_f | q_o(\mu) \leq q_f(\mu)) = E(q) \tag{2.17}$$

Using to the example above, this property implies that although there are more offline Waitrose consumers relative to online Amazon consumers (assuming going to Waitrose stores is more appealing than waiting for deliveries from Amazon), an online consumer does the same amount of grocery shopping on Amazon as an offline consumer does in Waitrose. These predictions derived from the assumption that matching quality follows the Fréchet distribution appear remarkably plausible. Given the total quantity consumed in channel m equals the number of consumers that choose m multiplied by the average quantity consumed in channel m, equation 2.17 implies that the quantity consumed through channel m is,

$$Q_{mij} = \pi_{mj} L_j E(q_m | q_m(\mu) \geq q_{n \neq m, n \in \{o,f\}}(\mu)) = \pi_{mj} L_j E(q) = \pi_{mj} Q_{ij} \tag{2.18}$$

Equation 2.18 shows that the expected quantity consumed through channel m equals the share of consumers that choose channel m multiplied by the expected overall demand from both channels[29]. Importantly, these properties hold when prices change. The change in channel-specific demand can be approximated by a derivative formula. During the online shopping

[26] Another way to think about this is, the perceived price, or the price adjusted by the matching quality $\frac{p}{z_m}$, is a random variable. Thinking z as the productivities draw in the EK model, then the perceived price is the same as the purchase price in the EK model. Quantity is a function of the perceived price, so it is independent of channels (origin region in the EK model).

[27] Waitrose is a chain of British supermarkets, similar to Whole Food in the US.

[28] If we think the share of offline consumers for a product as the frequency of offline shopping for a particular consumer, then this property implies that although the consumer does offline shopping more frequently assuming offline shopping is more convenient (higher s_f), the distribution of the number of goods the consumer purchased does not depends on the channel.

[29] Note that this property relies heavily on the Fréchet distribution assumption.

event, price shock k increases from 1 to a higher value. Taking derivatives of Q_{mij} on k gives

$$\frac{dQ_{mij}}{dk} = \underbrace{\pi_{mj}\frac{dC_{ij}A}{dk}}_{\text{mean effect}} + \underbrace{C_{ij}A\frac{d\pi_m}{dk}}_{\text{share effect}} \tag{2.19}$$

I denote the first term as the mean effect and the second term as the share effect. Three interesting results emerge. First, the ratio of the mean effect to the share effect of the offline channel is constant $-\frac{\sigma-1}{\theta}$. Second, the ratio of the share effect in the online channel to the share effect in the offline channel is -1. Third, the ratio of the mean effect to the share effect of the online channel is,

$$\gamma = \frac{\sigma-1}{\theta}(k^{\frac{\sigma}{\sigma-1}}s)^{\theta}$$

which is an increasing function of s and k. Specifically, $\gamma = \frac{\sigma-1}{\theta}$ when $s = 1$ and $k = 1$. In this case, the online and offline channels are symmetric in terms of preference and price differences. Above linkages between these terms allow me to express these effects by one of these effects. I choose to normalize these effects based on the share effect of the online channel. Define ν as the share effect of the online channel. The other parts can be shown as below,

$$\frac{dQ_{oij}}{dk} = (\underbrace{\pi_o\frac{dA}{dk}}_{\gamma\nu} + \underbrace{A\frac{d\pi_o}{dk}}_{\nu})C_{ij} \tag{2.20}$$

$$\frac{dQ_{fij}}{dk} = (\underbrace{\pi_f\frac{dA}{dk}}_{\frac{\sigma-1}{\theta}\nu} + \underbrace{A\frac{d\pi_f}{dk}}_{-\nu})C_{ij} \tag{2.21}$$

Consumption in city j can then be obtained by aggregating through origin cities $i \in I$. Therefore, the online-offline substitution ratio λ_j, defined by the ratio of the change in online shopping to the change in offline shopping in city j, is,

$$\lambda_j = \frac{\sum_i \Delta Q_{oij}}{\sum_i \Delta Q_{fij}} = -\frac{\theta + (\sigma-1)(k^{\frac{\sigma}{\sigma-1}}s)^{\theta}}{\theta-\sigma+1} \equiv \lambda \tag{2.22}$$

The online-offline substitution ratio is the ultimate goal of the demand model. It has following properties:

(i) This ratio is the same across cities due to the symmetry assumption of s.

(ii) Given the assumption $\theta > \sigma - 1$, it follows that $\lambda < 0$, which guarantees that offline consumption decreases when online shopping increases.

(iii) As along as $\sigma > 1$, $|\lambda|$ is always greater than 1. If varieties are substitutable, then $\sigma > 1$

is satisfied. This indicates that the switch from offline shopping to online shopping is not one-to-one. In the trade literature, σ is frequently estimated to be greater than one.

(iv) The inverse of the absolute value of the ratio $\frac{1}{|\lambda|}$ determines the amount of offline shopping that consumers are willing to substitute with a one unit increase in online shopping, and therefore dictates the amount of traffic related to offline shopping that can be saved.

Especially, it turns out that λ can be conveniently estimated using the elasticity of online consumption quantity to the relative price of online to offline channel,

$$\frac{1}{|\lambda|} = 1 - \frac{\sigma}{\rho} \tag{2.23}$$

where ρ is the elasticity of online consumption quantity to the relative price of online to offline channel. To be brief, this follows from equation 2.12 and equation 2.20, by expressing s^θ using the information on the share of online shopping π_{oj},

$$\rho = \frac{(\sigma - 1)s^\theta + \theta}{s^\theta + 1} \frac{\sigma}{\sigma - 1} \tag{2.24}$$

where $\rho \approx \frac{\widetilde{Q}_{oj}}{\Delta k}$[30]. $\widetilde{Q}_{oj} = \frac{\Delta Q_{oj}}{Q_{oj}}$ denotes the growth rate of online shopping quantity. Δk is the change in the relative price shock. This holds when Δk is small[31]. ρ can be calculated using data on $\widetilde{Q}_{oj}$ and Δk. See Appendix B.2.3.1 for a derivation.

Equation 2.23 shows that $\frac{1}{|\lambda|}$ increases with the elasticity of online consumption quantity to the relative price of online to offline channel ρ, while decreases with the elasticity of substitution between varieties σ. Intuitively, if consumers are more sensitive to the price difference between channels (a higher ρ), consumers reduce more offline shopping with one unit increase in online shopping; if products are highly substitutable (a higher σ), then consumers reduce less offline shopping with one unit increase in online shopping. Note that σ measures the substitutability across products while λ measures the substitutability across channels. As certain approximations are involved in the derivation, Appendix B.2.3.2 provides a simulation to validate the formula for λ. The simulated λ is slightly smaller than its theoretical value in equation 2.22 due to an omitted higher order component in equation 2.19, but are almost identical to the value obtained using equation 2.23.

[30] ρ_j could vary across cities as s_j is potentially heterogenous and vary across cities. As s_j is assumed to be the same across products and thus across cities, ρ_j reduces to a scalar ρ.

[31] When Δk is small, $\frac{\widetilde{Q}_{oj}}{\Delta k} = \frac{\frac{\Delta Q_{oj}}{Q_{oj}}}{\frac{\Delta k}{k}}$ approximates the elasticity of online consumption quantity to the relative price of online to offline channel. $k = 1$ due to the assumption of equal price across channels. Note that k increases when the online price decreases in the set-up of my model, so ρ is positive and is the absolute value of the elasticity of online consumption quantity to the relative price of online to offline channel. Denote $p = \frac{1}{k}$ as the relative price, $\frac{\Delta k}{k} = p\Delta\frac{1}{p} \approx -\frac{\Delta p}{p}$.

2.2.2 Traffic Congestion and Online Consumption

Assuming the two ways of shopping have different levels of traffic efficiency, the overall traffic due to shopping will then change during the event. The following section combines insights from transportation engineering literature and accounting assumptions on vehicle demand and traffic density to derive the elasticity of traffic congestion index to the quantity of online shopping. The elasticity includes three components: traffic density, the importance of e-commerce, and a traffic-saving factor, which is per-unit online good traffic saving. Note that the model ignores the effect of traffic congestion on vehicle demand given the unique context of the event, for which I will provide more details.

2.2.2.1 Online Consumption and Vehicle Demand

Given a time interval in city j, shopping vehicle demand D_j can be calculated as the sum of online shopping vehicle demand and offline shopping vehicle demand,

$$\begin{aligned} D_j &= t_o \sum_i Q_{oij} + \zeta_j t_f \sum_i Q_{fij} \\ &= t_f \delta Q_{oj} + \zeta_j t_f \frac{\pi_f}{\pi_o} Q_{oj} \end{aligned} \tag{2.25}$$

where t_o is the per-unit good vehicle demand for online shopping and t_f is the per-unit good vehicle demand for offline shopping. $\delta = \frac{t_o}{t_f}$, which I call vehicle-saving ratio. It is smaller than one if online shopping is more vehicle-efficient relative to offline shopping. ζ_j is the average share of shopping made through private vehicles[32]. Thus, $\zeta_j t_f \sum_i Q_{fij}$ is the vehicle demand for offline shopping. Now, denoting the share of shopping vehicle demand to the overall vehicle demand (including vehicles for other purposes such as commute) on the road as ψ_j, then the overall vehicle demand in a city is $\frac{D_j}{\psi_j}$. Denoting the capacity of roads (for example, the total length of roads) in the city as R_j, then the traffic density on the roads (vehicle/km),

$$n_j = \frac{D_j/\psi_j}{R_j} \tag{2.26}$$

Given the change of online consumption ΔQ_{oj}, the change in shopping vehicle demand is the sum of the change of vehicle demand in online shopping and that in offline shopping.

$$\begin{aligned} \Delta D_j &= t_o \sum_i \Delta Q_{oij} + \zeta_j t_f \sum_i \Delta Q_{fij} \\ &= t_f (\delta - \frac{\zeta_j}{|\lambda|}) \Delta Q_{oj} \end{aligned} \tag{2.27}$$

[32]Bus travel can be discounted into car travels, which is ignored here for simplicity. Other forms of transporting shopping goods include walking or taking public subways.

Similarly, the change of traffic density is

$$\Delta n_j = \frac{\Delta D_j}{R_j} \tag{2.28}$$

Combining equation 2.25, 2.26, 2.27, 2.28 gives,

$$\frac{\Delta n_j}{n_j} = \frac{\psi_j(\delta - \frac{\zeta_j}{|\lambda|})}{(\delta + \zeta_j \frac{\pi_f}{\pi_o})} \frac{\Delta Q_{oj}}{Q_{oj}} \tag{2.29}$$

Equation 2.29 presents the relationship between the changes in traffic density with the changes in online shopping.

Before moving on to introducing the relationship between traffic congestion index and online shopping, it is worth noting that the change of shopping vehicle demand for intercity logistic is,

$$\Delta D_{ij} = t_h \Delta Q_{oij} \qquad \text{if } i \neq j \tag{2.30}$$

where t_h is the per unit of good vehicle demand for online shopping on the intercity roads. This equation simply states that the increase in online shopping increases vehicle demand for intercity travel owning to offline retail logistics. Of course, the reduction in offline shopping may reduce intercity traffic; however, this negative adjustment may materialize much slower than the sharp increase in the demand for online shopping. For this reason, I assume that possibility away.

2.2.2.2 Vehicle Demand, Traffic Density, and Traffic Congestion Index

Now the task is to provide a mapping from vehicle demand to traffic congestion index using traffic density. Note that the model assumes that vehicle demand increases traffic density, and thus increases traffic congestion, while ignores the effect of traffic congestion on vehicle demand. The reason is that empirical evidence shows that the change in traffic congestion is very small. Traffic congestion index reduces by 4%, which is about 2 minutes time reduction for a one hour travel. Such small change is arguably undetectable by commuters, at least in the one week post-event time window that this research focuses on. Therefore, I model a one-way relationship between vehicle demand and traffic congestion.

As summarized by Yang et al. (2018b), traffic speed and density follow a monotonic relationship. Density further reflects vehicle demand monotonically because a decision to use road transport is essentially a decision to add a vehicle on the road (Else, 1981; Walters, 1961). I follow the functional form of speed and density as in Adler et al. (2017)[33] and derive the

[33]This functional form was proposed by Underwood (1961). See Brilon & Lohoff (2011) for how well this functional form fits with real-world data and other possible functional forms.

relationship between the traffic congestion index and density,

$$lnT_j = \frac{n_j}{n_{mj}} - ln(u) \tag{2.31}$$

where T_j is the traffic congestion index for a given road segment. n_{mj} is the density of vehicles on the roads when maximum flow is achieved, and u is a constant[34]. Using equation 2.29, the marginal change in traffic congestion index can be expressed as

$$\Delta lnT_j = \frac{\Delta n_j}{n_{mj}} = \frac{n_j}{n_{mj}} \frac{\psi_j \pi_o}{\delta \pi_o + \zeta_j \pi_f} (\delta - \frac{\zeta_j}{|\lambda|}) \frac{\Delta Q_{oj}}{Q_{oj}} \tag{2.32}$$

When the change of ΔQ_{oj} is very small, $\frac{\Delta Q_{oj}}{Q_{oj}} \approx \Delta lnQ_{oj}$[35]. Hence, I derived the elasticity of traffic congestion to online consumption,

$$\epsilon_j \approx \frac{n_j}{n_{mj}} \frac{\psi_j \pi_o}{\delta \pi_o + \zeta_j \pi_f} (\delta - \frac{\zeta_j}{|\lambda|}) \tag{2.33}$$

ϵ_j is a variable that varies across cities. ϵ_j can be decomposed into three parts: the first part $\frac{n_j}{n_{mj}}$ is the ratio of actual density to the optimal density, which measures the congestion level. The intuition for this term to appear in the equation is that the impact of the reduction of vehicles is stronger when roads are more congested. The second part $\frac{\psi_j \pi_o}{\delta \pi_o + \zeta_j \pi_f}$ captures the importance of online shopping relative to offline shopping in the city. Intuitively, a higher share of online shopping in a city implies a larger scope for e-commerce to impact the city's traffic congestion. The third part $\delta - \frac{\zeta_j}{|\lambda|}$ determines the traffic saved by per-unit online good, which I term "traffic-saving factor". Importantly, as the first two parts of ϵ_j are always positive, the condition for online shopping to reduce traffic congestion is simply,

$$\delta < \frac{\zeta_j}{|\lambda|} \tag{2.34}$$

The intuition of the condition is that, for the increase in online shopping to result in a reduction in vehicles, the vehicle-saving ratio is sufficiently low, or the amount of offline shopping that consumers are willing to substitute with online shopping is sufficiently large.

In the following sections, I first conduct a quantitative analysis of the elasticity ϵ_j, then focus

[34]Omitting city index j, Underwood density-speed equation is

$$n = n_m ln(\frac{v_r}{v})$$

where v is speed, and v_r is the "reference" speed, which is estimated to be 300km/h for typical motorways conditions in the transport engineering literature. Assuming that the reference speed is u times to the free-flow speed v_f, traffic congestion index $T = \frac{time}{time_f} = \frac{1/v}{1/v_f} = \frac{1/v}{u/v_r} = \frac{1}{u} e^{n/n_m}$ where $time$ is the actual passing time of a road segment and $time_f$ is the free-flow passing time. See Notley et al. (2009) for details.

[35]It is easy to show that $\Delta lnQ = ln(1 + \frac{\Delta Q}{Q}) \approx \frac{\Delta Q}{Q}$ when ΔQ is very small.

on estimating the mean of the elasticity $\epsilon = \bar{\epsilon_j}$ empirically. Note, I use equation 2.32 as the estimation equation for ϵ instead of equation 2.33 because ΔQ is large[36].

2.2.3 Quantitative Analysis of the Model

In order to quantify the elasticity of traffic congestion to online shopping using the model, I begin by assuming values of parameters or sample means of some variables for the three components in equation 2.33. Table 2.1 lists the parameter values and sample statistics used to estimate the elasticity in equation 2.33. I first estimate λ and the vehicle-saving ratio for per unit of good, with reasonable assumptions of σ and using the sample moment of ρ. Note that there are some complications in transferring the spike in the Baidu Index to the growth rate of online shopping quantity between the two weeks surrounding the event[37]. A very crude procedure is to use the average daily online consumption (estimated using yearly online consumption) and consumption on the event day (estimated using reported overall sales on the event day) to calculate the weekly online consumption growth rate. The extra online consumption due to the event in the first week can be evenly added to the normal online consumption stream in the second week. If the numbers from the two sources do not align, I can adjust the growth rate of the Baidu Index, accordingly, to better approximate the weekly growth rate of online shopping. The weekly growth of consumption is estimated at about 160%, which turns out to coincide with the sample mean of the growth rate of the Baidu Index. For this reason, I just use the growth rate observed from the Baidu Index without any adjustments. Admittedly, estimation errors may arise due to a lack of information. Because the online prices are reported to be about 80% of the online price on the event day as mentioned earlier, the change in price shock is set to 0.25 ($= \frac{1}{0.8} - 1$). The estimate of ρ is then 6.4 ($= 1.6/0.25$). The elasticity of substitution between varieties σ is assumed equal to four, which is in line with figures frequently used in the international trade literature (see Redding & Sturm (2008)). Using equation 2.23, the online-offline substitution ratio λ is estimated to be -1.9, which suggests that offline shopping reduces a half unit when online shopping increases one unit. Because a fraction ζ of offline consumption is made by using private vehicles or taking taxis, the decrease in offline shopping vehicle demand is $frac\zeta 1.9$ units assuming that the offline consumption reduces proportionally for consumers who use private vehicles and for consumers who do not use private vehicles. ζ is estimated to be 0.91, derived by calculating the share of shopping trips in private vehicles or taxies to the total shopping trips during peak hours in urban areas in the United States using the NHTS 2017

[36]The estimation of the elasticity ϵ based on equation 2.32 is therefore an approximated value.

[37]For simplicity, I assume that the impact of the shopping event only lasts for a week.

data[38]. It is reported[39] that a typical Amazon driver can delivery about 150-200 parcels a day (about 10 hours, according to the report.). Given that it takes consumers about one hour for a round trip for retail shopping[40], and assuming the amount of good that consumers buy is equivalent to a parcel in the trip, then online delivery is about 15 times more efficient than offline retail[41]. Thus, δ is assumed to be 0.067. For a one unit increase in online shopping, about 0.48 (= 0.91/1.9) units of vehicles used for offline shopping are saved, while additional online shopping vehicles is only 0.067 unit of vehicles. Taken together, the vehicle-saving ratio per unit online shopping quantity is -0.413(=0.067-0.48)[42].

Next, I set the values of parameters or estimate sample statistics in $\frac{\psi\pi_o}{\delta\pi_o+\zeta\pi_f}$. π_o is estimated equal to be 12.6% according to national statistics on the share of online sales to overall retail sales. I do not have city-level data on the share of online consumption, which is the reason for assuming homogeneity in $s(\omega)$. The share of online shopping grows dramatically since online shopping increases by 160% on average in the week after the online shopping event. I postulate the value of the share of online shopping in the week after the event using its initial value and the growth rate of online shopping, and then take the mean of both values for calculating the elasticity. ψ is estimated based on the US NHTS 2017 data as well. I calculate the share of shopping trips using vehicles to total trips using vehicles during peak hours in urban areas. Finally, I estimate the value for $\frac{n}{n_m}$. This part can be expressed as $\frac{n}{n_m} = ln\frac{v_f u}{v} = lnu + lnT$. lnT can be estimated using the sample mean of the traffic congestion index, which is 1.65 in peak hours in the two weeks surrounding the event. u is estimated using the ratio of the reference speed, which I set to 300 km/h, and free-flow speed, which I set to 60 km/h. Collectively, the elasticity of traffic congestion index to online shopping quantity is estimated to be -0.06. However, the estimates based on the quantitative model vary substantially when the share of online shopping and the growth rate of online shopping change. If increasing the values of the share of online shopping to 0.2, which is the statistic for Beijing in 2016, the elasticity is estimated to be -0.11. If I also increase the values of the growth rate to the maximum value observed in the sample (2.7), then the elasticity is -0.25. Therefore, while the model predicts a negative elasticity of traffic congestion to online consumption quantity, it does not identify exact magnitudes. For this reason, I turn to empirical estimates.

[38]The reason to use the US data is that I do not have access to Chinese household traffic survey data. The figure observed from the US data is likely to be larger than that in Chinese cities as the number of road vehicles per capita is much higher in the US. See https://en.wikipedia.org/wiki/List_of_countries_by_vehicles_per_capita

[39]See https://www.bbc.co.uk/news/uk-england-37912858.

[40]According to the US NHTS 2017 data, an average shopping trip using cars take 27 minutes one-way.

[41]Amazon driver delivers about 15(=150/10) parcels in an hour, while a consumer buys a parcel in an hour.

[42]Note I have only considered vehicle savings from the change of logistics from consumers to stores. There may also be vehicle savings from the change of logistics from warehouses to stores in the long term, which is, however, unlikely to be an issue for this study as I focus on a short term sales event.

2.3 Data and Descriptives

2.3.1 Data on Traffic Congestion and Pollution

This study collected traffic congestion data from a GPS navigation company in China. The traffic congestion index is the ratio of the actual passing time to the free-flow passing time for a given road segment recorded from the company's millions of GPS navigation service users. The data contains 94 major Chinese cities. An average city in the sample has a population of 3 million[43] and an average annual GDP per capita of 86 thousands *yuan*.

Figure B.1 shows the time series of traffic congestion index in the period that is close to the shopping event in 2016. The green line shows daily average traffic congestion; the red line shows traffic congestion during peak hours from 7 am to 9 am and from 5 pm to 7 pm; the blue line shows traffic congestion during off-peak hours. The unit in the horizontal axis is the number of days away from the online shopping event on 11 November 2016. The big drop of these lines around day 77 shows the Chinese New Year holiday, when people enjoy the holiday and commute much less[44]. The days between the vertical lines in the left of the graph are the weeks surrounding the online shopping event of 11 November and an offline shopping event on 12 December in 2016. Traffic congestion levels start high on Monday, drop in the middle of the week, and bounce back to another peak on Friday, before plummeting over the weekend. Given the high volatility in the time series of congestion data, I restrict the sample to a narrow band around the event: weekdays of one week before the event and two weeks after the event. This restriction reduces the unobserved time trends of traffic in the study period and highlights the impact of the online shopping event on traffic congestion[45].

Inspired by Akbar et al. (2018), I utilized Baidu Map API to query hourly travel time based on real-time road traffic condition across the cities in my sample in one week before and two weeks after the Singles' Day shopping Event in November 2018. There are 9,216 records of travel time information recorded from 7 am to 11 pm for each day. Routes between cities remain the same. The lengths of routes are known and remain constant. This dataset shows how the Singles' Day shopping event shifts intercity traffic pattern. The data were collected from 5 November to 23 November. I was unable to collect the data for some hours of the day. The data are mostly complete on the Thursday and Friday in the week before the event and in

[43]The population data is based on household (Hukou) registration, which does not include migration population. The number is usually smaller than the actual population in cities.

[44]Many migrant workers return to their hometowns, often in rural areas. Therefore, the population temporarily drops in cities.

[45]Ideally, I would like to have the traffic congestion index in two weeks before the online shopping event to study the pre-event trend carefully; however, I started to code the program to track the hourly update of traffic congestion index on the website of navigation company on 6 November 2016, and the website does not provide data retrospectively. The data collection was also interrupted during the first weekend after the event due to technical reasons, so I cannot study the effects over the weekends.

the two weeks after the event. In the related graphs and regressions, I added hour × weekday fixed effect to address the missing variable issue. In the regressions, I restrict sample size to Thursday and Friday only.

Air pollution data are published online hourly by China's Ministry of Environmental Protection. I collected the data for about 1,563 monitoring stations across 337 prefectures. The measures include air quality index (AQI), carbon monoxide (CO), nitrogen dioxide (NO2),ozone (O3), particulate matter with a diameter of 2.5 μm or less (PM2.5), particulate matter with a diameter of 10 μm or less (PM10), and sulfur dioxide (SO2). NO2 concentration is the primary outcome variable of interest as it is a major vehicle exhaust. Figure B.4 shows the daily cycle of the shock of NO2 and traffic congestion index[46]. The timing of the generation of NO2 in each hour follows the change of traffic congestion index closely in the daytime. I further plot daily NO2 level and traffic congestion index in a longer time horizon as shown in Figure B.5. NO2 follows traffic congestion index closely across days, especially when traffic congestion index plummeted during important holidays such as the Chinese New Year Festival. NO2 concentration is a good predictor of traffic congestion. Other pollutants do not correlate with traffic congestion as well as NO2.

Table 2.2 reports summary statistics for our dataset in the two weeks window surrounding the Singles' Day Shopping Event. Each observation is city-by-hour. The first two columns report the outcomes of interest one week before and one week after the event for the peak hours, and the last two columns report those for off-peak hours. The average traffic congestion index for peak hours is 1.74 before the event, and dropped to 1.67 after the event. The hourly concentration of pollution appears to be much more volatile than traffic congestion. Pollution levels increase substantially in the week after the shopping event; however, its increase does not correlate with the change of online shopping as shown later in Section 2.6.5.

2.3.2 Data on Online Shopping

To measure the change in online shopping, I use the frequency of searching the name of the online shopping platform in each city based on Baidu Index. Baidu is the largest search engine service provider in China and dominates the market since Google search engine exited the Chinese market in the year 2010. Baidu Index is a publicly available web service and bears a number of similarities to Google Trends. The query index is based on the frequency of the search keywords within a day as the minimal unit. Importantly, it provides the IP addresses of its users to city level. This feature allows tracking the trends of search frequency of the online shopping platform for each city. Web search engine index has been previously adopted to track economic activities in real time (Choi & Varian, 2012). Vosen & Schmidt (2011) show that forecasting of monthly private consumption based on Google Trends outperforms

[46]I estimated hourly shock of NO2 following procedures in Henderson (1996).

survey-based indicators.

Figure 2.1 depicts an example output from Baidu Index for the query of the two possible names of the online shopping platform: Taobao and Tmall. Taobao is a consumer-to-consumer (C2C) online retail platform for small businesses and individual entrepreneurs to open online stores. In contrast, Tmall runs a business-to-consumer (B2C) online platform for local Chinese and international businesses to sell brand name goods. Tmall shops are required to have established physical stores, and often have national offline distribution channels. In short, Taobao operates like eBay while Tmall operates like Amazon. Unlike either eBay or Amazon, Alibaba provides both services and the product searches in either platform give results from both Taobao and Tmall[47]. Assuming that a constant share of the population entering the online shopping website through the search engine in cities, the index can serve as a proxy of the number of online consumers and captures the increase of online shopping during the shopping event day. As shown in Figure 2.1, daily searches on the online shopping platform escalated from around 900 thousand times before the event to around 2.7 million times at the peak on 11 November 2016. I extracted the index for the day of 11 November 2016 which is the event day, and 11 October 2016 which is one month before the event. I only used the index on the event day instead of a cumulative sum of the index around the event, given the fact that the sale only lasts for one day and consumers have to complete the order on the day of the event[48]. The daily value in a month before the event shows the event-free frequency of visiting the online platform. I choose 11 October 2016 to represent the event-free average search as it is far enough from the event in time but not too far, and using the same day as the event day in the last month could avoid possible bias in the seasonal trend of searches within a month[49].

The online shopping platform also publishes an index that measures e-commerce development in cities: Alibaba E-commerce Development Index (AEDI). AEDI is the weighted average of an online shopping index and an online selling index. The online shopping index is a weighted average of the number of online buyers and average online consumption; The online selling index is a weighted average of the number of online sellers and average online sale. The two indices thus measure the intensity of online shopping and online selling in each city, respectively. Equations in Appendix B.2.2 show how the two indices are constructed. The Baidu index and AEDI cover about 277 cities while the traffic congestion index only covers

[47]Consumers can easily search goods from the two sources in the same search entry box in either of the two domains of the platform. Therefore, I use the sum of the search of both platforms in Baidu Index instead of only using Tmall, despite the fact that the online shopping event is only for Tmall stores.

[48]The index is above average in the few days around the event day. Consumers may browse products before the event and check the status of delivery after the event. Using the cumulative sum of these searches may double count the actual number of transactions.

[49]It is possible to obtain an average of Baidu Index over a period; however, there is a daily cap for querying these indices, which makes obtaining more data time consuming and difficult. As shown in 2.1, the index was very flat before the event. Obtaining more days is unlikely to change the pre-event average of the index.

94 major cities. The overlap cities form the main sample of the study. I occasionally use the expanded sample with 277 cities when I do not need the traffic congestion index.

To validate the e-commerce indices, I collect actual online consumption or online sales data from news and online reports. In Figure B.2(a), I record the first-hour online consumption data by provinces in the Singles' Day event in 2016. I took the average of the city-level online shopping index to create an index at the province level. The graph shows that log online consumption is positively correlated with online shopping index[50]. In Figure B.2(b), I collected monthly online sales data of Alibaba in May 2017, by cities. Plotting it against the online selling index shows that the log online selling index is a good predictor of log online sales in cities.

I extracted city-to-city postage fee per km from the website of a leading national logistics company and derived the average postage fee for each city[51]. Specifically, I take the simple average of postage fee for a destination city across all origin cities. I calculate the distance matrix of cities based on their centroids. Coupled with population data, I measure market access for each city. In addition to the overall trend in online shopping measured by the Baidu Index, and the cross-sectional variation measured by the online shopping indices, I obtain the number of online stores in different categories for both Tmall and Taobao in each city. I assign the online stores to their registration cities, despite that being online means that it can provide its products to all cities. This data allows me to examine the heterogeneity in the effects of online shopping on traffic congestion.

2.4 Econometric Models for Online Shopping and Traffic Congestion

This section discusses the econometric models to estimate the effect of online shopping on traffic congestion. First, I present simple regression models that quantify the changes in intracity traffic congestion index and intercity travel time surrounding the event. Then, I present the ordinary least squares (OLS) estimates and instrumental variable estimates of equation 2.32. Third, I present an event-study approach as a robustness check. Due to reasons mentioned in Section 2.3.1, I restrict the timeframe of the analysis to weekdays in one week before the online shopping event and one or two weeks after the online shopping event. Note that all mathematical notations in the regression models have different meanings compared to those in the theoretical model section unless explicitly specified otherwise[52].

[50]The relationship holds when controlling for GDP and population.
[51]Logistics is a highly competitive industry in China, so the price should be very similar across firms
[52]This relieves the burden of finding new Greek letters.

2.4.1 Quantifying the Changes in Travel Time and Traffic Congestion Surrounding the Event

To quantify the change of intracity traffic congestion, I estimate equation 2.35,

$$lnT_{iht} = \sum_t \gamma_t Week_t + \sum_t \beta_t Week_t lnO_i + \sum_t \delta_t Week_t lnS_i + \sum_t Week_t X_i \zeta_t + \iota_i + \lambda_h + \psi_w + \varepsilon_{iht} \tag{2.35}$$

where t indexes weeks, with $t = 0$ indexes the week before the event and $t = 1, 2$ indexes the first and second week after the event. The dependent variable lnT_{iht} is the log traffic congestion index in city i at hour h (in week t). $Week_t$ are dummies indicating the first and second week after the shopping event, with the reference group being the week before the shopping event. ι_i is city fixed effect, λ_h is the hour of the day dummy, and ψ_w is the day of week dummy. γ_1 and γ_2 capture the average change of the traffic congestion index in the first and second week after the shopping event, respectively. I further expect $|\gamma_2| < |\gamma_1|$ as the impact of the event would fade away. To relate the change of traffic congestion to heterogeneous shocks experienced by cities during the online shopping event, I include the interaction of log online shopping index O_i and log online selling index S_i with the week dummies in the regression model. X_i are control variables such as income and the number of internet and mobile users in cities. The interactions of X_i with week dummies control for potential city-specific trends that correlate with income and the number of online consumers.

Similarly, I quantify the change of intercity travel time with equation 2.36,

$$\begin{aligned} Time_{ijht} = \alpha Distance_{ij} + \sum_t \gamma_t Week_t + \sum_t \beta_t Week_t lnO_i + \sum_t \delta_t Week_t lnS_i \\ + \sum_t Week_t X_i \zeta_t + \iota_i + \kappa_j + \lambda_h + \psi_w + \varepsilon_{ijht} \end{aligned} \tag{2.36}$$

where $Time_{ijht}$ is the intercity travel time from city i to j at hour h (in week t). $Distance_{ij}$ is the length of each route between city i and j in km. ι_i and κ_j are origin and destination city fixed effects. As predicted by equation 2.30, roads between cities are expected to be filled with trucks that deliver goods from manufacturers or warehouses to distributors, therefore I expect at least $\gamma_1 > 0$ and $\gamma_1 > \gamma_2$ as the effects are expected to weaken over time.

2.4.2 Ordinary Least Square Estimation of the Effect of Online Shopping on Traffic Congestion

This section explores the variation in the increase of online shopping index across cities to estimate the relationship between online shopping and traffic congestion. Estimating equation 2.32 derived from the theory section requires estimating below first-differences regression

model,

$$\Delta lnT_i = \beta \tilde{B}_{it} + \varepsilon_i \tag{2.37}$$

where i indexes cities. ΔlnT_i is the change of log traffic congestion index. $\tilde{B}_{it}$ denotes the growth rate of Baidu Index for the online shopping platform $\frac{\Delta B_i}{B_i}$, which is the proxy for $\frac{\Delta Q_{oj}}{Q_{oj}}$ in equation 2.32. β estimates the mean elasticity ϵ in the theoretical model.

There are occasional missing values in the traffic congestion index. If the missing values happen on hours or day of week with particularly high or low traffic congestion, then the difference of the outcome variable between weeks may arise due to missing values. Therefore, I add hour of the day and day of week fixed effects to avoid potential biases. For these reasons, I estimate below level specification,

$$lnT_{iht} = \beta \frac{B_{it}}{B_{i0}} + \lambda_h + \psi_w + \mu_{iht} \tag{2.38}$$

where $t = 0, 1$ with $t = 0$ indicating the week before the event and $t = 1$ indicating the week after the event. The dependent variable lnT_{ih} is log traffic congestion index in city i at hour h. B_{it} is the value of Baidu Index in time t, and B_{i0} is the Baidu Index for the week before the event. In $t = 0$, the value of $\frac{B_{it}}{B_{i0}}$ is always 1. The construction of the regressor allows the coefficient from the level specification have the same interpretation as in the change specification as $\Delta \frac{B_{it}}{B_{i0}} = \frac{\Delta B_i}{B_i}$ given there are only two periods. The value of B_{i0} is taken on the day of 11 Oct 2016, and B_{i1} is measured by the peak value on the day of the event, as explained in Section 2.2.3. λ_h represents hour dummies, and ψ_w represents the day of week dummies. The residual μ_{iht} in equation 2.38 can be further decomposed into three components,

$$\mu_{iht} = \iota_i + \tau_t + \varepsilon_{iht}$$

where ι_i represent city-specific unobserved components that are fixed over time. τ_t represents general time effects due to the seasonality in traffic congestion, and error term ε_{iht}. I include city fixed effects to account for ι_i and time trends to account for τ_t in the regression model.

There are three potential issues with OLS estimation of equation 2.38. The first is measurement error. I measure the change of online shopping by the number of searches of the online shopping platform. If the measurement error is "white noise", then it biases the OLS estimates toward zero, which is the well-known attenuation bias[53]. The second is reverse causality. The estimate of β will be biased if the change in traffic congestion can affect the change of online shopping. Consumers in the cities with a higher level of traffic congestion may prefer online shopping. The city fixed effects can alleviate this concern of the effect of the level of traffic congestion on the level of online shopping since I essentially regress the

[53] I cannot assess whether the measurement error complies with the classic measurement error (CME) model. See Angrist & Krueger (1999) for the consequences of other types of measurement errors.

percentage increase in traffic congestion on the percentage increase in online shopping. However, the change in traffic congestion might affect the change in online shopping. Consumers that observed a higher reduction in traffic congestion have a higher chance to choose offline shopping or other types of travels by cars, which leads to an overestimation of the effect (or an underestimation of the absolute value of the effect if it is negative). As argued in Section 2.2.2.2, the temporary reduction in traffic congestion is unlikely to be detected by commuters, especially as the study limits the timeframe to a narrow time band[54]; however, this cannot rule out the risk completely. The third issue is omitted variable bias (OVB). Although the city fixed effects and common time trend have eliminated the possible correlation between the level of online shopping with the residual, the change of online shopping may correlate with the city-specific trend in the residual, that is, $\Delta\varepsilon_i$ correlates with $\frac{B_{it}}{B_{i0}}$. For example, the true model may include the interaction terms of time trends τ_{it} with road network density presumably because cities with denser road network may experience less traffic congestion under the similar level of a travel demand shock. Road network density is also likely to negatively associated with the increase in online shopping as cities with denser road network may have more street shops within walking distance of consumers, which makes online shopping a less attractive shopping option. This leads to an overestimation of the effect. To tackle the potential endogeneity issues in the estimation above, I propose an instrumental variable (IV) identification strategy.

2.4.3 The Instrumental Variable Estimation of the Effect of Online Shopping on Traffic Congestion

The proposed IV is the interaction of the online event with the average postage fee between cities conditional on its market potential. The rationale behind the IV is that postage fee is a major factor in deciding the amount of online shopping in cities, and importantly the online shopping platform waived the postage fee on the day of the event. Hence, places that had higher postage fee are expected to consume more during the limited time window of free shipping as they have a higher opportunity cost for not participating in the sale event. However, postage fee is likely to correlate with other factors in deciding the trade volume of a city, which in turn may affect the change in traffic congestion during the event. First, the most significant confounding factor is the remoteness of a city in the trade network. Cities with higher postage fee are likely to locate further away from other cities. Further, the postage fee is likely to be a function of trade quantity, which is a function of remoteness. A higher trade volume means higher scale economy in the trade route, so the freight cost in each route can be reduced. Second, the importance of the size of the waived postage fee depends on the price index of a city. The model in Redding & Sturm (2008) shows that the price index is a

[54]Consumers have to change their travel behavior very fast in light of the change in traffic congestion, which seems unlikely. For example, Hall et al. (2019) shows that Uber drivers' earnings adjusted to fare cuts fairly slowly although the information of fare cuts is very clear given the digital platform context.

function of market access. Cities with higher market access have lower price index. Given these considerations, I control for polynomial terms of the market potential of a city in the IV specification. The market potential is measured by the weighted sum of the population in all destinations j that can be reached from origin i by incurring transport cost c_{ij} along a specific route between i and j. That is:

$$M_i = \sum_{j \neq i} \frac{N_j}{c_{ij}} \tag{2.39}$$

where M_i is the market potential of city i, N_j is the population in city j, and c_{ij} is the straight line distance from city i to city j. I use the simple inverse cost weighting scheme similar to Gibbons et al. (2019) and Couture et al. (2018). For robustness checks, I construct additional sets of market access variables using other measures of market potential instead of population. One measure is the overall number of online shops in each city listed in the online shopping platform of Alibaba. Another is the number of Tmall shops in each city. Tmall shops are certified online retailers with established brands and revenues above a certain threshold. It is the Tmall shops that are available for the online shopping event during the event day.

Specifically, I estimate below system of regressions:

$$\tilde{B_{it}} = \tau^{1st} + \gamma P_i + f(M_i) + X_i\theta^{1st} + \upsilon_i \tag{2.40}$$

$$\Delta lnT_i = \tau^{reduced} + \delta P_i + f(M_i) + X_i\theta^{reduced} + \epsilon_i \tag{2.41}$$

$$\Delta lnT_i = \tau^{IV} + \beta^{IV}\hat{\tilde{B}}_{it} + f(M_i) + X_i\theta^{IV} + \varepsilon_i \tag{2.42}$$

where P_i is the average of waived postage fee in city i[55], $f(M_i)$ is a polynomial function of the market potential in city i, X_i represents other control variables including GDP per capita, the number of internet users and the number of mobile users. Equation 2.40 is the first-stage regression that estimates the effect of waived postage fee on the changes in online shopping. Equation 2.41 is the reduced-form regression that estimates of the effect of waived postage fee on the changes in traffic congestion. Equation 2.42 provides the 2SLS estimate of β^{IV}, which identifies the causal effect of the change of online shopping on the change of traffic congestion. Above IV specifications can be written in a level specification similar to equation 2.38. In that case, the instrument will be the interaction of the online shopping event dummy with the pre-event average postage fee[56].

The key identification assumption is that conditional on the polynomial terms of market

[55] $P_i = \frac{\sum_j P_{ij}}{I}$, where P_{ij} is the postage fee from city i to j, I is the number of cities.

[56] For example, the first-stage regression model using the level of traffic congestion is:

$$lnB_{it} = \iota_i + \tau_{it}^{1st} + \gamma P_i \tau_{it}^{1st} + f(M_i)\tau_{it}^{1st} + X_i\tau_{it}^{1st}\theta^{1st} + \upsilon_{it}$$

potential and other possible controls:

1. $Cov(P_i, \tilde{B_{it}}) \neq 0$, that is, P_i affects the change of online shopping (relevance);
2. $Cov(P_i, \varepsilon_i) = 0$, that is, P_i only affects the change of traffic through online shopping (exclusionary restriction);
3. $Cov(P_i, \beta) = 0$ and $Cov(\tilde{B_{it}}, \beta) = 0$, that is, both the waived postage fee and the growth rate of Baidu Index are not correlated with the congestion relief effects of online shopping.

The validity of the first assumption can be tested in the first stage regression. The validity of the second assumption cannot be directly tested but is likely to be satisfied. Given the online shopping event is the only significant event in the short periods of two weeks, it is unlikely that the waived postage fee can affect the reduction of traffic through other intermediate factors other than the increase of online shopping. To account for the possibility that cities with different average income levels and size of consumers may respond differently to the change in postage fees, I control for the former using GDP per capita, and the latter with the number of internet users and the number of mobile users. Finally, for the third assumption, Heckman et al. (2006) show that when both instrument variable and endogenous treatment variable are uncorrelated with gains from the treatment, the IV estimator can obtain the mean treatment effect (the mean of the distribution of β given the heterogeneity). It is reasonable to believe that cities do not foresee the impact of online shopping on traffic congestion and consume more online products or choose to be more responsive to the waived postage fee, because there has not been common knowledge on the gains of traffic reduction from e-commerce. This assumption implies that even if consumers adjust their travel demand based on traffic congestion simultaneously, which is unlikely to be true, the IV can provide unbiased estimates as long as the gains of traffic congestion from online shopping are unclear to commuters.

2.4.4 Event Study Estimates

Above specifications use traffic in the week before the online shopping event as the counterfactual traffic had the share of online shopping not changed. The counterfactual might be contaminated if consumers hold up their consumption until the event day. As a robustness check, I use traffic congestion data following other shopping events of similar influence on consumption as the counterfactual outcome. If consumers do reduce their budgets after shopping events, this approach could cancel out part of the budget reallocation effect. Specifically, I consider a follow-up event one month after the Singles' Day shopping event: the Double Twelve Shopping Event on the day of 12 December each year. On the day of the event in 2016, the event generated $13.85 billion sales, which is along the same magnitude as the online shopping event ($17.6 billion). Prices in the offline channel dropped significantly due to discounts in using Alipay at the counters of the stores. The magnitude of price shocks in

this event is similar to that of the Singles' Day shopping event[57].

2.4.4.1 Traffic in the Weeks After Both Shopping Events

Denote the period dummy D_t, with $D_t = 1$ indicating the weeks after the online shopping event, and $D_t = 0$ indicating the weeks after the offline shopping event. I estimate below equation,

$$lnT_{it}^{W_t=1} = \iota_i + \theta D_t^{W_t=1} + \beta_1 lnO_i D_t^{W_t=1} + \beta_2 lnS_i D_t^{W_t=1} + \epsilon_{it}^{W_t=1} \tag{2.43}$$

where $W_t = 1$ indicates the weeks after both events. O_i is the online shopping index and S_i is the online selling index. I replace Baidu Index with the online shopping index because I cannot measure the change in offline shopping during the offline shopping event, I need a more flexible way to measure the change of online and offline shopping during both events. I use the interaction of the level of online shopping before the event with the period dummy to measure the change in online shopping (replace $\frac{\Delta B_{it}}{B_{i0}}$ with $O_i D_t$), given the fact that the cities with a higher online shopping index experienced a greater increase in online shopping during the event as shown in 2.4. The online selling index could potentially capture the change of traffic in the cities that sell products to other cites, so I include them in the regression. β_1 is a difference-in-differences in style estimator (Cooper et al., 2011), which reflects how traffic responds differently in cities with different intensity of online shopping. If online shopping reduces traffic, I expect $\beta_1 < 0$.

2.4.4.2 Traffic in the Weeks Before Both Shopping Events: A Placebo Test

Replicating the above specification in the weeks before the two shopping events would serve as a placebo test. Had both events not happened, we should not observe the correlation between the change of traffic and the change of online shopping.

$$lnT_{it}^{W_t=0} = \iota_i + \theta D_t^{W_t=0} + \beta_1 lnO_i D_t^{W_t=0} + \beta_2 lnS_i D_t^{W_t=0} + \epsilon_{it}^{W_t=0} \tag{2.44}$$

Here, I expect that $\beta_1 = \beta_2 = 0$.

2.4.4.3 Further Differencing Out Unobserved Trends: Triple Differences in Style

Combining equation 2.43 with equation 2.44 provides another difference-in-differences in style or triple-differences in style estimate of the effect of online shopping on traffic congestion.

[57]It used to be an online event for Taobao stores, which are mostly individual sellers like eBay merchants. As Tmall increasingly dominates the online marketplace and the Singles' Day shopping event (for Tmall) becomes exponentially far-reaching in recent years, the online impact of the Double Twelve event is reported to be negligible. In 2016, the event turned to the offline channel to promote the company's mobile payment product Alipay (similar to Apple Pay). Consumers can obtain up to 50% deals when paying using Alipay during the event.

The control group includes the week before the online shopping and the week before the offline shopping events (in different months), while the treatment group includes the two weeks after both events. The post online shopping event week in the treatment group is treated (by the online shopping event). Taking the difference of equation 2.43 and equation 2.44 can eliminate unobserved city-specific monthly trends $\iota_i D_t$. For example, online price tends to start low after Chinese New Year, and increase mildly throughout the year, reaching a peak before the following Chinese New Year, then plummeting to a low point again[58]. The triple difference in style specification is,

$$lnT_{it} = \iota_i D_t + \iota_i W_t + D_t W_t + \beta_1^{triple} lnO_i D_t W_t + \beta_2^{triple} lnS_i D_t W_t + X_i + \lambda_h + \varepsilon_{it} \quad (2.45)$$

β_1^{triple} in Equation 2.45 estimates a triple differences in style estimator. Again, if online shopping is more traffic-efficient, β_1^{triple} should be negative. Note that this is different from the classical difference-in-differences specification in the literature where the treatment and control groups contain different cross-sectional observations. The control group in this setting contains the same individual cities as in the treatment group, but at different times. The time interval between the two events is about two weeks, which could arguably insulate the impact of the first event from the second event. If that is true, the two weeks surrounding the offline event can be used as the counterfactual outcomes for the two weeks surrounding the online event.

2.5 Initial Evidence on the Connection Between Online Shopping and Traffic Congestion

This section provides initial evidence on the connection between online shopping and traffic congestion. First, I demonstrate that the trends of intracity and intercity traffic break around the online shopping event. Second, I show that there is a substantial change in online shopping patterns around the event. Third, I provide graphic evidence on the correlation between the change in online shopping and the change in traffic congestion.

2.5.1 The Traffic Congestion Trend Surrounding the Event

Figure 2.2 compares the intracity traffic and intercity traffic one week before and two weeks after the event. To highlight the difference of traffic congestion or traffic time in different time of a day, I divide a day into five segments: Morning off-peak (before 7 am), Morning peak (7 am-10 am), Day Off-peak (10 am-17 pm), Evening Peak (17 pm-20 pm), Evening Off-peak (20 pm-0 am). Figure 2.2(a) shows the intracity traffic congestion in one week

[58] See Alibaba Shopping Price Indices (aSPI) http://topic.aliresearch.com/market/aliresearch/aspi.php

before and two weeks after the Singles' Day shopping event. The dashed orange line provides the average traffic congestion index one week before the event and serves as the reference group. The solid blue line moves downward in most segments, which suggests that traffic congestion within cities eased in the first week after the event. The short-dashed green line shows average the traffic congestion index in the second week after the event, which tends to be regressive towards the week before the event. Figure 2.3 plots the de-trended traffic congestion index[59] in the three weeks and highlights the impact of the event on traffic congestion. These changes of traffic congestion index and travel time index in the upper panel and the bottom panel suggests that this short-term surge of online shopping reduces traffic within cities and increases traffic in the intercity roads, which provides suggestive evidence for the prediction from equations 2.27 and 2.30. Figure 2.2(b) shows the trend of intercity traffic congestion surrounding the event. The y-axis is the travel time index, which is obtained following two steps. First, I regress the raw intracity travel time data on hour $\times$ day of week fixed effects to obtain the residuals of the regression. I then add the mean of the raw travel time in the regression sample to the residual. Using the index instead of raw data is to address the missing variable problem. In contrast to Figure 2.2(a), the solid blue line that represents the average travel time index one week after the event moves up substantially in all time segments in a day, and falls to the levels closer to the pre-event week in the second week after the event.

Table 2.3 quantifies the change of traffic within cities following the specification of equation 2.35. The first three columns show the result for peak hours and the last three columns show the result for off-peak hours. The coefficient of $Week_1$ dummy in column 1 shows that peak hour traffic congestion is reduced by 3.3% in the first week. The coefficient of $Week_2$ dummy shows that the traffic congestion index bounces back to the level before the event in the second week after the event. Column 4 shows the corresponding result for the off-peak hour sample. The traffic reduction effect is about one-third of that in the peak hour sample. Interestingly, there is a small increase in traffic congestion in the second week. Given the seasonal trends in traffic congestion, it is difficult to interpret this slight bounce in traffic congestion. Columns 2 and 5 add the interaction terms of week dummies with the log online shopping and the log online selling index. I find that cities with a higher online shopping index experienced a larger reduction in traffic while cities with a higher online selling index experienced a higher increase in traffic, with stronger effects in the peak hour sample. Columns 3 and 6 further control for other city characteristics that might affect the change in traffic congestion due to the event by adding the interactions of the week dummies with these characteristics. Log GDP per capita is used to control for income. Log number of mobile users and internet users are used to control for the number of online consumers. Again the change in traffic is negatively correlated with the online shopping index while positively correlated with the

[59] I regress the traffic congestion index on hourly and day of week fixed effects first and then take the residual.

online selling index. Similar results are found in off-peak hours, although the estimate for the interaction term of $Week_1$ dummy and the log online shopping index reduces traffic and is less precise. These results provide strong evidence indicating that cities engaging in higher intensity of online shopping experience more reduction in traffic congestion. Appendix Table B.2.4.2 shows the regression results of the changes in intercity traffic[60] following equation 2.36.

2.5.2 The Trend of Online Shopping Surrounding the Event

Next, I turn to measure the change of online shopping in the weeks before and after the event using the Baidu Index as a proxy. The search index is consistent with the online shopping index provided by the online shopping platform. Figure B.3 shows that the logarithm of Baidu Index is positively correlated with the logarithm of online shopping index in an expanded sample consisting of 277 cities, both before and during the shopping event. Table 2.4 quantifies the linear relationship between the two indices, controlling for city characteristics. The table presents the results in both the regular sample where I have the traffic congestion data and the expanded sample where I have the online shopping index and the Baidu Index but not the traffic congestion data. In the regular sample, cities with a 100 percent higher online shopping index have, on average, a 63 percent higher value of the Baidu Index in normal days (i.e., non-sales event days). This correlation is reduced by two-thirds after controlling for income and online consumers in Column 2. The Baidu Index increased by about 1.6 times ($exp(0.915) - 1$) during the event. Column 3 shows that the cross-sectional correlation between online shopping index and Baidu Index holds up during the event. The interaction of the event dummy and the log online shopping index is positive but not significant, indicating that cities with higher online shopping index have a higher increase in online shopping. Columns 4-6 show the same results in the expanded sample as Columns 1-3. The pattern remains and the interaction of the event dummy and the log online shopping index is larger and statistically significant as the sample size increases, which suggests that cities that are more adapted to online shopping spent even more during the sale. This is important for interpreting the results presented earlier in Section 2.5.1 and that I will show in Section 2.6.4, where I use the interaction of the event dummy and the online shopping index as the key regressors. This rules out the possibility that online shopping grows less in cities with higher online shopping index due to mean reversion. Thus, this interaction term contains the variation of differential growth of online shopping due to the event across cities.

[60]The intercity comparison uses data in the year 2018. The measure is travel time in minutes. The intracity comparison uses data in the year 2016. The measure is traffic congestion index. It would be ideal to know the intercity travel time in the year 2016, but I started to collect real-time intercity travel time data this year. I did not use intracity traffic congestion data in the years 2017 and 2018, because offline retailers also participated in the Singles' Day Shopping Event.

2.5.3 The Connection Between the Change in Online Shopping Activity and the Change in Traffic Congestion

Does the change in online shopping activity correlate with the change in traffic congestion? Figure 2.4 presents the scatter plot of the change of the logarithm of peak hour traffic congestion against the growth rate of the Baidu Index in one week before and one week after the event. Most cities experienced a drop in traffic congestion after the event. The magnitude of the reduction effect is larger for cities with a higher increase in online shopping as indicated by the dashed line with a modest negative slope. Given that the online shopping event is the only significant event in the narrow two-week window, this provides strong suggestive evidence on the connection between the increase in online shopping and the reduction in traffic congestion. The next section further investigates the potential causal link.

2.6 Regression Estimates of the Effect of Online Shopping on Traffic Congestion

2.6.1 OLS Estimates

Table 2.5 reports the OLS estimates for the effect of online shopping on traffic congestion. The dependent variable is log traffic congestion, and the key regressor of interest is the Baidu Index in time t divided by its value in $t = 0$. The OLS results indicate a negative association between online shopping and traffic congestion. Column 1 follows equation 2.32 and estimates a regression model without a common trend in traffic congestion. The following columns estimate equation 2.38, which accounts for common trends. The sample in Column 2 contains all hours, while Columns 3 and 4 look at peak hours and off-peak hours separately. The result from peak hours is much stronger than the effect estimated from off-peak hours. As the Singles' Day shopping event in 2016 falls on a Friday in the first week and some consumers may adjust their travel plans in order to have enough time to shop on the internet[61], Column 5 excludes Fridays from both pre-event and post-event weeks. The result remains. Finally, the last column excludes both Fridays and off-peak hours, which is the preferred sample specification and the sample will be used in the IV estimation. It suggests that a 10% increase in online shopping reduces traffic congestion by 0.13%. Since the traffic congestion index is likely to correlate within cities in the time dimension, I cluster the standard errors by cities in all columns[62].

[61]The company used live stream to engage consumers during the event. Consumers may go home early to participate in a series of interactive sales.

[62]Estimating the first-differences specification

$$\Delta lnT_i = \beta \tilde{B}_{it} + \tau_i + \varepsilon_i$$

gives very similar point estimates.

Given the granularity of the data in time, I stratify the data by hours and plot the β in equation 2.38 in Figure 2.5. The traffic reduction effect is most significant from 9 am to 10 am and around 7 pm. This is in line with the prediction of the theoretical model. Because the ratio of traffic density to free-flow density $\frac{n}{n_m}$ in peak hours is much higher than off-peak hours, the size of the effect in peak hours should be larger. Intuitively, traffic congestion is more likely to happen when the sum of different types of trips, such as shopping trips, commuting trips, leisure and other trips, exceeds a threshold where the road's maximum traffic capacity is reached[63]. Commuting trips have consumed most of the road capacity and left the roads in a congested or semi-congested situation in peak hours. Therefore the impact of the reduction in offline shopping trips on traffic congestion is much more notable.

2.6.2 The Instrumental Variable Estimates

Table 2.6 reports the results from estimating equations 2.40 and 2.41. The dependent variable is the change of the logarithm of the weekly average traffic congestion index. Specifically, I first take the mean of congestion index in peak hours from Monday to Thursday (excluding Friday to avoid the event day) for each city and week. I take the logarithm and then take the difference across weeks. As shown in Column 1 in the upper panel, the waived postage fee is a strong predictor of the increase of online shopping conditional on log market potential. I further add second-order and third-order polynomial terms of log market potential in Columns 2 and 3 to strip out any variation in the IV that is related to remoteness, trade quantity and price index. In Column 4, I add the control for average income measured by GDP per capita, and the coefficient of interest barely changes. Column 5 further controls for the number of mobile users and internet users. The combination of both variables controls for the number of online consumers. The waived postage fee remains highly significant through all specifications. The F-Statistics in Columns 1 and 2 indicate that the first-stage impact of waived postage fee is very powerful, despite that it drops when adding more potentially irrelevant controls in Columns 3-5. The second panel of Table 2.6 shows the reduced-form result of the effect of waived postage fee on traffic congestion. Waiving postage fee has a significant and consistent effect on the reduction of traffic congestion. Taking literally, these results imply that the interaction of the reduced postage fee and the online shopping event resulted in a surge in online shopping and a reduction in traffic congestion.

Table 2.7 contains the instrumental variable estimates of the effect of online shopping on traffic congestion with three different constructs of market access. In the first panel, the market access is the inverse distance weighted city population, so the reported β^{IV} is simply the ratio of the reduced-form estimates in Table 2.6. The estimates suggest that a 10% increase in online shopping reduces traffic congestion by 1.4%-1.7% in peak hours, which is

[63]See Braithwaite (2017) for Figure 4.3, which shows personal trips by start time and purpose in weekdays in England in the year 2011. Shopping trips are a non-negligible component of traffic in peak hours.

an elasticity of -0.14 to -0.17. Further, these estimates are largely insensitive to polynomial terms of different orders of market access and including control variables on income and the number of online consumers. The results are also robust to different measures of market access. The second panel uses the number of both Tmall and Taobao online stores by cities listed in the online shopping platform of Alibaba and the last panel uses the number of only Tmall shops as the market access. The results support the robustness of the IV estimates in the first panel. Although not reported here, the elasticity of traffic congestion to online consumption for all hours is -0.094 to -0.113.

Note that the IV estimate is much larger than the preferred OLS estimates in Table 2.5. As the study by Løken et al. (2012) shows, the difference between the OLS estimates and IV estimates can be decomposed into two parts: the difference in the *marginal effects* between OLS and IV estimates, and the difference in the *weights* between OLS and IV estimates. Reasons that could explain the differences in the marginal effects have been discussed earlier in Section 2.4.2, including measurement error, reverse causality and omitted variable bias. Here, I explore the possible difference between the OLS and IV estimates arising from the difference in the regression weights following the method proposed in Løken et al. (2012). The regression weights can be calculated as,

$$w_{gi}^{OLS} = \frac{Cov(d_{gi}, \tilde{B}_i)}{Var(\tilde{B}_i)} \tag{2.46}$$

$$w_{gi}^{IV} = \frac{Cov(d_{gi}, P_i)}{Cov(\tilde{B}_i, P_i)} \tag{2.47}$$

where $\tilde{B}_i$ is the Baidu Index growth rate, which is the endogenous variable in regression 2.42. P_i is the instrumental variable: the waived postage fee. d_{gi} are dummy variables constructed as $d_g = 1\{\tilde{B}_i >= b\}$. b are evenly distributed cutoffs with an interval of 0.1 that divides the range of $\tilde{B}_i$ into 20 groups, with g indicating each group. w_{gi}^{OLS} and w_{gi}^{IV} gives the weights of OLS and IV estimates in group g, respectively. As shown in Løken et al. (2012), OLS estimates give more weights to the marginal effects close to the sample median of the regressor, while IV estimates assign more weights to the marginal effects for the changes of the endogenous variable that are most affected by the IV. Given the heterogeneity in β, the IV weights lead to different estimates even if the marginal effects of OLS and IV are the same. Figure 2.6 shows the distribution of the weights IV and OLS estimates, respectively. The Figure reveals that the IV estimate assigns more weights to the marginal effects in the right tail of the distribution of the growth rate of Baidu Index relative to the OLS estimate. A following up question is what are the characteristics of the cities in the right tail. Appendix B.2.4.1 provides some suggestive evidence showing that the IV estimates are identified from the cities with higher market potential, income, and number of online consumers.

2.6.3 Heterogeneity

My theoretical model predicts that the elasticity ϵ_j include three components: congestion level $\frac{n_j}{n_{mj}}$, the importance of online shopping traffic $\frac{\psi_j \pi_o}{\delta_j \pi_o + \zeta_j \pi_f}$, and the traffic saving factor. All three parts could vary with cities, which results in heterogeneity in β. Among the three sources of heterogeneity, the comparison of the effect between peak hours and off-peak hours has shown the heterogeneity due to the first component $\frac{n_j}{n_{mj}}$. For the second component, I do not have data on the share of online consumption (π_o) or the share of shopping-related traffic (ψ), or the share of shopping made through private vehicles (ζ) by cities. This section focuses on the investigation of the heterogeneous effect caused by the third component: the traffic-saving factor.

The theory predicts that the reduction of traffic congestion is caused by higher traffic efficiency in online shopping relative to offline shopping. If that is correct, the more products delivery vans can carry, the more traffic-efficient is online shopping. The number of products that a van of standard size can carry depends on the size of goods; therefore, online delivery of bulk goods may have lowered traffic efficiency relative to smaller pieces of goods such as apparel. This implies that δ varies across product categories. I do not have data on the volume of purchase by categories in cities; however, I extracted the number of online shops by categories from the website of the online shopping platform. In the international trade literature, countries are found to be the net sellers of the products that they demand the most, which is known as the home market effect. Analogously, if such effect also exists in the trade across cites, then the number of online store of certain category could indicate the domestic demand for the products in that category in the registration city (Costinot et al., 2016; Coşar et al., 2018). Assuming that the number of online shops in certain categories registered in a city provides a proxy for the consumption of the category[64], I might obtain some suggestive evidence of how differential traffic-saving factors across product categories causing heterogeneous congestion relief effect. For that reason, I add interaction terms of the change in online shopping with the number of online shops in a certain category in cities to equation 2.42. Formally, I estimate

$$\Delta lnT_i = \alpha^{IV} + \beta^{IV} \hat{\hat{B}}_i + \beta^{IV_{Int}} \hat{\hat{B}}_i H_{ik} + f(M_i) + X_i \theta^{IV} + \varepsilon_i$$

, where H_{ik} is the number of online shops in category k in city i.

I adopt two strategies to estimate the endogenous interaction term $\hat{\hat{B}}_i H_{ik}$. The first IV strategy interacts the waived postage fee with the number of online shops in categories to instrument the endogenous interaction term. The second IV strategy first predicts the growth rate of the Baidu Index after regressing it on the waived postage fee and controls, and then use the

[64]Here, the location choice of retailers indicates local demand for certain products.

interaction of the predicted value with the number of online shops in category k as the IV for the endogenous interaction term. Table 2.8 shows the result for the heterogeneous effects estimated in both ways. The upper panel presents the results estimated with the first strategy, and the lower panel reports the second. There is little difference between the results of the two strategies. Cities with a higher number of online shopping in the categories of furniture, appliance and home products tend to benefit less from the increase in online shopping. This is consistent with the prediction from the theory: the products in these categories are bulky and have lower traffic saving potential ($|\delta - \frac{\zeta_j}{|\lambda|}|$ is lower).

2.6.4 Event Study Estimates

Table 2.9 shows the results for the event study estimation discussed in Section 2.4.4.3. The dependent variable is log traffic congestion index. As explained in Section 2.5.2, the interaction of online shopping index with period dummy captures differential growth of online shopping across cities. I control for city average income and number of online consumers by including the interactions of these variables with the period dummy. Standard errors are clustered at city level to address the serial correlation of traffic congestion in time within cities. The first panel compares the traffic one week *after* the online shopping event with the traffic one week *after* the offline shopping event following equation 2.43. Consistent with the earlier results, the cities with higher online shopping activities experienced a larger reduction in traffic congestion. A 100% increase in online shopping index reduces traffic congestion index by 1.9%, on average. The effect is stronger over peak hours (3.3%) and weaker (1.5%) during off-peak hours. Note that the magnitude is not comparable with the earlier results as the measure of online shopping is different. Interestingly, the online selling index appears to have a positive effect on traffic. I interpret this result as reflecting the increased traffic due to online delivery in the cities that are more actively engaged in online selling activities.

The second panel compares the traffic one week *before* the online shopping event with the traffic one week *before* the offline shopping event following the specification in equation 2.44. When the event is "turned off", the correlation between the change in traffic and the change in online shopping disappeared. This assures that the connection between traffic and online shopping event is not just a superficial correlation. The third panel estimates equation 2.45 which takes the difference of the first two panels to eliminate potential monthly trends. The main result reduces slightly, but remains significant. A 100% increase in online shopping index reduces traffic congestion index by 1.4%, with a larger effect for peak hours (2.5%) and a smaller effect for off-peak hours (1.1%). The online selling index continues to correlate with traffic positively, but the effect loses its statistical significance.

2.6.5 The Effect on Air Pollution

Another question of interest is whether online shopping leads to the reduction of traffic-related air pollution, which is a proposed mechanism in Gendron-Carrier et al. (2018). Table 2.10 shows the results of estimating equation 2.38 using air quality index and six other types of air pollutants as dependent variables. The first takeaway is that there is a common trend of increasing air pollution in the week after the event for the seven indicators except for Ozone. The point estimates for the effect of online shopping on the change of NO2 and CO suggest that online shopping reduces traffic-related air pollution; however, the effects are not statistically significant. Online shopping appears to have caused little changes in the other air pollution indicators. One possible explanation is that the surge of online shopping drives up products production, which could contribute to the common trend of increasing air pollution. It is also possible that delivery vans use diesel and produce higher per-vehicle air pollution relative to private vehicles, which may have offset the reduction effect of online shopping on traffic congestion.

2.7 Welfare Analysis

If we divide the purpose of travels into commuting, shopping and others, then the reduced travel demand for shopping will benefit the other two types of travels. I focus on the congestion relief benefit for commuters in peak hours, because I do not have data on the number of travels for specific purposes and commuting travels are the majority.

Figure 2.7 presents a classic static economic model of traffic congestion. The downward-sloping blue curve is the travel demand function. The upward-sloping green curve is the average cost of trips. The per-trip cost increases with the number of vehicles on roads due to the externalities of traffic congestion. Note this curve is flat when traffic density is sufficiently low. The constant marginal cost of each trip includes costs such as time cost, fuel cost, or fare of a bus trip. The slope of the curve becomes upward when traffic density is higher than free-flow density, when an additional vehicle causes delay for existing vehicles. Given the focus here is traffic congestion in peak hours, I assume that the average costs of trips increase with the number of trips. At the original equilibrium point (q_1, p_1), the total welfare of consumers from these travels is $A + B$. The welfare loss due to negative externality from traffic congestion is $C + D + E$[65]. When e-commerce takes away part of the shopping travel demand, the demand curve shifts inward to the orange line. A new consumer equilibrium will be reached at point (q_2, p_2). The total welfare of consumers from these travels is then $A + C$, and the welfare loss reduces to E. The welfare gain is then $C - B(=(A + C) - (A + B))$. Note that B actually represents the welfare of consumers who switched from offline shopping to

[65]This obtained by integrating the marginal negative externality along the upward-sloping green curve up to the point of the market equilibrium.

online shopping. Assuming offline shopping and online shopping provides the same level of utility, B is negligible. The welfare gain is then just C. If we know the decrease of time cost of travel $p_1 - p_2$, and the number of trips q' at the new equilibrium, then we can approximate part C using the shaded square in the graph (the region of p_1, a, b, p_2). The triangle part a, b, c represents the welfare gain due to the adjustment of travel demand in the long run. Given that I cannot clearly identify the long-term effects for reasons listed in Section 2.8, I will only focus on the short-term effect here.

I estimate the welfare improvement only for Beijing. Apart from data constraint, congestion relief in Beijing has been extensively studied and these prior estimates provide a benchmark to compare the effect of e-commerce with other potential policies such as congestion charge on the demand side and providing new subways on the supply side. According to Gu et al. (2019), the average commuting time is 56 minutes one-way in Beijing in the year 2016. The average peak hour traffic congestion index before the event is 2. This implies that the free-flow travel time is 28 minutes for a typical commuting trip. If the congestion index goes down by 1.4%, then the actual travel time will be 55.2 minutes, and the travel time saving is 0.78 minutes for each way. Therefore a 10% increase in e-commerce saves 1.57 minutes per workday. The value of commuting time can be derived from the annual wage and working hours, discounted by a factor of 0.5 (Anderson, 2014), which is 0.77 *yuan*/minute. Beijing has 5.7 million people commuting by car and 5.2 million people commuting by bus in a workday. Assuming cars and buses are affected by traffic in the same way, and there are 250 workdays in a year, the estimated welfare gain C will be 1.65 billion *yuan* for 239 million US dollars for peak hours. This is about a third of the size of the gain estimated from the supply of new subway lines in Gu et al. (2019). However, the cost of this "policy" appears to be much smaller than the heavy investment in infrastructure, because this is achieved by improving the traffic efficiency in shopping goods within cities and the change of shopping behavior. There is potentially rising cost of traffic congestion in the intercity roads as shown in this paper; however, its aggregated value is likely to be small given the share of population traveling on the highway for commuting is much smaller relative to commuting on intracity roads. There are obviously many other gains and losses in a broader sense, such as benefits from the convenience of online shopping, losses due to the structural change of the economy. However, these factors are not directly related to traffic congestion, and thus they are out beyond the scope of this paper.

2.8 Discussion on the Long Run Effect

Above results show that traffic congestion was reduced in the short-term event. Does the growing popularity of online shopping improve traffic congestion in a longer term? Through the lens of the theoretical model, the empirical results suggest that the traffic-saving factor is

smaller than zero. The interpretation of the result is that the vehicle-saving ratio is sufficiently low, and the amount of offline shopping that consumers are willing to substitute with online shopping is sufficiently large. In the long-run, the vehicle-saving ratio is likely to become even lower. As online shopping is increasingly convenient, the amount of offline shopping that consumers are willing to give can be even larger. Therefore online shopping is likely to continue to reduce the overall shopping vehicle demand. However, adaptation may reserve this result. Similar to the proposition of the fundamental law of road congestion by Duranton & Turner (2011), the reduction in shopping-related vehicle demand might be offset by vehicle demand for other purposes, such as commuting and leisure trips. Without knowing the long-term travel demand, I cannot predict the general equilibrium effect of online shopping. It is possible to answer the question empirically using longer-term changes in online shopping and traffic congestion. Unfortunately, I currently do not have such data. For the change in online shopping, the Baidu Index does not appear to be measuring the long-term change of online shopping. The official statistics indicate that the overall online sale increased by 32.2% between the year 2016 and 2017[66]; however, there is little increase in the Baidu Index. Nevertheless, Gu et al. (2019) provides some positive perspectives on the long-term effect. They find a persistent congestion relief effect of new subway lines on traffic congestion over the timeframe of one year. This suggests that the reduction of traffic achieved through diverting peak-hour road traffic to other transportation modes can be larger than the latent travel demand. E-commerce could at least be a complementary soft policy that reduces traffic congestion along with many other congestion relief policies such as congestion charge and provision of the fast transit systems in dense urban areas.

2.9 Conclusions

The paper has aimed to make three contributions to our understanding of the congestion relief effects of e-commerce in the context of the Singles' Day Shopping Event. First, I derived the elasticity of traffic congestion index to the quantity in online shopping. The elasticity includes three components: traffic density, the importance of e-commerce, and a traffic-saving factor, which is per-unit online good traffic-saving. The condition for online shopping to reduce traffic is the traffic-saving factor being negative. Quantifying the traffic-saving factor depends on knowing the substitution between online and offline channels, which relies on assumptions of channel choices on the demand side.

My second contribution is to develop such a model that can predict the online-offline substitution of quantities. Portraying e-commerce as trade across cities, I develop a trade model with heterogeneous consumers and two shopping channels. Inspired by the mechanism that the lowest price wins the market introduced by the work of Eaton & Kortum (2002), I assume

[66] See http://www.ec.com.cn/article/dssz/scyx/201801/24827_1.html

that consumers have Fréchet distributed matching quality with varieties, and purchase from the channel that gives the higher matching quality. The model thus inherits the beauty in the price distribution properties from Eaton & Kortum (2002) model: the distribution of quantity consumed in a specific channel is independent of the channel. The channel specific quantity consumed can be expressed as the mean quantity consumed and the share of consumers that choose that channel. The effect of the price change on quantity in each channel can then be decomposed into a mean effect and a share effect. Particularly, the online-offline substitution is found to be a concise function of the elasticity of online consumption quantity to the relative price of online to offline channel, and the elasticity of substitution between varieties.

My third contribution is exploiting the traffic congestion reduction induced by the event to provide the first available empirical evidence on the congestion-relief effect from e-commerce. The evidence suggests that the reduction in the traffic congestion index surrounding the event is related to the increase in online shopping. The temporarily waived postage fee during the event day provides exogenous incentives for consumers to switch to online shopping. Using the waived postage fee as an instrument, I estimate that a 10% increase in online shopping reduces traffic congestion by about 1.4%, which is about a -0.14 elasticity. While the anticipation of the event may encourage consumers to hold up consumption until the event day, and thus introduces bias, I find the congestion relief effects of online shopping remain significant in a difference-in-differences specification, where the counterfactual traffic congestion level is the week after another a large scale shopping event. The effect is stronger in peak hours when traffic density is high, and in cities that engage in more online sales of bulk product, which confirms the embedded heterogeneity of the elasticity as predicted by the theoretical model. Welfare calculations suggest that the reduced shopping vehicle demand leads to a welfare gain of about 239 million US Dollars for peak hours, which is about a third of the size of the welfare gains estimated from the supply of new subway lines in a city.

Admittedly, at least three limitations are worth noting. First, I use the trend in the searches of the online shopping platform on the internet monitored by Baidu Index as a proxy of the growth rate of online shopping quantity. The growth rate of the Baidu Index may be systematically higher or lower than the actual growth rate of online shopping quantity and leads to bias in the estimate of the elasticity. Obtaining confidential online shopping data by cities recorded by the online shopping platform can improve the estimate. Second, because the event creates a spike in online shopping, the empirical results may not reveal the effects of the marginal changes in the adaptation to e-commerce. Third, it is difficult to generalize the effect from the day-long shopping event to effects over longer periods of time, without knowing the travel demand elasticity over a longer horizon. Online shopping can reduce shopping vehicle travel demand but may not necessarily reduce traffic congestion due to potential long-run adaptations as predicted by the fundamental law of traffic congestion. Exploring the long-term congestion reduction effects of e-commerce may be an area of future

research.

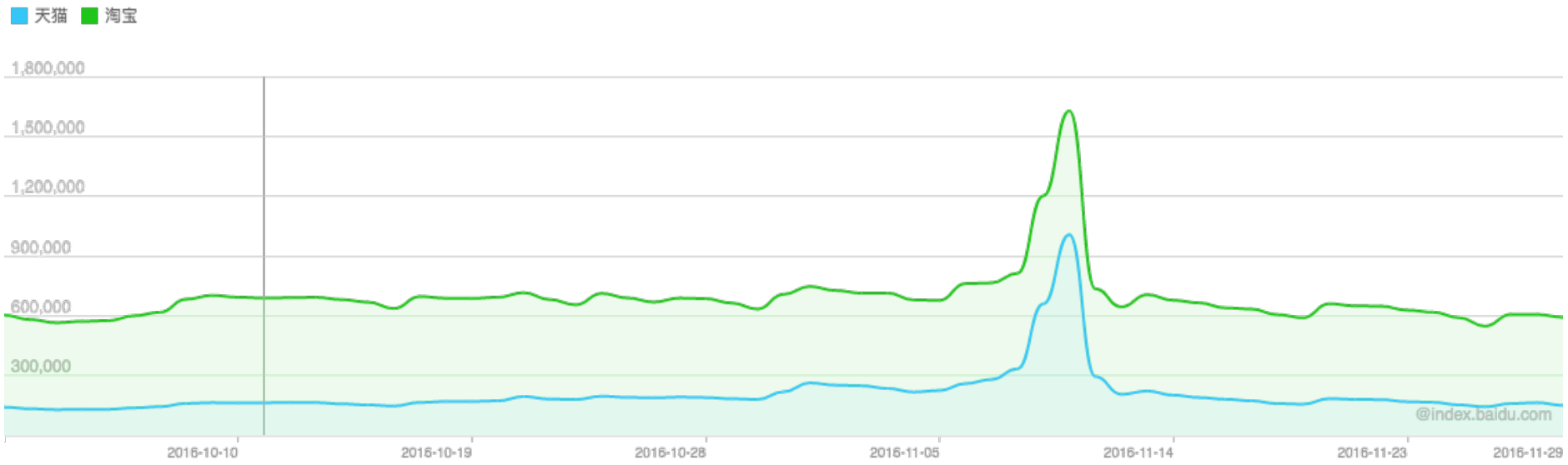


Figure 2.1: Baidu Index between October and November 2016

Note: The y-axis shows the value of Baidu Index, and the x-axis shows date. The green line on top shows the index for Taobao shops, and the blue line in the bottom shows the index for Tmall shops.

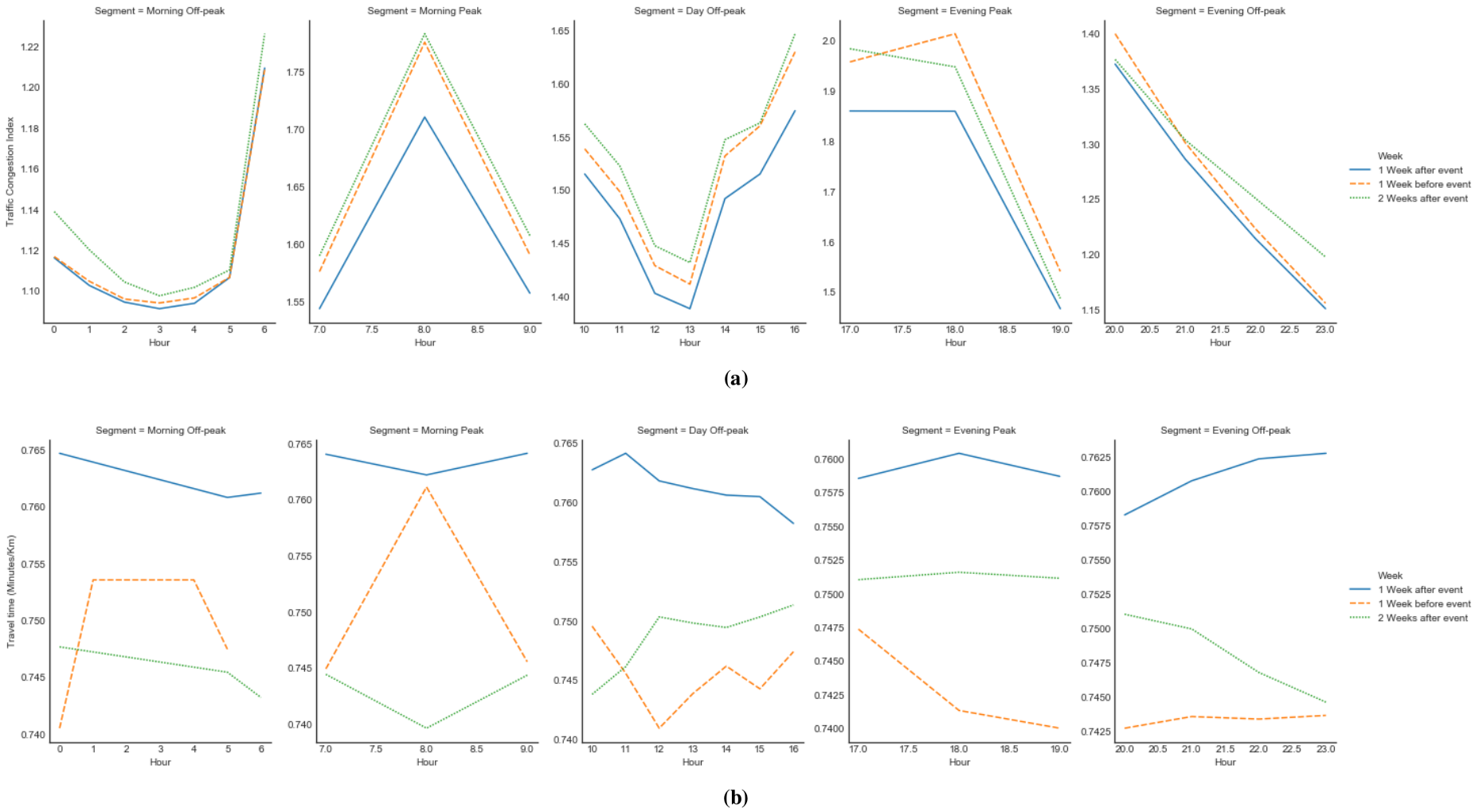
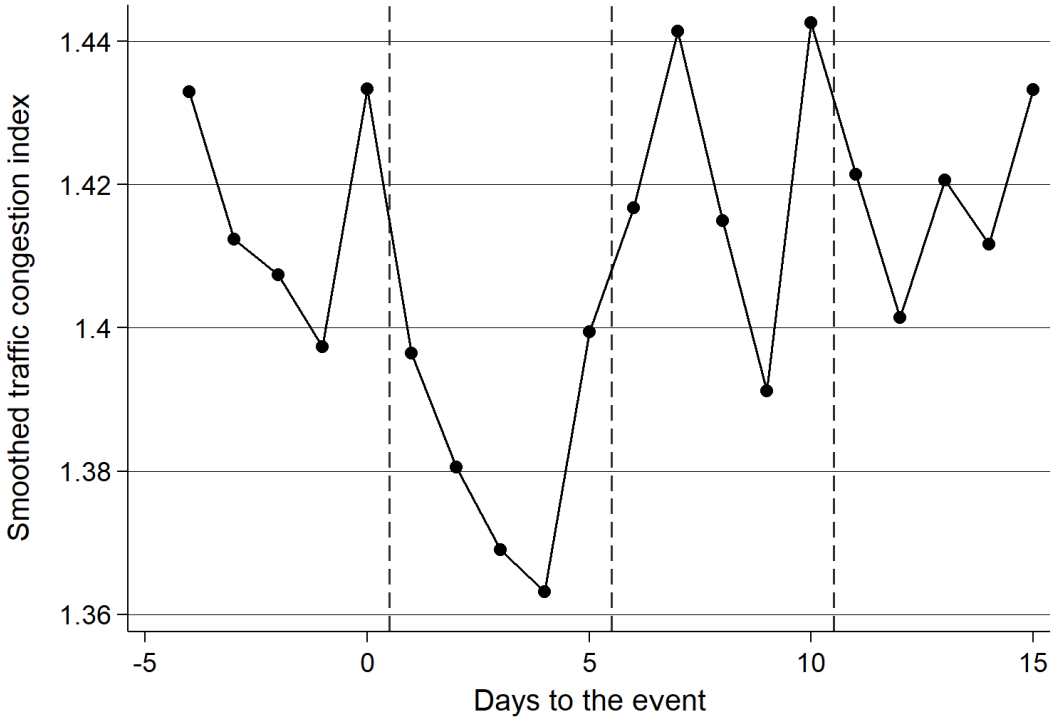


Figure 2.2: The changes in intracity and intercity traffic congestion surrounding the event

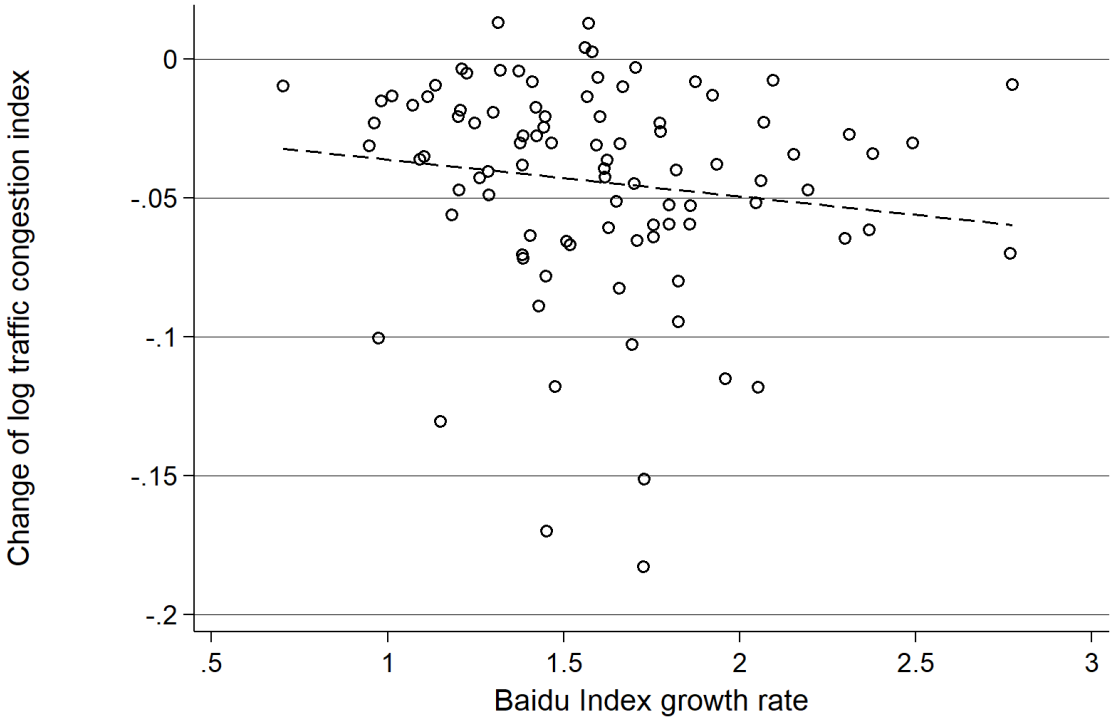
Note: Figure (a) shows intracity traffic congestion during weekdays in one week before and two weeks after the Singles' Day shopping event. Figure (b) shows smoothed intercity travel time in one week before and two weeks after the Singles' Day shopping event. The smoothed measure is obtained through regressing the raw traffic data on the hour × day-of-week fixed effects and adding the mean of the raw travel time in the regression sample to the residual. The graph includes observations in all weekdays in one week before and two weeks after the event. Smoothing is used to address missing observations. Travel time per km shifted upwards in the week after the event and further reduced in the following week.

Figure 2.3: The trend of traffic congestion index surrounding the event



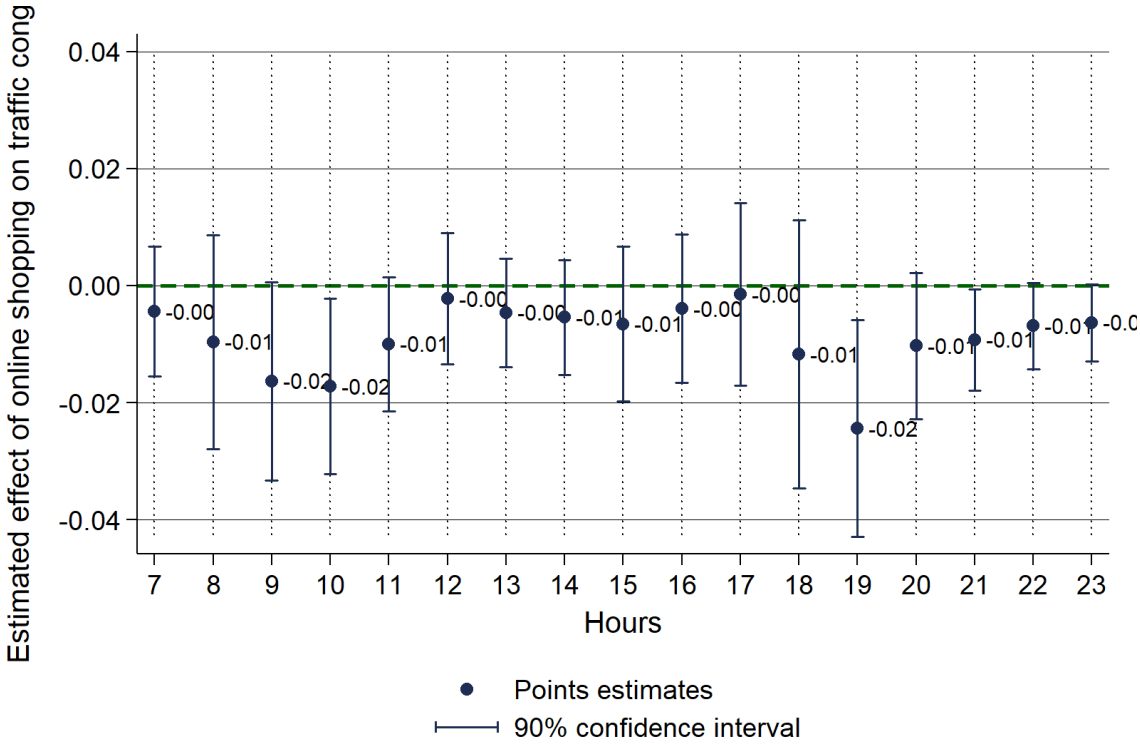
Note: Y-axis shows the residual after regressing congestion index on city, hour and weekday fixed effects. X-aixs shows the days until the event. Dashed vertical lines separate weeks.

Figure 2.4: Change in traffic congestion versus change in online shopping



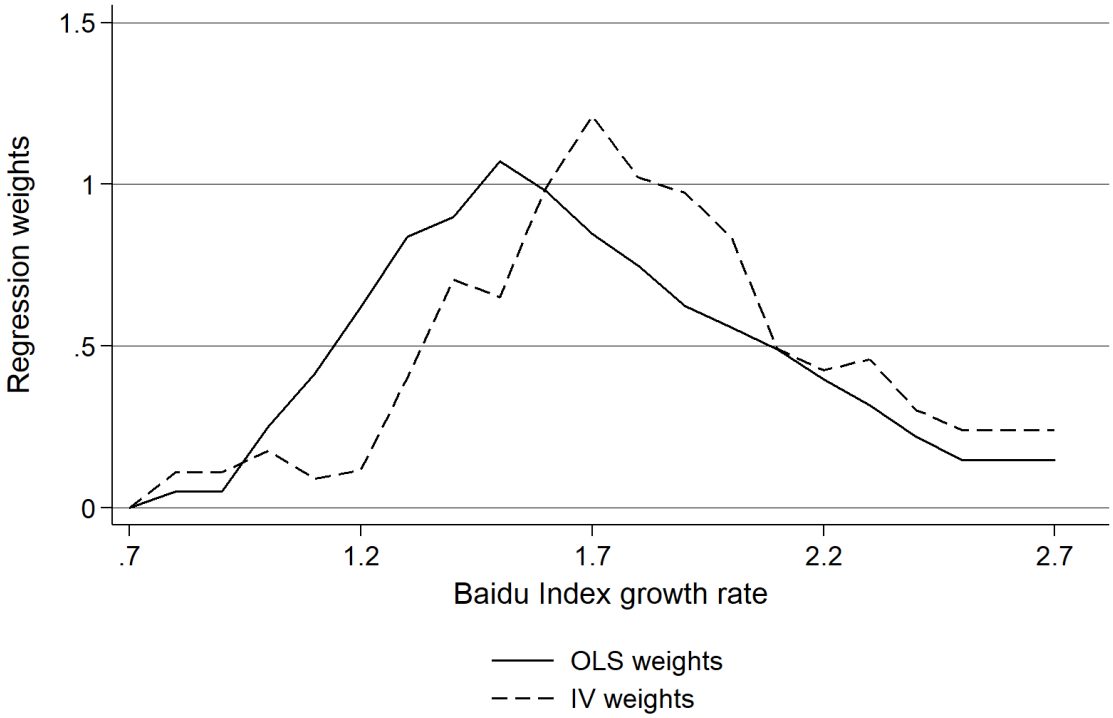
Note: The dependent variable is changes in log traffic congestion index. For each week, I first take the mean of the congestion index in peak hours from Monday to Thursday (excluding Friday to avoid the event day) for each city and then take the logarithm. The key regressor of interest is the growth rate of the Baidu Index.

Figure 2.5: The effects of the increase of online shopping on traffic congestion, by hour



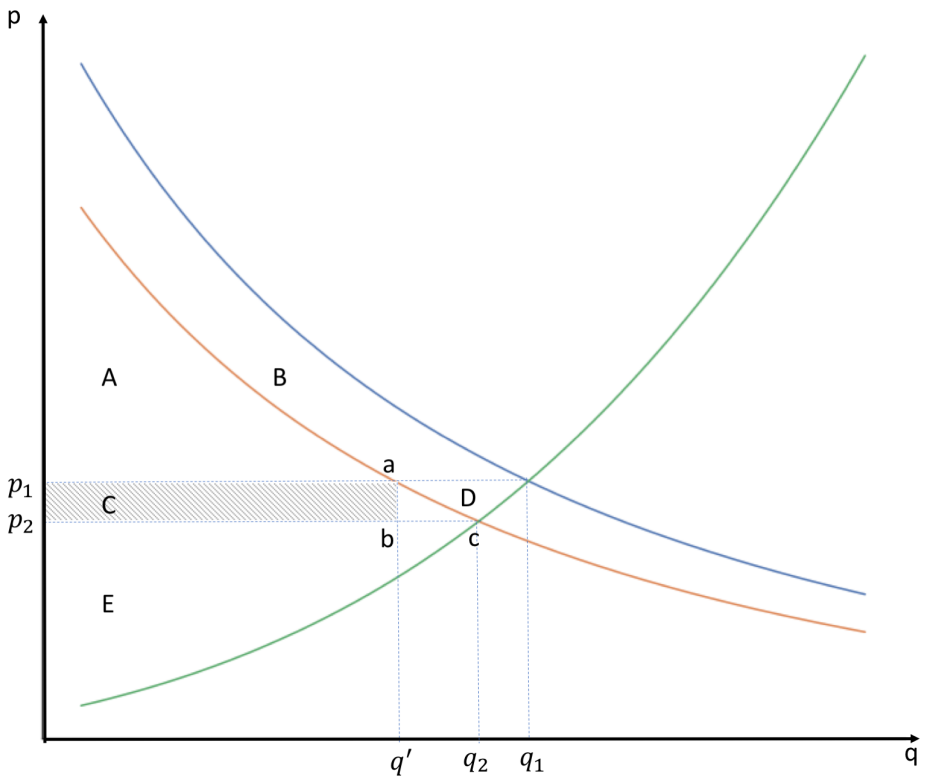
Note: This figure stratifies the data by hours and plots the coefficient β in equation 2.38.

Figure 2.6: OLS estimates and IV estimates weights



Note: The method to produce the regression weights follows Løken et al. (2012).

Figure 2.7: The welfare effects of e-commerce



Note: This Figure illustrates the congestion relief effect due to the increase in online shopping. The blue downward-sloping curve is the travel demand function (willingness to pay for various quantities of trips). The demand curve shifts inward to the orange line when e-commerce reduces shopping vehicle demand. The green upward-sloping curve is the average cost of trips. The per-trip cost increases with the number of vehicles on roads due to traffic congestion when there are enough vehicles so that the traveling speed is below free-flow speed. Given that I focus on peak hours, I assume that the average costs of trips increase with the number of trips. The original equilibrium point is (q_1, p_1). The total welfare of consumers from these travels is $A + B$. The welfare loss due to negative externality from traffic congestion is $C + D + E$. The changes in travel demand result in a new equilibrium at (q_2, p_2). The total welfare of consumers is $A + C$, and the welfare loss reduces to E. The welfare gain is then, $C - B$.

Table 2.1: Parameter values for quantitative analysis

Parameter	Variables	Mean	Source	Year
Statistics				
Sample mean of Online shopping growth rate	B	1.6	Baidu Index data	2016
Mean change in price shock	k	0.25	CCX Credit Technology Online Report	2016
Sample mean of traffic congestion index	T	1.65	Traffic congestion data	2016
The ratio of traffic density to optimal traffic density	$\frac{n}{n_m}$	2.3	Traffic congestion data	2016
Share of online shopping	π_o	0.126	National Bureau of Statistics of China	2016
Share of shopping made through private cars	ζ	0.86	US National Household Travel Survey (NHTS)	2017
Share of shopping vehicles to the overall number of vehicles on the road	ψ	0.31	US National Household Travel Survey (NHTS)	2017
Parameters				
Elasticity of substitution	σ	4	Redding & Sturm (2008)	
Per unit good vehicle-saving ratio	δ	0.067	BBC News	
The ratio of reference speed to free-flow speed	u	5	Notley et al. (2009)	

Table 2.2: Summary statistics for traffic congestion and air pollution by peak hours

	Peak hours		Non-peak hours	
	Before	After	Before	After
Traffic congestion index	1.74	1.67	1.31	1.29
	(0.33)	(0.25)	(0.22)	(0.19)
NO2	42.86	56.44	38.33	52.40
	(20.88)	(26.84)	(21.47)	(27.45)
CO	1.05	1.44	1.00	1.40
	(0.62)	(0.90)	(0.62)	(0.91)
AQI	73.26	105.90	73.58	109.42
	(51.37)	(66.74)	(54.21)	(70.45)
O3	36.11	37.89	40.09	40.40
	(25.31)	(35.91)	(26.67)	(36.04)
PM10	82.85	125.93	83.67	129.36
	(70.85)	(92.78)	(77.42)	(92.47)
PM2.5	46.50	74.09	47.47	77.43
	(36.84)	(55.79)	(38.61)	(58.97)
SO2	20.87	26.30	20.36	26.89
	(21.62)	(28.24)	(21.43)	(29.35)
Observations	2820	2820	8207	8460

Note: The observation is the city-hour. Parentheses contain standard deviations.
* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.3: Estimates of the changes in intracity travel time before and after the event

	(1)	(2)	(3)	(4)	(5)	(6)
Week 1	-0.033***	-0.117***	-0.047	-0.011***	-0.048***	-0.017
	(0.003)	(0.029)	(0.055)	(0.002)	(0.015)	(0.043)
Week 2	-0.005	-0.038	0.147	0.008***	0.009	0.113
	(0.004)	(0.050)	(0.109)	(0.003)	(0.027)	(0.083)
Week 1 × Ln online shopping index		-0.016**	-0.009		-0.009**	-0.009
		(0.007)	(0.006)		(0.004)	(0.005)
Week 2 × Ln online shopping index		-0.004	0.002		-0.002	-0.000
		(0.013)	(0.013)		(0.006)	(0.008)
Week 1 × Ln online selling index		0.045***	0.036**		0.021**	0.020**
		(0.016)	(0.014)		(0.009)	(0.009)
Week 2 × Ln online selling index		0.016	0.016		0.001	0.004
		(0.028)	(0.028)		(0.015)	(0.017)
Week 1 × Log GDP per capita			0.004			0.002
			(0.005)			(0.003)
Week 2 × Log GDP per capita			-0.008			-0.007
			(0.009)			(0.008)
Week 1 × Log mobile users			-0.026***			-0.020***
			(0.007)			(0.006)
Week 2 × Log mobile users			-0.028*			-0.011
			(0.016)			(0.010)
Week 1 × Log internet users			0.013**			0.016***
			(0.006)			(0.006)
Week 2 × Log internet users			0.016			0.008
			(0.016)			(0.010)
R^2	0.614	0.612	0.614	0.842	0.841	0.841
N	11287	10807	10567	33278	31870	31159

Note: The dependent variable is ln congestion index. The omitted group is the week before the Singles' Day shopping event. Columns 1-3 show results for peak hours, and columns 4-6 show results for off-peak hours. City fixed effects, day-of-week fixed effects, and hour fixed effects are included. Standard errors are clustered at the city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.4: The relationship between web searches and the online shopping index

	Regular Sample			Expanded Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Log online shopping index	0.632***	0.198***	0.167**	0.707***	0.285***	0.243***
	(0.057)	(0.056)	(0.064)	(0.032)	(0.032)	(0.037)
Online event dummy	0.915***	0.915***	0.829***	0.803***	0.802***	0.760***
	(0.094)	(0.055)	(0.091)	(0.050)	(0.029)	(0.028)
Online event dummy × Log online shopping index			0.063			0.084**
			(0.065)			(0.038)
Income		Yes	Yes		Yes	Yes
Online consumsers		Yes	Yes		Yes	Yes
R^2	0.575	0.857	0.857	0.646	0.890	0.892
N	184	182	182	552	538	538

Note: The dependent variable is log Baidu index. The key regressor of interest is log online shopping index, online event dummy and their interaction. Income control includes log GDP per capita. Internet controls include log number of mobile users, and log number of households with internet connection. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.5: Ordinary least square estimates of the effect of online shopping on traffic congestion

	(1)	(2)	(3)	(4)	(5)	(6)
	No common trend	Base line	Peak hours	Non-peak hours	Excluding Friday	Excluding Friday Peak hours
$\frac{B_{it}}{B_{i0}}$	-0.011***	-0.009**	-0.012*	-0.007*	-0.008*	-0.013*
	(0.001)	(0.004)	(0.007)	(0.004)	(0.004)	(0.007)
Week 1		-0.004	-0.021**	0.001	-0.006	-0.018
		(0.007)	(0.010)	(0.006)	(0.007)	(0.012)
R^2	0.855	0.855	0.647	0.852	0.857	0.643
N	22307	22307	5640	16667	17795	4512

Note: The dependent variable is log traffic congestion. The key regressor of interest is the Baidu Index in time t divided by its initial level. Column 1 compares the change of traffic congestion in one week before and one week after the event without common time trend. Column 2 allows for common time trend. Columns 3 and 4 show results for peak hours and off-peak hours, respectively. Column 5 excludes Fridays. Column 6 shows peak hour results with samples excluding Friday. City, hour and day-of-week fixed effects are included. Standard errors are clustered at the city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.6: Estimates of the impact of the waived postage fee on changes in online shopping and traffic congestion

	(1)	(2)	(3)	(4)	(5)
Online shopping changes					
Log avg postage fee	1.896***	1.838***	1.807***	1.808***	1.524***
	(0.486)	(0.468)	(0.472)	(0.470)	(0.395)
Traffic congestion changes					
Log avg postage fee	-0.265***	-0.286***	-0.292***	-0.285***	-0.259***
	(0.077)	(0.079)	(0.080)	(0.076)	(0.075)
F-Statistic	29.51	20.19	15.43	14.3	13.72
Market potential	Yes				
Market potential square		Yes			
Market potential cube			Yes	Yes	Yes
Income				Yes	Yes
Online consumers					Yes
N	91	91	91	90	90

Note: The upper panel presents the reduced-form regression results for the changes in online shopping. The dependent variable is the growth rate of the Baidu index, and the key regressor of interest is log average postage fee. The lower panel reports the reduced-form regression results for the changes in traffic congestion index. The dependent variable is the change in log traffic congestion in one week before the event and one week after the event. For each week, I first take the mean of congestion index in peak hours from Monday to Thursday (excluding Friday to avoid the event day) for each city and then take the logarithm. The key regressor of interest is log average postage. Income control includes log GDP per capita. Online consumers controls include log number of mobile users and log number of internet users. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.7: The instrumental variable estimates of the effect of increased online shopping on traffic congestion

	(1)	(2)	(3)	(4)	(5)
Population					
Baidu Index growth rate	-0.140***	-0.156***	-0.162**	-0.157**	-0.170**
	(0.053)	(0.059)	(0.063)	(0.064)	(0.074)
Tmall plus Taobao					
Baidu Index growth rate	-0.103***	-0.118***	-0.123***	-0.123***	-0.143**
	(0.037)	(0.038)	(0.041)	(0.042)	(0.067)
Tmall					
Baidu Index growth rate	-0.096***	-0.097***	-0.096***	-0.095***	-0.096**
	(0.035)	(0.034)	(0.032)	(0.033)	(0.040)
Market potential	Yes				
Market potential square		Yes			
Market potential cube			Yes	Yes	Yes
Income				Yes	Yes
Online consumers					Yes
N	91	91	91	90	90

Note: The first row of each panel indicates which market potential variable construction is used. Market potential in the first panel is constructed with inverse distance weighted city population; The second panel uses the overall number of online shops in cities; The last panel uses the number of Tmall shops in cities. The dependent variable is the changes in log traffic congestion index. For each week, I first take the mean of congestion index in peak hours from Monday to Thursday (excluding Friday to avoid the event day) for each city and then take the logarithm. The key regressor of interest is the growth rate of the Baidu Index. Income control includes log GDP per capita. Online consumers controls include log number of mobile users and log number of internet users. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.8: The heterogeneous effects of online shopping on traffic congestion by product categories

	(1) Office	(2) Clothing	(3) Furniture	(4) Appliance	(5) Home	(6) Electronics	(7) Baby	(8) Cosmetics
Interacted instrument								
Baidu Index growth rate	-0.118**	-0.113**	-0.123**	-0.113**	-0.094***	-0.120**	-0.113**	-0.113**
	(0.049)	(0.052)	(0.052)	(0.047)	(0.035)	(0.052)	(0.049)	(0.050)
Interaction with product categories	0.002	-0.001	0.006**	0.024**	0.009***	0.002	0.003	-0.003
	(0.002)	(0.001)	(0.003)	(0.012)	(0.003)	(0.001)	(0.013)	(0.019)
Predicted instrument								
Baidu Index growth rate	-0.118**	-0.112**	-0.123**	-0.113**	-0.094***	-0.119**	-0.113**	-0.113**
	(0.049)	(0.053)	(0.052)	(0.047)	(0.035)	(0.052)	(0.049)	(0.050)
Interaction with product categories	0.002	-0.001	0.006**	0.023*	0.009***	0.002	0.002	-0.004
	(0.002)	(0.001)	(0.003)	(0.012)	(0.003)	(0.001)	(0.015)	(0.021)
N	90	90	90	90	90	90	90	90

Note: The coefficients are estimated using 2SLS. The dependent variable is the changes in log traffic congestion index. For each week, I first take the mean of the congestion index in peak hours from Monday to Thursday (excluding Friday to avoid the event day) for each city and then take the logarithm. The key regressors of interest are the growth rate of the Baidu Index and its interaction with the number of online sellers of the product category in the column titles. Controls variables are the same as in the last column of table 2.7. Income control includes log GDP per capita. Online consumers controls include log number of mobile users and log number of internet users. Controls also include third-degree polynomials of market potential. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.9: Difference-in-differences in style estimation of the effect of online shopping on traffic congestion

	All hours (1)	Peak hours (2)	Non peak hours (3)
Post week			
Period × Ln online shopping index	-0.019**	-0.033**	-0.015**
	(0.008)	(0.014)	(0.006)
Period × Ln online selling index	0.033**	0.056**	0.025**
	(0.016)	(0.027)	(0.013)
N	12672	3168	9504
Prior week			
Period × Ln online shopping index	-0.005	-0.008	-0.003
	(0.009)	(0.015)	(0.008)
Period × Ln online selling index	0.015	0.024	0.011
	(0.018)	(0.033)	(0.016)
N	12436	3168	9268
Difference			
Period × Post × Ln online shopping index	-0.014**	-0.025**	-0.011**
	(0.006)	(0.012)	(0.006)
Period × Post × Ln online selling index	0.018	0.031	0.014
	(0.013)	(0.028)	(0.011)
R^2	0.881	0.683	0.880
N	25108	6336	18772

Note: The dependent variable is log traffic congestion index. The first panel compares the traffic one week after the online shopping event with the traffic one week after the offline shopping event; The second panel compares the traffic one week before the online shopping event with the traffic one week before the offline shopping event; The third panel estimates the differences between the estimates in the first two panels. Controls variables include the interaction of period dummy with log GDP per capita, log mobile users, log internet users. The first and second panel includes period-fixed dummy and period × city fixed effects. The third panel includes the interaction of the period dummy and post-event week dummy, period × city and post-event week dummy × city fixed effects. Standard errors are clustered at the city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 2.10: Ordinary least square estimates of the effect of online shopping on air pollution

	(1) ln aqi	(2) ln co	(3) ln no2	(4) ln o3	(5) ln pm10	(6) ln pm25	(7) ln so2
$\frac{B_{it}}{B_{i0}}$	0.086	-0.035	-0.039	0.140	0.089	0.085	0.040
	(0.068)	(0.065)	(0.046)	(0.125)	(0.075)	(0.093)	(0.063)
Week 1	0.268**	0.333***	0.348***	-0.335*	0.335***	0.338**	0.156
	(0.109)	(0.106)	(0.076)	(0.191)	(0.119)	(0.152)	(0.106)
R^2	0.552	0.540	0.606	0.534	0.573	0.501	0.648
N	20084	20084	20083	20054	19975	20084	20084

Note: The dependent variables are AQI and another six types of pollutants. The key regressor of interest is the Baidu Index in time t divided by its initial level. City, hour and day-of-week fixed effects are included. Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Chapter 3

Colonial Legacies: Shaping African Cities

3.1 Introduction

This paper shows that the spatial structures of a large set of cities in Sub-Saharan Africa are strongly influenced by the type of colonial rule experienced. Our main findings are based on a sample of 318 cities in 15 former British and 13 former French colonies in Sub-Saharan Africa (excluding South Africa). The extent and nature of land development are based on satellite measures of built cover which span 1990-2014 at fine spatial resolution. Compared to Anglophone cities, Francophone cities are more spatially "compact". There are several complementary dimensions to compactness which we investigate. First, we look at "sprawl" in cities taken as a whole today. Otherwise similar Anglophone cities cover 29% more area and have 17% more "openness" - the measure of sprawl from Burchfield et al. (2006). Second, we look at regularity and density of layout in older sections of the city likely to be physically influenced by how cities were laid out in colonial times. Francophone cities are laid-out in a more Manhattan-like gridiron fashion with lineal road systems, 4-way intersections and rectangular blocks, with higher density development.

Third, in the main body of evidence, we look at disconnectedness of new developments at the extensive margin of cities built well after the colonial era, to see if there is persistence of compactness beyond a margin where colonial influenced physical layout matters. Such persistence would then be due to persistence in colonial planning and land use management practices decades after the end of the colonial era. We focus on the degree to which new developments "leapfrog", or are scattered and spatially disconnected from existing developments, employing a novel measure of leapfrogging. New patches of development can be broken into in-fill and extensions of prior developments versus disconnected, or leapfrog

developments. We define leapfrog developments to be new patches of built cover emerging in a city which are at least 300 meters (or other minimum) away from the edge of any existing development. Anglophone cities have 54% more leapfrog patches compared to otherwise similar Francophone cities. In summary, we find Anglophone cities both sprawl more overall and have more leapfrog new developments than Francophone ones.

Taking our results for the moment as correct, we will argue that the Anglophone-Francophone differences arise because of differences in planning and land management policies and institutions under the two colonial rules. As such, a key lesson from the paper is that major planning procedures and policies put in place at a point in time can leave a huge footprint on urban form decades later. That lesson still leaves two important questions, both of policy relevance. First speaks to motivation: why does sprawl matter? Second speaks to mechanisms. By what means would colonial institutions have both a historical and on-going impact? We address both questions now in the introduction and return to them later in the paper.

Why does sprawl matter? The paper presents some suggestive evidence in an Appendix. For a sample of 45,000 household in 193 cities, we show using Demographic and Health Survey (DHS) data that, conditional on socio-economic characteristics, families have worse connections to electricity, phone landlines, piped water, and city sewer systems, if they live in neighborhoods of a city which are more sprawling, presumably because of increased cost of infrastructure provision. However attempting a focused and full analysis of why sprawl matters would be a full-length paper on its own, contributing to a literature of over 180 papers reviewed in Ahlfeldt & Pietrostefani (2019)'s new meta-study on the "compact city policy paradigm".

On the direct public policy side, planners argue that compactness lowers the cost of providing public services and urban infrastructure. Compact cities require less infrastructure per person in the form of roads and utilities and the opportunity to operate mass transit systems more effectively, with the planning literature offering assessments of the savings from compactness (Trubka et al., 2010; Calderón et al., 2014). Hortas-Rico & Solé-Ollé (2010) provide econometric evidence on public budget cost savings from increased density for Spain, and Carruthers & Ulfarsson (2003) do the same for the USA. The literature also argues that how cities are shaped and sprawl affects how we live: whether we attain efficient density for production in the face of communication externalities (Rossi-Hansberg, 2004; Helsley & Strange, 2007); how much we pollute (Glaeser & Kahn, 2010); how much time we spend commuting (Harari, 2017); and how we interact socially (Putnam, 2000; Burley, 2016), with sprawl argued to lower positive density externalities, increase pollution and commuting times, and enhance social isolation[1].

[1]Using data on German neighbourhoods, Burley (2016) corroborates Putnam's hypothesized correlation between socialization and neighbourhood density, but also presents panel data evidence to suggest that sorting may explain most of this: more social people move to denser neighbourhoods which facilitate socialization. Of

Ahlfeldt & Pietrostefani (2019)'s meta-study after factoring in the credibility of different studies gives recommended elasticities (Table 6) for the impact of increased density. Results of course differ across studies and methodologies. However, the recommended and best founded estimates from a causal point of view suggest increased density significantly reduces energy consumption and vehicle mileage, and, in particular lowers public sector costs, as well as raising wages and improving other production type outcomes. On the other hand, they find evidence on crime rates and pollution to be more mixed. While the evidence overall suggests positive impacts of increased density, regardless of one's take on the normative implications of compactness, there is no question that, for hundreds of millions of Africans, as they urbanize, the institutions and history under which this happens will affect how they live their lives. Colonial origins can have a strong lasting influence on the way people live today in African cities, and most likely other parts of the world.

Why are Francophone cities more compact? From the literature, the answer is that the different institutions imposed under the two colonial rules affected and continue to affect urban spatial structure, including road layouts, sprawl and leapfrogging. The literature on the history of urban planning in Africa argues that the French compared to the British adopted more centralised and standardised urban institutions within cities. Much of this literature is based on contrasting the "indirect rule" strategy of the British with French "direct (and assimilative) rule" (Silva, 2015; Njoh, 2006; Crowder, 1964)[2]. Following an economically-oriented style of rule, the British operated under a dual mandate system and dual structure of local government (Oto Peralias & Romero-Ávila, 2017). Home (2015) develops the dual mandate theme in detail for Anglophone Africa and some other parts of the British Empire: "Native authorities would continue to govern the native population, while townships, largely based on the cantonment model, accompanied the colonizers ... Land laws distinguished between on the one hand, the plantation estates and townships of the European colonizers, and, on the other hand, indigenous or customary land under the dual mandate approach...." (p.55, 57). In summary, generally there was no overall integrative land use plan for a city. Any planning focused on the British sections of the city.

Driven to establish dominance over their colonies and to promote cultural assimilation, local French rule sought to bring all urban land under one control, supplanting all indigenous institutional structures and practices with French varieties, and bringing all public service provision under the local colonial government. Njoh (2006) provides significant details of the French style of rule and city planning for Benin, Cameroon, Cote d' Ivoire, Guinea, Mali, Madagascar, Mauritania, Niger and Senegal, and Togo, with other details in Oto Peralias & Romero-Ávila (2017). As part of maintaining control over the landscape, the French wanted

course that means greater density is of benefit to social people.

[2]In dealing with local chiefs, British rule was advisory while French rule was supervisory, sapping "the traditional powers of chiefs in the interest of uniformity of the administrative system" (Crowder, 1964).

the different neighbourhoods spatially integrated and linked in a lineal pattern so that from one intersection an official could see 2 km in four directions (Njoh, 2016, chap. 1). Different chapters in Silva (2015) also detail how the French adopted centralised and standardised grid systems. Durand-Lasserve (2005) writes about the urban dimension to the direct control strategy: "Customary land management is not recognized.....In former French colonies, this situation is clearly linked with....the French Centralist political model. It is characterised by state monopoly on land, and state control over land markets and centralized land management system....". We interpret these writings as indicating that, at the local level, the French imposed more centralised city planning and land use management than the British. Even in contexts where there are strong pre-colonial institutions especially in countries with a heavy Islam influence and urban history such as Senegal, researchers such as Ross & Bigon (2018) acknowledge the strong French imprint not just in the main colonial centres but throughout urban centres in the country.

How might these differences affect spatial layout? Ross & Bigon (2018) argue that grid structures of neighborhood road systems are found in many places across the globe historically. What is more distinct is attempting to impose such a structure on a whole city so that grid-like neighborhoods are linked together in a common gridiron, not as more disconnected gridded pieces. We expect Francophone cities in the older colonial sections are more likely to have lineal road and rectangular block structures throughout. As such the literature argues that imposing a road grid pattern leads to greater contiguity (Libecap & Lueck, 2011; Ellickson, 2012; O'Grady, 2014), where the aggregation properties of rectangles without gaps or overlaps promotes contiguity of the spatial structure, reducing sprawl. Second leapfrogging seems less likely at least historically in Francophone cities. British dual mandate already allows for disconnected parts of the city; and, then in the disconnected native parts, the British had a hands-off approach. The specifics of French centralized city planning allow for neither[3].

Methodologically, in terms of identification of causal effects of colonial rule, Africa gives an experiment in which initial institutions are given by the happenstance of what colonial power a city fell under, with country borders argued to be imposed from above with no regard for local conditions, history, or past governance (Michalopoulos & Papaioannou, 2016). Even if we think the colonial origin of a city is happenstance, the train of events could have left Francophone versus Anglophone cities ending up on average with very different geographies, which we know affects sprawl and city shape (Burchfield et al., 2006; Harari, 2017). We have a large set of geographic controls to account for these effects and results are robust to many experiments. However there could still be unobserved geographic features of cities or pre-colonial histories which differ systematically between Anglophone and Francophone cities.

[3]However both regimes did advocate racial segregation, the French as part of an integrated physical city layout.

To meet this challenge to identification of causal effects, we conduct a border experiment in West Africa identifying and matching cities within 100 km of borders between different pairs of Anglophone and Francophone countries to ensure more equality in geographic and historical unobservables. We find as strong effects for this border sample.

The context also lends itself to a second experiment, separating out the effects of colonial influences due to the persistence of physical infrastructures and historical city layouts versus persistence in influence in new areas of the city of colonial planning and land management practices and norms. Historical spatial layout of cities affects urban form for many decades. Public capital stocks are long-lived; and rights of way for roads which are key to laying out a city once established are usually followed, given the high costs of acquiring new rights of way in an already built-up area. Persistence due to prior infrastructure could go beyond the older sections of cities to some degree, since postcolonial lay-outs may initially follow existing types of patterns as accretions of older developments. An interesting paper by Huillery (2009) argues that, for Africa, colonial investments in infrastructure had strong impacts on people's access to infrastructure 30 years after the end of colonial rule, potentially reflecting persistence in the infrastructure itself and in post-colonial policies governing infrastructure investments.

However, our context allows for a clear extensive margin where the primary influence cannot be persistence of the physical. For 111 cities of the 318 of our cities which are included in the World Cities Data set[4], population grew by 550% from 1960 to 2000, with approximately the same growth rate of Anglophone and Francophone cities. For 249 of our cities for which we have a 1975 measure of the built cover area, built cover grew by 145% from 1975 to 2014. We will focus on incidence of leapfrogging in areas of cities developed after 1990, in sections of cities built well beyond the colonial physically influenced city. Here differential effects must arise from colonial institutions and city planning and land management practises which persist decades after the end of colonial rule, as well as post-colonial influence of former colonial powers. Home (2015) and Scholz et al. (2015) argue with examples including Ghana and Tanzania that generally Anglophone countries in the post-colonial era maintained colonial era planning practices through at least the end of the 20th century. Individual countries made some adjustments; and, in the 15 years, some have started to experiment with new paradigms. Njoh (2004) makes the same argument for French West African countries, as does Silva (2015, chapter 2): general inertia and continuation of colonial planning practises in the post-colonial era.

This discussion leads to an organization of the paper and presentation of results in four sections, after review of other literature, and introduction of the context and data. First we look overall for cities in 2014 at the degree of sprawl as measured by openness and by overall

[4]http://www.econ.brown.edu/Faculty/henderson/worldcities.html

area of the city. Then we turn to the more likely older and colonial influenced parts of a sample of larger cities for which there is reliable Open Street Map (OSM) data, where persistence of physical layout and infrastructure and its extensions will drive outcomes. There we examine whether these sections of cities have more of a "Manhattan-like" gridiron structure, with lineal roads, 4-way intersections, and rectangular blocks in Francophone compared to Anglophone cities; and, for the full sample, we compare the intensity of built cover in older parts of cities. Next we turn to our main results, where we examine a clear extensive margin where persistence due to existing layout and road structures should not be important. We look beyond the 1990 built area of the city, to identify patches of new development. We examine whether Anglophone cities have higher counts of leapfrog patches and a higher ratio of leapfrog to total number of patches of new development. We then conduct a number of robustness tests and examine the issue of mechanisms and persistence. In the conclusions we return to the question of why sprawl matters.

3.2 Related Literature

The related literature not already covered has two distinct veins: the more general literature on colonialism and the literature on local governance and urban form. For the latter we divide the review into theory versus empirical papers. Our review is brief.

3.2.1 Colonialism and Institutions

The literature on institutions and persistence (e.g., Banerjee & Iyer (2005) and Guiso et al. (2016)) argues that historical institutional accidents can have a strong impact on modern day outcomes[5]. Historical colonial rule and associated institutional choices have been documented to be significant for contemporary institutions, economic development and political stability (Acemoglu et al., 2001; La Porta et al., 2004). La Porta et al. (2008) show that having French civil law as opposed to British common law imposed resulted in differences in regulatory outcomes, banking procedures, property rights enforcement and the like. They argue that French civil law operates to control economic life, while Mahoney (2001) argues that given the ideological differences underlying the two legal systems, French civil law is more "comfortable with the centralized activist government"[6]. Oto Peralias & Romero-Ávila (2017) point out that in Africa with its limited extractive opportunities and large indigenous populations compared to some other British colonial regions, even the imposition of British common law was limited, while the French tended to impose direct centralist rule in all

[5]There is the specific work on Francophone and Anglophone colonial legacy within a small area of Cameroon (split into parts which are former British and French colonies) as affecting wealth and water outcomes Lee & Schultz (2012).

[6]In a different vein, Acemoglu et al. (2011) argue that the imposition of French civil law in the 19th century on areas of Germany which had remnants of feudalism and elitist extractive institutions improved subsequent economic growth.

colonies. For this paper, it is helpful to note the parallels between statements about French civil law and our characterization of Francophone cities as being managed top down by a central city authority with an eye to imposing regularity in design and layout.

3.2.2 Theory Literature on Local Governance, Urban Structure and Sprawl

Theory papers examine the potential impact of an authority with overall control in metropolitan area governance, as opposed to there being either no control or decentralised governance. An older literature dating back to the 1960s on the benefits of centralized city governance does not deal with compactness per se[7]. However recent work focused on specific externalities argues that cities with no centralized planning will have insufficient density. Rossi-Hansberg (2004) and Helsley & Strange (2007) show that, in the face of communication or social interaction externalities which decay with distance, absent appropriate regulation by a single city authority, cities will lack efficient density of activity near the city centre. More informally, Brueckner (2001) and Brueckner (2005) note that uncoordinated developers will take advantage of the fact that congestion is unpriced and public infrastructure may be subsidised which will lead to sprawl, for example, through ribbon developments sited along government built arterial roads. These are strong arguments that uncoordinated and decentralised land development will result in cities that are less compact, offering an empirical prediction in comparing Anglophone and Francophone cities. That said, these papers do not directly model the political forces driving the decisions of a central authority, treating that authority as a benevolent dictator.

Our main empirical results concern leapfrogging, which is examined in two theory papers. Turner (2007) examines whether neighbourhoods on the urban fringe will have leapfrog commercial developments. Henderson & Mitra (1996) consider a city with spatially decaying communication externalities across firms and strategic competition by developers setting up new developments on the city fringe. Such developments may be contiguous to old ones or leapfrog. Both papers argue that higher intensity of development in the core city is associated with a lower likelihood of land developers engaging in leapfrog development at the extensive margin. In summary, the theory papers suggest that centralized control and more intensive core city development which are aspects of Francophone rule can each reduce sprawl.

[7]See Davis & Whinston (1964); Hochman et al. (1995) There is also the huge literature on decentralization of governance within countries. See a general summary in Oates (1999) and for within cities see Helsley (2004) and Epple & Nechyba (2004).

3.2.3 Empirical Literature on Land Use Regulation and Urban Form

A key paper by Libecap & Lueck (2011) uses a border methodology to study the allocation of rural land in Ohio under a 'metes and bounds' system versus a rectangular survey system. The former is a decentralised system with plot alignments and shapes defined by the individuals claiming the land and topographic constraints, while the latter involves centralised and regularised demarcation of surveyed plots. The authors find subsequent strong coordination benefits and reduced transaction costs due to regularity, which they show metes and bounds is less likely to achieve. Their exploration of land demarcation systems in rural areas suggests land institutions in urban areas may be distinguished by their degree of centralisation and standardisation. The parallels to colonial land demarcation systems were extended more directly to cities by O'Grady (2014), focusing on an example comparing a centralised and standardised rectangular grid demarcation with more ad hoc demarcation systems. O'Grady (2014) shows that, for New York City, neighborhoods with a greater fraction of rectangular grids imposed centrally and historically then experienced higher future land values and more compactness, or higher density of use.

Other papers examine persistence in spatial outcomes driven by historical infrastructure investments (e.g. Bleakley & Lin (2012) and Brooks & Lutz (2014))[8]. The key paper on sprawl by Burchfield et al. (2006) analyses geographic and historical influences on the degree of land use sprawl in US cities. Shertzer et al. (2016) argue that 1923 Chicago zoning ordinances have a bigger effect on the spatial distribution of economic activity today than geography or transport networks in Chicago.

3.3 Context, Data, Specification and Identification

In this section we describe the context and data, with more details given in the parts of Appendix C.3.1 and in Table C.1. Then we turn to a base specification and issues of identification.

3.3.1 Colonial Countries

Our classification of African countries by colonial origin is shown in Figure 3.1 (a) along with the cities in our sample. The division is not always straightforward. World War I changed the colonial map, with former German colonies being split among the French (e.g., most of Cameroon) and British (e.g. Tanzania), with many complex splits vis-à-vis modern countries (e.g., Togo). If we think governance procedures and urban plans were developed near the end of the 19th century and early 20th before the end of World War I, those procedures could

[8] Michaels & Rauch (2017) look at the differential influence of the fall of the Roman Empire in France versus England on urban population size and growth centuries later, based on the notion that French Roman settlements persisted after the fall, while British ones due to political upheavals disappeared almost immediately.

set the tone for decades to come. We would then face the problem of German influences confounding the picture. Omission of these countries in robustness checks has no impact on results. While some approaches to governance and land allocations are in place well before World War I, many cities were in infancy potentially limiting the pre-World War I influence.

3.3.2 Data on Land Use and Cities

We utilize three epochs of land cover data - 1990, 2000, and 2014 - which classify pixels of 38m spatial resolution into different uses where our general focus is on built cover (impervious surface) versus non-built cover (water, various vegetation and crop, barren water and so on). These data are constructed from the Global Human Settlement Layer (GHSL) – a new global information baseline describing the spatial evolution of the built environment, a project which is part of the Global Human Settlement Project by the European Commission and Joint Research Centre (Pesaresi et al., 2013). It is the most spatially global detailed dataset on built cover available today. While the data are based on open access Landsat satellite imagery, other information from publicly available and validated coarse-scale global urban data (MODIS Global Urban Extents, MERIS Globcover and Landscan among others) to more fine-scaled and volunteered geographic information (Open Street Maps and Geonames) are incorporated[9] (Pesaresi et al., 2016). Available since 1972 (Ban et al., 2015), the GHSL estimates the presence of built-up areas in different epochs (1975, 1990, 2000 and 2014)[10]. For built-up cover we have two types in any year, the stock of built cover from the prior period (defined to also be covered in the current and subsequent time periods) and new cover built since the last period, which we use to analyse the nature of new development.

In applying these data, we have a base sample of 333 cities, of which 106 are former Francophone cities and 227 former Anglophone cities, with the latter including 122 Nigerian cities. These cities are reported in Table C.1 and shown in Figure 3.1 (a). These 333 cities are all cities in the relevant colonial origin countries which are over 30,000 in estimated population in 1990,[11] which have built cover data for years of 1975, 1990, 2000, and 2014.

[9]Landsat data is typically available at 30m spatial resolution. GHSL employs an information fusion operating procedure based on a tiling schema to combine the source Landsat imagery with other data, imposing further restrictions on effective data resolution - the GHSL project adopts a nominal spatial resolution at the equator of 38.21m which best approximates the native 30m of the Landsat imagery.

[10]Pre-processed Landsat scenes were collected for the epochs (1975, 1990 and 2000) from the Global Land Survey (GLS) at the University of Maryland (Giri et al., 2005) and were combined with Landsat scenes for the 2014 epoch to create the spatiotemporal composite. The epochs that characterise the built-up GHSL data approximate the temporal dimension of the GLS data. Epochs signify a time-period range around a given year from which the best available Landsat scene is drawn. For instance, the 1990 epoch for a city i may be drawn from 1988, while it may be 1992 for city j. Processing uses supervised and unsupervised classification based on a combination of data-driven and knowledge-based reasoning. Spectral, textural and morphological features are extracted and a supervised classification method relying on machine learning is employed using a global training dataset derived from various sources at different scales.

[11]These are based on population censuses around 1990 and with growth to 1990 generally based on city population growth rates between two relevant population censuses.

We use *Citypopulation.de* to get city population numbers (based on Censuses), supplemented with data from *Africapolis* for Nigeria. We set 30,000, because across countries and time there is a difference in population cut-off points for reporting on city populations; a 30,000 cut-off provides more consistency in reporting. We also wanted cities to have some degree of maturity to urban spatial development and planning (or lack thereof). We then apply criteria on the extent of persistent cloud cover to get cloud free city-year observations for 2000 and 2014[12]. Removing cities with cloud cover and hence only partial coverage for land cover, in 2000 we have 299 city observations and in 2010 we have 307, with a total of 318 out of 333 cities in one year or the other.

From the base sample, in robustness checks, we explore various sub-samples. One is West Africa which is distinct as seen in Figure 3.1 (a), in that it contains most of the Francophone countries. Another sub-sample excludes Nigeria which is a third of the sample, to make sure it is not driving the results. There is a sample for Open Street Map analysis of all Francophone cities over 300,000 in 2012, with the size bound imposed to ensure more reliable Open Street Map data which are new to Africa. These 20 cities are then propensity matched to 20 Anglophone cities over 300,000 (out of 68) which have similar attributes like populations, growth and coastal location, among others. We will use these to analyse differential urban structure and road layouts in the colonial portions of larger Francophone versus Anglophone cities. These cities are listed in the Appendix and mapped in Figure 3.1 (b). Finally there is a sub-sample of newer cities founded after 1800, whose origins are more likely to be colonial. These cities are denoted in Table C.1.

3.3.3 Data on Geography and the Extent of the City

In applying these data, we must define the spatial extent of cities. Since outcome measures involve aspects of the built environment, we do not want to use a measure based on built cover per se to define the extent of the urban area. We will note later how that biases results, by tending to omit extensive margin developments which are more leapfrog in nature as opposed to infill and extension. We rely on night light readings (Elvidge et al., 1997) for Africa and define the city to be the area within the outer envelope of all areas lit for at least two of the last 5 years from 2008-2012 (Donaldson & Storeygard, 2016; Henderson et al., 2017). African cities generally have low light levels, so we do not threshold the lights to be above some cut-off. For smaller cities thresholding tends to exclude obvious built areas (looking at Google Earth) and even some entire cities. In very big cities, blooming is an issue and the

[12]We require the city-year to be 95% cloud free in 1990 for initial stock variables and 100% free in 2000 and 2014 for flow variables. We lose 49 city-year observations from imposing the 0 cloud cover restriction and 11 more from requiring no more than 5% cloud cover in 1990. If we imposed a 0 cut-off in 1990 for cloud cover, there would be a loss of another 65 cities. We use the 1990 built cover within our cities at times as a control variable, when looking at flows to 2000 or 2014. Since 1990 defines 2000 pre-built area, in the 2000 analysis any 1990 cloud cover areas in a city are dropped from the calculations for that city.

lights boundary can include large undeveloped areas and cover satellite towns. In various robustness checks, some reported in Baruah et al. (2017) and some later in the paper, we experiment with imposing light thresholds, setting distance limits over which we look, and trimming the cities with high maximal and low minimum distances from the centre to the farthest edge. Results are robust to these experiments. We use night lights to define the city centre, as the brightest lights pixel (about 0.8 x 0.8 km square near the equator) in 1992/93. We also defined smoothed built cover boundaries for cities as described in the Appendix for 1975, 1990, 2000, and 2014. The 1990 measure gives an urban core, beyond which, in the extensive margin, we will find over 98.5% of our post-1990 leapfrog patches.

3.3.4 Specification

Throughout the paper, regressions have the following general form:

$$Y_{ijt} = X_{ij}\beta + Z_{ijt}\theta + \delta Anglophone + \epsilon_{ijt} \tag{3.1}$$

where i is city, j is country and t is time. The initial regressions are cross section. The later leapfrog ones will be flow measures for 1990 to 2000 and 2000 to 2014; there we will add a time dummy to the error structure, d_t. X_{ij} are city i factors which are either time-invariant or for which we want a base period measure. Z_{ijt} are time-varying factors where relevant in the leapfrog regressions. The coefficient of interest is δ - the Anglophone differential. For the border experiment in Section 3.6.2 we will also adjust the error structure to have fixed effects for 14 cross-border clusters of cities in close spatial proximity. In addition, for all relevant tables we do an Oaxaca (1973) decomposition to obtain the differential for Anglophone cities based on the differential in outcomes not explained by differences in characteristics. That is, we estimate

$$\begin{aligned} Y^A_{ijt} &= X^A_{ijt}\beta^A + Z^A_{ijt}\theta^A + \varepsilon^A_{ijt} \\ Y^F_{ijt} &= X^F_{ijt}\beta^F + Z^F_{ijt}\theta^F + \varepsilon^F_{ijt} \end{aligned} \tag{3.2}$$

and calculate the unexplained part $\bar{X}^F(\beta^A - \beta^F) + \bar{Z}^F(\theta^A - \theta^F)$, as well as the explained part.

A basic identification issue is whether Anglophone cities differ from Francophone because of colonial origins or because of differential underlying geographic conditions of cities which influence urban layout, regardless of colonial origins, noting that Burchfield et al. (2006), Saiz (2010) and Harari (2017) all show that geography influences urban form[13]. Our X_{ij} controls on geography reflect this concern. We use measures found in different literatures, starting

[13]There are also social conditions and in a developed country context we might worry about differential attitudes towards use of the automobile and the development of sprawl. First we note that even in seven major Sub-Saharan African cities, automobiles presently account for under 15% of trips (Consortium, 2010). Second, that fraction would have been even smaller in the colonial era.

with Burchfield et al. (2006). First there is terrain where hilly areas spread out developments around inaccessible topography. We have measures of ruggedness as defined by Nunn & Puga (2012) done here at a 30m x 30m resolution and of the range of elevation within the city. Water is another constraining feature - we have whether a city is coastal, distance to the coast from the city centre; and, if the city is coastal, the length of coastline (in km) within its boundary. More extensive coast means more inlets and bays again influencing city shape (Harari, 2017) and thus the possibility of gridiron structures. For the same reason, within the city, we control for the fraction of pixels that are inland water (lakes, rivers, wetlands). We control for average rainfall and average temperature for 1950-2000 and for an index of soil suitability for cultivation from Ramankutty et al. (2002), all reflecting in part the opportunity cost of land at the city edge. We have base 1990 land cover in the city which will appear in leapfrog specifications.

Finally we control for local ethnic-linguistic fractionalization (ELF), where increased fractionalization and the potential for conflict may affect sprawl, for example inducing people to spatially separate more within cities for example. The 19th edition of the World Language Mapping System gives polygons showing the extent of any language group circa the early 1990s. Polygons of different languages may overlap and languages are defined in trees by different levels from 1 up to 15. So a level 1 is Indo-European; level 3 is Indo-Aryan; level 7 includes Punjabi and Urdu; and below level 7 are local dialects such as "Eastern Punjabi". For details on the use of these data see Desmet et al. (2018). Within each city we define ELF as

$$ELF^{l}_{mi} = 1 - \sum_{m} (\pi^{l}_{mi})^2 \tag{3.3}$$

where π^{l}_{mi} is the share of group m in city i at language level l. Ethnolinguistic fractionalization starts at 0 for unilingual cities and rises towards 1. Following Desmet et al. (2018) we generally use level 15 for languages although results do not change with a coarser resolution. We populate the city and construct these shares using the GHS data, with resolution at the 1km grid square.

The hardest items to deal with are growth and economic opportunities for the city. We have initial population size (based on circa 1990 population) and lagged country level GDP per capita, a Z_{ijt} variable. For a city specific growth control, we experimented with various measures of local and national growth of night lights and national urbanization rates which yield similar results, but decided to directly condition outcomes on individual city population growth rates, despite the potential endogeneity issues. Faster growing cities will be more likely to sprawl per se in the short run and Appendix Table C.8 hints at modestly faster average growth among Anglophone cities. We thus decided this was a critical control, and viewed

any feedback of sprawl on growth to be second order. Fortunately, results are essentially the same with and without the economic controls. The growth control is specifically the annualized city population growth rate from 1990 to 2012. This loses us about 9% of the sample due to lack of circa 2012 population numbers. We also control for whether the city is a national capital in 1990 or not; and, if not, its distance from the national capital and the center of power as in Campante & Do (2014). As part of opportunities, we control for the malaria index from Anthony et al. (2004).

3.4 Overall Patterns in the Data for Cities as a Whole

Using the GHSL Landsat based data, first, we correlate an accepted measure of sprawl with Anglophone colonial origins to see motivating patterns in the data. We examine the *openness index* from Burchfield et al. (2006) for the overall city. For openness, following Burchfield et al. (2006), for each built-up 38m x 38m pixel in a city in a year we calculate the fraction of unbuilt pixels in the immediate 1 sq km neighbor. These fractions are then averaged across all built pixels in the city. The measure reflects the extent of open space around the typical built pixel in a city. Second we look at the lights area of cities, to see if, *ceteris paribus*, Anglophone cities occupy larger areas.

What correlations do we see in the raw data? First, we compare distributions of openness for Francophone versus Anglophone cities,[14] based on graphs in Burchfield et al. (2006). Figure 3.2 (a) and Figure 3.2 (b) show the probability distribution function (pdf) for the distribution of built-up pixels in 1990 and 2014 by the percent of land not built in the surrounding one square kilometer (i.e., openness). In both years, the dotted line for Francophone relative to the solid line for Anglophone shows the Francophone pdf shifted left. Francophone cities tend to have a greater fraction of built pixels in areas with very low openness and a smaller fraction of pixels in areas which are very open, suggesting that Francophone development is more compact and Anglophone more sprawled. Visually it may appear that the differential is smaller in 2014, raising the possibility of some convergence. We did explore this issue, but regression work suggests that, conditional on 1990 openness, there is no distinct Anglophone convergence in openness between 1990 and 2014 (Baruah et al., 2017).

Table 3.1 examines the Burchfield openness index in 2014 in regressions controlling for geography and other city characteristics and focusing on the Anglophone city effect. Column 1 has no controls. Column 2 adds in all geographic and situational controls including distance to the national capital, a malaria index, and a local measure of linguistic fractionalization. Column 3 adds the country lagged GDP per capita control and the listed controls on 1990 population size and growth in Table 3.1. Results on these controls are in Table C.2. The

[14]These are 307 cities where the Landsat images used are 95% cloud free in 1990 and 100% cloud free in 2014.

Anglophone effect in column 1 is an increase in openness of 23 %. It drops modestly and insignificantly when geographic and situational controls are added and is little affected by economic controls. The end result is that Anglophone countries have 17 % more openness. Some critical controls variables in Appendix Table C.2 have expected impacts in Burchfield et al. (2006): bigger cities have less openness and cities with greater elevation differentials have more. We note that it could be that French centralised land use control may have responded to differential geography of cities differently than the more decentralised British approach. The Oaxaca decomposition yields unexplained effects which are similar in magnitude to the base regression coefficients in columns 2 and 3.

Columns 4-6 of Table 3.1 turn to the area of cities as measured by the lights boundary, following the same format as columns 1-3. In column 6 with full controls, Anglophone cities occupy 29% more land than their otherwise similar Francophone counterparts. Coefficients for control variables are in Table C.2. The Oaxaca unexplained effect of Anglophone at 27% is close to the base estimate of 29%. One should worry about lights being a noisy measure of larger and smaller city areas. We ran a robustness check in Table C.9 in the Appendix trimming the sample by dropping the bottom and top 2.5% each of the sample by distance from the center to the nearest boundary point. Trimming has little effect on magnitudes, whether we measure sprawl by openness or total area. In summary in Table 3.1, Anglophone cities as a whole have significantly more sprawl. The task now is to look at the two margins, intensive in the colonial influenced sections of the city and extensive in effectively the areas beyond the 1990 built part of the city.

3.5 The Colonial Portions of Cities

We define portions of cities as possibly being under direct colonial influence if they are within 5 km of the city centre. These include the old colonial sections and, through road extensions, the potentially physically colonial influenced sections of the city. For larger cities we know the spatial layouts of the colonial sections, today and some historically from maps. For all cities we know their current intensity of land use, or built cover.

3.5.1 Road Layouts: Anglophone Versus Francophone Cities

To better understand aspects of colonial influence we start with an example, which first compares Bamako to Accra, and then Brazzaville to Harare. Examples are difficult to construct since the key is to have city-pairs for which we could obtain detailed road maps from about the same year near the end of the colonial era. We also want cities with similar initial and final sizes. All four cities emerged as cities in the late 19th century, Bamako and Brazzaville under French rule and Accra and Harare under British. Starting with Bamako and Accra, their populations were similar in the early 20th century: Bamako at 16,000 in 1920

and Accra at 18,574 in 1911[15]. They retain that modest population difference with Accra at roughly 2.3 million and Bamako at 1.8 million today. While Accra is a coastal city, Bamako is on a major river with the initial city on just one side (like a coastline). Bamako had its first (apparently implemented) road plan in 1894 (Njoh, 2006, p. 92), replacing spontaneously prior developed roads with a street network on a classic gridiron with streets intersecting perpendicularly (Njoh, 2001).

Bamako's urban land was under state control by 1907 with the *"Plan d'une cite administrative - un quartier de Bamako"*, with the state supreme in land allocations and assignment of set plots (Bertrand, 2004). Accra proceeded under the usual British dual mandate without a comprehensive plan until The Town and Country Planning Ordinance of 1945 when, according to Grant & Yankson (2003), "zoning and building codes were strictly enforced to maintain an orderly European character and ambience", with a focus on the European Central Business District (Ahmed & Dinye, 2011; Grant & Yankson, 2003).

Figure 3.3 (a) and (b) show the road layout in the older sections of these cities, roughly up to 5km out from the city centre. For Bamako we show the 1963 road layout from tracings of road maps and the road layout today from Open Street Maps. For Accra we show the roads for 1966,[16] as well as today. Visual inspection suggests several takeaways. First there is physical persistence in both cities - roads that were in place 50 years ago generally remain in place today. Second Bamako presents as having large sections of intense dense, gridiron-like road structures where neighborhoods are interconnected by mostly long lineal roads. Also, 1963 fringe roads that appear to meander to the northeast have in some cases been replaced by grid-like structures. New sections of the city generally are on a rectangular grid structure. In contrast, Accra shows much less grid-like structure with fewer lineal connecting roads between developments even in the colonial parts of the city. Moreover, new developments on the fringes of the colonial parts of the city appear to have much less rectangularity and lineal connections than Bamako.

Figure 3.3 (c) and (d) show a comparison between Brazzaville and Harare. Both cities had a population of about 50,000 in 1945[17]. These populations have grown to about 2 million or more based on available information. The Figure shows current OSM roads and the 1958 and 1954 mapped roads in respectively Harare and Brazzaville. The comments we made on rectangularity and lineal connected or gridiron roads systems for Accra versus Bamako

[15]For Bamako: "France: Africa: French West Africa and the Sahara". Statesman's Year-Book. London: Macmillan and Co. 1921. pp. 895–903 – via Internet Archive. Colony of French Sudan. For Accra "Population Studies: Key Issues and Continuing Trends in Ghana" S.N.A. Codjoe, D.M. Radasa, and S.E. Kwankje, Sub-Saharan Publishers, Accra, 2014, p.115

[16]The source of both old maps is Bodleian Library of Oxford University. It is digitized by Ramani Geosystems, a firm based on Nairobi.

[17]From respectively Robert Edmond Ziavoula, ed. (2006), Brazzaville, une ville à reconstruire, Paris: Éditions Karthala; and the Demographic Yearbook 1955 of the UN

apply equally well in this comparison. These illustrative mappings suggest evidence of more regular layout and centralised planning in Francophone cities compared to their Anglophone counterparts.

To test whether these differences hold more generally, we took all 20 Francophone cities in Sub-Saharan Africa over 300,000 in 2012, to analyse road layouts from OSM. Since OSM data are relatively new for Sub-Saharan Africa, we restricted our sample to larger cities and to mapping within 3-5 km of the city centre to try to ensure better reporting. We then chose 20 corresponding Anglophone cities over 300,000 out of the 68 in that size range, using a one to one Mahalanobis distance based matching approach without replacement. The covariates include initial city population in 1990, city annual estimated population growth from 1990 to 2012, average rainfall from 1950 to 2000, coastal dummy, and absolute elevation, noting the comment earlier on including potentially endogenous variables. With matching, means of the matching variables show insignificant differences between Francophone and chosen Anglophone cities (See Table C.3 in the Appendix). Also in the end there are 11 Nigerian out of 20 Anglophone cities, effectively matching Francophone ones concentrated in West Africa. Other samples drawn to reduce the Nigerian count show similar if not stronger results[18].

For this matched sample we ask if the Francophone colonial sections of cities and immediate extensions have different structures than Anglophone ones, with a more regular and connected road system, which would guide the complementary layout of private investments. Here we give quantitative evidence of the more standardised grid system of Francophone cities. Figure 3.4 illustrates the process followed and derivation of measures. In part A we have the raw OSM road network data for part of a city and part B shows the derived roadblocks. Roadblocks are categorised by their degree of rectangularity using the minimum bounding rectangular method of Žunić et al. (2012) and Rosin (1999). The minimum bounding rectangle is a rectangle which minimally encloses the actual block polygon. Rectangularity of a block is the ratio of the area of the block to the area of its minimum bounding rectangle - a perfectly rectangular roadblock would be 1, and the ratio tends to fall as it takes on more complex shapes. Part C of Figure 3.4 ranks all the blocks in the shot - the dark blocks with rectangularity measures equal to or greater than 0.9 are ones we call rectangular blocks. We chose a cut-off of 0.9 to allow for measurement error and topography in approximating perfect rectangles.

Part D of Figure 3.4 shows how we define *gridiron blocks*, which is the basis for our main measure and captures contiguity in rectangularity of layout in the city. To be a gridiron block, a block must be rectangular blocks, be devoid of dangles, and be connected to all neighbouring

[18]For example, for another project, we had a sample of 55 cities generally over 240,000 for which we obtained SPOT data which was weighted against having too many Nigerian cities and towards greater country (Francophone) coverage. We also matched with Anglophone cities without an explicit requirement that they be over 300,000 which again weighs against Nigeria.

blocks by 4-way intersections. Dangles are roads off the regular road network which lead to no connection (i.e., dead-end), or blocks with a *cul-de-sac*, dead-end, or T-intersection; and they are illustrated in Part E of Figure 3.4. Part D of Figure 3.4 shows in yellow the subset of rectangular blocks which qualify as gridiron. For analysis we calculate the share of gridiron blocks to all blocks in the area in question, to capture the degree to which there is an overall, and connected gridiron structure.

We believe OSM data pretty comprehensively maps roads in these 40 African cities up to about 5 km from city centres, covering both the colonial parts of the city which generally lie within 3 or fewer km of the centre and post-colonial immediate extensions. Further out, mapping is expected to be of poorer quality because of the incomplete nature of volunteered OSM information. In Figure 3.5, for each of these cities we show the fraction of gridiron blocks out to 5 km with Anglophone cities represented by the darker shades. Francophone cities generally have higher shares of gridiron blocks. The Anglophone outlier, Bur-Sudan (Port Sudan), was a new "planned city" from scratch, like for example Canberra. The visual impression is confirmed by a regression coefficient giving the average Anglophone differential. Anglophone cities average 20 percentage points fewer gridiron blocks, from a mean of 17. The sample means are almost the same at 3 and 5 km, so there is no overall diminishing of regularity with the 150 percent increase in area covered. We also looked at the share of dangles. Anglophone cities have 3.5% higher shares of blocks (for a mean of 10.7) with at least one dangle to all blocks of the area in question, but the coefficient is only significant at the 11% level.

Overall the results suggest a strong colonial influence of centralised control and grid planning, as suggested by Njoh (2016) and Durand-Lasserve (2005), which persists until today[19].

3.5.2 Intensity of Land Use in the Colonial Portions of Cities and Immediate Extensions

We now return to the full sample. Corresponding to grid-like structures of roads is much greater intensity of land use in the colonial portions and their extensions for Francophone cities compared to Anglophone, indicating much greater compactness. In Table 3.2, for the full sample of cities, we show ring by ring intensity regressions for 2014, as we move out from the city centre in 1 km increments. The dependent variable is the log of the total number of built pixels in each ring. Shown are the coefficients for Anglophone and for an additional covariate beyond those in Table 3.1: the log of number of available pixels (built or not) in each ring by city, which also allows for differential ring counts based on geography (e.g.,

[19]One issue is whether Anglophone cities were regularly laid out but just not on a rectangular grid, using more diagonal roads with roundabout intersections. We checked the count of roundabouts within 5 km of the centre. On average there is no difference between Francophone and Anglophone cities.

rings intersecting a coastline or river).

For rings 1-2, 2-3, and 3-4 km Anglophone cities have 46-64% fewer built pixels, with similar Oaxaca unexplained portion effects. Anglophone rings near the city center are much less intensely developed. After 4km, the sample starts to drop as we lose smaller cities with no area beyond the given radius. At 4-5 km and beyond out to 11-12 km, there are no significant differences for Anglophone countries although coefficients are generally negative. A full ring set of results and many other intensity specifications are in Baruah et al. (2017). We report two here. The first is in column 7 of the table, where all rings are pooled out to 20km and we estimate the height and slope to the intensity gradient as we move away from the city centre. Anglophone cities have 73% less intensity at the centre; Francophone cities have a steep slope to the intensity gradient of -0.29; and Anglophone cities have a significantly flatter slope of -0.19. In Table C.3, when we estimate the gradients separately for the two samples, we get essentially the same slopes of -0.29 and -0.20. Given the height and slope differences, the two gradients cross at about 8 km from the center. This raises an issue of how to examine the comparative extensive margin of cities, which we will discuss below.

The second related aspect concerns the one cross-section for 2014 where, besides the count of built pixels, there is a measure of the intensity of building in those built squares for built pixels: an estimate of the fraction of the grid square that is covered with built surface from the GHSL. Table C.5 shows this for the rings 0-1 to 5-6 km. Anglophone built grid squares are less intensely developed than Francophone ones and significantly so from 1 to 4 km. The bottom line is always that Francophone cities have much more intense land use in the older colonial physically influenced portions of cities than Anglophone cities.

The evidence so far indicates that Francophone cities are more compact overall and in particular in colonial physically influenced sections of cities. We want to look beyond these sections of the city, to examine persistence through more than persistence of physical colonial layout. Moreover openness and intensity raise two problematic measurement issues. First in looking at these measures, it is hard to define the comparative extensive versus intensive margins of cities in a simple way. For the same city populations, Francophone cities are smaller and extend less from the city centre. At, say, 7-9 km, many middle size Francophone cities will be ending, or at their edges and extensive margins. Note above, estimates of simple intensity gradients suggest a cross point of the steeper and higher intercept Francophone gradient with the Anglophone gradient at just under 8 km, indicating Francophone cities will typically be ending near that point. In contrast, corresponding size Anglophone cities will have city edges further from the centre; and 7-9 km will still be the intensive margin. Second, our data measure built cover or footprint of impermeable surface on the ground. They do not measure building volumes since we have no data on heights. It could be Anglophone cities near the centre are built higher with more open ground cover, but a high intensity of

building volume. Of course that would not be consistent with Anglophone cities covering more area overall for the same population. However given these two measurement issues and the desire to focus on the extensive margin, our main results are based on a measure of disconnectedness, or leapfrogging, which we can accurately measure and which operates almost exclusively at the extensive margin of any city.

3.6 Compactness in the (Vast) Post-Colonial Extensive Margins of Cities

Leapfrogging is a concept well established in the literature. Burchfield et al. (2006) in analyzing leapfrogging show that, between 1976 and 1992, much new development in the USA is in areas with little prior development. However, the stock distribution of the fraction of the area within 1 km of a pixel which is not built does not change over time, so this style of development is not changing overall USA sprawl. Here we have a novel and we think better conceptually based measure of leapfrogging. The measure is a flow measures of actual new developments, that are spatially separate from existing developments, for developments from 1990 to 2000 and then again from 2000 to 2014. Using the 1990 to 2000 period as an example, in 1990 we have a set of built pixels, which are typically in clusters. We define the boundary or outer envelope of each cluster of contiguously developed pixels, which we call patches. In the illustrative Figure 3.6, the 1990 developed areas are in light shaded orange. The focus is on newly developed pixels. These also appear as patches of contiguous newly built pixels, which have boundaries. Around each bounded patch (or singleton) of newly built pixels we draw a 300m buffer, effectively including all pixels or parts of which lie within 300m of the nearest border of the new patch. Then we focus on the areas *within (just) these buffers* around new patches to define three types of new development. If that buffer area is generally contained within an existing development it is called infill (red area in the figure). If it only marginally intersects the existing cover (or is within 300m of it), it is called extension (blue in figure). It does not intersect (within 300m) any existing 1990 development it is called leapfrog (green patch).

Our buffer choice of 300m is guided by the literature on "walkable neighbourhoods". Most notably, Barton et al. (2003) claim a theoretical circular catchment of radius 300m (corresponding to walking time of 5 minutes) as a planning goal for urban amenities and interactions. Thus, leapfrogging occurs when a new urban patch development arises beyond the walkable distance of an existing urban patch. Of course, walkable distance is in the eye of the beholder and we experimented with a different size buffer as reported later under robustness checks.

Controlling for the size of the 1990 developed land area of the city, as well as the usual controls including initial city population and population growth rate, we will look at the

comparative count of leapfrog patches in Anglophone versus Francophone cities and the ratio of leapfrog to all new patches of development to ascertain whether Anglophone cities have more scattered development.

3.6.1 Primary Results

We now turn to statistical analysis and look at the absolute and relative count of leapfrog patches in a city and the area they encompass. Most critical to our claim that we are looking at the extensive margin is the fact that over 98.5% of all leapfrog patches in the sample lie outside the smoothed land cover boundary of the city in 1990. These are developments in areas new to the city since 1990. Leapfrogging patches average about 12% of all patches but have high variation across cities (the standard deviation on the variable is 11). While the estimating equation is based on equation 1, we have two periods for each city. The time dummy, d_t, captures the fact that the second time interval (2000-2014) for LF patches is 4 years longer than the first (1990-2000). We also control for initial built cover of the city in 1990. We cluster errors at the city level to deal with serial correlation. Focusing on flows and the extensive margin may help difference out the influence of key unobserved geographic factors. A finding of greater leapfrogging in Anglophone cities would suggest colonial patterns of disconnected and independent developments in Anglophone compared to Francophone cities persist under today's inherited institutions at a margin well beyond the colonial city circa 1965.

We show two sets of results. First is for the count of LF patches. However, a city can have more total patches of greenfield development overall, either LF or contiguous (infill or extension in Figure 3.6). So a second issue is whether a city has a higher ratio of LF to total new patches, indicating less contiguity of new patches to existing ones. Note if $\frac{\partial ln(\text{count LF})}{\partial \text{Anglophone}} = \delta_1$ and $\frac{\partial [ln(\text{count LF}) - ln(\text{count total patches})]}{\partial \text{Anglophone}} = \delta_2$ then

$$\frac{\partial ln(\text{count total patches})}{\partial \text{Anglophone}} = \delta_1 - \delta_2$$

Columns 1-3 in Table 3.3 show basic results for the logarithm of the count of LF patches in a city. We follow the format of Table 3.1, where column 1 has no controls other than the time dummy and initial 1990 cover; column 2 adds the geographic and situational controls; and column 3 adds the economic controls. Geographic controls matter as do growth controls in the sense of reducing the magnitude of the Anglophone coefficient from 0.90 to 0.54. Coefficients on controls are given in Table C.6.

In the main column 3 of Table 3.3, Anglophone cities have 54% more leapfrog patches. In the specification, there is a small count of about 5% of observations which are zeros which we set to the minimum of 1 (so the log is zero). Results in the Appendix Table C.7 show OLS results

excluding these zeros, Tobit results, and Poisson count results. The Anglophone magnitude is very similar across these specifications. We also note the issue again that Francophone cities may respond to differential geography of cities differently than the Anglophone ones. In this case the Oaxaca unexplained portion effect is modestly larger than the base Anglophone coefficient in column 3. We also note that the effects for the two time episodes of leapfrogging are similar if separately analyzed.

In columns 4-6, we show results for the ln (count LP patches/ count total patches). The coefficient on the ratio in column 6 is 0.33, which is little different than the coefficient of 0.34 without controls in column 4. As noted above, columns 3 and 6 together imply that the marginal effect of Anglophone over Francophone for total patches is about 0.21 (0.539 - 0.325). Anglophone cities develop more by building in greenfield areas (all new patches), rather than intensifying already built cover. While the GHSL data do not measure changes in intensification, we know from the one cross section in Table C.5, that built cover is more intensively built in Francophone cities. Given that Anglophone cities have more greenfield development, they are even more prone to these added patches being leapfrog ones. That is, apart from utilizing greenfield development Anglophone cities have less focus on contiguity of new developments to old. In column 7 we show results for ln (average area of LP patches), which checks whether Anglophone patches are somehow bigger, so, for example, they might be easier to service with infrastructure. There is no average size difference in leapfrog patches between the two types of cities. In summary Anglophone cities have more patchy development at the extensive margin, especially leapfrog development, where these leapfrog patches are no bigger than their Francophone counterparts.

3.6.2 Identification

Are the effects in Table 3.3 causal? Qualitatively, causality is suggested through the weight of different pieces of evidence and the use of a large set of controls and flow data, but biases obviously may remain. Although the insertion of many controls has limited impact on the Anglophone "treatment", the characteristics between the Anglophone and Francophone sets of cities are not balanced in all cases (column 1, Table C.8), suggesting there could be unobservables affecting outcomes which are also unbalanced. To deal with this we turn to a border experiment, to try to compare Anglophone versus Francophone cities facing identical circumstances.

Figure 3.7 shows West Africa where 5 Anglophone countries share borders with a number of Francophone countries. At these borders there are no significant waterways. We show cities within a 100 km buffer of the borders involved. Results are almost the same if using a 125 km or 150 km buffer. We choose the smaller buffer, but dropping below 100 km loses too many cities. To refine the border experiment, we break border segments into 14

finer portions, grouping nearby cities into natural clusters, to try to control for unobserved geographic or other influences. These clusters are given in Figure 3.7. Clusters around Ghana and Gambia are country-pair clusters, while those for Sierra Leone and Guinea split naturally into an east and west group. For Nigeria with its huge border, ignoring the green border cluster in southwest Cameroon for the moment, we first match all Anglophone cities to the nearest Francophone city (outside of Anglophone-Francophone Cameroon) creating 8 clusters and then we assign any so far unmatched Francophone cities to the nearest cluster. This algorithm for Nigeria also gives the other cluster groupings just mentioned, except the cities in Burkina Faso have no Ghana counterpart, so these cities are neutralized by their cluster fixed effect. An issue for country borders is that part of the 100 km Cameroon (green border) buffer was under British control after World War I through to the mid-1960s. We did the analysis both excluding and including this area, which is one cluster with cities that are in both Francophone and Anglophone Cameroon, with no Nigerian counterpart. Clearly the Anglophone Cameroon cities have conflicting effects: British heritage versus French rule for 50 years. We think it is better to exclude the area, but results do not vary significantly (Baruah et al., 2017).

With these groupings, have we attained balance? Table C.8 shows our key covariates from column 3 of Table 3.1 regressed on a constant and the Anglophone indicator. Column 1 shows that for the full sample there is a lack of balance for several covariates. All of that disappears for the border sample overall (column 2). When we control for the 14 clusters in column 3, one of the ten covariates has a significantly different mean. That is rainfall, which is not significantly different for Anglophone cities in general nor for the border sample. We believe true rainfall within clusters must be almost identical and that the column 3 difference reflects cross border measurement error based on placement of weather stations and interpolation.

In Table 3.4 for the border sample, we run the same leapfrog regressions as in Table 3.3 but with the smaller sample, we limit the controls to eight variables in Table C.8 plus the time dummy[20]. We show the results for a base case without fixed effects in row 1 and then in row 2, we add the city cluster fixed effects, which matter, to control for unobservables. Results compare 35 Anglophone and 23 Francophone cities generally having two growth incidents, without the Anglophone-Francophone Cameroon segment. In the three columns we show the outcomes: log (count of LF), and log (ratio of LF to total count) and log (average area of leapfrog patches) for each of the two rows. In Table 3.4, the Anglophone

[20]Given the small size of the border sample, we want to consolidate our list of controls to avoid potential low statistical power due to a long list of independent variables. Geographic controls are less important than economic controls as we have minimized the geographic difference in the border and by including city cluster fixed effects, so we keep all economic controls. We kept rainfall as it was unbalanced in column 3 in Table C.8. We chose coast dummy over coast length, and chose ruggedness over elevation range. We dropped fraction of rivers and lakes as there are not significant inland rivers and lakes observed in the cities along the shared borders.

degree of leapfrogging is significantly higher in both specifications, with somewhat larger point estimates than in Table 3.3: 0.80 for the city-cluster fixed effects in Table 3.4 versus 0.54 in Table 3.3. For the ratio of leapfrog to all patches, again magnitudes are higher than in Table 3.3 (0.53 versus 0.33). The net effect on total patches is also higher than in the overall sample: 0.27 (vs 0.21). As is the case in general, areas of patches do not differ by colonial origin. Table 3.4 gives strong evidence that it is colonialism and not other factors driving our results for our key measure, leapfrogging[21].

3.6.3 Robustness

The next issue is robustness of Table 3.3 leapfrog results to other considerations. For that we turn to Table 3.5. In Table 3.5 in column 1 we show the base case. In each of the three sets of rows we show the outcomes: log (count of LF), and log (ratio of LF to total count), and log (average area of leapfrog patches). In columns 2 and 3, first we experiment with types of leapfrog measures. Column 2 removes from the counts and areas any developments that are just one (isolated) pixel (38 m x 38m), as an attempt to deal with an obvious source of mis-measurement of built cover. Column 3 uses a buffer around newly built areas of 60m rather than 300m in defining leapfrog developments. In both cases, the impact on point estimates is fairly minimal and the differential in coefficients between rows 1 and 2 which captures the marginal impact of Anglophone on total patches remains about 0.20. The rest of the columns deal with sampling issues.

Columns 4 and 5 worry about defining city boundaries by night light readings. One issue is blooming of night lights in bigger cities, which then add non-urban areas within the lights boundary where Anglophone versus Francophone differentials might exist for other reasons. Another issue is small cities with minimal electricity provision. In column 4 we define city boundaries not by going to zero lights but by pixel light readings falling below a reading of 5. In column 5 we trim the top and the bottom 2.5% of cities, in terms of maximum distance from the centre to any part of the lights boundary. These experiments reduce coefficients but not significantly so. We note however (but not shown in the table), that if we defined the area of the city as a smoothed 2014 built area cover, that would bias our results. For the leapfrog count outcome, the coefficient is minimally affected, but the ratio of LF to all patches then has a coefficient of zero. By cutting on smoothed cover, we mechanically tend to exclude areas with more leapfrogging relative to other patches.

Columns 6 and 7 turn to different country samples which are potentially more problematic. Column 6 removes countries which were initially German colonies before being assigned to Britain or France after World War I. Dropping those countries (Namibia, Tanzania, Togo, and

[21]To complete the picture we also reran Table 3.1 specifications on openness and area. However the sample size is very small (54) and results are insignificant.

Cameroon) reduces coefficients, but the leapfrog coefficient is still 0.49 (vs 0.54). Dropping Nigeria which is a big portion of the sample, if anything, strengthens results. Column 8 focuses on the sample of 40 cities for which we assembled OSM data. Here the point estimates on absolute and relative LF counts are in line with the rest of the data. That is reassuring for the applicability of the gridiron results. Finally in column 9, we drop national capitals to show they are not driving results.

3.7 Persistence at the Extensive Margin: Mechanisms

Exploring mechanisms and persistence is difficult for Africa. An issue in the analysis is that most relevant data on African cities are only available with intensive fieldwork with use of local archives and contacts. That is a reason the literature on the colonial legacy of urban planning focuses on case studies of individual cities as in the chapters in Silva (2015). For our 318 cities or even a sub-sample of 40, there is no general data source on current or historical cadaster records, shapefiles for municipal boundaries or when city plans were first formulated or amended. For example, we conducted a two-week search of the internet focusing on JSTOR and Google scholar (as well as Wikipedia) by city and country using keywords such as founding, history, plans, maps, planning and the like. Those searches typically lead to books such as *Africa South of the Sahara 2004* from Europa (2003) and ones we have referenced above. We only found 13 cities which specifically listed a first city plan and date from 1885 to 1960. Founding was more useful and gave us a founding date for a large set of cities. We then divided the sample into those with a founding date from 1800-1965, using a longer time frame recognizing early French influence in Africa, versus either no date or a date outside that range. Internet searches of planning departments in Africa universities to see where faculty are trained is fruitless, due to lack of website information. All this is the reason why the literature on colonial and post-colonial planning we have cited is so crucial. It gives the information available, based on scholars' extensive fieldwork in specific cities and countries.

We thought of a number of inferences one might try to make based on the data at hand. One example is that ethnolinguistic fractionalization might mean more sprawl in unplanned Anglophone cities, than planned Francophone cities; but our ELF measure has no impact in general or in interaction terms. Another example involves colonial origin of a city. In Table 3.6 we look at our leapfrog measures in Tables 3.3-3.5. In columns 1 and 2 we try to separate out colonial origin from non-colonial origin cities. An initial thought was that, for colonial origin cities, Francophone cities would be more comprehensively planned while Anglophone less so under the dual mandate. What the columns suggest is more nuanced. Colonial and non-colonial origin cities have similar Anglophone effects for the count of LF patches, both having similar percentages more than their Francophone counterparts. The

difference is in the ratio of LF to total new patches which is only significant for non-colonial origin cities. Recalling our decomposition into total new patches versus the ratio of LF to total patches in Section 3.5.2, the implication is that Anglophone colonial origin cities have more total patches (LF or otherwise) compared to Francophone colonial origin ones, while only Anglophone non-colonial origin have relatively more LF to total patches. This indicates that there is less focus on contiguity of new development in Anglophone non-colonial origin cities, compared to Francophone. The arguably greater Anglophone differential in lack of contiguity for non-colonial cities echoes Ross and Bigon (2018), who argue that French planning was imposed throughout the urban hierarchy, while the British typically had little involvement in cities in which they had little presence (Home, 2015).

One final example looks at other historical influences on spatial layout, in particular the pre-colonial governance situation. Similar to Michalopoulos & Papaioannou (2013), we measure pre-colonial political institutions using Murdock (1967)'s "Jurisdictional Hierarchy Beyond the Local Community Level" index. We assign African cities the index of 0-4, where 4 is very strong pre-colonial statehood[22]. Some of our cities include different indexed areas and we did an area weighted average, so the measure is somewhat continuous up to 4. The idea is that the French might have had less success in imposing their institutions and practices in areas where there was prior strong statehood; but, as with colonial origins, the result is more nuanced. We run separate regressions for the Anglophone and Francophone samples, with the index added in, with results in column 3. For LF patches in rows 2 and 3, the marginal effect of increased pre-colonial statehood on increased count of LF patches is the same under both regimes: more LF development in both situations in cities where there was a stronger statehood presence before colonialism. However only for the British does stronger pre-colonial statehood raise the ratio of LF to total patches. While development in stronger precolonial statehood cities is more patchy within the Francophone context so there is less overall imposition of development through intensification of existing built cover, the French are still able to impose contiguity of the patches to existing developments, so there is not a greater ratio of LF to total new patches in Francophone cities as prior strength of pre-colonial statehood increases.

3.8 Conclusions

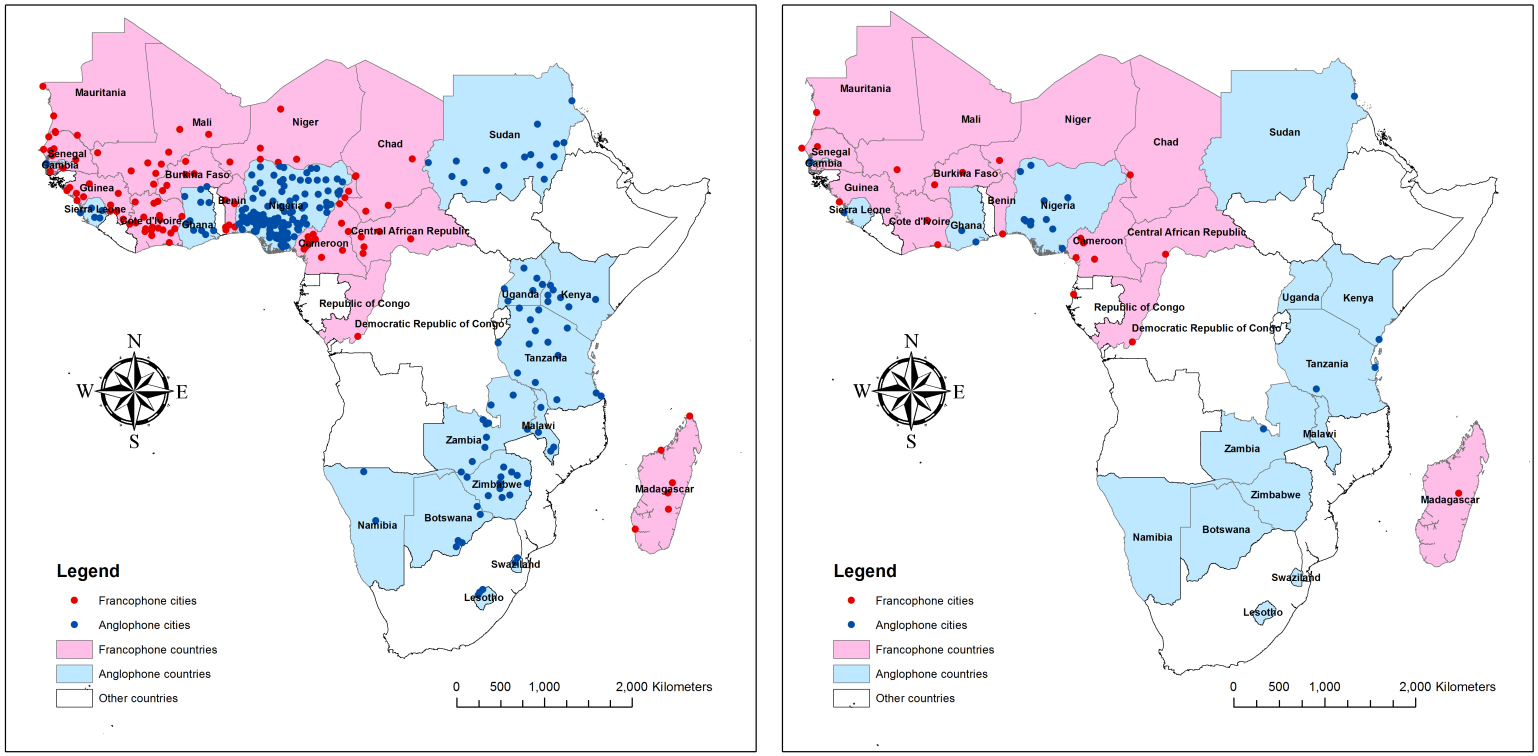
The literature on colonialism in Africa suggests that, compared to the British, the French imposed more comprehensive citywide land use planning, including the layout of roads. The literature suggests these planning practices persisted well into the post-colonial era. The African context of colonialism provides an experiment to show that choosing institutions

[22]Index of value 0 indicates stateless societies "lacking any form of centralized political organization"; 1 indicates petty chiefdoms; 2 indicates paramount chiefdoms; 3 and 4 indicate part of large states.

which involve more centralized land use control within each city, as in Francophone compared to Anglophone cities, leads to more compact cities at both the intensive and extensive margins.

Specifically the paper shows that Francophone African cities have more grid-like structures in their core areas. Anglophone cities have a citywide index of openness which is 17% higher and cover 29% more land. Anglophone intensity of land use is much lower at the centre and, in contrast to Francophone cities, the intensity of land use gradient is flatter. Anglophone cities are more sprawled. Correspondingly, with new development, Anglophone cities have about 54% more leapfrog patches, a number that is robust to a border experiment and many experiments with definitions and relevant cuts on the data in terms of samples.

The question is whether there is a consequence to having greater leapfrog development and more sprawl. Such areas are more expensive to service and potentially less likely to receive connections to public utilities, such as electricity, phone landlines, piped water, and city sewers in an African context. In Appendix C.3.3, we examine this issue using DHS data on whether a family has a piped water connection, an electricity connection, a telephone landline connection, or a (flush) toilet connected to a public sewer system. We find that areas with more sprawl in a city (less intensity of land use and more leapfrogging) in general have poorer connections.



(a) Full sample

(b) 40 cities

Figure 3.1: Spatial distribution of sample cities

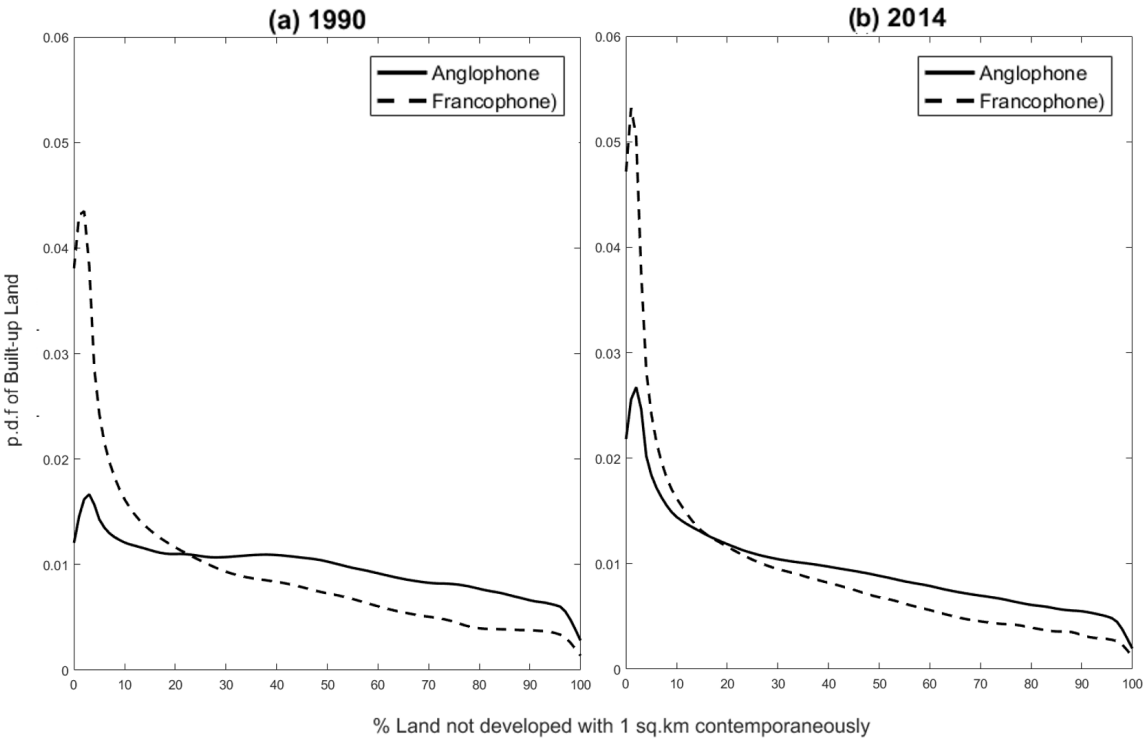


Figure 3.2: Probability function of Anglophone and Francophone built-up land across areas with different degrees of sprawl for (a) 1990 and (b) 2014

(a) Accra **(b)** Bamako

(c) Harare **(d)** Brazzaville

Figure 3.3: Road blocks in Accra, Bamako, Harare and Brazzaville

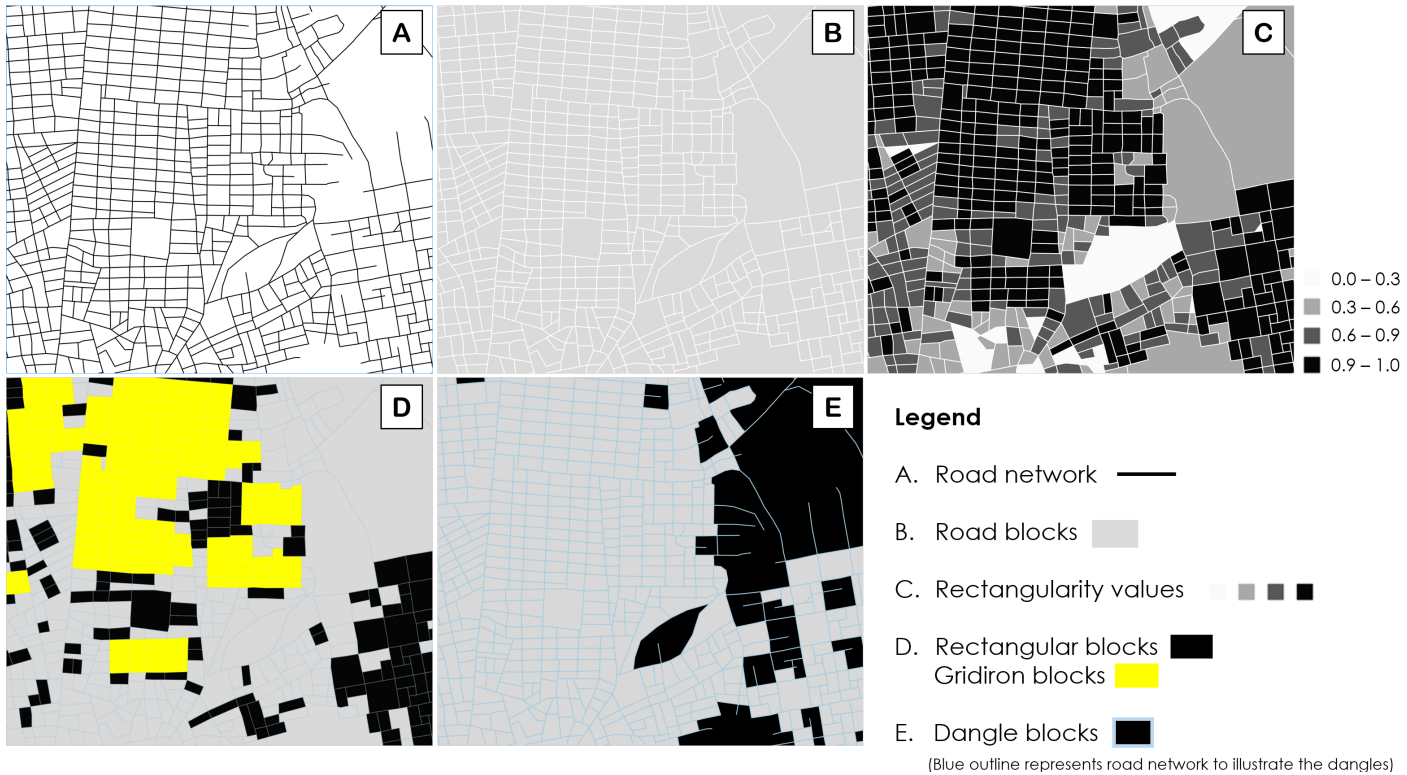


Figure 3.4: Road blocks and rectangularity

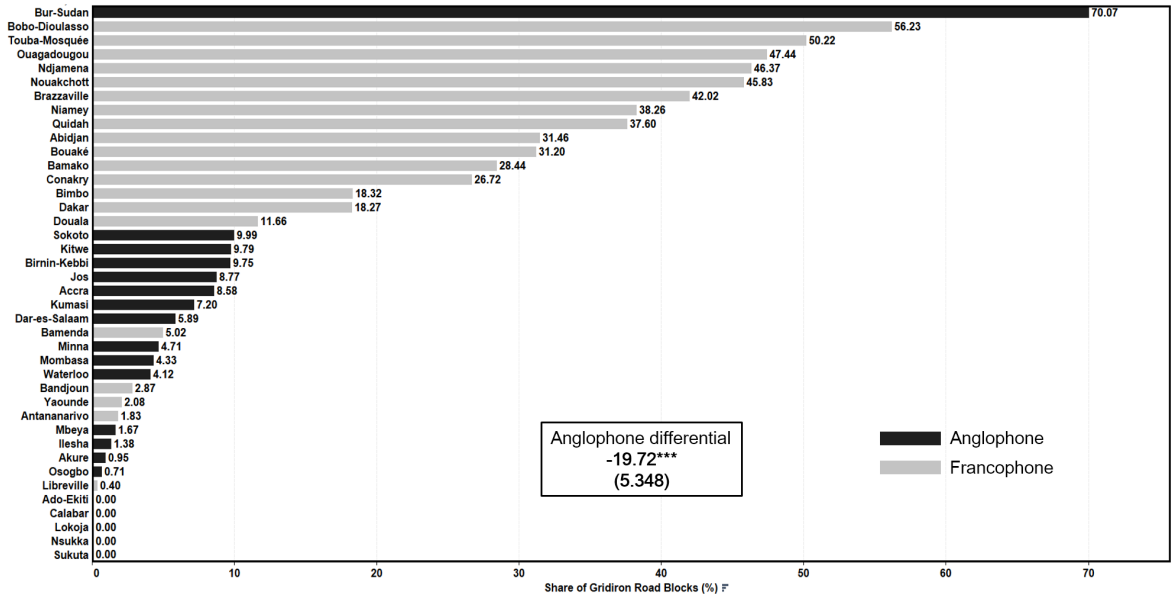


Figure 3.5: Share of gridiron road blocks within contemporary 5km

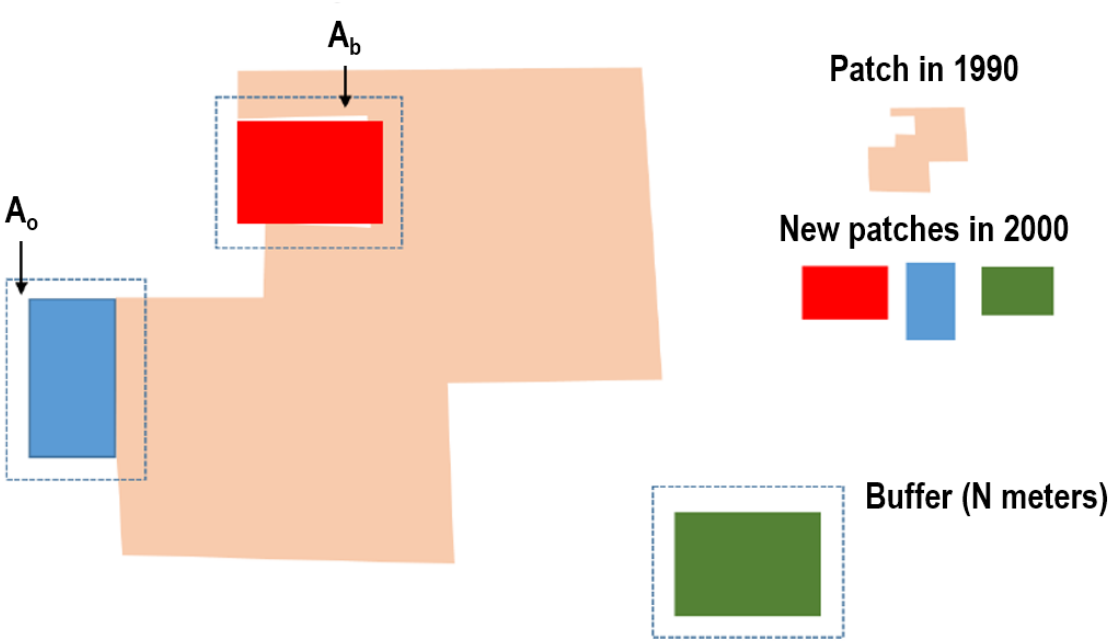


Figure 3.6: Illustration of using the landscape expansion index (Liu et al., 2010) for defining leapfrog patches

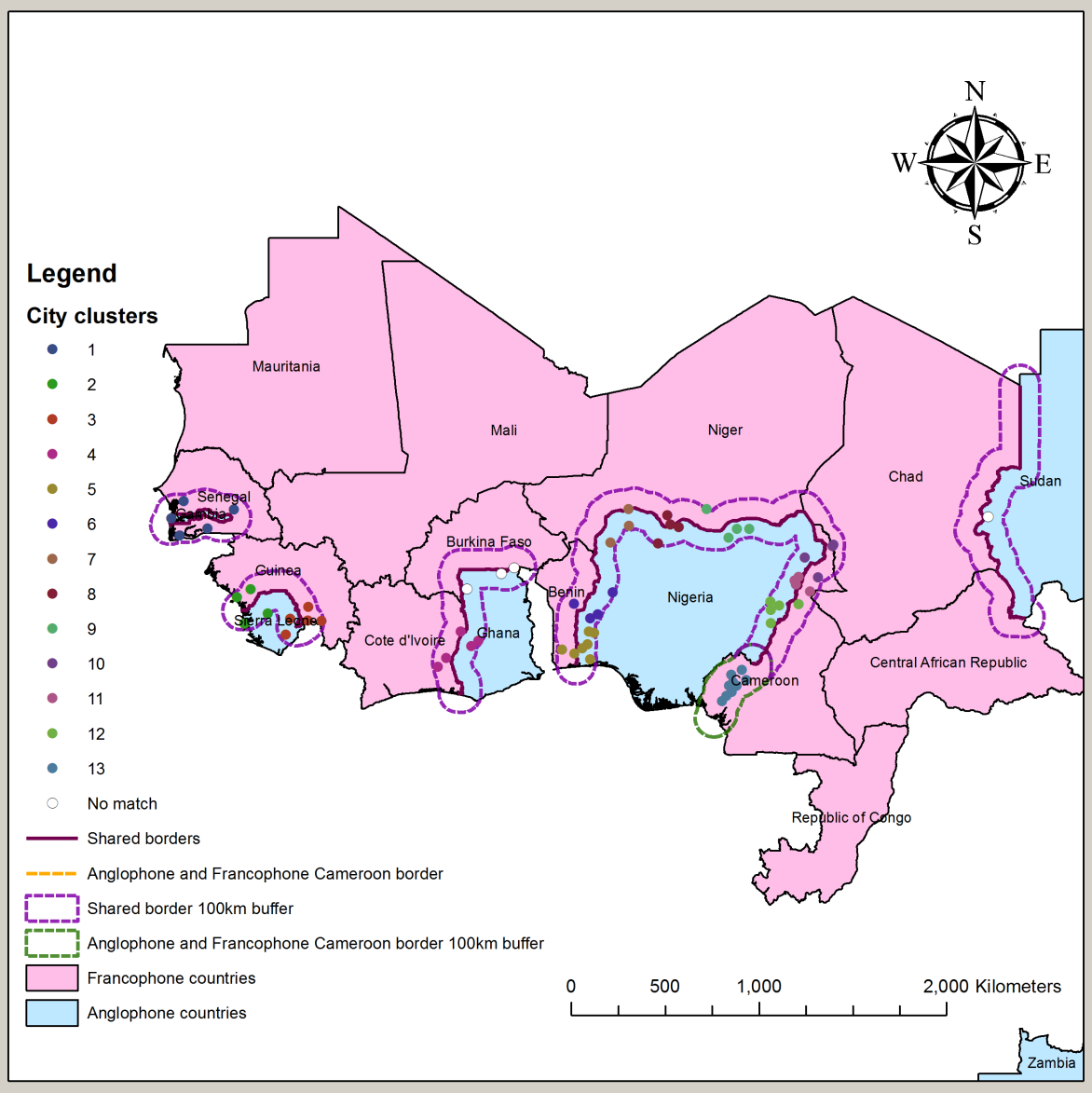


Figure 3.7: Shared borders

Table 3.1: Sprawl: Openness and area

	Openness			Area		
	(1)	(2)	(3)	(4)	(5)	(6)
Anglophone dummy	0.229***	0.172***	0.173***	0.355**	0.282**	0.285***
	(0.045)	(0.044)	(0.055)	(0.151)	(0.116)	(0.093)
Ln annual population			-0.522			8.453***
growth 90 to 12			(1.217)			(2.241)
Ln projected city			-0.174***			0.857***
population 1990			(0.025)			(0.042)
Geographic and Situational Controls	No	Yes	Yes	No	Yes	Yes
Economic Controls	No	No	Yes	No	No	Yes
R^2	0.077	0.212	0.341	0.014	0.492	0.786
N	307	307	281	307	307	281
Oaxaca decomposition						
Explained		0.063*	0.119*		0.052	0.228
		(0.035)	(0.061)		(0.131)	(0.162)
Unexplained		0.166***	0.142*		0.303**	0.267**
		(0.056)	(0.073)		(0.145)	(0.104)

Note: Dependant variable is ln openness in the year 2014 in columns 1-3, and ln area in columns 4-6. Geographic and situational controls include ln ruggedness, ln rainfall, ln elevation range, coast dummy, interaction of ln coast length with coast dummy, interaction of ln distance to coast with coast dummy, malaria index, land suitability index, ln temperature, ethnic fractionalization index at level 15 in year 1975, distance to national capital in the year 1990 and non-national capital dummy. Economic controls include ln annual population growth from 1990 to 2012, ln projected city population in 1990, lag $t-1$ ln country GDP per capita. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 3.2: Intensity by rings in 2014

	(1) 1km	(2) 2km	(3) 3km	(4) 4km	(5) 5km	(6) 6km	(7) Intensity Gradient
Anglophone Dummy	-0.202	-0.456***	-0.615***	-0.636**	-0.212	-0.136	-0.726***
	(0.136)	(0.125)	(0.180)	(0.269)	(0.321)	(0.325)	(0.204)
Ln ring total pixel	1.321***	1.569***	0.848***	0.750***	0.953***	0.777***	0.628***
	(0.354)	(0.430)	(0.180)	(0.153)	(0.144)	(0.167)	(0.056)
Ring distance							-0.287***
							(0.033)
Ring distance × Anglophone							0.093***
							(0.034)
Anglophone mean	1.734	3.865	4.170	3.916	3.720	3.874	
Francophone mean	1.889	4.279	4.409	3.800	3.403	3.409	
Geographic and Situational Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Economic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.181	0.348	0.419	0.447	0.556	0.536	0.481
N	284	286	283	273	251	216	3178
Oaxaca decomposition							
Explained	0.070	0.143	0.205	0.384	0.486	0.432	
	(0.144)	(0.177)	(0.227)	(0.323)	(0.372)	(0.422)	
Unexplained	-0.185	-0.320*	-0.446**	-0.534*	-0.177	0.177	
	(0.161)	(0.182)	(0.225)	(0.318)	(0.343)	(0.407)	

Note: Dependant variable is ln built-up area in 2014 in a ring. First 6 columns stratify sample by rings. Column 7 show intensity gradient for rings up to 20km. City characteristics control variables include ln projected city population in 1990, ln country GDP per capita in 1990, and city population growth from 1990 to 2014. Geography controls include ln ruggedness, ln rainfall, ln elevation range, coast dummy, interaction of ln coast length with coast dummy, interaction of ln distance to coast with coast dummy, malaria index, land suitability index, ln temperature, ethnic fractionalization index at level 15 in year 1975, distance to national capital in the year 1990 and non-national capital dummy. Economic controls include ln annual population growth from 1990 to 2012, ln projected city population in 1990, lag $t-1$ ln country GDP per capita. Anglophone mean and Francophone mean report mean built-up area in both groups in square kilometers. Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 3.3: Leapfrogging

	Ln count of LF			Ln LF minus ln total patches			Ln avg. LF area
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Anglophone dummy	0.907***	0.685***	0.539***	0.339***	0.292***	0.325***	-0.013
	(0.139)	(0.138)	(0.167)	(0.099)	(0.101)	(0.118)	(0.061)
Ln initial cover 1990	0.656***	0.533***	0.295***	-0.081**	-0.151***	-0.295***	0.022
	(0.049)	(0.049)	(0.058)	(0.033)	(0.034)	(0.040)	(0.020)
Year dummy 2014	0.517***	0.505***	0.495***	0.119**	0.117**	0.119**	0.137***
	(0.065)	(0.065)	(0.067)	(0.053)	(0.054)	(0.056)	(0.035)
Ln annual population growth 90 to 12			12.114***			5.842**	2.215*
			(3.338)			(2.585)	(1.287)
Ln projected city population 1990			0.712***			0.443***	0.036
			(0.101)			(0.072)	(0.035)
Geographic and Situational Controls	No	Yes	Yes	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	No	No	Yes	Yes
R^2	0.446	0.536	0.602	0.053	0.131	0.230	0.109
N	606	606	551	606	606	551	525
Oaxaca decomposition							
Explained	-0.181	0.079	0.124	0.035	0.105*	0.133	0.063
	(0.113)	(0.159)	(0.201)	(0.023)	(0.060)	(0.102)	(0.043)
Unexplained	0.894***	0.635***	0.613***	0.328***	0.258**	0.314**	-0.022
	(0.139)	(0.150)	(0.183)	(0.099)	(0.104)	(0.129)	(0.061)

Note: Geography controls include ln ruggedness, ln rainfall, ln elevation range, coast dummy, interaction of ln coast length with coast dummy, interaction of ln distance to coast with coast dummy, malaria index, land suitability index, ln temperature, ethnic fractionalization index at level 15 in year 1975, distance to national capital in the year 1990 and non-national capital dummy. Economic controls include ln annual population growth from 1990 to 2012, ln projected city population in 1990, lag $t-1$ ln country GDP per capita. Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 3.4: Identification based on border sample

	(1) Ln count of LF	(2) Ln LF minus ln total patches	(3) Ln avg. LF area
Basic controls			
Anglophone dummy	1.033**	0.820***	-0.131
	(0.432)	(0.302)	(0.169)
City cluster FE's			
Anglophone dummy	0.796**	0.529***	-0.200
	(0.306)	(0.184)	(0.131)
R^2	0.669	0.425	0.264
N	108	108	103

Note: Controls include ln initial cover 1990, year dummy 2014, lag t-1 ln country GDP per capita, ln annual population growth 90 to 12, ln projected city population in 1990, ln average ruggedness, ln rainfall and coast dummy. Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 3.5: Leapfrogging: Robustness

	(1) Base	(2) No single pixel patches	(3) 60 meter buffer	(4) Light 5	(5) Distance trim	(6) No German colonies	(7) No Nigeria	(8) 40 cities	(9) Non capital
Ln count of LF									
Anglophone	0.539***	0.503***	0.414***	0.449***	0.390**	0.486**	0.579***	0.646***	0.504***
dummy	(0.167)	(0.166)	(0.139)	(0.169)	(0.163)	(0.207)	(0.215)	(0.229)	(0.184)
Ln LF minus ln total patches									
Anglophone	0.325***	0.292**	0.201***	0.288**	0.256**	0.185	0.300**	0.292	0.370***
dummy	(0.118)	(0.122)	(0.068)	(0.118)	(0.117)	(0.151)	(0.150)	(0.218)	(0.131)
Ln avg. LF area									
Anglophone	-0.013	-0.031	-0.067	-0.028	-0.027	-0.154**	0.007	0.136	-0.006
dummy	(0.061)	(0.053)	(0.045)	(0.062)	(0.062)	(0.078)	(0.067)	(0.113)	(0.066)
R^2	0.602	0.604	0.700	0.605	0.573	0.609	0.615	0.596	0.532
N	551	551	551	551	529	489	330	49	518
Oaxaca decomposition (for count of LF patches)									
Explained	0.124	0.190	0.113	0.278	0.111	0.078	0.343	-0.846**	0.355*
	(0.201)	(0.204)	(0.177)	(0.201)	(0.208)	(0.238)	(0.262)	(0.402)	(0.183)
Unexplained	0.613***	0.522***	0.493***	0.450**	0.479**	0.591***	0.587**	0.886**	0.556***
	(0.183)	(0.178)	(0.145)	(0.185)	(0.187)	(0.222)	(0.248)	(0.354)	(0.203)

Note: Columns 1-7 and 9 include same controls as columns 3, 6 and 7 in Table 3.3. Columns 8 include same controls as in Table 3.4. Standard errors are clustered at city level in columns 1-7 and 9, and robust standard errors are applied in column 8. Adjusted R^2 and N are reported for ln count of LF.

Table 3.6: Colonial origin and pre-colonial institutions

	(1) Colonial origin 1800	(2) Non-colonial origin 1800	(3) Pre-colonial instutions
Ln count of LF			
Anglophone dummy	0.584**	0.677***	
	(0.265)	(0.181)	
Pre-colonial institutions index: Anglophone sample			0.269***
			(0.099)
Pre-colonial institutions index: Francophone sample			0.252*
			(0.128)
Ln LF minus ln total patches			
Anglophone dummy	0.091	0.279**	
	(0.160)	(0.134)	
Pre-colonial institutions index: Anglophone sample			0.152**
			(0.071)
Pre-colonial institutions index: Francophone sample			-0.062
			(0.093)
R^2	0.561	0.470	
N	141	465	

Note: Adjusted R^2 and N are reported for ln count of LF for the first two columns. Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Chapter 4

Valuing the Environmental Benefits of Canals Using House Prices

4.1 Introduction

The Britain has an extensive canal and navigable river network, which played a vital role in transporting goods from the Industrial Revolution through the 18th, 19th and early part of the 20th Century. The use of canals and waterways for transporting freight had all but disappeared by the mid-20th Century, and many had fallen into disrepair or been abandoned. Since then, the canal and waterway network has been restored and developed into a valuable environmental and recreational amenity, providing the venue for an extensive range of tourism and leisure activities. These canals[1] also provide transport corridors for walkers and cyclists along the towpaths formerly used by horses for drawing boats. It is estimated that in 2016 there were more than 4.3 million individuals, making a total of 396 million visits to the canals for various purposes including walks, hikes, boating, fishing, and cycling (Canal and River Trust, 2016).

This paper applies revealed preference, ”hedonic” property value methods to estimate the value of canals and waterways in England and Wales. Standard cross-sectional regression methods are supplemented with a ”difference-in-differences” estimation method that looks at the changes in property prices occurring around the time of a canal restoration project in the mid 2000s. How much money people spend on housing, and hence what they sacrifice in terms of the value of other forms of enjoyable consumption, in order to benefit from living near it, reflects the willingness to pay for the amenity of canals. Assuming people are free to move and are well-informed, they will end up living in places where the benefits to them

[1]We use the term “canals” to refer to waterways that were dug out where there was no previous waterway, and rivers that were canalised to make them navigable.

of doing so are at least equal to the costs - otherwise they would move somewhere else. The market price of houses with similar size and structural characteristics but in different places, adjusts to trace out the value of those places to the population. In turn, the value of a place can be unpacked into its constituent components – proximity to transport, proximity to jobs, crime, quality of schooling, quality of environment, recreational facilities and so on. The price premium for each component measures the marginal willingness to pay for an amenity. The level of exposure to an amenity of a place, such as school quality in the catchment area or distance to closest train station, is measured by the distance from the amenity. Hence, we measure the access to canals based on the distance of a property from its nearest canal and estimate how housing values change with distance, holding all other differences constant. Living near a canal not only reduces the travel cost to waterways, but also allow households to enjoy a canal frontage[2].

There are, however, potentially many confounding factors that vary with distance to a canal and affect the price directly. These include the physical characteristics of the housing, amenities like distance to employment or proximity to public transportation nodes. For example, canals in urban areas are usually found in old industrial areas, and properties near these areas are typically older and smaller. Furthermore, industrial buildings could be a dis-amenity to residents and all these factors could affect housing values but are not related to canals. We adopt several strategies to avoid this type of bias. First, we control for a rich list of observable housing and location characteristics in our analysis to reduce observable differences between properties. We further partial out time-invariant unobservables with the inclusion of fixed effects at a small geographical scale – either Middle Layer Super Output Areas (MSOAs) or Lower Layer Super Output Areas (LSOAs) – and for differing price trends at Local Authority Districts (LADs) level. We also exploit the localized environmental benefits associated with waterways and constrain our analysis to properties within 1500m from the canal to mitigate unobserved neighbourhood differences. Confounding factors that vary at a higher geographical level between LSOAs/MSOAs – such as access to labour markets – are eliminated. Our sample, which covers almost more than 2 million property transactions in England and Wales, is sizable to ensure precise estimation on the coefficients of the model. The main issue is that it is impossible for researchers to observe and control for all the differences between properties.

Our second strategy is difference-in-differences approach that exploits the variation in the amenity value from canal restoration overtime. We focus on the restoration of an abandoned canal – the Droitwich Canal in the West Midlands of England, which was closed since 1939. By the early 2000s, bulk of the canal was overgrown, drained of water, non-navigable

[2]The values that can be elicited through house prices are therefore what environmental economists refer to as use value, as opposed to, say, the satisfaction one might get from just knowing that such a resource exists without ever intending to visit or use it.

or completely destroyed. The Canal underwent a major restoration in 2007 and was re-opened in 2011. The restoration provides an avenue for recreation activities such as boat navigation, improved the environment and provided a habitat for aquatic life. In our study, we compare price changes for properties close to the canal after the canal is restored (treatment group) with price changes for comparable properties that are unaffected by canal restoration (control group). This strategy hinges on the fact that prices trends in the treatment group are similar to price trends in the control group if restoration had not taken place. This depends on the comparability between properties in the treatment and control group. Hence, we select (1) properties close but slightly further away from the Droitwich canal, and (2) properties near to an existing neighbouring canal – the Worcester and Birmingham canal – that remain in continuous use and are not affected by the restoration over this period. These comparisons allow us to estimate the value of the restoration and the enhanced recreational and environmental amenities it provides, in so far as this value shows up in different price changes in the treatment and control groups.

Our findings from examining the entire canal system in England and Wales suggest that proximity to canals increases house prices, although the effect appears highly localised. Houses within 100 metres of a canal have a price premium of around 5% relative to those beyond 1km (estimated on the whole 2002-2016 study period) falling to 3.4% in 2016. There is no impact on prices in the 100m-1km range. The localized effect suggests it might be associated with canal-side properties and others which have immediate access or views of these waterways. The effect is bigger – around 10% - in dense urban areas. We also investigate the association between canal proximity, and the share of new-build homes sold, as a proxy for housing construction. We observe that the proportion of new-build sales is a 6 percent higher within 100m of a canal compared to further away, representing a 75% relative increase. Similar effects are detected for the re-opening of the Droitwich canals. The restoration leads to a 10% increase in values for properties within 100m of the restored canals. Back of the envelope calculations indicate the land value uplift from the canal network was around £0.8-£0.9 billion in 2016.

The remainder of the paper is as follows: Next, we outline the existing evidence on the effect of waterways on housing prices and explain how our paper contributes to the existing literature. Section 4.3 describes the data and the methodology in detail. Section 4.4 presents and discusses the results of the analysis and Section 4.5 concludes.

4.2 Existing Evidence on Waterways and House Prices

In this section, we review the existing literature measuring the economic value of canals. We highlight the limitations associated with the existing papers, before recommending some strategies to improve the estimation.

Previous studies estimating the economic value of waterbodies examine the impact of a wide range of features that include seaside, rivers, streams, lakes and canals on home prices. For a comprehensive overview of the existing research on the capitalization of different inland waterways on home prices, refer to Nicholls & Crompton (2017). Overall, most studies show that home-owners pay more to reside near canals. Garrod & Willis (1994) documented that properties in London that are located along canals are sold at a premium of 2.9%, while properties further away but within 200 metres are sold at a premium of 1.5%. Extending the analysis to a sample of property sales in Milan, Italy, Bonetti et al. (2016) examine the difference in the willingness to pay for artificial (canals) and natural waterways (streams). They report that every metre away from canals reduces home prices by 0.074%. Conversely, they do not find a house price premium for streams. Examining the effects of canals on housing values in Texas, US, Nelson et al. (2005) finds that homes with a canal frontage are sold at a premium of 11%, around $16,298. Although the literature consistently reports that households pay a premium to reside near canals, the magnitude of the estimates varies across studies, suggesting that the estimates are susceptible to the econometric framework adopted.

Another stream of literature related to this paper is on how waterways restoration affects home prices. Streiner & Loomis (1995) estimate the value of stream restoration projects using a sample of property sales from 1983 to 1993 in California, USA. These projects, conducted by the Department of Water Resources, reduce damages from flooding, improve bank stability and restore aesthetic and environmental value of streams. Breaking down the effects of different restoration projects, they report that flood prevention restoration increases home values by 5%, at around $7,804, while stabilizing streams improves housing prices by $4,488. The authors conclude that the heightened property taxes from the increased home values outweigh the cost of these restoration projects. Mooney & Eisgruber (2001) investigate the impact of riparian buffers on housing values for a sample of 705 property sales in Oregon, USA. Although these buffers provide a more conducive habitat for aquatic species by reducing stream temperature, they obstruct the views of the river. As a result, home prices fell after these buffers are erected. From these results, it is evident that homeowners value both the scenic views and the stability of waterways.

Notable limitations are observed in these studies. Most of these studies are conducted on a small sample of sales in a specific city. In this paper, we draw inferences from a larger sample of sales of more than 2 million transactions across England and Wales from 2002 to 2017. Furthermore, most of these papers exploit the cross-sectional variation in housing prices and compare property prices close to canals with those further away to recover the price premium paid to reside near canals. The main issue is that there could be unobserved confounding factors that correlate with the proximity from canals, affect home prices and bias the estimates. In this paper, other than improving the traditional hedonic framework by adding a rich set of covariates and micro-geographic fixed effects, we further exploit the

natural experiment of the restoration of the Droitwich Canal and compare home prices before and after the restoration to estimate how much homeowners value canals.

4.3 Methods and Data

4.3.1 Estimation Methods

4.3.1.1 Regression Specifications for National Analysis

We estimate the house price premium associated with proximity to canals with a standard hedonic property price regression estimated using ordinary least squares (OLS) that takes the following form:

$$lnp_{ijkt} = \alpha_j + \sum_{k \in K} \beta_k D_{ik} + X'_{it}\theta + \tau_t + \varepsilon_{ijkt} \tag{4.1}$$

where lnp_{ijkt}, the dependent variable, is the natural logarithm of the price of property i located in neighbourhood j and sold at time t. The key variable of interest is D_{ik}. It is a set of distance band indicators at 100 metres interval and up to 1km ($K = \{1, 2, ..., 10\}$) computed base on the euclidean distance of property i from the nearest canal. D_{i1} is a binary variable that takes the value of 1 if a property is between 0-100m from a canal, and 0 otherwise, and D_{i2} flags out properties that are between 100-200m from a canal and so on, up to 900-1000m. The key parameters of interest β_1, β_2,...β_{10}, capture the average percentage difference in transacted prices for properties in given distance band relative to the baseline (properties beyond 1000m from canals). We expect to observe the price premium decreasing with k as the distance and travel cost to the canal increasing and the benefits from canals decaying. Figure 4.1 maps the canals and canalised rivers used in this analysis.

We further control for a rich set of time-varying housing and neighbourhood characteristics denoted by X_{it} to mitigate the risk of unobserved confounding factors from biasing our estimates. For more information on the list of controls included in our analysis, refer to Table D.1. We also control for neighborhood fixed effects that partials out time-invariant unobservables at neighborhood j as denoted by α_j. Depending on specifications, we could define j as MSOA or LSOA. τ_t denotes year-quarter fixed effects that control for general property price trends across areas over time. The assumption is that, conditional on controls, $E[\varepsilon_{ijkt}|D_{ik}] = 0$.

In reality, this assumption is likely to be violated if there are unobserved confounding factors that are correlated with distance from canals (D_{ik}) and affect house prices. To minimize unobserved neighbourhood differences between properties, we further restrict our analysis to properties within 1500 meters from the canals. The notion is to compare properties in the same neighbourhood but enjoy different exposure to the environmental benefits with varying

distances from canals. This is possible since the amenity values of canals, such as waterfront view, are highly localized. In addition, we control for the interaction of Local Area District (LAD) Fixed Effects with year-quarter fixed effects to account for differential property price time trends across space. Put differently, we are comparing property sales within the same LAD in a given year-quarter but with varying proximity from canals.

Demand to live near canals could also affect developers actions. If households are attracted to reside near canals and are willing to pay more for properties near canals, developers could be enticed to redevelop older developments for larger profit margins. Hence, we examine whether properties closer to canals are more likely to be new builds. To implement this analysis, we simply replace the dependent variable in equation 4.1 with an indicator of whether a sale is a new build.

4.3.1.2 Difference-In-Differences Estimation

Even with the inclusion of a robust set of controls and geographic fixed effects, and the restriction of property sales around canals, researchers could miss out on unobserved confounding factors when estimating these cross-sectional hedonic regressions. Therefore, we adopt a difference-in-differences design to examine how much households are willing to pay (WTP) to reside near canals. The advantage of this approach is that it allows us partial out confounding factors correlated with the location of canals by exploiting both the spatial and temporal variation in the environmental benefits of canals. This method is frequently applied in valuing amenities/disamenities such as proximity to transportation nodes (Gibbons & Machin, 2005), wind farms (Gibbons et al., 2015), exposure to air pollution (Currie et al., 2015), crime risk (Linden & Rockoff, 2008) and traffic (Tang et al., 2016).

Self-evidently, the limiting factor in applying this approach to the evaluation of the environmental benefits of canals is that, in general, accessibility to canals and their environmental benefits is rarely changing. One exception is where there have been substantial canal restoration projects, bringing disused, buried and derelict canals back into use as functioning recreational waterways. Canal restoration projects have occurred throughout Britain over many decades, often carried out by volunteers, but only one significant project lines up with the time period of our data on housing transactions – the restoration of the Droitwich Canal in the West Midlands in the late 2000s.

The Droitwich Canal is a canal formed from two canals – the Droitwich Barge Canal and the Droitwich Junction Canal – linking the River Severn and the Worcester and Birmingham canal, and passing through the centre of Droitwich (formally Droitwich Spa), a town of 25,500 people in the county of Worcestershire. The canals were abandoned in 1939 after an Act of Parliament and fell into decline. Parts of the canals had been restored on a

voluntary basis, organised by the Droitwich Canals Trust, formed for this purpose in 1973. As a result, a section of the canal in the centre of Droitwich and three locks at the eastern end had been restored by the mid-2000s. Full restoration began in 2007, a major project with a cost of £11 million funded by National Lottery grants, local councils and charitable donations. All of the canals required dredging, repair of locks and other structures. The most significant works were complete reconstruction of a section by canalising 550 metres of the River Salwarpe through Droitwich, a new tunnel under a main road to link the Barge Canal to the River Severn, improvement to a bridge on the M5 motorway, a complete new cut with four new locks, plus extensive environmental mitigations and enhancements. The project was coordinated by British Waterways, the public corporation that managed canals and waterways at that time and was scheduled to start in 2007, with planning applications were submitted in May 2007. The work was due to be completed by 2009, although the canals were not fully restored and opened for navigation until July 2011. The history can be traced through various web sources[3]. The non-technical summary of the project published by Waterways (2010) describes the purpose of the project thus:

"This project will bring the canals into navigable use and will create a unique 21-mile cruising ring linking Droitwich Spa to Worcester, which can be completed in a weekend by boat. The project is not solely about navigation as it includes many works to enhance the canal corridors as a recreational and environmental resource for local people as well as visitors to the area. Canal restoration will provide a stimulus to the local economy by encouraging tourism-related businesses and will provide many benefits to the local community. It is intended that the vision will be delivered through a series of objectives including: To restore the canals to good navigable condition; To use the canals as a catalyst to stimulate sustainable regeneration in Droitwich Spa and the surrounding area; To create an environment in which a visit to the waterways is an educational and interpretative experience of the canals' history and environment; To conserve, enhance & promote the built heritage & environmental assets of the canal; To achieve high levels of public accessibility for all; To sustain harmony between environmental, heritage & recreational uses." (British Waterways 2010).

The project was evidently very ambitious in its environmental and recreational aims, and so potentially provides a useful experiment for estimating the value of these benefits to local homeowners. To implement this idea in a difference-in-differences design we need to define treatment and control groups. The control group needs to be carefully chosen such that it is likely to have followed the same "counterfactual" trends in outcomes as the treatment group would have done in the absence of the policy. This is also known as the parallel trend assumption. In particular, we select properties that (1) are slightly further away that

[3]For more information, one can refer to http://www.droitwichcanals.co.uk, https://www.waterways.org.uk/waterways/history/historic_campaigns/droitwich_canals/droitwich_canals and https://en.wikipedia.org/wiki/Droitwich_Canal

are between 1000 and 1500 meters from the Droitwich Canal and (2) properties along the Worcester and Birmingham Canal. Figure 4.2 presents a map of the Droitwich Canal and Worcester and Birmingham Canal overlaid on a satellite photograph, making the general layout and similarity in the landscape crossed by each canal clear. The latter is selected because they are also close to a canal, and share a similar geographical landscape being part of the same general local economy.

Another key element in our setup is the definition of the "treatment" date when the benefits from the restoration of the canals to start to materialise – which we refer to as the post-restoration date. The project is extended over a number of years from the mid-2000s and there was some restoration activity well before that. There are two plausible choices of this post-restoration date in relation to the major restoration scheme that started in 2007. One date is the submission of planning applications around May 2007. A second is the completion and opening around September 2011. We explore the impacts using one or the other, or both of these dates by estimating the following regression specification:

$$lnp_{ikpt} = \sum_{k \in K} \beta_k D_{ik} + \sum_{k \in K} \delta_k D_{ik} Post_t + \sum_{k \in K} \eta_k C_{ik} + \sum_{k \in K} \gamma_k C_{ik} Post_t + X'_{it}\theta + \iota_p + \tau_t + \varepsilon_{ikpt} \tag{4.2}$$

where C_{ik} represents distance to canal indicators as for equation 4.1, where the distance is to either the Droitwich or Worcester and Birmingham Canals. D_{ik} represent equivalent indicators for the Droitwich Canal and are the main “treatment” variables. $Post_t$ represents the period after the treatment. The variable is an indicator that the observed property sale is occurring in the post-intervention period. The key parameters δ_1, δ_2,...δ_{10} measure difference between the price change occurring within each distance band close to the Droitwich Canal, and the price change occurring in the same distance band close to the Worcester and Birmingham Canal. The price changes are estimated before and after restoration. In addition, the price change occurring in each distance band is estimated in relation to the price change occurring in properties between 1000m and 1500m (reference group). In other words, the parameters δ_1, δ_2,...,δ_{10} estimate how house price premium decays with distance from the Droitwich Canal, as compared to the way the price changes decay with distance from the Worcester and Birmingham Canal.

The main advantage of the DID estimation is that it permits the inclusion of postcode fixed effects (ι_p). In other words, we are exploiting the changes in house prices and access to canals within postcode p over time. This is possible because the restoration of canals generates variation in the access of canals for a particular property over time and negates the risk of unobserved factors correlated with distance from canals from biasing our key estimates. The rest of the variables are similar to equation 4.1.

4.3.2 Data Sources

The main source of data for the analysis set out above is the Land Registry "price-paid" dataset that provides detailed information on transaction prices and some basic characteristics. This dataset has been linked to information from Energy Performance Certificates (EPC), which are required for all properties bought and sold in England and Wales[4]. The EPC data provides a much richer description of the structure of the property. Although the EPC information only dates back to 2008, the information can be used for properties with EPCs, when they were sold in earlier periods (assuming the basic structure of the property has not changed). Given this limitation, we do not go back beyond 2002, although the price-paid data extends back to 1995. Our full dataset covers more than 11 million property transactions from 2002 to 2017, falling to around 2 million when we restrict to 1500m buffers around canals.

For each property, we observe the postcode, floor area, number of rooms, number of heated rooms, energy efficiency, house type (flat, semi-detached, terrace house) and whether the property is new build and has a fireplace. Other characteristics are available in the EPC data, but much of this is incomplete. We geographically locate each property based on its full postcode – which typically corresponds to around 17 houses. Although the coordinates are accurate to 1m for the postcode centroid, there is a degree of approximation in terms of the exact location of a property due to the potential size of each postcode, particularly in sparse rural areas. Geo-referenced information of the 371 canals across England and Wales comes from the Canal & River Trust. The total length of canals spans across 3,530 km[5].

Using geographical information system software (ArcGIS), we compute the straight line distance between each property postcode and its nearest canal. This is the main variable of interest in this study. We further measure the proximity of each postcode from features of canals (also provided by Canal & River Trust) that could affect home prices through channels other than the environmental and local recreational benefits. These features include bridges (benefits as crossing points), docks and wharves (industrial areas), embankments, lakes, overflow outfall and reservoirs (signifying possible flood risk). From Ordnance Survey Strategi data Survey (2015), we also compute the distance between each postcode from the nearest train lines and stations, as we are concerned that properties closer to canals could be more or less accessible to these transportation modes, given that railroad and canals often follow the same transport corridors. Distance to rivers and distance to green space is taken from the OS Open Rivers and Open Greenspace datasets (Survey, 2018a,b). Land use comes from Landcover map Landsat remote sensed data (Rowland, 2017), each postcode assigned the land use at its centroid, and categories aggregated up to 9 major groups, urban, suburban,

[4]This data linking was done for another project by colleagues at LSE.

[5]For more details, refer to https://data.gov.uk/dataset/660ab8be-2912-4ef5-a8a9-7ed3111e34d1/canal-centre-line

and a rural land cover types.

Using the location of each sale, we further map each postcode to Census data units, the Middle Layer Super Output Areas (MSOA), Lower Super Output Areas (LSOA) and Output Areas (OA). There are around 180,000 OAs and 35,000 LSOAs and 7,200 MSOAs across England and Wales. OAs are the smallest geographical area in which Census data from the Office of National Statistics is collected at every decade. There are in total two waves of Census collected in 2001 and 2011 over the sample period, though we use data from the 2001 Census only. To control for neighbourhood differences between properties, we account for a wide array of characteristics, specifically unemployment rate, proportions owning cars, social renting, home-owning, with no education, ethnic minority residents, non-EU residents, share of lone-parent households, population and population density, all at OA level. The LSOA codes are also used to merge in employment data and employment industry sector shares at LSOA-level. These data come from the Business Register and Employment Survey supplied via the Nomis UK data service (www.nomisweb.co.uk). The earliest comprehensive data readily available at a small area level is 2015 and we only use this year of data (matched to all years of transaction data). The data sources are set out in Table D.1.

4.4 Results

4.4.1 Descriptive Statistics

Our main estimation sample contains 2,048,723 transactions from 159,788 postcodes, 6,979 LSOAs, 1,861 MSOAs and 160 Local Authority Districts. The means and standard deviations of the variables in our main estimation dataset of transactions are summarised in Table D.2. Since our analysis compares house prices in places close to canals with prices in places further away, the table splits the information into three groups 0-100m from a canal, between 100 and 1000 metres of a canal, and between 1000 and 1500m of a canal. We do not report the figures for the full set of distance variables, but report those for rail, town centres and rivers. A key thing to note is that there are differences in the characteristics of properties sold close to canals and those further away on many dimensions, but on others the areas seem quite similar and it is hard to observe systematic patterns.

Evidently, simply looking at mean prices is not very informative. On average, in these unadjusted figures, property prices are slightly higher in the 100m zone than the 100-1000m zone, but both of these zones are slightly cheaper on average than the zone beyond 1500m. The estimated gap between prices in the 100m zone and the 100-1000m depends on how it is measured, around 1% in the simple means, around 5% when based on the average differences in log prices (0.05), and around 10% when looking at price per square metre. At the same time, properties within 100m of canals are smaller, more likely to be new builds, and much

more likely to be flats (37% as compared to 16.5% elsewhere). Population density is lower, there are more social renters and more unqualified people in OAs within 100m of canals, but otherwise the demographic characteristics look similar across all the groups. Canals tend to follow paths of least resistance and natural lines of communication, so properties close to canals tend to be close to railways, rail stations, close to other rivers and closer to town centres. Given the canals' original purpose for transporting goods, it is not surprising to find that there is more employment on average in MSOAs close to canals, slightly more heavily represented by manufacturing, mining/utilities, accommodation/food, and business administration, and less represented by health and education services. Interestingly, residential properties within 100m of canals are 52% urban and 45% suburban, whereas the rest of the sample is split 65-69% suburban, 28-32% urban. This presumably reflects that if a canal passes through a town, it typically passes through its centre, again because of their historical transportation role. Only a small proportion of properties within 1500m of a canal are in places with non-urban/suburban of land cover. It is important to correct for all these structural and geographical differences when comparing prices in the various distance zones, and the results from the regression analysis we use to do this are reported in Section 4.4.2 below.

The sample for the analysis of the Droitwich Canal restoration is much smaller, as it is restricted to properties within 1500m of either the Droitwich or Worcester and Birmingham Canals. A selected set of descriptive statistics for this group are reported in Table D.3. Here, we report means and standard deviations for the three distance groups related to the Droitwich Canal (<100m, 100-1000m and 1000-1500m) and for the overall sample for the Worcester and Birmingham control group (<1500m from the Worcester and Birmingham canal). Again there are dissimilarities along some dimensions when we compare these groups. However, the patterns are different from those in the full England and Wales sample and even less systematic. Properties 100m from the Droitwich canal are marginally smaller than those 100m-1000m away, and considerably smaller than those near the Worcester and Birmingham canal. There is a higher proportion of terraced houses close to the Droitwich Canal than elsewhere and more social renters. In general, statistical tests of the difference between these groups indicate that only a few of the differences are statistically significant. The simple mean price differences are not revealing of any strong patterns. The results of the difference-in-differences analysis using these data are presented below in Section 4.4.3.

4.4.2 Regression Estimates for National Analysis

The results from the regression analysis discussed in 4.3.1.1 are presented in Table 4.1. Column 1 of Table 4.1 shows the results with no control variables, other than a set of LAD-year-quarter indicator variables (to capture general variation between LADs and over time), and basic house structure variables, house type (detached, semi, terraced, flat), new/old, leasehold/freehold, floor area, number of rooms, heated rooms, fireplace, energy performance

rating (a 10-point scale). Column 2 retains these control variables, but adds in controls for geographical location, specifically the distances to various features, predominant land cover, employment, and a set of MSOA fixed effects to eliminate price variation between MSOAs, as discussed in Section 4.3. Column 3 replaces MSOA with LSOA fixed effects (the employment variables are now excluded as these do not vary within LSOA). Column 4 includes additional controls for neighbourhood (OA) demographics. A full set of regression coefficients and standard errors for an example specification is provided in the Appendix, Table D.2.

The striking feature of the table is the 3-5% price premium for properties within 100m of canals. Beyond this distance threshold, the effects in column 1 become slightly negative before becoming near zero and insignificant at around 600m. This pattern of negative effects between 200 and 600m is evidently related to confounding factors near canals because, when we control for geographical factors in the remaining columns, these effects disappear. Likely explanations are, as discussed earlier, that canals often pass through industrial areas in towns, and these areas are likely to be less attractive to residents. The difference between the first and remaining columns illustrates the importance of carefully controlling for these kinds of geographical influences. In column 3 and 4, when we control for LSOA fixed effects and neighbourhood demographics – including education, ethnicity, and unemployment – it is likely we are over-controlling, and that the estimated price premium is an underestimate. The estimates are also less precisely measured (wider confidence intervals). The reasons for this are firstly that LSOAs are relatively small spatial units, so within each LSOA there is relatively little variation in distance to canals, particularly in dense locations. Also the problem with controlling for demographic characteristics is that these will respond to the local housing price, because people chose where to live based on the housing costs. Poorer, less educated and ethnic minorities tend to live in lower cost places. This implies that including controls for these demographics may eliminate some of the price effects we intend to estimate. We therefore regard column 3 and 4 as robustness checks, and our preferred estimate is that in column 2.

How should we interpret the key result of Table 4.1 a 5% premium for living within 100m of a canal? The result implies that people are willing to pay up to this amount to live within 100m of a canal, relative to what they are prepared to pay to live elsewhere. The short distance range of this effect suggests that the value is primarily associated with canal-side properties and others with immediate access or views of the canals. There is no premium for living near a canal other than right up close to it. This lack of a price premium for moderate proximity suggests that residents are not, on average, paying to save the time it takes to walk the additional distance from home that is, say, 1500m rather than 500m away. If canal users are doing so only occasionally, or if their primary motivation is to exercise, this finding is not too surprising. It is worth noting that people likely differ in the value they place on canal-side properties and immediate access to canals. Because properties with this access are scarce, the

values estimated here cannot safely be generalised to the whole population, because residents with the highest willingness to pay are those who end up owning the homes, and it is their willingness to pay which determines the market price. See Bayer et al. (2007) for discussion of these issues. The values should thus be seen as upper bounds to the value of canal-side locations to the average person in the population.

These results do not identify any specific feature of canals that might be attractive. In additional analysis we looked at the effects of specific features – locks, aqueducts, wharves, and canalised rivers – alongside the basic effects of canal proximity. We found no interesting patterns related to aqueducts or wharves, but there is a significant (at 10% level) price premium associated with canal locks, and an insignificant effect of canalised rivers within 100m, of a similar magnitude to that for canals[6]. This pattern for locks is illustrated in Figure 4.3, in which we separate out the basic canal effect (top panel) and the additional lock effect (bottom panel). Note, the distance scale for the locks plot is in 100m, but has a different range, because the nearest lock can be much further than 1500m away, even though the sample restriction means that the nearest canal is within 1500m. There is evidently an additional effect from locks, of around 4.5% within 100m falling to 3% at 200m, although the estimates are only statistically significant at the 10% level. Some of this effect may be driven by former lock keeper's canal-side cottages, but there may be some heritage value associated with locks in general.

In the next analysis in this section, we look at how the price effects from canal proximity vary by type of location. Here we focus only on the effects of being within 0-100 metres, given the lack of any effects elsewhere. Table 4.2, column 1 shows the differences by built-up urban and non-urban locations (using the land cover categories described in section 4.3.2). The first row of column 1 indicates that outside urban areas, the price premium for the 0-100m band is 2.7%. This increases by an additional 7% in urban areas, making the total effect in urban areas around 10%. A plausible explanation for this finding is that canals offer particular environmental and recreational benefits in urban areas, where there is limited green space available, and canal-side locations may be particularly coveted. Urban in this land cover data refers to the densest parts of cities. Column 2 repeats the analysis for suburban and urban areas, which represent over 95% of the sample. Here we can see that all of the basic premium for canal proximity is driven by urban and suburban locations, and the effect in rural places (given by the first row) is insignificantly negative. The implied premium for living within 100m in urban and suburban areas in these estimates is 5.9% (this is slightly higher than in Table 4.1, because here we are comparing 0-100m, with 100-1500m). We also double checked for effects at higher distance bands in the urban/suburban sample, but found none

[6]The coefficients on our control variables indicate that there is also a premium of a similar magnitude for living near other natural rivers that extends over a wider range of distance, but again these are not statistically significant, and not the primary focus of this analysis.

(see the Appendix, Figure D.1). Column 3-4 look at differences by whether a property is close to other rivers or green space which might provide alternative recreational and environmental services, but we find no evidence that this matters in general in the national sample, even if it matters to urban populations as evidenced by column 1.

It is useful to translate the percentage premium on house prices into monetary equivalents, which represent willingness to pay for canal-side amenities – i.e. how much households are willing to give up on other expenditure in order to enjoy homes close to canals. Some care is needed in doing this, as we have estimated an average percentage premium over the whole period, but average house prices have doubled over the period from 2002 to 2017 so it is not necessarily appropriate to apply the percentage uplift to current prices to get the monetary equivalent. Instead, we first estimate the percentage price premium for properties within 0-100m in each year. Figure 4.4 plots these results. We do not report 2017 as our data only spans part of this year. The figures for each year after 2002 need to be added to the figure of 0.081 in 2002 to get the relevant percentage increase in that year. From the graph it is clear that the percentage premium remained stable from 2002 up until 2007. From then on it fell considerably, the obvious explanation being a shift in the housing market following the great recession in 2008. It is well known that the character of the housing market has changed since then, with much lower transaction volumes. Table 4.3 reports the monetary equivalents for each year, obtained by multiplying the percentage canal premium for each year by the mean price in the sample of properties 0-1500m from a canal in each year. The table shows the amounts in nominal terms, converted to 2016 prices using the Consumer Price Index, and the annual equivalents assuming a discount rate of 3.5% (obtained by multiplying the real capitalised value by the discount rate). Evidently, willingness to pay has declined substantially post-recession. Prior to 2008, households were willing to pay around £520 per year to live within 100 metres of a canal (the mean in the 2002-2007 period). From 2008 onwards, this figure has fallen to half that at £260. The average overall is £370 in 2016 prices.

As a final step in this analysis, we look at the availability of new build homes, as explained at the end of Section 4.3.1.1. The results are reported in Table 4.4, which shows the effect of canal distance on the proportion of new build sales. As can be seen from the table, the proportion of new builds is significantly higher closer to canals. Without any control variables, other than MSOA and LAD-year-quarter fixed effects, we find effects of 8.5 percentage points within 100m, falling to zero by around 600m, in column 1. The most likely explanation for this higher proportion, is number of new builds being constructed. Such a relationship will arise because of demand in these places and/or greater supply due to lower construction costs due to the availability of land for building or industrial premises for conversion (e.g. former warehouses). As a basic step to control for factors affecting the supply side, column 2 introduces more control variables for location, land use and employment share (as for Table 4.1, column 2). Doing so reduces the estimates slightly, and – as with the price analysis –

indicates that any effects are constrained to within 100m, where we find a 5.9 percentage point higher proportion of new builds. For the reason mentioned in Table 4.1, we control for LSOA fixed effects and neighbourhood demographics in column 3 and 4 as a robustness check. The effects remain strong.

What does this mean in terms of the number of new homes attributable to the canal? There are around 63,700 unique homes in the 100m buffer sold between 2002 and 2017, and 10,500 of these are newly built over this period. The rate of new building in the area outside the 100m zone is 7.8%, so the additional 5.9% means the rate of new building in the 100m zone is 76% higher. This means that we would expect 0.078 ×63,700 = 4967 new homes in the 100m zone if the new-build rate was the same as elsewhere. Our estimates attribute an additional 0.059 × 63,700 = 3758 homes to the existence of demand for a canal-side location (rather than other features of the land near the canal; the remaining 10,500 – 3758 = 6742 is presumably due to these other factors). Although it is impossible to rule out that this effect is still partly due to the kind of land and existing buildings available, i.e. is driven by the supply side of the market, the combination of more new builds and positive price effects from the previous analysis suggests, fairly unambiguously, that these effects are demand driven.

4.4.3 Difference-In-Differences Estimates

In this section, we report the results of the difference-in-differences analysis of the Droitwich Canal restoration described in Section 4.3.1.2. As discussed in that section, these results relate to the impact that the restoration had on the relationship between canal distance and price in the Droitwich area, compared to a control area near the Worcester and Birmingham Canal. The presentation of the results is otherwise similar to the main results in Section 4.4.2 above. Figure 4.5 summarises our key estimates graphically, with point estimates and 90% confidence intervals. Panel (a) shows the impact of the restoration using a post-intervention date of May 2007, the date the main restoration period began, Panel (b) shows the additional effects – on top of those related to the start of the renovation – occurring around the official opening date in 2011. The impact shown in panel (a) is thus a short run effect from 2007 to 2011. Panel (c) simply reports the effect of the start of renovation in May 2007, without controlling for opening, to give a clearer picture of the overall change before and after this time. Note, the regressions used to derive these estimates control for full postcode fixed effects – i.e. eliminate all fixed over time differences in prices between postcodes – hence do not include the controls for distance to transport and other features or employment, since these do not vary within postcode. We include interactions between neighbourhood (OA) 2001 census demographics and the post-intervention indicator to control for possible spurious price trends related to these characteristics.

The plots in Figure 4.5 (a) and (c) bear a similarity to those from the national estimates in

Table 4.1, although the methods used to estimate them are substantially different. Here we are estimating only from the changes in prices over time near the Droitwich Canal around the time of the start of the major restoration, compared to the changes over time occurring over the same time in the control group. The effect of the restoration within 0-100m is large before opening, at around 15%, although there is a marked decline after opening. Taken together the overall impact reported in panel (c) is around 10%, which is substantially larger than the 5% found on the national cross-sectional analysis in Table 4.1, although given the wider confidence intervals the figures are statistically similar. These patterns of distance decay in these estimates are not so clear cut, with some evidence of price uplift in 400-1000m bands. It is possible that the effects are spuriously related to confounding factors specific to the Droitwich area compared to the control Worcester area. Nevertheless, the sharp distance decay between 0-100m and the rest provides some assurance that the 0-100m effect can be treated as a "causal" impact of the canal restoration on immediately proximate property prices.

The estimates from this difference-in-differences evaluation are less precise than those from the cross-sectional analysis, and are based on a single case study area and much smaller sample. There are risks in looking at a single case like this, in that the estimates may be influenced by local price trends specific to the area. The number of affected properties is small – around 289 sales occur in 36 postcodes within 100m band between 2002 and 2017. It is, however, reassuring that this methodology arrives at results which point in the same direction as the national cross-sectional analysis. The likely interpretation is that households value the environmental amenities associated with living very near to a canal, or alongside the canal, and the Droitwich Canal restoration increased the quality of these amenities as the project intended. A back-of-the-envelope calculation, multiplying the number of unique properties transacted since 2007 within 100m of the Droitwich Canal (176), by the mean price in 2007 in the 0-1500m sample area (£195,000) and the percentage increase implied by Figure 4.5 (10%), suggests that the total gain in value for these homes was £3.4 million. This figure of course ignores the homes that have not yet sold, the value uplift to land that has yet to be developed, and the value ignores any benefits not captured in the housing market.

4.5 Conclusions

Canals potentially provide a desirable recreational and environmental amenity. In this paper, we estimate the monetary-equivalent value of this amenity to local residents using house prices. The revealed preference framework adopted in this study is a standard approach to valuing non-market goods in the environmental and urban economics literature. Analysis of the effects of canal proximity for the whole of the England and Wales network indicates that households are willing to pay a 5% premium to live within 100m of a canal, on average

over the 2002-2017 period. The price premium falls substantially after the great recession from about 8.1% down to 3.4% in 2016, corresponding to annual monetary willingness to pay of around £520 pre-recession, and £260 post-recession, in 2016 prices. We find no price premium for living close to a canal but beyond 100m, which suggests that the effect is driven predominantly by canal-side properties, and others with a direct outlook on the canals or immediate access. We further observe a higher proportion of new-build sales within 100m of canals relative to elsewhere - a 5.9% increase on a 7.8% baseline, so around 75% higher - suggesting considerable response in construction to this demand for canal-side homes. A unique application of a difference-in-differences evaluation methodology to the restoration and environmental rehabilitation of the Droitwich Canal in the West Midlands supports the key findings on prices.

As an interesting, if very imprecise exercise, we calculate the potential implied land and property value uplift from the canal network. The length of the network covered in this analysis is 3500km. The price effects extend over 100m either side of the canal, so the affected area is 0.2 × 3500km = $700km^2$, which is just under half the area of Greater London. Though we do not have the exact figure in our data, around 10% of the land of England is urban/suburban and so developed or hypothetically developable, so the price uplift from canals would affect about $70km^2$, or 70 million m^2 of residential or potential residential land. Price per square metre of residential floor space in our sample of postcodes with 1.5km of the canals in 2016 is around £2700. If residential land prices are around two-thirds of this, they would be around £1800 per square metre on average. The 3.4% premium for living close to canals in 2016, thus implies a land value uplift of 0.045 × 1800 × 70 million = £4.3 billion pounds.

Of course not all of this urban land is built on for housing or ever likely to be. The proportion built on is more like 2.2%, so the implied increase in value of developed land is closer to £0.9 billion[7]. A similar figure can be obtained by aggregating the implied increased in value in the housing stock in our data. There are around 100,000 unique properties within 100m of a canal that transacted at least once over the entire 1995-2017 period on which we have data. The average price outside this distance band is £235,000 in 2016. The 3.4% uplift to property prices therefore implies a total increase in value of around £0.8 billion (0.034 ×235,000 ×100,000) aggregating across all the affected homes[8].

[7]These urban land cover figures come from NEA (2011).

[8]This value is relative to other places, so is not necessarily an addition to the total value of the land or housing stock in England and Wales.

Figure 4.1: Map of waterways managed by the Canal and River Trust and used in this analysis
Source: National Geographic, Esri, Garmin, HERE, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, increment P Corp

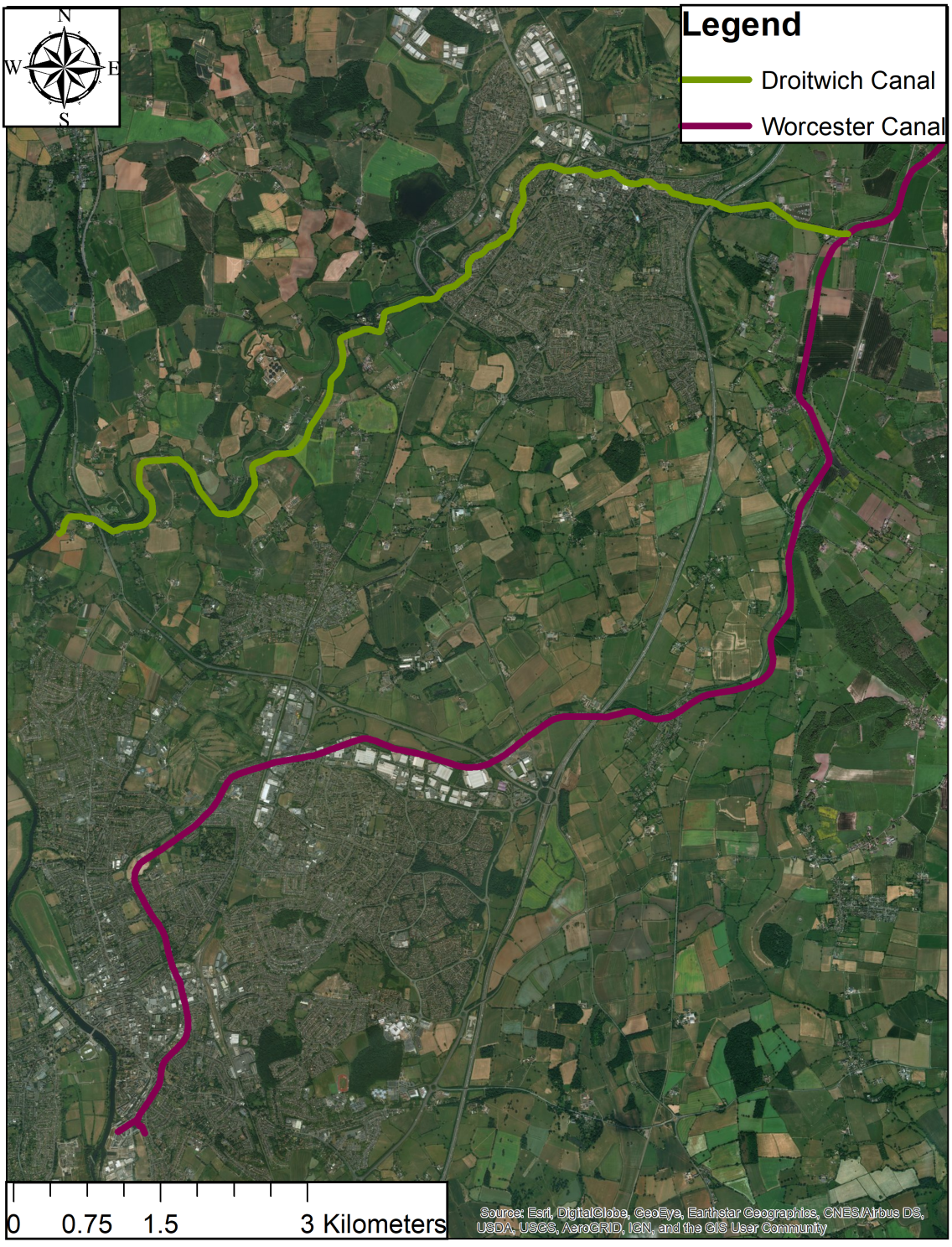
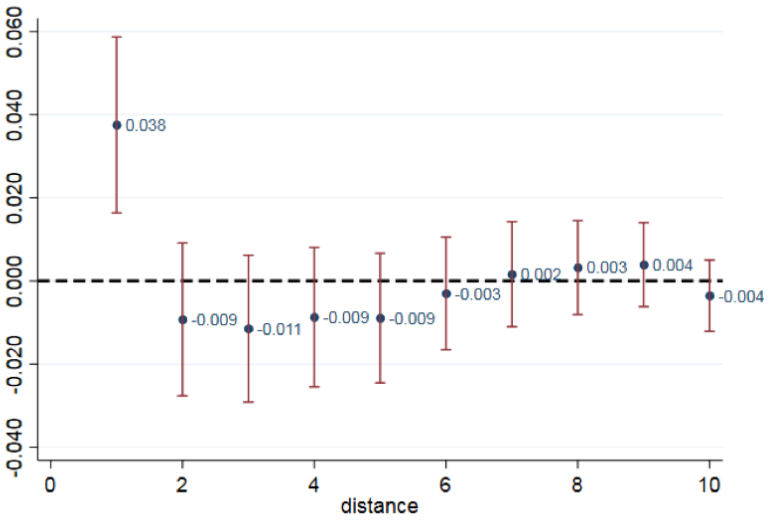
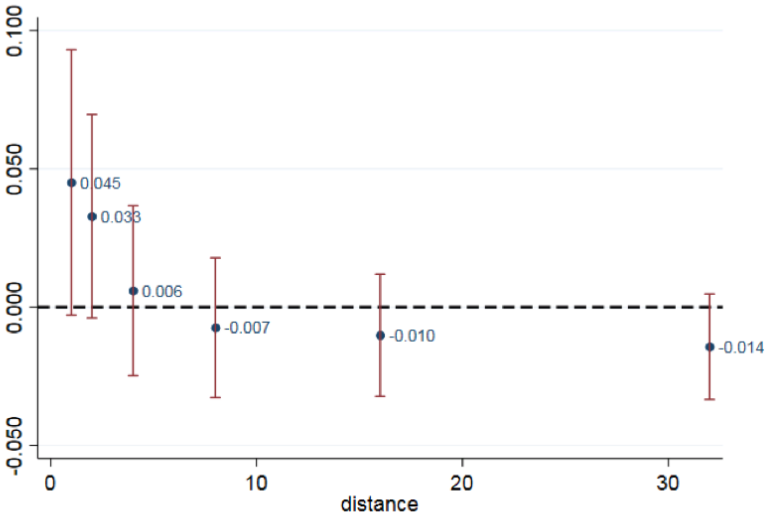


Figure 4.2: Droitwich Canals and Worcester Canal
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



(a)

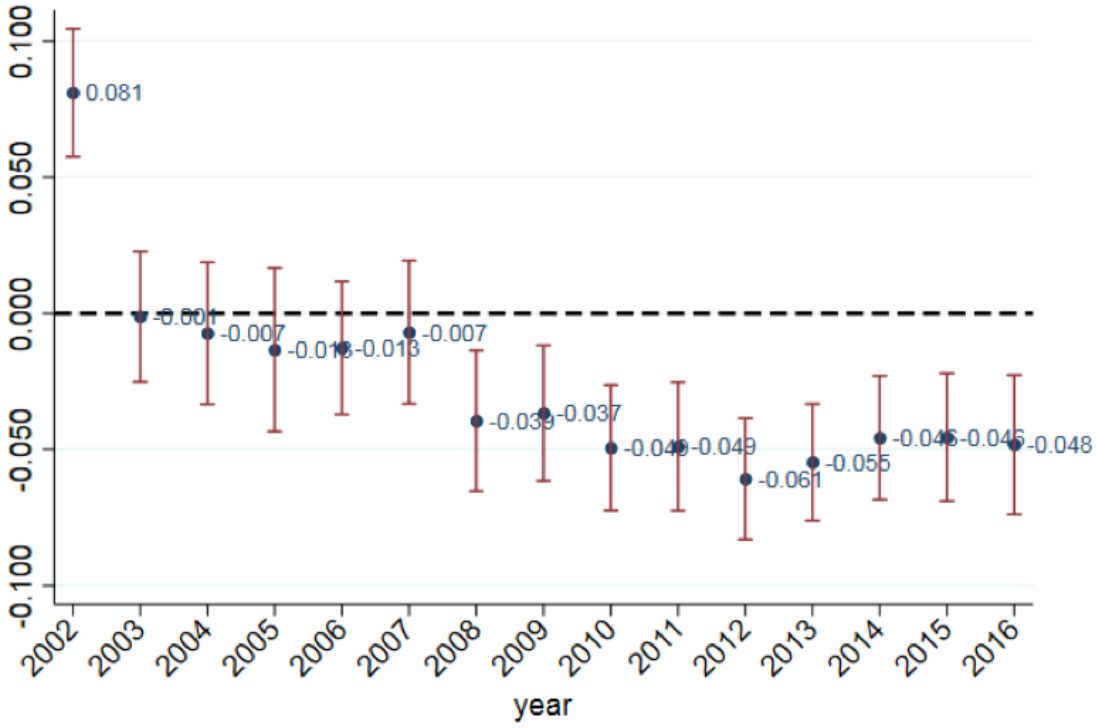


(b)

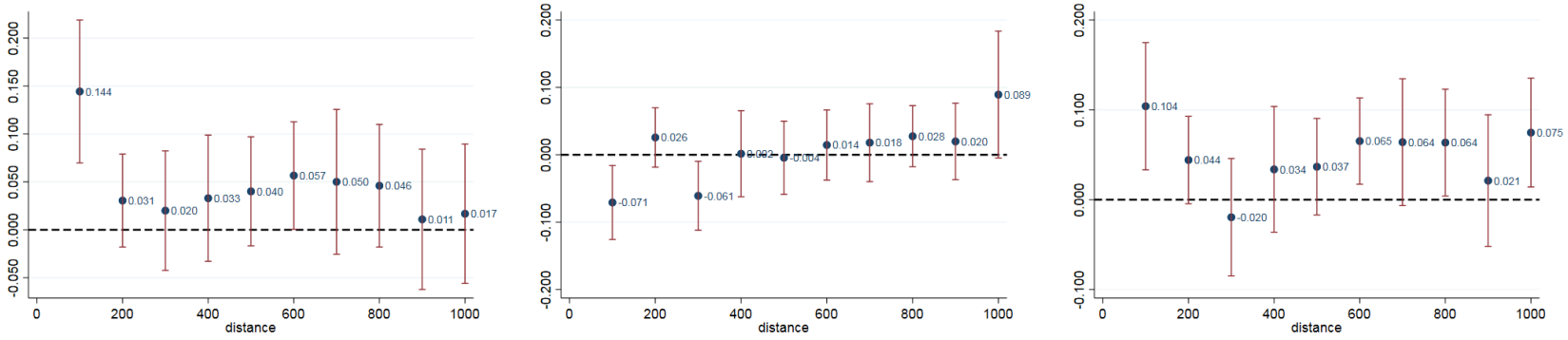
Figure 4.3: Effects of proximity to locks on house prices

Note: Lock effects significant at 10% level from 0-200m. Regressions control additionally for distance to wharves and aqueducts, where no significant effects were found. Regressions which restrict to effects from canals and locks within 0-100m show significant effects: canals 0.052 (0.008); locks 0.041 (0.021). Distances in 100m. Distance scale for locks differs from canals, because sample is restricted to 1500m from canals, but not 1500m from locks. In the regression, dependent variable is natural logarithm of transacted house prices. The sets of control variables include house structure, location, land cover, employment and LADs × Years × Quarters fixed effects and MOSA fixed effects. Standard errors are clustered at MSOA level.

Figure 4.4: Differences in percentage price effects by year



Note: The figure shows the estimates of the percentage price premium for properties within 0-100m in each year. The number for each year after 2002 need to be added to the number of 0.081 in 2002 to get the relevant percentage increase in that year.



(a) **(b)** **(c)**

Figure 4.5: Price effects from Droitwich Canal restoration at different distances

Note: The figures show distance on the x-axis, measured in 100m units. The y-axis represents the difference in log house prices relative to the baseline group, the group of properties beyond 1000m from a canal up to the 1500m limit of the estimation sample. Each dot is a coefficient (corresponding to the estimates in equation 4.2) and its value is shown alongside. The vertical capped bars indicate confidence intervals. Panel (a) shows a short run effect from the restoration date in 2007 to the opening date in 2011; Panel (b) shows the additional effects occurring around the official opening date in 2011. Panel (c) shows the effect of the start of renovation in 2007 without controlling for opening. In the regression, the sets of control variables include house structure, location, neighbourhood and postcode fixed effects. Standard errors are clustered at postcode level.

Table 4.1: Main results for effects from canal proximity on house prices

	(1)	(2)	(3)	(4)
Distance band 100 meters	0.0522***	0.0489***	0.0400***	0.0317***
	(0.0107)	(0.0107)	(0.0100)	(0.0091)
Distance band 200 meters	-0.0052	-0.0045	0.0009	-0.0003
	(0.0079)	(0.0088)	(0.0086)	(0.0079)
Distance band 300 meters	-0.0201***	-0.0106	-0.0038	-0.0025
	(0.0070)	(0.0084)	(0.0078)	(0.0072)
Distance band 400 meters	-0.0182***	-0.0083	-0.0043	-0.0013
	(0.0067)	(0.0078)	(0.0072)	(0.0065)
Distance band 500 meters	-0.0247***	-0.0106	-0.0060	-0.0009
	(0.0068)	(0.0073)	(0.0067)	(0.0061)
Distance band 600 meters	-0.0124*	-0.0051	-0.0005	0.0031
	(0.0064)	(0.0064)	(0.0061)	(0.0055)
Distance band 700 meters	-0.0069	-0.0011	-0.0005	0.0028
	(0.0063)	(0.0060)	(0.0057)	(0.0050)
Distance band 800 meters	-0.0104*	0.0006	0.0007	0.0039
	(0.0062)	(0.0054)	(0.0047)	(0.0042)
Distance band 900 meters	-0.0036	0.0029	0.0037	0.0035
	(0.0059)	(0.0048)	(0.0040)	(0.0037)
Distance band 1000 meters	-0.0144**	-0.0054	0.0003	0.0003
	(0.0057)	(0.0042)	(0.0036)	(0.0034)
LADs × Years × Quarters	X	X	X	X
House Structure	X	X	X	X
Location		X	X	X
Land Cover		X	X	X
Employment		X	X	X
MSOA Fixed Effets		X		
LSOA Fixed Effects			X	X
Neighbourhood				X
N	2048723	2048723	2048723	2048723
R^2	0.72	0.78	0.80	0.81
Absolute Price Change	12579.47	11764.38	9581.50	7573.25

Note: Dependent variable is natural logarithm of transacted house prices. See Table D.2 for the exact list of variables included for each set of controls. Absolute price change is calculated at the mean of the regression sample of year 2016 using the estimates for the 100 meters distance band. In columns 1, 3 and 4, standard errors are clustered at LSOA level. In column 2, standard errors are clustered at MSOA level.
* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 4.2: Heterogeneous effects from canal proximity on house prices

	(1) Urban	(2) Urban or suburban	(3) No rivers	(4) No green space
Canal within 100m	0.0268***	-0.0136	0.0583***	0.0567***
	(0.0058)	(0.0164)	(0.0086)	(0.0090)
Canal within 100m in area specified	0.0710***	0.0726***	-0.0151	-0.0015
	(0.0168)	(0.0181)	(0.0174)	(0.0128)
N	2048723	2048723	2048723	2048723
R^2	0.78	0.78	0.78	0.78
Absolute Price Change	24122.28	14274.93	10384.62	13324.18

Note: Dependent variable is natural logarithm of transacted house prices. Column headings: (1) Urban land cover predominant; (2) Urban or suburban landover predominant; (3) No rivers within 870 metres (top quartile); (4) No green space within 250 metres (top quartile). Specification controls for structural characteristics, distances to other water features, rail and town centres, land cover categories, employment variables at LSOA level, MSOA fixed effects, LAD x year x quarter fixed effects. See Table D.2 for the exact list of variables included for each set of controls. Absolute price change is calculated at the mean of the regression sample of year 2016 using the additional effect (main effect plus interaction effect) for the 100 meters distance band. Standard errors are clustered at LSOA level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table 4.3: Willingness to pay for property 0-100m from canals, by year

Year	£ Willingness to pay nominal	£ Willingness to pay 2016 prices	Annual equivalent at 3.5% discount rate
2002	9660	13058	457
2003	10886	14520	508
2004	11622	15298	535
2005	11278	14542	509
2006	12180	15350	537
2007	14211	17494	612
2008	7755	9220	323
2009	8226	9566	335
2010	6447	7262	254
2011	6393	6892	241
2012	4096	4292	150
2013	5612	5738	201
2014	7836	7891	276
2015	8155	8212	287
2016	7826	7826	274
Mean	8812	10477	367

Table 4.4: Effect of proximity to canals on new-build sales

	(1)	(2)	(3)	(4)
Distance band 100 meters	0.0851***	0.0593***	0.0490***	0.0405***
	(0.0093)	(0.0129)	(0.0101)	(0.0099)
Distance band 200 meters	0.0295***	0.0080	0.0076	0.0020
	(0.0066)	(0.0109)	(0.0087)	(0.0085)
Distance band 300 meters	0.0154***	-0.0030	0.0022	-0.0019
	(0.0059)	(0.0096)	(0.0080)	(0.0078)
Distance band 400 meters	0.0113*	-0.0033	0.0004	-0.0030
	(0.0059)	(0.0090)	(0.0073)	(0.0072)
Distance band 500 meters	0.0063	-0.0060	-0.0011	-0.0041
	(0.0051)	(0.0081)	(0.0066)	(0.0065)
Distance band 600 meters	0.0065	-0.0032	-0.0016	-0.0037
	(0.0050)	(0.0071)	(0.0060)	(0.0059)
Distance band 700 meters	0.0055	-0.0024	-0.0010	-0.0028
	(0.0047)	(0.0064)	(0.0054)	(0.0054)
Distance band 800 meters	-0.0041	-0.0100*	-0.0077*	-0.0087*
	(0.0041)	(0.0053)	(0.0046)	(0.0046)
Distance band 900 meters	0.0053	0.0012	0.0012	-0.0004
	(0.0045)	(0.0054)	(0.0042)	(0.0041)
Distance band 1000 meters	-0.0048	-0.0076*	-0.0067**	-0.0076**
	(0.0036)	(0.0039)	(0.0034)	(0.0033)
LADs × Years × Quarters	X	X	X	X
Location		X	X	X
Land Cover		X	X	X
Employment		X	X	X
MSOA Fixed Effets		X		
LSOA Fixed Effects			X	X
Neighbourhood				X
N	2048723	2048723	2048723	2048723
R^2	0.17	0.18	0.25	0.25

Note: Dependent variable is dummy variable indicating whether a sale is a new build. See Table D.2 for the exact list of variables included for each set of controls. Absolute price change is calculated at the mean of the regression sample of year 2016 using the estimates for the 100 meters distance band. In columns 1 and 2, standard errors are clustered at MSOA level. In column 3 and 4, standard errors are clustered at LSOA level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

A.1 Does Subway Improve Employment?

Table A.1: The effect of subway on employment: Difference-in-difference

	All survived Firms			Excluding Movers			New firms	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Manufacture firms								
Treated	0.415***	0.423***	0.389***	0.310**	0.313**	0.298**	0.049	-0.629*
	(0.116)	(0.115)	(0.113)	(0.135)	(0.137)	(0.138)	(0.040)	(0.330)
Adjusted R^2	0.06	0.06	0.06	0.03	0.02	0.02	0.47	0.04
N	162	162	162	103	103	103	264	56
Producer service firms								
Treated	0.418**	0.412**	0.436**	0.337	0.269	0.285	0.029	-0.143
	(0.198)	(0.178)	(0.178)	(0.265)	(0.249)	(0.277)	(0.036)	(0.177)
Adjusted R^2	0.02	0.01	0.02	0.00	-0.00	-0.03	0.61	0.02
N	119	119	119	80	80	80	357	148
Consumer service firms								
Treated	-0.231***	-0.250***	-0.198**	-0.159**	-0.174**	-0.108	0.049**	-0.386***
	(0.085)	(0.088)	(0.090)	(0.079)	(0.079)	(0.082)	(0.020)	(0.086)
Adjusted R^2	0.01	0.02	0.04	0.01	0.01	0.03	0.58	0.05
N	414	414	414	311	311	311	1401	705

Note: The dependent variable is changes in log employment. The key regressor of interest is treatment group dummy. The sample in Columns 1-3 includes all survived firms that are observed in both year 2008 and 2013. The sample in Column 4-6 excludes movers. Column 1 and 4 do not include any control variables. Column 2 and 5 control for log revenue in the year 2008. Column 3 and 6 include other control variables including age, SEZ dummy, plant dummy, and distances to the terminal station of the Line 2 original part. Column 7 and 8 include all control variables except for lag revenue. All samples include firms located up to 10 km from the district center and within 1km to the equidistance line on both sides. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

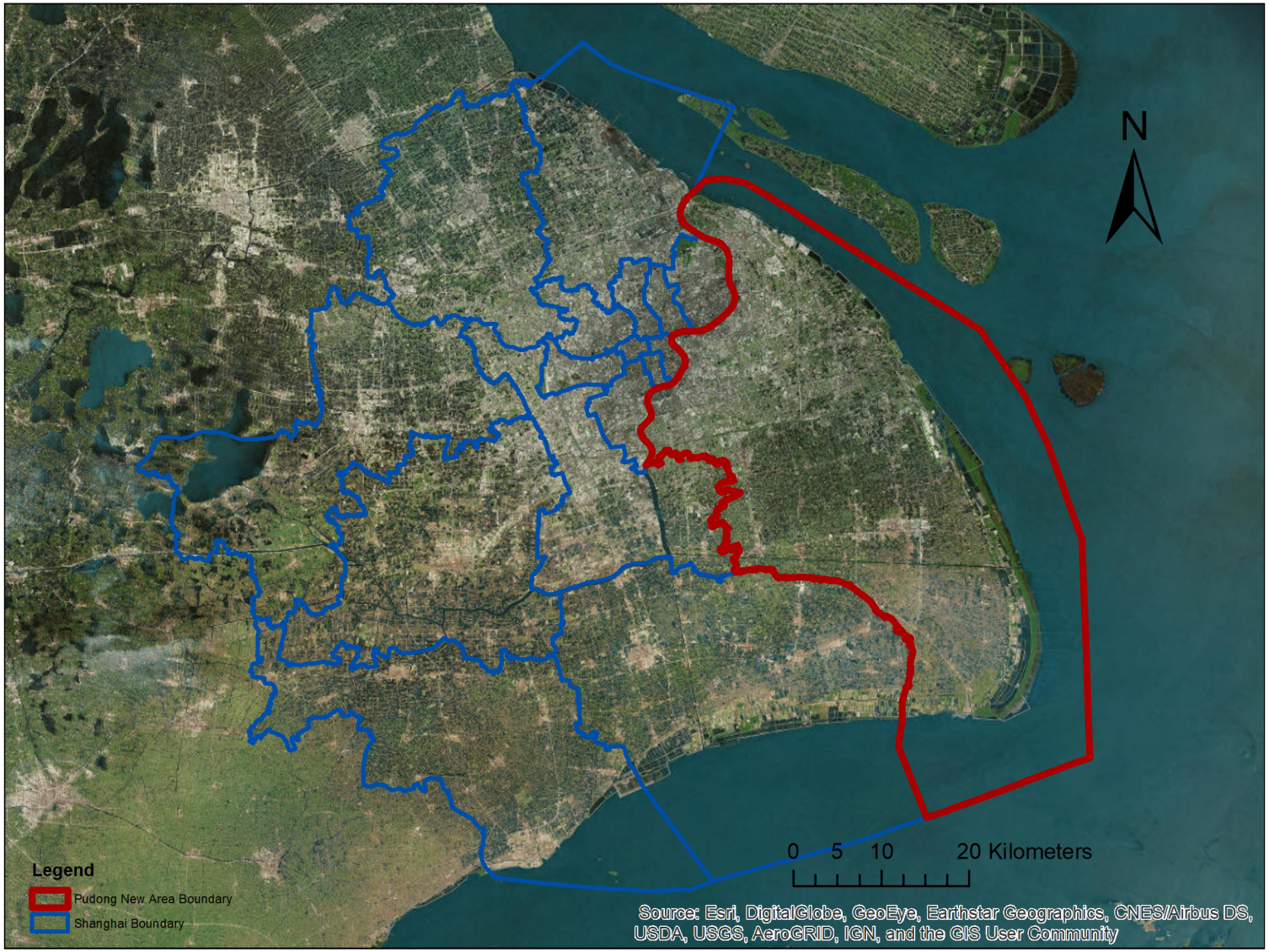
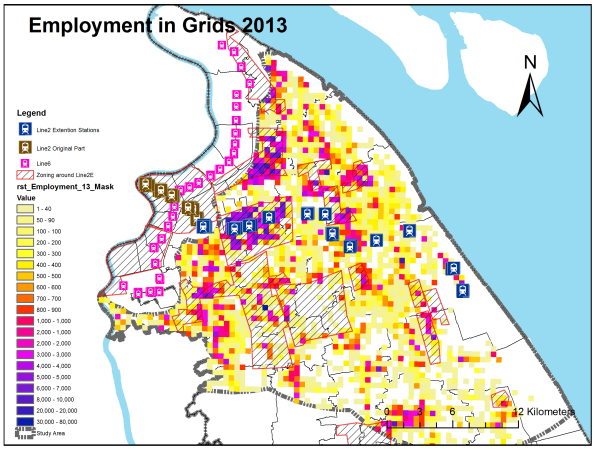
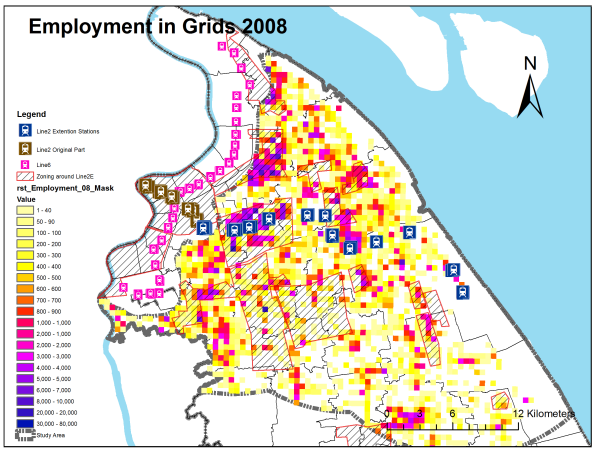


Figure A.1: The geography of Pudong New Area

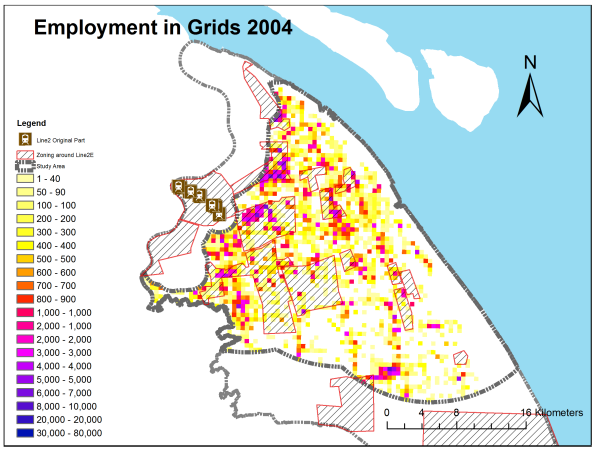
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community



(A.2.1) Year 2013



(A.2.2) Year 2008



(A.2.3) Year 2004

Figure A.2: Spatial Distribution of Employment

B.2 Does E-Commerce Reduce Traffic Congestion?

B.2.1 Supplementary Figures

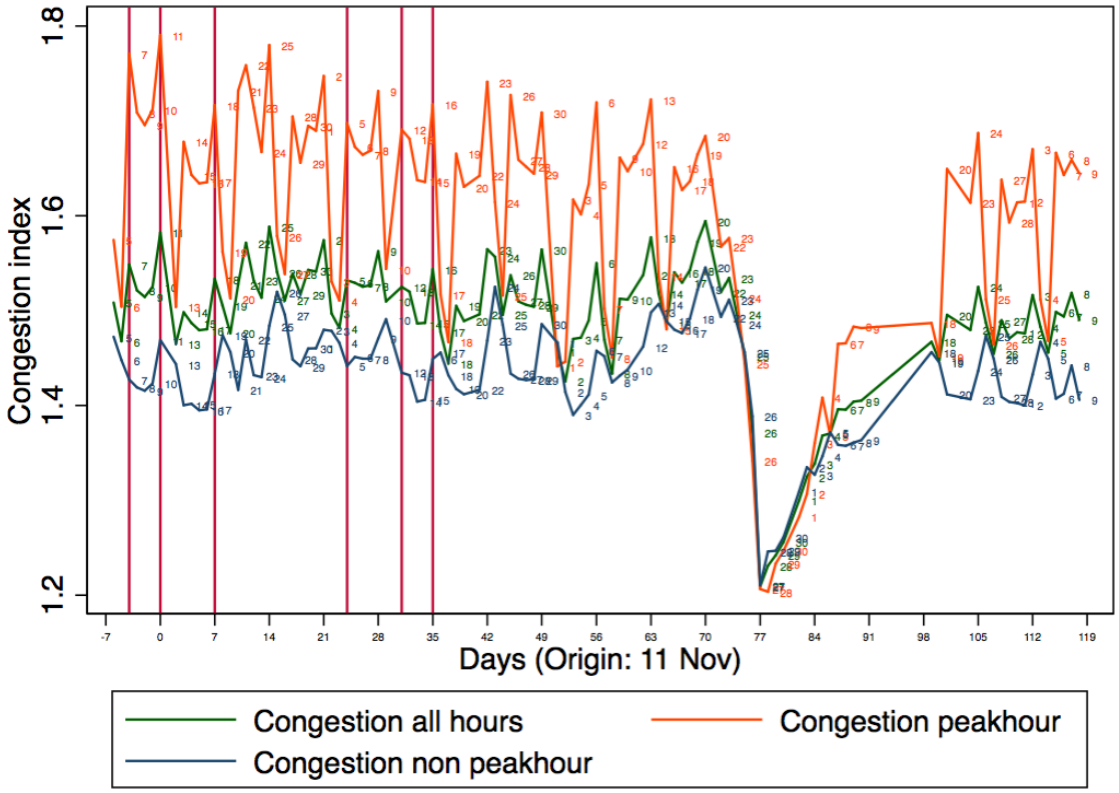
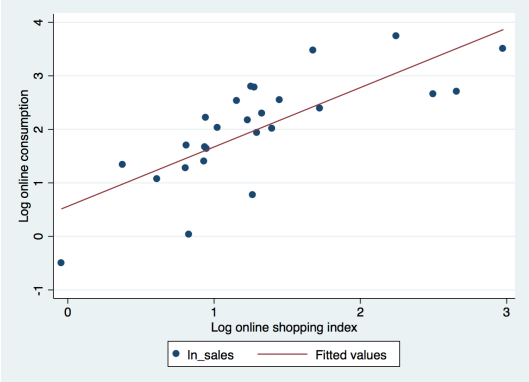
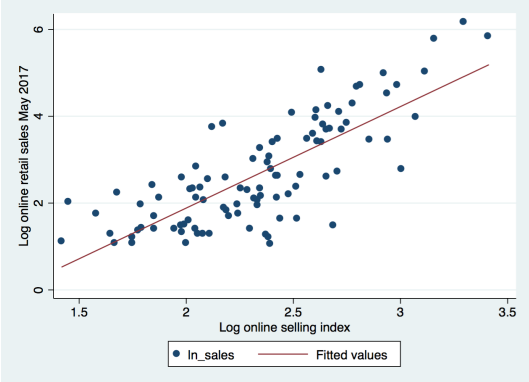


Figure B.1: The average daily level of traffic congestion

Note: Figure shows the daily average of traffic congestion, peak hour traffic congestion, and off-peak hour traffic congestion from 4 Nov 2016 to 9 March 2017. 0 in the x-axis marks Single Day Shopping Event day. The annotated numbers along the lines are dates of the observation. The first three vertical lines mark the week before and the week after the online shopping event day. The second three vertical lines mark the week before and the week after the offline shopping event day. The plummet of the index in the right shows traffic during Chinese New Year. The traffic congestion index is calculated using average actual travel time to free flow travel time of millions of GPS trails collected map navigation company.



(B.2.1)

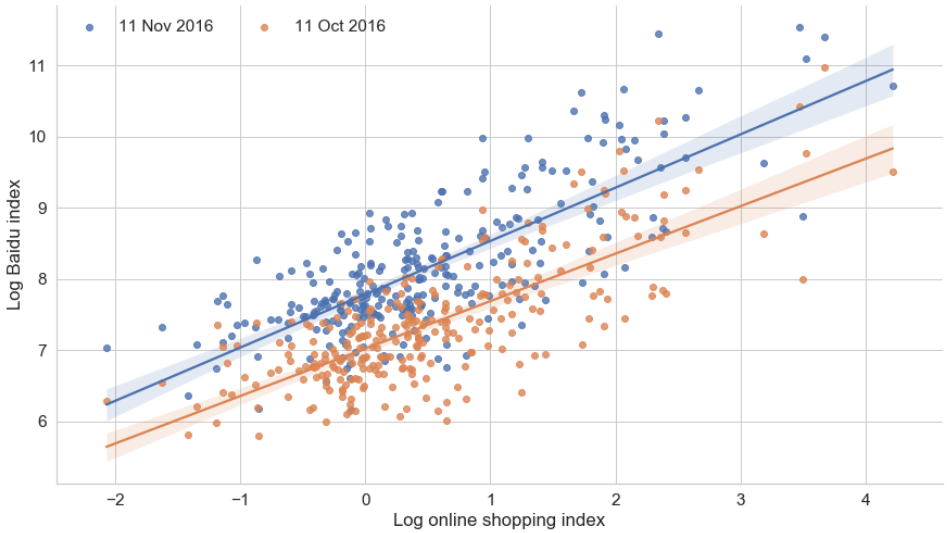


(B.2.2)

Figure B.2: Validation of e-commerce indices

Note: Figure (a) shows that log online consumption during Singles' Day Shopping event day is predicted by the online shopping index. Online consumption data is released by Alibaba and aggregated to province level. The online shopping index is the average of the province. Figure (b) shows that log online selling index is a predictor of log online sales. Online sales data is a overall predicted online sale in May 2017, released by an independent e-commerce research institute.

Figure B.3: Validation of the Baidu index and the online shopping index



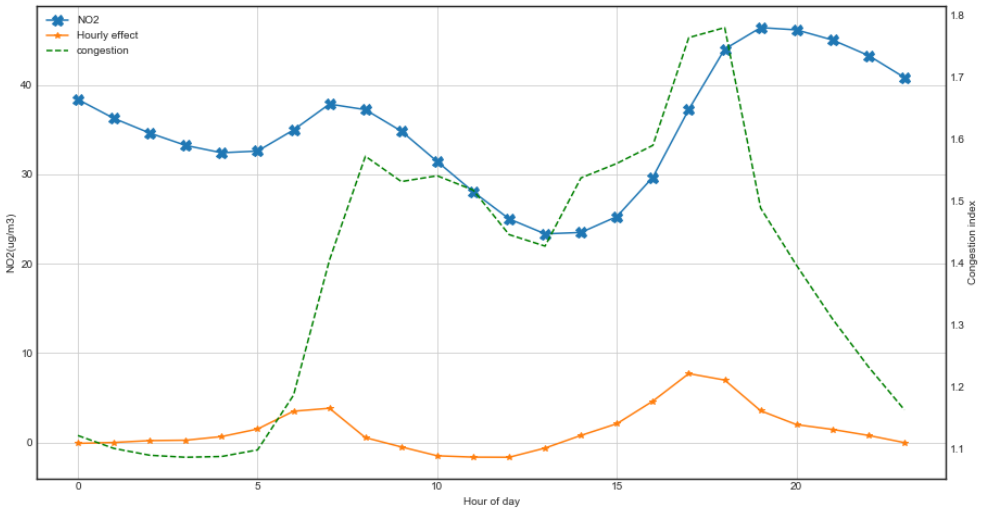


Figure B.4: Correlation between level of NO2 and traffic congestion in a day

Note: The figure overlays hourly NO2 shock and cumulative NO2 concentration with the traffic congestion index. Hourly NO2 shock is derived through solving a system of equations as explained in the text. The left Y-axis shows the concentration of NO2 in ug/m3 and the right Y-axis shows the traffic congestion index.

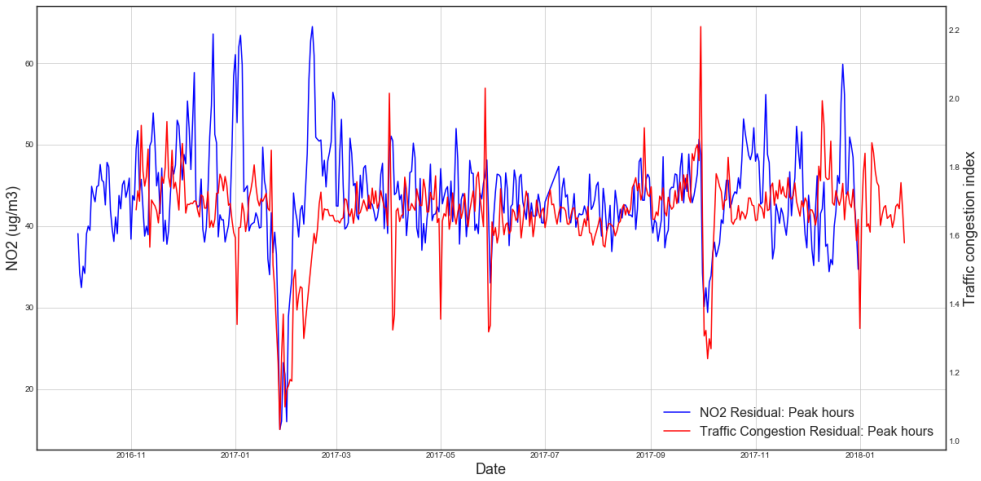
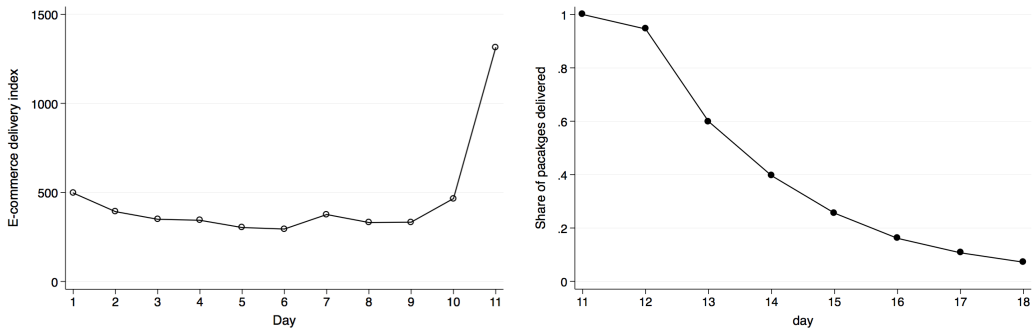


Figure B.5: Correlation between level of NO2 and traffic congestion across days

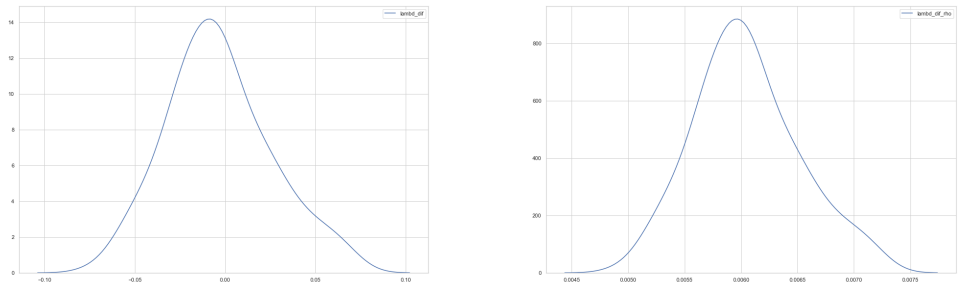
Note: The figure overlays smoothed daily NO2 level of concentration with smoothed traffic congestion index between Nov 2016 and Jan 2018. Both series smoothed by regressing on year, month and day-of-week dummies.



(B.6.1) **(B.6.2)**

Figure B.6: Time span of online shopping delivery

Note: Figure (a) shows the number of packages delivered before the Singles' Day Shopping Event. Figure (b) shows the time span of online delivery after the Singles' Day Shopping Event



(B.7.1) **(B.7.2)**

Figure B.7: Simulation results for a validation of the formula for λ

Note: Figure (a) show the distribution of the percentage difference between the λ^{theory} and $\lambda^{simulated}$. Figure (b) shows the distribution of the percentage difference between the λ^{ρ} and $\lambda^{simulated}$.

B.2.2 Construction of E-Commerce Indices

$$
\begin{aligned}
Online\ Shopping\ Index_i = \quad & E-Consumer\ Density\ Index_i \times 0.6 \\
& + Average\ Online\ Consumption\ Index_i \times 0.4 \\
E-Consumer\ Density\ Index_i = \quad & \frac{Density\ of\ E-Consumer_i}{Expected\ Density\ of\ E-Consumer} \\
= \quad & \frac{Count\ of\ E-Consumer_i \div Population_i}{63.26\%} \\
Average\ Online\ Consumption\ Index_i = \quad & \frac{Average\ Online\ Consumption_i}{Expected\ Average\ Online\ Consumption} \\
= \quad & \frac{Average\ Online\ Consumption_i}{32.5\ thousand\ RMB} \\
Online\ Selling\ Index_i = \quad & E-Seller\ Density\ Index_i \times 0.6 \\
& + Average\ Online\ Sale\ Index_i \times 0.4 \\
E-Seller\ Density\ Index_i = \quad & \frac{Density\ of\ E-Seller_i}{Expected\ Density\ of\ E-Seller} \\
= \quad & \frac{Count\ of\ E-Consumer_i \div Population_i}{13.47\%} \\
Average\ Online\ Sale\ Index_i = \quad & \frac{Average\ Online\ Sale_i}{Expected\ Average\ Online\ Sale} \\
= \quad & \frac{Average\ Online\ Sale_i}{1\ million\ RMB}
\end{aligned}
$$

B.2.3 Mathematical Derivations

B.2.3.1 Derivation of the Elasticity of Online Consumption to the Relative Price Between Two Channels

Given equation 2.20, and the mean effect is γ times of the share effect, the derivative of online consumption quantity from city i in city j to the price shock can be written as $\gamma + 1$ times the share effect. I take advantage of $k = 1$, which is the key assumption for the result in equation 2.23 to hold, before the event to simplify the derivatives calculation.

$$
\frac{dQ_{oij}}{dk}|_{k=1} = (\underbrace{\pi_o \frac{dA}{dk}}_{\gamma\nu} + \underbrace{A\frac{d\pi_o}{dk}}_{\nu})C_{ij} \tag{B.21}
$$

$$
= (\frac{\sigma-1}{\theta}s^{\theta}+1)\frac{d\pi_o}{dk}AC_{ij} \tag{B.22}
$$

Summing across all origin cities i,

$$
\frac{d\sum_i Q_{oij}}{dk}|_{k=1} = (\pi_o\frac{dA}{dk} + A\frac{d\pi_o}{dk})C_{ij} \tag{B.23}
$$

$$
= (\frac{\sigma-1}{\theta}s^{\theta}+1)\frac{d\pi_o}{dk}A\sum_i C_{ij} \tag{B.24}
$$

which gives the derivative of online consumption quantity in city j from all source cities including itself,

$$\frac{dQ_{oj}}{dk}|_{k=1} = (\pi_o \frac{dA}{dk} + A\frac{d\pi_o}{dk})C_{ij} \tag{B.25}$$

$$= (\frac{\sigma-1}{\theta}s^\theta + 1)\frac{d\pi_o}{dk}A\sum_i C_{ij} \tag{B.26}$$

Using equation 2.11, $A\sum_i C_{ij} = Q_j$ and $Q_{oi} = \pi_o Q_j$, equation B.27 can be written as,

$$\frac{dQ_{oj}}{Q_{oj}}|_{k=1} = (\frac{\sigma-1}{\theta}s^\theta + 1)\frac{d\pi_o}{dk}\frac{1}{\pi_o}dk \tag{B.27}$$

Given the derivative of the share of online shopping on price shock is,

$$\frac{d\pi_o}{dk}|_{k=1} = \frac{\theta s^\theta}{(s^\theta+1)^2}\frac{\sigma}{\sigma-1} \tag{B.28}$$

The growth rate of online shopping can be simplified to,

$$\frac{dQ_{oj}}{Q_{oj}}|_{k=1} = dk\frac{(\sigma-1)s^\theta+\theta}{s^\theta+1}\frac{\sigma}{\sigma-1} \tag{B.29}$$

Given that $\frac{\Delta Q_{oj}}{Q_{oj}}/\frac{\Delta k}{k}$ approximates the elasticity of the online consumption quantity to price shock when Δk is small, I obtain,

$$\rho|_{k=1} = \frac{(\sigma-1)s^\theta+\theta}{s^\theta+1}\frac{\sigma}{\sigma-1} \tag{B.210}$$

Plugging equation B.210 into equation equation 2.22 gives,

$$\lambda|_{k=1} = -\frac{\theta+(\sigma-1)s^\theta}{\theta-\sigma+1} \tag{B.211}$$

$$= -\frac{\frac{\sigma-1}{\sigma}\rho(s^\theta+1)}{\frac{\sigma-1}{\sigma}\rho(s^\theta+1)-(\sigma-1)s^\theta-\sigma+1} \tag{B.212}$$

$$= -\frac{\frac{\sigma-1}{\sigma}\rho}{\frac{\sigma-1}{\sigma}\rho-\sigma+1} \tag{B.213}$$

$$= -\frac{\rho}{\rho-\sigma} \tag{B.214}$$

Taking the inverse of λ and the absolute value gives,

$$\frac{1}{|\lambda|} = 1-\frac{\sigma}{\rho} \tag{B.215}$$

B.2.3.2 Simulation Procedure and Results

One of the key predictions from the model is that the online-offline substitution is

$$\lambda^{theory} = \frac{\sum_i \Delta Q_{oij}}{\sum_i \Delta Q_{fij}} = -\frac{\theta + (\sigma - 1)k^\theta s^\theta}{\theta - \sigma + 1}$$

The exercise below verifies this result.

First, I set the number of products as 1000, and the number of consumers as 4000. For simplicity, I assume there are two cities, with one city having 1600 consumers and the other having 2400 consumers. I set that one city produces 300 types of products and the other city produces 700 types of products, which are sold in both cities. Denote vector $\vec{x}$ as the values of city characteristics x in cities i and j. I set hourly wage $\vec{w} = (10, 20)$ and productivity $\vec{p} = (10, 20)$. I set intercity transportation cost $\tau = 1$ if the firm and the consumer are in the same city, and set $\tau = 2$ if the firm and the consumer are in the different cities. Income in each city is assumed to be $\vec{w} \times 40$, assuming workers work 40 hours a week. Consumption quantity is thus calculated on a weekly basis. I create two set of draws of z_o or z_f for each pair of product and consumer, from two Fréchet distributions with an online channel preference parameter $s_o = 0.8$ and an offline channel preference parameter $s_f = 1$. The shape parameter is set as the same in both channels, $\theta = 9$. I set the elasticity of substitution between varieties $\sigma = 4$ following the trade literature. These results hold for other sets of assumed values.

In normal days without a sales event, I set $k_o = k_f = 1$. For a specific variety, I assign consumers who satisfy the condition $k_o z_o > k_f z_f$ to online type and assign the rest to offline type. I then calculate the consumption quantity based on equation 2.3 for online type and offline type, respectively. The price index is calculated based on equation 2.4.

In the event, I set $k_o^{event} = 1.1$ while maintaining the values of initial draws of z_m. Based on the old condition $k_o z_o > k_f z_f$ and a new condition $k_o^{event} z_o > k_f z_f$, there are three types of consumers for each variety: consumers that remain online, consumers that switch from offline to online, and consumers that remain offline. Consumers switch from offline to online because for them $k_o z_o < k_f z_f$ while $k_o^{event} z_o > k_f z_f$. Then, I recalculate the consumption quantity for each consumer.

I sum the consumption quantity for online and offline consumers, respectively, to obtain Q_{mj} and Q_{mj}^{event} for city j, which gives the ΔQ_{mj}. I use them to calculate $\lambda^{simulated}$ and compare it with λ^{theory}. I construct a statistic $\frac{\lambda^{simulated} - \lambda^{theory}}{\lambda^{theory}}$ to measure the relative difference between the two values. Given that $\lambda^{simulated}$ is a random variable, I replicate the above procedures 100 times and plot the distribution of the statistic in Figure B.7 (a).

Similarly, I calculate $\rho_j = \frac{\Delta Q_{oj}}{Q_{oj}}$ to obtain the online-offline substitution λ^ρ following equation

2.23. Again, I construct a statistic $\frac{\lambda^{simulated} - \lambda^{\rho}}{\lambda^{\rho}}$ to measure the relative difference between the two values. I replicate the above procedures 100 times and plot the distribution of the statistic in Figure B.7 (b).

B.2.4 Additional Analysis and Results

B.2.4.1 The Characteristics of the Cities That Contribute Higher Variations to the IV Estimates

A heuristic approach would be looking at heterogeneous effects of the IV on the endogenous variable in the first stage. Table B.1 shows the result when adding interaction of the IV with city characteristics. Column 1 presents the results without any interaction terms as the benchmark. The rest of the columns add the interaction term with the variable specified in the column title. The effect of postage fee on the change of online shopping clearly increases with market potential, income, and the number of online consumers. This suggests that the IV estimates are identified from the cities with such attributes.

Table B.1: Heterogeneity in the first-stage

	(1) No interaction	(2) Market potential	(3) Income	(4) Internet	(5) Mobile
Log avg postage fee	0.612***	0.132	0.000	-0.031	-0.014
	(0.223)	(0.134)	(0.008)	(0.038)	(0.028)
Interaction of postage fee					
with the variable in column title		0.471***	0.088***	0.194***	0.152***
		(0.047)	(0.001)	(0.004)	(0.003)
Market potential	-0.017	-0.319***	0.005	0.024	0.013
	(0.115)	(0.112)	(0.004)	(0.017)	(0.012)
Log GDP per capita	-0.060**	-0.044**	-0.089***	-0.007*	-0.005
	(0.028)	(0.019)	(0.005)	(0.004)	(0.003)
Log internet users	0.016	-0.001	0.000	-0.193***	-0.007
	(0.043)	(0.023)	(0.002)	(0.007)	(0.005)
Log mobile users	0.078*	0.014	-0.001	0.004	-0.137***
	(0.045)	(0.022)	(0.002)	(0.006)	(0.006)
R^2	.38	.79	1	.99	.99
N	90	90	90	90	90

Note: The dependent variable is the growth rate of the Baidu index. The second row of each column shows the coefficients of the interaction term of the log average postage fee with the variable in the column title. Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

B.2.4.2 The Changes in Intercity Traffic

Table B.2 quantifies the change in intercity traffic following equation 2.36. The outcome variable of interest is the travel time in minutes for any routes between cities. The regression sample contains Thursdays and Fridays in the weeks before and after the event. All regressions control for route length. As shown in the first row of the table, the estimated average travel time per km is stable at 0.7 minutes (or 86 kmh) during peak hours and 0.69 during off-peak hours. Intercity roads congestion appeared not to vary much with peak hours. The first three columns report the results for peak hours and the last three columns report the results for off-peak hours. Columns 1 and 4 report the average change in traffic in the first week γ_1 and that in the second week γ_2 for peak hours and off-peak hours, respectively. Column 1 shows that a typical traveler is slowed down by about 24 minutes in peak hours in the first week after the large-scale online sale. The effect continues in the second week, though nearly halved in magnitude. The increase in traffic in the second week is only a third of the first week for off-peak hours. As expected, the delay is more significant during peak hours. Columns 2 and 4 show the results for the interaction specifications, which explores how the characteristics of the route's origin city affect travel time. Cities with a higher online selling index have a bigger spike in traffic congestion, while cities with a higher online shopping index experience, comparatively, less increase in traffic. A possible explanation is that the cities that are involved in more online selling activity may ship more goods out of the city, with the traffic on the intercity roads originating from these cities expected to be more congested. The effect is more salient in the second week. Similar patterns are observed for off-peak hours. Columns 3 and 6 add in the interaction terms of the online shopping index and the online selling index in the destination cities. The characteristics of the route destination cities appear irrelevant. It is desirable to have a longer time series to pin down the diminishing trend of traffic congestion after the event. Unfortunately, I only have the data for three weeks[9]. Nevertheless, this regressive pattern in the post-event travel time is in line with the e-commerce delivery index and share of packages delivered by days surrounding the event, as shown in Figure B.6. The pattern of the change in travel time emerging from the three weeks in the 9,126 routes between cities provides strong evidence of the impact of the online sale on traffic congestion on intercity roads.

[9]I only employed the Baidu map API for Thursday and Friday in three weeks due to budget constraints because of data collection cost.

Table B.2: Estimates of the changes in intercity travel time before and after the event

	(1) Peak hour	(2) Peak hour	(3) Peak hour	(4) Non peak hour	(5) non peak hour	(6) Non peak hour
Distance	0.70***	0.70***	0.70***	0.69***	0.69***	0.69***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Week 1	24.06***	29.35	42.11	27.60***	19.82	24.97
	(8.83)	(44.07)	(51.37)	(7.84)	(38.16)	(41.87)
Week 2	7.84**	-12.06	2.30	11.83***	-2.25	4.87
	(3.16)	(18.29)	(24.74)	(2.24)	(12.22)	(17.77)
Week 1 × Online Selling Origin		3.49	3.43		7.02	7.00
		(22.09)	(22.10)		(19.42)	(19.42)
Week 2 × Online Selling Origin		19.23*	19.15*		10.67	10.63
		(10.44)	(10.45)		(6.96)	(6.97)
Week 1 × Online Shopping Origin		-9.75	-9.76		-6.23	-6.24
		(8.31)	(8.31)		(7.34)	(7.34)
Week 2 × Online Shopping Origin		-18.12***	-18.11***		-7.84**	-7.84**
		(5.42)	(5.42)		(3.45)	(3.46)
Week 1 × Online Selling Destination			-4.66			-1.62
			(12.99)			(8.13)
Week 2 × Online Selling Destination			-6.33			-3.45
			(9.10)			(7.36)
Week 1 × Online Shopping Destination			-1.25			-0.95
			(5.35)			(3.39)
Week 2 × Online Shopping Destination			0.41			0.73
			(4.18)			(3.21)
R^2	0.96	0.96	0.96	0.98	0.98	0.98
N	263302	263302	263302	680925	680925	680925

Note: The dependent variable is travel time in minutes. The reference group is the week before the Singles' Day shopping event. Columns 1-3 show the results for peak hours, and columns 4-6 show results for off-peak hours. Origin city fixed effects, destination city fixed effects, day-of-week fixed effects, and hour fixed effects are included. Standard errors are clustered at the origin cities × destination cities level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

C.3 Colonial Legacies: Shaping African Cities

C.3.1 Data

C.3.1.1 Data Sources

Built-up cover of 38 meter resolution in year 1990, 2000, and 2014 is from Gobal Human Settlement layer (GHSL). Pesaresi, M.; Guo Huadong; Blaes, X.; Ehrlich, D.; Ferri, S.; Gueguen, L.; Halkia, M.; Kauffmann, M.; Kemper, T.; Linlin Lu; Marin-Herrera, M.A.; Ouzounis, G.K.; Scavazzon, M.; Soille, P.; Syrris, V.; Zanchetta, L., "A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results," Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.6, no.5, pp.2102,2131, Oct. 2013

Historical road blocks data for the 40 Open Street Map (OSM) cities is derived from digitalising historical maps from Oxford and Cambridge library. Current road blocks data is extracted from OSM. https://www.openstreetmap.org

Population of cities is from two sources: Citypopulation.de (Census data, 39 countries) http://citypopulation.de and Africapolis database (for Angola and Nigeria) http://www.oecd.org/swac/ourwork/africapolis.htm. The population of an urban area is the sum of the population of all "cities" falling within the lights boundary of an urban area.

National GDP per capita is from Penn World Table 9.0. https://www.rug.nl/ggdc/productivity/pwt/

National urban population is from World Bank. http://wdi.worldbank.org/table/3.12

River and lake GIS data is from Global Lakes and Wetlands Database. https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database

Elevation and ruggedness variables are derived from Aster DEM elevation by NASA. https://search.earthdata.nasa.gov/search?m=24.1875!3.234375!3!1!0!0%2C2

Rainfall variables are from "The Climate Data Guide: Global (land) precipitation and temperature: Willmott & Matsuura, University of Delaware." https://climatedataguide.ucar.edu/climate-data

Fringe geographic controls are from Land use system of the World. Nachtergaele, F & Petri, Monica. (2008). Mapping Land Use Systems at global and regional scales for Land Degradation Assessment Analysis Version 1.1. http://www.fao.org/geonetwork/srv/en/graphover.show?id=37048&fname=lus_ssa.png&access=public

Public utility connections variables including electricity, phone land line, piped water and flush toilet are from Demographic and Health Surveys Program (DHS). https://dhsprogram.com

C.3.1.2 Statistics on Variables

	mean	sd	min	max
Ln count of LF	3.17	1.67	0.0	7.6
Ln LF minus ln total patches	-2.46	0.90	-7.0	0.0
Ln average LF area	8.26	0.44	7.3	10.9
Ln openness index 2014	3.75	0.38	2.4	4.6
Ln light area	4.89	1.20	1.6	8.9
Anglophone dummy	0.73	0.44	0.0	1.0
Ln initial cover 1990	15.60	1.60	9.6	20.8
Year dummy 2014	0.51	0.50	0.0	1.0
Lag t-1 ln country GDP per capita	7.64	0.47	6.6	9.2
Ln annual population growth 90 to 12	0.03	0.02	-0.0	0.1
Ln projected city population 1990	11.31	0.92	10.3	15.7
Ln ruggedness	6.89	1.19	3.1	8.7
Ln rainfall	4.48	0.59	1.1	5.6
Ln elevation range	5.22	0.71	3.5	7.3
Coast dummy	0.04	0.20	0.0	1.0
Interaction ln coast length	0.42	2.07	0.0	12.0
Interaction ln distance to coast	12.29	2.65	0.0	14.4
Fraction of river area	0.01	0.03	0.0	0.2
Fraction of lake area	0.01	0.03	0.0	0.3
Fraction of forrest	0.20	0.30	0.0	1.0
Fraction of shrubs	0.17	0.27	0.0	1.0
Fraction of crops	0.40	0.37	0.0	1.0
Fraction of wetlands and water	0.01	0.05	0.0	0.4
Fraction of sparse vege and bare land	0.03	0.15	0.0	1.0
Observations	551			

Note: The sample is the same as column 4 in Table 3.3

C.3.1.3 City Built Cover Boundary

We adopted a smoothing algorithm to define the city built cover boundary. First, we measured the area of total built cover for each 500m $\times$ 500m grid. Then the smoothing algorithm gives each grid the average built cover value of its neighbor grids and itself. The neighborhood is all queen and rook neighbors on the grid. Note if there is any grid in a neighborhood that has no built-up cover, the averaged built-up is set to be zero. This condition helps to eliminate scattered built-up and obtain continuous built cover area. Finally, we select the grids with neighborhoods which average over 10% built cover, and use them to form the final built cover boundary of cities.

C.3.2 Supplementary Tables

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Bohicon	Benin	0	89,553	166,611	Yes	
Djougou	Benin	0	47,383	81,341		
Lokossa	Benin	0	30,328	70,048		
Parakou	Benin	0	96,206	216,706		
Pobè	Benin	0	35,163	67,425		
Quidah	Benin	0	921,859	1,922,874		
Toviklin	Benin	0	35,688	66,505		
Francistown	Botswana	1	65,935	109,269	Yes	
Gaborone	Botswana	1	215,068	487,079	Yes	
Kanye	Botswana	1	30,552	47,698	Yes	
Molepolole	Botswana	1	35,517	67,791		
Selebi-Phikwe	Botswana	1	45,446	61,570		
Banfora	Burkina Faso	0	41,261	97,859	Yes	
Bobo-Dioulasso	Burkina Faso	0	262,478	645,198		
Koudougou	Burkina Faso	0	60,177	99,187		
Ouagadougou	Burkina Faso	0	578,653	2,213,074		
Ouahigouya	Burkina Faso	0	44,462	89,579		
Bafang	Cameroon	0	37,503	33,806		
Bamenda	Cameroon	0	129,657	413,538	Yes	
Bandjoun	Cameroon	0	129,500	359,215		
Bertoua	Cameroon	0	48,871	116,686	Yes	
Douala	Cameroon	0	935,407	2,691,721		
Dschang	Cameroon	0	39,347	80,013		
Edéa	Cameroon	0	52,976	74,076		
Foumban	Cameroon	0	60,988	96,722		
Garoua	Cameroon	0	154,400	287,668		
Guider	Cameroon	0	35,432	62,750		
Kousséri	Cameroon	0	58,443	108,520		
Kumbo	Cameroon	0	38,606	112,836		
Loum	Cameroon	0	40,726	60,213		
Maroua	Cameroon	0	133,940	243,578		
Mbouda	Cameroon	0	37,434	50,758		
Meiganga	Cameroon	0	32,793	40,856		
Ngaoundéré	Cameroon	0	87,298	198,223	Yes	
Nkongsamba	Cameroon	0	88,275	112,347	Yes	
Yaounde	Cameroon	0	771,858	2,744,390	Yes	
Bambari	Central African Republic	0	38,985	43,081		
Berbérati	Central African Republic	0	45,426	110,757		
Bimbo	Central African Republic	0	492,970	995,932		
Bossangoa	Central African Republic	0	32,124	39,833		
Bouar	Central African Republic	0	39,766	40,765		
Carnot	Central African Republic	0	32,915	56,765		
Abéché	Chad	0	48,962	109,300		
Moundou	Chad	0	93,710	145,775	Yes	
Ndjamena	Chad	0	475,961	1,061,368	Yes	
Sarh	Chad	0	71,999	101,946		
Abengourou	Cote d'Ivoire	0	61,400	nan		
Abidjan	Cote d'Ivoire	0	2,312,639	4,395,000		

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Akoupé	Cote d'Ivoire	0	38,495	nan		
Bondoukou	Cote d'Ivoire	0	35,283	nan		
Bouaflé	Cote d'Ivoire	0	37,918	nan		
Bouaké	Cote d'Ivoire	0	352,785	536,719	Yes	
Daloa	Cote d'Ivoire	0	130,708	nan	Yes	
Danané	Cote d'Ivoire	0	34,582	nan		
Dimbokro	Cote d'Ivoire	0	39,581	nan		
Ferkéssédougou	Cote d'Ivoire	0	40,675	nan		
Gagnoa	Cote d'Ivoire	0	112,890	nan		
Issia	Cote d'Ivoire	0	30,922	nan		
Katiola	Cote d'Ivoire	0	34,581	nan		
Korhogo	Cote d'Ivoire	0	115,302	nan		
Man	Cote d'Ivoire	0	94,435	nan		
Odienné	Cote d'Ivoire	0	31,202	nan		
Sinfra	Cote d'Ivoire	0	37,773	nan		
Séguéla	Cote d'Ivoire	0	31,517	nan		
Yamoussoukro	Cote d'Ivoire	0	139,062	nan		
Libreville	Gabon	0	394,152	694,622	Yes	
Banjul	Gambia	1	357,893	460,450	Yes	Yes
Accra	Ghana	1	2,004,164	3,689,581		Yes
Bawku	Ghana	1	39,747	63,318		
Bolgatanga	Ghana	1	37,953	69,431		
Dzodze	Ghana	1	52,458	nan		
Ho	Ghana	1	45,396	116,172		
Koforidua	Ghana	1	68,148	129,122	Yes	
Kumasi	Ghana	1	836,568	2,382,130		Yes
Nkawkaw	Ghana	1	35,816	48,870		
Sunyani	Ghana	1	46,279	76,966		
Tamale	Ghana	1	177,409	409,675		
Techiman	Ghana	1	34,094	69,700		
Wa	Ghana	1	45,405	71,967		
Yendi	Ghana	1	34,652	54,365		
Boké	Guinea	0	35,332	58,679		
Conakry	Guinea	0	942,708	1,824,765	Yes	
Fria	Guinea	0	41,303	53,703	Yes	
Guéckédou	Guinea	0	85,391	64,617		
Kamsar	Guinea	0	55,242	82,002		
Kankan	Guinea	0	80,409	180,127		
Kindia	Guinea	0	85,776	129,993		
Kissidougou	Guinea	0	59,539	86,954		
Labé	Guinea	0	40,570	84,218		
Macenta	Guinea	0	44,266	56,709		
Mamou	Guinea	0	45,178	63,059	Yes	
Nzérékoré	Guinea	0	88,082	181,799		
Eldoret	Kenya	1	116,456	285,187		
Garissa	Kenya	1	32,881	161,277		
Kisii	Kenya	1	47,004	74,984		
Kisumu	Kenya	1	194,711	326,009	Yes	
Kitale	Kenya	1	56,884	80,007	Yes	
Mombasa	Kenya	1	491,834	1,167,440		Yes

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Nairobi	Kenya	1	1,516,055	5,044,352	Yes	
Nakuru	Kenya	1	170,002	336,431	Yes	
Maputsoe	Lesotho	1	59,779	103,567		
Maseru	Lesotho	1	117,442	178,016	Yes	
Teyateyaneng	Lesotho	1	42,583	61,599	Yes	
Antananarivo	Madagascar	0	675,058	1,300,000		
Antsirabe	Madagascar	0	117,026	nan	Yes	
Antsiranana	Madagascar	0	54,808	nan		
Fianarantsoa	Madagascar	0	101,428	nan	Yes	
Mahajanga	Madagascar	0	99,126	nan		
Toliara	Madagascar	0	75,032	nan		
Blantyre	Malawi	1	372,552	738,274		
Lilongwe	Malawi	1	268,767	799,762		
Mzuzu	Malawi	1	59,752	159,233		
Zomba	Malawi	1	48,517	99,277		
Bamako	Mali	0	758,125	2,452,195		
Gao	Mali	0	54,413	99,059		
Kayes	Mali	0	55,029	149,909		
Koutiala	Mali	0	55,163	167,010		
Mopti	Mali	0	76,285	134,933		
San	Mali	0	34,466	73,915		
Sikasso	Mali	0	87,024	261,123		
Ségou	Mali	0	92,519	188,365		
Tombouctou	Mali	0	31,338	64,488		
Kaédi	Mauritania	0	31,104	47,803		
Nouadhibou	Mauritania	0	61,209	113,789		
Nouakchott	Mauritania	0	418,294	938,154		
Rosso	Mauritania	0	30,530	50,861		
Oshakati	Namibia	1	34,552	83,432	Yes	
Windhoek	Namibia	1	140,410	358,996		
Arlit	Niger	0	36,261	78,651		
Birni-N'Konni	Niger	0	31,023	63,169		
Maradi	Niger	0	115,144	292,762		
Niamey	Niger	0	427,540	978,029		
Tahoua	Niger	0	52,951	117,826		
Zinder	Niger	0	126,517	235,605		
Aba	Nigeria	1	444,346	1,091,560		
Abakaliki	Nigeria	1	158,289	439,893	Yes	
Abraka	Nigeria	1	119,940	259,762		
Abuja	Nigeria	1	384,364	3,028,556		
Ado-Ekiti	Nigeria	1	291,866	647,182		Yes
Afikpo	Nigeria	1	74,524	141,516	Yes	
Agbor	Nigeria	1	67,857	129,551		
Aiyetoro	Nigeria	1	43,862	49,195	Yes	
Ajaokuta	Nigeria	1	57,702	82,522		
Akure	Nigeria	1	356,210	675,366	Yes	Yes
Akwanga	Nigeria	1	41,705	91,050	Yes	
Ankpa	Nigeria	1	39,291	70,006		
Argungu	Nigeria	1	40,367	87,700		
Auchi	Nigeria	1	72,986	147,505		

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Azare	Nigeria	1	65,234	124,820	Yes	
Bama	Nigeria	1	64,076	107,727		
Bauchi	Nigeria	1	232,939	435,001	Yes	
Bida	Nigeria	1	85,084	233,626		
Birnin-Kebbi	Nigeria	1	142,795	347,188		Yes
Biu	Nigeria	1	49,067	105,096		
Calabar	Nigeria	1	159,490	436,394		Yes
Damaturu	Nigeria	1	36,386	85,027	Yes	
Doma	Nigeria	1	42,091	83,383		
Dutse	Nigeria	1	152,198	193,025		
Egbe	Nigeria	1	34,188	89,210		
Egume	Nigeria	1	71,733	133,130		
Ejigbo	Nigeria	1	31,525	92,402		
Ekehen	Nigeria	1	30,566	57,101		
Emure-Ekiti	Nigeria	1	67,364	78,826		
Enugu	Nigeria	1	503,384	912,182	Yes	
Funtua	Nigeria	1	89,954	183,064	Yes	
Ganye	Nigeria	1	58,710	102,167		
Gashua	Nigeria	1	52,963	82,391		
Gboko	Nigeria	1	184,658	362,100	Yes	
Gombe	Nigeria	1	191,795	372,804	Yes	
Gusau	Nigeria	1	135,788	242,556		
Hadejia	Nigeria	1	45,276	94,181	Yes	
Ibadan	Nigeria	1	1,711,452	2,911,228	Yes	
Idah	Nigeria	1	82,520	161,370		
Idanre	Nigeria	1	49,885	97,053		
Ife	Nigeria	1	263,879	491,656		
Igbo-Ora	Nigeria	1	31,519	76,914		
Igboho	Nigeria	1	31,854	62,311		
Ihiala	Nigeria	1	96,474	nan		
Ikare	Nigeria	1	147,132	364,228		
Ikirun	Nigeria	1	215,476	427,992		
Ikole	Nigeria	1	56,932	100,183		
Ikom	Nigeria	1	40,718	52,109		
Ikot-Ekpene	Nigeria	1	146,477	nan	Yes	
Ikot-Etim	Nigeria	1	87,282	165,044		
Ila	Nigeria	1	43,213	59,975		
Ilesha	Nigeria	1	139,202	332,008		Yes
Ilorin	Nigeria	1	538,446	833,589		
Ilutitun	Nigeria	1	45,214	70,917		
Iseyin	Nigeria	1	47,732	174,531		
Iwo	Nigeria	1	88,314	240,838		
Jalingo	Nigeria	1	83,219	176,451	Yes	
Jega	Nigeria	1	32,799	69,227		
Jibia	Nigeria	1	35,397	56,556		
Jimeta	Nigeria	1	238,746	567,818		
Jos	Nigeria	1	487,013	789,950		Yes
Kaduna	Nigeria	1	849,035	1,139,643	Yes	
Kafanchan	Nigeria	1	41,236	132,111		
Kano	Nigeria	1	1,385,370	3,734,597		

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Katsina	Nigeria	1	189,505	425,669		
Katsina-Ala	Nigeria	1	43,751	74,895	Yes	
Kontagora	Nigeria	1	60,584	108,312	Yes	
Lafia	Nigeria	1	152,660	312,263		
Lagos	Nigeria	1	6,327,849	14,564,075		
Langtang	Nigeria	1	65,532	121,295		
Lokoja	Nigeria	1	63,547	375,656	Yes	Yes
Maiduguri	Nigeria	1	490,729	694,554	Yes	
Makurdi	Nigeria	1	179,494	301,249	Yes	
Malumfashi	Nigeria	1	46,775	58,968	Yes	
Maya-Belwa	Nigeria	1	30,627	42,151		
Michika	Nigeria	1	48,163	74,898		
Minna	Nigeria	1	98,628	459,441	Yes	Yes
Mubi	Nigeria	1	80,666	127,945		
Nasarawa	Nigeria	1	30,873	57,046	Yes	
New-Bussa	Nigeria	1	40,675	83,317		
Nguru	Nigeria	1	44,872	103,062		
Nkume	Nigeria	1	129,318	nan		
Nsukka	Nigeria	1	638,402	1,918,146		Yes
Numan	Nigeria	1	72,049	77,368		
Obudu	Nigeria	1	59,422	167,241		
Ogbomosho	Nigeria	1	134,065	383,364		
Oguma	Nigeria	1	35,039	72,981		
Ogwashi-Uku	Nigeria	1	42,955	67,482		
Okeho	Nigeria	1	41,304	105,183		
Okenne	Nigeria	1	85,307	376,128	Yes	
Okigwi	Nigeria	1	33,699	83,387		
Okitipupa	Nigeria	1	68,819	113,745		
Okpakeke	Nigeria	1	31,662	58,191		
Okpo	Nigeria	1	30,700	59,740		
Omu-Aran	Nigeria	1	47,679	81,069		
Omuo-Ekiti	Nigeria	1	31,118	99,172		
Ondo	Nigeria	1	228,481	426,176		
Onitsha	Nigeria	1	956,207	8,290,100		
Ore	Nigeria	1	45,689	102,651		
Oro-Esie-Iludin	Nigeria	1	46,096	75,454		
Osogbo	Nigeria	1	497,049	774,670		Yes
Otun-Ekiti	Nigeria	1	33,762	41,416		
Oturkpo	Nigeria	1	79,827	147,733		
Owo	Nigeria	1	103,021	186,305		
Oye-Ekiti	Nigeria	1	60,751	80,981		
Oyo	Nigeria	1	188,026	363,371		
Potiskum	Nigeria	1	46,192	241,243	Yes	
Saki	Nigeria	1	74,705	253,572		
Shendam	Nigeria	1	34,042	42,405		
Sokoto	Nigeria	1	310,603	606,753		Yes
Takum	Nigeria	1	31,065	53,909		
Uba	Nigeria	1	55,350	70,447		
Ugep	Nigeria	1	34,279	149,847		
Umuahia	Nigeria	1	116,720	nan		

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Uromi	Nigeria	1	182,758	365,049		
Uyo	Nigeria	1	197,529	2,513,616		
Vande-Ikya	Nigeria	1	35,671	64,535		
Wukari	Nigeria	1	43,003	83,693		
Yelwa	Nigeria	1	35,055	72,400		
Zaki-Biam	Nigeria	1	54,169	83,361		
Zaria	Nigeria	1	375,845	747,127		
Zuru	Nigeria	1	49,083	110,647		
Brazzaville	Republic of Congo	0	731,625	1,652,847	Yes	
Dakar	Senegal	0	1,975,856	3,435,250	Yes	
Diourbel	Senegal	0	79,063	104,578		
Kaolack	Senegal	0	153,840	199,066		
Kolda	Senegal	0	36,624	71,134		
Richard-Toll	Senegal	0	36,610	67,954		
Saint-Louis	Senegal	0	118,992	188,160		
Tambacounda	Senegal	0	44,844	90,956		
Touba-Mosquée	Senegal	0	168,853	781,727		
Ziguinchor	Senegal	0	128,061	168,198		
Bo	Sierra Leone	1	76,138	220,890		
Freetown	Sierra Leone	1	561,004	1,049,768		Yes
Kenema	Sierra Leone	1	66,406	187,158		
Makeni	Sierra Leone	1	48,170	108,671		
Torgbonbu	Sierra Leone	1	95,889	98,014		
Ad-Damazin	Sudan	1	58,786	255,340		
Ad-Duwaym	Sudan	1	53,580	79,009		
Al-Fashir	Sudan	1	130,226	244,208		
Al-Junaynah	Sudan	1	80,450	229,835		
Al-Manaqil	Sudan	1	60,108	111,669		
An-Nuhud	Sudan	1	52,539	69,668		
Atbara	Sudan	1	121,082	330,905		
Bur-Sudan	Sudan	1	293,338	421,429	Yes	Yes
El-Duein	Sudan	1	64,709	161,998		
El-Obeid	Sudan	1	211,433	384,829	Yes	
Gedaref	Sudan	1	178,488	295,201		
Kaduqli	Sudan	1	61,151	68,492		
Kassala	Sudan	1	223,586	318,335	Yes	
New-Halfa	Sudan	1	52,391	66,386		
Nyala	Sudan	1	194,574	606,114		
Sannar	Sudan	1	58,718	266,989		
Mbabane	Swaziland	1	82,878	nan	Yes	
Tabankulu	Swaziland	1	30,730	nan		
Arusha	Tanzania	1	122,068	416,442		
Bukoba	Tanzania	1	31,826	128,796		
Dar-es-Salaam	Tanzania	1	1,333,413	4,520,658	Yes	Yes
Dodoma	Tanzania	1	90,565	213,636	Yes	
Kigoma	Tanzania	1	80,568	215,458		
Lindi	Tanzania	1	39,534	78,841		
Mbeya	Tanzania	1	144,556	385,279		Yes
Mtwara	Tanzania	1	68,149	100,626		
Musoma	Tanzania	1	68,356	134,327		

Continued on next page

Table C.1: List of African cities in the sample

City name	Country	Anglophone dummy	Projected population 1990	Projected population 2012	Colonial origin sample	40 cities sample
Mwanza	Tanzania	1	193,317	706,453		
Shinyanga	Tanzania	1	49,960	103,795		
Singida	Tanzania	1	41,807	85,242		
Songea	Tanzania	1	57,908	203,309		
Sumbawanga	Tanzania	1	51,038	124,204		
Tabora	Tanzania	1	96,935	160,608		
Tanga	Tanzania	1	142,799	221,127		
Zanzibar	Tanzania	1	174,467	501,459		
Fort-Portal	Uganda	1	32,130	51,795		
Gulu	Uganda	1	34,535	146,233		
Kampala	Uganda	1	803,069	2,269,969		
Masaka	Uganda	1	47,671	112,864		
Mbale	Uganda	1	51,446	117,706		
Mbarara	Uganda	1	39,119	164,150		
Njeru	Uganda	1	96,824	219,039		
Soroti	Uganda	1	40,903	48,069		
Chipata	Zambia	1	52,213	128,045	Yes	
Choma	Zambia	1	30,143	54,492		
Kabwe	Zambia	1	154,318	207,909	Yes	
Kasama	Zambia	1	47,653	108,492		
Kitwe	Zambia	1	355,793	1,066,992	Yes	Yes
Livingstone	Zambia	1	76,875	143,249	Yes	
Luanshya	Zambia	1	118,143	133,187		
Lusaka	Zambia	1	813,154	2,000,916	Yes	
Mansa	Zambia	1	37,882	88,890		
Ndola	Zambia	1	329,228	468,324	Yes	
Bulawayo	Zimbabwe	1	611,307	653,337		
Chinhoyi	Zimbabwe	1	41,969	68,273		
Gweru	Zimbabwe	1	125,626	154,825	Yes	
Harare	Zimbabwe	1	1,405,753	2,133,801	Yes	
Hwange	Zimbabwe	1	44,297	19,870		
Kadoma	Zimbabwe	1	66,150	91,633	Yes	
Kwekwe	Zimbabwe	1	101,681	136,804	Yes	
Marondera	Zimbabwe	1	37,277	61,998	Yes	
Masvingo	Zimbabwe	1	48,780	87,886	Yes	
Mutare	Zimbabwe	1	124,697	186,208	Yes	
Zvishavane	Zimbabwe	1	32,571	45,230		

Notes: Two cities are only included in the 40 cities sample, but not included in the 333 cities full sample. They are Bimbo in Central African Republic, Libreville in Gabon.

Table C.2: Coefficients of geographic controls for openness index and area

	(1) Openness	(2) Area
Anglophone dummy	0.173***	0.285***
	(0.055)	(0.093)
Ln ruggedness	-0.062*	-0.172**
	(0.034)	(0.080)
Ln rainfall	-0.150***	-0.912***
	(0.049)	(0.119)
Ln elevation range	0.137***	0.358***
	(0.038)	(0.077)
Coast dummy	0.776	-0.234
	(1.184)	(3.173)
Interaction ln coast length	0.027	0.007
	(0.107)	(0.273)
Interaction ln distance to coast	0.098***	0.064
	(0.037)	(0.086)
Fraction of river area	-0.142	-0.263
	(0.679)	(1.030)
Fraction of lake area	-0.664	0.831
	(0.970)	(1.174)
Malaria index	-0.002	0.003
	(0.004)	(0.006)
Land suitability	0.235**	0.267
	(0.094)	(0.179)
Log temprature	-0.270	-1.458***
	(0.273)	(0.537)
Non-capital dummy	0.186**	0.038
	(0.090)	(0.164)
Distance to national capital	-0.000**	-0.001***
	(0.000)	(0.000)
ELF level 15 1975	-0.038	0.117
	(0.061)	(0.128)
Lag t-1 ln country GDP per capita	0.082	0.071
	(0.061)	(0.078)
Ln annual population growth 90 to 12	-0.522	8.453***
	(1.217)	(2.241)
Ln projected city population 1990	-0.174***	0.857***
	(0.025)	(0.042)
R^2	0.341	0.786
N	281	281

Note: Robust standard errors are applied.
* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.3: Balance test for 40 cities sample

	(1) Project city population 1990 (Thousands)	(2) Ln annual city population growth 90-12	(3) Rainfall	(4) Coast Dummy	(5) Elevation
Anglophone dummy	-214.631	-0.018	-7.533	-0.050	-1.969
	(163.192)	(0.108)	(19.921)	(0.152)	(143.559)
Constant	692.799***	0.882***	120.676***	0.350***	378.372***
	(125.665)	(0.066)	(15.943)	(0.109)	(99.903)
Adjusted R^2	0.018	-0.026	-0.022	-0.023	-0.026
N	40	40	40	40	40

Note: Robust standard errors are applied.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.4: Built-up cover intensity gradient 2014

	(1) Anglphone Cities	(2) Francophone Cities
Ring distance	-0.195***	-0.285***
	(0.011)	(0.035)
Ln ring total pixel	0.646***	0.664***
	(0.054)	(0.104)
Lag t-1 ln country GDP per capita	-0.039	-0.039
	(0.174)	(0.308)
Ln projected city population 1990	0.933***	1.194***
	(0.098)	(0.201)
Ln annual population growth 90 to 12	1.011	-4.779
	(4.467)	(6.302)
Ln ruggedness	0.001	0.557***
	(0.180)	(0.180)
Ln rainfall	0.521***	-0.542**
	(0.198)	(0.210)
Ln elevation range	0.038	-0.303
	(0.169)	(0.185)
Coast dummy	-0.331	-1.825
	(2.587)	(3.657)
Interaction ln coast length	0.011	-0.460
	(0.212)	(0.349)
Interaction ln distance to coast	-0.005	-0.655***
	(0.175)	(0.185)
Fraction of river area	-0.454	-3.752*
	(4.063)	(1.980)
Fraction of lake area	3.398	-1.006
	(4.174)	(5.017)
Malaria index	0.030	0.061***
	(0.019)	(0.015)
Land suitability	-0.975**	1.088**
	(0.454)	(0.414)
Log temprature	-1.932**	-5.473**
	(0.979)	(2.370)
Non-capital dummy	-0.740**	-0.558
	(0.312)	(0.516)
Distance to national capital	-0.001**	0.001
	(0.000)	(0.001)
ELF15 1975	0.161	-0.142
	(0.285)	(0.363)
R^2	0.479	0.551
N	2475	703

Note: Sample includes rings up to 20km. Standard errors are clustered at city level.
* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.5: Intensity of built pixels 2014

	(1) 1km	(2) 2km	(3) 3km	(4) 4km	(5) 5km	(6) 6km
Anglophone Dummy	-0.109*	-0.215***	-0.403***	-0.283**	-0.312**	-0.139
	(0.061)	(0.057)	(0.095)	(0.134)	(0.147)	(0.175)
Ln ring built pixel	1.489***	1.333***	1.361***	1.362***	1.324***	1.310***
	(0.057)	(0.045)	(0.054)	(0.056)	(0.042)	(0.048)
Anglophone mean	1.196	2.514	2.539	2.425	2.229	2.299
Francophone mean	1.463	3.153	3.024	2.569	2.586	2.501
Geographic and Stiational Controls	Yes	Yes	Yes	Yes	Yes	Yes
Economic Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.907	0.930	0.920	0.920	0.929	0.915
N	283	285	279	259	236	209

Note: Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.6: Coefficients of geographic controls for leapfrogging

	(1) Ln count of LF	(2) Ln LF minus ln total patches	(3) Ln avg. LF area
Anglophone dummy	0.539***	0.325***	-0.013
	(0.167)	(0.118)	(0.061)
Ln initial cover 1990	0.295***	-0.295***	0.022
	(0.058)	(0.040)	(0.020)
Year dummy 2014	0.495***	0.119**	0.137***
	(0.067)	(0.056)	(0.035)
Ln ruggedness	-0.190*	-0.204***	-0.034
	(0.107)	(0.076)	(0.032)
Ln rainfall	-0.279**	-0.053	-0.007
	(0.137)	(0.087)	(0.048)
Ln elevation range	0.223**	0.069	0.100***
	(0.110)	(0.082)	(0.036)
Coast dummy	-3.614	-3.707**	-1.156
	(2.854)	(1.572)	(0.980)
Interaction ln coast length	0.539**	0.409***	0.142*
	(0.221)	(0.122)	(0.083)
Interaction ln distance to coast	0.245**	0.113	0.046
	(0.116)	(0.080)	(0.037)
Fraction of river area	-0.365	-0.270	-0.929
	(1.716)	(1.436)	(0.786)
Fraction of lake area	-1.057	-1.455	-0.466
	(1.921)	(2.022)	(0.740)
Malaria index	0.030***	0.013*	-0.001
	(0.010)	(0.008)	(0.004)
Land suitability	0.375	-0.005	-0.112
	(0.291)	(0.193)	(0.106)
Log temprature	-3.506***	-1.736***	-0.040
	(0.795)	(0.542)	(0.282)
Non-capital dummy	0.012	-0.152	-0.044
	(0.203)	(0.148)	(0.091)
Distance to national capital	-0.001***	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)
ELF level 15 1975	-0.280	-0.159	0.004
	(0.184)	(0.131)	(0.068)
Lag t-1 ln country GDP per capita	-0.063	-0.087	0.059
	(0.144)	(0.117)	(0.050)
Ln annual population growth 90 to 12	12.114***	5.842**	2.215*
	(3.338)	(2.585)	(1.287)
Ln projected city population 1990	0.712***	0.443***	0.036
	(0.101)	(0.072)	(0.035)
R^2	0.602	0.230	0.109
N	551	551	525

Note: Standard errors are clustered at city level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.7: Leapfrogging other specifications

	Count of LF	Ln count of LF		Ln LF minus ln total
	(1)	(2)	(3)	(4)
	Poisson	OLS	Tobit	OLS
Anglophone dummy	0.661***	0.472***	0.577***	0.283***
	(0.018)	(0.122)	(0.140)	(0.091)
Ln initial cover 1990	0.299***	0.298***	0.325***	-0.272***
	(0.006)	(0.040)	(0.046)	(0.030)
Year dummy 2014	0.541***	0.502***	0.515***	0.135**
	(0.010)	(0.082)	(0.095)	(0.061)
Ln ruggedness	-0.136***	-0.130*	-0.203**	-0.129**
	(0.012)	(0.075)	(0.086)	(0.056)
Ln rainfall	-0.271***	-0.233**	-0.283**	-0.026
	(0.015)	(0.100)	(0.114)	(0.075)
Ln elevation range	0.180***	0.246***	0.222**	0.069
	(0.011)	(0.083)	(0.096)	(0.062)
Coast dummy	-7.481***	-5.206*	-3.788	-4.731**
	(0.412)	(2.669)	(3.156)	(1.991)
Interaction ln coast length	0.875***	0.571**	0.582**	0.415**
	(0.035)	(0.246)	(0.292)	(0.183)
Interaction ln distance to coast	0.248***	0.140	0.268***	0.026
	(0.012)	(0.086)	(0.099)	(0.064)
Fraction of river area	-5.020***	0.898	-0.410	1.007
	(0.303)	(1.752)	(1.982)	(1.307)
Fraction of lake area	1.460***	-1.520	-1.064	-2.043
	(0.245)	(1.673)	(1.963)	(1.248)
Malaria index	0.040***	0.022**	0.034***	0.006
	(0.001)	(0.009)	(0.010)	(0.007)
Land suitability	0.130***	0.115	0.423*	-0.165
	(0.028)	(0.220)	(0.255)	(0.164)
Log temprature	-2.680***	-2.877***	-3.793***	-1.189**
	(0.070)	(0.723)	(0.824)	(0.539)
Non-capital dummy	-0.120***	0.001	0.044	-0.168
	(0.019)	(0.232)	(0.272)	(0.173)
Distance to national capital	-0.001***	-0.001***	-0.001***	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)
ELF level 15 1975	0.026	-0.299**	-0.276*	-0.168
	(0.019)	(0.145)	(0.166)	(0.108)
Lag t-1 ln country GDP per capita	0.064***	-0.057	-0.066	-0.049
	(0.012)	(0.112)	(0.127)	(0.083)
Ln annual population growth 90 to 12	5.178***	12.764***	12.200***	6.481***
	(0.278)	(2.558)	(2.948)	(1.908)
Ln projected city population 1990	0.532***	0.628***	0.705***	0.365***
	(0.008)	(0.076)	(0.087)	(0.056)
Constant	-1.956***	-0.867	-0.494	2.098
	(0.280)	(2.806)	(3.231)	(2.094)
R^2		0.633		0.231
N	551	525	551	525

Note: Columns 2 and 4 show OLS results excluding cities with zero LF patches. Standard errors are clustered at country year level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.8: Balance test for leapfrogging regressions

	(1) Full sample	(2) Border sample	(3) Border sample
Ln initial cover 1990	-.282	-.5	-.119
	(.187)	(.311)	(.369)
Ln projected city population 1990	.115	-.104	.038
	(.116)	(.253)	(.282)
Ln annual population growth 90 to 12	.003	.004	.001
	(.002)	(.004)	(.004)
Ln ruggedness	.581***	.418	.011
	(.156)	(.28)	(.137)
Ln rainfall	.037	-.011	.129***
	(.081)	(.137)	(.041)
Ln elevation range	.29***	.019	.022
	(.084)	(.19)	(.174)
Coast dummy	-.041	.042	.1
	(.028)	(.065)	(.089)
Fraction of river area	-.007**	-.006	-.004
	(.003)	(.007)	(.004)
Fraction of lake area	.001	.001	.001
	(.003)	(.002)	(.002)
Interaction ln coast length	-.428	.49	1.116
	(.292)	(.711)	(.971)
City cluster FE	No	No	Yes
Observations	318	58	58

Note: Border sample does not include cities in Anglophone and Francophone Cameroon border area. Robust standard errors are applied in column 1 and 2, and are clustered at city level in column 2 and 3.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.9: Openness and area: Robustness

	(1) Base	(2) Distance trim	(3) No German colonies	(4) No Nigeria	(5) Colonial origin 1885	(6) Non capital	(7) 40 cities
Ln openness index 2014							
Anglophone dummy	0.173***	0.159***	0.188**	0.156**	0.332***	0.166***	0.302**
	(0.055)	(0.057)	(0.077)	(0.069)	(0.078)	(0.063)	(0.134)
Ln area							
Anglophone dummy	0.285***	0.278***	0.227*	0.551***	0.104	0.260***	0.352
	(0.093)	(0.098)	(0.120)	(0.120)	(0.318)	(0.098)	(0.301)
R^2	0.786	0.717	0.790	0.821	0.693	0.727	0.481
N	281	266	248	172	54	261	26

Note: Columns 1-4 and 6 include the same controls as in column 4 and 8 of Table 1. Columns 5 and 7 include the controls of ln country GDP per capita in 2000, ln projected city population in 1990, ln average ruggedness, and coast dummy. Columns 5 includes ln annual urban population growth 90 to 14 in country level due to severe missing data problem in city population in the colonial origin sample. Column 7 includes ln annual population growth 90 to 12 in city level. Robust standard errors are applied. R^2 and N are reported for ln area.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

C.3.3 Additional Analysis

C.3.3.1 The So-What Question: Public Policy Relevance

The planning literature argues that having less compact and more irregularly laid out cities raises the cost of infrastructure provision. In Africa we further argue that such higher costs will lower the likelihood of receiving public infrastructure provision. To assess the reduced form impact, we use data from the Demographic and Health Survey (DHS) on whether a family has a piped water connection (with the alternatives being a shallow or deeper well or having no water connection), an electricity connection, a telephone land line connection, or a (flush) toilet connected to a public sewer system. About 63% and 75% of households in our sample are connected to water and electricity respectively, while having a flush toilet connected to a sewer system or having a landline are at 13% and 6% respectively. We tend to focus on the first 3 outcomes.

DHS uses cluster sampling of 20-30 households in a neighbourhood and we restrict attention to clusters defined by DHS to be in urban areas. We cover about 45,000 households in 60 Francophone cities in 7 countries and 133 Anglophone cities in 11 countries. We constrain DHS surveys to be within 2 years each of 2000 and 2014, the base years for which we measure openness and leapfrogging, and we control for which year before and after the base years a survey is.

The specification differs in two ways relative to equation (1), apart from the outcomes involving infrastructure connections and the specification being a LPM. First, the treatment variables are measures of leapfrogging and openness in the local area around where a household lives. Second identification is based on city-year fixed effects, or within city variation in servicing. We are not measuring an Anglophone (versus Francophone) effect per se, since such colonial influences may reflect other public policy elements. Note colonial origin would be neither a valid instrument nor reduced form measure. We are trying to represent a cost effect of lack of compactness on public service provision. However we can and do look at whether there is a differential in impact of openness and leapfrogging from being in a Francophone versus an Anglophone city. Note we will examine the effects of both leapfrogging and openness measures on the likelihood of a connection. While the two measures are related, greater openness might indicate greater distance on average between residences, while leapfrogging indicates that some developments are scattered. Both measures will matter and both suggest connecting households will require more piping, wiring or lines per household entailing more costs.

The challenge in implementation concerns location. Within an urban area, cluster locations are randomized within 2 km by randomly picking a directional ray (angle) from the true

cluster centre and then choosing a location randomly along that ray within 2 km of the cluster centre. Under this algorithm, while locations near the true location are more likely to be chosen, the randomized location is equally likely to be in any ring out from the true location up to the 2 km. We draw a 2 km circle around the specified location and look at the effect of more leapfrog patches and openness in that circle on provision, conditional on far it is from the city centre, and other controls. One could view this as a measure of, say, how likely a cluster is to be in a leapfrog patch, but we interpret it as a measure of the overall degree of leapfrogging and openness in the surrounding area. To exploit economies of scale in construction, cities roll out public utilities in large spatial zones. The higher the degree of leapfrog and open development in an area, the less likely it is to be serviced, because roll-out is more costly. Regardless, because of the randomization of location, the variables of interest are measured with error. We could not think of an instrument which both met the exclusion restriction and had power.[10] Given the measurement error involved we did not anticipate seeing strong patterns in the data, but were surprised.

Results are in Table C.10 covering 42,748-44,561 households for each attribute. In a linear probability formulation, each attribute has two columns. In the table's reduced form specification, both columns have basic supply and demand controls, with the count of leapfrog developments as the cost factor in the first and that and the share of built cover in the surrounding neighborhood (lack of openness) in the second. An increased count of leapfrog patches significantly reduces the likelihood of connections in 3 of the 4 outcomes, having no effect on piped water. Effects are fairly small. A one standard deviation (5.6) increase in count of LF patches reduces the probability of an electricity connection by 0.018 from a mean of 0.74, although for flush toilets connected to a sewer system the decrease is 0.025 from a mean of 0.12. In the second set of columns we add in the extent of built cover, which has significant impacts on all connections except a landline. Again effects are fairly small. A one standard deviation (0.31) in built cover increases the likelihood of an electricity connection by 0.016, water by 0.025 (from a mean of 0.64), and flush toilet connected to a sewer system by 0.036. In Table C.11 we examine whether for either leapfrogging or openness, there is a differential Anglophone effect. In only one of the eight cases (built cover for flush toilets to a sewer system) is the Anglophone interaction term significant. Overall, effects are consistent, suggesting that, despite measurement error, we indeed find a negative relationship between increased sprawl and public utility provision in both Anglophone and Francophone cities.

[10] e.g. Using the propensity of surrounding areas to have LF development for whether cluster is recorded in a leapfrog patch does not.

Table C.10: Public utility connection

	Has electricity		Has piped water		Has flush toilet		Has phone land line	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Count of LF	-0.0033**	-0.0026*	-0.0002	0.0008	-0.0045***	-0.0031***	-0.0019**	-0.0020**
	(0.0014)	(0.0014)	(0.0013)	(0.0014)	(0.0010)	(0.0010)	(0.0008)	(0.0009)
Share of built cover		0.0526*		0.0797**		0.1168***		-0.0026
in buffer		(0.0298)		(0.0364)		(0.0266)		(0.0128)
Ln buffer center	-0.0727***	-0.0666***	-0.0700***	-0.0608***	-0.0383***	-0.0248***	-0.0124***	-0.0127***
distance	(0.0066)	(0.0075)	(0.0088)	(0.0096)	(0.0062)	(0.0073)	(0.0035)	(0.0038)
Ln buffer ruggedness	0.0553***	0.0507***	0.0565***	0.0495**	-0.0054	-0.0157	-0.0012	-0.0010
	(0.0165)	(0.0171)	(0.0204)	(0.0210)	(0.0120)	(0.0124)	(0.0062)	(0.0064)
Buffer has river of	0.0209	0.0232	0.0361	0.0395	-0.0057	-0.0007	0.0041	0.0040
lake	(0.0214)	(0.0214)	(0.0272)	(0.0270)	(0.0174)	(0.0175)	(0.0122)	(0.0121)
Household size	0.0086***	0.0086***	0.0024***	0.0024***	0.0026***	0.0026***	0.0069***	0.0069***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
Sex of household	-0.0075	-0.0076	-0.0092*	-0.0094**	-0.0159***	-0.0161***	-0.0096***	-0.0096***
head: Male	(0.0050)	(0.0050)	(0.0047)	(0.0047)	(0.0041)	(0.0040)	(0.0028)	(0.0028)
Highest educational	0.0367***	0.0374***	0.0117	0.0128*	-0.0052	-0.0036	0.0126***	0.0126***
level of head: Primary	(0.0072)	(0.0072)	(0.0073)	(0.0073)	(0.0045)	(0.0045)	(0.0034)	(0.0034)
Highest educational	0.1739***	0.1740***	0.0519***	0.0521***	0.0407***	0.0410***	0.0424***	0.0424***
level of head: Secondary	(0.0070)	(0.0070)	(0.0071)	(0.0071)	(0.0047)	(0.0047)	(0.0038)	(0.0038)
Highest educational	0.2877***	0.2877***	0.0578***	0.0577***	0.1652***	0.1652***	0.1124***	0.1124***
level of head: Higher	(0.0088)	(0.0088)	(0.0094)	(0.0094)	(0.0089)	(0.0089)	(0.0072)	(0.0072)
Highest educational	0.1417***	0.1417***	0.0457**	0.0457**	0.0124	0.0124	0.0027	0.0026
level of head: Unknown	(0.0247)	(0.0247)	(0.0225)	(0.0225)	(0.0165)	(0.0165)	(0.0150)	(0.0150)
Period dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
City FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.320	0.321	0.436	0.437	0.255	0.259	0.126	0.126
N	44517	44517	44561	44561	44500	44500	42748	42748

Note: Period dummies control the difference between the DHS survey years with year 2000 and 2014 when satellite data is available. Period dummies include 1 year before dummy, 1 year after dummy, 2 years before dummy, and 2 years after dummy. Standard errors are clustered at DHS cluster level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table C.11: Public utility connection: Interaction effects

	(1) Has electricity	(2) Has piped water	(3) Has flush toilet	(4) Has phone land line
Share of built cover	0.039	0.052	0.005	-0.011
in buffer	(0.043)	(0.047)	(0.028)	(0.017)
Count of LF	0.000	-0.003	-0.001	-0.001*
	(0.004)	(0.005)	(0.001)	(0.001)
Anglophone ×	0.028	0.041	0.197***	0.015
Share of built cover in buffer	(0.055)	(0.070)	(0.045)	(0.023)
Anglophone ×	-0.003	0.004	-0.002	-0.001
Count of LF	(0.004)	(0.005)	(0.002)	(0.001)
Ln buffer center	-0.066***	-0.060***	-0.021***	-0.012***
distance	(0.008)	(0.010)	(0.008)	(0.004)
Ln buffer ruggedness	0.051***	0.052**	-0.008	-0.001
	(0.017)	(0.021)	(0.013)	(0.006)
Buffer has river of	0.023	0.040	0.001	0.004
lake	(0.021)	(0.027)	(0.017)	(0.012)
Household size	0.009***	0.002***	0.003***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)
Sex of household	-0.008	-0.009**	-0.016***	-0.010***
head: Male	(0.005)	(0.005)	(0.004)	(0.003)
Highest educational	0.037***	0.013*	-0.004	0.013***
level of head: Primary	(0.007)	(0.007)	(0.004)	(0.003)
Highest educational	0.174***	0.052***	0.040***	0.042***
level of head: Secondary	(0.007)	(0.007)	(0.005)	(0.004)
Highest educational	0.287***	0.057***	0.164***	0.112***
level of head: Higher	(0.009)	(0.009)	(0.009)	(0.007)
Highest educational	0.141***	0.046**	0.012	0.003
level of head: Unknown	(0.025)	(0.023)	(0.017)	(0.015)
Periods dummies	Yes	Yes	Yes	Yes
City fixed effect	Yes	Yes	Yes	Yes
Adjusted R^2	0.321	0.437	0.262	0.126
N	44517	44561	44500	42748

Note: Period dummies control the difference between the DHS survey years with year 2000 and 2014 when satellite data is available. Period dummies include 1 year before dummy, 1 year after dummy, 2 years before dummy, and 2 years after dummy. Standard errors are clustered at DHS cluster level.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

D.4 Valuing the Environmental Benefits of Canals Using House Prices

Table D.1: Definitions of variables used and the respective data sources

Variable	Source	Definition
Dependent Variable		
Ln Price	Land Registry	Natural Log of Transacted House Price
Structure		
Floor area (m2)	Land Registry	Size of transacted unit
Number rooms	Land Registry	Number of rooms in transacted unit
Number heated rms	Land Registry	Number of heated rooms in transacted unit
Fireplace	Land Registry	Binary variable =1 if transacted unit has a fireplace, = 0 otherwise
Energy efficiency	Land Registry	Overall Energy Efficiency (scaled from 1-100)
Leasehold	Land Registry	Binary variable =1 if transacted unit is leasehold, = 0 otherwise
Tenure missing	Land Registry	Binary variable =1 if tenure variable is missing, = 0 otherwise
New	Land Registry	Binary variable =1 if transacted unit is new build, = 0 otherwise
Terraced	Land Registry	Binary variable =1 if transacted unit is terrace house, = 0 otherwise
Flat	Land Registry	Binary variable =1 if transacted unit is flat, = 0 otherwise
Semi-detached	Land Registry	Binary variable =1 if transacted unit is semi-detached, = 0 otherwise
Land cover		
Arable	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on arable land, = 0 otherwise
Freshwater	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on freshwater, = 0 otherwise
Improved grass	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on improved grassland, = 0 otherwise
Urban	Ordnance Survey	Binary variable =1 if the centroid of the postcode is urban land, = 0 otherwise
Heather, bog, rock	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on heather, bog or rock land, = 0 otherwise
Grassland	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on grassland, = 0 otherwise
Sediment/marsh	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on sediment or marsh land, = 0 otherwise
Woodland	Ordnance Survey	Binary variable =1 if the centroid of the postcode is on woodland, = 0 otherwise
Dist to greenspace	UCL	Euclidean distance from the nearest green space
Green area	UCL	Size of nearest green space
Other river		
100m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is below 100m, = 0 otherwise
200m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is between 100m and 200m, = 0 otherwise
400m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is between 200m and 400m, = 0 otherwise
800m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is between 400m and 800m, = 0 otherwise
1600m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is between 800m and 1600m, = 0 otherwise
3200m	Ordnance Survey	Binary variable =1 if the distance to the nearest other river is between 1600m and 3200m, = 0 otherwise
Other distances (km)		
Dist lakes	Ordnance Survey	Euclidean distance from the nearest lake
Lakes >10km	Ordnance Survey	Binary variable =1 if the distance to the nearest lake is above 10km, = 0 otherwise

Continued on next page

Table D.1: Definitions of variables used and the respective data sources

Variable	Source	Definition
Dist docks	Canal Trust	Euclidean distance from the nearest dock
Docks > 10km	Canal Trust	Binary variable =1 if the distance to the nearest dock is above 10km, = 0 otherwise
Dist bridges	Canal Trust	Euclidean distance from the nearest bridge
Bridges > 10km	Canal Trust	Binary variable =1 if the distance to the nearest bridge is above 10km, = 0 otherwise
Dist embankments	Canal Trust	Euclidean distance from the nearest embankment
Embankments > 10km	Canal Trust	Binary variable =1 if the distance to the nearest embankment is above 10km, = 0 otherwise
Dist reservoirs	Canal Trust	Euclidean distance from the nearest reservoir
Reservoirs > 10km	Canal Trust	Binary variable =1 if the distance to the nearest reservoir is above 10km, = 0 otherwise
Dist rapid rail	Ordnance Survey	Euclidean distance from the nearest rapid rail line
Rapid rail > 10km	Ordnance Survey	Binary variable =1 if the distance to the nearest rapid rail lines above 10km, = 0 otherwise
Dist railways	Ordnance Survey	Euclidean distance from the nearest railway
Railways > 10km	Ordnance Survey	Binary variable =1 if the distance to the nearest railway is above 10km, = 0 otherwise
Dist town centre	UCL	Euclidean distance from the nearest town centre
Town centre > 10km	UCL	Binary variable =1 if the distance to the nearest town centre is above 10km, = 0 otherwise
Dist rail stations	Ordnance Survey	Euclidean distance from the nearest rail station
Rail stations > 10km	Ordnance Survey	Binary variable =1 if the distance to the nearest rail station is above 10km, = 0 otherwise
Dist rapid rail stat.	Ordnance Survey	Euclidean distance from the nearest rapid rail station
Rapid rail stat > 10km	Ordnance Survey	Binary variable =1 if the distance to the nearest rapid rail station is above 10km, = 0 otherwise
Dist outfall	Canal Trust	Euclidean distance from the nearest outfall
Outfall > 10km	Canal Trust	Binary variable =1 if the distance to the nearest outfall is above 10km, = 0 otherwise
Employment		
Total employment (1000s)	Nomis	Number of employment in thousands
No employment	Nomis	Number of unemployed
Agriculture share	Nomis	Share of employment in agriculture in LSOA
Mining utilities share	Nomis	Share of employment in mining utilities share in LSOA
Manufacturing share	Nomis	Share of employment in manufacturing in LSOA
Construction share	Nomis	Share of employment in construction in LSOA
Motor industry share	Nomis	Share of employment in motor industry in LSOA
Wholesale share	Nomis	Share of employment in wholesale in LSOA
Retail share	Nomis	Share of employment in retail in LSOA
Transport share	Nomis	Share of employment in transport in LSOA
Accom/food share	Nomis	Share of employment in accommodation and food services in LSOA
Financial services share	Nomis	Share of employment in financial service in LSOA
Property services share	Nomis	Share of employment in property service in LSOA
Prof, science, tech share	Nomis	Share of employment in professional, science and technical activities in LSOA
Business admin share	Nomis	Share of employment in business administration in LSOA
Public admin share	Nomis	Share of employment in public administration in LSOA
Education share	Nomis	Share of employment in education in LSOA
Health share	Nomis	Share of employment in health in LSOA
Arts entertainment share	Nomis	Share of employment in arts, entertainment and recreation in LSOA
IT share	Nomis	Share of employment in information and communication in LSOA
Demographics		
No education share	Census 2001	Share of residents with no education qualifications

Continued on next page

Table D.1: Definitions of variables used and the respective data sources

Variable	Source	Definition
No car share	Census 2001	Share of households without cars
Unemployment rate	Census 2001	Share of unemployed for the economically active
Lone parent household share	Census 2001	Share of single parent households
Non EU residents share	Census 2001	Share of non EU residents
Social renters share	Census 2001	Share of households who are social renters
Owners share	Census 2001	Share of households who are property owners
Non-white share	Census 2001	Share of non white residents
Population	Census 2001	Population size
Population density	Census 2001	Population size per unit area

Table D.2: Descriptive statistics for the England and Wales sample

	0-100 metres		100-1000 metres		1000-1500 metres	
	mean	sd	mean	sd	mean	sd
Dependent Variable						
Natural log of price	11.867	0.618	11.815	0.673	11.886	0.667
Price	175685.100	236664.800	173885.900	245630.600	187898.700	273691.900
House Structure Controls						
Price per metre squared	2313.448	3251.735	2111.266	13422.470	2231.633	3303.369
Size(sqm)	79.684	39.543	86.250	40.247	87.285	42.968
No.of Rooms	4.036	1.569	4.465	1.535	4.504	1.565
Fireplace	0.115	0.308	0.149	0.344	0.149	0.345
Energy Efficiency	64.070	13.801	59.564	13.157	59.870	12.827
Freehold	0.545	0.498	0.740	0.438	0.748	0.434
New Built	0.169	0.375	0.066	0.248	0.061	0.239
Terrace House	0.306	0.461	0.381	0.486	0.351	0.477
Flats	0.368	0.482	0.165	0.371	0.166	0.372
Semi-Detached	0.194	0.395	0.292	0.455	0.307	0.461
Location Controls						
Distance to rail (m)	785.977	1071.971	941.253	1135.310	1128.576	1167.526
Distance to town centre (m)	1582.133	1640.985	1665.240	1655.186	1653.508	1583.840
Distance to rail station (m)	1626.303	1800.961	1808.862	1870.892	1934.246	1835.618
Other river 100m	0.192	0.394	115577.000	0.065	0.246	1168074.000
Other river 200-100m	0.202	0.401	115577.000	0.096	0.295	1168074.000
Other river 200-400m	0.220	0.414	115577.000	0.219	0.414	1168074.000
Other river 400-800m	0.238	0.426	115577.000	0.370	0.483	1168074.000
Other river 800-1600m	0.120	0.325	115577.000	0.210	0.407	1168074.000
Other river 1600-3200m	0.027	0.163	115577.000	0.037	0.190	1168074.000
Arable	0.003	0.052	0.003	0.054	0.003	0.057
Freshwater	0.002	0.043	0.000	0.022	0.001	0.029
Improved grass	0.018	0.131	0.021	0.143	0.024	0.153
Suburban	0.516	0.500	0.653	0.476	0.686	0.464
Urban	0.448	0.497	0.317	0.465	0.280	0.449
Heather bog rock	0.000	0.000	0.000	0.004	0.000	0.004
Grassland	0.001	0.034	0.001	0.031	0.001	0.032
Sediment marsh	0.000	0.005	0.000	0.009	0.000	0.010
Woodland	0.012	0.110	0.005	0.071	0.005	0.070
Distance to green space	191.093	184.947	172.989	137.552	176.873	141.505
Area of nearest green space (m2)	58174.780	262329.900	76117.790	574594.600	64821.440	346820.000
Employment Controls						
Total employment	1655.350	3172.158	1144.229	2852.941	1038.609	3386.612

Continued on next page

Table D.2: Descriptive statistics for the England and Wales sample

	0-100 metres		100-1000 metres		1000-1500 metres	
	mean	sd	mean	sd	mean	sd
No employment	0.000	0.013	0.001	0.027	0.002	0.046
Non-farm agriculture	0.002	0.012	0.001	0.011	0.001	0.009
Mining & utilities	0.010	0.038	0.008	0.037	0.007	0.035
Manufacturing	0.095	0.136	0.086	0.142	0.065	0.128
Construction	0.067	0.079	0.077	0.101	0.080	0.106
Motot	0.023	0.040	0.023	0.051	0.020	0.045
Wholesale	0.047	0.069	0.041	0.075	0.036	0.073
Retail	0.107	0.121	0.101	0.125	0.100	0.132
Transport	0.046	0.087	0.046	0.093	0.039	0.093
Accommodation & food	0.083	0.094	0.075	0.096	0.070	0.095
Financial and insurance	0.019	0.057	0.013	0.052	0.013	0.052
Property	0.018	0.040	0.015	0.040	0.016	0.040
Prof science technical	0.079	0.088	0.077	0.096	0.081	0.099
Communications	0.036	0.051	0.034	0.065	0.035	0.063
Business admin	0.077	0.113	0.068	0.101	0.068	0.104
Public admin	0.024	0.072	0.021	0.071	0.021	0.077
Education	0.094	0.161	0.126	0.188	0.141	0.205
Health	0.117	0.145	0.134	0.173	0.147	0.185
Arts and entertainment	0.055	0.080	0.053	0.078	0.056	0.084
Neighbourhood Controls						
Low qualifications	0.275	0.133	0.296	0.130	0.289	0.128
Households no car	0.287	0.180	0.278	0.170	0.267	0.171
Unemployment Rate	0.055	0.047	0.055	0.044	0.052	0.042
Lone Parent HH	0.061	0.051	0.064	0.048	0.063	0.049
Non EU Residents	0.078	0.099	0.071	0.098	0.077	0.102
Social Renters	0.155	0.193	0.146	0.184	0.143	0.184
Property Owners	0.679	0.247	0.724	0.222	0.738	0.219
Non-white Residents	0.112	0.168	0.106	0.164	0.109	0.162
Population	291.622	76.550	301.933	71.815	304.437	68.526
Pop Density	34.705	36.889	52.758	52.610	56.005	67.125

Figure D.1: Distance to canal effects on prices, for urban and suburban areas only

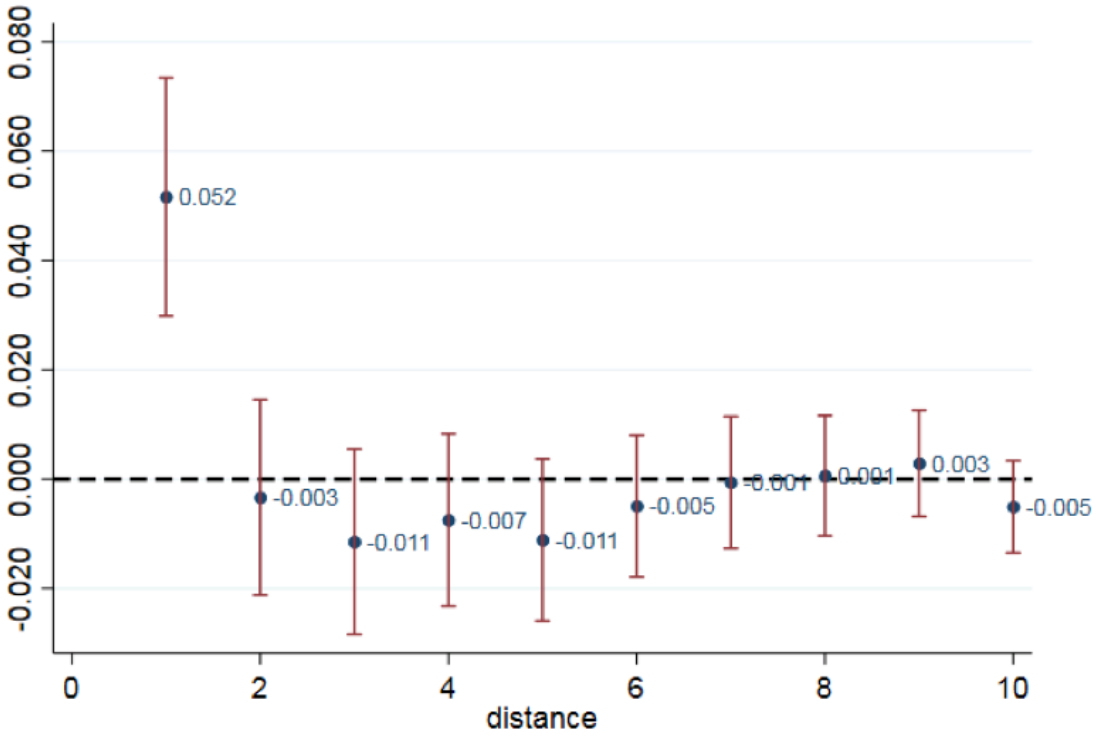


Table D.3: Descriptive statistics for the Droitwich sample

	(a) 0-100 m Droit.		(b) 100-1000m Droit		(c) 1000-1500m Droit		(d)<1500m W&Birm.	
	mean	sd	mean	sd	mean	sd	mean	sd
Natural log of price	11.928	0.475	11.954	0.485	12.265	0.456	11.932	0.449
Price	173564.9	130386.2	178095.9	150709.7	234656	111832	169158.4	123812.6
Price per m2	1984.42	804.16	2027.83	1223.39	2202.61	1010.34	2021.58	1890.03
Size(m2)	91.318	49.206	91.606	46.269	112.179	57.611	87.863	40.671
No.of Rooms	4.525	1.699	4.623	1.703	5.495	1.843	4.579	1.66
Fireplace	0.068	0.233	0.112	0.307	0.17	0.366	0.169	0.366
Energy Efficiency	63.88	12.284	62.91	12.152	59.219	11.578	58.935	13.478
Freehold	0.744	0.437	0.775	0.417	0.871	0.336	0.809	0.393
New Built	0.034	0.18	0.064	0.245	0.026	0.16	0.056	0.23
Terrace House	0.388	0.488	0.287	0.453	0.082	0.275	0.309	0.462
Flats	0.176	0.381	0.211	0.408	0.043	0.202	0.171	0.377
Semi-Detached	0.147	0.355	0.256	0.437	0.307	0.461	0.307	0.461
Low qualifications	0.325	0.072	0.307	0.119	0.199	0.077	0.251	0.111
Households no car	0.223	0.138	0.222	0.139	0.078	0.058	0.215	0.144
Unemployment Rate	0.05	0.029	0.041	0.033	0.029	0.013	0.038	0.029
Lone Parent HH	0.04	0.033	0.056	0.047	0.035	0.028	0.05	0.036
Non EU Residents	0.019	0.02	0.026	0.032	0.02	0.011	0.035	0.033
Social Renters	0.146	0.138	0.209	0.198	0.037	0.08	0.099	0.154
Property Owners	0.743	0.146	0.724	0.188	0.918	0.093	0.774	0.193
Non-white	0.013	0.008	0.012	0.013	0.015	0.013	0.04	0.061
Population	275	66.561	282.672	55.054	304.461	43.953	291.599	53.937
Pop Density	24.128	18.375	38.827	25.857	35.302	18.31	45.577	26.077
N	387		310		2047		20427	

Note: Means and standard deviations are reported for three distance groups related to the Droitwich Canal (<100m, 100-1000m and 1000-1500m) and for the overall sample for the Worcester and Birmingham control group (<1500m from the Worcester and Birmingham canal).

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

List of Tables

List of Figures

Bibliography

Acemoglu, D., Cantoni, D., Johnson, S., & Robinson, J. A. (2011). The consequences of radical reform: The French Revolution. *American Economic Review*, 101(7), 3286–3307.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401.

Adler, M. W., Liberini, F., Russo, A., & van Ommeren, J. N. (2017). *Road Congestion and Public Transit*. Itea conference working paper.

Ahlfeldt, G. & Pietrostefani, E. (2019). The economic effects of density: A synthesis. *CEPR Discussion Paper 13440*.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., & Wolf, N. (2015). The economics of density: Evidence from the Berlin wall. *Econometrica*, 83(6), 2127–2189.

Ahmed, A. & Dinye, R. D. (2011). Urbanisation and the challenges of development controls in Ghana: A case study of wa township. *Journal of Sustainable Development in Africa*, 13(7), 210–235.

Akbar, P. A., Couture, V., Duranton, G., & Storeygard, A. (2018). *Mobility and Congestion in Urban India*. Working Paper 25218, National Bureau of Economic Research.

Anderson, M. L. (2014). Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *American Economic Review*, 104(9), 2763–96.

Angrist, J. D. & Krueger, A. B. (1999). Chapter 23 - empirical strategies in labor economics. volume 3 of *Handbook of Labor Economics* (pp. 1277 – 1366). Elsevier.

Anthony, K., Andrew, M., Andrew, S., Pia, M., Ehrlich, S. S., & Jeffrey, S. (2004). A global index representing the stability of malaria transmission. *The American Journal of Tropical Medicine and Hygiene*, 70(5), 486–498.

Antràs, P., Fort, T. C., & Tintelnot, F. (2017). The margins of global sourcing: Theory and evidence from u.s. firms. *American Economic Review*, 107(9), 2514–64.

Armington, P. S. (1969). A theory of demand for products distinguished by place of production (une théorie de la demande de produits différenciés d'après leur origine) (una teoría de la demanda de productos distinguiéndolos según el lugar de producción). *Staff Papers (International Monetary Fund)*, 16(1), 159–178.

Ban, Y., Gong, P., & Giri, C. (2015). Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 1–6.

Banerjee, A. & Iyer, L. (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *American Economic Review*, 95(4), 1190–1213.

Barton, H., Grant, M., & Guise, R. (2003). *Shaping neighbourhoods: A guide for health, sustainability and vitality*. Taylor & Francis.

Baruah, N. G., Henderson, J. V., & Peng, C. (2017). Colonial legacies: Shaping African cities. *SERC working paper*.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2014). Roads, railroads and decentralization of chinese cities.

Bayer, P., Ferreira, F., & McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4), 588–638.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4), 841–890.

Bertrand, M. (2004). Land management and urban development projects: A comparison of experiences in French-speaking and English-speaking West Africa. *International Development Planning Review*, 26(1), 83–96.

Bleakley, H. & Lin, J. (2012). Portage and path dependence. *Quarterly Journal of Economics*, 127(2), 587–644.

Bonetti, F., Corsi, S., Orsi, L., & De Noni, I. (2016). Canals vs. streams: To what extent do water quality and proximity affect real estate values? a hedonic approach analysis. *Water*, 8(12), 577.

Braithwaite, A. (2017). The implications of internet shopping growth on the van fleet and traffic activity. (May), 1–6.

Briant, A., Lafourcade, M., & Schmutz, B. (2015). Can tax breaks beat geography? lessons from the french enterprise zone experience. *American Economic Journal: Economic Policy*, 7(2), 88–124.

Brilon, W. & Lohoff, J. (2011). Speed-flow models for freeways. *Procedia - Social and Behavioral Sciences*, 16, 26 – 36. 6th International Symposium on Highway Capacity and Quality of Service.

Brooks, L. & Lutz, B. (2014). Vestiges of transport: Urban persistence at a micro scale. *GWU working paper*.

Brueckner, J. K. (2001). Urban sprawl: Lessons from Urban Economics. *Brookings-Wharton papers on urban affairs*, 2001(1), 65–97.

Brueckner, J. K. (2005). Transport subsidies, system choice, and urban sprawl. *Regional Science and Urban Economics*, 35(6), 715–733.

Burchfield, M., Overman, H. G., Puga, D., & Turner, M. A. (2006). Causes of sprawl: A portrait from space. *The Quarterly Journal of Economics*, 121(2), 587–633.

Burley, J. (2016). *The Built Environment and Social Interactions: Evidence from Panel Data*. Working papers, University of Toronto.

Cairns, S. (2005). Delivering supermarket shopping: More or less traffic? *Transport Reviews*, 25(1), 51–84.

Cairns, S., Sloman, L., Newson, C., Anable, J., Kirkbride, A., & Goodwin, P. (2004). Smarter choices-changing the way we travel final report of the research project: 'the influence of soft factor interventions on travel demand' disclaimer and copyright accompanying report. *Final report of the research project: 'The influence of soft factor interventions on travel demand'*.

Calderón, F., Oppenheimer, J., & Stern, N. (2014). Better growth, better climate–the new climate economy report–the synthesis report. *Technical Report*.

Campante, F. R. & Do, Q.-A. (2014). Isolated capital cities, accountability, and corruption: Evidence from US states. *American Economic Review*, 104(8), 2456–81.

Canal and River Trust (2016). Canal and river trust annual report 2016-2017. Retrieved from url: https://canalrivertrust.org.uk/media/original/33176-annual-report-2016-17.pdf.

Carruthers, J. I. & Ulfarsson, G. F. (2003). Urban sprawl and the cost of public services. *Environment and Planning B: Planning and Design*, 30(4), 503–522.

Cavallo, A. (2017). Are online and offline prices similar? evidence from large multi-channel retailers. *American Economic Review*, 107(1), 283–303.

Chandra, A. & Thompson, E. (2000). Does public infrastructure affect economic activity?: Evidence from the rural interstate highway system. *Regional Science and Urban Economics*, 30(4), 457–490.

Choi, H. & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(SUPPL.1), 2–9.

Consortium, T.-A. (2010). *Public transport in Sub-Saharan Africa: Major trends and case studies.* International Association of Public Transport (UITP), Brussels.

Cooper, Z., Gibbons, S., Jones, S., & Mcguire, A. (2011). Does hospital competition save lives? evidence from the english nhs patient choice reforms. *Economic Journal*, 121(554), 228–260.

Costinot, A., Donaldson, D., Kyle, M., & Williams, H. (2016). *The More We Die, The More We Sell? A Simple Test of the Home-Market Effect*. Working Paper 22538, National Bureau of Economic Research.

Couture, V., Faber, B., Gu, Y., & Liu, L. (2018). *E-Commerce Integration and Economic Development: Evidence from China*. Working Paper 24384, National Bureau of Economic Research.

Coşar, A. K., Grieco, P. L., Li, S., & Tintelnot, F. (2018). What drives home market advantage? *Journal of International Economics*, 110, 135 – 150.

Crowder, M. (1964). Indirect rule—French and British style. *Africa*, 34(03), 197–205.

Currie, J., Davis, L., Greenstone, M., & Walker, R. (2015). Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings. *American Economic Review*, 105(2), 678–709.

Davis, O. A. & Whinston, A. B. (1964). The economics of complex systems: The case of municipal zoning. *Kyklos*, 17(3), 419–446.

Desmet, K., Gomes, J., & Ortuño-Ortín, I. (2018). *The Geography of Linguistic Diversity and the Provision of Public Goods*. NBER Working Papers 24694, National Bureau of Economic Research, Inc.

Dolfen, P., Einav, L., Klenow, P. J., Klopack, B., Levin, J. D., Levin, L., & Best, W. (2019). *Assessing the Gains from E-Commerce*. Working Paper 25610, National Bureau of Economic Research.

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

Donaldson, D. & Storeygard, A. (2016). The view from above: Applications of satellite data in Economics. *Journal of Economic Perspectives*, 30(4), 171–98.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(4), 795–813.

Durand-Lasserve, A. (2005). *Land for housing the poor in African cities: Are neo-customary processes an effective alternative to formal systems?*, chapter 12, (pp. 160–174).

Duranton, G. & Turner, M. A. (2011). The fundamental law of road congestion: Evidence from us cities. *American Economic Review*, 101(6), 2616–52.

Eaton, J. & Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5), 1741–1779.

Ellickson, R. C. (2012). The law and economics of street layouts: How a grid pattern benefits a downtown. *Alabama Law Review*, 64, 463.

Else, P. K. (1981). A reformulation of the theory of optimal congestion taxes. *Journal of Transport Economics and Policy*, 15(3), 217–232.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., & Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6), 1373–1379.

Epple, D. & Nechyba, T. (2004). Fiscal decentralization. In J. V. Henderson & J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics*, volume 4 chapter 55, (pp. 2423–2480). Elsevier, 1 edition.

Europa (2003). *Africa South of the Sahara 2004*. AFRICA SOUTH OF THE SAHARA. Routledge.

Fan, J., Tang, L., Zhu, W., & Zou, B. (2018). The alibaba effect: Spatial consumption inequality and the welfare gains from e-commerce. *Journal of International Economics*, 114, 203 – 220.

Forman, C., Ghose, A., & Goldfarb, A. (2009). Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Science*, 55(1), 47–57.

Fujita, M., Krugman, P. R., & Venables, A. J. (2001). *The spatial economy: Cities, regions, and international trade*. MIT press.

Gabaix, X., Laibson, D., Li, D., Li, H., Resnick, S., & de Vries, C. G. (2016). The impact of competition on prices with numerous firms. *Journal of Economic Theory*, 165, 1 – 24.

Garrod, G. & Willis, K. (1994). An economic estimate of the effect of a waterside location on property values. *Environmental and Resource Economics*, 4(2), 209–217.

Gendron-Carrier, N., Gonzalez-Navarro, M., Polloni, S., & Turner, M. A. (2018). *Subways and Urban Air Pollution*. Working Paper 24183, National Bureau of Economic Research.

Gibbons, S., Lyytikäinen, T., Overman, H. G., & Sanchis-guarner, R. (2019). New road infrastructure : The effects on firms. *Journal of Urban Economics*, 110(March 2017), 35–50.

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

Gibbons, S., Overman, H. G., & Patacchini, E. (2015). Chapter 3 - spatial methods. In J. V. H. Gilles Duranton & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics* (pp. 115 – 168). Elsevier.

Gibbons, S. & Wu, W. (2017). *Airports, Market Access and Local Economic Performance: Evidence from China*. SERC Discussion Papers 0211, Spatial Economics Research Centre, LSE.

Giri, C., Zhu, Z., & Reed, B. (2005). A comparative analysis of the Global Land Cover 2000 and MODIS Land Cover data sets. *Remote sensing of environment*, 94(1), 123–132.

Glaeser, E. L. & Kahn, M. E. (2010). The greenness of cities: Carbon dioxide emissions and urban development. *Journal of Urban Economics*, 67(3), 404–418.

Goldfarb, A. & Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1), 3–43.

Goldmanis, M., Hortaçsu, A., Syverson, C., & Emre, O. (2010). E-commerce and the market structure of retail: E-commerce and market structure. *Economic Journal*, 120(October), 651–682.

Gonzalez-Navarro, M. & Turner, M. A. (2018). Subways and urban growth: Evidence from earth. *Journal of Urban Economics*, 108, 85 – 106.

Goolsbee, A. & Klenow, P. (2018). Internet rising, prices falling: Measuring inflation in a world of e-commerce. (January 2014), 488–492.

Grant, R. & Yankson, P. (2003). City profile: Accra. *Cities*, 20(1), 65–74.

Gu, Y., Jiang, C., Zhang, J., & Zou, B. (2019). Subways and road congestion. *Working paper*.

Guiso, L., Sapienza, P., & Zingales, L. (2016). Long-term persistence. *Journal of the European Economic Association*, 14(6), 1401–1436.

Hall, J. V., Horton, J. J., & Knoepfle, D. T. (2019). Pricing efficiently in designed markets: The case of ride-sharing.

Harari, M. (2017). Cities in bad shape: Urban geometry in India. *Processed, Wharton School of the University of Pennsylvania*.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3), 389–432.

Helpman, E., Melitz, M. J., & Yeaple, S. R. (2004). Export versus FDI with heterogeneous firms. *American Economic Review*, 94(1), 300–316.

Helsley, R. (2004). Urban political economics. In J. V. Henderson & J. F. Thisse (Eds.), *Handbook of Regional and Urban Economics*, volume 4 chapter 54, (pp. 2381–2421). Elsevier, 1 edition.

Helsley, R. W. & Strange, W. C. (2007). Urban interactions and spatial structure. *Journal of Economic Geography*, 7(2), 119–138.

Henderson, J. V. (1996). Effects of air quality regulation. *The American Economic Review*, 86(4), 789–813.

Henderson, J. V., Storeygard, A., & Deichmann, U. (2017). Has climate change driven urbanization in Africa? *Journal of Development Economics*, 124(C), 60–82.

Henderson, V. & Mitra, A. (1996). The new urban landscape: Developers and edge cities. *Regional Science and Urban Economics*, 26(6), 613–643.

Hochman, O., Pines, D., & Thisse, J.-F. (1995). On the optimal structure of local governments. *The American Economic Review*, 85(5), 1224–1240.

Home, R. (2015). *Colonial urban planning in Anglophone Africa*. Routledge.

Hortas-Rico, M. & Solé-Ollé, A. (2010). Does urban sprawl increase the costs of providing local public services? Evidence from Spanish municipalities. *Urban Studies*.

Huillery, E. (2009). History matters: The long-term impact of colonial public investments in French West Africa. *American Economic Journal: Applied Economics*, 1(2), 176–215.

Kortum, S. S. (1997). Research, patenting, and technological change. *Econometrica*, 65(6), 1389–1419.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, 70(5), 950–959.

La Porta, R., Lopez-de Silanes, F., Pop-Eleches, C., & Shleifer, A. (2004). Judicial checks and balances. *Journal of Political Economy*, 112(2), 445–470.

La Porta, R., Lopez-de Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of Economic Literature*, 46(2), 285–332.

Lee, A. & Schultz, K. A. (2012). Comparing British and French colonial legacies: A discontinuity analysis of Cameroon. *Quarterly Journal of Political Science*, 7(4), 365–410.

Libecap, G. D. & Lueck, D. (2011). The demarcation of land and the role of coordinating property institutions. *Journal of Political Economy*, 119(3), 426–467.

Linden, L. & Rockoff, J. E. (2008). Estimates of the impact of crime risk on property values from megan's laws. *American Economic Review*, 98(3), 1103–27.

Liu, X., Li, X., Chen, Y., Tan, Z., Li, S., & Ai, B. (2010). A new landscape index for quantifying urban expansion using multi-temporal remotely sensed data. *Landscape Ecology*, 25(5), 671–682.

Løken, K. V., Mogstad, M., & Wiswall, M. (2012). What linear estimators miss: The effects of family income on child outcomes. *American Economic Journal: Applied Economics*, 4(2), 1–35.

Mahoney, P. G. (2001). The common law and economic growth: Hayek might be right. *The Journal of Legal Studies*, 30(2), 503–525.

Mayer, T. & Trevien, C. (2017). The impact of urban public transportation evidence from the paris region. *Journal of Urban Economics*, 102, 1 – 21.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York, NY, USA: Academic Press New York.

Michaels, G. (2008). The effect of trade on the demand for skill: Evidence from the interstate highway system. *The Review of Economics and Statistics*, 90(4), 683–701.

Michaels, G. & Rauch, F. (2017). Resetting the urban network: 117-2012. *The Economic Journal*, 128(608), 378–412.

Michalopoulos, S. & Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, 81(1), 113–152.

Michalopoulos, S. & Papaioannou, E. (2016). The long-run effects of the scramble for Africa. *American Economic Review*, 106(7), 1802–48.

Mooney, S. & Eisgruber, L. M. (2001). The influence of riparian protection measures on residential property values: the case of the oregon plan for salmon and watersheds. *The Journal of Real Estate Finance and Economics*, 22(2-3), 273–286.

Murdock, G. P. (1967). *Ethnographic Atlas*. University of Pittsburgh Press.

Nelson, G., Hansz, J. A., & Cypher, M. L. (2005). The influence of artificial water canals on residential sale prices. *Appraisal Journal*, 73(2).

Neumark, D. & Simpson, H. (2015). Chapter 18 - place-based policies. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics* (pp. 1197 – 1287). Elsevier.

Nicholls, S. & Crompton, J. (2017). The effect of rivers, streams, and canals on property values. *River research and applications*, 33(9), 1377–1386.

Njoh, A. (2001). *Planning Rules in Post-colonial States: The Political Economy of Urban and Regional Planning in Cameroon*. Nova Science Publishers.

Njoh, A. (2004). The experience and legacy of French colonial urban planning in Sub-Saharan Africa. *Planning Perspectives*, 19(4), 435–454.

Njoh, A. (2006). *Planning power: town planning and social control in colonial Africa*. UCL Press.

Njoh, A. (2016). *French Urbanism in Foreign Lands*. Springer.

Notley, S., Bourne, N., & Taylor, N. B. (2009). *Speed, flow and density of motorway traffic*. Trl insight report ins003, Crowthorne: Transport Research Laboratory.

Nunn, N. & Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *The Review of Economics and Statistics*, 94(1), 20–36.

Oates, W. (1999). An essay on fiscal federalism. *Journal of Economic Literature*, 37(3), 1120–1149.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14, 693–709.

Oto Peralias, D. & Romero-Ávila, D. (2017). *Colonial theories of institutional development: Toward a model of styles of Imperialism*. Contributions to Economics. Netherlands: Springer.

O'Grady, T. (2014). Spatial institutions in Urban Economies: How city grids affect density and development. *Harvard University*.

Parry, B. I. W. H. & Small, K. A. (2009). Should urban transit subsidies be reduced? *American Economic Review*, 20036, 700–724.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., & Syrris, V. (2016). *Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014*. Technical report, JRC Technical Report.

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., et al. (2013). A Global Human Settlement Layer from optical HR/VHR RS data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), 2102–2131.

Pogonyi, C. G., Graham, D. J., & M Carbo, J. (2018). *Growth or Displacement? A Metro Line's Causal Impact on the Spatial Distribution of Business Units and Employment: Evidence from London*. Technical report.

Punakivi, M. (2003). *Comparing alternative home delivery models for e-grocery business*. PhD thesis, Helsinki University of Technology.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Ramankutty, N., Foley, J. A., Norman, J., & McSweeney, K. (2002). The global distribution of cultivable lands: Current patterns and sensitivity to possible climate change. *Global Ecology and Biogeography*, 11(5), 377–392.

Redding, S. J. (2016). Goods trade, factor mobility and welfare. *Journal of International Economics*, 101, 148 – 167.

Redding, S. J. & Sturm, D. M. (2008). The costs of remoteness: Evidence from german division and reunification. *American Economic Review*, 98(5), 1766–97.

Redding, S. J. & Turner, M. A. (2015). Chapter 20 - transportation costs and the spatial organization of economic activity. In J. V. H. Gilles Duranton & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics*, volume 5 of *Handbook of Regional and Urban Economics* (pp. 1339 – 1398). Elsevier.

Redding, S. J. & Weinstein, D. E. (2016). *Measuring Aggregate Price Indexes with Demand Shocks: Theory and Evidence for CES Preferences*. Working Paper 22479, National Bureau of Economic Research.

Relihan, L. E. (2017). *Is online retail killing coffee shops?* Working paper.

Rosin, P. L. (1999). Measuring rectangularity. *Machine Vision and Applications*, 11(4), 191–196.

Ross, E. & Bigon, L. (2018). The urban grid and entangled planning cultures in Senegal. *Planning Perspectives*, 0(0), 1–27.

Rossi-Hansberg, E. (2004). Optimal land use and zoning. *Review of Economic Dynamics*, 7, 69–106.

Rowland, C.S.;Morton, R. L. G. A. C. (2017). Land cover map 2015 (vector, gb).

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

Scholz, W., Robinson, P., & Dayaram, T. (2015). *Colonial Planning Concept and Post Colonial Realities: The Influence of British Planning Culture in Tanzania, South Africa and Ghana*. Routledge.

Seim, K. & Sinkinson, M. (2016). Mixed pricing in online marketplaces. *Quantitative Marketing and Economics*, 14(2), 129–155.

Shertzer, A., Twinam, T., & Walsh, R. P. (2016). Race, ethnicity, and discriminatory zoning. *American Economic Journal: Applied Economics*, 8(3), 217–246.

Silva, C. N. (2015). *Urban planning in Sub-Saharan Africa: Colonial and post-colonial planning cultures*. Routledge.

Streiner, C. F. & Loomis, J. B. (1995). Estimating the benefits of urban stream restoration using the hedonic price method. *Rivers*, 5(4), 267–278.

Survey, O. (2015). Strategi. Downloaded from EDINA Digimap Ordnance Survey Service http://digimap.edina.ac.uk.

Survey, O. (2018a). Os open greenspace. Downloaded from EDINA Digimap Ordnance Survey Service http://digimap.edina.ac.uk.

Survey, O. (2018b). Os open rivers. Downloaded from EDINA Digimap Ordnance Survey Service http://digimap.edina.ac.uk.

Tang, C. K. et al. (2016). *Traffic externalities and housing prices: evidence from the London congestion charge*. SERC, Spatial Economics Research Centre.

Tikoudis, I., Verhoef, E. T., & Ommeren, J. N. V. (2015). On revenue recycling and the welfare effects of second-best congestion pricing in a monocentric city. *Journal of Urban Economics*, 89, 32–47.

Trubka, R., Newman, P., & Bilsborough, D. (2010). The costs of urban sprawl-infrastructure and transportation. *Environmental Design Guide*.

Tsivanidis, N. (2019). *The Aggregate and Distributional Effects of Urban Transit Networks : Evidence from Bogotá ' s TransMilenio*. Working paper.

Turner, M. A. (2007). A simple theory of smart growth and sprawl. *Journal of Urban Economics*, 61(1), 21–44.

Underwood, R. T. (1961). *Speed, Volume and Density Relationships*. Yale University Bureau of Highway Traffic.

Varian, H. R. (1980). A model of sales. *The American Economic Review*, 70(4), 651–659.

Vosen, S. & Schmidt, T. (2011). Forecasting private consumption survey based indicators vs google trends. *Journal of Forecasting*, 30(6), 565–578.

Walters, A. A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica*, 29(4), 676–699.

Waterways, B. (2010). Restoration of the droitwich canals. archived Wayback Machine, https://web.archive.org/web/20100924192725/http://britishwaterways.co.uk/media/documents/Droit_Restoration_Non_Technical_Summary.pdf.

Yang, J., Chen, S., Qin, P., Lu, F., & Liu, A. A. (2018a). The effect of subway expansions on vehicle congestion : Evidence from beijing. *Journal of Environmental Economics and Management*, 88, 114–133.

Yang, J., Liu, Y., Qin, P., & Liu, A. A. (2014). A review of beijing's vehicle registration lottery short term effects on vehicle growth and fuel consumption. *Energy Policy*, 75, 157–166.

Yang, J., Purevjav, A.-o., & Li, S. (2018b). The marginal cost of traffic congestion and road pricing : Evidence from a natural experiment in beijing. *American Economic Journal Economic Policy (Forthcoming)*.

Žunić, D., Martinez-Ortiz, C., & Žunić, J. (2012). Shape rectangularity measures. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(06), 1254002.