THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Essays in Urban & Development Economics

Nathalie Picarelli

A thesis submitted to the London School of Economics for the degree of Doctor of Philosophy, London, August 2017.

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 50,000 words.

Statement of conjoint work

I certify that Chapter 3 of this thesis was co-authored with Pascal Jaupart and Ying Chen. I contributed 33% of the work.

To my mother, from whom I learned to never stop trying.

Acknowledgements

This PhD would not have been possible without the time and support of many people.

I am deeply indebted to my supervisors Olmo Silva and Henry Overman. In very different ways, they have always been there to support me throughout this journey, going beyond what I could have expected from research advisors. I am grateful for their continued encouragement, trust, guidance and thoughtful advice that has helped me navigate the difficult ups and downs of applied empirical research. Their rigor and enthusiasm for research have been a great source of motivation and I am forever grateful.

This work has also benefited from the comments and time of faculty and colleagues at LSE and other universities. In particular, I wish to thank Vernon Henderson who provided helpful advice and inspiration throughout the PhD. And Murray Leibbrandt, who gracefully hosted me at the University of Cape Town for half a year to conduct part of this research.

Many more people outside of academia have made this PhD possible thanks to laughter, love and friendship. I am deeply thankful to Laura Quintero, Leona Verdadero, Nuria Tolsá Caballero, Monica Chavez, and Alex Girón Gordillo for their unconditional friendship. I would never have made it without them. Gina Andrade, Pilar Lopez Uribe, and Maria Sanchez-Vidal, have been my partners in crime for many adventures under the grey skies of London, and a big thank you goes for their patience and friendship. To Gabrielle Sanson, I am grateful to have shared immemorable moments and conversations in Cape Town. And to Ana Moreno-Monroy for being not only a close friend but a source of inspiration. Martin Heger, CK Tang, Piero Montebruno and Ying Chen have also been indispensable along our shared road.

I was lucky to share this journey with Pascal Jaupart: for his unfailing support and friendship, for always listening and bearing with my never-stopping brain, for his jokes and sweet sarcasm, this PhD owes much to him and I am forever indebted.

Finally, I want to express my deepest gratitude to the two most important persons in my life: my mother and my sister. This PhD is for my mother, whose love, wisdom, and support are behind every step of my life.

I gratefully acknowledge the generous financial support from the ESRC and the LSE, as well as the IGC for the work in chapter 3.

Abstract

This thesis consists of four independent chapters on urban and development economics.

Chapter 1 looks at the issue of distance and labour outcomes in urban areas of a developing country. It studies the effect of a housing relocation program on the labour supply and living conditions of low-income households across major cities in South Africa. For this, I use four waves of panel microdata collected between 2008 and 2014, and I exploit the arbitrary eligibility rules of the policy with a fuzzy regression discontinuity design to obtain causal estimates. In the short-term of two to four years following relocation, I find that the labour supply of recipient households decreases by one standard deviation, driven mostly by a decrease in female hours. I find evidence of a large increase in distance (km) to economic opportunities. This is likely to be an important factor behind the decline, directly or indirectly through within-family shifts in livelihood strategies. Evidence is limited regarding improvements in housing and neighbourhood quality.

Chapter 2 examines how neighbourhoods where children grow up can play a significant part in shaping their opportunities later in life. It provides unique evidence in a developing country context by using the random allocation of households to ethnically segregated residential areas during apartheid in South Africa. The main observations come from a panel of young adults aged 14 to 22 at baseline and residing in the city of Cape Town. It covers 5 periods of their life between 2002 to 2009. I focus on black children in families living in former black-only residential areas. I find compelling evidence of neighbourhood effects on labour and educational outcomes in adulthood across deprived neighbourhoods. The differences are more marked for young women, suggesting a stronger hold of social norms and institutions for young men. Location, both in terms of access to jobs and access to higher quality public amenities (schools), social networks and the underlying human capital composition of the neighbourhood are positively correlated to having better socioeconomic outcomes in adulthood.

Chapter 3 moves beyond socioeconomic outcomes, to study the relationship between extreme weather events and disease in developing cities. As climate change is making extreme weather events more frequent around the world, urban residents in developing countries have become more vulnerable to health shocks due to poor sanitation and infrastructure. The chapter empirically measures the relationship between weather and health shocks in the urban context of sub-Saharan Africa. Using unique

high-frequency datasets of weekly cholera cases and accumulated precipitation for wards in Dar es Salaam, we find robust evidence that extreme rainfall has a significant positive impact on weekly cholera incidence. The effect is larger in wards that are more prone to flooding, have higher shares of informal housing and unpaved roads. We identify limited spatial spillovers. Time-dynamic effects suggest cumulated rainfall increases cholera occurrence immediately and with a lag of up to 5 weeks.

Chapter 4 addresses questions related to the local impact of economic policies in developing countries. Specifically, I provide evidence on the local effect of a popular trade policy: export processing zones. The chapter examines the impact of their establishment on the levels of per capita expenditure across Nicaraguan municipalities for the period 1993 to 2009. Using the time and cross-section variation of park openings in a difference-in-differences framework, I find that on average consumption levels increased by 10% to 12% in treated municipalities. Yet, average effects mask significant disparities across the expenditure distribution. The results suggest that the policy benefited the upper-tail the most: expenditure levels increased by up to 25% at the 90th percentile. At the opposite end of the distribution, only the bottom decile registered a positive increase in expenditure levels of close to 10% across the period.

Contents

Introduction

This thesis consists of four independent chapters that fall on the realm of urban and development economics. A central theme explored in this thesis is how the inherently uneven spatial distribution of economic activities and people affects lower-income populations. Sorting of firms and households within cities and across regions is one of the fundamentals of urban economics. Agglomeration economies push firms to cluster (Duranton & Puga 2004); household sort in cities according to income levels and land prices (Alonso 1964; Mills 1967 and Muth 1969), transport mode choices (Glaeser et al. 2008), and amenities (Brueckner et al. 1999). Consequently, there are significant spatial variations in housing and infrastructure quality, environmental and transport externalities, public service provision and amenities across and within regions and cities.

In developing countries, poor households are disproportionately penalized by these spatial differences. They not only have limited access to efficient markets and quality infrastructure (Banerjee & Duflo 2007), they have to attribute disproportionately larger shares of their incomes and time to *mobility* (Baker et al. 2005). Under these conditions, geography, in the sense of location and distances, disproportionately affect the lives of the poor. This thesis contributes to improving our understanding of how and when this happens.

Relatedly, the process of urbanization in developing countries holds a fundamental place across these chapters. Rapid urbanization has often led to unplanned spatial expansions that put significant strain on housing and serviced land, resulting in congestion, contagion, insufficient public service provision, and limited formal employment opportunities. How these "demons of density" (Glaeser 2011) affect the urban poor relates to the fundamental relationship between space and poverty. Poor city dwellers often locate at the peripheries where land is cheaper but economic chances are scarcer, or settle in slums typified by inadequate physical infrastructure and high densities (Barnhardt et al. 2016), prone to become poverty traps (Marx et al. 2013). Limited urban connectivity and high levels of employment informality means that they face complex trade-offs between location-specific characteristics (such as public amenities and distances to jobs) and housing quality.

How distances and locations affect the poor through labour markets is a key dimension explored in this thesis (*chapters 1,2 & 4*). Distances create frictions and inefficiencies in urban labour markets (i.e. the spatial mismatch hypothesis; Kain 1968;

Zenou 2009; Gobillon et al. 2011). The emergence of slums testifies to the premium on proximity that poor households are willing to pay by living in substandard housing conditions to be close to employment opportunities (Galiani et al. 2016). Unequal access to infrastructure and public services and amenities is another mechanism explored (*chapters 1,2&3*). The under-provision of public services such as waste collection or public transport heavily burdens the poor in their daily transactions. It also makes them more vulnerable to weather shocks, and related negative income shocks (Burgess et al. 2017). Because of its relationship with labour markets, public amenities and infrastructure, location is also intrinsically related to socioeconomic mobility (Chetty et al. 2015).

Government and city authorities can affect the relationship between space and poverty. They can do this either by directly addressing market failures and negative externalities, or indirectly by impacting local economies. Some of the chapters (chapters *1&4*) of this thesis look at these policy options.

Three chapters focus on these questions by looking at cities in sub-Saharan Africa. The fourth focuses on larger regions in one of the poorest countries in Latin-America.

Chapter 1 looks at the fundamental question of how distances affect employment opportunities of low-income households in contexts of high commuting costs, low public transit provision and high frictions to access urban labour markets. These features imply that poor households obtain a higher utility from being closer to central areas (Glaeser et al. 2008) where public services and transit provision is higher. They are inversely more heavily burdened by living far. For instance, the disconnect from job opportunities can introduce search inefficiencies either by restricting the size of the effective labour market or because of information distortions; it can increase reservation wages and jobs search costs due to high commuting costs; and employers may discriminate against workers living far away, as they may consider it deteriorates their productivity (Zenou 2009; Gobillon et al. 2011).

I explore these relationships by examining a large relocation housing program across main metropolitan areas of South Africa. The chapter provides quasi-experimental evidence on the effect the program has on labour outcomes and living conditions of recipient low-income households in the country largest cities. To do this, I use four waves of panel micro data collected between 2008 and 2014, and I exploit the arbitrary eligibility rules of the policy with a fuzzy regression discontinuity design to obtain causal estimates. I find evidence that in two to four years following relocation to public housing, households are not better off on a majority of outcomes, including a

deterioration of their labour supply. A large increase in distance (km) to economic opportunities is likely to be an important factor behind the decline, directly or indirectly through within-family shifts in livelihood strategies. Evidence is limited regarding improvements in housing and neighbourhood quality.

Findings in this chapter offer significant insights into the response mechanisms of low-income households to relocation programs, particularly by inspecting the trade-offs they are willing to make between location-specific characteristics and housing quality. They suggest that in contexts of high commuting costs and high frictions to access urban labour markets, the location of dwellings can have a large effect on poor families' wellbeing. Much of the effect also depends on their possibility to implement alternative livelihood strategies.

Chapter 2 continues to investigate the relationship between location and poverty by exploring how it affects socioeconomic mobility. To do this, the chapter examines neighbourhood effects for children growing up in disadvantaged urban *ghettos*. This chapter is more descriptive in nature.

Neighbourhoods where children grow up can play a significant part in shaping their opportunities later in life (Glaeser & Cutler 1997; Oreopoulos 2003; Chetty et al. 2015). Growing up in deprived or 'bad' neighbourhoods, i.e. areas with high levels of unemployment, poverty, criminality and low-quality local public goods, can be detrimental in adulthood (Case & Katz 1991; Crane 1991). How they affect socioeconomic mobility is central to understanding the persistence of inequality over time (across generations) and space (poverty traps). This question is particularly relevant in developing countries, where cities display high levels of income inequalities, emphasized by stark spatial differences in the access to public services, formal housing, formal jobs and consumption amenities. There is extremely little empirical evidence of neighbourhood effects in this context. This chapter addresses this gap, by using the quasi-natural social experiment of apartheid in South Africa to measure the effects of growing up in disadvantaged neighbourhoods or ghettos. It also provides suggestive evidence on the main channels through which some neighbourhoods may be better than others for the life trajectories of children.

The analysis uses the random allocation of households to ethnic specific residential areas during apartheid in South Africa to disentangle the fundamental problem of measuring neighbourhood effects: observed differences in neighbourhoods may just reflect the spurious correlation induced by the likelihood that the same factors that lead to a given location choice also lead to certain socioeconomic outcomes. The main

observations come from a panel of young adults aged 14 to 22 at baseline and residing in the city of Cape Town. It covers 5 periods of their life between 2002 to 2009. I focus on compliers in black ghettos only; that is black children in families living in black-only residential areas (townships) by 1991 (when residential segregation was formally lifted). I provide strong evidence that for this group of young adults, residential decisions were exogenous and sorting is very limited.

I find compelling evidence of neighbourhood effects on labour and educational outcomes in adulthood across ghettos. The differences are more marked for young women, suggesting a stronger hold of social norms and institutions for young men. Location, both in terms of access to jobs and access to higher quality public amenities (schools), social networks and the underlying human capital composition of the neighbourhood are positively correlated to having better socioeconomic outcomes in adulthood.

This paper is limited in the extent to which it can provide causal evidence on the channels behind the persistence of spatial poverty traps. Still, it provides interesting findings regarding the importance of neighbourhoods on outcomes later in life in a large city of a developing country.

Chapter 3 completed with Pascal Jaupart and Ying Chen, looks at the relationship between *geography* and poverty within cities from a different angle. Departing from the fact that populations are affected differently by the same weather variations according to their income and locations (Dasgupta, 2010; Burgess et al. 2017), it focuses on how weather shocks can affect health within urban areas.

Climate change is expected to have a significant impact on the lives of the poor in the years ahead as extreme weather events such as floods, heavy precipitation and droughts become more frequent (Harrington et al. 2016). The question of how urban dwellers in developing cities are impacted by weather shocks is thus becoming increasingly relevant. Rapid urbanization has meant that many cities have poor infrastructure leaving populations vulnerable to vector-borne diseases.

Weather shocks such as rainfall can affect health through two main channels (Burgess et al. 2017): direct mechanisms via increased contact with pathogens; or indirect ones, through the effect they may have on real incomes, i.e. the negative income-shock may in turn lower the consumption of health-improving goods increasing the exposure to pathogens. Looking at health outcomes is important. Contagion is one "downside of density" (Glaeser & Sims 2015). Throughout history, cities with low-quality infrastructure and poor sanitation have been pockets of epidemics (for instance

19th century London or Paris, Kesztenbaum & Rosenthal 2016). Poor health and disease not only lower productivity in the short-term, they also hinder long-term economic growth (Well 2007).

Empirical evidence is scant concerning the impact of weather shocks in developing country cities. This chapter makes progress on this issue by looking at the effect of rainfall and flooding on cholera incidence in Dar es Salaam. The empirical analysis uses finely disaggregated ward-level panel data containing weekly recorded cholera cases and weekly accumulated precipitation for all the municipalities in the city. It is therefore testing whether exogenous weekly rainfall variation at the ward-level affects cholera occurrence. We find robust evidence that extreme rainfall has a significant positive impact on weekly cholera incidence. The effect is larger in wards that are more prone to flooding, have higher shares of informal housing and unpaved roads. Remarkably, we find little to no spatial spillovers from precipitation in neighbouring wards. Only when considering the relative elevation, there is a small significant effect from precipitation in neighbouring downhill wards. Time-dynamic effects suggest cumulated rainfall increases cholera occurrence immediately and with a lag of up to 5 weeks.

Results in the paper emphasize the key role of local infrastructure. Neighbourhoods with low-quality infrastructure are likely to be more exposed to the cholera bacteria when surfaces are washed and drains are overflown by severe precipitation. Vulnerable populations in these wards are also more likely to suffer from negative income shocks during extreme weather events. This chapter emphasizes the need for more evidence on large-scale policy interventions in poor urban areas to increase resilience and prevent contagion of treatable diseases in developing cities.

Chapter 4 of this thesis abandons the urban-specific setting to look at inequalities within municipalities in one of the poorest Latin-American countries. It focuses on understanding how the spatial variation in economic activity may affect differently low and high-incomes households in contexts of labour market frictions. It does so by *evaluating* the establishment of export processing zones (EPZ) in Nicaragua.

While having a wider reach than traditional place-based policies, EPZ programs are prone to have a significant influence on the local economies (Wang 2013) as they operate with incentives to hire and create economic activity in or near the areas where they locate, fostering agglomeration economies (Combes et al. 2011). Local welfare effects might differ from those at the aggregate level, particularly in cases with labour market frictions. Labour mobility and land-price responses may be such that the jobs created go to non-poor residents and that the gains from land prices benefit higher-

income households (Neumark and Simpson 2014). As stressed by Kline and Moretti (2014) and shown by Reynolds and Rohlin (2015) for the case of US federal empowerment zones, positive average effects of spatially-tied policies can mask significant disparities in terms of the actual beneficiaries across the income distribution in treated areas. There are heterogeneous effects according to whether individuals are homeowners or renters, or more generally by skill and initial income levels, that are not necessarily captured by looking at the average effect on local wages (Neumark and Simpson 2014). In this sense, disentangling in what way the establishment of an EPZ profits the different segments of the income distribution within concerned areas helps to shed light on the mechanisms of the policy and ultimate local beneficiaries.

In this chapter, I construct a unique municipal-level panel that allows for the examination of different moments of the expenditure distribution before and after the policy implementation. It covers five years between 1993 and 2009. By focusing on the aggregate outcomes at the municipal level and on levels of household per capita expenditures, the study captures general equilibrium effects. The analysis reveals interesting conclusions concerning average and distributional dynamics of EPZ policies at the local level. It confirms that there is a positive average treatment effect on the levels of real per capita expenditure in treated municipalities. It also reveals that the mean effect hides significant disparities across the income distribution, offering suggestive evidence that those at the upper-tail are the main beneficiaries of the policy over the period covered. The effects are however incremental in time, with median households benefitting after eight years of a plant establishment. Results are consistent with the existence of a skill-premium in the exporting sector that may be altering the cost of living (i.e. land price responses) in profit of the higher-skill/higher-income segments. I find that the average positive gains are concentrated in the high-skill working-age group.

This final chapter offers some interesting insight on the importance of analysing trade policies at the local level, considering local labour market dimensions and agglomeration economies. Goldberg and Pavcnik (2016) acknowledge the move towards this direction, and underline the interest of the empirical literature in influencing richer theoretical models.

This thesis contributes to the better understanding of the intrinsically unequal relationship between space and income levels in developing countries. It provides new insights into how and when spatial frictions affect poor households. Labour markets,

public services and amenities, and infrastructure are all centrally related to the location decisions of poor households and the effect of *distance and place* on their lives, in the short or longer term.

Many of the findings across these chapters highlight the importance of social networks. When *distance* matters, social networks are likely to play an important role, not only for jobs but also for enforcing norms and security, and providing informal safety nets. Gender-specific dimensions are also highlighted throughout the chapters. Males and females do not have the same relationships with location and distances. These dynamics matter for policies as poverty is rapidly urbanizing and cities are expected to host more than two-third of the world's population in the next decades.

Chapter 1

There is No Free House. Low-cost Housing & Labour Supply in Urban South Africa[1]

1.1 Introduction

The unprecedented pace of urbanization in developing countries has raised the need to improve our understanding of the policies that address "the demons of density" (Glaeser 2011). This is particularly true about housing policies. The fast growth of cities places significant strain on housing and serviced land, forcing poor households to choose substandard housing conditions and hostile geographical environments to be close to economic opportunities (Galiani et al. 2016). Far from temporary (Marx et al. 2013), informal settlements have proliferated as a direct response to the lack of affordable housing. Governments have often responded with relocation programs by which subsidized housing provision is directly allocated to eligible households through lotteries or waiting lists (Barnhardt et al. 2016). The success of these programs is closely related to the way they affect the livelihood strategies of recipients. Yet, despite popularity, there is still little robust empirical evidence about their benefits. Projects are often located at the peripheries of cities, which in contexts of low public transit provision may increase frictions in urban labour markets (Franklin 2017). Voluntary take-up may just reflect the lack of alternative options, and the costs of relocation (i.e. loss of social networks, higher commuting costs) may overcome its benefits (i.e. safer environments and improved housing).

[1] I further thank Stuart Rosenthal for his comments that greatly improved the final version of this paper, as well as two anonymous referees. Conversations with Nathaniel Baum-Snow, Pascal Jaupart, Liza Ciriola, Simon Franklin, Vernon Henderson, Maria Moreno-Monroy and Maria Sanchez-Vidal were also very helpful. I thank Michael Oberfichtner, participants at the 10th meeting of the UEA, and the IV Workshop in Urban Economics at IEB, for their comments. I also thank Michelle Chinehma for her support to access secured sections of the dataset.

This paper makes progress on this issue by providing the first quasi-experimental evidence of the effect of a large public housing program on labour outcomes and living conditions of recipient households across main metropolitan areas of a developing country. It further elucidates the response mechanisms of households in the presence of high frictions in urban labour markets.

The program I study here is South Africa's national Housing Subsidies Scheme rolled-out at the end of Apartheid under the Reconstruction and Development Program (RDP). Similar to other relocation programs, it consists of one-off capital subsidies for the construction of free-standing housing units on greenfield developments, managed by local authorities and allocated to eligible households on an ownership basis. Allocation is done through municipal waiting lists. Since 1994, 2.8 million houses were built under the program across South Africa, benefitting more than 12 million individuals (2015). The advantage of studying this program is its large national scale, permitting to look for the first time across the six main metropolitan areas[2] of a large country in sub-Saharan Africa[3]. A second advantage is that eligibility relies on an arbitrary income rule, with households earning less than R3500 per month eligible for free housing under the scheme. I exploit this discontinuity to estimate treatment effects that are as good as randomized in a neighbourhood around the discontinuity threshold. A final interesting feature concerns the country historical heritage. South Africa's urban structures have been marked by almost half a century of Apartheid spatial planning, which has accentuated constraints to access labour markets for the low-income households residing at cities' peripheries (Banerjee et al. 2008). Combined with very high unemployment rates, this issue has put housing policies at the top of the political agenda[4].

This study identifies the causal impact of receiving a subsidized housing under the RDP program by comparing households just below and above the income threshold, using being below the income threshold as an instrument for receiving subsidized housing. The probability of receiving subsidized housing decreases sharply above R3500 among otherwise eligible households. Under the assumption of no manipulation of the assignment variable this fuzzy regression discontinuity (FRD) approach ensures greater internal consistency than other quasi-experimental methods, and results are (locally)

[2] Here I look at the 6 largest metropolitan areas in South Africa, all with populations above 1 million.
[3] Sub-Saharan Africa is the region with the largest share of slum population in the world (199.5 million in 2015). It is estimated to be growing at 4.5% per year (UN-Habitat). Here I look at the 6 largest metropolitan areas in South Africa, all with populations above 1 million.
[4] Many argue that the housing scheme has helped reinforce the spatial logic of Apartheid by moving the poor to low-density areas often in old township locations (Lall et al. 2012).

comparable to that of randomized control trials (Lee and Lemieux 2010). I find no evidence of sorting at the threshold. To estimate treatment effect, I use four waves of panel data from the National Income Dynamics Study (NIDS) collected between 2008 and 2014, and documenting households' subsidy status for obtaining their dwelling. Overall, my results are stable across a range of specifications, bandwidths, controls and sample restrictions. With regard to the external validity of the estimates, the FRD estimand should be interpreted as a weighted average treatment effect for the subpopulation affected by the instrument (Local Average Treatment Effect or LATE), where the weights reflect the ex-ante likelihood that the households' income is near the threshold. Results are thus valid for the population induced to take-up treatment around the discontinuity.

I find that in the short period of two to four years following RDP allocation, households are not better off on a majority of outcomes. The labour supply of recipient households at cutoff decreases by between half to one standard deviation. Overall, the decline is driven by a large drop in total weekly hours of paid work of about 30 hours, consistent with a reduction of the number of employed members by almost one. The gender decomposition shows that the reduction is larger for female members through a reduction at their intensive margins, but male members also experience a non-significant drop. The share of working age members that are unemployed and discouraged also rises.

I find evidence that distances to the Central Business District (CBD) and employment nodes increase significantly following RDP allocation by 12 to 13 km. The fact that I find no significant effect on the monthly share of total household revenue spent on transport, further suggests that households are not compensating by longer (more expensive) commutes, but rather choose to reduce their overall travelling. It follows that the reduction in the labour supply could directly result from the more expensive commutes, supporting the existence of high frictions in South African urban labour markets in line with the spatial mismatch hypothesis (Gobillon et al. 2011). Females could be more severely affected. There is indication that women in developing countries rely on fewer transport modes, and are potentially more vulnerable to distance (Baker et al. 2005, Venter et al. 2007).

The decline in the labour supply could also reflect households shifting their livelihood strategies following the wealth shock. I explore within-household response mechanisms in the paper. I find no evidence that relocation severely disrupts household composition in terms of their labour capabilities. The age of the household head and

the age and number of children are insignificant, and there is no indication that less mobile members are being moved to the farther away house. RDP recipients display a 54 percentage points (pp) higher likelihood of receiving additional income from rent. The size is large. It presumably reflects a new earning strategy in response to the accrued distance to jobs (i.e. plausibly by constructing and renting backyard rooms). To a smaller extent, it is also coherent with a pure income and price-substitution effect. Other methods such as larger government transfers and subsistence agriculture are insignificant.

Despite the increase in distances, households that retrieve a higher utility from *better* housing may still be better-off with RDP housing. I find limited evidence in this regard. All characteristics of neighbourhood quality tested are insignificant. Revealing however is the insignificant but positive sign on the perception of insecurity which could result from isolation at cities' peripheries and the disruption of social networks. The loss of social networks has been raised as a major limitation of this type of relocation programs in developed and developing countries alike (Barnhardt et al 2016; Day and Cervero 2010; Mills et al 2006; Kling et al 2007). Conversely, I find a deterioration in housing amenities by half a standard deviation. This is mostly related to worse structures (fragile walls, number of rooms). The fact that mobile households still choose to relocate suggests that the improvement in housing consumption may be happening through intangible features of housing quality, such as increased tenure security and the possibility of becoming homeowners. Subjective wellbeing is an important component of housing policies (Galiani et al. 2016).

Findings in this paper remain valid for a short-term period of two to four years following allocation. Still, they offer significant insights into the response mechanisms of low-income households to relocation programs, particularly by inspecting the trade-offs they are willing to make between location-specific characteristics and housing quality. They suggest that in contexts of high commuting costs and high frictions to access urban labour markets, the location of dwellings can have a large effect on the wellbeing of poor families. Further, much of the effect also depends on the participants' possibility to implement alternative livelihood strategies. To succeed, housing policies need to consider these different dimensions.

By looking at these dynamics, the paper is closely linked to a large literature in urban economics that has tried to understand residential sorting within cities. These papers have put forward the role of commuting costs (Alonso 1964; Mills 1967 and Muth 1969), transport mode choices (Glaeser et al. 2008), amenities (Brueckner et al. 1999;

Lall et al. 2008), and the proximity to economic opportunities (Kain 1968) as key determinants for residential location choices. The mechanisms they put forward are key to understanding the results in this paper.

This study also relates to the large literature that has looked at the impact of housing lotteries and vouchers on the labour supply of beneficiaries in U.S. cities, with the Moving to Opportunity (MTO) program being the most notorious. Most of these papers find null (Jacob 2004; Kling et al. 2007) or negative impacts on the labour supply of recipient households (Jacob and Ludwig 2013; Mills et al. 2006 for the first year and null afterwards). Only Chetty et al. (2015) find long-term improvements in children under 13 at the time of moving. In developing countries, there are reasons to believe that labour supply responses to housing subsidies are more difficult to assess. Tenure insecurity, the predominance of informal employment, the greater constraints on transport infrastructure as well as residential inequalities being more closely related to the spatial layouts of cities, make housing a greater complement to work. These features are behind the existence of greater frictions in urban labour markets of developing countries cities.

In this sense, this paper adds to a nascent literature evaluating housing programs in cities of developing countries, which has mostly focused on titling and slum upgrading programs (Fields 2007, Takeuchi et al. 2008; Galiani & Schargrodsky 2010; Galiani et al. 2016). While still focusing on slum dwellers, the exception is Barnhardt et al. (2016)[5]. They provide the first experimental evidence of a housing lottery in Ahmedabad, and find robust evidence that the target population was worse off in the long term (14 years) on a variety of socioeconomic measures, including labour supply. Two other papers have examined the RDP policy before. Franklin (2015) estimates the causal effects of the South African RDP program on labour outcomes but focuses solely on the population of slum dwellers in the city of Cape Town[6], de facto studying the effect of titling on a subpopulation of the treated. Lall et al. (2012) remains relatively descriptive, but their exercise supports the results here. To the best of my knowledge, this is the first paper to estimate the effect of a large housing program for all low-income households across major cities of a developing country. It adds to the little evidence available to guide policy discussions.

[5] Galiani et al. (2016) also mention Cattaneo et al. (2006) but I could not find the paper. He reports that the paper analysed the performance of the Mexican "Iniciamos Tu Casa" program, which provided new houses to poor inhabitants located far from the city centre. A year later, the authors found that a large proportion of the participants had abandoned the houses, and those who remained complained that the new neighbourhoods provided them with poor access to public goods and general infrastructure.
[6] Only about 50% of beneficiaries come from informal settlements (GHS 2009).

The remainder of this paper proceeds as follows. The next section outlines the theoretical mechanisms through which the interactions between housing, location and labour supply are likely to occur. Section 1.3 describes the policy setting and the allocation mechanism. Section 1.4 outlines the empirical strategy, discusses the data and the validity of the research design. Section 1.5 presents the main results of the paper. Finally, Section 1.6 concludes.

1.2 Housing, Location, and Labour Supply

Many empirical and theoretical investigations have sought to elucidate the mechanisms by which low and high-income households sort within cities through the relative premiums they are willing to pay for intra-city differentials in location-specific characteristics (i.e. such as the level of public services and amenities, commuting times, housing quality, social networks, and access to employment opportunities).

The canonical model in urban economics, the "monocentric city" model associated with Alonso (1964), Mills (1967) and Muth (1969) (AMM), predicts that utility-maximizing households will sort according to their income elasticity of demand for housing and their opportunity cost of time (commuting costs). Spatial variation in land (house) prices arises out of competition for the most desirable sites which gives rise to an equilibrium price-gradient as well as a sorting equilibrium with low and high-income households living in different parts of the city. In its standard form[7], the model assumes that higher income households prefer consuming more housing, and thus benefit from living farther from the CBD.

Since AMM, different papers have put into question the dominance of the housing-force. LeRoy & Sonstelie (1983) and Glaeser et al. (2008) build on this idea and provide compelling evidence that the income elasticity of marginal commuting costs exceeds the income elasticity of housing demand, implying an indeterminate or reverse natural pattern of location by income. They put emphasis on transport mode choices as key determinants for residential sorting; arguing that better public transit provision in central cities attracts lower-income households. Brueckner & Rosenthal (2009) and Rosenthal (2014) have also shown that the quality of the housing stock (affordable housing supply) is a key determinant of residential location choices, while others have put emphasis on amenities (Brueckner et al. 1999; Albouy & Lue 2015).

[7] The model assumes that jobs are located in the CBD ("monocentric"), which has been one of its main criticisms. The main mechanism is then related to the dynamics between the cost of commuting – which increases with distance to CBD - and housing costs - which inversely fall with distance to CBD.

With this in mind, the overall effect of housing relocation programs will depend on the trade-offs poor households are willing to make when choosing between location-specific characteristics (such as amenities and distance-related costs and needs) and housing quality. These programs usually relocate poor households to the peripheries of cities where land costs are cheaper. If we consider the standard AMM model the first order source of compensation is through a price substitution effect. Mobile households that choose to relocate are compensated for higher commuting costs by lowering the cost of housing and increasing the quality of their housing consumption. In the case of developing countries, we can imagine that quality also comprises security of tenure and healthier environments (Galiani et al. 2016). For this to hold, the structures need to be supported by complementary physical infrastructure and social services such as roads and transport services, drainage, street lighting, electricity, together with policing, waste disposal and healthcare (Brueckner & Lall 2015).

This will not be true if poor households derive a higher marginal utility from being closer to the city centre. Distance results in higher frictions and inefficiencies in cities of developing countries, particularly in sub-Saharan Africa (World Bank 2017; Franklin 2017). Partly, these barriers result from the low provision of public services and transit, which exacerbates the farther away from central cities. This feature implies that poor households obtain a higher utility from being closer to central areas as they disproportionately benefit from public transit (in line with Glaeser et al. 2008). It also means they are more heavily burdened by living far. A large literature has put forward the importance of distance in creating frictions in urban labour markets (i.e. the spatial mismatch hypothesis; Kain 1968; Zenou 2009). The disconnect from job opportunities can introduce search inefficiencies either by restricting the size of the effective labour market or because of information distortions; it can increase reservation wages and jobs search costs due to high commuting costs; and employers may discriminate against workers living far away, as they may consider it deteriorates their productivity (Gobillon et al. 2011). As emphasized by Galiani et al. (2016), the emergence of slums testifies to the fact that the poor are willing to live in substandard housing conditions if this allows them to be close to employment opportunities in the city centre.

Under these conditions, we can imagine that the effectiveness of housing relocation programs will be closely related to the way they affect households labour supply. In a context where distance-related frictions are high, labour supply will be highly elastic with distance; which may in turn reduce the marginal utility poor households derive from better housing. The inverse is likely to be true in cases where distance-related-

frictions are low, and even when they are not, in cases in which households can easily adapt to residential relocation. This can happen when entire communities are moved (and social networks are kept intact). Takeuchi (2008) and Barnhardt et al. (2016) point to the importance of social and ethnic networks for employment and social services and the risks of distorting close-knit communities when relocation occurs. It can also happen if families can easily implement alternative livelihood strategies; for instance, if households have the possibility of shifting intensive and extensive margins between members or if they can easily switch to alternative sources of livelihood such as subsistence agriculture, rental income or more heavily rely on government transfers. The time dimension plays a key role here: the effects that may be true in the short to medium term could no longer hold in the long term, because of heterogeneous effects by age (Chetty et al. 2015) or different labour supply responses to the wealth effect (Moffit 2002, Jacob & Ludwig 2013)[8].

In the case of the South African program, we can imagine both forces exerting a pull. First, the design of the program combines a fixed ceiling for the total grant and minimum costs for the construction of the dwelling (dictated by a fixed standard), that squeezes the land cost and pushes developers to build in peripheral areas with low land prices (Lall et al. 2012). Because of this, relocation almost always increases distance to jobs. Still, households that derive a higher marginal utility from housing quality may still be better-off (AMM). The affordable housing sector is almost inexistent outside of the public sector, which makes benefitting from the subsidy virtually the only way of obtaining tenure security and accessing home-ownership for low-income households[9] (Lall et al. 2012). Close to 50% of beneficiaries come from informal settlements[10].

On the other hand, distance-related frictions and inefficiencies are accentuated by the spatial layout of South African cities. Largely, this results from path dependence in South African city-forms. South African apartheid policies constrained non-whites to

[8] Economic theory is ambiguous regarding the expected sign of any labour supply response to means-tested housing programs, mainly for at least two reasons. The first one involves how the dynamics of these programs relate to the life cycle model of labour supply; the second relates to its in-kind nature that offers a 'take it or leave it' level of housing consumption. The effect depends on how the program constrains consumption and whether the subsidized good is a complement or substitute to leisure (Jacob & Ludwig 2013).

[9] Limited access to land and housing finance, a highly regressive land taxation, and low supply elasticity of subsidized housing, has made it difficult for poor (and middle class) households to enter the formal housing market. The informal sector housing is a response to these failures (Lall et al. 2012). RDP corresponds to close to 25% of the housing stock in the country and virtually all the affordable formal housing sector. Further, there is a very limited rental market (25% in 2011) (Rust 2006).

[10] Ownership in the informal sector does not secure tenure. However, it might de facto not imply high insecurity due to the limited risk of eviction in South Africa. Informal dwellers are protected by the Constitution and past court rulings have always been in favor of slum dwellers.

live at the peripheries of cities far from work opportunities, often separated by landfill and geographical obstacles. While land-use restrictions have been abolished since the early 1990s, urban fragmentation remains the norm with the former white city concentrating most of the formal jobs, higher quality public services and high-income residential areas (Rospabe & Selod 2006). Commuting times and costs are extremely high in South African metropolitan areas. The average commuting time for employed workers was 84 minutes in 2013, but was much lower for the top income quintile (30 minutes, Kerr 2015). The latter also disproportionately reported using private vehicle (65% vs. 35% on average). Reflecting inadequate provision, less than 9% of commuters used public transport (train and buses) across urban areas, with nearly 50% choosing cars and minibus taxi (informal transportation). The geographical persistence of segregation and limited public transportation is one of the main causes behind the high levels of unemployment for the black population since the end of Apartheid (Banerjee et al. 2008).

Further, the possibility of shifting earning and livelihood strategies may be limited. Unlike other countries with similar official unemployment rates, the informal sector is small overall, partly due to legal barriers and job search mechanisms (Kingdon and Knight 2006). Subsistence agriculture is also a limited resource, particularly in urban areas. Whether by choice or by restriction, unemployment appears to be the default outcome of most individuals who do not find formal sector jobs (Magruder 2010). Location may thus disproportionately matter for poor households in South African metropolitan areas.

1.3 Policy Framework

1.3.1 History & main elements

South Africa's national housing policy was formally put in place following the end of Apartheid, to respond to the housing needs of the until-then formally excluded population. In 1997, the National Housing Department estimated that 2.2 million households were facing severe housing requirements, and included the right to adequate housing in the South African constitution (Sect. 26). A long list of regulations and laws[11] further defined the building blocks of today's housing strategy, and though the standards and implementation mechanisms have been revised throughout the years, the modality of allocation and final housing products have remained largely unchanged. By

[11] These include the National Housing Accord and the White Paper on Housing promulgated in 1994, the Housing Act of 1997, the Breaking New Ground Paper (2004), and the National Housing Code, 2009.

2014, 2.8 million dwellings had been constructed (table 1.1), amounting to 24% of the formal housing stock in the country[12]. According to the General Household Survey (GHS), 15.3% of South African households lived in a RDP or state-subsidised dwelling in 2014. There was a substantial increase in delivery during the period of this study, the number going from only 5.5% in 2002 to 9.4% in 2009, and 13.5% of households residing in RDP in 2012.

Its main components and the focus of this paper are the Housing Subsidy Schemes, popularly referred to as RDP housing projects, which provide qualifying households with the opportunity of owning their first house. The schemes comprise one-off capital subsidies[13] that are used for the construction of freestanding houses in new developments, administered by municipalities and later transferred to beneficiaries on an ownership basis. The purchase of land and construction of the houses are done by private sector operators, hired by local authorities (Lall et al. 2012). Because of this, the system is based on a fixed minimum cost for the house construction (R100-150 thousands, including land), tied to a minimum standard. Since 2009, the National Housing Code stipulates that all stand-alone RDP houses must at least have a minimum gross floor area of 40m^2, two bedrooms, combined living area and kitchen, and a separate bathroom with a toilet[14] (See Appendix A figure A1). The overall development needs to be 200-250m^2 with paved roads and electrical connexion (Tissington 2011). The design of the program has thus the effect of squeezing the price of land, forcing developers to build in low cost areas at the peripheries of cities.

The program uses means-testing to screen for eligible beneficiaries and allocation is made through waiting lists managed by municipal housing departments, ultimately responsible for transferring property deeds. RDP houses cannot be sold until after eight years since the time of procurement (Department of Human Settlements, DHS 2012). Since early 2000s, 6% of registered houses have been sold, but informal transactions (often without transferring titles) are estimated to be higher: for the period 2005-2011,

[12] The housing backlog increased since 1994, fuelled by a reduction of the housing supply at the lower-end of the affordable market and a rise in housing prices triggered by the high demand and the small size of the formal rental market (DHS 2012).

[13] There are in theory different types of subsidies through which it is possible to access subsidized housing (Individual subsidies, Consolidation subsidies, Institutional and Project-linked subsidies, and the People's Housing Process establishment grants). In practice, they are determined according to the same set of criteria, and because of the limitations of the low-cost housing market, more than 90% is delivered as a project-linked subsidy, the ones we focus on here. The rest is classified as an individual subsidy. In 1999 only 5% of RDP houses were in situ upgrading, the rest corresponded to greenfield developments (Khan and Thring 2003). Information is not available since, but conversations with officials at the municipality lead to believe the ratio remained largely unchanged.

[14] Before then, the typical size was a 25-35m^2 structure, with no separations until early 2000s, and a possible flush toilet (UN-Habitat 2008).

owners unofficially traded approximately 11% of all RDP houses (Urban Landmark 2011).

1.3.2 Allocation & waiting lists

Eligibility to receiving RDP housing is determined by two categories of conditions. The first ones are socio-demographic criteria: the recipient must be above 18 years of age, married, living with a partner, or otherwise have financial dependents, be of South African nationality (or permanent resident), and be a first time-owner. Additionally, it may not have benefited from a housing subsidy in the past. The second condition is an income-band criterion: the monthly household income must fall below R3501, making the housing program a means-tested benefit. Given the socio-demographic trends in South Africa, I consider the income criterion to be the main condition (figure A3)[15]. Yet, even in this case, this includes a large proportion of the population. In 1994 more than 80% of households fell under the eligible RDP income category (≤R3501). In 2012, the number decreased to 60% (DHS 2012). This dimension is reflected in the sample, overall in the period considered 45.44% of households have an income below R3501 (18.25% ≤R1500).

Since 2005, there are two official income bands for receiving RDP housing. The first one, consists of households with monthly incomes below or equal to R1500, and entitles the recipient to a full subsidized house without any contribution. The second band concerns households with monthly incomes between R1501-3500, and stipulates a one-off contribution of R2479 towards the purchase price of the property, payable to the municipality or to the provincial Housing Department. This condition was added to engender a "sense of ownership" and prevent sales below the market value. In practice, the contribution is not enforced and was progressively abandoned due to the difficulty for qualifying individuals to come up with the amount required (HDA 2011). This paper uses the discontinuity in the income criterion to identify the causal effect of receiving subsidized low-cost housing (see section 1.4.2 for more details on the income measure). I do not distinguish between both bands and use the R3500 threshold for the discontinuity. The main reason for this relates to the lack of enforcement of the distinction between bands in the period covered. This is reflected in the data as I do not observe a significant jump in the probability of receiving treatment at the R1500 (figure A5)[16].

[15] All figures and tables referenced with A are in Appendix A.
[16] The decline is only of 2-3% at R1500.

Generally, three phases can be identified in the allocation process. The first phase concerns the completion of a form at the provincial or municipal housing department, during which eligible households present proof of qualification. Screening is done by government officials based on the documentation supplied by the potential beneficiary. Details are checked against the National Housing Subsidy Database (NHSDB) as required by the National Housing Code. This database keeps records of all subsidy applicants approved by provinces across the country, with the purpose of preventing households from receiving more than one subsidy allocation. The information is recorded against the ID number of each individual (SERI 2013).

Once approved, individuals are allocated to what is commonly known as the "RDP waiting list" (second phase). The waiting list is the main mechanism through which qualifying households are granted a house. The list is administered at the municipal level, and allocation is made in date order of registration, as well as location, once a project is completed in the municipality where the application was made. Some municipalities have recently digitized their databases, and individuals can check whether they are part of one (but no information on time or rank is given). According to the General Household Survey, in 2009 the average time on the waiting list was 5 years. Because of the long waiting periods, eligibility criteria are re-assessed after successful allocation to a project. If still compliant, households are given a property (third phase).

The application procedure for accessing RDP housing is often characterized as long and cumbersome. This is partly due to a generalized take-up, with demand far exceeding supply throughout the entire existence of the program. In 2009, 13.5% of households had at least one member on a demand database. The proportion of households from informal settlements with a member in the RDP waiting list is higher, at 39% for the same year. I discuss the implications for identification related to the allocation mechanism and how I deal with these in the next section.

1.3.3 Implications for identification

The allocation mechanism has two main implications for identification.

The first one concerns the existence of crossovers, i.e. households benefiting from the program above the income cutoff. This implies that the probability of receiving treatment at the threshold does not jump from zero to one. One reason for this relates to the discretion that municipalities have in practice to allocate a proportion of the new houses to qualifying households from priority groups and catchment areas around the projects, depending on special needs (i.e. fires in informal settlement, community

decision)[17]. In principle, these will still be from the waiting list database and still in date order (Western Cape Government 2013), but the proportion varies by municipality and because of urgency, checks are not always conducted. A second reason for observing crossovers relates to the likelihood of subversion. The authorities have recorded different methods as means to 'jumping the queue'. These include land invasions, protest actions and violence, as well as bribes and connexions. While potentially more worrisome for identification - as subversion implies selection into treatment -, random subversion does not compromise the validity of the RD design (Lee and Lemieux 2010). Further, random subversion and crossovers are easily accommodated with Wald-type estimators of fuzzy RD designs, as used here. There is no reason to believe systematic subversion through the manipulation of the income receipts exists. While incentives to cheat are high, the relatively low level of the threshold – with two-thirds of the income distribution falling within the qualifying band, reduces the probability of large manipulations. Evidence in the dataset supports this: I find no bunching in the assignment variable at cutoff. I discuss this further in section 1.4.3.

The second implication concerns the time dimension, related to the long waiting period from application to assignment into new RDP houses. This can complicate identification in two ways. The first one through behavioural mechanisms that could bias the results, such as the anticipation of receiving a house in the future or the strong disincentive to earn above the income threshold to remain eligible once a house is assigned. The long waiting periods and the strong uncertainty related to actually benefitting from the subsidised program reduce the likelihood of an anticipation bias. The same logic is true with the disincentive to apply effort. The second issue with the time dimension is due to the fact that I do not observe the date at which households applied and registered in the waiting list of the housing database, and because the dataset is built with a 2 year-lag between observations, I do not know the exact moment when they move to the new house. To tackle this problem, I only use a subsample of the entire dataset available: I keep at baseline[18] only those households who fall under the socio-demographic conditions to receive RDP in the subsequent time periods. This also supposes the exclusion of any household that received RDP by the first year (i.e. I exclude all households treated at baseline). I then use the lagged monthly household income as assignment variable to best proxy for the time of application (and moving), which is likely to have happened prior to the period at which I observe a change in

[17] The proportions are ad-hoc and not publicly available.
[18] Baseline is defined as the first time a household is registered in dataset, 2008 for 92% of households. Only 3% in 2012.

RDP status. Doing this also allows me to limit any confounding effect of the policy on income, reducing endogeneity. As a robustness check I run regressions separately for households whose income during the entire period crosses-over the cutoff and for those whose income always remain on one side[19]. The proportion of income crossovers is small. Only 185 households cross over above the RDP threshold, among these only 45 are RDP beneficiaries. The opposite dynamic only concerns 25 households.

1.4 Empirical Strategy & Data

This section presents the methods and data employed to estimate the FRD. It also discusses the validity of the research design, potential sources of error and their implication.

1.4.1 The FRD design

The main threat to empirically identifying the causal impact of low-cost housing on labour outcomes is related to the non-random selection into the program. To address this issue, I exploit the arbitrary discontinuity income rule with a fuzzy regression discontinuity (FRD) design.

Given that I only observe the change in RDP status every two years, I use the two-year lagged household monthly nominal income (X_{hmt-1}) as the assignment variable to account for the uncertainty in the time of allocation and avoid any confounding problems with treatment; with X=3500 at cutoff (c). If common RD assumptions hold, and all and only eligible households before the cutoff obtained RDP housing, then the causal effect of subsidized housing would be given by the difference in outcomes Y_{hmt} between those with an income just above and just below R3500:

$$\lim_{x \to c-} E[\,Y_{hmt} | X_{hmt-1} = c] - \lim_{x \to c+} E[\,Y_{hmt} | X_{hmt-1} = c] = E[Y_{hmt}1 - Y_{hmt}0 | X_{hmt-1} = c] \quad (1.1)$$

In the case considered here, receiving a subsidy D_{hmt} is not deterministically related to crossing the threshold. For reasons discussed (i.e. crossovers), the jump in the probability of treatment at c does not go from zero to one (figure 1.1). In this 'fuzzy' RD setting, the causal effect is retrieved by dividing the jump in the relationship

[19] Results are unchanged in terms of sign of the point estimates but statistical power is significantly reduced when partitioning the sample. These results should be considered with caution as I impose selection into the sample and could potentially be excluding households that are better off as a result of treatment. They were circulated in an earlier version and are now available upon request.

between Y_{hmt} and the assignment variable $X_{hm,t-1}$ at c by the fraction induced to take-up the treatment at the threshold (Hahn et al. 2001) [20] :

$$T_{FRD} = \frac{\lim_{x \to c-} E[\, Y_{hmt} | X_{hmt-1} = c] - \lim_{x \to c+} E[\, Y_{hmt} | X_{hmt-1} = c]}{\lim_{x \to c-} E[\, D_{hmt} | X_{hmt-1} = c] - \lim_{x \to c+} E[\, D_{hmt} | X_{hmt-1} = c]} \quad (1.2)$$

The ratio (1.2) can be estimated using both parametric and non-parametric Wald-type estimators as long as the order of polynomial in the forcing variable and the data window are the same for the first and second stage outcomes. I estimate (1.2) by two-stage least squares, with the basic reduced-form and first-stage estimating equations respectively given by:

$$Y_{hmt} = \beta_0 + \beta_1 . Below_{hmt-1} + \sum_{s=1}^{S} \beta_s . (\tilde{X}_{hmt-1})^s + Below_{hmt-1} . \sum_{s=1}^{S} \pi_s . (\tilde{X}_{hmt-1})^s + v_{hmt}$$

(1.3)

$$D_{hmt} = \alpha_0 + \alpha_1 . Below_{hmt-1} + \sum_{s=1}^{S} \alpha_s . (\tilde{X}_{hmt-1})^s + Below_{hmt-1} . \sum_{s=1}^{S} \gamma_s . (\tilde{X}_{hmt-1})^s + \varepsilon_{hmt}$$

(1.4)

where, (as in the above) *h* indexes for households in metro-area *m* (1…6) at period *t* (1…4). The forcing variable is now $\tilde{X}_{imht-1}$ - the normalized value of the lagged monthly income with respect to the cutoff – so that the discontinuity occurs at zero; the variable Below is a dummy variable equal to one when the household is below the threshold in the previous period. I include *s-th* order polynomials in the assignment variable, and I allow the relationship between Y_{hmt} and $\tilde{X}_{hmt-1}$ to have different slopes on either side of the discontinuity. Other parameters are as previously defined. Following Gelman and Imbens (2014), I focus on first and second order polynomials in $\tilde{X}$ in the main specifications[21]. To avoid extrapolation bias in these global polynomial regressions, I restrict the sample by trimming off the tails of the income distribution (at 1 and 5 percent). Baseline estimates include time and metropolitan-area fixed effects, and I test the results to the inclusion of standard controls. These include the gender,

[20] They show that the FRD can be conceptualized as a local IV, and that the interpretation of the ratio for a causal effect requires the same assumptions as a regular IV at the local threshold, i.e. monotonicity and excludability.

[21] The Akaike information criteria (AIC) confirms that linear and second order polynomial are the best specifications at nearly all bandwidths considered. Results using larger order polynomials are available upon request.

age, ethnic group and education level of the household head, the number of members below 14 and above 68 years old. I cluster standard-errors at the household level.

As is best-practice, I also compute (1.2) using a non-parametric local linear regression specification with rectangular kernel weights and smaller bandwidths about the discontinuity[22]. For these, the property of the estimator depends crucially on the choice of bandwidth. I use both Imbens and Kalyanaraman (2012) and Cattaneo et al. (2014) bandwidths. These results are provided in section A.III of the online appendix. Since compliers are not located very near the threshold, the global polynomials approximations are my preferred estimates. The median income of compliers in the sample at baseline is R2435 and only 5.8% of compliers are within 10% of the income threshold. Including more observations below the cutoff reduces asymptotic size distortions (see figure A6 for estimates sensibility to the choice of bandwidths).

1.4.2 Data

This study utilizes data drawn from the first four waves (2008, 2010, 2012, and 2014) of the National Income Dynamics Study (NIDS) panel, conducted by Southern Africa Labour and Development Research Unit (SALDRU) at the University of Cape Town[23]. The panel collects information on demographic characteristics, dwellings, income, employment, health and wellbeing of the respondents. A two-stage cluster sample design was used to randomly select near 7300 households across 400 primary sampling units for the first wave, stratified by district council.

NIDS is a panel of continuing sample members (CSMs, i.e. individuals that were residents in participating households in wave 1); co-residents from wave 2 onwards are also re-interviewed if they do not abandon the household. Attrition is moderate for continuing sampling members (19% from wave 1 to 2 and 16% from wave 2 to 3), with the primary reason for no response being refusal and no contact. Tables A1 and A2 show the original sample of households' residents, and the redefinition of the sample with respect to the socio-demographic qualifying criteria of the RDP policy, as well as to all urban areas in the six largest metropolitan areas for which I conduct the analysis.

[22] Lee and Lemieux (2010) argue that ultimately there is no much difference between both methods, as they both have their source of bias and the best fit ultimately depends on the data. Here, I do not weight the data differently according to distance from the discontinuity, using rectangular kernels for the non-parametric estimations. The setting justifies this approach, as 25% of RDP recipients have an assignment variable below R1500.

[23] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008-2014, Waves 1 to 4 [dataset]. Cape Town: Southern Africa Labour and Development Research Unit, 2016 [producer]. Cape Town: Data First [distributor], 2016.

These are the city of Cape Town, Ekurhuleni, eThekwini, Johannesburg, Nelson Mandela Bay, and Tshwane (see figure A4). These cities are the only cities with populations above 1 million (Census 2011). The choice of restricting the analysis to these large urban areas is made for simplicity reasons. Firstly, the mechanisms are likely different in rural zones with different commuting behaviours and a higher likelihood of home agricultural production. Second, it would be very difficult and time consuming to obtain information on employment nodes and calculate distances from the home addresses to CBD for every secondary city in the sample. South Africa's six principal metro areas amount to 40% of the working age population, and receive more than 50% of RDP housing subsidies (Urban Landmark 2011). It is nonetheless important to keep in mind that results may well differ in other contexts.

My final sample is at the household level[24]. It leaves me with a total of 2,984 observations. I use this level of analysis given that the treatment occurs at the household level. Further, members' decisions to participate in the labour market are collective family decisions and possible shifts in livelihood strategies may occur following relocation. Relatedly, I use households' self-reported income as the forcing variable. This measure reflects regular nominal income received by all household members on a monthly basis, net of taxes. It includes both formal and informal incomes such as income received from helping friends, self-employment or subsistence agriculture[25]. Using this measure has the advantage of capturing all types of income of households. Drawbacks are discussed in the next section.

Information on whether a household received RDP housing in any of the subsequent periods is derived from the Household Questionnaire, which asks households if they received a government-housing subsidy to obtain the dwelling at each wave of NIDS. To correct for possible measurement error, I further refine recipients by crosschecking with the reported estimated subsidy amount and the estimated market value of the dwelling. I exclude those that report the market value of the dwelling above R300-450 thousands depending on district councils' differences in the distribution of housing prices[26]. The official estimated value of RDP houses is R200

[24] I consider a household to be the same across time when the household head does not change, or when after they die, their spouse becomes the new head. Given this, I construct household identifiers that are unique in time by keeping unchanged the identifiers of the first wave when this condition holds. Using different household identifiers every year slightly changes the standard errors but estimates remain unchanged.

[25] In cases of non-responses standard imputations were made. Refer to NIDS User Guides Public Release 2008-2014 for details.

[26] The benchmark is the value for the 75th percentile in the housing price distribution in each district council, as declared by households.

thousands, though in practice they can sell at 10% of that value. While the upper-end of the 'affordable' housing market in South Africa is capped at dwellings below R250 thousand, and some noise may exist from the larger ceiling chosen, it also prevents Type II errors (i.e. the exclusion of treated observations due to misreported households' estimations). As discussed, RDP eligibility is derived at the baseline. Nearly 15% of households in my final sample received RDP housing following the baseline period, in line with national averages. The proportions are stable across years.

Distance to the CBD and employment nodes are calculated as simple Euclidean distances between the household's address and the CBD or the main economic nodes' centroids. In the absence of supply-side data across cities, I use the 2013 South African National Household Travel Survey (NHTS) to identify main employment nodes by metropolitan area. The survey is designed to assess travel patterns in South Africa by Transport Analysis Zones (TAZ)[27]. I identify the main economic nodes by the density of daily commutes to each TAZ by metro area. I then create the weighted average distance to the primary (65%) and secondary (35%) 'destination for work' TAZ. I use the same weights across all metros, except for eThekwini where three nodes have the same weight as per the more polycentric nature of the city.

Table 1.2 contains basic summary statistics of households' socio-demographic and residential characteristics. Worth mentioning is the fact that most residential characteristics reflect the higher living standards of large urban areas, with 88% of households having access to electricity, and only 13% lacking weekly refuse collection. Housing informality is high and in line with national averages with close to 30% of dwellings identified as informal (it includes backyard dwellings). The same can be said for distances to CBD in minutes by mode of commute[28] and km (26.35 km).

Table 1.3 displays outcome means at baseline. I consider as employment any type of paid work (30% of the employed have no written contract and can be considered informal). Less than 50% of working age individuals are employed per household, which is also reflected in the low weekly labour hours (30 hours on average). This is not surprising given South Africa's high unemployment (25% across metro areas in the period) and relatively low labour force participation (55% in 2010). Only 7% of households report engaging in subsistence agriculture, and 35% receive a type of

[27] These were defined by Statistics South Africa as relevant travel micro-areas, and are created from the aggregation of census EAs (2011).
[28] I proxy time of commute by the average minutes per km of the main mode of transport used at corresponding TAZ of the household's residence, conditional on the household head's ethnic group and income-decile.

government grant. Regarding composition, 40% of households have children younger than 10 years old.

I follow Barnhardt et al. (2016) and group outcomes into thematic indices (table A3): labour supply and labour supply cost, amenities and neighbourhood quality. Each index is the simple average of the z-scores of their respective components. The labour supply cost index aims at quantifying the cost for households to participate in the labour market. The measure is imperfect, but the sign of the estimated coefficient will give an idea of the commuting burden following the assignment to an RDP dwelling. I will focus on actual distances and times in the analysis. The housing amenities index is computed to understand additional wellbeing dynamics related to low-cost housing. On average, the number of rooms per household is 3.6 and households rate their dwellings as structurally sound but requiring maintenance; 77% of households have access to flush toilets. Neighbourhood quality is defined by dummies of functioning streetlights and regular refuse collection, as well as a measure on the frequency of robberies and thefts. 71% of urban households declare having functioning streetlights and half of households perceive thefts as being 'common' in their neighbourhoods.

1.4.3 Discussion of validity

Identification requires three conditions. First is the absence of manipulation of the assignment variable around the cutoff, which can be formally tested by examining the density of its distribution about the discontinuity. Figure 1.2 shows the McCrary density plots (McCrary 2008) for wave one and subsequent waves of the panel. The test runs kernel local linear regressions of the log of the density separately on both sides of the threshold. I find no evidence of sorting, with the discontinuities statistically insignificant. Further, I compute local polynomial density estimates (Cattaneo, Jansson and Ma 2015) to test for the null hypothesis that the density of the assignment variable is continuous at cutoff. I confirm that there is no evidence of manipulation (table A4). Appendix Box A.1 addresses concerns regarding the fact that the assignment variable relies on self-reported income data. Even absent intentional misreporting or manipulation, self-reported income data is noisy and it is important to discard biased estimates that could result from bunching in the assignment variable due to its self-reported nature, i.e. it is easier for individuals to round the income they report (Barreca, Lindo and Waddell 2015). I find no evidence of systematic bunching at round numbers.

The second related key condition of a valid RD design is that 'all other factors' are continuous with respect to the threshold. If there was non-random sorting, we could

expect some of these characteristics to differ systematically between households immediately above and immediately below a given income threshold. Graphical tests using local linear polynomial regressions on lagged household characteristics are supportive of the identifying assumption (figure 1.3). Dashed lines show 95% confidence intervals, each point plots an average within a bin. I conduct a formal balance test by replacing the dependent variable in equations (1.3) and (1.4) with relevant observed lagged households' characteristics. The results indicate that these are well balanced on both sides of the cut-offs. The coefficients on below are typically small and statistically insignificant, with the few exceptions notably on informal dwelling which I include as control in relevant specifications.

Third, identification requires a strong relationship between the assignment and treatment variables, i.e. the conditional probability of receiving treatment should change discontinuously at the threshold. I examine graphically the first-stage relationship between being below the threshold and receiving RDP housing. Figure 1.1 shows local averages and linear fits at narrow bandwidths to plot RDP assignment in *t* against the monthly household income in *t-1*. I do not include controls or fixed-effects to transparently display the raw data. The likelihood of obtaining RDP decreases discontinuously at zero. Table 1.5 displays strong first-stage estimates for the preferred specifications, with F-Stats of up to 30.26. Figure 1.4 graphically explores the reduced-forms of selected outcome variables repeating the exercise done for assessing the continuity of control variables.

Some limitations of the identification strategy deserve discussion. As with many RD studies statistical power is an issue and weak identification when decreasing the sample size remains a problem due to the loss of efficiency and asymptotic size distortions in the standard errors (Marmer et al. 2014). While the presence of crossovers is well accommodated with the fuzzy design, it weakens the power of the identification with non-parametric estimates at narrow bandwidths. Figure A6 plots estimated non-parametric coefficients for different bandwidths: asymptotic size distortions are graphically visible but signs and sizes are consistent. Because of these limitations, I consider non-parametric estimates to be robustness checks. A second limitation concerns the external validity. As is the case with LATE, the estimated effects are based on *compliers* at cutoff and results are valid for the subpopulation affected by the instrument.

1.5 Results

1.5.1 Households labour supply

The paper's main findings concern the effects of obtaining RDP housing on the labour supply of households. Tables 1.6 and 1.7 contain these results. Each row reports estimates of equation (1.2) using parametric regressions and a polynomial of degree two. For robustness, I report estimates with linear polynomials (see section A.II in appendix A) and non-parametric local linear regressions (section A.III in appendix A). For guidance, OLS results and baseline means can also be found in appendix A, section A.I. The same applies for subsequent outcome variables.

Table 1.6 is divided in two parts. The first part (columns 1 to 3) contains the results on the labour supply index and its components, measured as standard deviations. Columns 4 to 9 display estimated coefficients on individual variables. I begin by examining the latter. I find a large decline in total hours of about 30 hours per week in treated households for the period considered. The sign and size are robust across specifications, with statistical significance ranging between 10 to 5% levels in parametric regressions. This is consistent with a decline in the overall number of employed household members (column 7) by almost one member, statistically significant at 5% level. The labour supply index reflects the same negative effect. Calculated as the average of the z-scores of weekly labour hours and the number of employed per working age members, it shows that households receiving RDP see their overall labour supply decline by almost one standard deviation. Point estimates are statistically significant at 10% level and driven by the negative effect on the intensive margin.

I also study the variation in impact across household members by gender (columns 5-6 for females and 8-9 for males). These regressions are identical to the pooled sample except that they now control for male and female working age members. The estimates reveal that most of the decline in labour supply happens through a reduction of total female hours. The effect is also larger (twice or three times as large) for females on the extensive margin but never statistically significant. While insignificant, male coefficients are also always negative.

Table 1.7 contains the results on unemployment. I use both a strict (columns 4 to 6) and large definition (columns 1 to 3). The strict measure excludes discouraged

searchers, meaning that it only includes those actively looking for a job. The coefficients for the total number of unemployed (large) are positive but non-significant across specifications; strict unemployment shows opposite signs but the same statistical insignificance. The size of both coefficients fails to compensate for the overall decline in labour supply. Still, their opposite dynamic suggests that the reduction in labour supply is more likely to end up in discouraged unemployment than in active search. The number of unemployed per working age member does register an increase of 0.26 pp for households below the threshold, statistically significant at 10% level. Despite the small size, it suggests that previously employed members are dropping out of the active labour market.

These findings are in line with Barnhardt et al. (2016). They also align with the conclusions of the U.S. literature that finds temporarily negative or no effect on the labour market outcomes of households benefitting for public housing programs (Mills et al 2006; Ludwig and Jacob 2012). As previously discussed, several hypotheses could explain the reduction in labour supply following relocation to RDP houses. The first possibility relates to the existence of high frictions in urban labour markets, which are exacerbated by moving farther away from city-centres. This could be amplified in cases where shifts in the livelihood strategies of families are limited. The fact that the bulk of the negative effect is driven by the reduction in the intensive margin of female members is telling. It could reflect both, the higher burden of distance on females (that might depend more on walking or social networks for work), or presumably also because of increased distance, a within household re-organization of the labour supply of members that pushes females to home-based work.

The alternative hypothesis is a price substitution or income effect arising from the wealth shock as predicted by standard labour supply models. According to the basic AMM model, the first order source of compensation for households choosing to relocate is through lower housing costs. Renters will presumably benefit the most from a potential income effect in the short-term. 74% of households at baseline pay rent (65% among dwellers of informal settlements). If this effect dominates, the re-allocation of female work to within the household may reflect South African social values. Caution is warranted when considering a possible income effect on labour supply here. Non-linearities may result from the in-kind nature of housing subsidies in the budget constraint (Moffit 2002). For instance, for many households relocating to RDP houses implies de facto higher levels of consumption of housing, as they now have to pay for utilities such as water and electricity. Because of this, while the income effect may

dominate in the short-term, it might not hold in the medium to long-term, and will depend on the differential between both past and present costs. I explore these alternative hypotheses further in the following sections.

1.5.2 Urban distances & commuting costs

Next, I examine the impact of obtaining RDP housing on commuting distances measured in time of commute (minutes)[29] and kilometres to CBD and main employment nodes. As discussed in previous sections, RDP houses are generally built at the peripheries of cities. On average, in the province of Gauteng where three of the metropolitan areas in the study are located, RDP houses were between 10 to 45 km from employment centres (figure A2). Average distance to CBD at baseline was already very high (26.8 km) so that any increase would potentially amplify distance-related frictions for relocating households.

Table 1.8 reports coefficients from equation (1.2) using global quadratic polynomial specifications on different measures of distance and commuting. Metropolitan-area and time fixed-effects should absorb any differences and changes in infrastructure and city-structures that are likely to affect commuting patterns. Column 1 contains the results for a composite index, resulting from the average of z-scores of distances to CBD in km and minutes and household monthly expenditure on transport. Columns 2 to 5 display estimated coefficients on each initial variable. I focus my analysis on the latter.

Overall, I find a large and positive increase in distances to CBD and employment nodes in kilometres and time of commuting. The surge in distances from work opportunities is large even for South African standards, between 12 to 13 km for households below the threshold. This surge is statistically significant at one percent level. The increase in commuting times is commensurate (close to 35 minutes), and close to 10% statistical significance. The sign and size are robust across global specifications, but double in size in nonparametric estimates suggesting some level of measurement error when reducing the sample size[30]. These positive values are translated into half a standard deviation increase of the composite index (labour supply cost), significant at 10% level. I find no effect on the monthly share of household revenue

[29] As explained earlier, to obtain the time of commute I proxy the household mode of transport by the preferred mode in the travel zone (TAZ) of the household residence, conditional on their monthly income and ethnic group. While it may seem a rough approximation, 71% of public transport users commute using minibus taxis across South Africa (Kerr 2015). I observe an extremely small variability in terms of choice by ethnic group, irrespective of their income. For instance, white South African disproportionately prefer to drive, while black South Africans predominately use minibus taxi.

[30] The small sample size prevents me from running the regression with distance stratifications. Approximations show no difference between groups originally further or closer to employment nodes.

spent on transport, and estimated coefficients are fairly small. The fact that increased distances do not translate into larger monthly expenses on transport suggests that households are not compensating by longer (more expensive) commutes, but rather choose to reduce their overall traveling.

These large and statistically significant increases in distance to city-centres suggest distance-related frictions might be an important factor behind the reduction of the labour supply of households. As postulated in the spatial mismatch hypothesis, the spatial disconnect between places of residence and the location of economic opportunities can increase search frictions and information distortions that rises both search costs for the unemployed, and reservation wages for the employed. This is consistent with findings in section 1.5.1. On the other hand, the reduction in the labour supply could also be indirectly related to distance, as households shift their earning strategies in response to the increased barriers to access labour markets. Both channels would explain the larger effect on the intensive margins of females. A nascent literature has shown that in poor settings the cost of commuting is higher for females who disproportionately choose to walk, limiting the size of their accessible labour market (Baker et al. 2005)[31]. Females might also depend more strongly on social networks, which may be disrupted with relocation. Due to increased commuting costs, it might thus be cheaper for women to stay at home, with households shifting their livelihood strategies in response. I explore this possibility next.

1.5.3 Family shifts in livelihood strategies

As discussed in section 1.5.2, in the presence of distance-related frictions, households may still be better-off when relocating if they can adapt their livelihood strategies. This section explores this hypothesis.

Table 1.11 reports estimated coefficients of equation (1.2) using global quadratic parametric specifications on different measures of the age composition of households. In the South African context, some papers have found that social benefits do alter household dynamics (Ardington et al. 2009). Examining the effect of RDP on overall household structures would go beyond the scope of this paper. Here, I rather focus on understanding whether household compositional changes in response to the increase in distances might be biasing the results on labour supply. This could arise in cases were

[31] Again, the small size of the sample prevents me from carrying a stratified analysis by gender without losing the first stage. At baseline, there are no indications that female-headed households face significantly longer distances and commute compared to male-headed households. On average distance to CBD for female vs. male headed households was 0.6 km higher (26.62 km), reflected by 5 minutes longer daily commuting times (120.44 minutes).

less mobile members (senior, children or disabled) are relocated to public housing while more mobile members remain in previous dwellings. Summary statistics do not suggest large differences in household composition for RDP-recipients and non-RDP recipients at baseline. The exception is the larger proportion of dependents and female-headed households in future RDP households, which reflects the policy of favouring vulnerable populations, and is the reason I include the gender of the household head and the proportion of members below 15 and above 68 years old as basic controls in all specifications.

Overall, I find no important changes on household composition when estimating LATE on dependency ratios (columns 1 and 2), the age and number of children living at home (columns 3, 4 and 7), the age of the household head (column 6), and the number of adults above 68 years old (column 5). The latter is the only statistically significant coefficient at 10% level, but displays a negative sign, in line with that of the aged dependency ratio. The interest of looking at dependency ratios is that they inform on overall changes in household members' age and indirect labour market abilities. None is statistically significant.

These results do not support the hypothesis according to which less mobile members are disproportionately relocated to RDP houses in response to accrued distances. Further, it does not support the possibility of women dropping out of the labour market because of pregnancy and having a larger number of young children. While some degree of household reconfiguration may still exist, they do not concern the overall age and work abilities of households, at least in the medium term considered here.

Table 1.12 explores alternative shifts in livelihood strategies. Columns (1) to (3) report point estimates of equation (1.2) with global quadratic polynomial specifications on three possible coping mechanisms: rental income, government transfers and subsistence agriculture. Only rental income is positive and statistically significant at 10% level. The size of the coefficient is large, with households below the threshold exhibiting an increase of 54 pp in their likelihood of receiving rental income. This result is interesting and should be considered carefully. There are two main ways through which low-income households in RDP houses could be extracting rental income. The first one is through the practice of backyard rooms, a distinct South African phenomenon by which low-income households build informal rooms in their backyards to rent to family

members or other individuals[32]. This is coherent with a change in income-generating strategies in response to relocation. The second is through the rental of their previous informal property. 20% of households in the sample report 'owning' their house in informal areas. These households could choose to relocate without giving up their previous dwelling. Lall et al. (2012) find evidence of very active informal housing markets in South African metropolitan areas. In this case, the decline in labour supply can presumably result from the price substitution and income effects. Unfortunately, the small size of the sample does not allow me to disentangle this channel further.

1.5.4 Amenities & neighbourhood quality

An important component to understand the overall effect of housing relocation programs is the actual quality of the new houses consumed. Despite the increase in commuting times and distances, households could still be better off when relocating if they derive a higher marginal utility from consuming more housing (AMM) or in this case, better housing (related to the possibility of home-ownership and tenure security, as well as less harmful geographical environments). Yet, for this condition to hold, the structures need to be supported by complementary physical infrastructure and social services such as roads and transport services, drainage, street lighting, electricity, together with policing, waste disposal and healthcare (Brueckner & Lall 2015).

To test if households are better-off in terms of living conditions, table 1.9 replicates the above regressions on an index of neighbourhood quality. The index is calculated as the average of the z-scores of three characteristics of neighbourhood quality: the perceived frequency of robberies, functioning street lighting and weekly refuse collection. These are partial measures of neighbourhood quality, but are still indicative of the general environment in relocation projects. Unfortunately, distances to bus and train stops were only collected at baseline and cannot be used to infer the effect of the policy on accessibility to public transport. Coefficients in columns (2) to (5) control for households previously living in informal housing.

Overall, I find no effect on neighbourhood quality, and none of its components are statistically significant. There are, however, telling differences regarding their signs, which suggest a consistent improvement in refuse collection across specifications (column 5). The effect on functioning street lights is ambiguous across specifications, and large standard-errors do not allow me to draw any conclusive remark. More

[32] For anecdotal and visual evidence see: http://informalcity.co.za/learning-from-backyard. (last visited May 17th 2017).

interesting, pulling the average in the other direction, is the steady perceived increase in thefts and robberies of about half a standard deviation (column 3). Though insignificant, the perception of security matters. The *deterioration* is revealing of the higher insecurity that may result from isolation, but also from the disruption of social networks in close-knit communities. While the evidence here is insufficient, part of the decline in labour supply could be related to the loss of informal business partners, day-care providers and security mechanisms put in place in previous locations. The loss of social networks is one of the main constraints on these types of relocation programs in both developed and developing countries (Barnhardt et al. 2016; Day and Cervero 2010; Field et al. 2008; Mills et al. 2006; Kling et al. 2007).

Table 1.10 presents the results of obtaining RDP housing on a housing amenities index. The index is calculated as the average of the normalized values of four measures of housing quality: the number of rooms, access to electricity, a categorical variable of dwelling quality and the type of toilet facility[33]. Results for the overall index are in columns (1) and (2), with the different components in columns (3) to (6). Columns (2) to (6) additionally control for the informal nature of the previous dwelling. Counterintuitively, I find a decline in housing amenities for households below the threshold of up to 0.68 standard deviations, statistically significant at 5% levels. While all coefficients in columns (3) to (6) have a negative sign, only the number of rooms and the perceived quality of the structure are statistically significant at 10% level.

Two remarks can be derived from these findings. First, while surprisingly negative, results on housing amenities are in line with anecdotal evidence. Officially, 14.5% of RDP residents have complained about the quality of the walls, while 13.9% regard the roofs as weak or very weak (GHS 2014). The negative perceptions are higher in press reports. Khan & Thring (2003) report that peripheral sites often fail to include the range of necessary public facilities and amenities. The high cost of services, which many RDP recipients previously did not pay, has also led to cases of up to 80% non-payment where entire projects have been disconnected from electricity grids (SERI 2013). Short-term mechanisms could also be at play by which delays in servicing in recently constructed areas or expectations about the new dwelling negatively bias perceptions. Still, the fact that RDP houses and informal shacks show no significant difference in terms of their valuation is indicative (Lall et al. 2012). It suggests the quality of amenities do not compensate for distance.

[33] A measure of access to water sources would have been ideal but the variable in the NIDS panel is highly imprecise.

The second remark concerns the hypothesis tested here. It is not clear that despite of the bad quality of the housing good, *intangible quality* has not increased. For a large majority of households, RDP housing offers the only option for becoming a homeowner and obtaining tenure security. In line with this, Galiani et al. (2016) find positive effects on subjective wellbeing for slum upgradings in South America. While I cannot measure subjective appreciation for tenure and ownership in the dataset, the mechanism may still explain why households choose to relocate despite not being better-off in terms of tangible quality and facing increased frictions to access labour markets.

To approximate a measure of subjective wellbeing, I test the effect of RDP on the household head's preference to continue living in the current area. Overall, the preference to stay is statistically significant and negative below the threshold (column 5). Families have an 88 pp higher preference to leave their current location at cutoff, significant at 10% level. However, when grouping the results by age groups (columns 6 and 7) it becomes clear that the preference to leave is only driven by heads aged 30 or below, for which point estimates are large, consistently negative across specifications and statistically significant at 10% levels. These results could suggest that older individuals obtain a higher marginal utility from ownership and security of tenure with respect to young adults, for whom the disutility of being far from economic opportunities dominates. In their paper, Barnhardt et al. (2016) find suggestive evidence that greater average distances to employment opportunities discourage adult children from staying in their parents' household. In the relatively short-term period analysed here, one could imagine similar dynamics.

1.6 Conclusions

This study has focused on identifying the reduced-form impacts of a large public housing program on the labour supply and living conditions of beneficiary low-income households across South Africa's six largest metropolitan areas. Albeit incomplete, findings here provide compelling causal evidence about a controversial type of housing policy and contribute to the better understanding of the response of low-income households to housing policies in developing countries.

Overall, I find that in the medium to short-term (two to four years) following the relocation to RDP housing, households are not better-off on several outcomes. Their labour supply is reduced at cutoff both in terms of hours and participation. The effect is

mostly driven by a reduction at the intensive margin of female members, though male members also experience a non-significant reduction. The share of working age members that are unemployed and discouraged also rises, suggesting members are dropping out of the active labour market. The surge in distances from the CBD and employment nodes of up to 12 to 13 km, offer a possible explanation for the negative impact on labour outcomes. It supports the existence of high frictions in South Africa urban labour markets that are amplified by distance. I find suggestive evidence that households may be shifting their livelihood strategies in response to the accrued distance, by increasing their rental incomes. This could also result from a price-substitution effect.

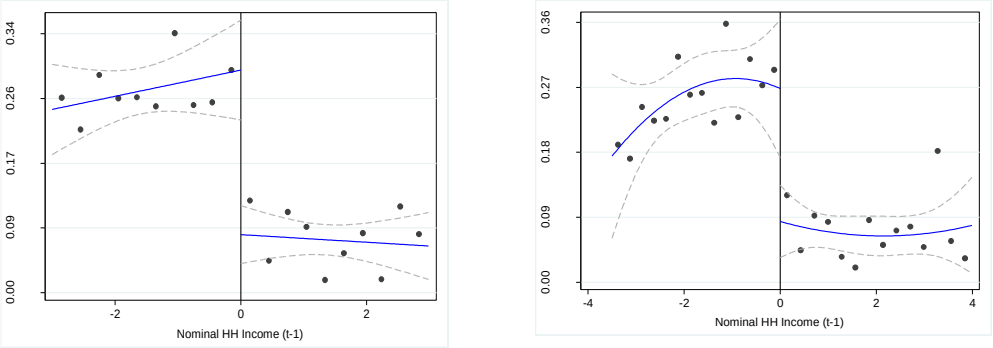
Whether the labour supply decline is due to the direct or indirect effect of longer distances is impossible to tell in the current setting. Still, both possibilities underline one of the main policy conclusions of the paper: location matters. Poor households are highly penalized in contexts of low public transit provision and inefficient labour markets as is often the case in most developing country cities. Under these conditions, housing policies should prioritize the reduction of distance-related frictions that push households to having to choose between housing quality and proximity to employment opportunities. Many options exist: relocation programs could be accompanied by complementary connectivity policies; local authorities could prioritize in-situ upgradings (Takeuchi et al 2008); or include monetary compensations (Lall et al. 2012). The possibility to implement alternative livelihood strategies in response to relocation is also important. This can be better achieved when social networks are not disrupted (Barnhardt et al. 2016).

An additional interesting finding of the paper concerns housing amenities and neighbourhood quality. Overall, I find a deterioration of the former and no effect on the latter. At first, this seems counterintuitive and contrary to the predictions of the standards monocentric model of urban economics, with households not being compensated by the increase in distance by a higher consumption of housing. The fact that mobile households still choose to relocate suggests that the improvement in housing consumption happens through intangible features of housing quality, such as increased tenure security and the possibility of becoming homeowners. Subjective wellbeing has been found to be an important component of housing policies (Galiani et al. 2016). This dimension adds to the complexities of the design of such policies, and the need to continue understanding household responses to different housing policies in developing countries. The difference between short and long term effects should also

be a priority for further research, as dynamics could well differ in the long run, particularly when considering the effect on young children and possible positive externalities on health outcomes.

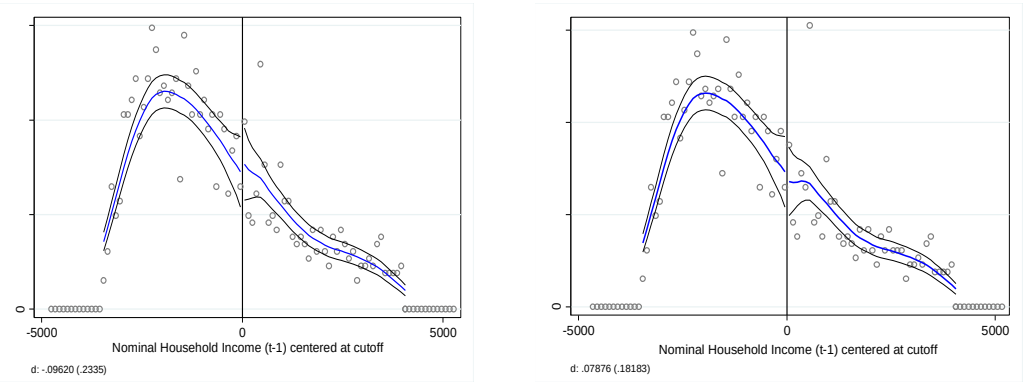
1.7 Tables & Figures

Figure 1.1. First Stage



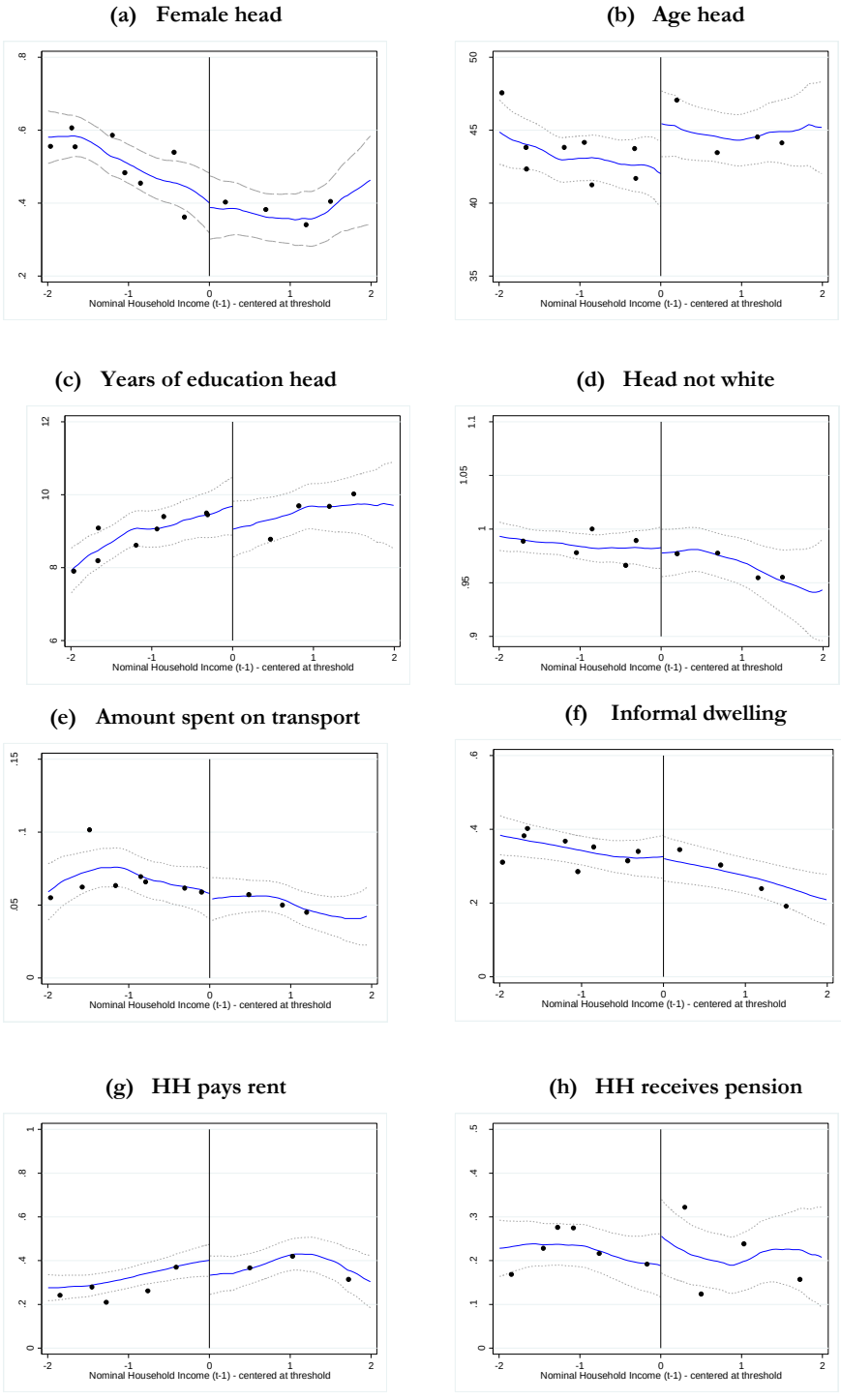
Notes: Each point plots an average value within a bin. The solid line plots a quadratic fit on the right-quadrant and a linear fit on the left-quadrant. Dashed lines show 95% confidence intervals.

Figure 1.2. McCrary Density Tests of Non-Manipulation of the Assignment Variable



Notes: Nominal Household Income, Wave 1 on the right quadrant. All other waves in left quadrant. The d. estimate fails to reject the null hypothesis of non-discontinuity at cutoff. Generated using the Stata program developed by McCrary (2008)

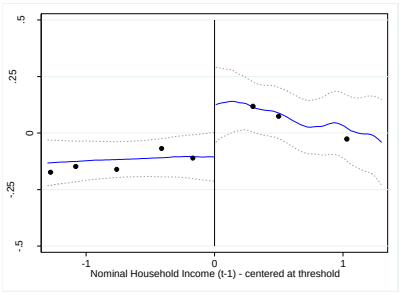
Figure 1.3. Balance Checks (t-1)



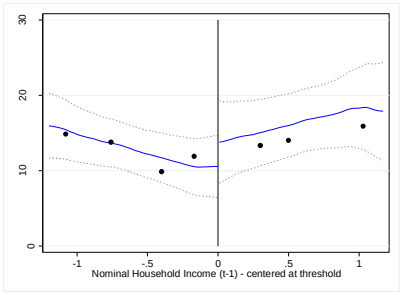
Notes: Each point plots an average value within a bin. The solid line plots local polynomial fits of degree 2. Controls and fixed-effects have been partialed out. Dashed lines show 95% confidence intervals. For variables definitions see table 2. All variables are lagged household characteristics.

Figure 1.4. Reduced Forms

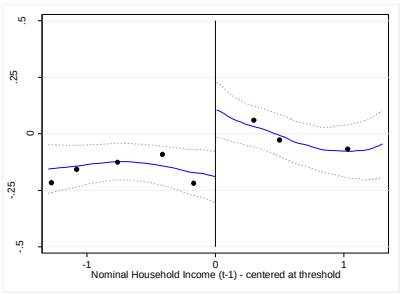
(a) Labour Supply Index



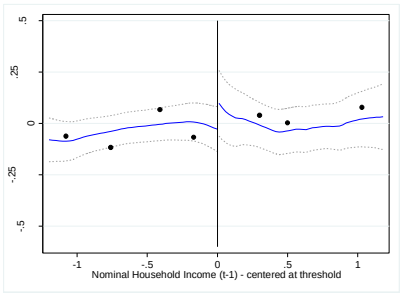
(b) Weekly hours of Females



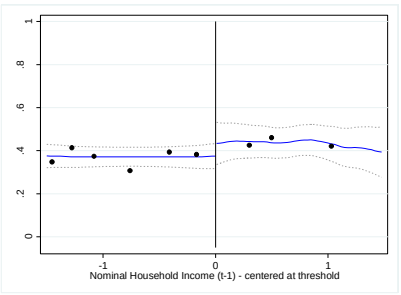
(c) Housing Amenities Index



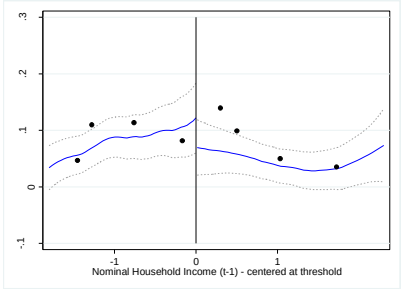
(d) Neighbourhood Quality Index



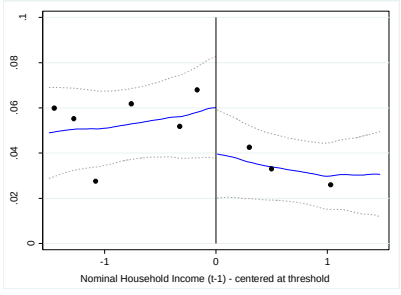
(e) Young children (<10 years)



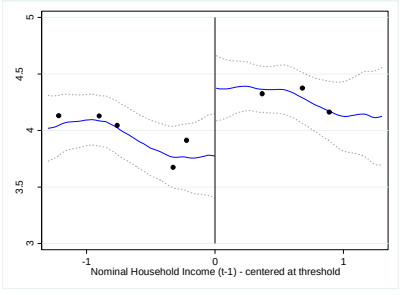
(f) Receives Rental Income



(g) Unemployed (wam)



(h) Preference to Stay (<30 years)



Notes: Each point plots an average value within a bin. The solid line plots local polynomial estimates. Controls and fixed-effects have been partialed out. Dashed lines show 95% confidence intervals. For variables definitions see table 3.

Table 1.1 RDP Houses Delivered since 1994

Period	Houses/Units Completed
1995-2000	870,629
2000-2005	745,023
2005-2010	756,119
2011-2014	463,504
Total	**2,835,275**

Notes Data from National Department of Human Settlements (2014).

Table 1.2. Summary Statistics at HH Level, All waves

	Mean	*Std. Dev.*	*N*
I. Households Socio-Demographics			
Female head	0.438	0.496	2,984
Age Head	46.863	14.782	2,984
Household Size	3.555	2.614	2,984
Head Not White	0.863	0.344	2,984
Head Completed Matric	0.336	0.472	2,984
Proportion of children (<15 years old)	0.197	0.221	2,984
Proportion of senior members (>68 years old)	0.054	0.177	2,984
Dependency ratio	0.477	0.627	2,984
Dependency ratio children	0.413	0.576	2,984
Dependency ratio - aged	0.065	0.228	2,984
Fraction of employed HH members	0.472	0.405	2,984
Received Housing Subsidy (RDP)*	0.142	0.364	2,180
Receives Government Grant	0.405	0.491	2,984
HH monthly expenditure (nominal Rands)	6,492	12,277	2,984
HH engages in non-commercial agriculture	0.041	0.198	2,029
II. Residential Characteristics (Households)			
Refuse Collection	0.879	0.327	2,454
Access to Electricity	0.875	0.331	2,430
Informal Dwelling (large definition)	0.278	0.448	2,566
Not Common to have thefts in neighbourhood	0.494	0.500	2,434
Amount Spent on Transport in last 30 days (% of total)	0.070	0.102	1,180
Time taken to nearest bus by foot (wave 1)	13.331	17.960	602
Time taken to nearest mini taxi by foot (wave 1)	10.159	8.564	602
Average Distance to CBD (km)	26.347	16.237	1,793
Average Distance to CBD (min. by mode of transport)	117.943	63.586	1,793
Average Distance to Main Employment Nodes(km)	27.286	15.713	1,793
Pays Rent	0.340	0.474	1,986
Has a Home Loan	0.104	0.306	2,245

Notes: Descriptive statistics for final subsample of RDP eligible households in urban areas for which the assignment and treatment variables are not empty. The proportion of children and senior members is calculated as their share over the household size. RDP excludes wave 1. Government grants include all kind of grants from disability to child support. Electricity refers to access to electricity in the dwelling. Refuse collection refers to having weekly removal by local authorities. I use the large definition of informal dwelling that includes backyard-living. Average distance to CBD in minutes is calculated proxying the mode of transport by the preferred mode used in the area by race and income level according to the National Travel Survey 2013. All distance variables are only available for period 1-3.

Table 1.3. Summary Statistics - Main Outcomes (Baseline)

	Mean	*Std. Dev.*	*N*
I. Household Labour Supply			
Number of employed HH members	0.815	0.776	723
Fraction of employed over wam	0.410	0.381	723
Fraction of female employed	0.465	0.601	723
Fraction of male employed	0.493	0.566	723
Weekly labour hours (per household)	30.311	34.310	723
Weekly labour hours per working age member	25.724	15.571	723
Weekly labour hours for female members	13.541	22.924	723
Weekly labour hours for male members	16.770	25.396	723
Adult Labour Supply (index)	0.098	1.000	692
Number of unemployed	0.472	0.776	723
Fraction of unemployed over wam	0.179	0.287	723
Number of unemployed (strict)	0.316	0.643	723
Fraction of unemployed over wam (strict)	0.117	0.224	723
II. Other Outcomes			
Adult Labour Supply Cost (index)	-0.082	1.000	692
Average distance to CBD (km)	26.823	16.897	723
Average distance to CBD (min)	116.370	64.580	723
Average distance to Employment Nodes (km)	27.809	16.389	723
Amount spent on transport last 30 days	0.108	0.136	366
Street Light	0.714	0.452	723
Refuse Collection	0.777	0.417	723
Not Common to have thefts in neighbourhood	0.570	0.495	723
Access to Electricity	0.847	0.361	723
# of children	0.962	1.327	723
# of adults > 68 years old	0.118	0.366	723
HH has children younger than 10 years old	0.399	0.49	723
HH engages in non-commercial agriculture	0.070	0.256	723
HH receives rental income	0.072	0.258	723
HH receives government grant	0.355	0.479	723
HH head prefers to stay at current location	0.666	0.472	723
Happier than 10 years ago	0.442	0.497	723

Notes: This is the final subsample of RDP eligible households in urban areas in period 1. Employed correspond to the large definition that considers any type of paid work outside of the home. It includes informal work. Unemployed strict excludes discouraged individuals, and only includes those actively looking for a job. Wam stands for working age members, as those between 15 and 68 years old. The Adult labour supply index and Labour Supply Cost index are as define in table A.3 in Appendix I. Average distances to CBD and main employment nodes are calculated as Euclidean distances. Average distance to CBD in minutes is estimated by proxying the time taken by the preferred mode of transport in the commuting zone of the household residence, conditional on income and ethnic group. Amount spent on transport is calculated as a share of total monthly expenditures. These have been extracted from the National Travel Survey 2013. Variables of neighbourhood amenities and government grant are as defined in table 2.

Table 1.4. Balance of Households Characteristics in the preceding years of RDP allocation

	Global Polynomials		Local Polynomials	
	Linear	Quadratic		
	(1)	(2)	(3)	(4)
A. Each Baseline Characteristics Separately				
Age of head	-0.21	-0.81	-1.60	0.15
	(0.72)	(0.76)	(1.14)	(1.03)
Head Not White	-0.02	0.084	-0.059	-0.079
	(0.03)	(0.11)	(0.06)	(0.10)
Head years of Education	-1.48	1.37	1.97	5.81
	(1.07)	(3.40)	(3.78)	(5.15)
Female head	0.25	-0.31	-0.11	-0.47
	(0.23)	(0.36)	(0.42)	(0.54)
Informal dwelling	0.42*	1.00**	0.23	-0.13
	(0.23)	(0.44)	(0.39)	(0.45)
Female members (total)	-0.33	-0.60	-1.01**	-1.58
	(0.33)	(0.43)	(0.59)	(1.08)
Amount spent on transport (monthly)	0.11	0.45	0.25	0.06
	(0.09)	(0.34)	(0.80)	(0.21)
Number of Adults>68 years	-0.31	-0.09	-0.07	-0.31
	(0.20)	(0.21)	(0.30)	(0.42)
Number of children living at home	-0.45*	-0.60*	-0.43	-0.67
	(0.25)	(0.35)	(0.46)	(0.59)
Household Pays Rent	-0.49	-0.56	0.01	-0.73
	(0.31)	(0.44)	(0.54)	(0.65)
Time &City fixed-effects	Y	Y	Y	Y
First Stage F-Statistic	28.84	14.84	9.66	9.95
Bandwidth				
IK (2012)			√	
CCT (2014)				√
Obs.	1,946	1,946	1,040	498

Notes: This table contains parametric and nonparametric regression discontinuity estimates to assess the difference of baseline characteristics of RDP eligible households. Panel A evaluates separately each characteristic. Informal dwelling includes backyard dwellings. The monthly amount spent on transport is a share of monthly income. All outcomes are lagged characteristics. Robust standard errors clustered at household level in parenthesis. *p<0.10; **p<0.05; ***p<0.01

Table 1.5. First Stage

HH Level dependent variable is RDP				
	Global RD		Local RD	
	(1)	(2)	(3)	(4)
HH income≤3500 in t-1, [Below]	0.1662***	0.1654***	0.1658***	0.2141***
	(0.0425)	(0.0301)	(0.0539)	(0.0662)
First Stage F-Statistic	15.86	30.26	9.46	10.46
Polynomial Order	2	1	1	1
Time & City fixed-effects	Y	Y	Y	Y
Controls	Y	Y	Y	Y
Obs.	1,960	1,960	1,041	489

Notes: The table reports the estimated discontinuity (2SLS) first stage. The dependent variable is a dummy for allocation of RDP to the household. Below is an indicator equal to one if the household nominal income is below the threshold in t-1. The regression also includes in columns 1 and 2, the interaction of below and the assignment variable. Robust standard errors, clustered at household level in parentheses. Controls include age, population group, education level and gender of the household head, as well as the proportion of adults>68 & children <14. *p<0.10; **p<0.05; ***p<0.01

Table 1.6. Discontinuity Effect of RDP on Main Labour Market Outcomes (1)

	Household Level dependent variable is:								
	Labour Supply Index			Main Outcomes					
	Composite Index	Employed per wam	Weekly hours per wam	Weekly Total Hours	Weekly Hours Female	Weekly Hours Male	Employed Members	Employed Females	Employed Males
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RDP	-0.94*	-1.09	-0.76*	-33.35*	-26.97**	-6.04	-0.96**	-0.76	-0.21
	(0.5255)	(0.7534)	(0.4475)	(19.2935)	(13.3795)	(15.0422)	(0.4353)	(0.4942)	(0.3390)
First Stage F-Statistic	15.86	15.86	15.86	15.86	15.86	15.86	15.86	10.91	12.42
Polynomial Order	2	2	2	2	2	2	2	2	2
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	N	Y	Y	Y
Obs.	1,960	1,960	1,960	1,960	1,960	1,960	1,960	1,728	1,819

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on labour market outcomes of beneficiary households. Dependent variables are measures of the labor supply of households. Columns (1) to (3) contain the labour supply index and its two components, measured as z-scores. All intensive measures are weekly total hours by household members, and extensive margins are the number of employed members. Wam stands per working-age members. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Columns (4) and (5) control for the number of working age members in the household, while columns (6) to (9) control for the number of female and male working age members in the households, separately. *p<0.10; **p<0.05; ***p<0.01

Table 1.7. Discontinuity Effect of RDP on Main Labour Market Outcomes (2)

	Household Level dependent variable is:					
	Unemployed Members	Unemployed Members	Unemployed per wam	Unemployed - strict	Unemployed - strict	Unemployed per wam-strict
	(1)	(2)	(3)	(4)	(5)	(6)
RDP	0.67	0.49	0.26*	-0.53	-0.41	-0.08
	(0.3744)	(0.4001)	(0.1358)	(0.3904)	(0.4346)	(0.1589)
First Stage F-Statistic	15.86	15.32	15.86	15.86	15.32	15.86
Polynomial Order	2	2	2	2	2	2
Time & City FE	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Obs.	1,960	1,501	1,960	1,960	1,501	1,960

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on labour market outcomes of beneficiary households. Dependent variables are measures unemployment. Columns (1-3) include discouraged members, while the reminder is only strict unemployment defined as those actively looking for a job. Wam stands per working-age members. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. In columns (1) and (3) I control for the number of working age members in the household. Results in columns (2) and (5) I control for female working age members. *p<0.10; **p<0.05; ***p<0.01

Table 1.8. Discontinuity Effect of RDP on Household Commuting Distances & Time

	Household Level dependent variable is:				
	Composite Index	Distance (km) to CBD	Distance (km) to main nodes	Distance (min) to CBD	Amount spent on transport
	(1)	(2)	(3)	(4)	(5)
RDP	0.53*	13.41***	12.19***	34.72	0.041
	(0.272)	(4.417)	(4.285)	(22.972)	(0.476)
First Stage F-Statistic	12.43	12.43	12.43	12.43	12.84
Polynomial Order	2	2	2	2	2
Time & City fixed effects	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Obs.	2,741	2,741	2,741	2,741	736

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on elements related to the labour supply cost of households. Dependent variables are measures of the labour supply cost of households, with in column (1) the labour supply cost index. Columns (2-4) are the components of the index, and column (5) contains the monthly amount spent on transport as a share of monthly expenditures. Very few households have answered this question, limiting the sample size. Distance to CBD and main nodes are Euclidean distances to CBD (and the secondary node) from the suburb's centroid of the household residence. Distances measured in time of commute are calculated using the time of the most frequently used mode in the commuting area of residence, conditional on income and population group. Column (1) is an average of the normalized values of columns (2) and (4). Here I display the non-normalized value to provide the explicit km and minutes. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. Due to limitations to access secure address identifiers, these regressions are run on a different sample - i.e., for periods 1-3 only. They control only for household size. *p<0.10; **p<0.05; ***p<0.01

Table 1.9. Discontinuity Effect of RDP on *Wellbeing* (1): Neighbourhood Quality

	Household Level dependent variable is:				
	Composite Index	Composite Index	Robberies uncommon	Street Light	Refuse collection
	(1)	(2)	(3)	(4)	(5)
RDP	-0.32	-0.16	-0.50	-0.25	0.19
	(0.3932)	(0.3517)	(0.5864)	(0.5858)	(0.5344)
First Stage F-Statistic	12.72	14.38	14.38	14.38	14.38
Polynomial Order	2	2	2	2	2
Time & City fixed effects	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Obs.	1,947	1,947	1,947	1,947	1,947

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include the neighbourhood quality index and its components in columns (3 to 5). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Columns 2-5 also include a dummy variable for living in informal dwelling in the previous period. *p<0.10; **p<0.05; ***p<0.01

Table 1.10. Discontinuity Effect of RDP on *Wellbeing* (2): Housing Amenities

	Household Level dependent variable is:					
	Composite Index	Composite Index	# of rooms	Dwelling Quality	HH has electricity	Toilet inside
	(1)	(2)	(3)	(4)	(5)	(6)
RDP	-1.06**	-0.68*	-0.85*	-1.21*	-0.31	-0.32
	(0.5140)	(0.3816)	(0.5181)	(0.6314)	(0.6396)	(0.5324)
First Stage F-Statistic	14.74	15.86	15.86	15.86	15.86	15.86
Polynomial Order	2	2	2	2	2	2
Time & City fixed effects	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Obs.	1,960	1,960	1,960	1,960	1,960	1,960

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include a housing amenities index and its components in columns (3) to (6). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Columns 2-6 also include a dummy variable for living in informal dwelling in the previous period. *p<0.10; **p<0.05; ***p<0.01

Table 1.11. Discontinuity Effect of RDP on Household Shifts Strategies (1): Compositional Changes

	Household Level dependent variable is:						
	Dependency Ratio Children	Dependency Ratio Aged	# Children living at home	Young Children	# of >68 years old	Age of Head	Age of children
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RDP	0.19	-0.07	0.21	-0.20	-0.10*	0.08	2.63
	(0.2879)	(0.0969)	(0.2728)	(0.2757)	(0.0535)	(0.6256)	(3.2253)
First Stage F-Statistic	13.09	13.09	14.44	14.44	14.44	14.44	14.44
Polynomial Order	2	2	2	2	2	2	2
Time & City fixed-effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	1,914	1,914	1,949	1,949	1,949	1,949	1,949

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on household shifts strategies. Dependent variables are measure different measures of household composition. Columns (1) and (2) are dependency ratios calculated as the share of aged <14 & >68 over the working age members. Young children is a dummy variable for household with children below the age of 10. Columns (6) and (7) are the average age of children and the age of the household head, respectively. All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head. *p<0.10; **p<0.05; ***p<0.01

Table 1.12. Discontinuity Effect of RDP on Household Shifts Strategies (2): Coping Mechanisms & Preferences

	Household Level dependent variable is:						
	Coping Mechanisms			Preferences			
	HH Receives Rental Income	Government Grant	Non-commercial Agriculture	Happier than 10 years ago	HH prefers to stay	HH prefers to stay (age<30)	HH prefers to stay (age>30)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RDP	0.54*	-0.09	0.11	-0.82*	-0.82*	-0.67*	-0.18
	(0.2966)	(0.2743)	(0.1205)	(0.4329)	(0.4618)	(0.3990)	(0.3571)
First Stage F-Statistic	15.86	14.44	14.20	11.56	9.86	9.86	9.86
Polynomial Order	2	2	2	2	2	2	2
Time & City fixed-effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	1,960	1,949	1,909	1,824	1,612	1,612	1,612

Notes: The table reports 2SLS estimated coefficients of the effect of RDP subsidies on household shifts strategies. Dependent variables are different measures of household related to possible coping strategies and preferences. Government Grand in column (2) is a dummy variable for receiving any type of government grant, including disability grants. Non-commercial agriculture is a dummy variable equal to one if households participated in agricultural activities without monetary compensation. While the remainder variables are dummy variables for expressing happiness with respect to 10 years ago, and preference to stay in the current place, by age of the household head. Column (5) is the average of columns (6) and (7). All regressions control for a quadratic polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. *p<0.10; **p<0.05; ***p<0.01

Chapter 2

Is there one Ghetto University? Neighbourhoods & Opportunities in a Developing City[1].

2.1 Introduction

Neighbourhoods where children grow up can play a significant part in shaping their opportunities later in life (Glaeser & Cutler 1997; Oreopoulos 2003; Chetty et al. 2015). Growing up in deprived or 'bad' neighbourhoods, i.e. areas with high levels of unemployment, poverty, criminality, and low-quality local public goods, can be detrimental in adulthood (Case & Katz 1991; Crane 1991). How they affect socioeconomic mobility is central to understanding the persistence of inequality over time (across generations) and space (poverty traps). This question is particularly relevant in developing countries, where cities display high levels of income inequalities, emphasized by stark spatial differences in the access to public services, formal housing, formal jobs, and consumption amenities. There is extremely little empirical evidence of neighbourhood effects in this context. This paper addresses this gap, by using the quasi-natural social experiment of apartheid in South Africa to measure the effects of growing up in disadvantaged neighbourhoods or ghettos[2]. It also provides suggestive evidence on the main channels through which some neighbourhoods may be better than others for the life trajectories of children.

[1] I am thankful for discussions with Thomas Piketty, Yanos Zylberberg, Pascal Jaupart, Ana Moreno-Monroy, Felipe Carozzi, and Nicola Branson, as well as participants at LSE seminars. I also thank Timothy Brophy and Daniel de Kadt for assistance with geo-located census data, and the Municipality of Cape Town for granting me access to their library and data on apartheid in the city.

[2] I understand ghettos here as disadvantaged residential urban areas that are enclaves of high poverty and have a high concentration of ethnic minority (Zenou 2009b). In the case of South Africa, paradoxically, black/Africans are the majority of the South African population (80.8% in 2016, according to mid-year population estimates by Statistics South Africa (SSA)). In Cape Town, they are the second ethnic group after coloured (42%), at (38%) in the census 2011. They are nonetheless the most disadvantaged ethnic group in terms of socioeconomic outcomes.

Figures 2.1 and 2.2 plot real earning trajectories and completed years of formal education of young adults in the analysis sample by age, sorted by the former apartheid designation of their neighbourhood of residence. Both outcomes differ significantly across areas: at 30, residents of former black and coloured-only areas earn well-below residents of former white-only neighbourhoods. They also complete fewer years of formal education. Figures 2.3 and 2.4 repeat the exercise across black ghettos only. To a lesser extent, the trajectories also differ by neighbourhood. In this paper, I focus on the latter and ask whether observable differences in ghetto characteristics can explain divergent life trajectories for children growing up in these areas (i.e. specific ghetto or neighbourhood effects).

The geographic variation in social mobility and the persistence of pockets of high-poverty within cities could be driven by two very different sources. The first, is that neighbourhoods have causal effects on economic and social trajectories of residents. That is, someone growing up in a completely different area would have a very different life outcome. Location-specific social determinants can include the quality of schools, social networks and peer-effects (Durlauf 2004). Another possibility is that the observed differences in neighbourhoods are due to residential sorting, i.e. differences in the types of people living in each area. In this case, 'bad' neighbourhoods may just reflect the spurious correlation induced by the likelihood that the same factors that lead to a given location choice also lead to certain socioeconomic outcomes. Disentangling these two dynamics has been one of the main difficulties of this literature.

To address this fundamental problem, the analysis uses the quasi-exogenous allocation of households to ethnic specific residential areas during apartheid in South Africa. Following the Group Areas Act (1950), ethnic groups[3] were allocated to specific urban spaces separated by buffer-zones to create physical barriers between them. City centres and inner suburbs were zoned for the dominant white group. Apartheid considerably limited the mobility of non-whites through aggressively enforced legislation and spatial planning. Here I focus on compliers in black ghettos only; that is black children in families living in black-only residential areas (townships) by 1991 (when residential segregation was formally lifted)[4]. I provide strong evidence that for this group of young adults, residential decisions were exogenous and sorting is very limited. The dataset is compiled from different sources, including various census

[3] The main ethnic groups were the white, black, Indian and coloured populations. In the Group Areas Act, the latter corresponded to those 'neither white nor black'. In practice, coloured are mixed-race individuals with European, black and Asian ancestry. Indians comprised most Asian populations.
[4] Apartheid officially ends with the advent of democracy in 1994.

datasets and historical archives. The main observations come from a panel of young adults aged 14 to 22 at baseline and residing in the city of Cape Town. It covers 5 periods of the child's life between 2002 to 2009. I mostly look at labour, education and behavioural outcomes computed across the span of the panel. This is intended to give an overview of life trajectories by the time children are between 21 and 30 years old. Children, parental and household characteristics are generally balanced across ghettos at baseline. The metropolitan area of Cape Town is the second largest city in South Africa. It was one of the least segregated cities in the early 20th century but became the most segregated by 1991, with only 5.7% of residents living outside their designated area (8% national average) (Christopher 2001). It thus offers a particularly interesting setting to measure neighbourhood effects in the absence of residential sorting for the subset of compliers.

I find evidence of heterogeneous neighbourhood effects across ghettos. That is, not all 'bad' neighbourhoods have the same impact on outcomes in adulthood. Most differences concern labour and educational outcomes. Residents in ghettos that are closer to the Central Business District (CBD) and relatedly also closer to former white-only neighbourhoods, benefit from a better access to jobs and public amenities (including better schools). Children that grew up here achieve on average 30% higher real earnings in adulthood, are 15 percentage points (pp) more likely to work and 15 pp less likely to be idle. They also do better in school, completing on average 0.8 more years of education and being 11 pp more likely to attend college. These children are also more likely to work in semi-skill jobs compared to other ghettos, where low-skill work is dominant.

These average differences mask some gender disparities. Differences across ghettos tend to be more prevalent for young women. One possible explanation is the strong gender-segmentation of labour markets (Magruder 2010). Under such conditions, discrimination and red-lining may be stronger and similar for males across ghettos. Negative peer effects and gang affiliations may also deter young boys earlier from pursuing education (Freeman 1999), later penalized in job markets. While evidence is limited, this hypothesis underlines the importance of better understanding behavioural and social norms and institutions in explaining neighbourhood effects.

When I examine the main neighbourhood characteristics associated with *better* outcomes in adulthood, I confirm that *proximity* is positively correlated with both better labour and educational outcomes. Access to better schools plays a significant part. Children attending former black-only schools are associated with on average 16% lower

earnings in adulthood. Longer distance to former white or coloured-only schools is correlated with higher delays in graduating their grade. Further, ethnic networks also seem to be important. A one percent increase in the black population of the neighbourhood is associated with a lower probability of being unemployed (6 pp) and fewer months spent searching for jobs (approx. 3.5 fewer on average). Results suggest that the positive effect of location is conditional on the human capital composition of the neighbourhood.

Several theories attempt to explain how neighbourhoods influence the life trajectories of children. Recent work by Chetty et al. (2014) and Chetty & Hendren (2016) have put forward the causal effect of place for intergenerational upward mobility and the life opportunities of children, with evidence across U.S cities. They have also underlined the fact that the time of exposure to a neighbourhood matters. While other empirical work is more mixed (Katz et al. 2001; Oreopoulos 2003; Ludwig e al. 2013; Gibbons et al. 2013), the question is important because of its relationship with the persistence of inequality in certain areas within cities (Piketty 2000; Durlauf 2004). Poor neighbourhoods can become spatial poverty traps (Kilroy 2007; Marx et al. 2013). This issue is even more relevant in developing countries where cities display high levels of income segregation and housing informality. Further, understanding the reach of local determinants of social upward mobility has important implications for policy (i.e. investment in local schools, place-based policies, mobility policies).

A related literature has examined the question of racial segregation and how it affects adults and children outcomes across cities. Racial segregation shapes the socioeconomic make-up of local communities. Prior research has typically proposed two set of explanations that could account for how racial context may affect someone's socioeconomic outcomes (Cutler & Glaeser 1997). The first mechanism relates to the neighbourhood effects literature. The costs of racial segregation being the same as those put forward for growing up in 'bad' neighbourhoods. The second however puts forward the benefits of ethnic density in creating group-specific networks that benefit from an income-mix (Cutler et al. 2008). Again, empirical evidence is mixed (Cutler & Glaeser 1997; Edin et al. 2003; Cutler et al. 2008; Ananat 2011; Bayer et al. 2014; Shertzer & Randall 2016), but some papers find that segregation impacts youths the most. Since Kain (1968) an empirical literature has also explored the question of segregation and access to jobs, putting forward the hypothesis that worse labour market outcomes in deprived neighbourhoods are driven by the lack of accessibility to jobs, related to the spatial mismatch between poor residential areas and job locations (Ihlanfeldt & Sjoquist 1998; Gobillon et al. 2007; Zenou 2009).

All these bodies of research matter for this work. Deprived neighbourhoods analysed here are not only disadvantaged residential urban areas with higher poverty incidence, a lower quality of public amenities, schools, and higher crime and unemployment. They are also ethnic enclaves – paradoxically not of minorities, but still of the most socioeconomically disadvantaged ethnic group in the country. The influence of apartheid spatial planning on Cape Town's urban form and the low provision of public transport reinforces the separation of jobs from the residential areas of the poorest dwellers. The hypotheses put forward in this paper relate to the spatial mismatch literature but go further in suggesting proximity also matters in terms of access to public amenities (particularly, higher quality schools). Poor households in developing country cities are more vulnerable to distances. The under-provision of infrastructure and public transport, the sprawling nature of the largest agglomerations, housing and labour informality, increase frictions related to space. Social networks, as imperfectly measured here, also seem to be an important factor. These two factors are likely to be related. In cities where distance matters, social networks play an important role, not only for jobs but also for enforcing norms and security.

Evidence of neighbourhood effects for developing countries remains largely qualitative and descriptive. This paper is to the best of my knowledge, the first to provide empirical evidence on the question of neighbourhood effects and the persistence of spatial inequality in this context. Identification benefits from the quasi-exogenous allocation of households across black townships during Apartheid. The main caveat remains the small number of ghettos, which limits the reach of the conclusions. While I cannot be definitive about what makes a ghetto better in the setting here, results provide compelling evidence on the importance of childhood neighbourhoods on outcomes later in life in a large city of a developing country. They contribute to the growing literature in development economics looking at the issues affecting poor urban dwellers. Understanding the local determinants of social mobility in emerging cities will be key to inform policy in the years ahead.

The remainder of this paper is organized as follows. Section 2.2 presents the South African context and discusses the policy used for identification. Sections 2.3 and 2.4 present the data and the empirical strategy, respectively. The main results are discussed in Section 2.5, while Section 2.6 looks at the channels. Section 2.7 concludes.

2.2 South Africa's Apartheid Heritage

The context for this paper is South Africa (RSA), and specifically the Cape Town metropolitan area. Both the country and the city have been severely marked by the policy of racial segregation put in place during most of the 20th century. RSA remains one of the most unequal countries in the world[5]. Cape Town had a Gini coefficient of 0.67 in 2011, higher than that of the most unequal U.S. cities. Income inequality has a racial dimension: the annual average income of black South Africans was half the national average in 2011, while it was more than three times higher for whites (census 2011)[6]. The persistence of inequalities across social and ethnic groups is also mirrored in space (Turok et al. 2017). The make-up of cities often reflects the heritage of apartheid's spatial planning. Urban areas have remained profoundly divided along income and racial lines (Pieterse 2009). This paper directly contributes to the understanding of the spatial persistence of inequality and the long-term effect of place on opportunities later in life. With this aim in mind, I discuss relevant institutional aspects next.

The formal enactment of apartheid took place in 1948[7]. All South Africans were officially classified according to skin colour, history, and language by the 1950 Population Registration Act. In the same year, the Group Areas Act (GAA) established spatial segregation by race. The apartheid system was put in place with the objective of reducing to a minimum the interactions between whites and other ethnic groups, particularly blacks (Turok 2012). There were three spatial levels of segregation during apartheid. First, the Grand Apartheid recognised ten homelands or Bantustan, where all Africans were supposed to live. These were considered independent[8], but were not recognized outside of South Africa. The second was urban apartheid: race-based residential segregation that allocated population groups into specific urban spaces separated by buffer-zones to create physical barriers between the different racial groups (Figures 2.5 & 2.6; Figure B1 & B2). City centres and inner suburbs were zoned for the dominant white group. The different black townships were created as need arose for

[5] See discussion in Piketty 2015 Nelson Mandela Annual Lecture.

[6] This is despite the large-scale expansion of social policies, including welfare grants, access to education, health facilities and basic services (Armstrong and Burger 2009).

[7] While the elections in 1948 marked the official start of apartheid, segregation had started in the late 19th century. The Natives Land Act of 1913 and the Natives Urban Act of 1923 were the first pieces of the legislation; the first one restricted the ownership of land to whites while the second ordered the removal of African from city centres to 'outside locations' (i.e. townships). A system of permits was then put in place.

[8] The Bantu Authorities Act (1951) created separate government structures for black and white citizens. Bantustan were consolidated native reserves in rural areas. Several acts later devolved administrative powers to the areas and removed the South African citizenship to *residents*. These areas suffered from severe underinvestment and widespread poverty that fuelled rural migration (Turok 2012).

cheap labour. Langa was the first black township in Cape Town formally created in 1921[9], followed by Nyanga (1945), Gugulethu (1958), Khayelitsha (1983) and lastly, Mfuleni in 1990. Last, was the separation to access public and private amenities (i.e. buses, shops, entrances, etc.)[10]. The latter included schools. School systems were completely separated in infrastructure and curricula by ethnicity, with each ethnic group represented by its own education department. White schools, and to a minor extent also coloured schools, had significantly more resources than black schools (Case & Deaton 1999; Kallaway 2002).

Identification of neighbourhood effects in this paper relies on the compliance of ethnic groups to reside within their designated areas. Compliance needs to be virtually perfect and residential choices must be severely constrained to be able to rule out residential sorting and with it, the spurious correlation induced by the likelihood that location choices also lead to certain socioeconomic outcomes (see discussion section 2.4.2). Several aspects of the legislations and policies during apartheid suggests this is virtually true. Lack of mobility was consciously built into the system. Until the 1980s, legislation enabled the resettlement, expropriation and forced removal of individuals to their designated urban area and rural homelands. Only blacks with formal jobs could officially reside within cities and towns, mostly as temporary migrants, and in clearly designated townships. Slum clearance, the prevention of land invasions, and widespread evictions were used to brutally resettle populations to their ethnically-designated townships (Turok 2012). Figure 2.7 supports an almost total compliance with this policy for black South Africans in Cape Town. Using 1985 census microdata, I plot the percentage of a neighbourhood's ethnic group (black in the upper-left quadrant) against the distance in km to former black townships. There seems to be virtually full compliance with the GAA. Coloured and Indian communities were less aggressively persecuted and residential compliance is much lower (Christopher 1997). Traditionally, these areas were more mixed with a minority of whites and blacks residing here. To limit selection-bias, the analysis is done using Cape Town's former black townships only. The late 1980s saw a progressive abandonment of these practices and more flexibility was introduced. In 1991, the GAA was repealed and the end of apartheid was officially achieved with the advent of democracy in 1994.

By the end of apartheid, the result was almost total segregation. Cape Town, one of the least segregated cities before 1923, became the most segregated one by 1994, with only 5.7% of residents living outside their designated area (8% national average). The

[9] Some back track its origin to 1903 when Africans were forced to live there during the bubonic plague.
[10] Regulated by the 1953 Reservation of Separate Amenities Act.

great majority of those contravening zoning legislations were servants living on the properties of their employees or Africans living in barracks (Christopher 1997, 2001). After more than 20 years from the end of apartheid, urban areas remain deeply divided by the inherited patterns of racial segregation. Population densities within cities are extremely uneven. In Cape Town, the average density in 2009 was 39 persons per hectare, varying between 100-150 in former townships and 4-12 in former white suburbs. The total population of the former white southern and northern suburbs combined was less than either of the two largest former coloured and black-only townships (Sinclair-Smith & Turok 2012). Further, racial *segregation* remains high: dissimilarity indices between black and white, and coloured and white were all above 0.7 in 2001 for both Cape Town and South Africa (Christopher 2001). Fragmented urban forms impose unequal access to jobs, amenities and public services. Former townships suffer from higher levels of unemployment, crime and worse socioeconomic conditions (Turok et al. 2017).

This paper focuses on the life trajectories of young adults. 37% of the working-age population in South Africa is between 15-34 years old (2011). This age cohort is particularly affected by unemployment. On average, youth unemployment was 36% in 2011 – ranging from 11.6% for white South Africans to close to 40% for blacks and 33% for coloured. Cape Town displays the same levels and ranges. A series of reasons have been put forward to explain these differences, including inequalities in schooling systems, resulting skills mismatches (Turok et al. 2017) and spatial mismatch (Banerjee et al. 2008). Apart from these, community, household, and personal factors also drive youth unemployment (Graham & Mlatsheni 2016). All these factors are related to the effect of place on social upward mobility.

2.3 Data

I draw information from several datasets. This section describes the different sources and key variables definitions. It also provides some descriptive statistics and test for balance across the different neighbourhoods of analysis.

2.3.1 Cape Area Panel Study (CAPS) & sample definition

The Cape Area Panel Study (CAPS) contains information on 4,758 randomly selected young adults aged 14-22 at baseline (2002) and living in the Cape Town

metropolitan area. At baseline 46% identify as black or African[11]. These young adults were interviewed in 2002 and re-interviewed four times after that (2003-2004, 2005, 2006 and 2009)[12]. The panel is a stratified random sample that was collected by the University of Cape Town, jointly with the University of Michigan for the first three waves[13].

I focus the analysis on *compliers*. That is, those young adults that in 2002 resided in previously designated black apartheid locations and identify as black or African. Further, I only keep in the sample those that were living in their neighbourhood of residence by 1991, the year at which residential restrictions was formally lifted (845 young adults at baseline). While apartheid ended in 1994, 1991 marks the repeal of the GAA. It is unlikely that major shifts occurred within cities between 1991-1994, but using 1991 as the cutoff removes selection-bias related to the fact that after that year blacks were *free* to choose where to reside. Figure 2.8 repeats the exercise of Figure 2.7 using the final sample of analysis and provides visual evidence of almost perfect residential compliance with the GAA. Using the same dataset, I regress living in former black and coloured-only apartheid areas on the probability of being black or coloured using the baseline year in the sample (Table 2.1). The positive and highly significant relationship is much larger for blacks; it is also larger for those living in the area by 1991, when residential choice was officially constrained. By restricting the analysis to this group, I can assume the allocation to ghettos was *almost* as good as random, allowing me to identify the effect of growing up in deprived neighbourhoods on outcomes in adulthood. The trade-off is that the final sample is much reduced to between 782 and 2,039 young adults depending on outcomes and data availability.

I look at two types of outcomes. The first ones (Table 2.2) are computed across the span of the panel, and intend to give an overview of the trajectory of young adults through labour, education, and behavioural outcomes. Monthly earnings equate to the

[11] I use these two interchangeably in the paper as per the official statistical definition.

[12] For this successive second wave, a subset was interviewed in 2003 (1,360) with the remainder re-interviewed in 2004.

[13] The Cape Area Panel Study Waves 1-2-3 were collected between 2002 and 2005 by the University of Cape Town and the University of Michigan, with funding provided by the US National Institute for Child Health and Human Development and the Andrew W. Mellon Foundation. Wave 4 was collected in 2006 by the University of Cape Town, University of Michigan and Princeton University. Major funding for Wave 4 was provided by the National Institute on Aging through a grant to Princeton University, in addition to funding provided by NICHD through the University of Michigan. Wave 5 was collected in 2009 by the University of Cape Town. Major funding for Wave 5 was provided by the Health Economics & HIV/AIDS Research Division (HEARD) at the University of KwaZulu-Natal, with additional funding from the Andrew W. Mellon Foundation, the European Union and the NICHD.

last monthly real earning[14] (2009) when youths are between 21 and 30 years old. Monthly earnings in the past 5 years are a measure of the average last observed real earning in 2006 and 2009. Both, months worked and months looking for work in 2009 measure the number of months for the activity during the entire lifespan of the panel up until 2009; while similar measures 'since last in school' are calculated from the last period at which the young adult was enrolled in any educational institution. The three measures of education include total years of completed education, a dummy for ever having attended college, and the total number of years delayed from graduating from high school. To measure the latter, I use a 2-year window from the expected age of graduation for the last observed grade (i.e., the last grade for which the young adult was enrolled). Finally, behavioural outcomes relate to smoking, consuming alcohol, or drugs in 2009, as well as ever having reported a live birth for females.

The second set of outcomes are the pooled cross-section of outcomes (Table 2.3). I follow Chetty et al. (2015) here. To avoid measuring earnings when children are still in school or tertiary education I only include observations in which a child is 20 or older, and not enrolled in school. This measure is thus relatively different from the above, for which I impose no schooling restriction. Here I look at labour outcomes: monthly earnings (constant), and dummy variables for doing any paid work (formal or informal), being unemployed (for more and less than 2 months), and being idle (i.e. neither working or in school).

Overall, summary statistics in Tables 2.2 and 2.3 show that young adults in the sample are quite economically disadvantaged. The average real monthly earnings oscillate always close to 2,200 Rands, which is well below the average real monthly earnings for the period 1995-2005 (R2,870). The strong link with employment is reflected in both the low number of average months worked for the entire period (15) and since leaving school (2.2); and the large number of average months looking for work (10-12 months). Only 30% are working between 2004 and 2009, and more than 40% are idle. Education levels are relatively high, with most children completing 11 years of formal education. Fewer are those that pursue a tertiary education: only 3% on average.

The CAPS dataset can be matched with the national registry of schools (SRN) in 2000. Doing this gives information on the school each matched individual in the sample attended at baseline. Because both datasets are geo-located I can also calculate the

[14]I use the CPI index from World Development Indicators (World Bank) to compute real earnings; base year is 2010.

average characteristics of the closest schools for each neighbourhood. During the first wave, a standardized test of literacy and numeracy was carried out for all children in the sample. The test was administered in English and Afrikaans and therefore indirectly penalises Xhosa speakers. Nonetheless, it provides a proxy for controlling for individual 'aptitude' in the absence of individual fixed effects conditional on the language spoken. I use the numeracy and overall tests scores as controls in the final specifications.

A final issue for identification concerns attrition. Attrition is a problem in the panel due to the migratory behaviour of these young adults. There is a risk that I am observing the outcomes of those that do not move and thus introduce selection bias. While I use sampling weights that adjust for young adults' non-response across waves[15], I cannot completely exclude this possibility. The most common cause for attrition among black African for the entire period was moving within South Africa (about 30 to 40% of the attritions by wave). Attrition is much higher for whites (44% lost by 2005) than for African and Coloured (20-23% in wave 4, and 25-30% in wave5). Movers that are successfully re-interviewed remain within the same neighbourhoods. The sign of the bias can go both ways: results could be upward biased if those remaining are exerting less effort to find jobs or complete their education; they could be downward biased if those leaving, on the contrary, use migration as an exit strategy from failing to find jobs and completing school in Cape Town. Both mechanisms are likely here. The city attracts significant migration from other Southern African countries. Competition for low and semi-skill jobs could be higher, and young adults failing to find jobs often choose to migrate. At the same time, the city is a regional hub with most economic opportunities in the Eastern Cape concentrated here.

2.3.2 Census data

This paper also uses census data to obtain general characteristics of neighbourhoods for the available years preceding baseline, notably 1985, 1996 and 2001 census datasets. These censuses are identified at the Enumeration Area (EA) level[16]. I aggregate this data to the neighbourhood level (2015 definitions by the City of Cape Town) and GAA zoned-areas for each ethnic group, see discussion below (Table 2.4). Caution is warranted with respect to the 1985 census dataset, since it was carried out before the end of apartheid and measurement error is high for blacks (Christopher 2001). These

[15] The results are generally unchanged when excluding sampling weights.
[16] I obtained EA identifiers for the census 1985 from Daniel de Kadt. He uses these in his paper with Melissa Sands, How Segregation drives voting behaviour: New theory and evidence (2015). I only use EA falling completely within neighbourhood boundaries and GAA areas.

three census years provide useful information on ghetto characteristics during the first 10 to 15 years of life of the young adults in the sample.

2.3.3 Apartheid locations

Former apartheid locations in Cape Town as designed by the GAA (Figure 2.5) were digitized from planning maps obtained at the library of the Municipality of Cape Town (Figure 2.6). These historical maps provided fine detailed demarcation of black, coloured, Indian/Asian and white segregated areas by streets. I attribute to each of Cape Town's neighbourhoods (smaller areas in Figure 2.6) their former GAA designation if they fall 90% within the area. Overall, 49 neighbourhoods fall within former black-only areas. I exclude townships for which I unfortunately do not have enough observations of young adults (N<30)[17]. There are four former black townships in the final sample (Langa, Nyanga, Gugulethu and Khayelitsha) composed of 19 neighbourhoods. It follows that the neighbourhoods of residence of the young adults in the CAPS sample are *split* into these former apartheid locations. These disadvantaged neighbourhoods are the four ghettos[18] or 'treatment arms'.

Table 2.4 provides summary statistics for each of them in 1985, 1996 and 2001. On average, all the areas are predominantly populated by black Africans (>98%) across the period. There is some heterogeneity in terms of the proportion of working age individuals and the proportion of adults born in Bantustan Republics in 1985. These reflect the larger share of single male migrant workers in locations that were specifically conceived as dormitories (i.e. Langa and Nyanga). The specificity dissipates progressively in 1996 and 2001, when areas are balanced in terms of age, gender and ethnic origin. Differences remain nonetheless (see discussion below). Further, they all vary in area and population sizes, partly reflecting the different periods at which they were formally designated.

Table 2.5 complements the summary statistics from the census datasets with descriptive characteristics obtained from CAPS and alternative sources. There are large differences in terms of distance to the CBD[19] and access to formal jobs[20]. Table 2.6 tests

[17] This only applies to Mfuleni. This area was also the last black-only township to be created in 1990.
[18] As previously defined, I understand ghettos here as disadvantaged residential urban areas that are enclaves of high poverty and have a high concentration of ethnic minority (Zenou 2009b). In the case of South Africa, the paradox is that Africans are the most the South African population (80.8% in 2016, according to mid-year population estimates by Statistics South Africa (SSA)). In Cape Town, they are the second ethnic group after Coloured (42%), at (38%) in the census 2011. They are nonetheless the most disadvantaged ethnic group in terms of socioeconomic outcomes.
[19] Distance to CBD is measured as the Euclidean distance in Km from the centroid of the young adults' neighbourhood of residence and Cape Town's CBD.

for balance of these characteristics across former GAA locations. All tested variables are statistically significant except for attending a school previously under the Black Department of Education (or DET). Largely, it is to be expected that children in the sample predominantly attend these schools located in former black-only locations (see discussion in section 2.6). Langa is a clear outlier here with a larger proportion of young adults attending former coloured or white only schools.

Overall, the following patterns emerge in terms of township specificities concerning neighbourhood characteristics (Tables 2.4, 2.5 and 2.6). Langa and Gugulethu are both closer to the CBD (10.5 km and 15.2 km vs 20 km on average, respectively) and to former coloured townships. Relatedly, they both display better accessibility to jobs and to former white or coloured only schools. On the opposite side, Khayelitsha - by far the largest black township in the city-, is located 25 km from the CBD. This reflects the fact that it was one of the last townships to be established in what was then vacant land at the periphery of the city. Access to jobs across periods is consequently the lowest. Former coloured or white only schools are the least accessible for Nyanga residents. But differences are small. In terms of the level of education in the census dataset, Gugulethu and Nyanga have the most adults with no formal education (13-14% in 2001), with Khayelitsha and Langa in the opposite of the spectrum (7-8%). Nyanga stands out in terms of insecurity. It displays the highest murder rates in 2007 (13.8%); it also registers the highest number of serious crimes and assaults committed in school premises in 1999. This reflects the fact that it is considered one of the most dangerous places in South Africa (and the world) with high levels of gang criminality. Housing informality is above 60% for all but Gugulethu where only half the residents report living in shacks.

To summarize these in a comparable way, I create an index of 'ghetto quality' (Tables 2.7 & B1 in Appendix B[21]). There is a limited use for this index in the analysis here, but it gives the reader an idea of rank between the different areas considered using census data characteristics. The index is the simple average of the z-scores of the characteristics in the table. They encompass measures for housing amenities, location, education, and poverty levels. Overall, Gugulethu ranks first and Langa second. Both driven by different dynamics: the first one because of better amenities (i.e. access to electricity and water) and housing conditions; the second one because of higher levels

[20] Accessibility to jobs is measured by the number of jobs in all neighbourhoods in Cape Town weighted by the inverse of their distance from the centroid of the neighbourhood of residence in km. Definition 1 uses the percentage of formal firms' turnover as a proxy for jobs. Definition 2 uses the number of employed in formal firms as a proxy for jobs. Data for 2001 and 2005 was obtained from Sinclair-Smith & Turok (2012). Data for 2009 was obtained from the Geospatial Analysis Platform (GAP) of the South African Council for Scientific and Industrial Research (CSIR).

[21] All tables numbered B# are in Appendix B.

of human capital, lower murder rates and accessibility to schools and the CBD. At the bottom are Nyanga and Khayelitsha. The first one ranking lower in all categories but centrality; in return, while still ranking at the bottom, Khayelitsha ranks first for access to housing amenities, and second after Langa for human capital.

The heterogeneity in terms of observable characteristics is key for the identification of neighbourhood effects: not all of the deprived neighbourhoods considered here are equal, largely as a result of historical paths, and I consider each ghetto as a different 'treatment arm'. Whether these differences are enough for children growing up in these areas to have different life trajectories is not straightforward. Certain areas may be better for social upward mobility and opportunities later in life than others (Figures 2.3 & 2.4), but overall children growing up in 'worse' environments may still be all similarly negatively affected through discrimination, red-lining, and lower social conditions (Figures 2.1 & 2.2). Understanding how these differences matter permits to better appreciate the mechanisms behind the spatial persistence of poverty.

Finally, Tables 2.8 and 2.9 report balance tests for selected baseline characteristics of children, parents, and households across areas. The idea here is to test for the absence of residential sorting in the sample, despite GAA locations differing on observables characteristics due to their history. Given that the sample of young adults was randomly selected to be representative at the city level, we can expect balance at baseline across ghettos if the randomization was successful and residents could not choose their location at the time of moving (prior to 1991). This should be the case under the apartheid setting described, and for our sample of compliers[22].

The condition is largely fulfilled. Children's characteristics are balanced at baseline in all but English and Xhosa as the main languages spoken in their households. These differences can reflect the geographical heterogeneity within families due to migration behaviour from Bantustan areas to Cape Town. I control for preferred spoken language in all specifications. The large majority of parental and household characteristics are also balanced at baseline. The exception concerns the higher percentages of uneducated parents in some areas, the number of members in the household at baseline, and growing up in informal areas. Differences are relatively small. Again, I control for these characteristics in all specifications. The general balance in baseline characteristics for children, parents and households in the sample validates the identification strategy. This allows me to test for neighbourhood effects and ghetto heterogeneity on a random

[22] Also, note that the age distribution of young adults in the pooled-cross section (waves 1 to 5) follows a normal distribution that is similar and balanced across ghettos (Figure B3).

sample of individuals that did not choose their place of residence. The main assumption is that allocation is *almost* as good as random across these different ghettos. This heterogeneity allows me to further investigates the channels through which some neighbourhoods may be better than others for social mobility.

2.4 Empirical Strategy

2.4.1 Methodology

This paper provides reduced-form evidence of neighbourhood effects for complier children growing up in former black-only apartheid ghettos. In the core analysis, I estimate the effect using OLS regression specifications of the form:

$$Y_{ihg} = \alpha + \sum_1^G \beta_g . Ghetto_g + X'_i . \gamma_1 + W'_h . \gamma_2 + \mu_{ihg} \quad (2.1)$$

where Ghetto are dummy variables equal to one for residing in one of the three ghettos (Khayelitsha is the excluded one)[23] and β_g are the coefficients of interest; X_i is a vector of individual characteristics and W_h is a vector of household baseline characteristics. All of the regressions are weighted to adjust for differences in sampling probabilities across sites and over time. I cluster the standard errors by household (allowing for common error components across siblings) because residential designation occurred at the family level and the number of neighbourhoods in the sample is too small to cluster at this level (N<30).

I systematically include the following individual and household characteristics: age, age squared, gender, and language spoken in the household. I also include dummy variables for the level of education of parents, the household size at baseline and a dummy variable for having grown up in an informal area. These last characteristics were unbalanced at baseline. I also test the sensitivity of the findings to the inclusion of the normalized result of the numerical test, as a proxy for general individual aptitude. I include the latter for precision as doing so does not alter the results.

Equation (2.1) measures childhood neighbourhood effects on outcomes in adulthood, i.e. outcomes defined for the lifespan of the panel (Table 2.2). I use an alternative specification that includes year fixed-effects to estimate the same effect for outcomes across the pooled cross section of young adults for the period after baseline (2004-2009), (Table 2.3). The intention is to measure outcomes in adulthood controlling

[23] Tables in section B1 of Appendix B display the result of the above specifications regressing outcomes on each ghetto separately (against all others).

for different year-specific shocks in the fashion of Chetty et al. (2015). The standard errors, which are clustered by original household (i.e. family), allow for common error components across siblings and adjust for the repeated observations of each child.

2.4.2 Main assumptions & limitations

The main strategy I employ for estimating how childhood neighbourhoods affect outcomes in adulthood relates to the assumption that children could not choose their neighbourhood of residence due to limited residential mobility during apartheid (Case & Deaton 2002). Because of this, I focus my analysis on compliers. That is, I only compare black children that grew up in different 'bad' or deprived neighbourhoods, all of which were designated as black-only areas during apartheid, and who were living in these areas by 1991. As discussed, apartheid severely constrained the residential choice of black South Africans, much more so than for any other ethnic group.

Another assumption behind identification here is that families could not choose between ghettos. This is not entirely true. On the one hand, it depends on the time their first relative started living in Cape Town, as black ghettos were progressively created as the need for low-skill workers arose. The larger concentration of working age males in Langa in 1985 is an example of this. Still, removals and the forced relocations of trespassers and informal settlements randomly assigned people to ghettos as capacity allowed[24]. This lessens residential sorting between ghettos. Supportive of this, I find that parental and household characteristics at baseline are mostly balanced (Table 2.9). Possible sorting is further mitigated by the focus on young children who did not choose their neighbourhood of residence (Cutler & Glaeser 1997; Dujardin et al. 2008). This setting allows me to assume that designation to one of the 'treatment arms' was *almost* as good as random.

Comparing children growing up in different black ghettos has the advantaged of answering two questions. Are there neighbourhood effects of growing up in a deprived neighbourhood? Second, are these effects different across ghettos (i.e. is there one 'ghetto university')? What makes a deprived neighbourhood better? Comparing black children growing up outside of the black-only designated area would have also allowed me to understand if life trajectories diverge from 'good' vs. 'bad' neighbourhoods. Unfortunately, selection-bias prevents me from looking at this question here. Parents sorting outside of their designated racial area could have done so to provide their children with better life opportunities (Noah 2016). These parents may also support

[24] South African History Online (SAHO). Online public library documenting the history of Apartheid.

their children in other unobservable ways. Further, discrimination in white or coloured areas could also bias the results.

A final assumption relates to the relevance and permeability of neighbourhoods. As underlined by Oreopoulos (2003), for interactions to matter at the neighbourhood level (or ghetto here), social contacts must depend significantly on where an individual resides, and neighbour relationships must be important enough to influence individual decisions. Further, many of these areas are relatively close to each other, sometimes bordering, and it is hard to imagine that they are fully self-contained administrative units. Any significant social interaction would violate the stable unit treatment value assumption (SUTVA). The design of apartheid mitigates the risk of large spillovers from other areas. Apartheid defined-ghettos were often separated by physical barriers and buffer zones, not only to avoid contact between ethnic groups but between them to limit the risk of uprisings (Figures B1 & B2). Further, apartheid reinforced township identities (Noah 2016). These large neighbourhoods are strongly associated with identities through religious, language and traditional practices, as is the case in Cape Town (Bekker & Leildé 2006).

This paper provides reduced-form evidence on neighbourhood effects and ghetto heterogeneity on complier children using the quasi-natural experiment of apartheid. The evidence here is conditional on the assumptions above holding. As discussed, several factors mitigate selection-bias and possible SUTVA violation. Only under these maintained assumptions, the main results can be considered as causal estimates.

2.5 Estimation Results

2.5.1 Education

Table 2.10 presents estimates of heterogeneous ghetto effects on children's overall education in adulthood, that is by the last period in the panel (2009) when children are between 21 and 30. They include all controls except the standardized results of the numerical aptitude test to avoid introducing endogeneity. The excluded ghetto here and in subsequent regressions is Khayelitsha, the one with the higher number of observations.

I begin in column (1) which shows the estimated effects on the total years of formal completed education. These include all possible levels from primary to tertiary and technical. The average years of formal education reached in adulthood is almost one year longer for children growing up in Langa (at 5% significance level) with respect to

those in Khayelitsha. Table B2 in Appendix B supports these findings against all other ghettos (panel 1). In contrast, the probability of attending college (column 2) is statistically significantly higher for children growing up in all ghettos other than Khayelitsha, by between 5 to 11 percentage points (pp). They are the largest in Langa. Separate regressions in Table B2 confirm that overall, children growing up in Khayelitsha are 6 pp less likely to attend college.

Table 2.11 presents estimates on the same outcomes when splitting the sample by sex. There is no strong evidence of any specific conditional effect here. The exception concerns girls growing up in Gugulethu and Langa, which are between 6 to 21 pp more likely to attend college than girls growing up in Khayelitsha, respectively. These are significant at 10% level.

Neighbourhoods do seem to matter for overall educational achievement in adulthood. Overall, children growing up in Langa display better achievements in education. Girls here are also more likely to attend college, substantially more so than in any other township (21 pp). In contrast, Khayelitsha is distinctively a worse place to grow up for the levels of education achieved in adulthood. These results are consistent with Langa boasting the highest human capital levels (i.e. the highest percentage of adults with at least secondary diplomas); but also, its proximity to better schools, at least when proxying quality by schools previously reserved to white or coloured children (see discussion in section 2.6). The negative results for Khayelitsha are more puzzling. It is second to Langa in terms of human capital levels. Yet, despite this, it displays the largest share of employed adults in construction, a sector with a smaller premium on education. It is also the ghetto where distances to schools are the largest. These factors are possible explanations.

The small gender heterogeneity in college attendance may reflect the general trend of young women achieving higher levels of education than their male counterparts, which manifests in neighbourhoods with better access to higher quality schools. However, it may also suggest that differences across ghettos might be less marked for young men. There are many possible explanations for this. Institutional perceptions and norms may be playing a part. Boys could be more affected than girls by negative peer effects (for instance they could be more pressured to join gangs, to skip school, etc.). While, both male and females are generally similarly distributed across sectors of occupation in the sample (Figure B4), there are marked gender-differences in terms of the skill levels of occupations in Langa (Figure B5). Young women here are over-represented in higher-skill occupations with respect to males, despite having the lower

proportion of women in this sector across ghettos. As education is determined earlier in lifer, the differences found here may be reflected in labour outcomes. Evidence while still incomplete, suggests female and males are affected differently by neighbourhood effects.

2.5.2 Labour outcomes in adulthood

Tables 2.12 and 2.13 present estimates of heterogeneous ghetto effects on labour outcomes in adulthood. Table 2.12 displays estimates of ghetto effects on real earnings and employment outcomes more generally across the period 2004-2009. These regressions are estimated with one observation per year per child. The standard errors, which are clustered by family, adjust for the repeated observations for each child. Table 2.13 only shows measures of real earnings and job search by the last period in the panel (2009) when children are between 21 and 30. They include all controls.

I begin in Table 2.12. Overall, children that grew up in Langa and Nyanga are 15 pp less likely to be idle with respect to Khayelitsha, and by the same magnitude are more likely to be working in young adulthood. Statistical significance is much higher for Nyanga at 1% level. However, it only holds for Langa when excluding all other ghettos (Table B4). There are no significant differences on real earnings or long term unemployment for the pooled cross section. Children growing up in Gugulethu are however 7 pp more likely to be unemployed for at the last 2 months before the survey. This is significant at 10% level.

Results in Table 2.13 complete the picture. Column (1) shows the estimated effects on average real earnings for the last two periods in the panel. Notably, only children growing up in Langa display 30% higher earnings in adulthood with respect to Khayelitsha, significant at 10% levels. Consistent with results in Table 2.12, column (3) shows that children in this former black township worked longer since leaving school; coefficients are not statistically significant but close to 10% levels. The largest difference with respect to the other ghettos concerns job search. That is, children that grew up in Langa had to spend substantially less months searching for jobs overall (columns 4 & 5). This is statistically significant at 5 and 10% levels. Other patterns that emerge are more mixed. Both the two other ghettos, display between 4 to 5 months longer search periods for jobs than Khayelitsha and double the number with respect to Langa. These differences are statistically significant at 1 and 5% levels. Only, children growing up in Nyanga worked significantly more months overall in adulthood (5 months) with respect

to Khayelitsha, significant at 5% level. The number of months is substantially smaller since leaving school (column 3), suggesting it reflects part-time employment.

Tables 2.14 and 2.15 repeat the above exercises but again I separate the sample by sex. In both tables the first panel displays the results for young women. Interesting patterns emerge here. First, the differences related to outcomes in adulthood (Table 2.15) indicate almost no differences across ghettos for males. This is consistent with findings on educational outcomes. The exceptions are the longer search months for boys growing up in Gugulethu of up to 7 months. Focusing on the first panel of the same table, we see that the pooled results on earnings for Langa may be driven by the effect on young women. They have on average 53% higher earnings in adulthood than girls in the omitted ghetto. They also display lower search periods with respect to all other ghettos. This is again in line with findings on educational achievements[25]. This gender heterogeneity seems to hold only for Langa. Outcomes in Table 2.14 show that for Nyanga the effect is even for both boys and girls. These overall patterns are largely unchanged when I run regressions for each ghetto separately in Tables B6 and B7.

There are three key facts that emerge from the analysis of labour outcomes. The first concerns the fact that differences across ghettos are small, with only Langa standing out in terms of better earnings, a higher likelihood of working, and lower search periods for jobs. Gugulethu, which was ranked first in terms neighbourhood quality, does not seem to display any specific differences for labour outcomes, and if anything, young adults suffer relatively more from shot-term unemployment. This finding supports differential effects being driven by exposure to higher levels of human capital (Cutler & Glaeser 1997), higher educational achievements and possible networks associated with it (Magruder 2010). Second, the almost inexistent differences across ghettos for young men together with the distinct differences for young women in Langa, are suggestive of gender-specific mechanisms in labour and educational achievements. A possible explanation concern higher levels of discrimination for males in highly gender-segregated jobs (Magruder 2010). Alternatively, social norms might be more penalizing for males through negative peer-effects already in school. Because of worse educational outcomes they may be later penalized in job markets. Finally, Nyanga shows interesting dynamics given how bad the ghetto ranked in terms of basic observable characteristics. The fact that young adults that grew up here show longer working periods overall (but not since finishing school), and a higher likelihood of

[25] Tables B13 and B16 contain the results of equation (2.1) on outcomes in tables 2.12 to 2.15, conditional on years of education. Controlling for education introduces endogeneity concerns and these tables are indicative. While coefficients vary slightly, results are unchanged.

working in adulthood is counterintuitive. Possible explanations could relate to a higher prevalence of informality or youths choosing gang-related activities as an alternative to formal employment (Freeman 1999).

To some extent, results in Tables 2.16 support these findings. Here, outcomes variables are dummy variables equal to one if young adults are employed in high, semi or low-skill occupations[26]. Because of the small sample size (i.e., only employed adults each year), I only look at the pooled cross-section across 2004-2009. Youths in Langa seem to have a 17 pp higher likelihood of working in a semi-skill occupation, compared to Khayelitsha, and almost 10 pp higher with respect to youths in Nyanga. Children that grew up in the latter and Gugulethu seem more likely to be in employed in low-skill occupations by between 7 to 8 pp. When running separate regressions however (Table B8), both youths in Langa and Nyanga seem overall more likely to work in semi-skill occupations; but those in Langa twice as much.

2.5.3 Behavioural Outcomes

I next examine the effects on basic behavioural and health-related outcomes in Table 2.17. Specifications are as before. Columns (1) to (3) measure the probability of engaging in the behaviour at endline, that is consuming alcohol, drugs or smoking when reaching adulthood. Only children growing up in Gugulethu are more likely to both drink alcohol and smoke in adulthood when compared to children growing up in Khayelitsha. They are 15 pp more likely to engage in these behaviours in adulthood, with coefficients statistically significant at 1 and 5% levels. There are no differences in the probability of consuming drugs (defined as cannabis and other illegal drugs). In column (4) I study the fertility behaviour of females but find no differences between women growing up in different areas. There is little that can be concluded from Gugulethu's specificity in terms of alcohol and drugs consumption. The similar fertility behaviour for young females supports the idea of the persistence of certain institutional norms and behaviours that are likely to be gender-biased.

[26] The definition is from Statistics South Africa. They define in South Africa as skilled (or high-skilled) occupations in standard occupations codes (SOC) 1-3, i.e. managerial, professionals and technicians; as semi-skilled (or medium-skilled) as those working in SOC 4-7, i.e. clerks, sales and services, skilled agriculture, crafts, and machine operators; low-skilled are elementary occupations and domestic workers (SOC 8). Here I include in column (4) a broader definition of unskilled because of the lack of precision in CAPS. It includes 'armed forces and others'; while in column (3) it is only composed of those employed in elementary occupations.

2.6 When Are Ghettos Better?

This section investigates the main channels that could explain the differences in outcomes. If not all deprived neighbourhoods have the same effect on lifetime opportunities, then what makes a 'bad' neighbourhood better? Understanding why children may have better labour outcomes in adulthood and achieve better education in certain ghettos opens the door to having a better grasp of the mechanisms behind the spatial persistence of poverty.

2.6.1 Setting

I look at this question by creating variables at the neighbourhood-level within ghettos. Doing this allows me to increase the 'treatment' variation[27] without compromising too much on the quasi-random allocation of households across areas. Sorting across neighbourhoods within ghettos is highly likely. While households were allocated to particular townships, they were free to move within these, and the SUTVA is probably violated. However, neighbourhoods within ghettos are relatively homogenous. There is no evidence to support they have different identities, or should be considered as anything more than just smaller administrative sub-divisions[28]. Despite these mitigating factors, results here are descriptive findings; they provide a useful start into the discussion of poverty traps and the spatial transmission of poverty.

To this aim, I generally estimate the following equation using an OLS specification of the form:

$$Y_{ihn} = \delta + Z'_n.\rho_n + X'_i.\pi_1 + W'_h.\pi_2 + \tau_{ihn} \quad (2.2)$$

Where Z_n are neighbourhood-level characteristics described below, and ρ_n are the coefficients of interest. All other parameters are specified as in equation (2.1). The regressions are weighted to adjust for differences in sampling probabilities across sites and over time, and I also cluster the standard errors by household (allowing for common error components across siblings). I focus on three sets of channels.

First, social channels. For this I use the percentages of ethnic composition of the neighbourhood in 1996 and 2001, from census data. I exclude 1985 due to the limitations of the census (see data section). I use the percentage of black and coloured

[27] There are 19 neighbourhoods overall in the final sample. Langa and Gugulethu, which are small townships are comprised only of one neighbourhood.

[28] See anecdotal evidence here: http://mapping.wm.edu/2014/01/04/post-apartheid-identity-in-cape-town-townships/ (last accessed August 7th 2017).

in the neighbourhood only, since the percentage of whites is almost zero in former black-only areas. The idea here is to measure the degree of ethnic concentration with a straight-forward statistic. Cutler & Glaeser (1997) stressed the advantage that ethnic concentration may have by fostering social networks that provide access to jobs. They find that the impact is highly correlated with the group's average human capital (Cutler et al. 2008). Social networks could also help 'minorities' through community enforced norms. Here this could happen through security mechanisms to guard informal houses for instance. Using the same dataset Magruder (2010) finds that networks are important job allocation mechanisms in Cape Town. The measure of work here includes any type of paid-work (i.e. including informal), and social networks could be particularly important for low-skill jobs.

Second, I look at measures of accessibility to jobs. As explained before I use three indices (see footnote 21) that compute the number of jobs accessible from one's neighbourhood of residence to all other neighbourhoods in the city, weighted by the inverse of their distance from the centroid of the neighbourhood of residence. I use data on employment for 2001, 2005 and 2009 (i.e., the baseline, midline and endline years of the CAPS panel). I also use a measure of distance to the CBD, calculated as the Euclidean distance in km from the centroid of the neighbourhood of residence. In 2009, 27% of jobs in Cape Town were within 10 km from the CBD, down from 30% of formal jobs in 2001. Centrality is likely to be correlated not only with better access to jobs but also amenities. These different accessibility measures test for the spatial mismatch hypothesis. Banerjee et al. (2008) underlined the importance of this mechanism in explaining youth unemployment in South Africa. The long and costly commuting trips in South African cities illustrate two of their major problems: urban sprawl and a high level of segregation of population groups. Rospabe & Selod (2006) have documented the problem for the city of Cape Town, finding evidence that the disconnect between jobs and residential locations are one of the main problems behind blacks' high unemployment.

Finally, I look at the quality of education. This mechanism has been widely considered in the neighbourhood effects literature (Durlauf 2004). Early models by Benabou (1996) show that segregation tends to amplify future human capital inequality. The fiscal channel is one of the main reasons for this, when funding for education comes from endogenous community decisions about how much fiscal revenues to allocate to schools (given that the quality of education is directly related to funding) (Piketty 2000). It is not the case in post-apartheid South Africa. In this context, the main differences across areas are due to the legacy of the apartheid system of education

that allocated different funding across ethnic groups. Black-only school were controlled centrally by a separate Department of Education and Training (DET) and had a separate budget. There were marked discrepancies in educational funding per pupil across racial groups and places of residence. Funding levels per pupil for whites, coloured, and blacks were, respectively, 1.85, 1.59, and 0.74. The curricula and standards were also different (Case & Deaton 2002). I include different measures that consider the quality and accessibility to schools. The distance in km from the neighbourhood's centroid to the closest school and to the closest former white or coloured-only schools; as better access to the latter may allow children living in former black townships to have access to better schools. I also include dummy variables for the former department of education of the school of each young adult (being a former black or coloured/white school). This measure may introduce bias since parents choosing to put their children in 'better' schools may also be different in other unobserved characteristics (i.e., for instance the amount of support they give to their children). They remain informative to understand how much education is central to explain the findings here. Finally, I also include a dummy variable for whether the school of the child offered transport subsidies. I ignore the pupil to teacher ratio since it has very little variation between areas.

2.6.2 Results

Tables B10-B12 in Appendix display the results from estimating the relationship of each channel separately, on the labour and educational outcomes of young adults. In Tables 2.18-2.20 I include the most relevant channels for each of these three categories together. The sample remains unchanged and is always of black compliers only.

Looking at which channels remain significant when including them all in the regressions is a more relevant exercise. I focus on these results. Table 2.18 examines educational outcomes. Here two main channels are associated with attaining a better education in adulthood: accessibility to jobs and the quality of the schools. The access index is positively associated with both the number of years of education completed and the probability of ever attending college. This relationship could reflect the endogenous mechanism between accessibility to jobs and employment, or the mechanical relationship between accessibility and closeness to the CBD, and with this to many other amenities. A school providing transport subsidies is also positively related to achieving more years of formal education. Finally attending a former black-only school is negatively related with the years of completed education. The coefficient is small but significant at 5% level.

Table 2.19 and 2.20 examine labour outcomes in adulthood. The percentage of blacks in the neighbourhood (1996) is correlated with lower job search and relatedly both a lower likelihood of long and short-term unemployment. These are significant at 5% level. The positive association is consistent with the importance of social networks for findings jobs in Cape Town (Magruder 2010). They support the theoretical models that put forward the positive effects of segregation. The fact that Langa and Nyanga have one of the largest share of black residents across census years, and both had the largest share of Bantustan migrants in 1985 (Table 2.4) could be a factor in explaining the better labour outcomes in adulthood. Accessibility also seems to matter. I use the (log) accessibility in 2009 since it is the most precise of the three accessibility indices. Both the probability of working and being idle are, respectively, positively and negatively related to access. These relationships are significant at 1% level. They give weight to the spatial mismatch hypothesis as a key factor behind higher unemployment in former townships. Langa is the most central of all four ghettos (the others were established much later). Finally, school quality and access measures seem to matter less. Only in Table 2.19, coefficients on earnings are negatively associated with attending a former black-only school and statistically significant. Only 13% of children in the sample attend other schools, but they do disproportionately so in Langa (30%).

Overall, these results are indicative of the importance of location. Ghettos that display closer proximity to jobs and presumably access to better public amenities (by being closer to former coloured and white only areas), are also better for the opportunities and outcomes of children in adulthood. This hypothesis relates to the spatial mismatch literature. It goes further here in that results offer suggestive evidence of the relationship of proximity and public amenities (such as schools) with lifetime trajectories of disadvantage children. Poor households in developing country cities are more vulnerable to distances. The under-provision of infrastructure and public transport, the sprawling nature of the largest agglomerations, housing and labour informality, increase frictions related to space. In contrast, private amenities such as better housing and public services do not seem to improve outcomes in adulthood (i.e. Gugulethu). Social networks, as imperfectly measured here, also seem to matter. These two factors are likely to be related. In cities where distance matters, social network play an important role, not only for jobs but also for enforcing norms and security.

2.7 Conclusions

This paper uses a unique dataset to provide evidence on neighbourhood effects in a large city of a developing country. The quasi-random allocation of families across former black-only townships during apartheid provides a natural experimental setting to study the heterogeneous effects of ghettos on compliers.

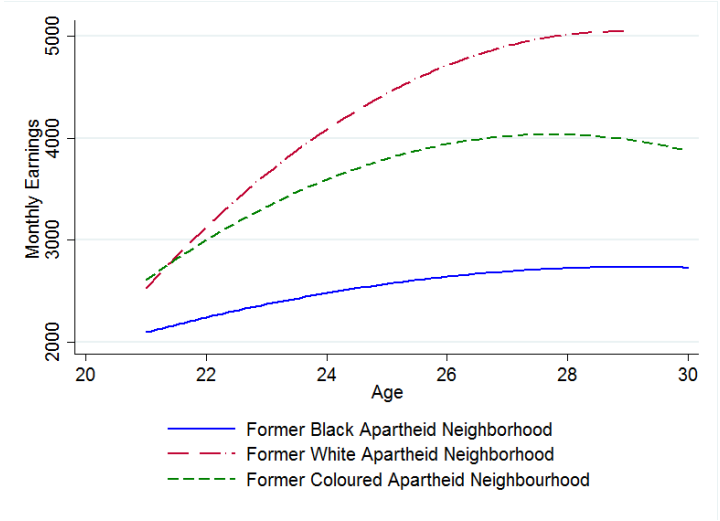
I find that not all 'bad' neighbourhoods are equal in how they influence the life trajectories of children. Differences are salient in both labour and educational outcomes. Ghettos that are closer to the CBD and relatedly, closer to former white-only areas, have better access to jobs and public amenities (including better schools). Children that grow up here achieve on average 30% higher real earnings in adulthood, are 15 pp more likely to work and 15 pp less likely to idle. They also do better in school, completing on average 0.8 more years of education and being 11 pp more likely of attending college. Proximity to jobs and better schools is not all that matters. Results are suggestive on the positive effect being conditional on the human capital composition of the neighbourhood. Further, ethnic networks seem to also be important channels.

I find that differences across ghettos are more marked for young women (in all but fertility). A possible explanation relates to the importance of social norms and institutions. Discrimination, red-lining, and peer-effects may work more negatively for males in gendered-segmented labour markets. Criminal and gang-related activities are male-dominated, and young boys could be deterred from pursuing education (Freeman 1999), later penalized in job markets. While I cannot be definitive with the evidence here, this hypothesis underlines the importance of better understanding behavioural and social norms in explaining neighbourhood effects. They have a strong policy dimension in that, if true, local public policies need to address social constructions to succeed. They also need to take into account specific gender dimensions.

This paper is limited in the extent to which it can provide causal evidence on the channels behind the persistence of spatial poverty traps. Still, it provides strong evidence regarding the heterogeneity of ghetto effects and the importance of neighbourhoods on outcomes later in life in a large city of a developing country. There is not one 'ghetto university'. Further research is needed to understand the factors behind the spatial persistence of poverty within cities, particularly in rapidly growing cities in emerging countries.

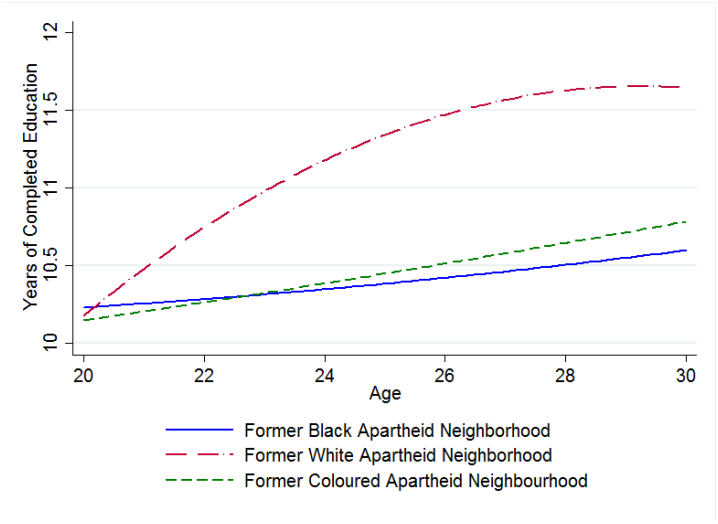
2.8 Tables & Figures

Figure 2.1 Total monthly earnings by age and former apartheid designation of neighbourhood



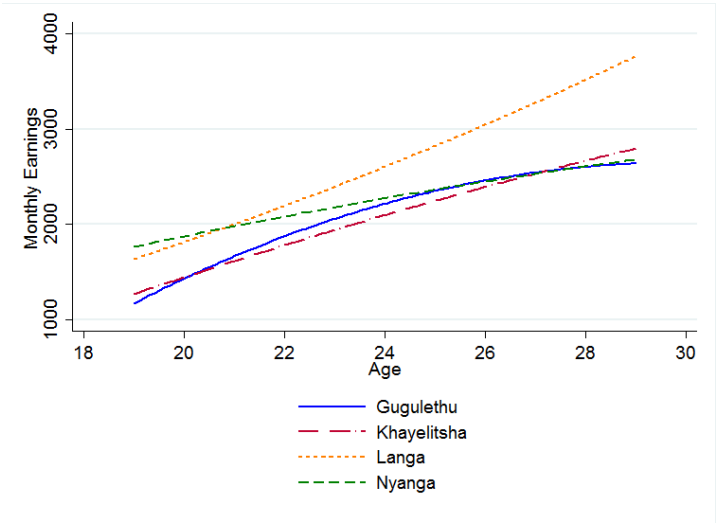
Notes: Plot of monthly earnings for youths in the analysis panel for ages between 22 and 30. The averages are computed according to the former apartheid designation of their neighbourhood of residence. For instance, former black apartheid neighbourhoods stand for neighbourhoods that were black-only before 1994. All earnings are in 2010 South African Rands.

Figure 2.2 Years of completed education by age and former apartheid designation of neighbourhood



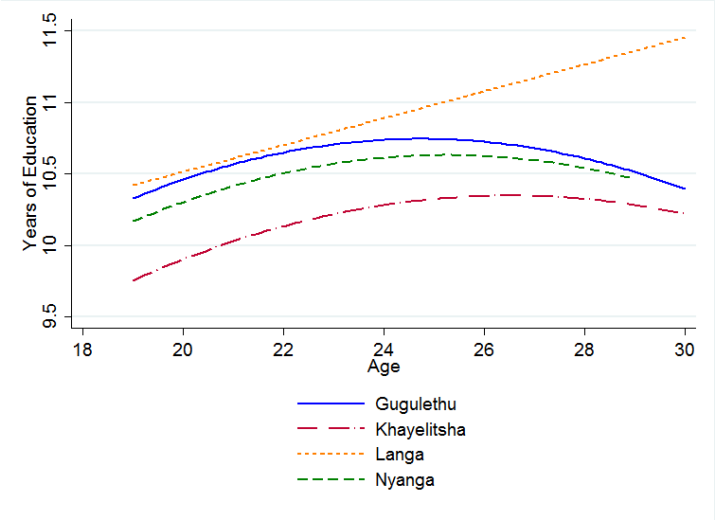
Notes: Plot of total years of education achieved for youths in the panel dataset for ages between 22 and 30. The averages are computed according to the former apartheid designation of their neighbourhood of residence. For instance, former black apartheid neighbourhoods stand for neighbourhoods that were black-only before 1994.

Figure 2.3 Total monthly earnings by age for former black-only neighbourhoods



Notes: Plot of monthly earnings for youths in the analysis panel for ages between 22 and 30. The averages are computed across former black-only ghettos (townships). All earnings are in 2010 South African Rands.

Figure 2.4 Years of completed education by age for former black-only neighbourhoods



Notes: Plot of total years of education achieved for youths in the analysis panel for ages between 22 and 30. The averages are computed across former black-only ghettos (townships).

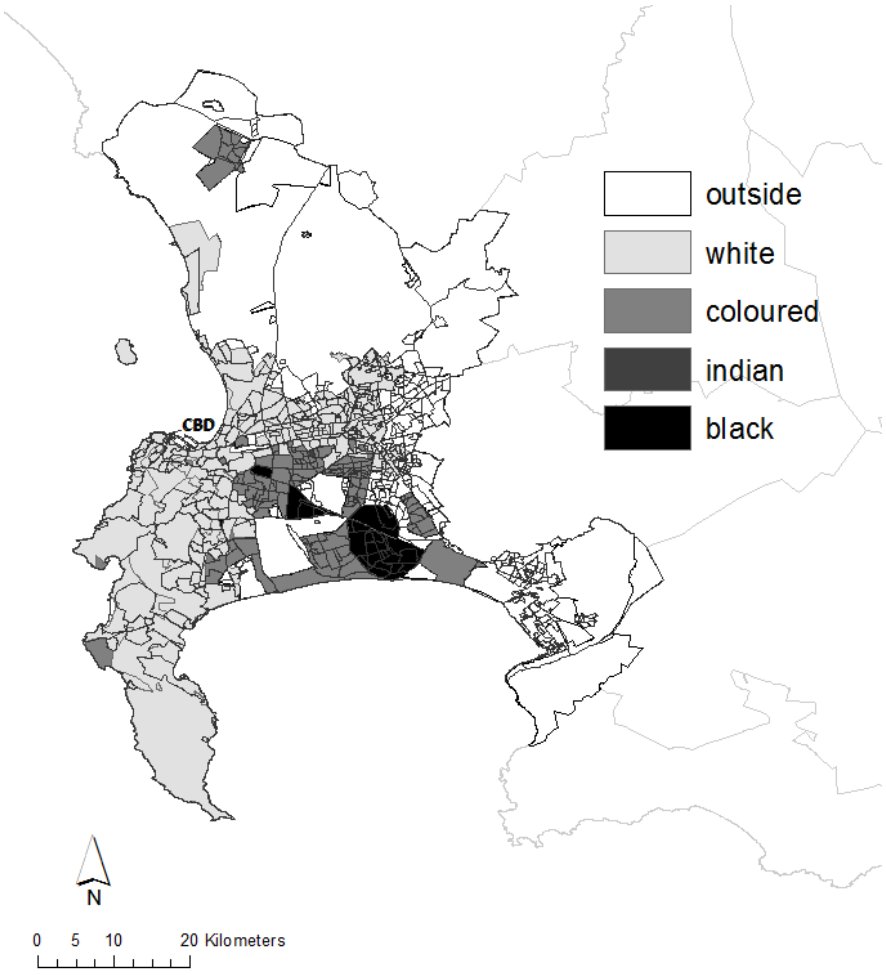
S

Figure 2.5 Maps of Group Areas Act 1950 in Cape Town.

[on personal website version of the paper]

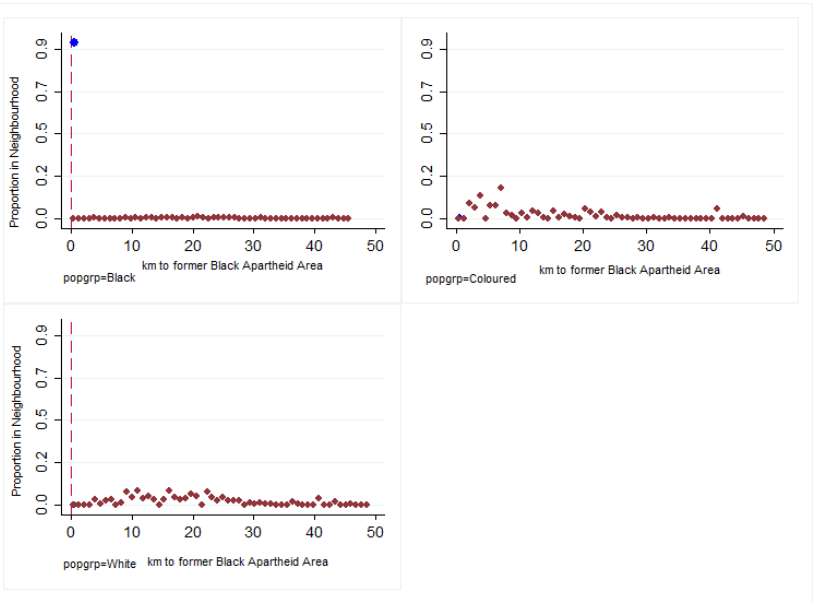
Notes: Maps obtained from the library of the Municipality of Cape Town. These maps show the division of neighbourhoods in Cape Town according to different ethnic groups defined by the Group Areas Act.

Figure 2.6 Digitized Map of Group Areas Act in the City of Cape Town



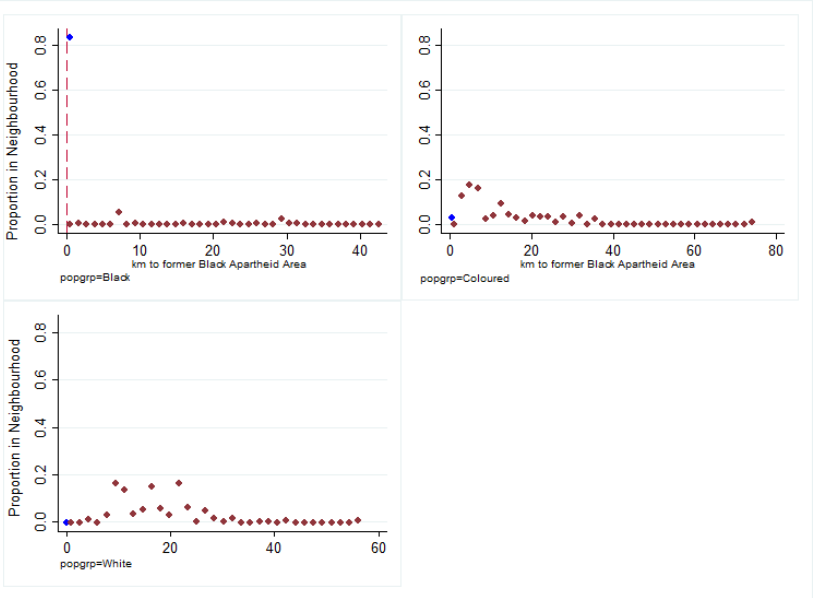
Notes: Digitized version of the maps in Figure 5 for the entire City of Cape Town by 1991. White, Coloured, Indian and Black areas were specifically designated areas where these ethnic groups could reside. Outside refers to areas that were not ethnically defined.

Figure 2.7 Compliance with Apartheid Designation, Census 1985



Notes: The dots are neighbourhood averages of the proportion of the ethnic group (y-axis), plotted against the distance in km to areas designated as black-only during Apartheid (x-axis). The averages are calculated using the Census 1985. The figure plots white, black and coloured ethnic groups, clockwise.

Figure 2.8 Compliance with Apartheid Designation in the sample (by 1991)



Notes: The figure is as figure 7. Here, the averages are calculated using the analysis dataset, restricting the sample to those younths that reside in the area by 1991. The figure plots white, black and coloured ethnic groups, clockwise.

Table 2.1: Ethnic Groups & Former Apartheid Locations

	Black Apartheid Ghetto		**Coloured Apartheid Ghetto**	
	(1)	(2)	(3)	(4)
	All sample	By 1991	All sample	By 1991
Black	0.814***	0.902***		
	(0.012)	(0.012)		
Coloured			0.443***	0.510***
			(0.016)	(0.020)
N	3913	2166	3913	2166
R-squared	0.685	0.822	0.273	0.315

Robust standard errors in parentheses
*** p<0.01, **p<0.05, * p<0.1

Table 2.2: Descriptive Statistics: Main Outcomes (by 1991 sample, *Compliers*)

Young Adult	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Monthly earnings	2455	1718	277	11470	356
Average monthly earning in the past 5 years	2116	1651	14	11470	526
Months worked by 2009	15.028	17.754	0	82	782
Months working since last in school	2.228	6.704	0	36	657
Months looked for work by 2009	10.853	13.503	0	68	782
Months looked for work since last in school	12.428	13.565	0	63	559
Years of education completed	10.392	1.906	3	16	773
Ever attended college	0.028	0.165	0	1	782
Years delayed from graduating grade	3.903	3.352	0	14	497
Smokes cigarettes	0.194	0.396	0	1	660
Consumes alcohol	0.325	0.469	0	1	659
Consumes drugs	0.050	0.218	0	1	660
Ever reported a live birth∗	0.390	0.488	0	1	410

Notes: Main outcomes variables are computed across the years in the panel dataset (2002-2009). All earnings are in 2010 South African Rands; average monthly earnings in the past 5 years corresponds to the average of the last observed earning between 2009 and 2006; all variables measured by 2009 are the # of months up until 2009 and thus across the period of the panel, and all variables "since left school" are the # of months since last attended any formal education institution; idle corresponds to not working nor in school; total years delayed from graduation are calculated using the upper-bound (+ 1 year) of the grade expected age of graduation for the last grade attended; ever reported a live birth is calculated for females only.

Table 2.3: Descriptive Statistics: Outcomes 2004-2009 (By 1991 Sample, *Compliers*)

Young Adult	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Monthly earnings	2232	1784	19	20950	849
Currently working (paid)	0.313	0.464	0	1	2039
Idle	0.424	0.494	0	1	2039
Unemployed - job search $<$ 2 months	0.177	0.381	0	1	2039
Unemployed - job search $>$ 2 months	0.231	0.422	0	1	2039

Notes: Main outcomes variables for all young adults (YA) older than 20 years old and not enrolled in school between 2004 and 2009. Currently working is a dummy variable if YA answers receiving payment for his work; monthly earnings are in 2010 South African Rands.

Table 2.4: Ghetto Characteristics (Census datasets)

Ghetto	Gugulethu (1)	Khayelitsha (2)	Langa (3)	Nyanga (4)
1985				
Black	0.983	0.996	0.991	0.996
Male	0.528	0.491	0.586	0.508
No Education	0.201	0.274	0.218	0.275
Working Age [15-65]	0.697	0.592	0.74	0.636
Bantustan birth	0.270	0.381	0.409	0.590
Population(thousands)	63.89	3.53	22.99	148.88
1996				
Black	0.978	0.989	0.983	0.983
Male	0.508	0.487	0.497	0.493
No Education	0.104	0.128	0.103	0.149
Speaks Xhosa	0.970	0.972	0.513	0.964
Population(thousands)	79.16	180.29	46.88	142.28
2001				
Black	0.988	0.995	0.997	0.991
Male	0.485	0.480	0.497	0.477
No Education	0.133	0.073	0.079	0.142
Working Age [15-65]	0.824	0.803	0.821	0.805
Speaks Xhosa	0.957	0.968	0.978	0.965
Informal Dwellings	0.518	0.639	0.665	0.633
Employed	0.315	0.344	0.357	0.295
Employed in manufacturing	0.089	0.101	0.127	0.098
Employed in Construction	0.071	0.126	0.096	0.137
HH below poverty line	0.673	0.706	0.754	0.773
Access to electricity	0.611	0.767	0.508	0.580
Population (thousands)	74.16	323.79	49.67	96.37
Population Density (tho. per km2)	13.29	10.10	13.17	6.47
Area (km2)	5.58	32.07	3.77	8.09
Murder rate (2007)	0.076	0.085	0.029	0.138

Notes: Data from Census 1985, 1996 and 2001 (Statistics South Africa). Poverty line defined as less than 20 thousands rands per year. Murder rate obtained from the City of Cape Town Crime Statistics Report (2007).

Table 2.5: Descriptives Ghetto Characteristics (Sample)

	Gugulethu	Khayelitsha	Langa	Nyanga
	(1)	(2)	(3)	(4)
Distance to CBD (km)	15.214	25.382	10.456	17.181
Distance to closest school (km)	0.183	0.431	0.247	0.230
Distance to closest coloured or white school	1.666	2.342	1.574	3.084
# of assault on school premises (1999)	11.571	4.753	5.917	16.175
# of serious crimes on school premises (1999)	1.00	1.079	0.273	1.259
Pupil to teacher ratio	60.00	51.589	60.00	47.013
Access to Employment (definition 1, 2001)	0.084	0.048	0.163	0.074
Access to Employment (definition 1, 2005)	0.107	0.064	0.182	0.098
Accesss to Employment (definition 2, 2009)	0.120	0.077	0.136	0.115
% of jobs within 5km (2009)	0.058	0.034	0.105	0.060
N	139	457	28	158

Notes: Data on schools was obtained from SNR 2000; Access to employment is measured as the number of jobs in all other neighbourhoods in Cape Town weighted by their distance from the neighbourhood's centroid in km (2001 and 2005 use definition 1 - where jobs are proxied with the percent of firm's turnover; 2009 uses definition 2 where jobs are proxied by the number of employed by firm.)

Table 2.6: Balance Tests of Ghetto characteristics [CAPS Sample]

	Gugulethu	Khayelitsha	Langa	Nyanga	Overall	p-values
	(1)	(2)	(3)	(4)	(5)	(6)
Distance to Coloured Ghetto (km)	2.000	4.672	2.000	4.557	4.078	0.000
	(0.000)	(0.060)	(0.000)	(0.072)	(0.054)	
% African-Black (2001)	0.990	0.994	0.997	0.992	0.993	0.000
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	
% Coloured (2001)	0.009	0.006	0.002	0.008	0.007	0.000
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	
Distance to CBD (km)	15.214	25.683	10.456	17.184	20.068	0.000
	(0.000)	(0.031)	(0.000)	(0.014)	(0.056)	
Acces to employment (def 1, 2001)	0.084	0.048	0.164	0.074	0.064	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	
Acces to employment (def 1, 2005)	0.107	0.064	0.182	0.098	0.083	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	
Acces to employment (def 2, 2009)	0.120	0.077	0.136	0.115	0.095	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	
Schools former black DOE*	0.547	0.536	0.500	0.576	0.545	0.805
	(0.042)	(0.023)	(0.096)	(0.039)	(0.018)	
Schools former white or coloured DOE*	0.094	0.066	0.179	0.139	0.090	0.013
	(0.025)	(0.012)	(0.074)	(0.028)	(0.010)	
Distance to closest school (km)	0.183	0.431	0.247	0.230	0.340	0.000
	(0.000)	(0.017)	(0.000)	(0.001)	(0.010)	
Distance to closest no-black school	1.666	2.342	1.574	3.084	2.344	0.000
	(0.000)	(0.091)	(0.000)	(0.049)	(0.057)	
Pupil to teacher ratio (YA school)*	64.273	54.063	64.273	48.058	55.030	0.000
	(0.000)	(0.442)	(0.000)	(0.804)	(0.360)	
N	139	457	28	158	782	

Notes: Standard errors in parentheses, p-values of joint-orthogonality test of treatment arms. *1999 from SNAPS dataset, former black, coloured and white DOE refers to DET, HOR and HOA schools during Apartheid. Distances are calculated as Euclidean distances from the centroid of an individual neighbourhood of residence.

Table 2.7: Ghetto Quality Index 2001, by components (z-scores)

	Gugulethu	**Langa**	**Khayelitsha**	**Nyanga**
Amenities	0.570	-0.726	0.604	-0.448
Formal dwellings	1.465	-0.788	-0.385	-0.293
Households above poverty line	1.181	-0.609	0.445	-1.018
Human capital	-0.109	0.710	0.677	-1.277
Inv. Km to CBD	0.302	1.050	-1.344	-0.008
Inv. Km to non-black schools	0.687	0.820	-0.153	-1.354
Inv. Murder rate	0.137	1.182	-0.062	-1.257
Total Index	0.605	0.234	-0.031	-0.808
Total Rank	1	2	3	4

Notes: All variables are standardized census variables from the 2001 census. Amenities include the proportion of households (HHs) with access to electricity, water inside their dwelling and refuse collection. HHs above the poverty are HHs with annual incomes above 20 thousands current Rands (2001). Human capital includes the proportion of adults that have at least some years of formal education, and the number of adults with higher education more than secondary (Matric). Km to CBD are the Euclidean distances to CBD from each neighbourhood centroid, Km to non-black schools are the Euclidean distances to the neighbourhood centroid to the closest white or coloured school. Both are the inverse distances so that proximity is considered as positive. The final variable is the inverse of the murder rate so that the higher values are attributed to the lowest murder rates. The higher the index the better the neighbourhood in terms of these characteristics.

Table 2.8: Balance Test of Youth Characteristics

	Gugulethu	Khayelitsha	Langa	Nyanga	Overall	p-values
	(1)	(2)	(3)	(4)	(5)	(6)
Year of birth	1984	1984	1984	1984	1984	0.690
	(0.219)	(0.123)	(0.489)	(0.192)	(0.090)	
Male	0.482	0.484	0.500	0.437	0.471	0.815
	(0.043)	(0.023)	(0.096)	(0.040)	(0.017)	
First Move (if not always)	1990	1990	1991	1991	1990	0.345
	(0.581)	(0.206)	(0.522)	(0.296)	(0.160)	
Speaks English	0.020	0.006	0.000	0.009	0.014	0.019
	(0.014)	(0.004)	(0.000)	(0.009)	(0.005)	
Speaks Xhosa	0.960	0.978	0.929	0.955	0.963	0.042
	(0.020)	(0.008)	(0.071)	(0.020)	(0.008)	
Speaks Afrikaans	0.000	0.006	0.000	0.009	0.007	0.641
	(0.000)	(0.004)	(0.000)	(0.009)	(0.003)	
Speaks Other	0.020	0.009	0.071	0.027	0.017	0.357
	(0.014)	(0.005)	(0.071)	(0.015)	(0.005)	
N	139	457	28	158	782	

Notes: Standard errors in parentheses, p-values of joint-orthogonality test of treatment arms.

Table 2.9: Balance Test of Parental and Household Characteristics

	Gugulethu (1)	Khayelitsha (2)	Langa (3)	Nyanga (4)	Overall (5)	p-values (6)
Mother:						
No education	0.035	0.083	0.000	0.043	0.067	0.078
	(0.017)	(0.014)	(0.000)	(0.017)	(0.009)	
Secondary incomplete	0.561	0.482	0.679	0.543	0.518	0.160
	(0.047)	(0.025)	(0.090)	(0.043)	(0.018)	
Secondary completed - Matric	0.096	0.063	0.071	0.058	0.068	0.765
	(0.028)	(0.012)	(0.050)	(0.020)	(0.009)	
Father:						
No education	0.052	0.159	0.050	0.122	0.139	0.086
	(0.029)	(0.022)	(0.050)	(0.035)	(0.016)	
Secondary incomplete	0.345	0.377	0.350	0.333	0.356	0.728
	(0.063)	(0.029)	(0.109)	(0.050)	(0.022)	
Secondary completed - Matric	0.172	0.091	0.200	0.144	0.122	0.139
	(0.050)	(0.017)	(0.092)	(0.037)	(0.015)	
Household:						
YA co-resident with a parent	0.628	0.674	0.704	0.713	0.673	0.712
	(0.046)	(0.025)	(0.090)	(0.041)	(0.018)	
Parent resident if < 22 yrs	0.719	0.812	0.750	0.766	0.783	0.198
	(0.038)	(0.018)	(0.083)	(0.034)	(0.014)	
HH size at baseline	7.590	5.835	6.107	6.937	6.357	0.000
	(0.298)	(0.108)	(0.510)	(0.256)	(0.099)	
Connected to electricity	0.956	0.954	0.963	0.975	0.960	0.879
	(0.019)	(0.011)	(0.037)	(0.014)	(0.008)	
Grew up in formal urban	0.870	0.353	0.857	0.841	0.556	0.000
	(0.029)	(0.023)	(0.067)	(0.029)	(0.017)	
Brackets of HH income	15.440	14.631	16.778	15.552	15.044	0.011
	(0.339)	(0.199)	(0.845)	(0.329)	(0.150)	
N	139	457	28	158	782	

Notes: Standard errors in parentheses, p-values of joint-orthogonality test of treatment arms.

Table 2.10: Results I: Education in young adulthood.

	Years of education	Ever attended college	Delay in graduating grade
	(1)	(2)	(3)
Gugulethu	0.316	0.047*	-0.216
	(0.280)	(0.026)	(0.374)
Langa	0.791**	0.112*	0.093
	(0.381)	(0.067)	(0.652)
Nyanga	0.299	0.056*	-0.062
	(0.240)	(0.029)	(0.350)
All Controls	Y	Y	Y
R-squared	0.001	0.041	0.018
N	600	592	600

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. The main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers 'spending most of their lives (formal and informal urban, formal and informal rural, with one excluded). The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991); blacks only. *** p<0.01, **p<0.05, * p<0.1

Table 2.11: Results II: Education in young adulthood by sex.

	Years of education	Ever college	Delay in graduating grade
	(1)	(2)	(3)
Females:			
Gugulethu	0.517	0.061*	-0.868*
	(0.348)	(0.037)	(0.450)
Langa	0.818	0.212*	0.265
	(0.616)	(0.117)	(0.970)
Nyanga	0.225	0.058	-0.607
	(0.261)	(0.039)	(0.447)
R-squared	0.016	0.021	0.324
N	319	323	183
Males:			
Gugulethu	0.033	0.035	0.653
	(0.423)	(0.039)	(0.576)
Langa	0.706	-0.010	-0.297
	(0.463)	(0.022)	(0.826)
Nyanga	0.354	0.055	0.782
	(0.408)	(0.044)	(0.532)
R-squared	0.030	0.006	0.222
N	273	277	157
All Controls	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. Controls are Table 10 (excluding sex). The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. Regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

Table 2.12: Results I: Labour Outcomes across period 2004-2009.

	Earnings (log)	Working	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
Gugulethu	-0.046	0.011	-0.014	0.074*	0.058
	(0.093)	(0.050)	(0.050)	(0.044)	(0.043)
Langa	0.146	0.156*	-0.157*	-0.008	-0.061
	(0.166)	(0.090)	(0.090)	(0.068)	(0.075)
Nyanga	0.023	0.133***	-0.153***	0.019	-0.021
	(0.083)	(0.048)	(0.049)	(0.043)	(0.046)
All Controls	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y
R-squared	0.129	0.069	0.072	0.010	0.005
N	711	1251	1251	1251	1251

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) are log monthly earnings, in column (2) working is a dummy variable for having any paid work, columns (3) and (4) measure short term (ST) and (LT) unemployment defined as not employed and looking for a job below and above 2 months, respectively. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school.The main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers spending most of their lives (formal and informal urban, formal and informal rural, with one excluded) and the results of an aptitude quantitative test administered on year one of the survey to all participants. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991); blacks only. . *** p<0.01, **p<0.05, * p<0.1

Table 2.13: Results I: Labour outcomes in young adulthood.

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Gugulethu	0.075	0.618	-0.089	4.929**	5.594***
	(0.100)	(2.130)	(0.758)	(1.991)	(2.036)
Langa	0.290*	4.855	3.496	-5.150**	-4.257*
	(0.175)	(4.464)	(2.220)	(2.021)	(2.282)
Nyanga	0.013	5.134**	0.491	3.935**	4.548***
	(0.088)	(2.052)	(0.888)	(1.730)	(1.684)
R-squared	0.094	0.216	0.008	0.115	0.112
All Controls	Y	Y	Y	Y	Y
N	446	590	543	590	538

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in a formal education institution, earnings corresponds to average log monthly earnings for the last two periods. The main explanatory is a dummy variable for the ghettos of residence. Controls and sample are as in Table 12. *** p<0.01, **p<0.05, * p<0.1

Table 2.14: Results II: Labour outcomes across period 2004-2009 by sex.

	Earnings (log)	Working	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
Females:					
Gugulethu	-0.145	0.042	-0.050	0.033	0.012
	(0.123)	(0.071)	(0.070)	(0.056)	(0.058)
Langa	0.222	0.202	-0.202	0.020	-0.050
	(0.202)	(0.125)	(0.125)	(0.093)	(0.091)
Nyanga	-0.036	0.133**	-0.155**	0.011	-0.005
	(0.104)	(0.067)	(0.066)	(0.049)	(0.056)
N	361	678	678	678	678
Males:					
Gugulethu	0.049	-0.022	0.025	0.128*	0.109
	(0.108)	(0.070)	(0.070)	(0.071)	(0.072)
Langa	0.034	0.120	-0.124	-0.042	-0.072
	(0.189)	(0.140)	(0.140)	(0.103)	(0.129)
Nyanga	0.034	0.133*	-0.153**	0.037	-0.037
	(0.107)	(0.069)	(0.070)	(0.069)	(0.073)
N	350	573	573	573	573
All Controls	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) are log monthly earnings, in column (2) working is a dummy variable for having any paid work, columns (3) and (4) measure short term (ST) and (LT) unemployment defined as not employed and looking for a job below and above 2 months, respectively. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school. Sample and controls are the same as in Table 12. Sex is excluded as regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

Table 2.15: II: Labour outcomes in young adulthood by sex.

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Females:					
Gugulethu	0.038	-2.659	0.551	3.737	4.285*
	(0.142)	(2.864)	(1.079)	(2.601)	(2.489)
Langa	0.531**	6.650	5.870	-6.305***	-5.955**
	(0.226)	(7.023)	(3.923)	(2.120)	(2.644)
Nyanga	-0.017	5.998**	1.421	4.305**	5.146***
	(0.118)	(2.831)	(1.281)	(1.944)	(1.870)
N	224	317	295	317	289
Males:					
Gugulethu	0.119	3.665	-0.721	6.295**	6.972**
	(0.124)	(3.007)	(1.001)	(3.012)	(3.098)
Langa	0.002	2.857	0.922	-3.834	-3.040
	(0.135)	(5.130)	(1.751)	(3.534)	(3.506)
Nyanga	0.019	4.029	-0.758	3.403	3.658
	(0.122)	(3.244)	(1.152)	(3.022)	(3.037)
N	222	273	248	273	249
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to average log monthly earnings for the last two periods. The main explanatory is a dummy variable for the ghettos of residence. Controls and sample are the same as in Table 12, sex is excluded as regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

Table 2.16: Results I: Skill Sectors across period 2004-2009.

	High-skill	Med-skill	Low-skill	Low-skill(b)
	(1)	(2)	(3)	(4)
Gugulethu	-0.020	0.010	0.081**	0.112***
	(0.027)	(0.045)	(0.037)	(0.040)
Langa	-0.041	0.178**	0.086	0.087
	(0.032)	(0.079)	(0.075)	(0.076)
Nyanga	-0.013	0.086**	0.071*	0.078*
	(0.023)	(0.044)	(0.042)	(0.043)
All Controls	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y
R-squared	0.005	0.035	0.037	0.030
N	768	768	768	768

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables are dummy variables for working in a high-skill occupation (column (1), medium skill occupation (column (2)), and low-skill occupations (columns (3) and (4)). The skill level of occupations is defined according to SSA definitions. Low-skill in column (4) also include armed forces and others. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school. Controls and sample are the same as in Table 12. *** p<0.01, **p<0.05, * p<0.1

Table 2.17: Results I: Health & behavioural outcomes in young adulthood.

	Cigarettes	Alcohol	Other drugs	Ever Pregnant
	(1)	(2)	(3)	(4)
Gugulethu	0.152***	0.158**	0.006	-0.071
	(0.055)	(0.067)	(0.026)	(0.075)
Langa	-0.030	-0.022	0.007	-0.011
	(0.073)	(0.087)	(0.052)	(0.136)
Nyanga	0.063	0.120**	0.015	-0.027
	(0.044)	(0.052)	(0.030)	(0.075)
All Controls	Y	Y	Y	Y
R-squared	0.218	0.201	0.041	0.164
N	535	534	535	317

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in columns (1) to (3) are dummy variable equal to one if YA smokes, drinks alcohol and consumes drugs in the last period. Ever pregnant is a dummy variable if the YA was ever pregnant in the period studied (for females only). Sample and controls are as in Table 12. *** p<0.01, **p<0.05, * p<0.1

Table 2.18: Results III: Channels, Labour outcomes across period 2004-2009 (all)

	Earnings (log)	Working	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
% Black 1996	0.037	0.047	-0.045	-0.062**	-0.060*
	(0.066)	(0.040)	(0.040)	(0.031)	(0.035)
Access Index 2009	0.035	0.087***	-0.091***	0.002	-0.018
	(0.051)	(0.028)	(0.028)	(0.025)	(0.025)
Former *black* school	-0.165**	0.058	-0.055	-0.020	-0.031
	(0.083)	(0.050)	(0.051)	(0.048)	(0.050)
Km to *white or coloured* school	-0.029	0.005	-0.006	0.003	-0.005
	(0.019)	(0.012)	(0.012)	(0.009)	(0.009)
R-squared	0.175	0.077	0.077	0.012	0.006
N	679	1183	1183	1183	1183
All Controls	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to average log monthly earnings for the last two periods. All regressions are separate regressions where the main explanatory is a dummy variable for the ghettos of residence. Controls are is in Table 11. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. *** p<0.01, **p<0.05, * p<0.1

Table 2.19: Results III: Channels, Labour outcomes in young adulthood (all)

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
% Black 1996	-0.088	0.437	0.134	-3.554**	-3.650**
	(0.088)	(1.853)	(0.903)	(1.601)	(1.613)
Access Index 2009	0.073	2.734**	0.770	1.380	1.824*
	(0.057)	(1.294)	(0.516)	(0.975)	(1.026)
Former *black* school	-0.086	-1.671	-0.071	-0.468	-0.678
	(0.076)	(1.625)	(0.706)	(1.450)	(1.456)
Km to *white or coloured* school	-0.033	-0.039	0.387	0.807	0.886
	(0.020)	(0.483)	(0.373)	(0.542)	(0.615)
R-squared	0.091	0.212	0.005	0.106	0.103
N	432	572	526	572	520
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to log monthly earnings in the final period (2009). All regressions are separate regressions where the main explanatory is a dummy variable for the ghettos of residence. Controls are as in Table 10. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. *** p<0.01, **p<0.05, * p<0.1

Table 2.20: Results III: Channels, on Education in young adulthood (all)

	Years of education	Ever college	Delay in graduating grade
	(1)	(2)	(3)
% Black 1996	0.141	-0.023	0.355
	(0.185)	(0.020)	(0.300)
Access Index 2009	0.339**	0.045***	-0.052
	(0.141)	(0.016)	(0.216)
Former *black* school	-0.491**	-0.003	-0.146
	(0.234)	(0.028)	(0.315)
Km to *white or coloured* school	0.019	0.004	0.188**
	(0.079)	(0.006)	(0.083)
Subsidy Transport	0.350*	0.012	-0.128
	(0.196)	(0.020)	(0.314)
R-square	0.090	0.004	0.202
N	393	397	271
All Controls	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. Controls are as in Table 12. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. Regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

Chapter 3

Cholera in Times of Floods[1]
Weather Shocks & Health in Dar es Salaam

3.1 Introduction

Climate change will have a significant impact on the lives of the poor in the years ahead as extreme weather events such as floods, heavy precipitation and droughts are expected to become more frequent (Harrington et al. 2016). Populations at different stages of development are affected differently by the same weather variations (Dasgupta, 2010; Burgess et al. 2017). As cities in developing countries continue to urbanize at an unprecedented pace, the question of how their urban dwellers are impacted by weather shocks is becoming increasingly relevant. On the one hand, urban residents may seem better prepared than their rural counterparts against weather extremes; their livelihoods are for instance less dependent on weather phenomena. Yet, rapid urbanization has often led to *unplanned* cities with poor infrastructure, limited public service provision, and with large segments of the population living in informal settlements.

Empirical evidence is scant concerning the impact of weather shocks in developing country cities. This paper tries to make progress on this issue by looking at the effect of rainfall and flooding on cholera incidence in Dar es Salaam. Looking at health outcomes is important. Contagion is one "downside of density" (Glaeser & Sims 2015). Throughout history, cities with low-quality infrastructure and poor sanitation have been pockets of epidemics (for instance 19th century London or Paris, Kesztenbaum &

[1] This research has benefited from valuable comments and conversations with Gharad Bryan, Ed Glaeser, Vernon Henderson, Aurélie Jandron, Henry Overman, Olmo Silva, as well as seminar participants at the LSE, the 7th European meetings of the UEA, and the 2017 Columbia University Ph.D. Workshop in Sustainable Development. We are also grateful to Mark Iliffe and Tanner Reagan for sharing datasets, and the Tanzanian Regional Medical Officers and Municipal Officers for assistance with accessing confidential cholera datasets, specifically Ndeniria Swai, Alex Mkamba, Michael Gweba, and Lawrent Chipatta, as well as Theophil Malibiche. Joshua Chipman and Claire Lwehabura were also very helpful during the process of this research.

Rosenthal 2016). Poor health and disease not only lower productivity in the short-term, they also hinder long-term economic growth (Well 2007).

We examine this question in the context of a cholera outbreak in Dar es Salaam in 2015 and 2016, during which almost five thousand cases were recorded. Cholera is a water and food-borne disease and its transmission is closely linked to inadequate access to clean water and sanitation facilities. Weather shocks such as rainfall can affect health through two main channels. The first one is related to direct mechanisms operating via human physiology and disease. In this case, heavy rainfall and floods can increase exposure to vibrio cholera bacteria, which survives better in wet environments (Lipp et al. 2002, Osei et al. 2010). Droughts have also been linked to cholera outbreaks, as population use unfit water for their needs (Sasaki et al. 2008, Taylor et al. 2015). The second sets of mechanisms are indirect ones, through the effect weather shocks may have on real incomes. In the case considered here, there are many ways through which floods and rainfall can reduce accessibility to work. The resulting lower income may in turn lower the consumption of health-improving goods (i.e. safe drinking water, medicines), increasing the exposure to the bacteria.

Our empirical analysis uses finely disaggregated ward-level panel data containing weekly recorded cholera cases and weekly accumulated precipitation for all the municipalities in the city. We are therefore testing whether exogenous weekly rainfall variation at the ward-level affects cholera occurrence. Sorting between neighbourhoods is taken care of with ward-fixed effects. The short timeframe considered excludes any concern related to changes in sorting patterns. The use of high-frequency data and the fine geographical detail allow us to estimate with precision our relationship of interest. We focus on reduced-form specifications; different spatial and time-lag models support our choice. We complement our data with ward-level infrastructure characteristics (i.e. roads, footways, drains, water wells and housing informality) to understand the relationship between infrastructure quality, precipitation, and cholera. Dar es Salaam is a city of more than 4 million people where close to 70% of its residents live in informal settlements. Access to improved sanitation is very low and only 37% of the city has regular refuse collection (World Bank 2017).

We find robust evidence that weekly accumulated rainfall and flooding leads to higher ward-level cholera occurrence. On average, a 10 mm increase in weekly accumulated precipitation leads to an increase of between 1.5% to 3.5% of weekly recorded cholera cases, significant at 1% level. The effect is found to be much larger when considering a more flexible quartile specification. On average, a single additional

week of rainfall falling above the 75th percentile of the total rainfall distribution (extreme rainfall), increases the number of effective cholera cases by up to 20.3% relative to a week with very light rain (<0.30 mm). Further, we find that the impact of heavy rainfall (>75th percentile) is close to 20 percentage points higher in wards at greater risk of flooding (i.e. higher flood-prone area), all else equal. A dry week (0 mm rainfall) is also positively related to cholera incidence, but the coefficient is not statistically significant. These findings are consistent with the direct and indirect mechanisms put forward earlier. Particularly, the inundation of drains, water systems, and pit latrines greatly enhances the risk of exposure to contaminated water and food. These results are robust to alternative estimators and specifications, including an instrumental variable strategy and controlling for the spatial autocorrelation of standard-errors.

Remarkably, we find little to no spatial spillovers from precipitation in neighbouring wards. Only when considering the relative elevation of contiguous wards there is a small significant effect from precipitation in downhill wards. That is, a 10 mm increase in weekly accumulated rainfall in downhill neighbouring wards increases cholera cases in the ward by 0.01% to 0.03%. This finding is consistent with water source contamination. In contrast, our results reveal moderate time-dynamic effects. Using a distributed lag model, we find significant positive effects of past rainfall on cholera incidence to up to five preceding weeks. Contemporaneous rainfall remains the largest determinant, with a stable size close to 3% and statistical significance at or above 5% level. All lags decrease in size the farther in time.

We explore the non-linear relationship of rainfall and cholera incidence with respect to various ward-level infrastructure characteristics as well. We find the effect of weekly rainfall on cholera cases to be consistently higher for wards with larger shares of informal housing and a higher density of footways (i.e. informal roads). These results are consistent with the two possible mechanisms outlined earlier. Neighbourhoods with limited access to sanitation and low-quality infrastructure are likely to be more exposed to the cholera bacteria when surfaces are washed and drains are overflown by severe precipitation. Vulnerable populations in these wards are also more likely to suffer from negative income shocks during extreme weather events.

Our findings relate to three different bodies of research. The first one is a large economics literature looking at weather shocks and economic events mostly in advanced economies (Munshi 2003; Miguel et al. 2004; Barrios et al. 2010), and including health outcomes of human populations (Deschenes & Greenstone 2007; Deschenes & Moretti 2009; Burgess et al. 2011; Deschenes 2011). Our paper follows

their empirical methods (for a review see Dell et al. 2014). The second is the large literature in development economics and public health studying policy interventions, mechanism of transmission and health outcomes in developing countries (Banerjee et al. 2004; Miguel & Kremer 2004; Dunkle et al. 2010; Penrose et al. 2010; Devoto et al. 2012). Here, we contribute to that literature by focusing on urban areas and by studying general mechanisms beyond specific policy interventions. Our findings use robust econometrics techniques to discern the relationship between disease, infrastructure, and weather shocks. Finally, we particularly contribute to the nascent literature in urban economics studying the effects of weather phenomena on urban areas in developing country cities (Kocornik-Mina et al. 2015; Glaeser & Henderson 2017; Henderson et al. 2017). To the best of our knowledge, our paper is the first to empirically study the effect of weather shocks on disease transmission within a city of a developing country. Our findings have important policy implications as extreme weather events become more frequent in the next decades. Cities in developing countries need to address infrastructure gaps to contain the risk of recurrent epidemic outbreaks in vulnerable populations and neighbourhoods. Investing in resilient infrastructure, with the proper servicing of informal settlements or related measures such a regulating waste dumping may prove to be more beneficial in the long-term than the use of short-term palliative measures during outbreaks. Evidence on large-scale policy interventions in urban areas is still limited and more is needed to understand how to prevent contagion of treatable diseases in developing cities.

This paper is organized as follows. The next section formalizes the relationship between health and weather shocks, specifically here cholera and rainfall in cities. Section 3.3 describes the context of Dar es Salaam, the cholera disease, and the data. In section 3.4 we present our empirical strategy. Section 3.5 presents the estimates of the effect of rainfall and flooding on cholera incidence and different extensions and robustness checks. Section 3.6 concludes.

3.2 Theoretical Framework: Weather & Health

In this section, we describe a simple theoretical framework to examine the various channels through which weather shocks (i.e. here, heavy precipitation) can affect cholera incidence in an urban setting. It relies on endogenous health models in a simplified fashion (for more details see Becker 2007; Deschenes 2012; Burgess et al. 2017).

Cholera is an acute diarrheal infection caused by ingestion of food or water contaminated with the bacterium vibrio cholera. It can affect both children and adults and can kill within hours if left untreated. The main reservoirs of the cholera bacterium are people and warm salt water bodies such as estuaries and coastal areas. Cholera transmission is closely linked to inadequate access to clean water and sanitation facilities (WHO 2017)[2]. Here we assume that weather shocks such as extreme rainfall, droughts and flooding can affect human health (i.e. cholera prevalence) both directly (through higher exposure to the bacteria, for instance), and indirectly (due to the negative income-shocks that may arise through the weather shock).

Consider a city with a large number of agents indexed by *i*. Agents seek to maximize their lifetime utility u_i depending on consumption c_{it} and health status h_{it}. These two are complements. This city is partitioned into several neighbourhoods and each agent lives in a given neighbourhood (or ward) indexed by *n*. We assume that agents are exogenously allocated to neighbourhoods with different characteristics such as better infrastructure, and the latter are exogenously determined. It follows that agents have limited scope for affecting local infrastructure, as well as other public goods and services, and take these as given (mobility is limited in the short-term scenario considered here).

Agent *i*'s health status, resident of neighbourhood *n* at time *t*, (h_{int}), is thus determined by her consumption c_{it} and random health shocks z_{int}. We do not specify a precise relationship between health outcomes and consumption, but it is assumed that an individual can improve her health by increasing her consumption, particularly by purchasing health-improving goods. All told, the agent's health status is thus given by:

$$h_{int} = h_i(c_{it}, z_{int}) \qquad (3.1)$$

where $h_i(\cdot)$ is increasing in c_{it} and decreasing in z_{int} (adverse health shocks). The function h_i is unrestricted and can differ across heterogeneous individuals. The vector z_{int} includes weather-shocks w_t such as flooding, droughts, and heavy rainfall or temperature extremes. We assume further that the effect of the weather shock is conditional on the quality of local neighbourhood infrastructures (q_n) such as access to drinking water and waste water systems, or road and pavement material. The adverse health shock is thus a function of the weather shock that varies with local infrastructure, $z_{int}(w_t(q_n))$.

[2] http://www.who.int/mediacentre/factsheets/fs107/en/, accessed on 28 June 2017.

Consumption[3] c_{it} is financed through labour income in period *t*, which depends on the agent's productivity a_{it}. Productivity is agent-specific but also depends on weather shocks w_t:

$$a_{it} = a_i(w_t) \qquad (3.2)$$

Here the effect on productivity can stem from the weather shock's impact on z_{int}, hampering the agent's ability to work efficiently, or deter accessibility to jobs or other factors of production (i.e. machinery, location)[4]. A given weather shock w_t thus affects an agent's health status through both consumption (via productivity) and health idiosyncratic shocks (via z_{int}). In other words, there are two fundamental mechanisms through which weather shocks (extreme precipitation for the purpose of our empirical analysis) can potentially harm an agent's health status here.

First, through direct health effects: random shocks z_{int}, enter the agent's health status directly as in equation (3.1). That is, holding constant the agent's income, location, and consumption decisions, we expect a negative weather shock to impact this agent's health adversely (w_t impacts z_{int} in the language of our model). In the case considered here, heavy rainfall and flooding can directly impact one's health status through greater exposure to and contact with contaminated water and food. An extensive public health literature discusses the potential for cholera prevalence in wet environments and in cases of heavy precipitation and flooding (see for example, Osei et al. 2012). Further, the magnitude of the effect can be expected to depend on the relationship the weather phenomena and the disease pathogens have with local infrastructure characteristics. Cholera strives in stagnant water and poor hygiene conditions as explained earlier.

The second, more indirect mechanism through which weather can affect health in this model is through the agents' productivity in equation (3.2). This term depends on weather shocks that may affect the agent's ability to work via z_{int}. Flooding and heavy rainfall may also significantly affect work-places and accessibility in contexts where poor roads and infrastructure is widespread (see footnote 4). Reduced productivity can

[3] We assume the consumption good is produced using an aggregate production function that requires capital and labour inputs; it exhibits decreasing return to scale. Goods can be bought and sold at the market price, which is exogenously determined. Agents are subject to budget constraints in each period which are a function of the labour income (in turn dependent on productivity and adverse weather shocks), as well as prices and quantities of goods consumed. We assume imperfect credit and savings markets which prevents agents from smoothing their consumption in time.

[4] We assume $a_{it} = a_i(z_{int}(w_t), Q_{nt}(w_t))$, where Q_{nt} refers to complements to work such as accessibility to jobs, machinery, or location, that can be affected by weather shocks.

translate in lower earnings and reduced consumption of healthier quality goods such as clean water or medicines. The dependency of productivity and hence, labour income, on this type of shocks is extremely likely in low-income countries where informal jobs dominate employment. It is estimated that close to 80% of jobs in the services sector are informal in Tanzania (UNDP 2015). Note that this assumes imperfect credit and savings markets preventing agents from smoothing their consumption when hit by economic hardship. Given the Tanzanian context, this assumption does not seem unrealistic.

The main implication of this exercise is to expect an increase in cholera cases due to a weather shock such as extreme precipitation and flooding. The increase should be larger in neighbourhoods with poorer infrastructure. Conversely, the impact of heavy rainfall should be mitigated in areas with a supply of higher-quality local public goods. Linking this section with our empirical analysis, we expect to see non-linear effects of rainfall on cholera occurrence.

Finally, one additional implication of this simple formalization concerns policy interventions. In the face of potential weather shocks, any agent *i* would seek to minimize the damage that the negative shock has on their utility, $u_i(c_{it}, h_{it})$ by consuming health-improving goods or potentially by reallocating resources between periods. This latter option here is limited due to credit and savings constraints. These potential shock-minimizing strategies have strong implications for policy. In the theoretical framework considered here, governments can reduce the adverse effect of weather on health outcomes by directly increasing the quality of infrastructure that is related to the pathogens' transmission (i.e. pavement, sanitation, water drainage, sewage), and thus directly limit the potential effect of an adverse health shock, z_{itn}. But they can also intervene by supporting the agent's shock minimization strategies through subsidized health goods or direct transfers.

3.3 Background & Data

This section provides further details on cholera-specific characteristics as well as Dar es Salaam's context. It also describes the data in detail and provides basic summary statistics.

3.3.1 Cholera

Cholera is an acute diarrhoeal infection of fecal-oral transmission. It is caused by the ingestion of food or water contaminated with the bacterium vibrio cholera. It takes between twelve hours and five days for a person to show symptoms after ingesting contaminated food or water. It can affect both children and adults and can kill within hours if left untreated; there is a 50% death rate if untreated, but all deaths are avoidable otherwise. Main treatments include antibiotics and Oral Rehydration Salts (ORS). Roughly 1.3 to 4.0 million cases are recorded worldwide every year, and the disease is endemic to many parts of sub-Saharan Africa and South Asia (WHO 2017).

There are multiple pathways for cholera transmission (Clasen et al. 2007). The disease is closely linked to inadequate access to clean water and sanitation facilities. Risk factors are also considered to be high population density and crowding, all of which are often common in urban slum areas (Penrose et al. 2010). Cholera incidence has been found to be highest in highly urbanized areas (Osei & Duker 2008; Sur et al. 2005). The main reservoirs of the cholera bacterium are people and warm salty water bodies such as estuaries and coastal areas. Global warming and rising sea levels are believed to create a favourable environment for cholera bacterium growth (WHO 2017). Heavy rainfall and flooding have all been associated with a higher likelihood of cholera outbreak. Surface runoff from point sources (pit-latrines, waste dump site, water wastes) may cause increased contamination of water sources, while stagnation and slow flowing of waterways may lead to increased exposure to cholera vibrios (Osei et al. 2010).

3.3.2 Dar es Salaam

Dar es Salaam is one of the largest cities in eastern Africa. It is located in the east of Tanzania by the Indian Ocean. Its urban population grew at 6.5% yearly between 2002 and 2012 (Wenban-Smith 2014), and today the city counts with more than 4.4 million people. Since 2016, it is divided in five municipal districts: Ilala, Temeke, Kinondoni, Kigamboni and Ubungo[5]. These municipalities are further divided up into 90 wards.

The rapid pace of urbanization has led to large infrastructure deficits. Close to 70% of Dar es Salaam's residents live in informal settlements without adequate access to clean water, proper drainage system and waste collection (UN-HABITAT 2010; Natty 2013). Only 13% of the city's residents have adequate sewage systems and 37% of the

[5] In the analysis, we only use 3 municipal districts as these were the ones that existed at the time cholera cases were recorded during the last outbreak. Ubungo and Kigamboni were created in 2016 from dividing Kinondoni and Ilala further so this does not impact our findings in any way.

solid waste is properly collected. The World Bank (2015) estimates that only 50% of residents have access to improved sanitation. The most common form of improved sanitation is improved pit-latrines (other forms are rare). About two-thirds of households in the city share their toilet facilities. Access to piped water is also very limited, with only 17% of city-centre dwellers having piped-water.

Dar es Salaam's geography and coastal location makes it vulnerable to climatic hazards, particularly floods, sea level rise and coastal erosion (Kebede and Nicholls 2010). There are two rainy seasons every year, the short (October to December) and long (March to May) seasons, and average annual precipitation is above 1,000 mm. The combination of high informality and climatic vulnerability makes flood risk one of the main challenges for sustainability, exposing infrastructure and residents to safety and health hazards from vector-borne diseases such as malaria and cholera (World Bank 2017).

Cholera has been endemic in Tanzania since the 1970s and Dar es Salaam has historically been the most affected region[6]. During the 2015-2016 outbreak, there were over 24,000 cases recorded nationally, with more than one fifth in Dar es Salaam (Figure C1 in appendix)[7]. Previous outbreaks occurring between 2002 and 2006 reported over 30,000 cases nationally, with nearly 18,000 in the capital city (WHO 2008). Given the city's poor sanitary conditions, high population density, lack of access to safe drinking water, and limited drainage, continuous heavy rainfall makes stagnant and unsanitary water a widespread health risk for common water borne diseases. The lack of storm water drains, frequently blocked by unregulated waste dumping, means that heavy rainfall quickly leads to flooding and contaminates water wells (Pan-African START 2011).

3.3.3 Data

To examine the relationship between weather variation and cholera incidence outlined in our theoretical framework, we collect data from several sources and put together a comprehensive ward-level panel dataset for each week between the first week of March 2015 and the first week of September 2016. The choice of the timeframe is data constrained – that is, we use the first week for which precipitation data is available and the last week for which cholera cases were recorded to avoid measurement error

[6] The largest cholera epidemic in Tanzania to date took place in 1997 where 40,000 cases were reported. The epidemic is said to have started in Dar es Salaam. Dar es Salaam has had the most cholera cases since 2002 of all regions of the country (Penrose et al. 2010).
[7] All tables and figures indexed by C# are in Appendix C.

from unrecorded cases. We cover all the 90 wards of the city[8]. The use of high-frequency data and the fine geographical detail allow us to estimate with precision our relationship of interest. The basic panel thus consists of weekly cholera cases registered according to the ward of residence. We combine this data with weekly accumulated precipitation and weekly air-temperature in these wards. Further, we add data on ward-level infrastructure, geographical characteristics (i.e., elevation, flood-prone surface) and population (census). We outline below the different data sources. Main summary statistics are in Tables 3.1 and 3.2.

3.3.3.1 Cholera cases

The key data in this analysis are the new ward-level cholera cases collected from the Regional Medical Office and Municipal Health Officers for all the wards of Dar es Salaam and covering the entire 2015-2016 outbreak. The data was registered daily for each individual presenting symptoms of severe diarrhoea in a medical facility. It includes basic socio-demographic characteristics (age, sex) of the individual, the ward and sub-ward of residence, as well as the date of the first symptom and registration at the hospital. Cases were tested for the vibrio cholera bacteria, and the dataset also includes lab results. We exclude all cases tested negative and focus on effective cholera cases only[9]. No positive case is reported earlier than mid-August 2015 (epidemic week zero). The outbreak officially lasted from August 2015 until May 2016. We aggregate the daily cases by week to account for the fact that the incubation period is between 12 hours to 5 days.

Measurement error is a potential problem. The biggest threat concerns the possibility that not all cholera cases are reported in the non-outbreak period. It is also possible that not all registered cases during the peak of the epidemic are effective cholera cases (see footnote 8). There are mitigating factors against both these risks. First, cholera is one of the few diseases that require reporting to the World Health Organization (WHO) by the International Health Regulations as it can quickly spread if left untreated and result in explosive outbreaks. This implies careful monitoring of the disease as well as frequent laboratory testing. Further, we focus our analysis during an outbreak where monitoring is more likely to be enforced. Finally, our baseline estimates

[8] Since mid-2016 there are 5 municipalities in Dar es Salaam as two municipalities were further sub-divided. We use the original administrative units at the time of the outbreak in our regression analysis for simplicity and coherence with the recorded cholera dataset. This should not affect any of the results.
[9] We include both positive and untested cases. Most untested cases are at the peak of the outbreak when all patients presenting symptoms are treated as cholera patients. Measurement error is possible but should be limited as tests are frequently carried, particularly at the beginning and end of the outbreak period.

are weighted by the population of the ward, to account for the difference in precision concerning cholera measurement from larger and smaller populated wards. While bias from measurement error in our dependent variable is still possible, it should not be large.

Overall, close to 5 thousand cases of cholera were reported positive in Dar es Salaam in the period analysed (4964 of total 5698 tested), with the bulk taking place during the first 10 weeks of the outbreak (Figures 3.1 & 3.2). On average, during the period covered there were 0.72 effective cases weekly per ward. The number is larger during the first 10 weeks of the epidemic (3.16) as well as the first 20 weeks of the epidemic (2.54) (Table 3.2). Cholera cases were more pronounced in Kinondoni and Ilala, reporting totals of 2428 and 1796, respectively. Temeke was the least affected (Figure C2). Most cases took place within 15 km from the Dar es Salaam CBD (Figure 3.3); only 2 of the 90 wards reported zero positive cases throughout the period.

3.3.3.2 Weather & geography

Rainfall - The weather datasets in this paper are from NASA. The daily precipitation measures by ward are derived from the Integrated Multi-satellite Retrievals (IMERG) for Global Precipitation Mission (GPM), where rainfall is comprehensively measured at the highest accuracy and finest spatial resolution to date (Huffman, 2016). We use the near-real-time total daily rainfall defined as precipitation accumulated in the past 24 hours by 23:59pm (Coordinated Universal Time) of each day. We calculate weekly accumulated precipitation from the daily data. In terms of the spatial resolution, rainfall is measured at squared pixels of $0.1 \circ \times 0.1 \circ$ (roughly $120 km^2$).

As ward boundaries are irregularly shaped, we compute ward-level daily rainfall accumulation by weighting recorded rainfall with the ward overlay with satellite pixels. We first union these two layers to create polygons at the ward-pixel level. These ward-pixel polygons all have consistent rainfall measurement, and their respective area is computed. We then sum up the ward-pixel rainfall measures for each ward by weighting by their ward area share. This gives us the area-weighted weekly rainfall accumulation at the ward level (Figure C3). The choice of focusing on rainfall accumulation (i.e. total weekly rainfall) stems from the fact that precipitation is 'readily stored' in the soil, tanks, or water wells. It is stagnant water that might breed cholera and thus, measuring average rainfall instead would fail to take this important dimension into account.

Because satellite data are subject to error (Dell et al. 2014), we also use an additional and independent gridded data set to address potential measurement issues

and obtain instrumental variables (IV) estimates. We use precipitation obtained from IMERG's predecessor technology, the Tropical Rainfall Measuring Mission (TRMM) (Goddard Earth Sciences Data and Information Services Center 2016). Despite the fact that TRMM is less accurate (Shari et al. 2016; Chen and Li. 2016; Wang et al. 2017) and its resolution coarser, it has been widely used since 1997. Its algorithm intercalibrates all existing satellite microwave precipitation measures, microwave-calibrated infrared satellite estimates, and precipitation gauge analyses. The near-real-time data is chosen over the production data as it is recommended for flood and crop forecasting (NASA Precipitation Measurement Missions 2016). The instrumental variables approach is motivated by the fact that both satellite measures assign weather variables to grid points and contain measurement error in their 'true' representation of rainfall. In that case, the IV estimates can correct for measurement error bias under the assumption that errors in both variables are uncorrelated (Burgess et al. 2017).

Temperature - The daily temperature data also comes from NASA. We obtained near-surface air temperature (i.e., temperature at the height of most human activities) from the FLDAS Noah Land Surface Model (McNally 2016). The spatial resolution of this dataset is also $0.1 \circ \times 0.1 \circ$, so ward-level daily temperature is computed similarly to rainfall above. Average weekly temperature is later computed at the weekly level.

Elevation – The elevation calculation is based on the Japan Aerospace Exploration Agency (JAXA) global digital surface model. The measurement is at 30-meter spatial resolution, based on the most precise global-scale elevation data at this time acquired by the Advanced Land Observing Satellite. Mean ward-level elevation is computed across all grids that fall inside each ward.

Flood-prone surface – To estimate the surface of a ward that is prone to flooding, we use data collected by the NGO Dar Ramani Huria (RH) in OpenStreetMap (OSM) format. Using community-based mapping RH is able to create highly accurate maps of infrastructure and flood-prone areas in Dar es Salaam. We complement their detailed mapping of drainage, waterways and wetlands with GeoFabrik's OSM data for missing wards. The data is less accurate but allows us to have a larger coverage. We then use InaSAFE[10] to model build-areas prone to inundation and calculate the total share of the ward area that is flood-prone. We compare our estimates to the more precise-ones of RH for available wards. The pairwise correlation is 0.81.

[10] InaSAFE is a free software that produces realistic natural hazard impact scenarios. It was developed by the government of Indonesia, the Australian government and the World Bank. For more details see http://inasafe.org/ (last accessed on July 21st 2017).

Basic summary statistics of weather and geographical variables are displayed in Table 3.2. The average weekly rainfall in Dar es Salaam according to the meteorological agency amounts to 20.6 mm. This is consistent with our weekly accumulation from both TRMM and GPM's measures. On average, in the period covered there were 20.1 mm of accumulated rainfall weekly, with a median of 2.9 mm. The rainiest month is usually April, which is seconded by our dataset. There is little spatial variation of temperature across the city's' wards, the average recorded weekly is 26.7° C with a standard deviation of 0.37° C. On average 10% of the area of a ward is prone to flooding, but there are significant disparities across wards (the standard deviation being 16%).

3.3.3.3 Infrastructure & population

Infrastructure - Infrastructure data at the ward-level is also obtained from data collected by RH's in OSM format, and complemented with GeoFabrik's for missing wards. We focus on the following characteristics which are likely to be correlated with cholera incidence: drains, roads, footways (i.e. unpaved roads) and water wells. For the first four variables, we use their density, calculated as the number of km per square km. Aside from roads where we can distinguish between roads and footways, we have no specific measure of quality of the infrastructure. A general assumption is to think that a higher density of roads and drains reflect higher-quality infrastructure, while a higher density of footways reflects lower-quality. The distinction in practice is hard to make, particularly for drains. Anecdotal evidence suggests drains often get clogged by unregulated waste dumping due to heavy rainfall and quickly contaminate surfaces. We are thus agnostic concerning the expected signs of these coefficients. We have unfortunately no data on sewerages[11].

We obtained a dataset of formal and informal plots from the municipalities' database of surveyed plots, and are then able to estimate the share of the ward's area that houses informal settlements. Not all municipalities have mapped their informal plots fully[12] which explains the smaller sample when using this data. For the wards for which we have information, 34% of the total areas are on average informal. The large number reflects the fact that 70% of the population of Dar es Salaam lives in informal settlements.

[11] Basic sanitation data in Table 3.1 is obtained from the 10% sample of the Census 2012. Unfortunately, these are only used in the descriptive section because of the lack of consistency in the sample.
[12] Only the Municipality of Kinondoni has.

Population- We make use of the population data from the Census 2012 to weight our regressions by ward population size. The interest in this is twofold. First, cholera incidence in wards with large populations is likely to be more precise, so weighting corrects for heteroskedasticity associated with these differences in precision (Burgess et al. 2017). Second, rather than on the average ward, the results reveal the impact on the average person, which is more meaningful here. We also use this data to calculate ward-level population density. The average ward of Dar es Salaam was populated with 48.5 thousand people in 2012; population density was 11.53 per square km (Table 3.1).

3.4 Empirical Strategy

In this section, we describe the econometric methods we use to estimate the effect of precipitation on cholera occurrence. As the relationship between rainfall and new cholera cases is expected to be non-linear, we adopt both parametric and flexible semi-parametric specifications. We begin by presenting specifications measuring the contemporaneous effect of precipitation. We then consider models allowing for the effect of rainfall to be associated with local public goods provision and other ward characteristics. We also assess the importance of the spatial spillovers of precipitation. Lastly, the last sub-section details a more general dynamic model including various precipitation time lags.

3.4.1 Contemporaneous effects

To quantify the contemporaneous effect of rainfall on cholera incidence in any given ward and week, we begin by estimating a baseline panel log-linear model relating the logarithm of cholera cases[13] to weekly rainfall accumulation for this ward:

$$C_{wmt} = \alpha . R_{wt} + \gamma . T_{wt} + \mu_w + \delta_t + \theta_m . t + \varepsilon_{wmt} \qquad (3.3)$$

where C_{wmt} is the outcome variable (log of total cholera cases) in ward *w* in week *t*. The key explanatory variable of interest is R_{wt} , measuring weekly accumulated rainfall. We also control for ward daily temperatures measured as weekly averages (T_{wt}) as temperature variation is likely to be correlated with rainfall variation. Since our focus is

[13] Since no cholera cases are recorded in several wards and weeks in our sample, we add one to all cholera cases and take the logarithm of that expression. In mathematical terms: $C_{wmt} = \ln(1 + C_{wmt})$. We test the results to linear regressions where the dependent variable is the ratio of cholera cases for every ten thousand people of the ward. Results are unchanged (section Appendix C, section C.IV) and we prefer the specification that considers the non-linear relationship between rainfall and cholera.

on precipitation and spatial variation in temperature in Dar es Salaam is limited, we model a linear temperature effect. The specification in equation (3.3) also includes a full set of ward fixed effects, μ_w , absorbing unobserved time-fixed ward idiosyncratic characteristics. Permanent differences in access to healthcare for instance will therefore not confound the estimates. Their inclusion also addresses the potential issue of sorting across neighbourhoods. We also include week fixed effects, δ_t, to control for time-varying influences common across wards. The equation also includes municipality linear time trends to account for time-varying factors that differ across administrative boundaries and affect health. We also estimate equation (3.3) with municipality-week fixed effects to flexibly control for unobserved municipality-wide time shocks. We use only three municipalities in the analysis as these were the administrative divisions existing at the time of data collection. Further, the main three hospitals are located in these municipalities. As shown later, our estimations across these specifications are consistent and robust. ε_{wmt} is an error term clustered at the ward level. Finally, we weight our regressions by ward population as explained earlier. Unweighted regressions are in Appendix C (section C.I). Results are unchanged.

To take into account non-linear relationships more rigorously, we also estimate contemporaneous rainfall effects using the following flexible model:

$$C_{wmt} = \sum_{k=1}^{4} \beta_k . 1\{R_{wt} \text{ in quartile k}\} + \phi . T_{wt} + \mu_w + \delta_t + \theta_m . t + \eta_{wmt} \quad (3.4)$$

where the independent variables we are mainly interested in capturing are the distribution of weekly rainfall in Dar es Salaam. The regressors $1\{R_{wt}$ in quartile k$\}$ calculate whether the total amount of rainfall R_{wt} in week *t* and ward *w* was in the first, second, third, or fourth quartile of the rainfall distribution of our study period. We estimate a separate coefficient on each of these quartile variables and treat the second quartile as the omitted reference category. The other regressors are as defined in equation (3.3). This approach has two benefits. The first one is to allow for more flexibility in the response function. The second one, more relevant here, is that it also allows us to specifically distinguish the effect of intense and light rain. The upper quartile (>75th percentile) is generally used as a proxy for flooding (Chen et al. 2017).

The parameters in equations (3.3) and (3.4) are thus identified from ward-specific deviations in rainfall from the ward average remaining after controlling for week fixed effects and municipality linear trends. Given the relatively short time period of analysis

we argue that this variation is as good as exogenous and uncorrelated with other unobserved determinants of cholera incidence.

Equations (3.3) and (3.4) make several important assumptions about the effect of rainfall on cholera. First, they assume that the impact depends on weekly accumulation alone. It ignores the possibility of within week variation in rainfall having an effect on health. In addition, equation (3.4) assumes that the impact of rainfall is constant within a given quartile. While this might be restrictive, we estimate separate quartile coefficients to improve on equation (3.3) and its parametric assumptions. Third, by estimating contemporaneous effects, we assume that past weekly rainfall does not affect health outcomes. We also ignore the possibility of neighbouring wards' rainfall influencing a given ward's cholera outcomes. We relax some of these assumptions in what follows.

A final concern is spatial dependence. In this case, within-cluster correlations in the specification of the error covariance matrix (i.e., standard-errors clustered at the ward level) may not be enough (Barrios et al. 2012). To account for this issue, we also compute equations (3.3) and (3.4) using Conley (1999) spatial standard-errors[14]. The implicit assumption is that spatial dependence is linearly decreasing in the distance from the wards centroids up to a cutoff distance, for which we chose 50 km based on Dar es Salaam's extent. This technique ensures that uncertainty in α and β is adjusted to account for heteroscedasticity, ward-specific serial correlation, and cross-sectional spatial correlation. Statistical significance is generally unchanged. We consider these results as robustness checks in Appendix C Section C. II.

We are interested in reduced-forms here. However, we are conscious that the true (unknown) relationship may include some time dependency in the dependent variable. That is, past cholera may determine contemporaneous cholera. To test the validity of our fixed-effects model, we compute a dependent-lagged model instead in Table C15 in Appendix C. While we find the effect of lagged cholera cases significant, and positive up to 5 weeks, the size of the coefficient for contemporaneous precipitation always remain stable and statistically significant. Further, contemporaneous rainfall is orthogonal to past cholera. Including the lags would only increase the precision of our point estimates but should not alter the identifying assumptions.

[14] We use the Stata code developed by Fetzer (2010) and Hsiang (2010).

3.4.2 Non-linear effects and spatial spillovers

Ward characteristics, such as population density or the number of water wells, may affect the impact of rainfall on health as outlined in our theoretical framework. To account for this possibility, we estimate variations of equation (3.3) that include interactions between rainfall and ward features. While local public goods are not exogenously allocated to wards, there are several reasons to believe this is not a problem here. First, the use of ward fixed effects should deal with neighbourhood sorting. Further, the lack of proper infrastructure is widespread in Dar es Salaam and public health evidence suggests households from all income-levels may be affected by cholera. Using the 2015-16 Tanzania Demographic and Health Survey and Malaria Indicator Survey (DHS) we test the relationship between income, wealth, and incidence of diarrhoeal diseases in the city (Appendix D). We find no evidence in favour of a wealth bias regarding the risk of contracting a diarrheal disease.

We also measure whether contemporaneous precipitation in neighbouring wards affect cholera cases in a given ward. We focus on first contiguity wards and consider total neighbouring accumulated rainfall to begin with. We then distinguish between rainfall recorded in uphill and downhill neighbouring wards. We calculate the average elevation of each unit and classify as uphill the neighbouring wards with a relatively higher elevation. Downhill neighbouring wards have a lower or equal average elevation. Formally the model we estimate is as follows:

$$C_{wt} = \rho_1 . R_{wt} + \rho_2 . UR_{wt} + \rho_3 . DR_{wt} + \pi . T_{wt} + \mu_w + \delta_t + \varsigma_{wt} \quad (3.5)$$

where UR_{wt} and DR_{wt} measure weekly accumulated rainfall in uphill and downhill neighbours, respectively. The other regressors are defined as in equations (3.3) and (3.4).

3.4.3 Dynamic effects

The empirical approaches discussed so far do not address the possibility of a dynamic relationship between precipitation and new cholera cases. Rainfall in one week might result in increased cholera incidence in the following weeks due its incubation period and the manner in which the disease spreads. This delayed response would imply that the contemporaneous estimates from equation (3.3) underestimate the true impact of rainfall. We investigate this possibility by including a distributed lag structure in our models:

$$C_{wt} = \sum_{j=0}^{J} \lambda_j . R_{wt-j} + \rho . T_{wt} + \mu_w + \delta_t + \zeta_{wt} \quad (3.6)$$

This model allows the effect of rainfall up to *J* weeks in the past to affect cholera incidence in a given week. In equation (3.6), the total dynamic effect of rainfall on cholera cases is obtained by summing the coefficients on the contemporaneous and lagged rainfall variables. Different lag structures potentially generate different estimates of the dynamic causal effect. As a consequence, we experiment with several time lags and use up to 5 lagged weekly accumulated precipitation in our regressions.

3.5 Main Results

This section presents our empirical results on the relationship between precipitation and cholera incidence. We begin with discussing baseline contemporaneous estimates of both rainfall and flooding. We then assess the importance of non-linear effects, spatial spillovers and measure dynamic effects last.

3.5.1 Baseline effects

Our baseline results concern the effect of rainfall and precipitation on weekly-ward cholera occurrence. Tables 3.3-3.6 report baseline estimates of population-weighted regressions. Unweighted regressions are in Appendix C Section C.I (Tables C.1-C4), while the same regressions with Conley HAC standard-errors are in section C.II (Tables C.7-C9). Conclusions remain unchanged irrespective of the specification.

Table 3.3 reports estimates based on equation (3.3). The first column shows coefficients obtained with ward and week fixed effects only. Precipitation is found to have a positive and statistically significant effect on cholera. The point estimate suggests that a 10 mm increase in weekly accumulated rainfall causes a 2% increase in recorded cholera cases in a given ward. Including municipality linear trends does not affect the results much (column 2). Municipality-week fixed effects are controlled for instead in column 3. While the impact of precipitation remains statistically significant at the 1% level, its magnitude increases; that is, there are 3.4% additional cholera cases per ward every 10 mm increases in rainfall. Overall, these reduced form estimates consistently show a positive impact of precipitation on cholera incidence.

All subsequent tables are organized in the same fashion, with municipality trends added in column 2 and municipality-week fixed effects added in column 3. To test the sensitivity of our results to measurement error in the recorded rainfall data, we instrument our main precipitation variable with rainfall recorded by the TRMM satellite as explained in section 3.3.3.2. The potential sources of measurement error in these two

datasets are likely to be unrelated, and therefore uncorrelated. Results are displayed in Table 3.4. The two satellite-based precipitation variables are strongly correlated, and first stage F statistics range between 24 and 37 across specifications (see fourth row). Our two-stages least squares coefficient estimates remain positive but become larger as attenuation bias theory would predict. Including municipality-week fixed effects results in a loss of statistical significance (column 3). The first stage F-statistic is also lower however, inflating standard errors to some degree. On the whole, our findings are supported by the IV results. There is a strong positive relationship between cholera occurrence in a given ward and precipitation. In the interest of proceeding conservatively we continue to stress the OLS results hereafter, but Table 3.4 suggests that the true impact of precipitation on cholera may be even larger.

Table 3.5 explores the impact of rainfall using the more flexible quartile specification detailed in equation (3.4). The second rainfall quartile (light rain or precipitation between 0 to 2.9 mm weekly) is used as omitted category. Notably, the semi-parametric relationship between weekly accumulated rainfall and cholera occurrence show particularly large effects at the upper-end of the rainfall distribution. Indeed, the estimated coefficients in the three columns consistently indicate that extreme precipitation has a strong impact on cholera incidence. For instance, a single additional week with recorded rainfall falling in the fourth quartile (>75th percentile, between 26.9 mm and 408.6 mm weekly), relative to a week with light rain, increases the number of new cholera cases by 20.3% (column 2). The first quartile coefficients, measuring loosely speaking the effect of a dry week relative to little rainfall, are positive but not statistically significant. These results are key findings in our paper. Clearly, extreme rainfall has a higher incidence on ward-level cholera occurrence than light rain, suggesting not all ranges of precipitation are necessarily related to cholera occurrence. Moreover, upper-quartile rain has been consistently used in the literature as a proxy for flooding (Chen et al. 2017), and implies water stagnation may be a likely mechanism.

To explore further the role of extreme precipitation, Table 3.6 puts the emphasis on flooding and attempts to measure its impact in various ways. We begin with assessing whether the impact of rainfall is non-linear and depends on the extent to which a ward is prone to flooding. We use our measure of the share of the ward that is subject to flooding and interact it with weekly accumulated rainfall. Our results presented in panel A show a positive interaction term as theory would predict. The interaction is non-statistically significant however. The coefficient of the uninteracted precipitation measure remains in the same order of magnitude as the coefficients of Table 3.3. In

panel B we measure the effect of the fourth quartile precipitation relative to the rest of the precipitation distribution. Here flooded is a dummy variable for weekly accumulated rainfall falling on the upper-quartile of the overall rainfall distribution. Our estimates are positive, significant at the 1% level, and stable across alternative specifications. In panel C we interact the flooded dummy with our flood-prone area share defined as above. The interaction term is now positive and significant at 5% level, implying that the impact of heavy rainfall is much higher in wards at greater risk of flooding all else equal.

Overall, the results of this section support the theoretical mechanisms described in section 3.2 and the channels put forward in the public health literature. There are various reasons why heavy rainfall and flooding could lead to an increase in cholera, as mentioned earlier. Not only the bacterium survives longer in wet humid surfaces, but the risk of increased contamination is higher. The inundation of drains, water systems, and pit latrines, greatly enhances the probability of exposure to contaminated water and food. Further, behavioural changes during periods of weather shocks may also increase the probability of contagion (WHO). Finally, indirect mechanisms through income-shocks due to the inundation of job locations or inaccessibility to the work-place may further contribute to the adoption of risky behaviour.

3.5.2 Non-linear effects: a story of infrastructure quality?

We now explore further the relationship between rainfall and cholera incidence and assess potential non-linearities related to ward-level characteristics. As explained above, the size of the weather shock in a given ward is likely to depend on the quality of the infrastructure such as the availability of well-functioning drains, paved roads and improved sanitation. This section concentrates on the correlation that rainfall and several indicators of ward infrastructure and 'neighbourhood quality' has with cholera incidence. Our choice of ward characteristics is in part dictated by data availability. We focus on population density, road density, as well as the density of drains and footways, and the number of water wells. We also include the percentage of the ward's area that hosts informal and formal housing.

As mentioned earlier, we are constrained when it comes to measuring the quality of infrastructure and focus on quantity when no distinction is possible. Because of this, while we expect higher population density to increase the measured effect of rainfall on cholera through a heightened risk of contagion, we are agnostic with respect to the influence of road and water infrastructure measures. On the one hand, greater physical supply of water wells and drains could be negatively associated with cholera by

efficiently evacuating used-water and rain. On the other hand, it could magnify the impact of heavy precipitation on cholera when the quality is low, for instance if because of unregulated dumping, drains and evacuation canals are clogged in times of heavy rain.

Table 3.7 reports baseline estimates of population-weighted regressions. Estimates of unweighted regressions are in Table C5, and results with Conley HAC standard-errors are in Tables C10, both in Appendix C. The size of the coefficients is stable across our different specifications. Further, contemporaneous precipitation remains consistently positive and statistically significant at between 1 to 5% levels.

The first seven columns of the tables separately estimate each interaction term, while in the last three columns we estimate all interactions jointly. Since we lose a large number of observations when we include certain interactions, we report results using three alternative samples. Overall, almost all characteristics considered individually are positive and significant at various levels of significance. Yet, only footway densities and housing informality are consistently so across the different specifications. The mechanisms here are intuitive. Footways are unpaved roads. Contaminated water might stagnate easier in muddy surfaces. At the same time, footways could just reflect informality. Indeed, informality displays the larger size, with on average weekly accumulated rainfall increasing cholera incidence by 2.3% to 4.5% more in wards with higher shares of informal housing.

Once we introduce all interaction terms together in columns (8) to (10) significance and signs are considerably changed suggesting individual interactions may be suffering from omitted variable bias. Nonetheless, some important patterns remain. First, the only consistently positive and statistically significant coefficient (at 5% level) across the various specifications is the non-linear informal housing correlation. While, the sample size is much smaller, results suggest living on informal housing *increases* cholera incidence due to weekly accumulated rainfall by between 2.3% and 2.7%. This finding is far from surprising. Informal settlements are usually located in flood-prone areas. They suffer from poor quality infrastructure and deprived water and sanitation conditions. Penrose et al. (2010) find similar patterns when investigating a previous cholera episode in Dar es Salaam. Two other infrastructure interactions display stable sizes and signs. The non-linear water wells correlation is negative but almost never statistically significant. Related to informality, wards with a higher density of footways have again a higher likelihood of accumulated rainfall affecting weekly cholera incidence. The size is small (between 0.2% to 2%), and fails to be statistically significant.

These results support the theoretical mechanisms outlined earlier. First, informal housing and unpaved roads increase the effect of the weather shock. They are thus likely to affect individuals directly and indirectly through health and productivity shocks. Our results so far, despite being imperfectly measured, suggest the quality of infrastructure is highly correlated with the detrimental effect accumulated rainfall and flooding have on cholera prevalence.

3.5.3 Spatial spillovers: Neighbours contagion.

Next, we focus on spatial precipitation spillovers as in equation (3.5). Precipitation recorded in adjacent wards might exacerbate pressure on water infrastructures. They might also contaminate common water sources, particularly if wards are at different levels of elevation. In this case, uphill rainfall may also wash down contaminated waste or soil material, harming wards below. Table 3.8 contains our baseline results, while Tables C6 and C13 in Appendix C contain the usual alternative specifications.

We first estimate the average effect of weekly accumulated rain in neighbouring wards on the cholera incidence of a given ward. We do not find evidence of an effect here. The estimated coefficients are almost no different from zero and insignificant across econometric specifications. In Table C11 of Appendix C we further look at the effect of one and two-weeks lags of neighbouring rainfall. None of these spatial lagged variables seem to matter.

We then differentiate between rainfall accumulation in uphill and downhill neighbouring wards. Results here are more nuanced. We register a small but significant effect from accumulated weekly precipitation in adjacent downhill wards. That is, a 10 mm increase in weekly accumulated rainfall in downhill neighbouring wards increases cholera cases in the ward by 0.01% to 0.03%. This finding is consistent with water source contamination from relatively lower wards. The size is negligible suggesting almost no spatial spillovers from rainfall in contiguous areas. Further, precipitation in one's own ward is almost always significant at 1% level, retaining the size of baseline estimates. Again, allowing for time lags yields no significant effect (Table C12).

Failing to detect any spatial spillovers of precipitation in contiguous wards is unexpected. It suggests only local contamination prevails. This is consistent with findings in Ambrus et al. (2015) on the 1854 cholera epidemic in London's Soho neighbourhood. Their identification strategy and results suggest cholera is contained within a very specific area. Implications concerning channels of transmission are many but go beyond the scope of this paper.

3.5.4 Time dynamic effects.

So far, we have not taken into account the possibility of a dynamic relationship between rainfall and cholera incidence. If cholera responds to precipitation with a delay, that is, if precipitation in previous weeks or days also impacts cholera in the current week, the estimates of Table 3.3 could underestimate the true effect. While cholera symptoms can manifest 12 hours after an individual being in contact with the bacteria, they can also take up to 5 days. These might not coincide with our weekly definition. Further, rainfall is easily stored and stagnation from previous weeks may contribute heavily to contagion.

To test for dynamic effects, we estimate distributed lag models (equation 3.6) and allow rainfall to affect health up to five weeks later. The sixth lag of rain (not shown) is not statistically significant. We also report the contemporaneous coefficient and the sum of the six week period. Table 3.9 displays our point estimates. We gradually introduce additional rainfall lags in our model, which includes municipality-week fixed effects.

First, this exercise allows us to confirm that including time-lags does not change our conclusions. The contemporaneous rainfall effect on cholera remains in the same order of magnitude as in Table 3.3, close to 3% and statistically significant at 1% level. Second, the results in the table clearly show that past rainfall up to five preceding weeks impact cholera incidence in the current week. All lags decrease in size the farther in time, suggesting the contemporaneous effect matters most. The total effect of precipitation is obtained by summing the coefficients on the contemporaneous and lagged precipitation variables. The total cumulated impact amounts to 0.12 points (last row). That is, six-week cumulated rainfall increases current cholera incidence in a ward by up to 12%. The key message of this table is that weekly cumulated rainfall promotes cholera occurrence immediately and with a lag of up to 5 weeks.

We repeat the exercise with our more flexible semi-parametric specification in Table 3.10 (quartiles). The table is as Table 3.9 except that we only include lags up to two weeks later. The third lags (not shown) are not statistically significant. Again, the predominant effect is that of extreme rainfall or flooding captured in the upper-quartile, and up to two prior weeks. As before, all lags decrease in size the farther in time, suggesting the contemporaneous effect matters most. Further, we confirm that the inclusion of the lags does not affect our conclusions as the size and statistical significance are unchanged for contemporaneous coefficients. The total cumulative effect of each quartile is also computed in the last row. The total cumulated impact amounts to 0.37 points for the upper-quartile. It is interesting to highlight that the size

of the two-week lag of the first quartile (no rain) remains positive but increases in size. It is even statistically significant at 10% level for one specification. This supports theories according to which dryness also matters for cholera incidence by increasing the risk of drinking unsafe water. The time lag is consistent with this type of behavioural changes.

3.6 Conclusion

Rapid urbanization in developing countries has often led to unplanned cities, particularly in sub-Saharan Africa, with large shares of the urban population living in informal settlements, with poor transport infrastructure and limited access to water and sanitation. Under these conditions, developing countries' city dwellers have become more vulnerable to disease transmission and epidemics. Global warming is expected to exacerbate these health-related risks. The World Bank (2016) estimates that climate change may push up to 77 million more urban residents into poverty by 2030. As extreme weather events become more frequent, understanding the relationship between disease transmission, infrastructure quality and weather shocks in urban areas is important. We make significant advances on this issue.

The key contribution of this paper has been to show that heavy rainfall has a strong positive effect on weekly cholera incidence within wards. We assemble a panel dataset defined at the ward level containing weekly information on cholera incidence, precipitation, and infrastructure quality from various sources. On average, we find that a 10 mm increase in weekly accumulated precipitation leads to an increase of up to 3.5% of weekly recorded cholera cases. Extreme rainfall has a larger impact: a single additional week of rainfall falling above the 75^{th} percentile of the total rainfall distribution increases the number of effective cholera cases by up to 20.3% relative to a week with very light rain. The impact is even higher in wards at greater risk of flooding.

Results in the paper also emphasize the key role of local infrastructure. We find the effect of weekly rainfall on cholera cases to be consistently higher in wards with larger shares of informal housing and a higher density of footways (i.e. informal roads). These results are in line with the mechanisms outlined. Neighbourhoods with low-quality infrastructure are likely to be more exposed to the cholera bacteria when surfaces are washed and drains are overflown by severe precipitation. Vulnerable populations in these wards are also more likely to suffer from negative income shocks during extreme weather events.

Findings here have important policy implications. Cities in developing countries need to address infrastructure gaps to contain the risk of recurrent epidemic outbreaks in fragile environments. Policies that improve the quality of local infrastructures and housing conditions should mitigate the negative impact of rainfall on health. Given the transmission channels of cholera, the proper servicing of informal areas, including sewerages connections and the pavement of informal roads, as well as the regulation of waste-dumping, may prove to be more beneficial in the long-term than the use of short-term palliative measures during outbreaks. Interventions improving access to drinking water as well as access to sanitation should also greatly reduce cholera risk. In the theoretical framework considered, governments can also reduce the adverse effect of weather on health outcomes by supporting households in periods of health-shocks through subsidized health goods or direct transfers. These policies also need to be taken into account given the large room for increasing social safety nets in urban areas. Evidence on large-scale policy interventions in urban areas are limited and more is needed to understand priority-investments that increase resilience and prevent contagion of treatable diseases in developing cities if these are to become engines of growth (Glaeser 2011).

3.7 Tables & Figures

Figure 3.1 Distribution of cholera effective cholera cases (epidemic weeks)

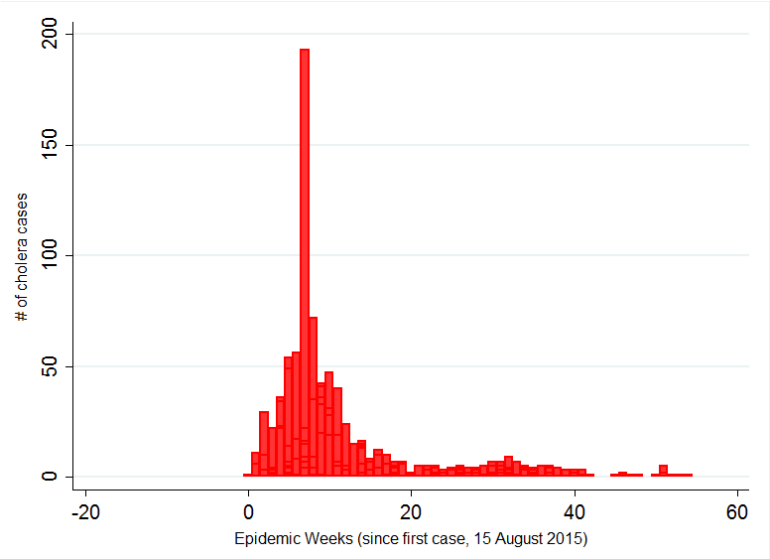


Figure 3.2 Distribution of effective cholera cases, by age (epidemic weeks)

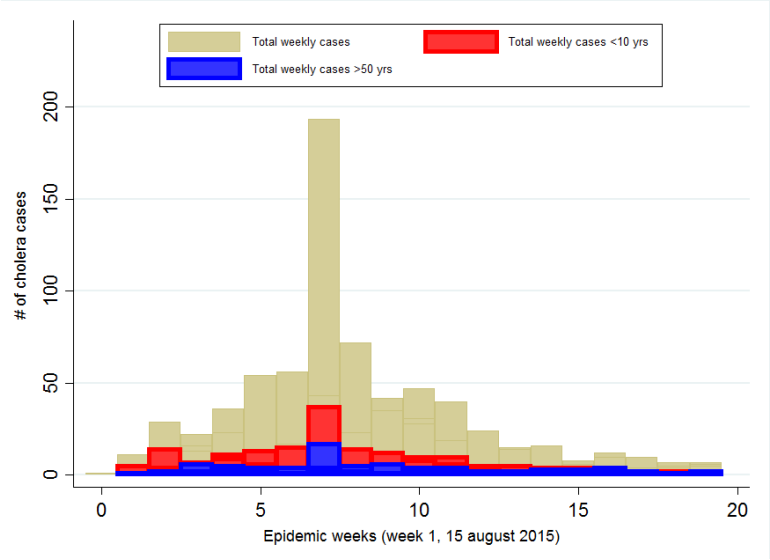


Figure 3.3 Spatial distribution of cholera indicidence, total outbreak (cases per 10,000 inhabitants)

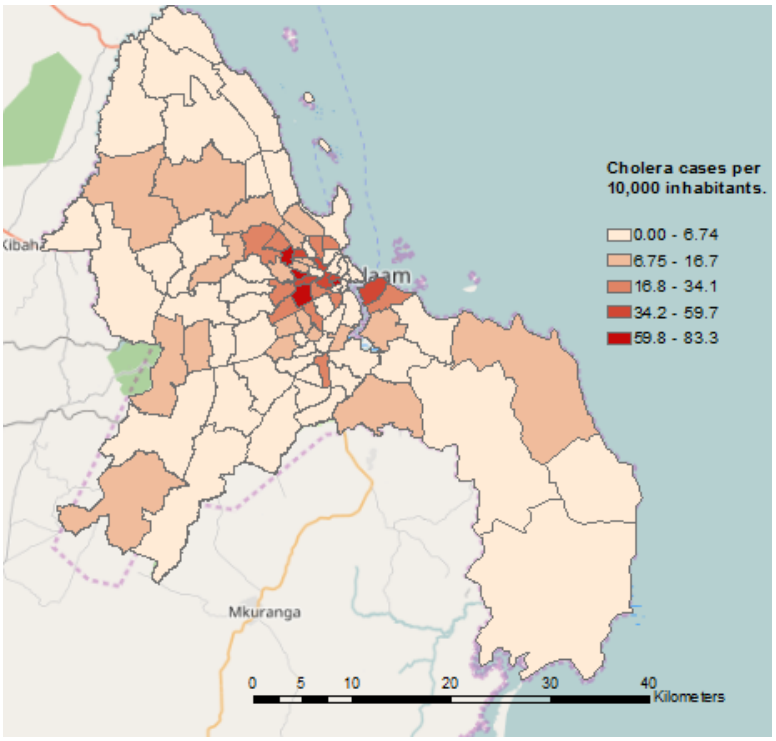


Table 3.1: Summary Statistics: Ward characteristics

	Mean	Std. Dev.	Min.	Max.	N
Area (km2)	18.12	31.15	0.414	209.55	90
Pop (c2012)	48,495	26,064	6,411	106,946	90
HH size (c2012)	4.00	0.21	3.60	4.40	90
Density (km2)	11.53	11.12	0.05	46.74	90
Improved sanitation	0.08	0.10	0.00	0.48	90
Electricity	0.06	0.08	0.00	0.37	90
Drinking water	0.07	0.08	0.00	0.42	90
Hospital per 10tho.	0.19	0.43	0.00	2.97	90
Density of roads*	4.00	3.84	0.00	18	88
Density of footways*	8.83	15.76	0.01	75.99	63
Density of waterways*	5.00	5.61	0.00	24.81	78
Density of drains*	3.28	2.66	0.02	11.04	45
# water wells	1.25	2.99	0.00	21	76
% area informal	34	28	0.00	88	23

Notes: c2012 referes to data from census 2012, all of the infrastructure density measures (*) are measured in km per square km

Table 3.2: Summary Statistics: Weather and Cholera

	Mean	Std. Dev.	Min.	Max.	N
Weather:					
% flooded area	10.00	16.00	0.00	73.00	90
Average temperature (C)	26.73	0.37	25.43	27.17	90
Total rainfall (10mm)	162.36	18.15	134.27	216.42	90
Average weekly temperature (C)	26.73	1.57	22.92	29.99	6930
Weekly rainfall accumulation (10mm), GPM	2.11	4.05	0.00	40.86	6930
Weekly rainfall accumulation (10mm), TRMM	2.67	5.53	0.00	44.62	6930
Cholera:					
Total cases 2015-2016	63.32	94.01	0.00	588	90
Total weekly cases per ward (excl. neg)	0.72	3.84	0.00	192	6930
Total weekly cases female	0.36	1.85	0.00	90	6930
Total weekly cases below 5 yrs	0.08	0.49	0.00	14	6930
Total weekly cases tested neg	0.11	0.56	0.00	20	6930
Total effective cases epiweek10	3.16	9.40	0.00	192	890
Total effective cases epiweek20	2.54	7.25	0.00	192	1780
Total effective cases epiweek30	1.77	6.03	0.00	192	2670

Notes: Temperature are degrees celsius; all measures of rainfall are accumulated rainfall (units: 10mm), cholera cases are total numbers. 88 of 90 wards where affected throughout the outbreak.

Table 3.3: Impact of Weekly Precipitation on Cholera Incidence

	Cholera cases (log)		
	(1)	(2)	(3)
Precipitation	0.0198***	0.0208***	0.0344***
	(0.0072)	(0.0073)	(0.0077)
N	6930	6930	6930
R^2	0.4491	0.4502	0.5254
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table 3.4: Impact of Weekly Precipitation on Cholera Incidence (IV Estimates)

	Cholera cases (log)		
	(1)	(2)	(3)
Precipitation	0.0689**	0.0779**	0.0451
	(0.0341)	(0.0356)	(0.0476)
N	6930	6930	6930
First Stage F-test	36.812	35.354	24.202
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016; they are weighted by the population of the ward (census 2012). Nasa GPM v3 precipitation measurement is instrumented with NASA TRMM measurement. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table 3.5: Impact of Weekly Quartiles of Precipitation on Cholera Incidence

	Cholera cases (log)		
	(1)	(2)	(3)
Quartile 1	0.0010	0.0040	0.0049
	(0.0197)	(0.0193)	(0.0177)
Quartile 3	-0.0360	-0.0286	-0.0536*
	(0.0324)	(0.0329)	(0.0317)
Quartile 4	0.1867***	0.2032***	0.1525**
	(0.0594)	(0.0610)	(0.0611)
N	6930	6930	6930
R^2	0.4510	0.4522	0.5254
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. The quartiles of the rainfall distribution are defined as follows: Q1 (0mm), Q2(0-0.29mm), Q3(0.29-2.69mm), Q4(2.69-40.86mm). *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table 3.6: Impact of Flooding on Cholera Incidence

	Cholera cases (log)		
	(1)	(2)	(3)
Panel A:			
Precipitation	0.0192***	0.0204***	0.0340***
	(0.0071)	(0.0073)	(0.0077)
Precipitation × % Flood-prone area	0.0091*	0.0076	0.0043
	(0.0048)	(0.0046)	(0.0047)
Panel B:			
Flooded (precipitation ≥75th p)	0.2189***	0.2280***	0.2029***
	(0.0485)	(0.0487)	(0.0498)
Panel C:			
Flooded (precipitation ≥75th p)	0.2007***	0.2103***	0.1970***
	(0.0470)	(0.0472)	(0.0517)
Flooded × % Flood-prone area	0.2262**	0.2176**	0.2076**
	(0.1019)	(0.1007)	(0.1028)
N	6930	6930	6930
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. All panels are independent regressions. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. Flooded is a dummy variable for weekly precipitation falling above the 75th percentile of the total rainfall distribution. Flood-prone area is the total area of the ward that is prone to flooding. All regressions control for weekly ward air temperature;they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table 3.7: Impact of Weekly Precipitation on Cholera Incidence: Infrastructure & Ward Characteristics

	Cholera cases (log)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Precipitation	0.0364***	0.0355***	0.0439***	0.0354***	0.0668**	0.0023	0.0525***	0.0518***	0.0696**	0.0058
	(0.0087)	(0.0080)	(0.0116)	(0.0079)	(0.0296)	(0.0135)	(0.0179)	(0.0126)	(0.0314)	(0.0244)
Precipitation × Pop. density	0.0002**							0.0000	-0.0002	-0.0007
	(0.0001)							(0.0002)	(0.0003)	(0.0004)
Precipitation × Roads density		0.0017*						0.0007	0.0019	-0.0054
		(0.0009)						(0.0016)	(0.0020)	(0.0066)
Precipitation × Footways density			0.0020*					0.0023	0.0033	0.0054
			(0.0011)					(0.0021)	(0.0028)	(0.0054)
Precipitation × # Water wells				0.0009				-0.0002	-0.0000	-0.0014
				(0.0007)				(0.0012)	(0.0013)	(0.0023)
Precipitation × Drains density					0.0018				0.0002	0.0039
					(0.0016)				(0.0021)	(0.0032)
Precipitation × % Informal housing						0.0168*				0.0264**
						(0.0083)				(0.0092)
Precipitation × % Formal housing							0.0032			
							(0.0038)			
N	6930	6776	4851	5852	3465	1771	2618	4004	2926	1463
R^2	0.5229	0.5335	0.5703	0.5370	0.6182	0.6606	0.5383	0.5860	0.6398	0.6768
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE										
Municipality × week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Separate regressions in columns (1-7). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Population density is the number of inhabitants per square km (census 2012), Roads density, footway density, and drains density are the meters of roads, footways and drains per km (OSM), % informal and formal houses in the ward are obtained from surveyed plots (not all plots are surveyed). *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table 3.8: Impact of Neighbours' Weekly Precipitation on Cholera Incidence: Spatial Spillovers

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation	0.0183**	0.0192***	0.0196***	0.0204***	0.0316***	0.0320***
	(0.0071)	(0.0070)	(0.0073)	(0.0072)	(0.0077)	(0.0076)
Neighbours precipitation	0.0004		0.0004		0.0009	
	(0.0007)		(0.0007)		(0.0008)	
Uphill neighbours precipitation		-0.0007		-0.0007		0.0001
		(0.0010)		(0.0009)		(0.0009)
Downhill neighbours precipitation		0.0011		0.0010		0.0013*
		(0.0007)		(0.0007)		(0.0007)
N	6930	6930	6930	6930	6930	6930
R^2	0.4491	0.4494	0.4502	0.4505	0.5255	0.5256
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes		
Municipal time trend			Yes	Yes		
Municipality × week FE					Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Uphill and downhill measures are for neighbouring wards at a higher or lower elevation than the given ward. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table 3.9: Dynamic Effects: Lags of Weekly Precipitation on Cholera Incidence

	Cholera cases (log)				
	(1)	(2)	(3)	(4)	(5)
Precipitation	0.0337***	0.0313***	0.0300***	0.0305***	0.0298**
	(0.0075)	(0.0071)	(0.0073)	(0.0073)	(0.0072)
Precipitation $_{(w-1)}$	0.0249***	0.0243***	0.0214***	0.0200**	0.0205**
	(0.0086)	(0.0085)	(0.0081)	(0.0082)	(0.0083)
Precipitation $_{(w-2)}$		0.0222***	0.0215***	0.0195***	0.0185**
		(0.0071)	(0.0071)	(0.0068)	(0.0070)
Precipitation $_{(w-3)}$			0.0256***	0.0252***	0.0240***
			(0.0077)	(0.0076)	(0.0073)
Precipitation $_{(w-4)}$				0.0182**	0.0180**
				(0.0076)	(0.0075)
Precipitation $_{(w-5)}$					0.0113*
					(0.0061)
Cumulative (6 weeks)					0.1221***
					(0.0288)
N	6840	6750	6660	6570	6480
R^2	0.5263	0.5267	0.5274	0.5275	0.5270
Ward FE	Yes	Yes	Yes	Yes	Yes
Municipality × week FE	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Precipitation $_{w-n}$ are the lags of weekly precipitation up to n weeks. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table 3.10: Dynamic Effects: Lags of Quartiles of Weekly Precipitation on Cholera Incidence

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Q1	0.0013	-0.0020	0.0039	0.0005	0.0061	0.0058
	(0.0194)	(0.0190)	(0.0190)	(0.0187)	(0.0175)	(0.0171)
Q3	-0.0335	-0.0372	-0.0262	-0.0301	-0.0519*	-0.0551*
	(0.0309)	(0.0300)	(0.0315)	(0.0307)	(0.0300)	(0.0296)
Q4	0.1825***	0.1753***	0.1982***	0.1908***	0.1504**	0.1442**
	(0.0559)	(0.0538)	(0.0578)	(0.0556)	(0.0574)	(0.0554)
$Q1_{w-1}$	-0.0034	-0.0066	0.0004	-0.0030	-0.0049	-0.0071
	(0.0218)	(0.0222)	(0.0220)	(0.0224)	(0.0198)	(0.0202)
$Q3_{w-1}$	-0.0198	-0.0215	-0.0137	-0.0154	-0.0227	-0.0271
	(0.0352)	(0.0346)	(0.0350)	(0.0344)	(0.0314)	(0.0302)
$Q4_{w-1}$	0.1212**	0.1139**	0.1355**	0.1277**	0.1368**	0.1256**
	(0.0590)	(0.0565)	(0.0592)	(0.0569)	(0.0605)	(0.0566)
$Q1_{w-2}$		0.0312		0.0347*		0.0158
		(0.0192)		(0.0192)		(0.0206)
$Q3_{w-2}$		0.0320		0.0372		0.0495
		(0.0356)		(0.0357)		(0.0352)
$Q4_{w-2}$		0.0903*		0.1029*		0.1052*
		(0.0543)		(0.0549)		(0.0568)
Q1 (Cumulative 3 weeks)						0.0145
						(0.350)
Q3 (Cumulative 3 weeks)						-0.0327
						(-0.490)
Q4 (Cumulative 3 weeks)						0.3750***
						(0.1348)
N	6930	6930	6930	6930	6930	6930
R^2	0.4522	0.4526	0.4535	0.4540	0.5267	0.5271
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes		
Municipal time trend			Yes	Yes		
Municipality × week FE					Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. The quartiles are defined as in Table 4. Quartile_{w-n} are the lags of the quartiles of weekly precipitation up to n weeks. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Chapter 4

Who Really Benefits from Export Processing Zones? Evidence from Nicaraguan Municipalities[1]

4.1 Introduction

The use of Export Processing Zones (EPZs)[2] as a policy for trade and economic development has exponentially grown in the past decades, particularly in developing countries. In 1986, the International Labour Organization (ILO) estimated that there were 176 zones in 47 countries. By 2008, the number reached more than 3,000 zones in 135 countries, accounting for over 40 million direct jobs and over US$200 billion in global exports (Farole & Akinci 2011). Despite the importance of the phenomenon, there is surprisingly little empirical evidence for evaluating these programs in the context of developing countries[3]. Further, the focus has remained on aggregate and traditional outcomes (FDI, exports, firm dynamics), and little attention has been paid to understanding the welfare and distributional effects resulting from the establishment of EPZs.

[1] I thank Vernon Henderson, Pascal Jaupart, three anonymous referees, and Rafael Lalive for their helpful comments to improve the work. I further thanks participants at LSE seminars, at the 4th SOLE/EALE World Meetings, and at the 2015 RIDGE Workshop on Trade and Firm Dynamics for useful comments. All errors remain my own. This work was supported by the Economic and Social Research Council.

[2] When discussing EPZs, a variety of terminologies, such as Industrial Free Zones, Free Trade Zones, Special Economic Zones and Maquiladoras are used interchangeably in the literature. Although each has their own particularity, we will consider the broad definition of them being "demarcated geographic areas contained within a country's national boundaries where the rules of business are different from those that prevail in the national territory. These differential rules principally deal with investment conditions, international trade and customs, taxation, and the regulatory environment; whereby the zone is given a business environment that is intended to be more liberal from a policy perspective and more effective from an administrative perspective than that of the national territory" (Farole & Akinci 2011:23).

[3]Aside notable exceptions, the literature is limited to case studies and macro-economic reviews. See Farole and Akinci 2011; Engman & Onodera 2007; Aggarwal 2006 and Glick & Roubaud 2006, for recent examples.

To fill this gap this study takes advantage of the gradual establishment of EPZs in Nicaraguan municipalities during the period 1993-2009 to assess the average effect and the distributional pattern of this spatially-bound policy within the host municipalities. Nicaragua provides an excellent setting to study this phenomenon. The number of firms operating under the regime has increased markedly since its inception in the early 1990s, across more than 20 of the 153 municipalities of the country. By 2010, EPZs accounted for 50% of total exports, and nearly 90% of manufacturing exports. The government considers that in the same year, EPZs jobs represented 25% of total formal work across the country, with on average firms directly employing around 7.5% of the total labour force of the municipalities where they locate[4]. Yet, no empirical assessment exists.

While having a wider reach than traditional place-based policies (i.e. diversify exports, increase FDI), EPZ programs are prone to have a significant influence on the local economies (Wang 2013) as they operate with incentives to hire and create economic activity in or near the areas where they locate fostering agglomeration economies (Combes et al. 2011). In this sense, local welfare effects might differ substantially from those at the aggregate level, particularly in cases with labour market frictions. Labour mobility and land-price responses may be such that the jobs created go to non-poor residents and that the gains from land prices benefit higher-income segments (Neumark & Simpson 2014). As stressed by Kline & Moretti (2014) and shown by Reynolds & Rohlin (2015) for the case of US federal empowerment zones, positive average effects of spatially-tied policies can mask significant disparities in terms of the actual beneficiaries across the income distribution in treated areas. There are heterogeneous effects according to whether individuals are homeowners or renters, or more generally by skill and initial income levels, that are not necessarily captured by looking at the average effect on local wages (Neumark & Simpson 2014). In this sense, disentangling in what way the establishment of an EPZ profits the different segments of the income distribution within concerned areas helps to shed light on the mechanisms of the policy and ultimate local beneficiaries. The question is of great policy relevance in developing economies with already large levels of overall income disparities.

For the analysis here, I exploit a repeated cross-sectional dataset based on official household surveys and censuses, and construct a unique municipal-level panel that allows for the examination of different moments of the expenditure distribution before

[4] Data obtained from the Comisión Nacional de Zonas Francas (CNZF) and Nicaragua's Central Bank (BCN).

and after the policy implementation. By focusing on the aggregate outcomes at the municipal level and on household per capita expenditures, the study captures general equilibrium effects of EPZ establishment within a locality[5]. The identification strategy is straightforward. I use both the time and cross-section variation of zones' establishment across municipalities to estimate their average effect on the levels of real expenditure per capita of working age individuals, and across all the deciles of the expenditure distribution in treated municipalities. In the main approach, I use a difference-in-differences (DID) strategy extended to quantiles (Athey & Imbens 2006). An important feature of the overall empirical strategy needs to be emphasized. The approach does not aim at measuring the impact of the EPZ policy on inequality across Nicaragua as a whole. Rather, it looks at the question of knowing whether within municipalities exposed to the establishment of EPZs, certain segments of the distribution capture more or less of the resulting gains or losses.

This analysis builds on two complementary bodies of research. Mostly, it borrows from the large literature that has looked at evaluating place-based programs in the US and Europe (Busso et al. 2013; Neumark & Kolko 2010), and adds to the nascent counterpart studying the impact of EPZ policies at the subnational level in developing countries. Findings in these papers are revealing of the extent to which local economies are influenced by the establishment of an exporting zone. Using employment data and census statistics on educational outcomes at the municipal-level for Mexico, Atkin (2012) finds an increase in school dropouts in concerned municipalities following the arrival of new better-paying export jobs. Similar in strategy to this paper, Wang (2013) shows that Chinese Special Economic Zones (SEZ) benefit local economies through higher levels and growth rates of per capita FDI and total factor productivity (TFP), as well as higher average wages that compensate any increase in the local cost of living. Additionally, this paper adds empirical evidence to the literature analysing the effects of trade policies on local labour markets in developing countries. This area of research has emphasized the importance of understanding the impact of trade policies at the local level, particularly in cases of limited factor mobility (see Goldberg & Pavcnik 2007 for an extensive review).

The validity of my findings relies on the assumption that the different empirical methods successfully account for the underlying differences in municipalities'

[5] In the absence of disaggregated property prices and firm-level data, using expenditure data remains the best strategy to capture the net effect of EPZ establishment. A similar strategy was used by Topalova (2010) to measure the local effect of trade liberalization within Indian districts.

characteristics that are likely to explain the non-random choice of EPZs location across time and space. In the absence of a credible instrument, I take care of unobservable confounders by allowing for time and municipalities fixed-effects across specifications. I also include province and region-time dummies to control in a flexible manner for the fact that EPZs concentrate in the western part of the country. Further, I use information on accessibility and socio-economic indicators to build alternative sets of control municipalities that are likely to be more meaningful comparable groups under the DID setting. My preferred control group is balanced on covariates and pre-treatment trends of outcomes, and the assumption of parallel trends of main outcomes holds.

Further threats to validity concern possible spillover dynamics and the relocation of individuals across treated and control areas. I address these concerns in several ways. First, I test for the existence of spillovers on the outcomes variables of neighbouring areas using varying degrees of distance. I find no evidence supporting large firm relocations or commuting flows. This is consistent with the literature registering small backward linkages generated by EPZ policies, and the low commuting of the Nicaraguan labour force (Jansen et al. 2010). The potential threat from labour mobility is harder to address. There are two dimensions to consider: in and out-migration. Both may have compositional effects that may alter the location decisions of firms (related on skills and local wages effects) as well as outcomes. Reassuringly, I find no evidence of large population readjustments across municipalities when measuring the effect of the policy on the likelihood of migrating for the working-age sample. Results are also unchanged when adding baseline population weights. Although, I find evidence of compositional changes towards an increase in the share of the high-to-low skill ratios in treated municipalities, the size is too small to challenge the identification validity, and if anything, contributes to elucidating the mechanisms of the policy. Finally, the results are also robust when excluding the capital city and municipality of Managua.

The analysis reveals interesting conclusions concerning average and distributional dynamics of EPZ policies at the local level. First, I confirm that there is a positive average treatment effect on the levels of real per capita expenditure in treated municipalities. The magnitude is small but non- negligible, with an average 10 to 12% increase in treated areas for the entire period. Second, I find that the mean effect hides significant disparities across the distribution. Results offer suggestive evidence that those at the upper-tail of the distribution are the main beneficiaries of the policy over the period covered. This effect is incremental with the more years of EPZ operation.

The impact is heterogeneous in time for the remaining segments, significant at middle-range deciles after eight years of an EPZ establishment. These results are consistent with the existence of a skill-premium in the exporting sector that may be altering the cost of living (i.e. land price responses) in profit of the higher-skill/higher-income segments. Indeed, I find that the average positive gains are concentrated in the high-skill working-age group.

The remainder of the paper is organized as follows. Section 4.2 describes the EPZ program in Nicaragua. Section 4.3 introduces and examines the data, and section 4.4 discusses the empirical strategy. Results are discussed in section 4.5, with extensions and robustness checks shown in section 4.6. The final section concludes.

4.2 Nicaragua's Export Processing Zone Program

4.2.1 Ley de Zona Franca: Legal procedures & tax exemptions

Resembling analogous initiatives, Nicaragua's EPZ regime ('Zonas Francas') was put in place with the aim of generating employment, attracting FDI, increasing non-traditional exports, acquiring new technologies, and expanding international trade. The program is relatively young compared to other Central American countries as an earlier initiative was abandoned for a decade during the civil conflict (1979-1990) and was only reintroduced following the shift to democracy and the market economy in 1990. In 1991, Decree N.46-91 formally established the existing regime, with the appropriate regulation passed in the following year (Decree N.31-92). Together, these two decrees created the legal framework concerning the organization of EPZs, the fiscal incentives, the rights and obligations of operators and users, as well as the regulatory body in the form of the National Commission of Free Trade Zones (CNZF). The law was amended over the years, with its most notable reform aimed at simplifying procedures taking place in 2005 (Decree N.50-2005).

Under the Nicaraguan legislation, EPZ status is granted to foreign or national private firms oriented towards the export of goods and services that meet a certain array of criteria and follow the official procedures with the CNZF. There have been many efforts to expedite the legal process of application in the recent years (i.e. establishment of a one-stop shop after 2005). Firms applying to be recognized under the EPZ regime need to provide detailed information of their project (which includes financial plans, market analysis, exporting markets, construction or building agreements, estimates of job creation, among others). They also need to comply with environmental, building

and health regulations, and the project must be approved by the mayoral office in the municipality where they intend to locate. There is no explicit rule regarding the age of the firm (new or pre-existing).

Following standard practice, the program is based on attracting and generating business by offering particular tax incentives to qualifying firms. These include the full exemption from payment of (1) Income Tax during the first ten years of operation (and 60% from the eleventh year onwards); (2) all taxes and custom duties of machinery, equipment and intermediate goods, as well as of transportation and support services for the zones (including for instance equipment needed for installation of health care or child care); (3) indirect taxes, taxes on selective sales or consumption; (4) export taxes on processed products made within the regime; (5) municipal taxes on property sales, including the tax on capital gains if any; and (6) municipal taxes. The regime also includes the unrestricted repatriation of capital. In exchange, firms have to pay an annual fee to the CNZF, determined according to the industrial sector and square meters of occupied area.

Under Nicaraguan law, EPZs are allowed to sell intermediate goods and inputs to other exporting firms in the country, provided they pay custom duties. However, EPZs are strictly forbidden from selling final goods in the local market. The legislation also provides conditions under which EPZ developers may benefit from tax exemptions, as well as the possibility for them to subcontract to local Nicaraguan firms which are then exempted from VAT payment. The restriction to sell in the local market and the barriers to sell intermediate inputs means that in terms of production the larger effect of EPZs is contained to the exporting sector, without hampering local producers. The possibility to outsource to local providers creates some incentives to generate backward linkages, and any firm relocation or diversion is likely to happen through this indirect mechanism. I expect these to be low. Backward linkages with local firms have been found to be very low in Latin America (Jenkins et al. 1998). The share of domestic expenditures as a share of total value of EPZ exports in Nicaragua has ranged between 25 to 30%, with wages constituting the bulk of the shares (ECLAC 2012). I formally address this issue in section 4.6.1.

Nicaraguan labour laws regulate work in EPZs and minimal wages have been in place since 1999. The country has the lowest manufacturing wages in the region, which is a source of its comparative advantage. Still, minimum wages in EPZs have traditionally been higher than the overall manufacturing and agricultural sectors throughout the period (20 to 70% higher on average per year).

4.2.2 Definition & characterization of EPZs

The law distinguishes three different categories: administrators (industrial park operators), users (situated within industrial parks) and 'single-factory' EPZs. The distinction is relatively artificial. Operators and users of industrial parks operate in the same area, with the former only providing a service to the using firms and most of time owning some firms within the park. The distinction serves administrative purposes for the payment of different types of fees to the CNZF. Further, the single-factory category is mostly used by large multinational corporations with a wide range of functions. Examples of these are New Holland Apparel, a large US manufacturing firm that produces sportswear for Nike, or Yazaki, a Korean firm in the area of light manufacturing considered to be the single largest employer in the country. Each of these firms employ about 2 to 8 thousand workers, respectively (2013). In this sense, despite the different legal definitions, industrial parks made of smaller firms are thus essentially equivalent in terms of employment and export volume to stand-along EPZ firms. As such, I do not distinguish between the different categories[6]. For the empirical analysis, I define EPZs as stand-alone firms or an industrial park hosting one or more firms, and reported by the CNZF at year end[7]. By this definition, the total number of EPZs in my sample is 64 by 2009 [8], located in 20 different municipalities across 10 provinces[9]. Figure 4.1 depicts the pattern of localization across municipalities according to the sequence of establishment.

In terms of numbers of firms and parks, a larger share is concentrated in the capital city Managua (37%), the adjacent provinces of Masaya and Carazo (20%), and the north-eastern provinces of Estelí and Chinandega (18%). These provinces also represent 55% of the country total population (2012). 76% of EPZs are foreign, the majority of which (40%) are from the US, followed by South Korea (22%). According to official statistics for 2010 (CNZF), American EPZs employed 35% of the total workforce in the sector, followed by South Korean which employed about 30%. Nicaraguan EPZs are attributed only 5% of the workforce in the sector, which implies little firm relocation from previously operating firms. There is no disaggregated statistics

[6] Because I do not have access to firm-level data (workers, exports, sales or value-added), I cannot corroborate or disentangle EPZ categories with a continuous treatment variable. Any attempt to categorize EPZs in terms of intensity would likely be misspecified.

[7] I compiled information on EPZs location by the end of each year, origin and sector of operation from the CNZF annual yearbooks. When possible, I cross-checked the information with the EPZ website profile.

[8] The discrepancy with the official figure of +150 stems from this definition, as I count industrial parks as one irrespective of the quantity of firms. Since the launch of the program, by this definition, the number of EPZs increased from 1 in 1993 to 64 (+150) in 2009.

[9] There are 153 municipalities, 15 provinces and 2 autonomous regions in Nicaragua.

on the number of employees within each EPZs. According to the World Bank Enterprise Surveys, only 12% of domestic firms exported in 2010. Near to 90% of these EPZs operate in the manufacturing sector, with a predominance of firms producing textile, apparel and light electronics, followed by cigar production and other agri-businesses. 25 firms are categorized as single factories, the majority of which are also in the apparel and light manufacturing sectors. All of the agri-business EPZs are stand-alone firms. Both the average years of stand-alone firms and industrial parks is close to 5 years, with standard deviations of 3.5 and 3.4 years, respectively. Attrition is relatively low, with about 5% of the EPZs having closed during the period considered, but some measurement error is possible due to the difficulty in tracking earlier EPZs.

In theory, eligible firms can locate anywhere in the country. The regime does not institute a particular area of the territory as being a free-trade zone, but rather grants the status to the firm, which needs to negotiate directly with the municipality (irrespective of locating as a stand-alone firm or within an industrial park, a letter of acceptance from the mayoral office of the relevant municipality is a pre-condition, CZNF 2015). In practice, not all municipalities are prone to receiving an exporting zone, and the export-oriented nature of the firms has made the western area of the country the natural host.

This dimension underlines the 'de-facto' place-based nature of the EPZ regime in Nicaragua. Place-based policies are defined as policies that particularly target geographic areas for some form of special treatment including special regulations or tax exemptions (Kline & Moretti 2014), in order to create incentives to hire or generate local economic activity. While in Nicaragua privileges are granted to the firm or industrial park, the location decision is subject to the municipal approval. Once established, EPZs are spatially-fixed, and their propensity to hire or outsource in the area where they operate implies that the local economy will be directly impacted through labour, capital, land and price channels. This is further emphasized in the case of industrial parks that remain spatially-bound but can increase the number of firms located within. In practice, the establishment of EPZs in a given municipality has the same local effects than a traditional place-based policy (Wang 2013, Atkin 2012). Limitations to factor mobility may emphasize this dimension. The trade literature has stressed the importance of analysing the effect of trade policies in local and regional labour markets (Autor, Dorn & Hanson 2013, Kovak 2013, Topalova 2010), particularly due to the lack of perfect factor mobility across geographical areas in both developed and developing countries.

For the empirical study, I set a general EPZ dummy variable, EPZ_{mt} , equal to one if an industrial park or stand-alone firm was authorized within the municipality, and

zero otherwise. I discuss the relevance of using this level of analysis further in the next section.

4.3 Data

4.3.1 Datasets

The data for this analysis were drawn from several sources. First, I compiled a repeated cross-sectional dataset for the period 1993-2009 based on five LSMS Household Surveys[10] (Encuesta Nacional de Hogares para la Medición del Nivel de Vida, EMNV) collected by the Nicaraguan Institute of Statistics (Instituto Nacional de Información de Desarrollo, INIDE), and representative at the national, regional, urban, and rural levels. These contain detailed information on living standards and various other households and individual socio-demographic characteristics. I use this information to construct a panel at the municipal-level and calculate expenditure deciles at this level of aggregation. These measures include a correction for the sample bias using sampling weights. Municipalities with less than 30 households per period were excluded to reduce measurement error. Overall, the sample comprises observations for an unbalanced panel of between 82 to 126 municipalities per year – of which 61 are observed for the full period, and 107 are observed four times (Table 4.1)[11]. The data on EPZs was compiled from various CNZF annual yearbooks, which only contain information on location and sector of operation of each zone at year end. Data on accessibility measures (road density, ports, airports and railways) was compiled from reports of the Ministry of Infrastructure and Transport (MTI)[12] and World Bank logistics' assessments (2004-2011).

4.3.2 Main definitions

This analysis follows common practice in referring to the measure of living standards as income, which should be interpreted broadly to encompass all the characteristics associated with geographical location including climate or local public good provision, in order to assume that individuals with the same level of income at different locations are equally well-off (Shorrocks & Wan 2004). Here, following best

[10] The Living Standards Measurement Surveys were developed by the World Bank to improve the type and quality of household data collected by statistical offices in developing countries. The years of each EMNV are: 1993, 1998, 2001, 2005, and 2009.

[11] 20 municipalities are omitted because of N<30 rule across the years. As a robustness check, I run the different specifications with a balanced panel, omitting observations observed less than 4 times. Conclusions remain unchanged.

[12] Ministerio de Transporte e Infraestructura (MTI) Red Vial de Nicaragua.

practice income is proxied by consumption expenditure (Goldberg & Pavcnik 2007). The choice for using consumption expenditure as a measure of income has been widely discussed in the literature. It is arguably the most appropriate variable for capturing lifetime wellbeing (Deaton 1997) as it better captures intertemporal shifts of resources and incorporates changes in purchasing power. Expenditure data are of better quality in LSMS Household Surveys (Deaton 2005; Banerjee & Duflo 2007), given that reporting problems are less pronounced and consumption is less affected by the redesigning of surveys, which hampers comparability across years. The large period considered in this analysis, and the existence of a large informal sector (73% of all jobs in 2010) in Nicaragua also motivate the use of expenditure levels instead of labour income. Further, wages are under-reported in the household surveys, particularly in 1993 and 2001 when only between 10 to 20% of working age members report a value. There is no information on the location of employment of workers in the household surveys, and I have no access to data on housing prices within municipalities. These data limitations further justify the use of expenditure per capita as the best approximation of a welfare measure that encompasses multiple dimension across time and unit of analysis.

I use constant annual per capita expenditure (in Córdobas of 1999, C$)[13], defined as the household annual net expenditure divided by the number of persons in the household. All households surveyed and all of their members are included (though consumption is adjusted depending on members being children or adults). Missing values and zero incomes have been excluded. Additionally, expenditure has been adjusted for difference in prices across provinces every year using local price indices in the households' surveys. Adjusting for spatial price differences will not alter inequality within regions, and are an important factor to take into account given the strong correlation of prices with welfare levels (Shorrocks & Wan 2005).

As mentioned, to analyse the moments of the expenditure distribution I follow Milanovic (2005, 2008) and partition the real per capita expenditure distribution by deciles at the municipal level using sampling weights to correct for sample bias. This method allows me to focus on the pattern of change across all the deciles shares of the distribution, and has the advantage of decomposing the average effect. This method is used for other household characteristics. There are limitations to using this approach

[13] I use the Managua CPI from the Central Bank of Nicaragua (CBN) as the overall country CPI is not available before 2000. The CPI is a Laspeyres-type (1999=100) index compiled on the basis of the Household Income and Expenditure Survey of 1998-1999. 15 departmental capitals and two autonomous regions were used to select the products to be included in the various baskets and the corresponding weights. These include durables and non-durables as well as residential rents. A summary methodology is available in the CBN Boletín Económico, Volume II, Number 2, April - June 2000 (IMF).

for small areas using household survey data. However, I take several measures to control for the quality of the values, such as excluding municipalities with too few observations, and comparing the municipal averages with census data available for 1993 and 2005. Results approximate well. Tables E1 and E2 in Appendix E contain descriptive statistics at the decile level according to treatment and control groups for the beginning and end of the period studied. These are reassuring in that all deciles display a similar evolution across the period, with the ratios between treated and control groups remaining constant in time. As expected individuals at higher deciles display higher levels of educational achievement, access to electricity and a smaller proportion is employed in the informal sector. Generally, the differences within deciles between treatment and control groups are small. Particularly, household size, average age and individuals at working age are consistently comparable and evolve at a similar ratio.

Because of the characteristics discussed in section 4.2, it seems logical to consider municipalities as the unit of analysis. They are the best approximation of the actual functional area at which firms and zones operate, and a fair representation of the local labour market. Labour mobility in Nicaragua is limited, not only because of the low-skill level of the labour force, but because commuting distances are found to be an important deterrent to participate in the labour market (Jansen et al. 2007). Excluding Managua, municipalities had a median population of 20 thousand in 2012. Additionally, they are the ultimate discrete policy unit for which welfare data is available and they remain the reference framework for delivering public services and distributing government funds (Lall & Chakravorty 2005).

4.3.3 Descriptive statistics

This section discusses key summary statistics for the main outcomes and control variables used in the analysis. Table 4.2 displays period averages for selected outcomes. They include both municipal-level variables, and averages for the working age population of the pooled cross-sectional dataset. Table 4.3 presents descriptive statistics of municipality characteristics according to treatment status and sequence of EPZ establishment. Values are period averages. The table examines whether or not early zones exhibit different characteristics relative to later zones, and both compared to municipalities never having authorized the establishment of an exporting zone. For this, I group the municipalities based on the establishment waves: group 1 [<1999] is composed of four municipalities that were exposed to EPZs by 1999, it includes Managua, which is always treated in the sample since the first industrial park was opened in 1994; group 2 [2000-2001] is composed of five municipalities that granted

authorization to new EPZs between 1999 and 2001; group 3 [2002-2005] is composed of seven municipalities that authorized establishment during this period; and group 4 [2006-2009] is composed of the final waves of municipalities receiving an exporting zone. Because data was of poor quality (too few observations) I had to exclude two municipalities treated in that last wave.

Overall, municipalities receiving EPZs display significantly better accessibility measures (i.e. density of paved roads, access to airports, ports or trade posts, and proximity to large water bodies), and closer proximity to the municipality of the capital city. Group 4 is a particular outlier, probably due to progress in infrastructure across the period. This trend is to be expected as the choice of location of exporting firms is unlikely to be exogenous and should be highly correlated with accessibility and infrastructure. Similarly, the sectoral compositions also diverge, with non-treated municipalities showing much higher levels of agricultural specialization and a predominance of rural areas. Nicaragua remains overall a rural country, with less than 60% of its population living in urban areas in 2010 (World Bank). The skill composition displays a better distribution across municipalities, differing only in the proportion of illiterate adults which is much lower in EPZs areas (20 vs 30%). On the other hand, all groups show similar levels of access to electricity, unemployment, share of working age populations, and migrants. In this sense, it would seem as if the geographical distribution of economic activities across the country is not necessarily correlated with different labour outcomes. This largely reflects the importance of agriculture for the Nicaraguan economy. Even in 2011, the sector was the main employer (32% of total employment) and represented 19% of total value added (above manufacturing).

Figures E1[14] provide a more detailed image of the distribution of the sector of economic activity across the deciles of per capita expenditure by years and treatment groups. Both EPZ and no-EPZ municipalities display similar patterns for services and agriculture, with a higher concentration at the upper-end and lower-end of the distribution, respectively. Discrepancies are noticeable however in manufacturing, construction, and transport. EPZ municipalities show a higher concentration in manufacturing at the upper-end of the distribution, despite a general shift towards the median during the period. The opposite is visible in the construction sector. Given the size of the change, it could well reflect the modernization of the sector rather than any trade-related effect.

[14] Figures and tables E# are contained in Appendix E.

4.4 Empirical Strategy

The main empirical analysis relies on the Difference-in-Differences (DID) framework. For the identification strategy, I rely on the variation provided by the timing of zone creation across the sample of municipalities.

4.4.1 Basic specifications

I follow Angrist & Pischke (2009) and define equation (4.1) and (4.2) as the extension of the simple DID model for multiple periods and groups, in order to estimate the Average Treatment Effect on the Treated (ATT) using both the pooled individual level dataset and municipal averages. As defined in section 4.2.2, I denote EPZ as a dummy variable that switches to one when a municipality *m* gets treated in year *t*. The basic specifications take the following forms:

$$Y_{imt} = \gamma_{0m} + \gamma_{1m}.t + \mu_t + \delta EPZ_{mt} + X'_{imt}\beta + \varepsilon_{imt} \quad (4.1)$$

$$Y_{mt} = \gamma_m + \mu_t + \varphi EPZ_{mt} + Z'_{mt}\theta + \omega_{imt} \quad (4.2)$$

Where *m* and *t* index for municipality (m=1... 126) and year (t=1993, 1998, 2001, 2005, and 2009) respectively, and *i* indexes for working-age individuals. Y_{mt} is the outcome variable of interest in municipality *m* at year *t*; Y_{imt} is the outcome variable of interest for individual *i* in municipality *m* at year *t*; γ_m are municipality-level fixed-effects and μ_t are time effects. The fixed-effects capture the permanent differences in the municipalities observed and unobserved characteristics that may influence the location of EPZs as well as any time varying shocks. X_{imt} and Z_{mt} are vectors for time-varying individual and household characteristics. In equation (4.1) they include measures for household size, age-squared, and dummies for gender, economic sector of employment, migrant status, education level, access to electricity and urban location. Equation (4.2) contains equivalent measures aggregated at the municipal level (i.e. shares of education levels, households with access to electricity, urban population, and migrants). $\omega_{imt}(\varepsilon_{imt})$ is the error term. Following Bertrand et al. (2004) I use robust standard errors clustered at the municipal level to prevent misleading inference due to serial correlation in the error term across years within a municipality. Results are robust to the clustering of standard errors at the province level to account for possible serial correlation at this level of aggregation.

The parameters of interest are δ and φ which measure the average effect of the establishment of an EPZ on the outcomes Y_{imt} and Y_{mt} (i.e. measures of real expenditure per capita). The effect is identified by comparing municipal outcomes among treated and non-treated groups before and after EPZ establishment. To control even more richly for differences across municipalities, I add a municipality-specific time trend in the form of interactions between a municipality dummy and a linear time trend, $\gamma_{1m}.t$ in equation (4.1). This specification allows for municipalities unobservable characteristics to follow different trends in time, which is particularly relevant in the case of a gradual treatment. The larger dataset allows introducing the trends without compromising degrees of freedom. I use province-time fixed effects and region-time fixed effects in alternative specification at both individual and municipal level of analysis, to account for sorting across Nicaragua and the concentration of EPZs in the western area of the country.

I measure the effect across the expenditure distribution by extending the DID method to each quantile (τ) instead of the mean (Athey & Imbens 2006). The specification takes the following form:

$$Y_{mt|\mathrm{EPZ}_{mt},\rho_{mt}}(\tau) = \gamma_m(\tau) + \mu_t(\tau) + \psi \mathrm{EPZ}_{mt}(\tau) + \mathrm{W'}_{mt}(\tau)\,\alpha + \xi_{(\tau,\rho_{mt})} \tag{4.3}$$

Where notations are the same as previously defined in the simple DID cases but τ now stands for each decile of the expenditure distribution estimated at the municipality level. All notations are as specified above. I cluster standard errors at both municipal and province levels. This method referred to as the QDID allows comparing individuals across both groups and time according to their quantile. Here, the quantiles are *fixed* and the DID estimator compares changes in the levels of expenditure around them, under the identifying assumption that the growth in expenditure from pre-EPZ to post-EPZ groups at each particular quantile would have been the same in both treatment and comparison groups in the absence of treatment. It relies on the assumptions of rank-preservation[15] and homogeneity of treatment effect across individuals (Athey & Imbens 2006).

[15] While rank-preservation is in theory a rather constraining assumption, social mobility in Latin America, and Nicaragua, in particular, is in practice very low. The World Bank (2013) estimates that since the 1990s more than one in five individuals remained chronically poor, and approximately the same proportion remained steadily in the middle class in the subcontinent, with Nicaragua being the second-worse country

4.4.2 Validity of the identification strategy

Given the setting here, there are three main threats to identification to address. As most spatially-bound policies these concern the reverse causality and omitted variable bias related to the non-random choice of EPZs location across time and space (i.e. sorting), and the limitations related to the fact that the experimented created here may not coincide with the administrative boundaries used (i.e. firm spillovers -creation and diversion -, and mobility of individuals). I will empirically evaluate the last two issues in sections 4.5.4 and 4.6.1.

I address the issue of the endogenous location of EPZs in several ways. Regarding unobservable factors, I allow for municipality and year fixed-effects. These allow me to sweep out time-invariant features of the municipality as well as time-specific dynamics. Additionally, I use province-time dummies and region-time dummies that control for omitted variables that change over time within these levels of aggregation. Some specifications also include municipality-level time trends to control for any omitted variable that varies over time within municipalities in an approximately linear fashion.

I further refine the identification strategy with the construction of alternative control groups based on relevant observable municipal characteristics. One important challenge in the evaluation of spatially-bound policies is the selection of appropriate counterfactuals. The ideal control municipalities here are areas economically similar to EPZ municipalities but lacking zones and firms' establishment. There are many reasons why using all never treated municipalities would yield bias estimates (i.e. different sectoral skill and age compositions for instance). An ideal approach would rely on areas that were targeted at some point in time but where a zone finally was not created, using the eligibility criteria to ensure similarity (Greenstone et al. 2010). An alternative, would be to use the temporal difference between treated municipalities to compare them against each other (Busso et al. 2013, Wang 2013). I do not have information on eligible municipalities that eventually did not receive an exporting zone. Further, because of the sequence and relatively small number of treated areas in earlier phases I cannot use the temporal sequence to compare municipalities against each other. Instead, I use municipalities' observable characteristics to define alternative control municipalities. While limitations to this approach remain, tests for the balance of covariates and pre-treatment outcome trends give me confidence on their validity.

---

in the region in terms of the proportion of individuals not displaying any type of upward mobility between 1998 and 2005 (80%). Intergenerational mobility shows similar patterns.

The export-oriented nature of EPZs and its human capital needs imply that both accessibility conditions (i.e. ports, airports, and roads), basic infrastructure and the educational attainments are likely to play an important role in the choice of location of the industrial parks and firms. The first approach is thus to match municipalities based on these and related observable elements, using propensity score matching to complement the DID strategy (Abadie 2005, Imbens & Woolridge 2009). This approach is widely used in the literature, and has generally been shown to produce sensible counterfactual groups. For this, I estimate the propensity score (logit) for a municipality to receive EPZs based on a series of pre-treatment variables for each of the treatment sequences using a k-nearest neighbour approach[16], and use only matched municipalities as the comparison group. As shown in Tables E3 and E4 there is a large overlap in p-scores between treated and matched municipalities, particularly for groups one and four. T-test and pseudo-R2 are both fairly low after matching which suggests that potentially important selection criteria become not significant after matching.

The second approach is to define the control municipalities based on the quality of their road infrastructure and distance to the capital city (Managua). Given the limited development of Nicaraguan ports, airports and railways, road networks remain the main mode of transport for most domestic movements and a large share of imports and exports (World Bank 2011). Good roads are not only likely to be one key determinant for EPZs location choices, but they may also be correlated with higher levels of regional development. I use the meters of paved roads every 1 km2 to account for road quality, and set the threshold at the minimum in treated municipalities (40 meters). At the same time, distance to the capital city is also likely to be a key factor in EPZ location, as firms are drawn towards it to benefit from agglomeration economies and higher levels of human capital. I set the threshold to the maximum distance for an EPZ municipality (<170km)[17]. Despite the fact this counterfactual is constructed in a more arbitrary way, it imposes less restrictions in terms of observables characteristics. Table 4.4 shows that there is significant overlap between this control group (group 2) and the treatment areas. While the propensity-score matched municipalities are more balanced in relation to the urban-rural dimension, the second group seems slightly closer in terms of

[16] Because of the small sample size, I use different values of k from 3 to 5. I note that the k-nearest neighbouring matching with replacement is likely to use some control municipalities multiple times in different periods.

[17] Earlier versions of this paper also included a control group that excluded the two autonomous regions of the Atlantic coast (RAAN and RAAS). These correspond to the poorest areas of the country and also have very particular ethnic and economic characteristic. The inclusion of province-fixed effects and region-time fixed-effects addresses this issue in a parametrical way.

educational attainment. In any case, with both cases, controlling for specific covariates will wipe-out any remaining difference between the two groups.

I choose group 2 as the preferred counterfactual as it imposes less restrictions in terms of sample size, makes fewer assumptions on the municipal characteristics linked to the program, and parallel trends in levels of pre-treatment outcomes are more consistent (Figures 4.2 and E2). I formally test the assumption of parallel trends for the main specification by including pre-treatment trends; these results are supportive of the identification strategy (Table E6)[18]. The common trend assumption that the two groups measured outcomes would have followed the same trends over time in the absence of policy, is essential to support the DID framework. Ultimately, I compare results with both methods and conclusions are unchanged (results with the matched municipalities are on appendix E).

4.5 Basic Results

4.5.1 Average Treatment Effect of EPZ establishment

Table 4.5 presents the results of the basic DID specifications (equation 4.1 & 4.2) using both the working-age population dataset and the municipal-level outcomes. The different columns explore the sensitivity of the results to different area fixed-effects and time trends. Overall, I find that the establishment of EPZs increases the average level of real per capita expenditure in the treated municipalities, with point estimates ranging between C$800-460 depending on specifications. The size is nontrivial but relatively small if we consider the large time span of the analysis. It amounts to about 10 to 12% of the average levels across the period. It is reassuring to see that the magnitude of the effect and its statistical significance remain stable across the specifications, particularly with the addition of municipality-time trends. Further, results are consistent irrespective of the sample used, though they lose precision with the panel dataset as observations are reduced. Using the alternative control group based on propensity-score matching does not alter the results (Table E5).

[18] I run a formal test of parallel trends in the main specifications of equations 4.1 and 4.2 by including pre-treatment trends in the fashion of Greenstone et al. (2010). For this, I augment equations 4.1 and 4.2 with a municipality-specific trend variable defined as the interaction between municipality specific linear-time trends and the treatment dummy variable EPZ as previously defined. Overall, coefficients on pre-treatment trends are not statistically significant supporting the conclusions of the effect not being the result of different pre-existent trends between treated and control municipalities (see Table E6 in Appendix E).

4.5.2 Deciles of the expenditure distribution.

Table 4.6 presents the results of DID estimates across moments of the real per capita expenditure distribution of treated municipalities (equation 4.3). In interpreting these results, it should be kept in mind that I estimate treatment effect on each decile of the expenditure distribution of treated municipalities strictly with respect to its respective decile in the control group. As noted, this implies the assumption of rank-invariance. While this is a strong assumption, social mobility in Latin America and particularly in Nicaragua has been very low since the early 1990s[19]. Caution is still warranted when interpreting these results, and I refrain from making any reference with respect to within-area inequality dynamics.

The most salient feature in Table 4.6 is the containment of the treatment effect at the top-end of the distribution (80th and 90th percentiles), with point estimates significant at 10% levels. The size of is much larger in this case, ranging at about 20 to 25% of the average levels of per capita expenditure of the top deciles for the period, suggesting that on average, the establishment of exporting zones profits the top deciles of the distribution in host municipalities the most. Point estimates are consistent to the inclusion of different time-area fixed-effects, as well as to the clustering of standard errors at the province level and against the alternative control group (Table E7). The coefficient at the 10th percentile is also positive and significant (at 10% level) in one of the specifications. The size is smaller, on average EPZs increase the expenditure levels of the bottom decile by 10% across the period.

The fact that such a strong concentration is visible when decomposing the average treatment effect across the expenditure distribution, provides support for two alternative but complementary theoretical intuitions. The first one, relates to the various models of urban economics that integrate land prices' responses and labour mobility dynamics for estimating welfare at the local level. Kline & Moretti (2014) show that mobility responses may lead the local cost of living to change, which in turn can lead landlords or high-skill incomers, to capture the benefits associated with the firm's arrival. Evidence of this on US empowerment zones is found by Reynolds & Rohlin (2015). The second relates to international trade models that put emphasis on the existence of a high-skill premium in the exporting sector stemming either from productivity gains or higher skill intensity (Goldberg & Pavcnik 2010 for review). Topalova (2010) emphasizes how these dynamics are likely to be larger in the case of

[19] The fact that individual characteristics across each decile remain relatively constant in both groups and with similar relative sizes, support the use of this assumption. See footnote 18.

rigid labour markets. Although information on the skill composition of workers within EPZs is not available, manufacturing firms in Nicaragua do exhibit higher high-to-low skill ratios, with on average only 34% of the workforce in the firms classified as unskilled (Nicaragua Enterprise Survey 2010). The positive coefficient of the first decile may also offer some support for polarization theories (Autor & Dorn 2013), according to which new technologies and the related inflows in skilled labour in local labour markets may have reallocated low-skill labour into lower-earnings services occupations. There is no solid evidence of polarization in developing countries yet, but it is plausible to imagine the changes happening between the formal and informal sectors, particularly in more urbanized areas in countries with small middle-classes.

The conclusions of the basic specifications are somewhat counter-intuitive with the mediatized image of the zones as prime generators of employment for the unskilled labour force. Data limitations prevent me from completely disentangling the mechanisms through which firms operate, however I address possible channels next by looking at the effect across the skill distribution and the effect of EPZ formation on labour mobility.

4.5.3 Heterogeneous time dynamics: an event study

Before introducing any additional extensions, I explore the time dynamics of the establishment of exporting zones. Considering the relatively long period covered, it is important to address the question of how the program impacts expenditure levels across time. The literature has emphasized that the impact of EPZs may take time to materialize. In some cases, it has also been found to diminish with the years, as low labour costs vanish.

To explore the time pattern, I compute a modified version of equation (4.3) adding a set of dummy variables for leads and lags of the year of EPZs authorization (Autor, 2003). Specifically, I add indicator variables for 11, 8 and 3 years before and after EPZ establishment[20]. The dummies are equal to one in only one year each per treated municipality. The treatment dummy EPZ_{mt} equals one only on the first year of EPZ establishment, or year zero. To help visualize the results figure 4.3 displays point estimates at different moments of the distribution with the 95% confidence intervals; estimates are reported in Table 4.7.

[20] Leads and lags of 4 and 8 years also include municipalities with 3 and 7 years, respectively, in order to avoid noise from dummies with few municipalities. Each lead contains 4 to 16 treated municipalities.

The positive sign and relatively large size of most of the coefficient seem to indicate a consistent incremental effect on the level of real expenditure per capita at all percentiles of the distribution with the more years of EPZ operation. Only the 90th percentile is statistically significant for year zero. The larger effect still occurs at the very upper-tail of the distribution, notably at the 90th percentile with statistically significant point estimates at 5 and 1% levels that imply a growing absolute effect of about 10 to 20% per period. However, the time decomposition indicates statistically significant effects at all deciles above the 30th after 8 years of EPZ operation, sustained only for the median deciles after that period. Different explanations are conceivable here. Classic trickle down mechanisms might be taking place, resulting from both medium-term consumption and production dynamics after the expansion of the zones, or the increases in productivity over the years. Negative values before treatment are too imprecise to make any conclusion[21], but they are supportive of the effect resulting from the arrival of new exporting firms.

Findings of this section shed light on the more complex dynamics of the EPZ policy. They highlight the predominance of the effect at the upper-tail of the distribution. At the same time, they are also suggestive of a diffusion of the treatment effect at the middle of the distribution and 10th percentile, after 8 and 11 years of zone establishment.

4.5.4 Skills distribution

I next use the distribution of per capita years of education to verify the predictions of the conceptual channels (i.e. a skill premium related to the establishment of EPZ). Ideally, I would want to break down the skill distribution by industry and test the effect of EPZ establishment across the full interactions of skill and industries (Atkin 2012). However, data limitations regarding industrial affiliations and labour earnings prevent me from disentangling the effect further.

Given the large proportion of unskilled individuals, testing for the effect of EPZ relative to skill-levels is not straightforward. 30% of the working age sample reports zero years of formal education across the period and only 5.4% reports having completed education at the tertiary level. For this reason, I use relative education levels. First, I calculate the distribution of education (years of completed education[22]) amongst

[21] The particular negative shock at -8 years coincides with the post-hurricane Mitch period hitting Nicaragua.
[22] Years of schooling or completed formal education span from 0 to 20 in the sample.

all working age individuals by year and provinces[23]. Following Atkin (2012) I then generate a three-level skill-specific measure. I code as low-skill, individuals below and at the 40th percentile of the schooling distribution, as mid-skill, individuals above the 40th and up until the 70th percentile, and as high-skill individuals with years of schooling above the 70th percentile[24]. With these measures in hand, I rerun specifications for equation (4.1), but augment the regression with 3 distinct dummies created from the interaction of the treatment variable and the binary variables for each skill level (low, mid and high)[25].

Table 4.8 shows the results of this regression for the average level of real per capita expenditures for the working-age individuals. Results are robust to the inclusion of province-fixed effects and the clustering of standard-errors at the province level. The ordering and signs of the coefficients are largely as expected due to findings in section 4.5.2. The highly significant and positive effect contained at the high-skill level supports the view that the gains from EPZ establishment are larger for the upper-deciles of the distribution. It is hard to be conclusive on what is driving the high-skill individuals to benefit the most from EPZ formation without decomposing by industry. These findings seem supportive of the channels previously mentioned in section 4.5.2. I find similar results using an alternative measure for skill-levels based on the actual level of education achievement[26] (column 2). Low-skill individuals also register significant positive point estimates here, which is consistent with the heterogeneous time dynamics and the bottom decile being significant in some specifications.

4.5.5 Relocation of workers

An important concern that remains relates to possible bias from internal mobility across areas during the period studied. Does low and high-skill labour relocate across space in response to the labour supply shock? The most important threat to validity in the place-based policy literature relates to compositional changes in the distribution of workers within units of analysis. This threat is greater during large time periods, like the one here.

[23] I use working-age individuals given that using the sample of employed instead does not allow me to generate a full distribution. Similarly, using municipalities provide an incomplete picture.
[24] The distribution by skill levels is skewed towards the lower tail in both treated and non-treated municipalities, but the proportion of low-skill is much higher in non-treated municipalities. This is largely due to the period 1993-1998 (Figure E3).
[25] Mid-skill is the omitted variable.
[26] Here low-skill corresponds to working age individuals reporting no years of formal education, mid-skill stands for those having completed primary school, and high-skill includes those having completed secondary education and/or university. I include secondary education as a high-skill given that only 5% of the entire sample has tertiary formation.

There are two dimensions to consider in this case. The first is possible internal migration across municipalities. This would impose an upward or downward bias on my results, depending on which type of migrants arrive or leave treated areas. This dimension is closely related to the skill composition of the migrants, as it could be foreseen that the higher skill may be drawn towards EPZ's employment, and on the contrary the lower-skill may be pushed to rural areas. Additionally, large exogenous inflows of migrants may alter both income levels and firms' location decisions related to the impact of the flows on local wages. During the entire period considered, 10.7% of the sample reports having moved from another municipality in the past five years[27]. The flow however seems to have been under-reported in 1998 and 2009, both years displaying extremely low percentages (2-5%). Excluding these, more than 7% of working age population is defined as an internal migrant (7.6 and 7.7% in treated and control municipalities, respectively). There is limited information regarding origin and destination of the flows, with data available only for 1998. For that year, 7.8% of the migrants moved to a municipality hosting an EPZ, with most flows taking place among non-treated areas.

To address this issue, I start by estimating the likelihood of migrating for the working age population following the establishment of a zone in a given area, using both OLS and logit estimators. Results are reported in Table E10. I find no evidence of a statistically significant impact. Additionally, point estimates are always very small, permitting to discard major bias from large displacement following the establishment of a zone in a given area. The absence of perfect factor mobility is not surprising. The World Bank (2012) estimates that during the 2000s, only 10% of the improvements in living standards were related to population shifts (from rural to urban), and find little evidence of changes in the structural distribution of employment in the last decade, with a continued dominance of the agricultural sector. As a second test, I estimate the effect of EPZ formation on changing the skill composition of the residing population. This may capture finer dynamics due to migration patterns that are not visible using the overall migrants' measure. Results in Table 4.9 are in line with theoretical intuitions. Although the sign of the coefficients on the share of high skill changes with respect to the definition, this is due to the larger weight given to the medium-skill (primary school) using the second definition. I focus on the first definition. Here point estimates on the high-skill are large and positive though not statistically significant. Conversely, there seems to be negative to zero effects on the low skill. Overall, the share of the high-to-

[27] As defined in the Household Surveys. This is the only measure consistent for the 5 different surveys.

low skill does seem to be impacted, increasing by 3 percentage points (pp) across the period. Though not robust across specifications, results are significant at 10% level. These changes are in line with the theoretical mechanisms put forward. The small size of the impact means the increase in the proportion of higher-skill workers in treated municipalities following the establishment of an EPZ is unlikely to be driving the average positive effects of the baseline results.

The second dimension to address concerns the general out-migration phenomenon of the poor in Nicaragua. Similar to internal migration, large out-migration flows would also alter the composition of the labour force. It is estimated that near half a million of Nicaraguan migrated to neighbouring Costa Rica since 1994. To address this issue, I run the previous equations holding constant population weights of 1993 (using census data and initial sampling weights). Results remain unchanged discarding any strong compositional changes from out-migration (Table E11).

4.6 Extensions & Robustness Checks

The final empirical section contains some alternative specifications that examine the robustness of the previous conclusions.

4.6.1 Spillover dynamics & commuting

This section addresses the concerns related to possible interactions taking place between an EPZ municipality and its neighbours. If these exist, not taking them into consideration would lead to a bias estimation of the EPZ effects.

I define spillovers as any form of business diversion or creation that could take place because of agglomeration forces between treated and control municipalities adjacent to each other. Backward linkages between EPZs and local producers are probably the main mechanism through which spillovers can be expected to happen. These could lead to neighbouring areas being indirectly 'treated' or on the contrary, to the relocation of economic activity closer to EPZs, in detriment of control municipalities. Additionally, I consider commuting behaviour as a form of spillover. Individuals working in a treated municipality but residing in a neighbouring area are likely to generate similar types of possible externalities. Both mechanisms are hard to measure. Information on commuting is only available for 2001, when it represented 3.56% of the total sample (5.05 and 3.02% in treated and control municipalities,

respectively). This is consistent with some studies finding limited labour mobility in Nicaragua[28].

In order to formally test for the possibility of these two dynamics, I measure treatment effect across the real expenditure distribution of municipalities adjacent to a treated municipality. For this, I compute a modified version of equation (4.3) adding a set of dummy variables that switch to one for municipalities neighbouring a treated area. Because of the particular small size of Nicaraguan municipalities in the west part of the country, considering the shared border as criteria for proximity would imply denoting the vast majority of control municipalities in the region as neighbours. Instead, I use the municipality's distance from its centroid to the centroid of the nearest treated one using two levels of distance, one set at between 5 to 15km and the other at 15 to 35km[29]. This differentiation allows accounting for both dimensions, commuting with the closer group, and more formal spillover effects with the one further away.

Table 4.10 provides the DID estimates augmented with the neighbour dummies on the deciles of the real per capita expenditure distribution. It is encouraging to find a close similarity in the size, sign and statistical significance of the EPZ coefficients to the ones in previous estimations. Point estimates on neighbouring areas seem to indicate low (negative) to no spillover effect from EPZ location. Only one coefficient shows statistical significance but the size is not higher than 2%. Despite the sign being consistently negative for the larger ring, bias from this phenomenon can be expect to be very small, validating the previous estimates.

4.6.2 Balanced panel & weight of capital city

Given that the identification strategy relies on the timing and cross-section variation of EPZ establishment across municipalities, this final empirical section addresses the possibility that the results are driven by the inclusion of particular outliers. First, I limit the sample to municipalities observed for at least 4 of the 5 periods considered. Second, I exclude the capital city (and municipality) of Managua from my sample to verify that this larger agglomeration is not pooling the results. Managua

[28] Related to the low-skill level of the labour force, the cost of transport with respect to minimum living standards and frequent transportation strikes, commuting distances are found to be an important deterrent for participating in the labour market (Jansen et al. 2007).

[29] The measure is of course imperfect. While using the centroid's distance offers the advantage of reducing the noise from larger municipalities where a large proportion would not be in contact with treated areas, it might also exclude some bordering zones. Smaller units of analysis and household's geo-references would be needed for a more precise measure. For simplicity, I refer to the measure as the direct distance between municipalities, but it always registers the distance between the centroids of two municipalities.

remains by far the largest municipality in the country in terms of economic output and population, with its average level of real expenditure per capita being twice the one of all other municipalities in the country. The range is almost as high across all the decile of the expenditure distribution. Results are reported in appendices E12-E13. Overall, conclusions remain unchanged.

4.7 Conclusion

The present analysis has exploited a rich dataset at both individual and municipal levels to estimate average and distributional effects of EPZ establishment within their host municipalities in Nicaragua for the period 1993-2009. Evidence amassed is consistent across different specifications and robustness checks.

Overall, I find robust evidence that EPZ establishment in a given municipality increased the average level of real expenditure per capita between 10 to 12% across the period. The decomposition of the effect across the expenditure distribution suggests that the upper-tail benefited the most both in terms of size and across time, with middle-range deciles benefiting only after eight years of a zone establishment. The exception is a small significant effect at the 10^{th} percentile. The bulk of the positive effect is concentrated on the higher-skill working-age individuals.

Data limitations do not allow me to be conclusive regarding the ultimate channels at play here. An obvious question is why EPZs would have mainly benefitted higher-income households. Is this due to the effect on the local cost of living, rent or land prices? Or is it the result of the skill premium in the exporting sector? The question remains open on whether the effect is really due to the exporting nature of the firms or the local dynamics on factor prices and possible production externalities.

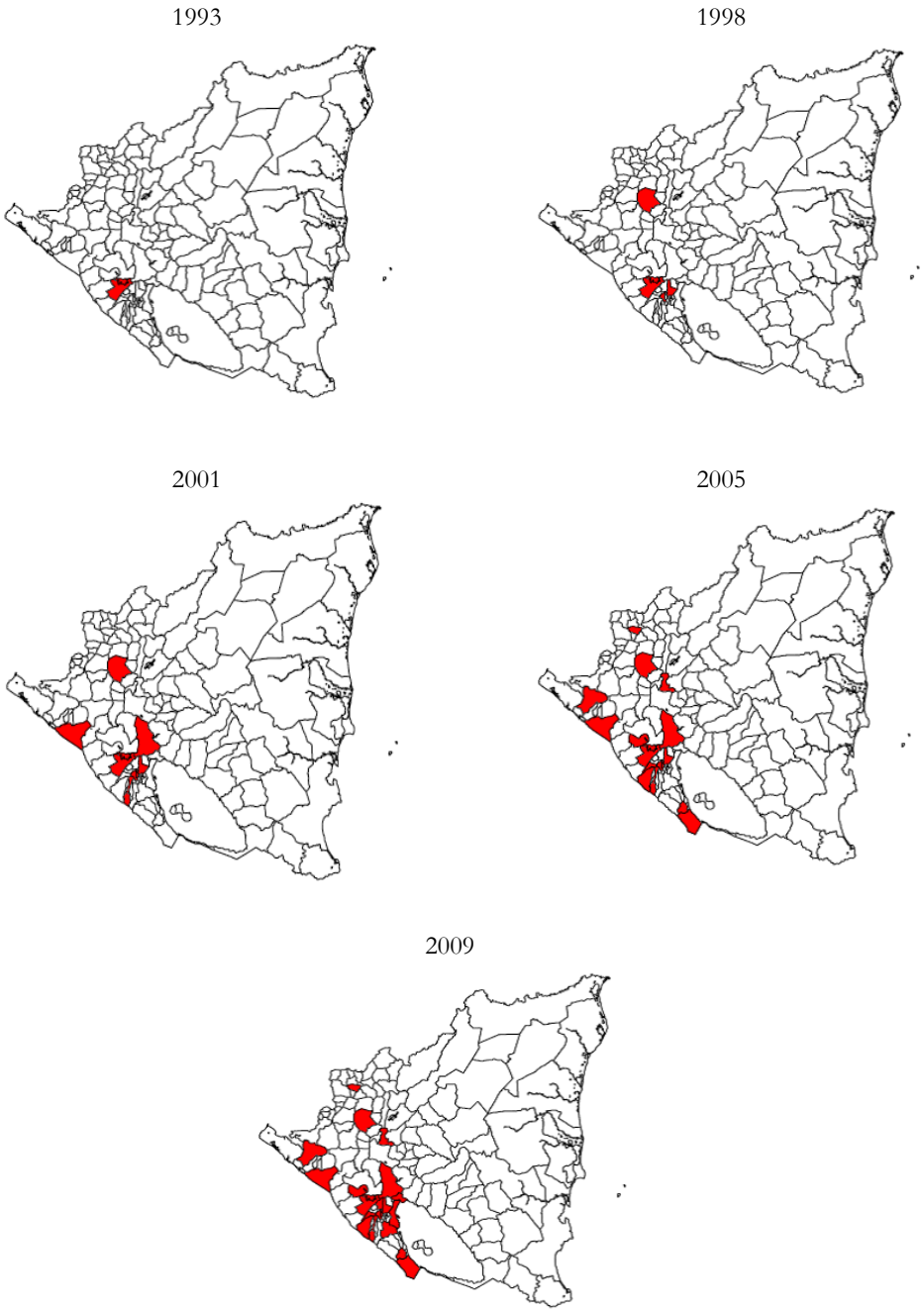
Theoretical models incorporating responses in land prices and rents (Neumark and Simpson 2014; Busso et al. 2013; Kline and Moretti 2014), as well as labour market heterogeneity (Topalova 2010; Kovak 2013) seem to be the best framework to understand both possibilities. In this sense, this paper offers some interesting insight on the importance of analysing trade policies at the local level, taking into account local labour market dimensions and agglomeration economies. Goldberg and Pavcnik (2016) acknowledge the move towards this direction, and underline the interest of the empirical literature in influencing richer theoretical models. Findings here encourage the honing of theoretical trade frameworks that incorporate labour market heterogeneity in

terms of skills and formality of employment, as well different mobility patterns between low and high income segments.

The findings in this study are also important from a policy perspective. They contribute to elucidate the dynamics of a popular policy tool, and contradict popular belief that EPZs are directly beneficial for the low-skill/low-income groups through local employment. The time lag registered for the policy to reach different segments of the distribution leaves a door open for policy intervention. There is an urgent need for further research in the area in order to identify with more precision the different mechanisms at play.

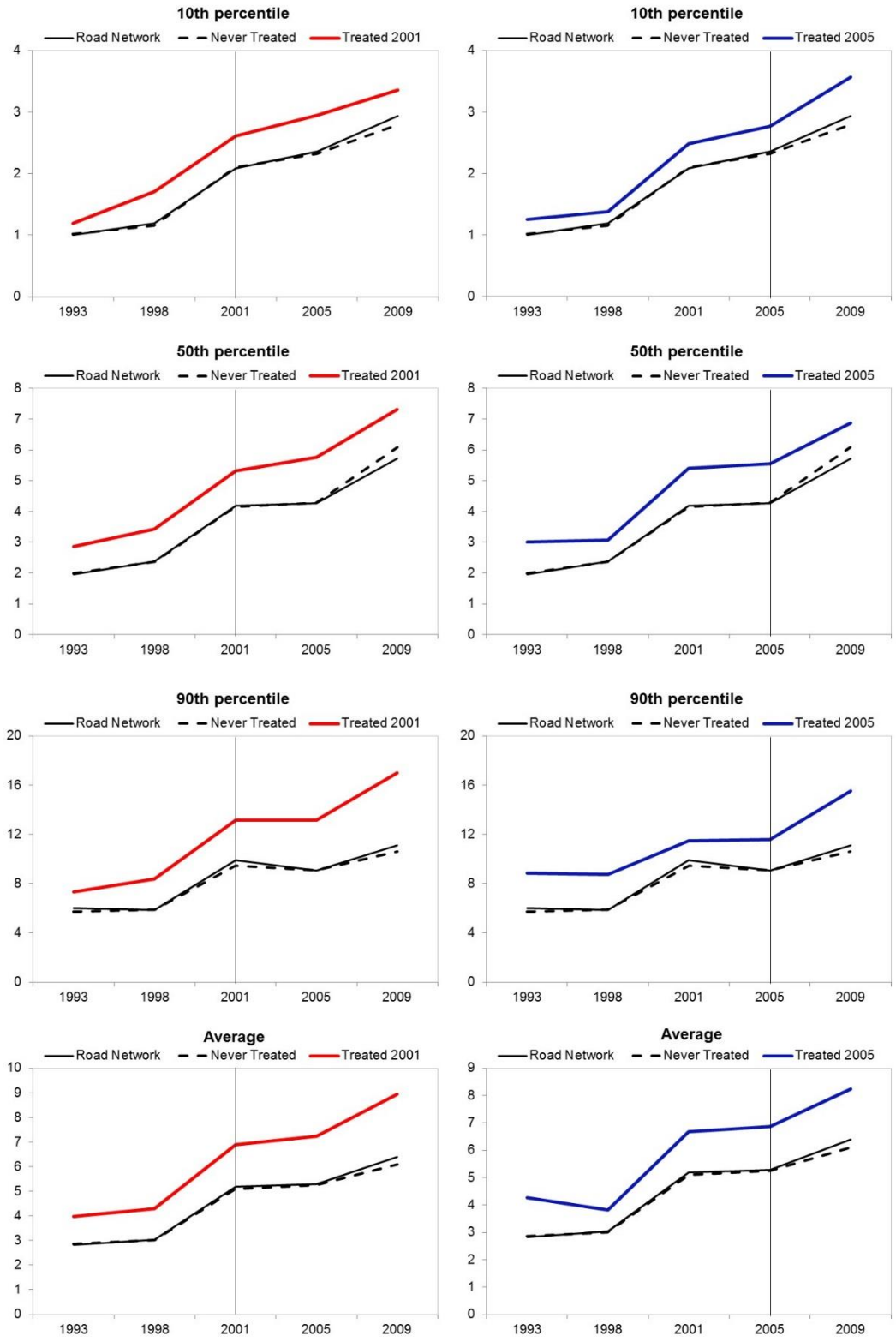
4.8 Figures and Tables

Figure 4.1. Gradual Establishment of EPZs in Nicaraguan Municipalities (1993-2009)



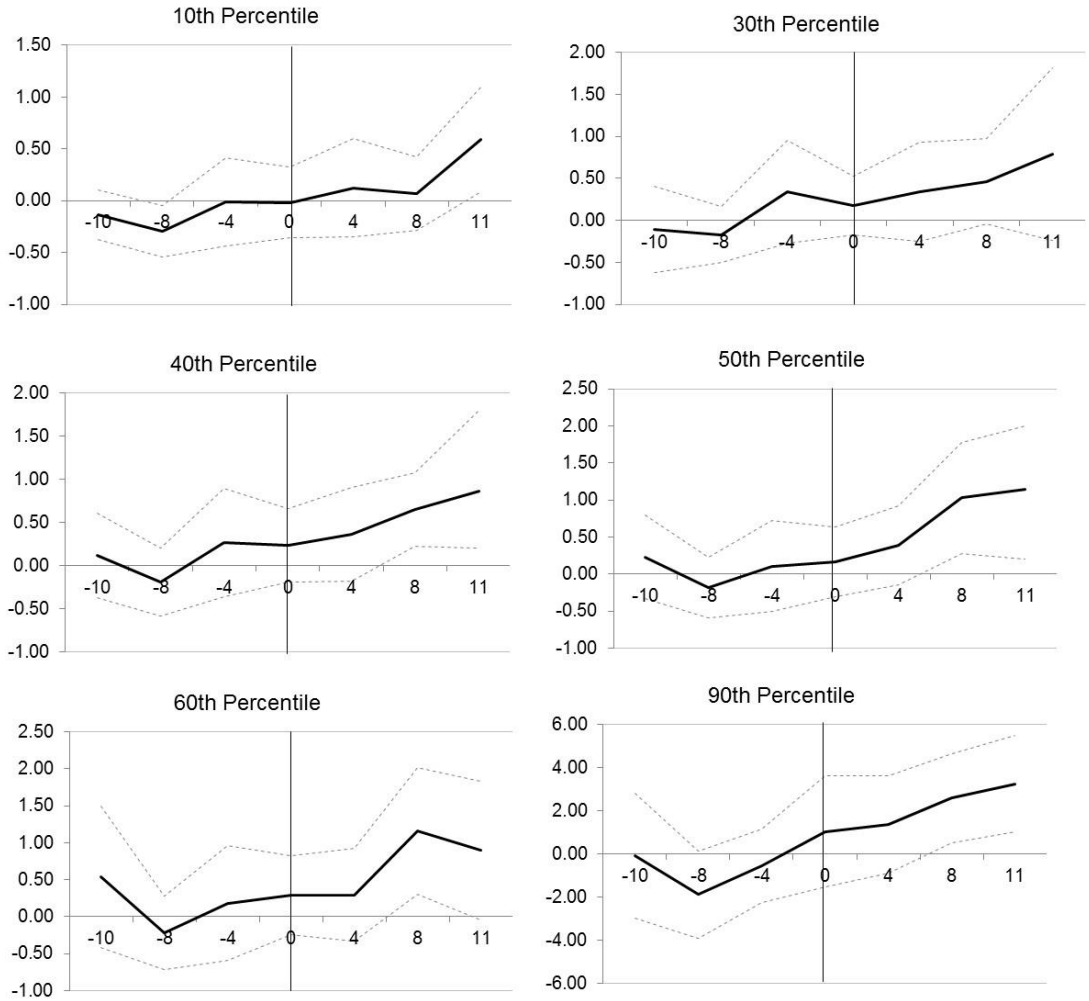
Notes: GIS digital maps. Municipalities are highlighted following the establishment of an industrial park or single factory under the EPZ regime. During the period considered some municipalities may contain more than one firm or industrial park. Data was compiled from the annual yearbooks of the Comision Nacional de Zonas Francas (CNZF).

Figure 4.2. Parallel Trends – Selected Outcome Variables, Preferred Control Group



Notes: Figures depict parallel trends of levels of real expenditure per capita according to treatment and control groups. Main control municipalities are selected according to the density of the road network. On the left side, the treated include only those treated in 2001, and on the right side the treated are only those treated in 2005. Values are in thousands of constant Cordobas (C$1999).

Figure 4.3. Events Study – Selected Outcome Variables



Notes: The horizontal axis measures the number of years since the EPZ program took place. The vertical line represents thousands of Cordobas. The plots connect by the solid line indicate the effect on the levels of real per capita expenditure compared to the period immediately before, the dotted lines are the 95% CI where the standard errors are clustered at the municipality level. See table 7 for point estimates.

Table 4.1. Sample Description (Cross Section and Panel Datasets)

Sample Level	1993	1998	2001	2005	2009	Total
Municipalities	80	115	114	126	92	527
with EPZs	1	3	9	16	18	47
Individuals	24,892	23,182	22,298	36,183	29,749	136,304
Individuals (≥15 years)	13,689	13,185	13,306	22,146	20,146	82,472
Households	4,412	4,125	4,075	6,779	6,375	25,766

Notes: Not all municipalities are observed every year. Municipalities with less than 30 households each year were dropped from the sample. The minimum number of households per municipality in a given year is 30 and the max excluding Managua is 408 (Managua's max. is 2004).

Table 4.2. Summary Statistics Main Outcome Variables (1993-2009)

Variable	Mean	Std. Dev	*N*
Average real per capita expenditure, annual	6,642.68	7,268.44	81,249
Individuals reporting having migrated in the past 5 years	0.15	0.36	82,472
Average real per capita expenditure, annual (municipality)	4,824.33	2,390.96	527
10th percentile - real per capita expenditure, annual	1,976.24	1,065.97	527
30th percentile - real per capita expenditure, annual	2890.912	1477.893	527
50th percentile - real per capita expenditure, annual	3,890.79	1,953.95	527
70th percentile - real per capita expenditure, annual	5,289.22	2,696.75	527
90th percentile - real per capita expenditure, annual	8,792.86	5,094.79	527
Share of low-skill (definition 1)	0.34	0.11	527
Share of high-skill (definition 1)	0.09	0.09	527
Share of low-skill (definition 2)	0.37	0.21	527
Share of high-skill (definition 2)	0.12	0.11	527
EPZ Neighbouring Municipalities (5 to 15 km)	0.05	0.21	527
EPZ Neighbouring Municipalities (15 to 30 km)	0.13	0.34	527
Migrant in the past 5 years (municipality share)	0.10	0.12	527

Notes: The first two variables are measured at the working-age individual level. Expenditure measures are expressed in constant Cordobas of 1999 (Exchange rate US$1=C$11.8). The share of low and high skill according to definition 1 are determined with respect to the skill distribution of the province; definition 2 is defined with respect to years of formal education. See section 6 for further details. The share of migrants is estimated by individuals reporting having migrated to the current municipalities in the previous 5 years. The EPZ neighbouring municipalities represents the share of municipalities at the given km distance to a treated municipality.

Table 4.3. Summary Statistics of the EPZ by Establishment Sequence (1993-2009)

Timing	Group 1 [<1999]		Group 2 [2000-2001]		Group 3 [2002-2005]		Group 4 [2006-2009]		Municipalities Never EPZ	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A. Granting Sequence										
Municipalities with newly established EPZs	4		5		7		2			
Total Municipalities with EPZs	4		9		16		18		108*	
B. Municipality Characteristics										
Distance To Managua (km)	56.70	51.44	43.09	76.25	84.84	48.07	35.88	13.28	142.77	77.85
Road Density (m)	91.30	28.51	92.33	24.50	80.68	29.66	225.42	109.39	46.24	40.69
Landlocked (=1)	0.53	0.51	0.38	0.50	0.57	0.50	0.36	0.49	0.46	0.50
Main Trade Access (=1 if main port, airport or trade post in municipality)	0.20	0.32	0.21	0.41	0.20	0.41	0.64	0.50	0.31	0.46
Electricity (% Households)	0.99	0.01	0.97	0.07	0.94	0.14	0.97	0.03	0.97	0.09
Completed Primary Education	0.27	0.03	0.29	0.03	0.28	0.44	0.29	0.03	0.26	0.06
Illiterate	0.13	0.06	0.18	0.10	0.20	0.12	0.18	0.08	0.31	0.16
Economically Active Population	0.61	0.05	0.59	0.42	0.58	0.07	0.57	0.05	0.54	0.07
Unemployment	0.07	0.06	0.10	0.07	0.11	0.09	0.11	0.09	0.08	0.07
Manufacturing (% employed)	0.23	0.05	0.21	0.08	0.17	0.08	0.17	0.03	0.12	0.08
Agriculture (%employed)	0.08	0.11	0.16	0.15	0.30	0.30	0.18	0.04	0.55	0.27
Urban	0.90	0.11	0.74	0.18	0.57	0.33	0.41	0.30	0.35	0.34
Migrants in past five years	0.10	0.14	0.08	0.97	0.07	0.09	0.07	0.08	0.09	0.12

Notes: Municipalities are grouped based on the sequence of EPZ establishment. Values are for sample averages across the period. Distance to Managua measures the centroids' distance of a municipality to the capital city, the measure excludes Managua; road density corresponds to meters of paved roads every 1 sq.km; main trade access is a composite dummy that equal one if a municipality has a major port, airport or trade post; economically active population corresponds to individuals betwene 13 and 68 years old as per the definition of the statistical office; percent employed in different sectors, urban and unemployment levels are census data (1995-2005).Migrants in the past five years is the share of individuals that migrated to the municipality in the last 5 years.*Never EPZ municipalities vary per year, as per the unbalanced nature of the panel. There are between 64 in 1993 to 108 in 2005.

Table 4.4 Balance of covariates for the pre-treatment period, by sequence of establishment – Preferred Control Group (Road Network)

Sequence of establishment	Group 1 [<1999]	RN Control	t-test	Group 2 [2000-2001]	RN Control	t-test	Group 3 [2002-2005]	RN Control	t-test	Group 4 [2006-2009]	RN Control	t-test
A. Municipality Characteristics												
Distance To Managua (km)	72.54	101.72	-1.70*	42.43	107.11	-3.9***	90.96	103.60	-1.00	35.65	107.92	-4.56***
Road Density (m)	84.67	71.67	0.47	92.00	71.88	1.26	78.42	72.28	0.51	225.42	156.85	3.01***
Landlocked (=1)	0.67	0.71	-0.18	0.40	0.74	-2.38**	0.57	0.74	-1.50	0.36	0.76	-2.98**
Main Trade Access (=1 if main port, airport or trade post in municipality)	0.27	0.43	-1.49	0.20	0.43	-1.48	0.42	0.42	-0.01	0.63	0.42	1.43
Electricity (% Households)	0.99	0.82	0.99	0.94	0.86	1.00	0.93	0.88	0.98	0.97	0.89	1.33
Completed Primary Education	0.27	0.29	-0.29	0.29	0.30	-0.03	0.27	0.27	-0.01	0.29	0.27	0.97
Illiterate	0.22	0.35	0.22	0.24	0.32	-1.42	0.24	0.31	-1.73*	0.21	0.29	-1.56
Economically Active Population	0.54	0.51	0.97	0.56	0.53	1.62	0.55	0.54	1.03	0.55	0.54	0.43
Unemployment	0.03	0.02	0.9	0.04	0.04	-0.04	0.04	0.04	1.0	0.12	0.03	1.20
Manufacturing (% employed)	0.18	0.13	1.10	0.16	0.13	1.17	0.16	0.13	1.41	0.17	0.13	1.64
Agriculture (%employed)	0.08	0.47	-2.29**	0.16	0.50	-3.62***	0.30	0.5	-2.16**	0.20	0.48	-2.11**
Urban	0.90	0.38	2.39**	0.78	0.38	3.42***	0.56	0.39	1.98**	0.35	0.40	-0.47
Migrants in past five years	0.01	0.02	-0.36	0.01	0.02	-1.05	0.07	0.07	-0.07	0.08	0.10	-0.47
B. Pre-trend, Selected Outcomes												
Average real per capita expenditure, annual	0.99	0.59	0.83	0.27	0.46	-0.42	0.48	0.61	-0.56	0.23	0.47	-0.99
10th p- real per capita expenditure, annual.	1.10	0.70	0.73	0.53	0.53	0.00	0.71	0.71	0.03	0.29	0.60	-1.13
50th p- real per capita expenditure, annual	1.13	0.67	0.78	0.32	0.61	-0.51	0.60	0.69	-0.32	0.25	0.53	-1.00
90th p- real per capita expenditure, annual	0.83	0.63	0.31	0.48	0.54	-0.10	0.57	0.64	-0.27	0.23	0.47	-0.83
Share of high-skill over low-skill (1)	4.30	4.02	0.07	2.28	4.21	-1.02	8.80	3.49	2.48**	2.46	2.61	-0.09
Share of high-skill over low-skill (2)	-0.71	-0.50	-0.26	-0.69	-0.48	-0.54	-0.56	-0.47	-0.21	-0.26	-0.09	-0.35
N=	4	101		5	91		7	90		2	93	

Notes: Municipalities are grouped based on the sequence of EPZ establishment. Values are for simple averages for the pre-treatment period. Definitions are as in Table 3. For municipalities treated before 2000, the historical trends denotes the average growth rate of the outcomes before 2001. I would otherwise have no pre-treatment trends for this group. A similar approach is done for the rest of the groups, but strictly considering only the pre-treatment period. The control group is defined with respect to the density of paved road.

Table 4.5. DID Estimates of the Effect of EPZs on Average Real Expenditure per capita

	(1)	(2)	(3)	(4)	(5)
EPZ	**800.5****	**639.6****	**525.8****	**477.4****	**457.9****
	(353.6)	(317.8)	(214.1)	(203.9)	(207.4)
Observations	52,997	52,997	375	375	375
R-sq.	0.341	0.339	0.722	0.818	0.914
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Municipality-year trend	✓	✓			
Province-year dummies			✓	✓	
Region-year dummies					✓
Covariates		✓		✓	✓
Municipalities	106	106	106	106	106

Notes: Dependent variable is levels of real expenditure per capita. Robust standard errors in parentheses, clustered at municipality. Covariates in column 2 include dummies for time-varying individual and household characteristics: educational level, age square, gender, urban, sector of work, household size and having migrated within the past five years. Individual observations are only for working age individuals. Covariates in column 4-5 include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the preferred control group using the road network buffer.

*** p<0.01, ** p<0.05, * p<0.1

Table 4.6. DID Estimates of the Effect of EPZs across the Expenditure Distribution in Treated Municipalities

Real Expenditure per capita	(1)	(2)
10th Percentile	**218.3**	**232.6***
	(145.1)	(138.0)
20th Percentile	143.8	149.8
	(175.7)	(169.9)
30th Percentile	169.2	173.1
	(219.7)	(211.4)
40th Percentile	213.8	209.5
	(228.2)	(218.5)
50th Percentile	258.4	245.2
	(264.8)	(251.5)
60th Percentile	271.3	250.9
	(317.7)	(302.5)
70th Percentile	463.2	414.9
	(359.9)	(345.4)
80th Percentile	**844.0***	785.6
	(506.5)	(475.0)
90th Percentile	**2138***	**2,061***
	(1,223)	(1,109)
Year FE	✓	✓
Municipality FE	✓	✓
Province-year trends	✓	
Region-year dummies		✓
Covariates	✓	✓
Observations	375	375
Municipalities	106	106

Notes: Dependent variables are deciles of real expenditure distribution. Robust standard errors in parentheses, clustered at municipality. Covariates include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the preferred control group using the road network buffer.
*** p<0.01, ** p<0.05, * p<0.1

Table 4.7. The Heterogeneous Time Dynamics of EPZ Establishment

	10th percentile (1)	20th percentile (2)	30th percentile (3)	40th percentile (4)	50th percentile (5)	60th percentile (6)	70th percentile (7)	80th percentile (8)	90th percentile (9)
EPZ establishment, t-11	-134.9	-85.65	-109.9	118.3	228.4	536.8	190.9	310.1	-99.58
	(122.3)	(216.1)	(260.9)	(249.7)	(289.2)	(487.8)	(551.6)	(491.2)	(1,475)
EPZ establishment, t-8	**-298.3****	-244.8	-171.0	-193.6	-184.0	-221.0	**-594.3***	-663.6	**-1,889***
	(127.1)	(174.7)	(170.4)	(201.1)	(211.4)	(253.4)	(341.9)	(448.8)	(1,034)
EPZ establishment, t-4	-10.84	140.8	337.8	266.8	105.9	180.5	-40.45	34.72	-559.7
	(215.9)	(230.5)	(310.7)	(321.5)	(313.5)	(396.0)	(440.5)	(510.6)	(871.0)
EPZ establishment, t0	-16.95	-20.45	172.4	231.8	163.3	287.7	391.4	537.1	1,025
	(175.1)	(189.8)	(177.5)	(216.2)	(239.5)	(272.0)	(511.7)	(548.0)	(1,316)
EPZ establishment, t+4	124.9	271.6	336.3	363.5	390.7	296.7	238.6	722.2	**1,032***
	(240.7)	(243.6)	(301.5)	(277.7)	(273.6)	(316.7)	(417.3)	(545.9)	(528.3)
EPZ establishment, t+8	66.13	399.8	**462.4***	**651.6*****	**1,029*****	**1,163*****	**1,149****	**1,942****	**2,591****
	(180.1)	(284.4)	(259.0)	(217.8)	(383.6)	(436.2)	(464.6)	(845.2)	(1,052)
EPZ establishment, t+11	**589.1****	492.9	782.7	858.8	**1,148*****	**895.0***	1,022	962.8	**3,237*****
	(258.7)	(344.7)	(523.6)	(532.5)	(433.8)	(478.1)	(680.5)	(634.0)	(1,140)
Observations	375	375	375	375	375	375	375	375	375
R-sq.	0.709	0.748	0.752	0.774	0.785	0.778	0.702	0.728	0.630
Municipalities	106	106	106	106	106	106	106	106	106

Notes: Dependent variables are deciles of the real per capita expenditure distribution. Robust standard errors in parentheses, clustered at municipality. All regressions include year and municipality fixed-effects and province-year dummies. Covariates include illiteracy rate, the share of population with primary and tertiary education, and the share of urban population, migrants and households with access to electricity at municipal-level. All leads are equal to one in only one year each per adopting state. EPZ dummy (t0) equals one only for year of establishment. EPZ t+/-4 includes t+/-3, and EPZ t+/-8 includes t+/-7, to have balanced dummies. Results are estimated against the preferred control group using the road network buffer. *** p<0.01, ** p<0.05, * p<0.1

Table 4.8. The Effect of EPZs By Skill-Level

	Skill-Distribution	*Education-Level*
	(1)	(2)
EPZ	486.1	204.0
	(466.7)	(485.1)
Low-Skill*EPZ	29.98	**209.1****
	(134.4)	(70.18)
High-Skill*EPZ	**1,810*****	**2,018*****
	(189.4)	(115.6)
Observations	52,997	52,997
R-sq.	0.338	0.346
Municipalities	106	106

Notes: Dependent variables are levels of real expenditure per capita. Robust standard errors in parentheses, clustered at municipality. All regressions include time and municipality fixed effects, municipality-time trends and province-time dummies. Covariates include education level, illiteracy rate, as well as the share of urban population, migrants and households with access to electricity at municipal-level. Column 1 show the results using the years of education to define the skill distribution. Column uses the second definition using the levels of formal education. Results are estimated against the preferred control group using the road network buffer. *** p<0.01, ** p<0.05, * p<0.1

Table 4.9. The effect of EPZ formation on skill composition

Definitions	Skills Definition 1		Skill Definition 2	
	(1)	(2)	(3)	(4)
Share of High Skill	0.253	0.134	-0.205	-0.242
	(0.176)	(0.150)	(0.228)	(0.221)
Share of Low Skill	-0.01	0.000	-0.018	-0.007
	(0.015)	(0.015)	(0.014)	(0.0112)
Proportion of High/Low Skill	**0.0369***	0.0251	0.005	0.002
	(0.022)	(0.021)	(0.016)	(0.011)
Observations	375	375	375	375
R-sq.	0.753	0.865	0.845	0.833
Year FE	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓
Municipality-year trend	✓		✓	
Province-year dummies		✓		✓
Covariates	✓	✓	✓	✓
Cluster Municipalities	106	106	106	106

Notes: Dependent variable are the share of high and low skill working age populations and the proportion of high to low skill populations at the municipal level. Robust standard errors clustered at municipality are in parentheses. Columns 1-2 define skills according to the distribution of years of education, while columns 3-4, and define skills according to the level of formal education achieved. Definition are in section 5.4. Covariates are as in table 6. Results are estimated against the preferred control group using the road network buffer.

*** p<0.01, ** p<0.05, * p<0.1

Table 4.10. The Effect of EPZs Establishment & Spillover Dynamics

	10th percentile	20th percentile	30th percentile	40th percentile	50th percentile	60th percentile	70th percentile	80th percentile	90th percentile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
EPZ	**232.8***	121.0	127.4	227.2	271.3	307.3	532.2	**942.5***	**2,123***
	(134.6)	(173.9)	(214.3)	(231.7)	(268.2)	(319.9)	(376.5)	(491.4)	(1,142)
Neighbouring Municipalities 1 (5 to 15km)	-147.1	-401.4	-362.3	109.8	12.05	74.33	454.2	699.8	-238.4
	(327.0)	(306.1)	(381.1)	(389.3)	(343.1)	(322.7)	(421.0)	(538.6)	(493.9)
Neighbouring Municipalities 2 (15 to 35km)	-139.6	-237.4	**-357.6***	-257.7	-239.9	-142.6	-170.1	-104.1	-435.0
	(127.5)	(159.3)	(184.7)	(176.0)	(222.3)	(248.9)	(300.7)	(415.8)	(568.3)
Observations	375	375	375	375	375	375	375	375	375
R-sq.	0.687	0.735	0.739	0.756	0.759	0.753	0.691	0.699	0.687

Notes: Dependent variables are deciles of the real per capita expenditure distribution. Robust standard errors in parentheses, clustered at municipality. All regressions include year and municipality fixed-effects and province-year dummies. Covariates include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and the share of households with access to electricity at municipal-level. The variables Neighbour Municipalities are dummy variables that equal one if a municipality's centroid is situated at 5 to 15km, and 15km to 35km distance from a treated municipality, and zero otherwise. I use the centroid's distance to avoid noise from larger municipalities. Results are estimated against the preferred control group using the road network buffer.

*** p<0.01, ** p<0.05, * p<0.1

Bibliography

Abadie, A. 2005 "Semiparametric difference-in-differences estimators." *The Review of Economic Studies* 72, no. 1, pp 1-19.

Aggarwal, A. 2006 "Special economic zones: revisiting the policy debate. *"Economic and Political Weekly,* pp.4533-4536**.**

Albouy, D. and Lue, B., 2015. Driving to opportunity: Local rents, wages, commuting, and sub-metropolitan quality of life. *Journal of Urban Economics*, *89*, pp.74-92.

Alonso, W., 1964. Location and land use. Toward a general theory of land rent. *Location and land use. Toward a general theory of land rent*.

Ambrus, A., Field, E. and Gonzalez, R., 2015. *Loss in the Time of Cholera: Long-run Impact of a Disease Epidemic on the Urban Landscape*. Working Paper.

Ananat, E.O., 2011. The wrong side (s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics*, *3*(2), pp.34-66.

Athey, S., and Imbens G. 2006. "Identification and Inference in Nonlinear Difference in Differences Models." *Econometrica* 74 (2): 431–97.

Atkin, D. 2012 Endogenous skill acquisition and export manufacturing in Mexico. No. w18266. National Bureau of Economic Research.

Autor, D. H, and Dorn D. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–97.

Baker, J., Basu R., Lall S., and Takeuchi A.. 2005. "Urban Poverty and Transport: The Case of Mumbai." Working Paper series, WPS3693, World Bank

Bayer, P., Fang, H. and McMillan, R., 2014. Separate when equal? Racial inequality and residential segregation. *Journal of Urban Economics*, *82*, pp.32-48.

Banerjee, A., Deaton, A. and Duflo, E., 2004. Health, health care, and economic development: Wealth, health, and health services in rural Rajasthan. *The American Economic Review*, *94*(2), p.326.

Banerjee, A.V. and Duflo, E., 2007. The economic lives of the poor. *The journal of economic perspectives*, *21*(1), pp.141-167.

Banerjee, Abhijit, Sebastian Galiani, Jim Levinsohn, Zoë McLaren, and Ingrid Woolard. 2008. "Why Has Unemployment Risen in the New South Africa? 1." *Economics of Transition* 16 (4): 715–40.

Barnhardt, S., Field, E. and Pande, R., 2017. Moving to opportunity or isolation? network effects of a randomized housing lottery in urban India. *American Economic Journal: Applied Economics*, *9*(1), pp.1-32.

Barrios, S., Bertinelli, L. and Strobl, E., 2010. Trends in rainfall and economic growth in Africa: A neglected cause of the African growth tragedy. *The Review of Economics and Statistics*, *92*(2), pp.350-366.

Barrios, T., Diamond, R., Imbens, G.W. and Kolesár, M., 2012. Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, *107*(498), pp.578-591.

Bekker, S.B. and Leildé, A., 2006. Reflections on identity in four African cities. African minds.

Benabou, R., 1996. Inequality and growth. *NBER macroeconomics annual*, *11*, pp.11-74.

Bertrand M., Duflo E., and Mullainathan S., 2004 "How much should we trust difference in differences estimates? "*The Quarterly Journal of Economics*; 119 (1) pp. 249-275.

Bleakley, H. and Ferrie, J., 2016. Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *The Quarterly Journal of Economics*, *131*(3), pp.1455-1495.

Brown, M., Daniels, R.C., De Villiers, L., Leibbrandt, M., & Woolard, I., eds. 2013, "National Income Dynamics Study Wave 2 User Manual", Cape Town: Southern Africa Labour and Development Research Unit

Brueckner, J.K., Thisse, J.F. and Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor?: An amenity-based theory. *European economic review*, *43*(1), pp.91-107.

Brueckner, J.K. and Rosenthal, S.S., 2009. Gentrification and neighborhood housing cycles: will America's future downtowns be rich?. *The Review of Economics and Statistics*, *91*(4), pp.725-743.

Brueckner, Jan K., and Somik V. Lall. 2015. *Cities in Developing Countries*. *Handbook of Regional and Urban Economics, Vol. 5*. 1st ed. Elsevier B.V

Burgess, R., Deschenes, O., Donaldson, D. and Greenstone, M., 2011. Weather and death in India. *Cambridge, United States: Massachusetts Institute of Technology, Department of Economics. Manuscript*.

Burgess, R., Deschenes, O., Donaldson, D., and Greenstone, M. (2017). *Weather , Climate Change and Death in India*.

Busso, M., Gregory J., and Kline P.. "Assessing the Incidence and Efficiency of a Prominent Place Based Policy." *American Economic Review*, 103(2): 897-947, 2013.

Case, A.C. and Katz, L.F., 1991. *The company you keep: The effects of family and neighborhood on disadvantaged youths* (No. w3705). National Bureau of Economic Research.

Case, A. and Deaton, A., 1999. School inputs and educational outcomes in South Africa. *The Quarterly Journal of Economics*, *114*(3), pp.1047-1084.

Chen, J.J., Mueller, V., Jia, Y. and Tseng, S.K.H., 2017. Validating Migration Responses to Flooding Using Satellite and Vital Registration Data. *American Economic Review*, *107*(5), pp.441-445.

Chen, F. and Xi L., 2016 "Evaluation of IMERG and TRMM 3B43 Monthly Precipitation Products over Mainland China," Remote Sensing.

Chernozhukov, Victor, and Christian Hansen. "The reduced form: A simple approach to inference with weak instruments." *Economics Letters* 100.1 (2008): 68-71.

Chetty, R., Hendren, N., Kline, P. and Saez, E., 2014. Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), pp.1553-1623.

Chetty, R., Hendren N., and Katz L.F.. *The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment*. No. w21156. National Bureau of Economic Research, 2015.

Chetty, R. and Hendren, N., 2015. The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates. *Harvard University and NBER*.

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

Christopher, A.J., 1997. Racial land zoning in urban South Africa. *Land Use Policy*, *14*(4), pp.311-323.

Christopher, A.J., 2001. Urban segregation in post-apartheid South Africa. *Urban studies*, *38*(3), pp.449-466.

Clasen, T., L. Haller, D. Walker, J. Bartram, and S. Cairncross. 2007, Cost-effectiveness of water quality interventions for preventing diarrhoeal disease in developing countries. Journal of Water and Health 05.4, 599 – 608

Combes, P-Ph., Duranton G., and Gobillon L. 2010 "The identification of agglomeration economies." *Journal of Economic Geography*.

Comision Nactional de Zonas Francas, Directorio Industrial, 2000-2011 - Nicaragua.

Conley, T.G., 1999. GMM estimation with cross sectional dependence. *Journal of econometrics*, *92*(1), pp.1-45.

Crane, J., 1991. The epidemic theory of ghettos and neighborhood effects on dropping out and teenage childbearing. *American journal of Sociology*, *96*(5), pp.1226-1259.

Cutler, D.M. and Glaeser, E.L., 1997. Are ghettos good or bad?. *The Quarterly Journal of Economics*, *112*(3), pp.827-872.

Cutler, D.M., Glaeser, E.L. and Vigdor, J.L., 2008. When are ghettos bad? Lessons from immigrant segregation in the United States. *Journal of Urban Economics*, *63*(3), pp.759-774.

Day, Jennifer, and Robert Cervero. "Effects of residential relocation on household and commuting expenditures in Shanghai, China." *International journal of urban and regional research* 34, no. 4 (2010): 762-788.

David, H., Dorn D., and Hanson G.. 2013 "The China syndrome: Local labor market effects of import competition in the United States." *The American Economic Review* 103, no. 6: 2121-2168.

de Kadt, D. and Sands, M., How segregation drives voting behavior: New theory and evidence from South Africa (working paper version 2016).

Deaton, A.. 1997 *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications, 1997.

Deaton, A. 2005 "Measuring poverty in a growing world (or measuring growth in a poor world)." *Review of Economics and statistics* 87, 1 : 1-19.

Deschenes, O. and Greenstone, M., 2007. The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather. *The American Economic Review*, *97*(1), pp.354-385.

Deschenes, O. and Moretti, E., 2009. Extreme weather events, mortality, and migration. *The Review of Economics and Statistics*, *91*(4), pp.659-681.

Deschenes, O., 2011 "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US," American Economic Journal: Applied Economics,2011, 3 (4), 142–85

Dasgupta,P., 2010 (5). "The place of nature in economic development," Handbook of Development Economics.

Dell, M., Jones, B. F., and Olken, B. A., 2012. Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, *4*(3), 66–95.

Dell, M., Jones, B. F., and Olken, B. A., 2014. What Do We Learn from the Weather ? The New Climate–Economy Literature. *Journal of Economic Literature*, *52*(3), 740–798.

Devoto, F., Duflo, E., Dupas, P., Parienté, W., and Pons, V., 2012. Happiness on tap: Piped water adoption in urban Morocco. *American Economic Journal: Economic Policy*, *4*(4), 68–99.

Dunkle, S.E., Mba-Jonas, A., Loharikar, A., Fouché, B., Peck, M., Ayers, T., Archer, W.R., De Rochars, V.M.B., Bender, T., Moffett, D.B. and Tappero, J.W., 2011. Epidemic cholera in a crowded urban environment, Port-au-Prince, Haiti. *Emerging infectious diseases*, *17*(11), p.2143.

Duranton, G. and Puga, D., 2004. Micro-foundations of urban agglomeration economies. Handbook of regional and urban economics, 4, pp.2063-2117.

Durlauf, S.N., 2004. Neighborhood effects. *Handbook of regional and urban economics*, *4*, pp.2173-2242.

Edin, P.A., Fredriksson, P. and Åslund, O., 2003. Ethnic enclaves and the economic success of immigrants—Evidence from a natural experiment. *The Quarterly Journal of Economics*, *118*(1), pp.329-357.

Engman, M., Onodera O., and Pinali E. 2007 *Export processing zones: Past and future role in trade and development*. No. 53. OECD Publishing.

ECLAC, Economic Survey of Latin America and the Caribbean, 2012.

Farole, T., and Akinci G., eds. *Special economic zones: progress, emerging challenges, and future directions*. World Bank Publications, 2011.

Freeman, R.B., 1999. The economics of crime. *Handbook of labor economics*, *3*, pp.3529-3571.

Fetzer, T., 2014. Can Workfare Programs Moderate Violence? Evidence from India, STICERD Working Paper.

Field, E.. 2007. "Entitled to Work: Urban Property Rights and Labor Supply in Peru." *Quarterly Journal of Economics*, no. November.

Field, E., .Levinson M., Pande R., and Visaria S.. "Segregation, Rent Control, and Riots: The Economics of Religious Conflict in an Indian City." *The American Economic Review* (2008): 505-510.

Franklin, S. Location, search costs and youth unemployment: A randomised trial of transport subsidies in Ethiopia. *Forthcoming*, Economic Journal.

Franklin, S.. *Enabled to Work: The Impact of Government Housing on Slum Dwellers in South Africa*. No. 2015-10. Centre for the Study of African Economies, University of Oxford, 2015.

Galiani, S., Gertler, P.J., Undurraga, R., Cooper, R., Martínez, S. and Ross, A., 2017. Shelter from the storm: Upgrading housing infrastructure in Latin American slums. *Journal of Urban Economics*, *98*, pp.187-213.

Galiani, S., and Schargrodsky E. 2010. "Property Rights for the Poor: Effects of Land Titling." *Journal of Public Economics* 94 (9-10). Elsevier B.V.: 700–729.

Gelman, Andrew, and Guido Imbens. *Why high-order polynomials should not be used in regression discontinuity designs*. No. w20405. National Bureau of Economic Research, 2014.

Gibbons, S., Silva, O. and Weinhardt, F., 2013. Everybody needs good neighbours? Evidence from students' outcomes in England. *The Economic Journal*, *123*(571), pp.831-874.

Glaeser, E.L., Kahn, M.E. and Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *Journal of urban Economics*, *63*(1), pp.1-24.

Glaeser, E., 2011. *Triumph of the city: How our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin.

Glaeser, E., and Sims, H., 2015. *Contagion, crime and congestion: overcoming the downsides of density*. (International Growth Center, Policy Brief).

Glaeser, E. and Henderson, J.V., 2017. Urban economics for the developing World: An introduction. *Journal of Urban Economics*, *98*, pp.1-5.

Glick, P., and Roubaud F. 2006 "Export processing zone expansion in Madagascar: What are the labour market and gender impacts?." *Journal of African Economies* 15, 4: 722-756.

Greenstone M. R. Hornbeck and E. Moretti (2010), "Identifying agglomeration economies: Evidence from winners and losers of large plant openings", *Journal of Political Economy*, Vol. 118, No. 3, 536-598.

Gobillon, L., Selod, H. and Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban studies*, *44*(12), pp.2401-2427.

Gobillon, L., Magnac, T. and Selod, H., 2011. The effect of location on finding a job in the Paris region. Journal of Applied Econometrics, 26(7), pp.1079-1112.

Goddard Earth Sciences Data and Information Services Center, "TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7," Goddard Earth Sciences Data and Information Services Center (GES DISC) 2016. Accessed: May 11 2017.

Goldberg P., and Pavcnik N. 2007,"Distributional Effects of Globalization in Developing Countries, *Journal of Economic Literature, American Economic Association*, vol. 45(1), pp 39-82.

Goldberg P., and Pavcnik N. 2016,"The Effects of Trade Policy" National Bureau of Economic Research, Working Paper No. 21957.

Gordon, R.; Bertoldi, A.; Nell, M. 2011. "Housing Subsidy: Exploring their Performance," *Urban Landmark Report*.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. "Identification and estimation of treatment effects with a regression-discontinuity design. *"Econometrica* 69, no. 1 (2001): 201-209

Harrington, L.J., Frame, D.J., Fischer, E.M., Hawkins, E., Joshi, M. and Jones, C.D., 2016. Poorest countries experience earlier anthropogenic emergence of daily temperature extremes. *Environmental Research Letters*, *11*(5), p.055007.

Henderson, J.V., Storeygard, A. and Deichmann, U., 2017. Has climate change driven urbanization in Africa?. *Journal of Development Economics*, *124*, pp.60-82.

Housing Development Agency. 2012. "South Africa : Informal Settlements Status."

Hsiang, S. M., 2010. Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America. PNAS, 107(35).

Huffman, G., 2016"GPM IMERG Late Precipitation L3 1 day 0.1 degree x 0.1 degree V03," Goddard Earth Sciences Data and Information Services Center (GES DISC). Accessed: April 20 2017.

Ihlanfeldt, K.R. and Sjoquist, D.L., 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Housing policy debate*, *9*(4), pp.849-892.

Imbens, G., and J. M. Wooldridge. 2009 "New developments in econometrics" *Lecture Notes, CEMMAP, UCL*.

Imbens G., and Kalyanaraman K.. 2011 "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of Economic Studies*.

Jacob, Brian A. *Public housing, housing vouchers and student achievement: Evidence from public housing demolitions in Chicago*. No. w9652. National Bureau of Economic Research, 2003.

Jacob, B.A. and Ludwig, J., 2012. The effects of housing assistance on labor supply: Evidence from a voucher lottery. *The American Economic Review*, *102*(1), pp.272-304.

Jansen, Hans G., Morely Samuel, Kessler Gloria, Pineiro Valeria, Sanchez Marco, Torero Maximo. "The impact of the Central America Free Trade Agreement on the Central American textile maquila industry" IFPRI Discussion paper, 2007

Jenkins, Mauricio, Gerardo Esquivel, and Felipe Larraín B. 1998. "Export Processing Zones in Central America." 646. Development Discussion Papers, Central America Project Series.

Kain, J.F., 1968. Housing segregation, negro employment, and metropolitan decentralization. *The Quarterly Journal of Economics*, *82*(2), pp.175-197.

Kallaway, P. ed., 2002. *The history of education under apartheid, 1948-1994: the doors of learning and culture shall be opened*. Pearson South Africa.

Katz, L.F., Kling, J.R. and Liebman, J.B., 2001. Moving to opportunity in Boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics*, *116*(2), pp.607-654.

Kilroy, A., 2007. Intra-urban spatial inequality: Cities as "urban regions". *World Development Report: Reshaping economic geography*.

Kerr, Andrew. 2015 "Tax(i)ing the poor? Commuting costs in South Africa." Southern Africa Labour and Development Research Unit Working Paper, no. 156.

Kebede, A.S., Nicholls, R.J., Hanson, S. and Mokrech, M., 2010. Impacts of climate change and sea-level rise: a preliminary case study of Mombasa, Kenya. *Journal of Coastal Research*, *28*(1A), pp.8-19.

Kesztenbaum, L. and Rosenthal, J.L., 2017. Sewers' diffusion and the decline of mortality: The case of Paris, 1880–1914. *Journal of Urban Economics*, *98*, pp.174-186.

Khan F.; Thring, P. 2003. "Housing policy and practice in post-Apartheid South Africa". Heinemann Educational Books.

Kline, P. and Moretti, E., 2014. People, places, and public policy: Some simple welfare economics of local economic development programs.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. "Experimental analysis of neighborhood effects." *Econometrica* 75, no. 1 (2007): 83-119.

Kocornik-Mina, A., McDermott, Th. K.J., Michaels, G. and Rauch, F., 2015. *Flooded cities.* CEP Discussion Paper, 1398. London School of Economics and Political Science, CEP, London, UK.

Kovak, B.K., 2013. Regional effects of trade reform: What is the correct measure of liberalization?. *The American Economic Review*, *103*(5), pp.1960-1976.

Lall, S. V., and Chakravorty S. 2005"Industrial location and spatial inequality: Theory and evidence from India." *Review of Development Economics* , 1 : 47-68..

Lall, S. V., Mattias KA Lundberg, and Zmarak Shalizi. "Implications of alternate policies on welfare of slum dwellers: evidence from Pune, India". *Journal of Urban Economics* 63.1 (2008): 56-73.

Lall, S. V., Rogier van den Brink, Basab Dasgupta, and Kay Muir Leresche. "Shelter from the storm--but disconnected from jobs: lessons from urban South Africa on the importance of coordinating housing and transport policies." *World Bank Policy Research Working Paper* 6173 (2012).

Lall, S. V. 2012. Shelter from the Storm — but Disconnected from Jobs Lessons from Urban South Africa on the Importance of Coordinating Housing and Transport Policies, Policy Research Working Paper 6173 (August), World Bank.

Lall, S.V., Henderson J.V, Venables A.J., "Africa's cities:Opening the door to the world"; World Bank 2017

Lam D., Ardington C., Branson N., Case A., Leibbrandt M., Maughan-Brown B., Menendez A., Seekings J. and Sparks M.. *The Cape Area Panel Study: A Very Short Introduction to the Integrated Waves 1-2-3-4-5 Data.* The University of Cape Town, October 2010.

Lee, David S, and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

LeRoy, S.F. and Sonstelie, J., 1983. Paradise lost and regained: Transportation innovation, income, and residential location. *Journal of Urban Economics*, *13*(1), pp.67-89.

Lipp, E.K., Huq, A. and Colwell, R.R., 2002. Effects of global climate on infectious disease: the cholera model. *Clinical microbiology reviews*, *15*(4), pp.757-770.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R. and Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity. *The American Economic Review*, *103*(3), pp.226-231.

Magruder, J.R., 2010. Intergenerational networks, unemployment, and persistent inequality in South Africa. *American Economic Journal: Applied Economics*, *2*(1), pp.62-85.

Marmer, V., Feir D., and Lemieux T. "Weak identification in fuzzy regression discontinuity designs." Available at SSRN 1608662 (2014).

Marx, Benjamin, Thomas Stoker, and Tavneet Suri. 2013. "The Economics of Slums in the Developing World." *Journal of Economic Perspectives* 27 (4): 187–210.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142 (2)

McNally, A., 2016. "FLDAS Noah Land Surface Model L4 monthly 0.1 x 0.1 degree for Eastern Africa (GDAS and RFE2) V001," Goddard Earth Sciences Data and Information Services Center (GES DISC). Accessed: June 18 2017.

Miguel, E. and Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), pp.159-217.

Miguel, E., Satyanath, S. and Sergenti, E., 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of political Economy*, *112*(4), pp.725-753.

Milanovic, B., 2005. Can we discern the effect of globalization on income distribution? Evidence from household surveys. *The World Bank Economic Review*, *19*(1), pp.21-44.

Milanovic, B. and Ersado, L., 2012. Reform and inequality during the transition: an analysis using panel household survey data, 1990–2005. In *Economies in Transition* (pp. 84-108). Palgrave Macmillan UK.

Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, *57*(2), pp.197-210.

Mills, Gregory, Daniel Gubits, Larry Orr, David Long, Judie Feins, Bulbul Kaul, Michelle Wood, and Amy Jones. "Effects of housing vouchers on welfare families." *Washington, DC: US Department of Housing and Urban Development, Office of Policy Development and Research. Retrieved October*8 (2006): 2010.

Moffitt, Robert A. "Welfare programs and labor supply." *Handbook of public economics* 4 (2002): 2393-2430

Munshi, K., 2003. Networks in the modern economy: Mexican migrants in the US labor market. *The Quarterly Journal of Economics*, *118*(2), pp.549-599.

Muth, R., 1969. Cities and housing: The spatial patterns of urban residential land use. *University of Chicago, Chicago*, *4*, pp.114-123.

NASA Precipitation Measurement Missions, "What is the difference between "Realtime" (RT) and "Production" (Prod) Data?," https://pmm.nasa.gov/data-access 2016. Accessed: April 20 2017.

National Department of Human Settlements, Republic of South Africa. 2012. "Human Settlements Review."

Natty, M., 2013 "Introduction: Key Urban Characteristics of Dar es Salaam Challenges and Opportunities for Resilient Development in the Times of Climate Change," Technical Report, GIZ and Dar es Salaam City Council 2013.

Neumark, D., and Kolko J.. 2010. "Do Enterprise Zones Create Jobs? Evidence from California's Enterprise Zone Program." *Journal of Urban Economics* 68 (1). Elsevier Inc.: 1–19.

Neumark, D. and Simpson, H., 2014. "Place-based policies" No. w20049. National Bureau of Economic Research.

Noah, T., 2016. Born a crime: stories from a South African childhood. Ed. Spiegel & Grau.

Oreopoulos, P., 2003. The long-run consequences of living in a poor neighborhood. *The Quarterly Journal of Economics*, *118*(4), pp.1533-1575.

Osei, F.B. and Duker, A.A., 2008. Spatial dependency of V. cholera prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *International Journal of Health Geographics*, *7*(1), p.62.

Osei, F.B., Duker, A.A., Augustijn, E.W. and Stein, A., 2010. Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. *International Journal of Applied Earth Observation and Geoinformation*, *12*(5), pp.331-339.

Pan-African START, 2011 "Urban Poverty & Climate Change in Dar es Salaam, Tanzania: A Case Study," Technical Report, Pan-African START Secretariat & International START Secretariat & Tanzania Meteorological Agency & Ardhi University.

Penrose, K., De Castro, M. C., Werema, J., and Ryan, E. T. 2010. Informal urban settlements and cholera risk in Dar es Salaam, Tanzania. *PLoS Neglected Tropical Diseases*, *4*(3).

Pieterse, E., 2009. Post-apartheid geographies in South Africa: Why are urban divides so persistent. *Interdisciplinary Debates on Development and Cultures: Cities in Development—Spaces, Conflicts and Agency. Leuven University*, *15*.

Piketty, T., 2000. Theories of persistent inequality and intergenerational mobility. *Handbook of income distribution*, *1*, pp.429-476.

Reynolds, C.L. and Rohlin, S.M., 2015. The effects of location-based tax policies on the distribution of household income: evidence from the federal Empowerment Zone program. *Journal of Urban Economics*, *88*, pp.1-15.

Rospabe, S. and Selod, H., 2006. Does city structure cause unemployment? The case of Cape Town. *Poverty and policy in post-apartheid South Africa*, pp.262-287.

Rust, Kecia. 2006. "Analysis of South Africa's Housing Sector Performance," *Finmark Trust*, December.

Sasaki, S., Suzuki, H., Igarashi, K., Tambatamba, B. and Mulenga, P., 2008. Spatial analysis of risk factor of cholera outbreak for 2003–2004 in a peri-urban area of Lusaka, Zambia. *The American Journal of Tropical Medicine and Hygiene*, *79*(3), pp.414-421.

Sharifi, E., Reinhold S., and Bahram S., 2016. "Assessment of GPM-IMERG and Other Precipitation Products against Gauge Data under Different Topographic and Climatic Conditions in Iran: Preliminary Results," Remote Sensing.

Shertzer, A., Twinam, T. and Walsh, R.P., 2016. Race, ethnicity, and discriminatory zoning. *American Economic Journal: Applied Economics*, *8*(3), pp.217-246.

Shorrocks, A. and Wan, G., 2005. Spatial decomposition of inequality. *Journal of Economic Geography*, *5*(1), pp.59-81.

Sinclair-Smith, K., & Turok, I. (2012). The changing spatial economy of cities: An exploratory analysis of Cape Town. Development Southern Africa, 29(3), 391–417.

Socio-Economic Rights Institute of South Africa (SERI). 2013. "'Jumping the Queue', Waiting Lists and Other Myths."

Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008-2012, Waves 1 to 3 [dataset]. Cape Town: Southern Africa Labour and Development Research Unit, 2015 [producer]. Cape Town: Data First [distributor], 2015.

Statistics South Africa. National Household Travel Survey 2013 [dataset]. Version 1. Pretoria: Statistics South Africa [producer], 2014. Cape Town: DataFirst [distributor], 2014

Sur D., Manna B., Deb A.K., Deen J.L., Danovaro-Holliday M.C., et al., 2004. Factors associated with reported diarrhoae episodes and treatment-seeking in an urban slum in Kolkata, India. Journal of Health, Population and Nutrition 22: 130–138.

Takeuchi, Akie, Maureen Cropper, and Antonio Bento. "Measuring the welfare effects of slum improvement programs: The case of Mumbai." *Journal of Urban Economics* 64, no. 1 (2008): 65-84.

Taylor, D.L., Kahawita, T.M., Cairncross, S. and Ensink, J.H., 2015. The impact of water, sanitation and hygiene interventions to control cholera: a systematic review. *PLoS one*, *10*(8), p.e0135676.

Tissington, K.. 2010. "A Review of Housing Policy and Development in South Africa since 1994." *SERI Report* September

Topalova, P., 2010. Factor immobility and regional impacts of trade liberalization: Evidence on

poverty from India. *American Economic Journal: Applied Economics*, *2*(4), pp.1-41.
Turok, I., 2012. *Urbanisation and development in South Africa: Economic imperatives, spatial distortions and strategic responses*. London: Human Settlements Group, International Institute for Environment and Development.
Turok, I., Budlender, J. and Visagie, J., 2017. *The Role of Informal Urban Settlements in Upward Mobility* (DPRU Working Paper No. 201701).
UN-Habitat. 2008. "Housing in South Africa." Vol. 4:2.
UN-HABITAT 2010, "Informal Settlements and Finance in Dar es Salaam, Tanzania". The Human Settlements Financing Tools and Best Practices Series.
UNDP 2015, Economic Transformation for Human Development, Human Development Report 2014 Tanzania.
Venables, A.J. and Venables, T., 2017. Breaking into Tradables: urban form and urban function in a developing city. *Journal of Urban Economics*.
Wang, R., Jianyao C. and Xianwei W., 2017 "Comparison of IMERG Level-3 and TMPA 3B42V7 in Estimating Typhoon-Related Heavy Rain," Water.
Well, David N., 2007. Accounting for the effect of health on economic growth. *The Quarterly Journal of Economics* 122, no. 3 (2007): 1265-1306.
Wenban-Smith, H., 2014. Population Growth, Internal Migration and Urbanisation in Tanzania, 1967-2012: A Census Based Regional Analysis," Technical Report, International Growth Centre 2014.
Western Cape Government. 2013. "Informal Settlements Status"
WHO, 2008 "Cholera Country Profile: United Republic of Tanzania," WHO Global Task Force on Cholera Control 2008.
World Bank (2009) Nicaragua Poverty Assessment, *Poverty Reduction and Economic Management Sector, Latin America and the Caribbean Region* Report No. 39736-NI, Washington DC.
World Bank (2010): *Logistics Performance Index*. World Bank Enterprise Surveys (2010)
World Bank (2011), Lopez H.J., Shankar R. (Eds). Getting the Most Out of Free Trade Agreements in Central America, Directions in Development (Trade), World Bank Publications, 2011
World Bank (2012) "Better Jobs in Nicaragua: the role of human capital, *Human Development Department, Latin America and the Caribbean Region* Report No. 72923-NI, Washington DC
World Bank (2013), Ferreira et al (Eds). Economic mobility and the rise of the Latin American middle class. World Bank Publications, 2013.
World Bank, 2015. Measuring Living Standards within Cities. Households Surveys: Dar es Salaam and Durban. Washington, DC: World Bank
World Bank, 2016. Investing in Urban Resilience. GFDRR, World Bank Group Publications.
World Bank, 2017, Eds. Lall, S.V., Henderson, J.V. and Venables, A.J. Africa's Cities: Opening Doors to the World. World Bank Publications.
Zenou, Y., 2009. Urban search models under high-relocation costs. Theory and application to spatial mismatch. *Labour Economics*, *16*(5), pp.534-546.
Zenou, Y., 2009. *Urban labor economics*. Cambridge University Press.

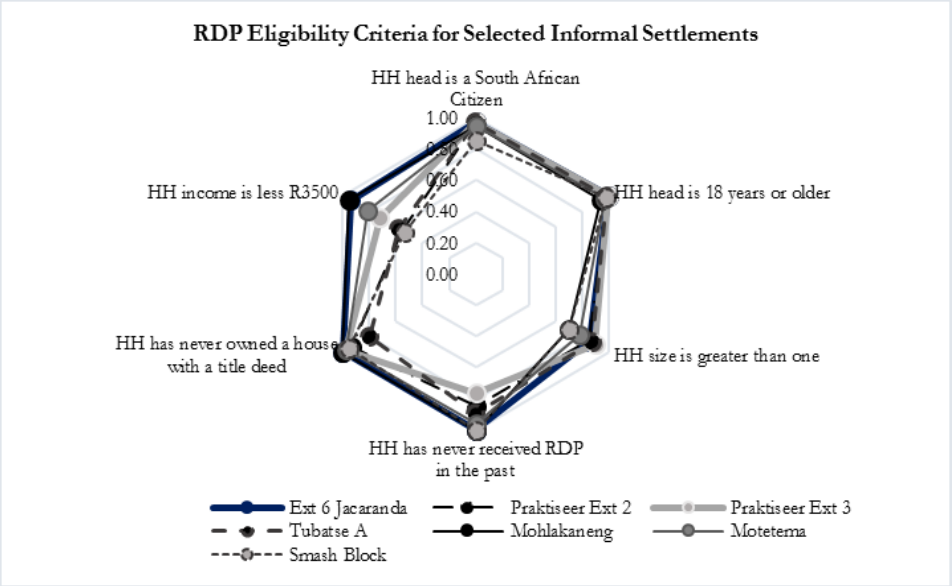
Appendices

Appendix A.

Appendices to Chapter 1.
There is No Free House. Low-Cost Housing & Labour Supply in Urban South Africa.

[LEFT BLANK]

Figure A3. Main Qualification Criteria for Subsidy Housing in Selected Informal Settlements



Notes: Data from HDA (2012)

Figure A4. Distribution of final sample by Metro-Municipality

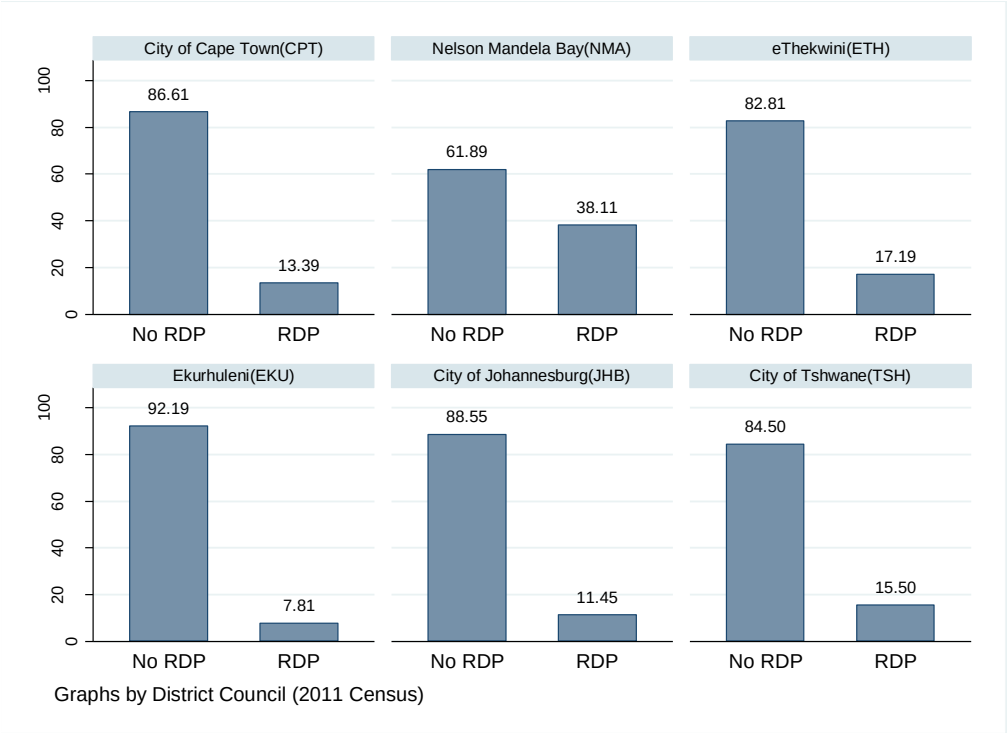
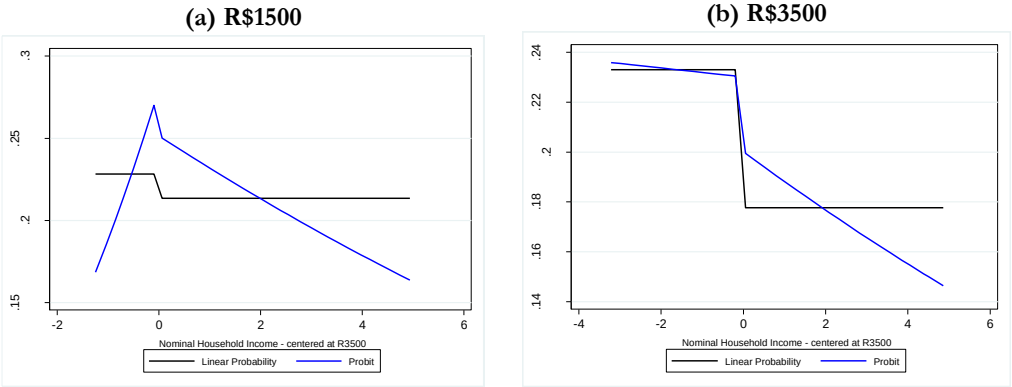
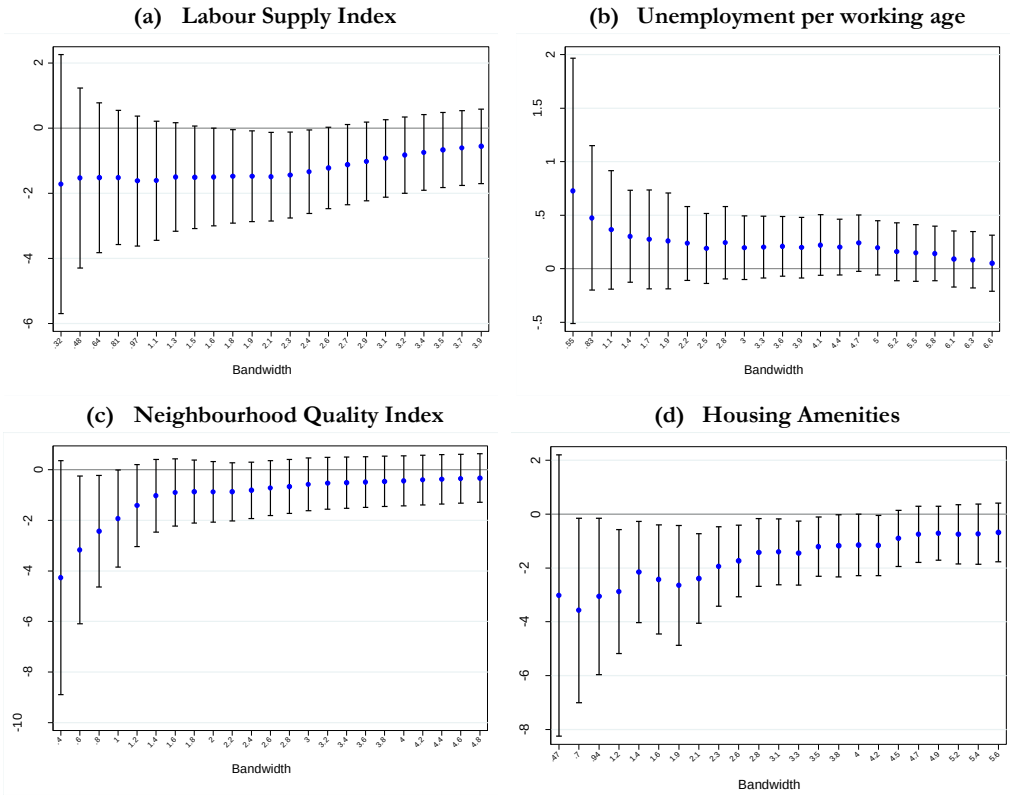


Figure A5. Probability of Receiving RDP Housing at R$1500 & R$3500



Notes: Probit and Linear probabilities of receiving RDP at cutoff.

Figure A6. Non-Parametric Estimates by bandwidths



Notes: The graphs show the 2SLS coefficients estimated in nonparametric regressions when varying the bandwidths between 0.5 and 6 thousand around the threshold. CCT and IK bandwidths are preferred at 1.11 and 2.48 respectively.

Table A1. Initial Sample (NIDS Panel)

	2008	2010	2012	2014	Total
I. Original Sample					
Individuals	28,226	35,116	38,209	43,236	144,787
Adults (14+ years)	18,603	24,012	26,034	29,380	98,029
Households	6,875	6,586	7,539	8,845	29,845
CSM	-	29,448	29,448	30,479	89,375
RDP recipients (HH)	396	997	1,297	1,587	4,277
RDP recipients (Adults)	1,492	3,953	6,835	6,835	19,115

Notes: CSM refers to continuing sample members as define per NIDS. RDP recipients in each wave are individuals that report living in household that has received a housing subsidy at the period interviewed.

Table A2. Sample Selection Criteria

	2008	2010	2012	2014	Total
II. Selected Sample					
Raw data	18,603	24,012	26,034	29,380	98,029
Step 1	18,017	23,215	24,388	27,621	93,241
Step 2	17,340	18,289	20,670	24,096	80,395
Step 3	8,831	10,385	12,465	13,593	45,274
Step 4	8,831	10,104	11,528	12,283	42,746
Step 5	1,269	1,290	1,741	2,031	6,331
Final Sample - Individuals	3,170	4,098	5,018	6,353	18,639
Households	923	940	1,240	1,595	4,698
% RDP HH	0.0	17.6	17.2	17.4	14.2

Notes: Rows in the table refer to different selection steps. Numbers reported refer to observations left after each step. Description of sample selection procedure:
Raw Data: Adults (14+ years old, as defined by SSA), household residents.
Step1: Keep only individuals with age ≥14 in wave 1, age≥16 in wave 2 & age≥17 in wave 3 with South African nationality
Step2: Keep only individuals that did not own a dwelling in previous period.
Setp3: Keep only individuals married/living with partner or with dependents (hh size>1) in previous period.
Step4: Keep only individuals that never received RDP at baseline
Step5: Keep only individuals in 6 main metro municipalities & urban.
Final Sample: Head of HH eligible for RDP subsidies and working-age household members, not subject to income condition yet.

Table A3. Summary Statistics – Indices (Baseline)

	Mean	*Std. Dev.*	*N*
A. Adult Labour Supply	**0.098**	1.00	692
Fraction of HH members employed over working age members	0.294	1.00	692
Weekly labour hours per working age member	-0.065	1.00	692
B. Adult Labour Supply Cost	**-0.08**	1.00	692
Distance to Main Employment Nodes (min by mode of transport)	-0.051	1.00	692
Distance to Main Employment Nodes (km)	-0.115	1.00	692
Monthly amount spent on transport as a fraction of HH income	0.058	1.00	692
C. Amenities Index	**-0.262**	1.00	692
HH has electricity	-0.225	1.00	692
Type of toilet facility in the HH [none to inside, 1-7]	-0.061	1.00	692
Dwelling Rating [1-7]	0.000	1.00	692
Number of Rooms	-0.520	1.00	692
D. Neighbourhood Quality Index	**0.0872**	1.00	692
Refuse collection	0.194	1.00	692
Not common to have robberies in the neighbourhood	0.003	1.00	692
Functioning street light	0.064	1.00	692

Notes: This is the final subsample of RDP eligible households in urban areas in period 1. Employment definition is as in Table 3. Indices' subcomponents are z-scores. Electricity is a dummy variable related to access to electricity within the dwelling; type of toilet is a categorical variable that goes from 1 to 7 for none to flush toilet within the household. The dwelling rating rates 1-5 the structure from dilapidated to in good condition. Refuse collection is a dummy variable for weekly refuse collection by local authorities.

Table A4. RD Manipulation Test using local polynomial density estimation

Running Variable - Household Nominal Monthly Income (t-1)		
Robust Bias-Corrected	T	p>|T|
2010-2014	-0.135	0.892
2008	0.873	0.383
2010	0.082	0.935
2012	0.824	0.410
2014	0.791	0.429
Order local poly (left, right)	2	2

Notes: Data-driven bandwidth selectors following Cattaneo, Jansson and Ma (2015). Tests the null hypothesis that the density of the forcing variable is continuous at cutoff, and its implementation requires the estimation of the density of observations near the cutoff, separately for observations above and below the cutoff, using a local polynomial density estimator. The latter does not require pre-binning of the data. Estimates use triangular kernel. Results are unchanged with uniform kernels or linear regressions. 2008 refers to current household income.

A.I OLS Results

Table A5. OLS Results: Main Labour Market Outcomes

	Household Level dependent variable is:										
	Labour Supply Index			**Main Outcomes**							
	Composite Index	Employed per wam	Weekly hours per wam	Weekly Total Hours	Weekly Hours Female	Weekly Hours Male	Employed Members	Employed Females	Employed Males	Unemployed members	Unemployed per wam.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Baseline Mean	0.148	0.410	25.724	30.311	13.541	16.770	0.815	0.465	0.493	0.472	0.179
	[1.00]	[0.381]	[15.571]	[34.310]	[22.924]	[25.396]	[0.776]	[0.601]	[0.566]	[0.776]	[0.287]
RDP	-0.19***	-0.25***	-0.15***	-6.05***	-2.29**	3.72***	-3.68***	-0.027	-0.10***	0.096***	0.0439***
	(0.0382)	(0.0496)	(0.0344)	(1.4549)	(1.0411)	(1.0421)	(0.8335)	(0.0297)	(0.0286)	(0.0268)	(0.0117)
Time & City FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050	2,050

Notes: The table reports OLS coefficients. Dependent variables are measures of the labour supply of households. Columns (1) to (3) contain the labour supply index and its two components, measured as z-scores. All intensive measures are weekly total hours by household members, and extensive margins are the number of employed members. Wam stands per working-age members. Columns (10-11) are measures of unemployment that include discouraged members. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68-year-old & children <14 years old. Columns (4-5) and (10-11) control for the number of working age members in the household, while columns (6) to (9) control for the number of female and male working age members in the households, separately. Standard deviations are in brackets. *p<0.10; **p<0.05; ***p<0.01

Table A6. OLS Results: Accessibility and *Wellbeing* Measures

	Household Level dependent variable is:				
	Commuting Times & Distances			**Neighbourhood Quality**	**Amenities**
	Composite Index	Distance (km) to CBD	Distance (min) to CBD	Composite Index	Composite Index
	(1)	(2)	(3)	(5)	(6)
Baseline Mean	-0.081 [1.00]	26.823 [16.897]	116.37 [64.58]	0.087 [1.00]	-0.262 [1.00]
RDP	0.154** (0.070)	0.170** (0.078)	0.156* (0.092)	0.098*** (0.0254)	0.239*** (0.0275)
Time & City FE	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Obs.	2,050	2,050	2,050	2,050	2,050

Notes: The table reports OLS estimates of the effect of RDP subsidies on indices of commuting costs, neighbourhood quality and housing amenities. Variables are as defined in main tables. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old, and the working age members. Standard deviations are in brackets. Robust standard errors, clustered at household level in parentheses. *p<0.10; **p<0.05; ***p<0.01

Table A7. OLS Results: Household Shifts Strategies (1): Compositional Changes

	Household Level dependent variable is:						
	Dependency Ratio Children	Dependency Ratio Aged	# Children living at home	Young Children	# of >68 years old	Age of Head	Age of children
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline Mean	0.396 [25.535]	0.041 [0.182]	0.962 [1.327]	0.399 [0.490]	0.118 [0.366]	46.974 [15.204]	8.920 [4.285]
RDP	0.0151 (0.0162)	0.0025 (0.0057)	0.069 (0.0550)	-0.009 (0.0241)	-0.017 (0.0206)	-0.169 (0.1895)	0.62 (0.6138)
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	2,049	2,049	2,049	2,049	2,049	2,049	2,049

Notes: The table reports the estimated OLS coefficients. Dependent variables are measures of household's age composition, as define in main tables. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head. Column 6 excludes age as control. Standard deviations are in brackets. *p<0.10; **p<0.05; ***p<0.01

Table A8. OLS Results: Household Shifts Strategies (1): Coping Mechanisms & Preferences

	Household Level dependent variable is:						
	Coping Mechanisms			Preferences			
	HH Receives Rental Income	Government Grant	Non-commercial Agriculture	Happier than 10 years ago	HH prefers to stay	HH prefers to stay (age<30)	HH prefers to stay (age>30)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Baseline Mean	0.072 [0.258]	0.355 [0.479]	0.070 [0.256]	0.442 [0.497]	0.666 [0.472]	0.615 [0.487]	0.744 [0.436]
RDP	0.065*** (0.0146)	0.082*** (0.0193)	0.006 (0.0120)	-0.034 (0.0224)	0.090 (0.0186)	0.086 (0.0221)	0.072*** (0.0182)
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	2,049	2,049	2,049	2,049	2,049	2,049	2,049

Notes: The table reports OLS coefficients of the effect of RDP subsidies on household shifts strategies. Dependent variables are different measures of household related to possible coping strategies and preferences. Government Grand in column (2) is a dummy variable for receiving any type of government grant, including disability grants. Non-commercial agriculture is a dummy variable equal to one if households participated in agricultural activities without monetary compensation. While the remainder variables are dummy variables expressing happiness with respect to 10 years ago, and preference to stay in the current place, by age of the household head. Column (5) is the average of columns (6) and (7) All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Standard deviations in brackets. *p<0.10; **p<0.05; ***p<0.01

A.II Global Linear Polynomials

Table A9. Discontinuity Effect of RDP on Main Labour Market Outcomes (1)
Global linear polynomial

	Household Level dependent variable is:					
	Unemployed Members	Unemployed Members	Unemployed per wam	Unemployed - strict	Unemployed - strict	Unemployed per wam - strict
	(1)	(2)	(3)	(4)	(5)	(6)
RDP	0.26	0.38	0.24**	-0.12	-0.41	0.04
	(0.2219)	(0.2713)	(0.1195)	(0.2722)	(0.4343)	(0.1086)
First Stage F-Stat	30.25	25.53	30.25	30.25	25.53	30.25
Polynomial Order	1	1	1	1	1	1
Time & City FE	Y	Y	Y	Y	Y	Y
Controls	N	Y	Y	Y	Y	Y
Obs.	1,960	1,626	1,960	1,960	1,626	1,960

Notes: The table reports 2SLS coefficients of the effect of RDP subsidies on labour market outcomes of beneficiary households. Dependent variables are measures unemployment. Columns (1-3) include discouraged members, while the remainder is only strict unemployment defined as those actively looking for a job. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. In columns (1) and (3) I control for the number of working age members in the household. Results in columns (2) and (5) I control for female working age members.
*p<0.10; **p<0.05; ***p<0.01

Table A10. Discontinuity Effect of RDP on Main Labour Market Outcomes (2) Global linear polynomial

	Household Level dependent variable is:								
	Labour Supply Index			Main Outcomes					
	Composite Index	Employed per wam	Weekly hours per wam	Weekly Total Hours	Weekly Hours Female	Weekly Hours Male	Employed Members	Employed Females	Employed Males
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RDP	-0.7376*	-0.49	-0.9343**	-40.26**	-22.15**	-11.82	-0.91**	-0.4527	-0.40
	(0.3830)	(0.5131)	(0.3490)	(14.8517)	(10.3345)	(11.1699)	(0.4144)	(0.3070)	(0.2611)
First Stage F-Statistic	30.25	30.25	30.25	30.25	30.25	30.25	30.25	23.95	26.67
Polynomial Order	1	1	1	1	1	1	1	1	1
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	N	Y	Y	Y
Obs.	1,960	1,960	1,960	1,960	1,960	1,960	1,960	1,620	1,716

Notes: The table reports 2SLS coefficients of the effect of RDP subsidies on labour market outcomes of beneficiary households. Dependent variables are measures of the labour supply of households. Columns (1) to (3) contain the labour supply index and its two components, measured as z-scores. All intensive measures are weekly total hours by household members, and extensive margins are the number of employed members. Wam stands per working-age members. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68-year-old & children <14 years old. Columns (4) and (5) control for the number of working age members in the household, while columns (6) to (9) control for the number of female and male working age members in the households, separately. *p<0.10; **p<0.05; ***p<0.01

Table A11. Discontinuity Effect of RDP on Household Commuting Distances and Times - Global linear polynomial

	Household Level dependent variable is:				
	Composite Index	Distance (km) to CBD	Distance (km) to main nodes	Distance (min) to CBD	Amount spent on transport
	(1)	(2)	(3)	(4)	(5)
RDP	0.47*	12.07***	10.89**	28.83	0.033
	(0.283)	(4.534)	(4.387)	(23.716)	(0.0375)
First Stage F-Stat	38.41	38.41	38.41	38.41	25.84
Polynomial Order	1	1	1	1	1
Time & City fixed effects	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Obs.	2,295	2,295	2,295	2,295	736

Notes: The table reports 2SLS estimates of the effect of RDP subsidies on elements related to the labour supply cost of households. Dependent variables are measures of the labour supply cost of households, with in column (1) the labour supply cost index. Columns (2-4) are the components of the index, and column (5) contains the monthly amount spent on transport as a share of monthly expenditures. Very few households have answered this question, limiting the sample size. Distance to CBD and main nodes are Euclidean distances to CBD (and the secondary node) from the suburb's centroid of the household residence. Distances measured in time of commute are calculated using the time of the most frequently used mode in the commuting area of residence, conditional on income and population group. Column (1) is an average of the normalized values of columns (2) and (4). Here I display the non-normalized value to provide the explicit km and minutes. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. Due to limitations to access secure address identifiers, these regressions are run on a different sample - i.e., for periods 1-3 only. They control only for household size. *p<0.10; **p<0.05; ***p<0.01

Table A12. Discontinuity Effect of RDP on *Wellbeing* (1): Neighbourhood Quality
Global Linear Polynomial

	Household Level dependent variable is:				
	Composite Index	Composite Index	Robberies uncommon	Street Light	Refuse collection
	(1)	(2)	(3)	(4)	(5)
RDP	-0.06	0.001	-0.48	0.05	0.39
	(0.2750)	(0.2609)	(0.4500)	(0.4263)	(0.4082)
First Stage F-Stat	28.57	28.57	28.57	28.57	28.57
Polynomial Order	1	1	1	1	1
Time & City FE	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Obs.	1,947	1,947	1,947	1,947	1,947

Notes: The table reports 2SLS c of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include the neighbourhood quality index and its components in columns (3 to 5). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Columns 2-5 also include a dummy variable for living in informal dwelling in the previous period. *p<0.10; **p<0.05; ***p<0.01

Table A13. Discontinuity Effect of RDP on *Wellbeing* (2): Housing Amenities Global Linear Polynomial

	Household Level dependent variable is:					
	Composite Index	Composite Index	# of rooms	Dwelling Quality	HH has electricity	Toilet inside
	(1)	(2)	(3)	(4)	(5)	(6)
RDP	-0.61* (0.3253)	-0.47* (0.2748)	-1.13** (0.4617)	-0.93** (0.4609)	-0.001 (0.4622)	0.20 (0.4014)
First Stage F-Statistic	30.25	30.25	30.25	30.25	34.05	30.25
Polynomial Order	1	1	1	1	1	1
Time & City FE	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	N	Y
Obs.	1,960	1,960	1,960	1,960	1,960	1,960

Notes: The table reports 2SLS estimates of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include a housing amenities index and its components in columns (3) to (6). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. Columns 2-6 also include a dummy variable for living in informal dwelling in the previous period. *p<0.10; **p<0.05; ***p<0.01

Table A14. Discontinuity Effect of RDP on Household Shifts Strategies (1): Compositional Changes Global Linear Polynomial

	Household Level dependent variable is:						
	Dependency Ratio Children	Dependency Ratio Aged	# Children living at home	Young Children	# of >68 years old	Age of Head	Age of children
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RDP	0.003	0.13	-0.35	-0.31	0.15	-0.17	1.61
	(0.0029)	(0.1001)	(0.4056)	(0.2019)	(0.1540)	(0.6501)	(2.9818)
First Stage F-Statistic	28.25	28.25	27.89	27.89	27.89	27.89	27.89
Polynomial Order	1	1	1	1	1	1	1
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	1,882	1,882	1,947	1,947	1,947	1,947	1,947

Notes: The table reports 2SLS estimates of the effect of RDP subsidies on household shifts strategies. Dependent variables are different measures of household composition. Columns (1) and (2) are dependency ratios calculated as the share of aged <14 & >68 over the working age members. Young children is a dummy variable for household with children below the age of 10. Average age of children and age of the household head are in columns (5) and (6). All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head. *p<0.10; **p<0.05; ***p<0.01

Table A15. Discontinuity Effect of RDP on Household Shifts Strategies (2): Coping Mechanisms & Preferences
Global Linear Polynomial

	Household Level dependent variable is:						
	Coping Mechanisms			**Preferences**			
	HH Receives Rent Income	Government Grant	Non-commercial Agriculture	Happier than 10 years ago	HH prefers to stay	HH prefers to stay (age<30)	HH prefers to stay (age>30)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RDP	0.26*	1.10	0.11	-0.59**	-0.24	-0.38	-0.04
	(0.1378)	(0.1975)	(0.0937)	(0.4329)	(0.2309)	(0.2545)	(0.2138)
First Stage F-Statistic	31.67	28.65	28.58	26.34	21.34	21.34	21.34
Polynomial Order	1	1	1	1	1	1	1
Time & City fixed effects	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y
Obs.	1,952	1,949	1,909	1,824	1,612	1,612	1,612

Notes: The table reports 2SLS coefficients of the effect of RDP subsidies on household shifts strategies. Dependent variables are different measures of household related to possible coping strategies and preferences. Government Grand in column (2) is a dummy variable for receiving any type of government grant, including disability grants. Non-commercial agriculture is a dummy variable equal to one if households participated in agricultural activities without monetary compensation. While the remainder variables are dummy variables expressing happiness with respect to 10 years ago, and preference to stay in the current place, by age of the household head. Column (5) is the average of columns (6) and (7). All regressions control for a linear polynomial of the assignment variable and its interaction with Z. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. *p<0.10; **p<0.05; ***p<0.01

A.III Local Linear Regressions

Table A16. Discontinuity Effect of RDP on Household Commuting
Local linear regressions

	Household Level dependent variable is:			
	Composite Index	Distance (km) to CBD	Distance (km) to main nodes	Distance (min) to CBD
	(1)	(2)	(3)	(4)
RDP	0.16	23.25*	12.61	40.49
	(0.617)	(12.878)	(9.772)	(52.103)
First Stage F-Statistic	21.19	21.19	21.19	21.19
Obs.	1,808	1,808	1,808	1,808
Bandwidth (IK 2012)	√	√	√	√
RDP	0.29	21.10	14.00*	34.56
	(0.564)	(14.852)	(8.260)	(43.052)
First Stage F-Stat	10.96	10.96	10.96	10.96
Obs.	638	638	638	638
Bandwidth (CCT 2014)	√	√	√	√

Notes: The table reports local linear regressions estimates of the effect of RDP subsidies on outcomes related to the labour supply cost of households. Dependent variables are measures of the labour supply cost of households, with in column (1) the labour supply cost index. Columns (2-4) are the components of the index, and column (5) contains the monthly amount spent on transport as a share of monthly expenditures. Distance to CBD and main nodes are Euclidean distances to CBD (and the secondary node) from the suburb's centroid of the household residence. Distances measured in time of commute are calculated using the time of the most frequently used mode in the commuting area of residence, conditional on income and population group. Column (1) is an average of the normalized values of columns (2) and (4). Here I display the non-normalized value to provide the explicit km and minutes. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. Robust standard errors, clustered at household level in parentheses. Due to limitations to access secure address identifiers, these regressions are run on a different sample - i.e., for periods 1-3 only. They control only for household size. IK (2012) and CCT (2014) refer to the rule for bandwidth choice as previously specified. Here optimal IK bandwidth is 3.453 and CCT 1.076. All regressions include city and time fixed-effects and basic controls. *p<0.10; **p<0.05; ***p<0.01

Table A17. Discontinuity Effect of RDP on Main Labour Market Outcomes - Local Linear Regressions

	Household Level dependent variable is:										
	Labour Supply Index			Main Outcomes							
	Composite Index	Employed per wam	Weekly hours per wam	Weekly Total Hours	Hours Female	Hours Male	Employed Members	Employed Females	Employed Males	Unemployed Members	Unemployed per wam
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
RDP	-0.96*	-1.19	-0.72	-30.57	-21.81	-8.25	-0.77	-0.56	-0.12	0.28	0.19*
	(0.5410)	(0.73105)	(0.4833)	(20.6477)	(14.4828)	(15.6408)	(0.5706)	(0.4988)	(0.2001)	(0.3468)	(0.1077)
First Stage F-Stat	9.45	9.45	9.45	9.45	10.15	9.45	9.45	9.45	9.45	9.45	9.45
Obs.	1041	1041	1041	1041	1077	1041	1041	1041	1041	1041	1041
Bandwidth (IK)	√	√	√	√	√	√	√	√	√	√	√
RDP	-0.52	-0.54	-0.48	-14.07	-21.68	6.66	-0.46	-0.86	-0.04	0.57	0.24
	(0.6678)	(0.9417)	(0.5703)	(26.6326)	(18.9268)	(21.5702)	(0.6286)	(0.8402)	(0.5338)	(0.4987)	(0.1585)
First Stage F-Stat	9.05	9.05	9.05	9.05	9.05	9.05	9.05	9.05	9.05	9.05	9.05
Obs.	489	489	489	489	489	489	489	489	489	489	489
Bandwidth (CCT)	√	√	√	√	√	√	√	√	√	√	√

Notes: The table reports local linear regressions estimates of the effect of RDP subsidies on labour market outcomes of beneficiary households. Dependent variables are measures of the labour supply of households. Columns (1) to (3) contain the labour supply index and its two components, measured as z-scores. All intensive measures are weekly total hours by household members, and extensive margins are the number of employed members. Columns (10) and (11) include discouraged unemployed. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. Wam stands per working-age members. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 year old & children <14 years old. Columns (4-5) and (10-11) control for the number of working age members in the household, while columns (6) to (9) control for the number of female and male working age members in the households, separately. IK (2012) and CCT (2014) refer to the rule for bandwidth choice as previously specified. All regressions include city and time fixed-effects and basic controls. *p<0.10; **p<0.05; ***p<0.01

Table A18. Discontinuity Effect of RDP on *Wellbeing* (1): Neighbourhood Quality - Local Linear Regressions

	Household Level dependent variable is:			
	Composite Index	Robberies uncommon	Street Light	Refuse collection
	(1)	(2)	(3)	(4)
RDP	0.015	-0.309	0.096	0.220
	(0.3730)	(0.6398)	(0.6110)	(0.5283)
First Stage F-Statistic	10.26	10.26	10.26	10.26
Obs.	1,036	1,036	1,036	1,036
Bandwidth (IK 2012)	√	√	√	√
RDP	-0.08	0.31	-0.64	-0.02
	(0.5349)	(0.8712)	(0.8894)	(0.7715)
First Stage F-Statistic	9.05	9.05	9.05	9.05
Obs.	489	489	489	489
Bandwidth (CCT 2014)	√	√	√	√

Notes: The table reports local linear regression estimates of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include the neighbourhood quality index and its components in columns (2 to 4). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. They also include a dummy variable for living in informal dwelling in the previous period. IK (2012) and CCT (2014) refer to the rule for bandwidth choice as previously specified. All regressions include city and time fixed-effects and basic controls.*p<0.10; **p<0.05; ***p<0.01

Table A19. Discontinuity Effect of RDP on *Wellbeing* (2): Housing Amenities Local Linear Regressions

	Household Level dependent variable is:				
	Composite Index	# of rooms	Dwelling Quality	HH has electricity	Toilet inside
	(1)	(2)	(3)	(4)	(5)
RDP	-0.93*	-1.018	-1.38*	-0.77	-0.07
	(0.5327)	(0.7623)	(0.8342)	(0.8434)	(0.5007)
First Stage F-Statistic	10.26	10.26	10.26	10.26	10.26
Obs.	1,036	1,036	1,036	1,036	1,036
Bandwidth (IK 2012)	√	√	√	√	√
RDP	-1.59**	-1.56*	-1.81*	-2.22**	-0.71
	(0.6677)	(0.8206)	(0.9325)	(1.0315)	(0.7346)
First Stage F-Statistic	9.05	9.05	9.05	9.05	9.05
Obs.	489	489	489	489	489
Bandwidth (CCT 2014)	√	√	√	√	√

Notes: The table reports local linear regression estimates of the effect of RDP subsidies on measures of urban wellbeing. Dependent variables include a housing amenities index and its components in columns (2) to (5). All variables are measured as z-scores. The first stage instrument Z is a dummy equal 1 for being below the assignment threshold. Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. They also include a dummy variable for living in informal dwelling in the previous period. IK (2012) and CCT (2014) refer to the rule for bandwidth choice as previously specified. All regressions include city and time fixed-effects and basic controls. *p<0.10; **p<0.05; ***p<0.01

Table A20. Discontinuity Effect of RDP on Household Shifts Strategies: Coping Mechanisms & Preferences
Local Linear Regressions

	Household Level dependent variable is:						
	Coping Mechanisms			**Preferences**			
	HH Receives Rent Income	Government Grant	Non-commercial Agriculture	Happier than 10 years ago	HH prefers to stay	HH prefers to stay (age<30)	HH prefers to stay (age>30)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
RDP	0.33	0.003	0.09	-0.68	-0.30	-0.95*	0.17
	(0.2577)	(0.3286)	(0.1530)	(0.4692)	(0.4027)	(0.5231)	(0.4088)
First Stage F-Statistic	9.45	10.23	9.45	8.72	8.52	8.52	8.52
Obs.	1,041	1,015	1,041	967	972	972	972
Bandwidth (IK 2012)	√	√	√	√	√	√	√
RDP	0.29	-0.83*	0.12	-0.80*	-0.88*	-0.42*	-0.36
	(0.2826)	(0.4910)	(0.1591)	(0.4811)	(0.5040)	(0.7393)	(0.4517)
First Stage F-Statistic	9.05	9.24	9.05	10.67	9.99	9.99	9.99
Obs.	489	451	489	419	464	464	464
Bandwidth (CCT 2014)	√	√	√	√	√	√	√

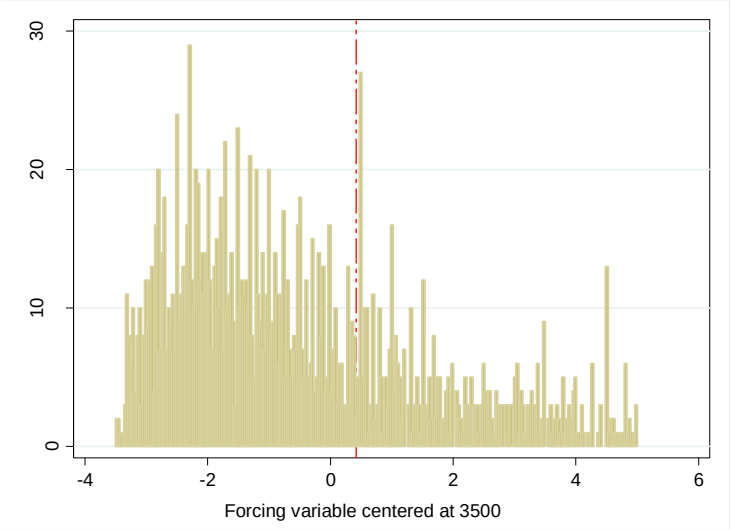
Notes: The table reports local linear regression estimates of the effect of RDP subsidies on household shifts strategies. Dependent variables are different measures of household related to possible coping strategies and preferences. Government Grant in column (2) is a dummy variable for receiving any type of government grant, including disability grants. Non-commercial agriculture is a dummy variable equal to one if households participated in agricultural activities without monetary compensation. While the remainder variables are dummy variables expressing happiness with respect to 10 years ago, and preference to stay in the current place, by age of the household head. Column (5) is the average of columns (6) and (7) Robust standard errors, clustered at household level in parentheses. All regressions control for the age, population group, education level and gender of the household head, as well as the proportion of adults>68 years old & children <14 years old. IK (2012) and CCT (2014) refer to the rule for bandwidth choice as previously specified. All regressions include city and time fixed-effects and basic controls. *p<0.10; **p<0.05; ***p<0.01

Appendix Box A.1

A final issue that could complicate the identification strategy concerns the fact that the assignment variable relies on self-reported income data. Even absent intentional misreporting or manipulation, self-reported income data is noisy. This section aims to evaluate possible bias that could arise from bunching in the assignment variable due to its self-reported nature, i.e. it is easier for individuals to round the income they report (Barreca, Lindo and Waddell 2015).

Below, I show that while there is heaping, it is random across round and non-round numbers. Figure A.5 plots the distribution of the data using one-unit bins, and shows that heaping is random across values with no particular higher density at any multiple of R$10. Particularly, it does not affect density at the cutoff of the value of the running variable (McCrary tests). Further, outcome variables do not seem to present significant variations by heaping types (Figure A.6). The scatter plots show similar distributions for selected outcome variables for heaping or non-heaping values of the running variable, with none being systematically higher.

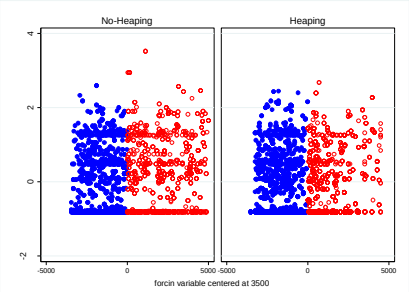
Figure A.7 Bunching in the Running Variable



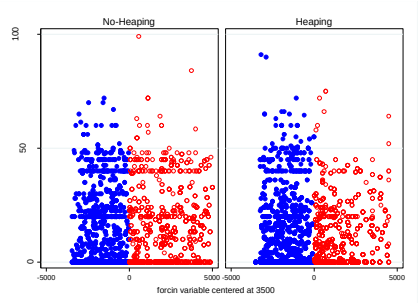
Notes: Obtained using Raj Chetty program bunch_count.ado

Figure A8. Excluding Heaps vs Non-Heaps – Selected Outcome variables

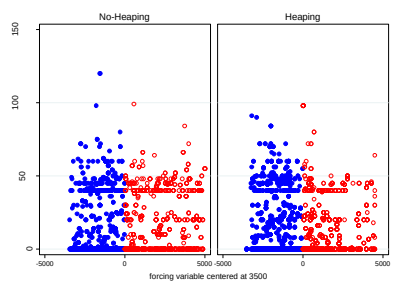
(a) Labour Supply Index



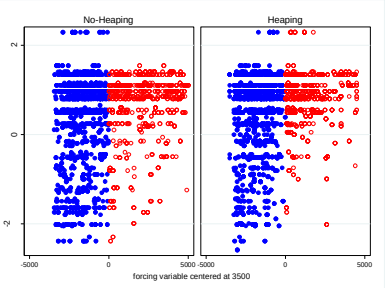
(b) Total Hours by Working Age Member



(c) Total Hours by Working Age Male Member



(d) Amenities Index



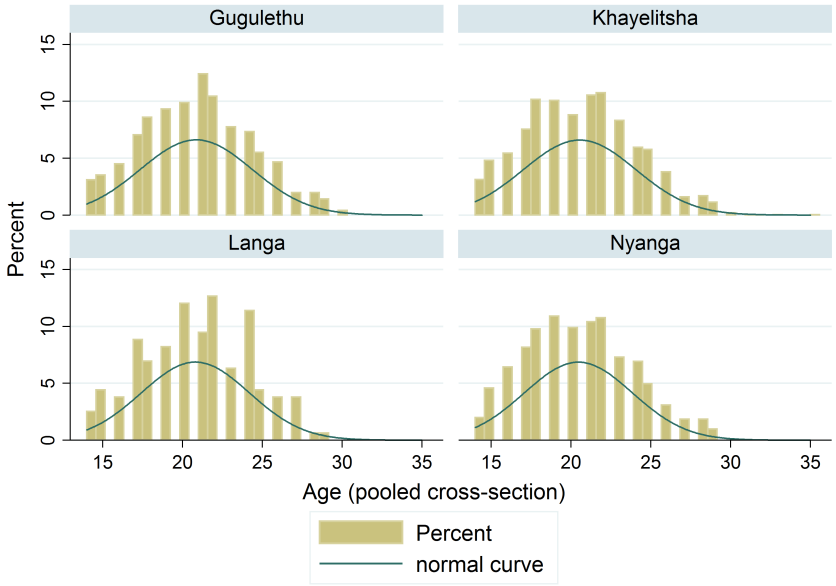
Appendix B.

Appendices to Chapter 2.
Is there one Ghetto University?
Neighbourhoods & Opportunities in a Developing City.

[Left BLANK]

[Left BLANK]

Figure B3. Age Distribution of Young Adults by Ghetto (pooled-cross section, all waves)



Notes: Age distribution across ghettos, final sample dataset of compliers in black townships; CAPS dataset.

Figure B4. Gender Distribution of Young Adults by Sector of Occupations (SIC Classifications)

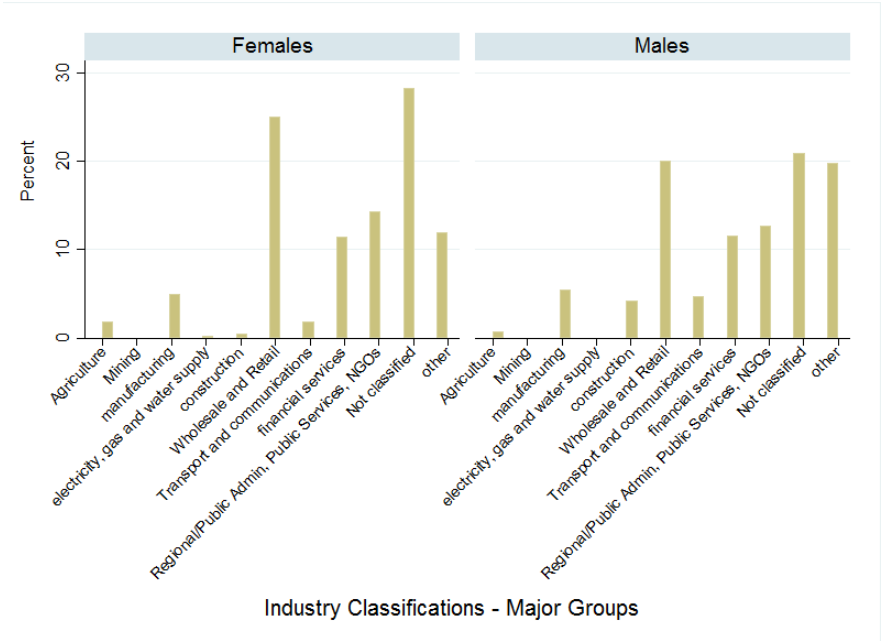
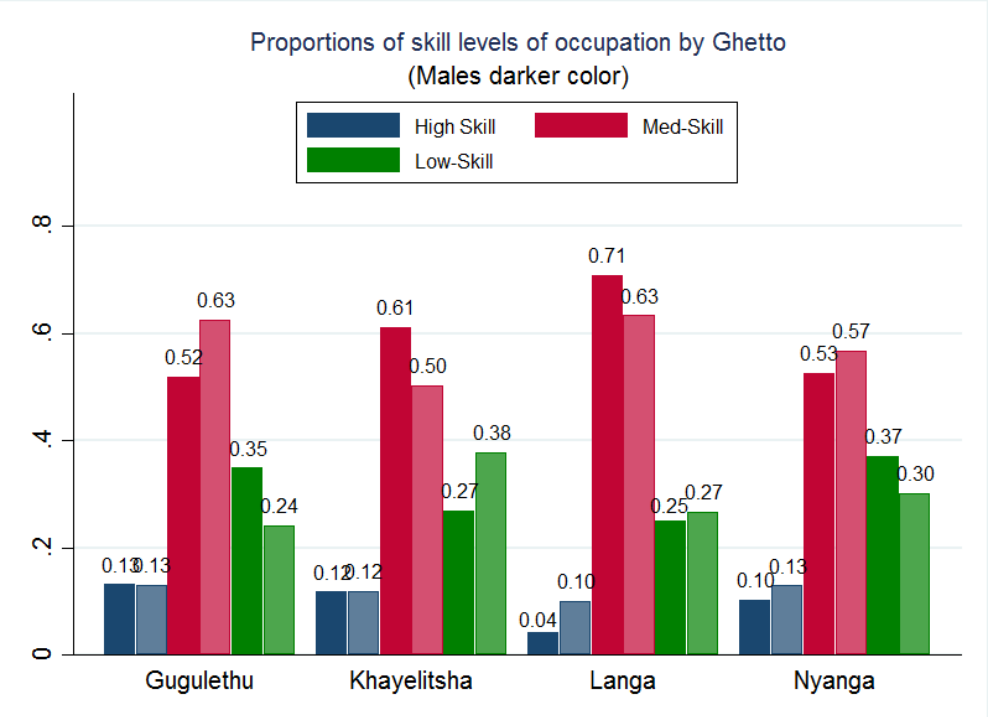


Figure B5. Proportion of skill levels of occupation by Ghetto



Notes: Skill defined according to the classification of SOC occupations by Statistics South Africa.

Table B1. Ghetto Quality Index 2001, all components (z-scores)

	Langa	**Gugulethu**	**Khayelitsha**	**Nyanga**
Electricity for light	-0.994	-0.048	1.377	-0.335
Water inside	-0.002	1.414	-0.738	-0.674
Refuse collection	-1.181	0.342	1.173	-0.334
Formal dwelling	-0.788	1.465	-0.385	-0.293
HHs above poverty line	-0.609	1.181	0.445	-1.018
Adults > matric	0.637	0.423	0.433	-1.493
Adults some school	0.783	-0.641	0.920	-1.062
Km to CBD (inv.)	1.05	0.302	-1.344	-0.008
Km to non-black schools (inv.)	0.820	0.687	-0.153	-1.354
Murder rate (inv.)	1.182	0.137	-0.062	-1.257
Total Index	0.234	0.605	-0.031	-0.808
Total Rank	2	1	3	4

Notes: All variables are standardized census variables from the 2001 census. They are defined as in Table 1. Inv. stands for inverse meaning that the values are substracted by their maximum so that the higher number reflects proximity and lower murder rates.

Section B.I: Separate Ghetto Regressions

Table B2. Results I: Education in young adulthood (separate regressions)

	Years of education	Ever college	Delay in graduating grade
	(1)	(2)	(3)
Langa	0.621*	0.083	0.168
	(0.372)	(0.066)	(0.626)
R-squared	0.040	0.007	0.271
Gugulethu	0.138	0.017	-0.198
	(0.267)	(0.025)	(0.327)
R-squared	0.036	0.001	0.271
Khayelitsha	-0.353*	-0.058**	0.108
	(0.199)	(0.023)	(0.302)
R-squared	0.042	0.017	0.271
Nyanga	0.126	0.030	0.021
	(0.227)	(0.027)	(0.308)
R-squared	0.036	0.004	0.270
All Controls	Y	Y	Y
N	592	600	340

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. All regressions are separate regressions where the main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, gender, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), main language spoken (English, Xhosa and Afrikaans, with one excluded), and the type of place where YA answers spending most of their lives (formal and informal urban, formal and informal rural, with one excluded). The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991); black only *** p<0.01, **p<0.05, * p<0.1

Table B3. Results II: Education in young adulthood by sex (separate regressions)

	Years of education	Ever college	Delay in graduating grade
	(1)	(2)	(3)
Females:			
Langa	0.625	0.179	0.640
	(0.600)	(0.116)	(0.930)
Gugulethu	0.378	0.026	-0.651
	(0.320)	(0.036)	(0.415)
Khayelitsha	-0.395	-0.073**	0.607
	(0.247)	(0.031)	(0.378)
Nyanga	-0.008	0.022	-0.340
	(0.238)	(0.039)	(0.411)
N	319	323	183
Males:			
Langa	0.607	-0.034**	-0.642
	(0.457)	(0.017)	(0.790)
Gugulethu	-0.178	0.012	0.319
	(0.403)	(0.032)	(0.505)
Khayelitsha	-0.262	-0.039	-0.570
	(0.314)	(0.034)	(0.466)
Nyanga	0.293	0.041	0.475
	(0.384)	(0.039)	(0.459)
N	273	277	157
All Controls	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. Controls and sample are as in A2. Regressions are run separately by sex. Each point estimate is from a separate regression where the main explanatory is a dummy variable for the ghettos of residence. *** p<0.01, **p<0.05, * p<0.1

Table B4. Results I: Labour outcomes across period 2004-2009 (separate regressions)

	Earnings (log)	Working (paid)	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
Langa	0.194	0.186*	-0.174*	-0.079	-0.164*
	(0.172)	(0.102)	(0.102)	(0.099)	(0.098)
R-squared	0.022	0.004	0.003	0.002	0.006
Gugulethu	0.015	-0.074	0.081	0.098	0.120*
	(0.105)	(0.063)	(0.064)	(0.066)	(0.067)
R-squared	0.018	0.002	0.003	0.003	0.001
Khayelitsha	0.005	-0.052	0.076	-0.022	-0.011
	(0.086)	(0.055)	(0.055)	(0.053)	(0.056)
R-squared	0.018	0.002	0.003	0.003	0.009
Nyanga	-0.063	0.083	-0.118**	-0.043	-0.057
	(0.090)	(0.057)	(0.057)	(0.054)	(0.058)
R-squared	0.020	0.004	0.008	0.002	0.007
All Controls	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y
N	711	1251	1251	1251	1251

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (1) are log monthly earnings, in column (2) working is a dummy variable for having any paid work, columns (3) and (4) measure short term (ST) and (LT) unemployment defined as not employed and looking for a job below and above 2 months, respectively. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school. All regressions are separate regressions where the main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, gender, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), main language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers spending most of their lives (formal and informal urban, formal and informal rural, with one excluded) and the results of an aptitude quantitative test administered on year one of the survey to all participants. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991); black only. *** p<0.01, **p<0.05, * p<0.1

Table B5. Results I: Labour outcomes in young adulthood (separate regressions)

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Langa	0.264	3.194	3.373	-7.570***	-7.187***
	(0.169)	(4.275)	(2.157)	(1.988)	(2.284)
R-squared	0.097	0.209	0.011	0.102	0.094
Gugulethu	0.046	-1.810	-0.541	3.696*	3.962**
	(0.094)	(1.928)	(0.665)	(1.946)	(1.989)
R-squared	0.091	0.209	0.004	0.101	0.096
Khayelitsha	-0.061	-3.262*	-0.516	-3.463**	-4.251***
	(0.076)	(1.749)	(0.722)	(1.391)	(1.392)
R-squared	0.092	0.213	0.004	0.102	0.101
Nyanga	-0.037	4.561**	0.279	2.519	2.800*
	(0.082)	(1.843)	(0.807)	(1.720)	(1.682)
R-squared	0.091	0.217	0.004	0.097	0.091
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to average log monthly earnings for the last two periods. Sample and controls are as in Table A4. *** p<0.01, **p<0.05, * p<0.1

Table B6. Results II: Labour outcomes across period 2004-2009 by sex (separate regressions)

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Females:					
Langa	0.281	0.153	-0.144	0.009	-0.051
	(0.202)	(0.121)	(0.121)	(0.090)	(0.088)
Gugulethu	-0.162	-0.026	0.028	0.027	0.017
	(0.126)	(0.061)	(0.061)	(0.052)	(0.053)
Khayelitsha	0.044	-0.104*	0.118**	-0.020	0.002
	(0.092)	(0.059)	(0.059)	(0.041)	(0.046)
Nyanga	-0.002	0.104*	-0.124**	-0.002	-0.005
	(0.107)	(0.057)	(0.056)	(0.046)	(0.052)
N	361	678	678	678	678
Males:					
Langa	0.007	0.092	-0.092	-0.089	-0.094
	(0.185)	(0.136)	(0.136)	(0.099)	(0.129)
Gugulethu	0.032	-0.088	0.100	0.116*	0.131*
	(0.098)	(0.066)	(0.067)	(0.064)	(0.068)
Khayelitsha	-0.042	-0.071	0.080	-0.065	-0.017
	(0.097)	(0.057)	(0.057)	(0.058)	(0.057)
Nyanga	0.011	0.133**	-0.155**	-0.008	-0.074
	(0.097)	(0.065)	(0.065)	(0.062)	(0.069)
N	350	573	573	573	573
All Controls	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to log monthly earnings in the final period (2009). Each point estimate is from a separate regression where the main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, gender, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers 'spending most of their lives (formal and informal urban, formal and informal rural, with one excluded) and the results of an aptitude quantitative test administered at the beginning of the survey on year one to all participants. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. *** p<0.01, **p<0.05, * p<0.1

Table B7. Results II: Labour outcomes in young adulthood by sex (separate regressions).

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Females:					
Langa	0.567**	5.224	5.257	-8.530***	-8.711***
	(0.226)	(6.690)	(3.860)	(1.935)	(2.531)
Gugulethu	0.011	-5.383**	-0.349	2.435	2.422
	(0.146)	(2.576)	(1.000)	(2.524)	(2.422)
Khayelitsha	-0.054	-2.644	-1.405	-3.125*	-4.094**
	(0.111)	(2.457)	(1.030)	(1.710)	(1.650)
Nyanga	-0.091	6.348**	0.836	3.581*	3.992**
	(0.123)	(2.556)	(1.207)	(1.904)	(1.855)
N	224	317	295	317	289
Males:					
Langa	-0.047	0.797	1.332	-6.504*	-6.103*
	(0.125)	(4.916)	(1.647)	(3.609)	(3.570)
Gugulethu	0.116	1.677	-0.484	5.122*	5.658*
	(0.110)	(2.759)	(0.812)	(2.912)	(2.973)
Khayelitsha	-0.066	-3.749	0.560	-3.797*	-4.217*
	(0.110)	(2.505)	(0.931)	(2.261)	(2.335)
Nyanga	-0.035	2.339	-0.544	1.069	1.210
	(0.110)	(2.968)	(0.974)	(2.956)	(2.959)
N	222	273	248	273	249
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings corresponds to average log monthly earnings for the last two periods. Each point estimate is from a separate regression where the main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers 'spending most of their lives (formal and informal urban, formal and informal rural, with one excluded) and the results of an aptitude quantitative test administered at the beginning of the survey on year one to all participants. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991), these are Black only. *** p<0.01, **p<0.05, * p<0.1

Table B8. Results I: Skill Sectors across period 2004-2009 (separate regressions)

	High-skill	Med-skill	Low-skill	Low-skill(b)
	(1)	(2)	(3)	(4)
Langa	-0.033	0.150*	0.046	0.036
	(0.029)	(0.077)	(0.072)	(0.071)
R-squared	0.006	0.032	0.033	0.023
Gugulethu	-0.012	-0.039	0.045	0.074**
	(0.024)	(0.041)	(0.034)	(0.036)
R-squared	0.005	0.029	0.034	0.027
Khayelitsha	0.019	-0.065*	-0.076**	-0.092***
	(0.020)	(0.037)	(0.033)	(0.034)
R-squared	0.006	0.031	0.039	0.031
Nyanga	-0.003	0.071*	0.036	0.032
	(0.020)	(0.040)	(0.039)	(0.040)
R-squared	0.005	0.031	0.033	0.023
All Controls	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y
N	768	768	768	768

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha.ll regressions are separate regressions where the main explanatory is a dummy variable for the ghettos of residence. Outcome variables are dummy variables for working in a high-skill occupation (column (1), medium skill occupation (column (2)), and low-skill occupations (columns (3) and (4)). The skill level of occupations is defined according to SSA definitions. Low-skill in column (4) also include armed forces and others. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school. Controls and sample are the same as in Table 10. *** p<0.01, **p<0.05, * p<0.1

Table B9. Results I: Health & Behavioural Outcomes in young adulthood (separate regressions)

	Cigarettes	Alcohol	Other drugs	Ever Pregnant
	(1)	(2)	(3)	(4)
Langa	-0.086	-0.096	0.002	0.0123
	(0.071)	(0.084)	(0.051)	(0.1301)
R-squard	0.218	0.204	0.062	0.167
Gugulethu	0.143**	0.117*	0.000	-0.0599
	(0.053)	(0.063)	(0.025)	(0.0670)
R-squared	0.231	0.210	0.062	0.169
Khayelitsha	-0.093**	-0.124***	-0.010	0.0433
	(0.039)	(0.047)	(0.022)	(0.0625)
R-squared	0.226	0.215	0.063	0.168
Nyanga	0.014	0.069	0.012	-0.0051
	(0.042)	(0.048)	(0.028)	(0.0677)
R-squared	0.217	0.205	0.063	0.167
All Controls	Y	Y	Y	Y
N	535	534	535	317

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in columns (1) to (3) are dummy variable equal to one if YA smokes, drinks alcohol and consumes drugs in the last period. Ever pregnant is a dummy variable if the YA was ever pregnant in the period studied (for females only). Controls and sample are as in Table A1. *** p<0.01, **p<0.05, * p<0.1

Table B10. Results III: Channels, Labour outcomes across period 2004-2009 (separate regressions)

	Earnings (log)	Working (paid)	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)
Social Channels:				
% Black 2001	0.001	-0.004	-0.014	-0.012
	(0.040)	(0.023)	(0.018)	(0.023)
% Black 1996	0.006	0.048	-0.058*	-0.058*
	(0.066)	(0.041)	(0.030)	(0.034)
% Coloured 2001	-0.000	0.004	0.015	0.013
	(0.041)	(0.024)	(0.018)	(0.024)
% Coloured 1996	-0.026	0.023	0.036	0.013
	(0.073)	(0.042)	(0.033)	(0.036)
Access to jobs:				
Access Index 2001	0.105	0.155**	0.014	-0.037
	(0.128)	(0.070)	(0.056)	(0.059)
Access Index 2005	0.119	0.191**	0.018	-0.043
	(0.141)	(0.078)	(0.064)	(0.066)
Access Index 2009	0.026	0.079***	0.013	-0.011
	(0.049)	(0.028)	(0.024)	(0.024)
Km to CBD	-0.004	-0.010***	-0.002	0.001
	(0.007)	(0.004)	(0.003)	(0.003)
Quality of education:				
Former white-col. school	0.230	-0.097	-0.071	-0.035
	(0.169)	(0.081)	(0.072)	(0.084)
Former Black school	-0.192**	0.067	-0.022	-0.035
	(0.081)	(0.049)	(0.047)	(0.048)
Km to closest school	-0.107	-0.119**	0.023	0.056
	(0.101)	(0.054)	(0.060)	(0.057)
Km to white-col. school	-0.032	0.004	0.001	-0.005
	(0.021)	(0.011)	(0.009)	(0.009)
N	711	1251	1251	1251
All Controls	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings are average log monthly earnings for the last two periods. Each point estimates is the result of a separate regression of the channel variable on the outcomes variable. Controls and sample are as in Table A4. *** p<0.01, **p<0.05, * p<0.1

Table B11. Results III: Channels, Labour outcomes in young adulthood (separate regressions)

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Social Channels:					
% Black 2001	-0.017	-0.737	-0.530	-1.560*	-1.569*
	(0.0477)	(0.8160)	(0.4446)	(0.8252)	(0.8172)
% Black 1996	-0.093	0.224	0.215	-3.795**	-3.941**
	(0.0873)	(1.8426)	(0.8907)	(1.5918)	(1.5980)
% Coloured 2001	0.019	0.766	0.530	1.624*	1.634*
	(0.0490)	(0.8418)	(0.4539)	(0.8464)	(0.8388)
% Coloured 1996	0.070	-0.962	1.168	5.427**	5.368**
	(0.1137)	(2.3076)	(1.4238)	(2.1580)	(2.1511)
Access to Jobs:					
Access Index 2001	0.241*	4.524	2.479	-1.239	-0.021
	(0.1398)	(3.3802)	(1.5503)	(2.0638)	(2.3196)
Access Index 2005	0.261*	5.502	2.680	-0.452	0.989
	(0.1550)	(3.7167)	(1.6582)	(2.4070)	(2.6625)
Access Index 2009	0.065	2.244*	0.671	1.363	1.868*
	(0.0552)	(1.2479)	(0.4959)	(0.9625)	(0.9998)
Km to CBD	-0.009	-0.280	-0.092	-0.148	-0.214
	(0.0075)	(0.1715)	(0.0679)	(0.1307)	(0.1373)
Quality of education:					
Former white-col. school	0.102	-3.269	0.134	-3.762**	-3.151*
	(0.1309)	(2.2771)	(1.1390)	(1.6697)	(1.8293)
Former Black school	-0.062	-1.590	0.028	-0.633	-0.804
	(0.0755)	(1.5927)	(0.6978)	(1.4275)	(1.4240)
Km to closest school	-0.166	-4.536*	0.560	-1.020	-1.008
	(0.1336)	(2.5294)	(1.3211)	(2.2097)	(2.0345)
Km to white-col. school	-0.023	0.091	0.436	0.680	0.777
	(0.0195)	(0.4611)	(0.3576)	(0.5227)	(0.5855)
N	446	590	543	590	538
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings are average log monthly earnings for the last two periods. Each point estimates is the result of a separate regression of the channel variable on the outcomes variable. Controls and sample are as in Table A4. *** p<0.01, **p<0.05, * p<0.1

Table B12. Results III: Channels, on Education in young adulthood (separate regressions)

	Years of education	Ever college	Delay in graduating grade
	(1)	(2)	(3)
Social Channels:			
% Black 2001	-0.015	-0.007	0.184**
	(0.0878)	(0.0090)	(0.0904)
% Black 1996	0.195	-0.003	0.036
	(0.1958)	(0.0168)	(0.2894)
% Coloured 200	0.019	0.008	-0.197**
	(0.0925)	(0.0096)	(0.0957)
% Coloured 1996	0.025	0.003	0.134
	(0.2222)	(0.0178)	(0.3686)
Access to jobs:			
Access Index 2001	0.741**	0.100**	-0.114
	(0.3112)	(0.0476)	(0.5078)
Access Index 2005	0.839**	0.116**	-0.183
	(0.3489)	(0.0509)	(0.5622)
Access Index 2009	0.282**	0.041***	-0.103
	(0.1316)	(0.0153)	(0.1925)
Km to CBD	-0.039**	-0.005**	0.013
	(0.0178)	(0.0021)	(0.0262)
Quality of education:			
Former white-col. school	0.902***	0.041	-0.194
	(0.2644)	(0.0412)	(0.3739)
Former Black school	0.282	-0.012	-0.463*
	(0.1796)	(0.0185)	(0.2696)
Subsidy Transport	0.236	-0.001	-0.052
	(0.1954)	(0.0235)	(0.2895)
Km to closest school	-0.185	-0.033**	0.097
	(0.2722)	(0.0150)	(0.3941)
Km to white-col.school	0.017	0.003	0.153**
	(0.0564)	(0.0044)	(0.0665)
N	583	590	338
All Controls	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses. Outcome variables in column (1) is a dummy variable for YA not working and not studying in 2009, years of education (2) is the total years of formal education attained by the end of the period, ever college (3) is a dummy variable for ever attending college over the years in the panel, and delay in graduating grade is the delay in years for completing last observed high-school grade against an upper bound (+1 year) of expected age of graduation. Controls and sample are as in Table A2. Each point estimates is the result of a separate regression of the channel variable on the outcomes variable. *** p<0.01, **p<0.05, * p<0.1

Table B13. Results IV: Labour Outcomes across period 2004-2009 (conditional education).

	Earnings (log)	Working	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
Gugulethu	0.010	0.021	-0.024	0.071	0.054
	(0.077)	(0.048)	(0.048)	(0.044)	(0.043)
Langa	0.133	0.142*	-0.141*	-0.003	-0.055
	(0.135)	(0.085)	(0.084)	(0.067)	(0.074)
Nyanga	0.032	0.127***	-0.148***	0.023	-0.017
	(0.074)	(0.045)	(0.046)	(0.043)	(0.046)
All Controls	Y	Y	Y	Y	Y
Years of Education	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y
R-squared	0.136	0.098	0.104	0.012	0.009
N	704	1238	1238	1238	1238

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) are log monthly earnings, in column (2) working is a dummy variable for having any paid work, columns (3) and (4) measure short term (ST) and (LT) unemployment defined as not employed and looking for a job below and above 2 months, respectively. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school.The main explanatory is a dummy variable for the ghettos of residence. Controls include age, age square, dummy variables for the education of mothers (secondary complete, primary and no education, with one excluded category), language spoken (English, Xhosa and Afrikaans, with one excluded), the type of place where YA answers 'spending most of their lives (formal and informal urban, formal and informal rural, with one excluded) and the results of an aptitude quantitative test administered on year one of the survey to all participants. They also include years of total education completed. The sample is only compliers - i.e. YA residing in a former Apartheid Black ghetto who moved to their neighbourhood of residence before the end of Apartheid's Group Areas Act removal (1991); blacks only. . *** p<0.01, **p<0.05, * p<0.1

Table B14. Results IV: Labour outcomes in young adulthood (conditional on education).

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Gugulethu	0.088	0.867	-0.101	5.270***	5.523***
	(0.090)	(2.166)	(0.762)	(1.970)	(1.993)
Langa	0.183	4.549	3.308	-4.141**	-3.521
	(0.154)	(4.446)	(2.208)	(2.093)	(2.309)
Nyanga	0.006	5.226**	0.442	4.181**	4.769***
	(0.082)	(2.051)	(0.887)	(1.704)	(1.658)
All Controls	Y	Y	Y	Y	Y
Years of Education	Y	Y	Y	Y	Y
R-squared	0.181	0.219	0.009	0.139	0.130
N	440	583	537	583	535

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings are average log monthly earnings in for the last two periods. The main explanatory is a dummy variable for the ghettos of residence. Controls and sample are as in Table A13. *** p<0.01, **p<0.05, * p<0.1

Table B15. Results IV: Labour outcomes across period 2004-2009 by sex (conditional on education).

	Earnings (log)	Working	Idle	Unemployed (ST)	Unemployed (LT)
	(1)	(2)	(3)	(4)	(5)
Females:					
Gugulethu	-0.097	0.034	-0.041	0.035	0.014
	(0.096)	(0.068)	(0.068)	(0.056)	(0.057)
Langa	0.208	0.171	-0.167	0.026	-0.042
	(0.159)	(0.110)	(0.108)	(0.092)	(0.089)
Nyanga	-0.008	0.127**	-0.148**	0.012	-0.003
	(0.097)	(0.064)	(0.063)	(0.049)	(0.056)
N	359	670	670	670	670
Males:					
Gugulethu	0.123	0.010	-0.007	0.115	0.091
	(0.101)	(0.066)	(0.067)	(0.072)	(0.070)
Langa	0.059	0.124	-0.128	-0.043	-0.074
	(0.157)	(0.142)	(0.142)	(0.104)	(0.130)
Nyanga	0.047	0.131**	-0.152**	0.041	-0.034
	(0.106)	(0.065)	(0.065)	(0.069)	(0.072)
N	345	568	568	568	568
All Controls	Y	Y	Y	Y	Y
Years of Education	Y	Y	Y	Y	Y
Time fixed-effects	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (1) are log monthly earnings, in column (2) working is a dummy variable for having any paid work, columns (3) and (4) measure short term (ST) and (LT) unemployment defined as not employed and looking for a job below and above 2 months, respectively. Outcomes are observed across the period 2004, 2005, 2006, 2009 for YA above 20 years old and not enrolled in school. Controls and samples are the same as in Table A13, excluding sex as regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

Table B16. Results IV: Labour outcomes in young adulthood by sex (conditional on education).

	Earnings (log)	Months worked	Months worked since school	Months search	Months search since school
	(1)	(2)	(3)	(4)	(5)
Females:					
Gugulethu	0.018	-3.093	0.472	3.763	4.349*
	(0.128)	(2.883)	(1.058)	(2.586)	(2.494)
Langa	0.352*	5.515	5.697	-6.181***	-5.863**
	(0.206)	(7.012)	(3.930)	(2.172)	(2.664)
Nyanga	-0.049	5.772**	1.367	4.262**	5.191***
	(0.111)	(2.844)	(1.262)	(1.952)	(1.879)
N	222	314	292	314	287
Males:					
Gugulethu	0.168	4.637	-0.743	6.425**	5.997**
	(0.119)	(3.070)	(1.048)	(2.779)	(2.927)
Langa	-0.024	3.328	0.767	-2.555	-1.805
	(0.133)	(5.176)	(1.726)	(3.699)	(3.682)
Nyanga	0.049	4.586	-0.824	3.947	3.640
	(0.114)	(3.199)	(1.164)	(2.919)	(2.988)
N	218	269	245	269	248
All Controls	Y	Y	Y	Y	Y

Notes: Robust standard errors clustered at household level in parentheses; the excluded ghetto is Khayelitsha. Outcome variables in column (2) and (4) are defined as the total over the entire period, in columns (3) and (5) the same variables are defined since last enrolled in formal education institution, earnings are average log monthly earnings for the last two periods. The main explanatory is a dummy variable for the ghettos of residence. Controls and sample are the same as in Table A15, sex is excluded as regressions are run separately by sex. *** p<0.01, **p<0.05, * p<0.1

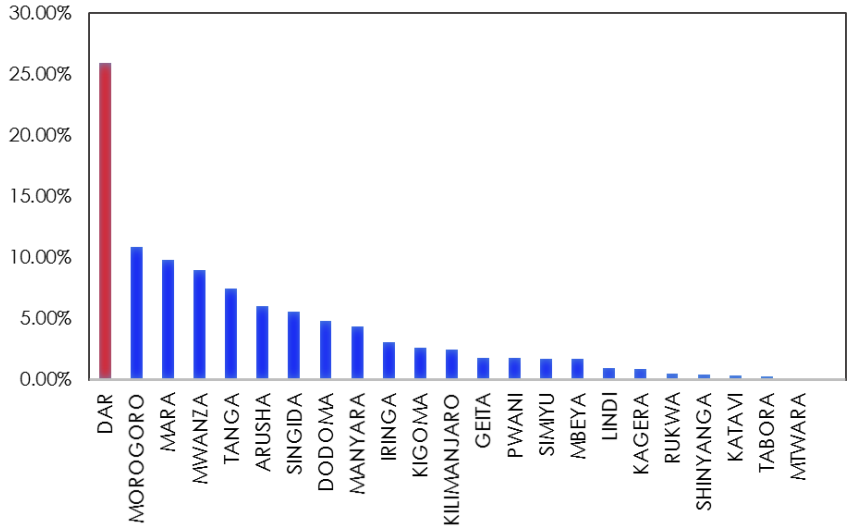
Appendix C.

First Appendices to Chapter 3.

Cholera in Times of Floods.

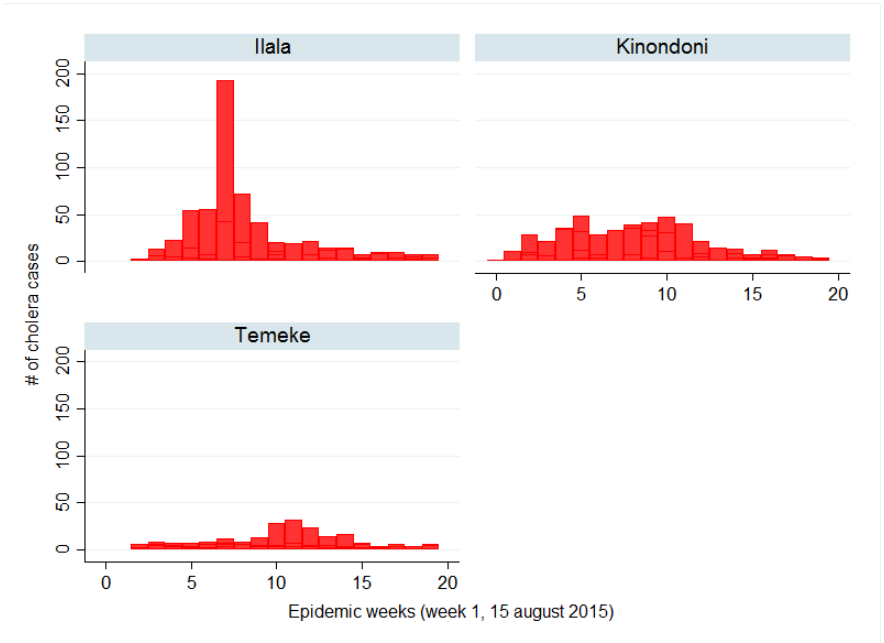
Weather shocks & health in Dar es Salaam.

Figure C1. Cholera cases during 2015-2016 outbreak, by region



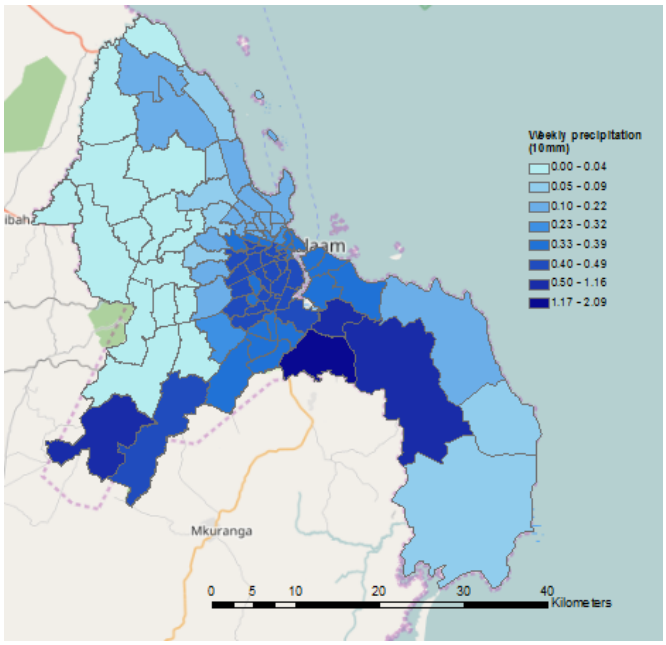
Notes: Data obtained from the Red Cross. Total cases (vs. effective in analysis) up until April 2016.

Figure C2. Distribution of effective cholera cases (epidemic week), by district municipality



Notes: Therea are currently 5 municipal districts in Dar es Salaam. Here we use the three that existed when the cholera outbreak started and at the levels at which the data was collected.

Figure C3. Ward-level weekly rainfall accumulation (area-weighted)



C.I Unweighed Main Regressions

Table C.1: Impact of Weekly Precipitation on Cholera Incidence (unweighted)

	Cholera cases (log)		
	(1)	(2)	(3)
Precipitation	0.0157**	0.0165***	0.0231***
	(0.0060)	(0.0061)	(0.0063)
N	6930	6930	6930
R^2	0.4049	0.4057	0.4638
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.2: Impact of Weekly Precipitation on Cholera Incidence (IV Estimates, unweighted)

	Cholera cases (log)		
	(1)	(2)	(3)
Precipitation	0.0431**	0.0477**	0.0394
	(0.0188)	(0.0194)	(0.0266)
N	6930	6930	6930
First Stage F-test	40.822	40.21	27.985
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Nasa GPM v3 precipitation measurement is instrumented with NASA TRMM measurement. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.3: Impact of Weekly Quartiles of Precipitation on Cholera Incidence (unweighted)

	Cholera cases (log)		
	(1)	(2)	(3)
Q1	-0.0089	-0.0058	-0.0030
	(0.0160)	(0.0157)	(0.0155)
Q3	-0.0301	-0.0250	-0.0359
	(0.0264)	(0.0265)	(0.0269)
Q4	0.1592***	0.1706***	0.1521***
	(0.0521)	(0.0528)	(0.0565)
N	6930	6930	6930
R^2	0.4065	0.4074	0.4645
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. The quartiles of the rainfall distribution are defined as follows: Q1 (0mm), Q2(0-0.29mm), Q3(0.29-2.69mm), Q4(2.69-40.86mm)*p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table C.4: Impact of Flooding on Cholera Incidence (unweighted)

	Cholera cases (log)		
	(1)	(2)	(3)
Panel A:			
Precipitation	0.0154**	0.0162***	0.0227***
	(0.0060)	(0.0061)	(0.0062)
Precipitation × % Flood-prone area	0.0079	0.0066	0.0059
	(0.0051)	(0.0050)	(0.0052)
Panel B:			
Flooded (precipitation ≥75th p)	0.1865***	0.1930***	0.1856***
	(0.0424)	(0.0425)	(0.0456)
Panel C:			
Flooded (precipitation ≥75th p)	0.1688***	0.1757***	0.1658***
	(0.0420)	(0.0421)	(0.0451)
Flooded × %Flood-prone area	0.2181***	0.2113***	0.2136**
	(0.0795)	(0.0786)	(0.0879)
N	6930	6930	6930
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. All panels are independent regressions. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. Flooded is a dummy variable for weekly precipitation falling above the 75th percentile of the total rainfall distribution. Flood-prone area is the total area of the ward that is prone to flooding. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016.*p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.5: Impact of Weekly Precipitation on Cholera Incidence: Infrastructure & Ward Characteristics (unweighted)

	Cholera cases (log)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Precipitation	0.0238***	0.0243***	0.0246**	0.0221***	0.0245	-0.0008	0.0509***	0.0242*	0.0258	0.0063
	(0.0069)	(0.0067)	(0.0121)	(0.0064)	(0.0225)	(0.0128)	(0.0173)	(0.0138)	(0.0253)	(0.0221)
Precipitation × Pop. density	0.0002***							0.0001	0.0000	-0.0008
	(0.0001)							(0.0002)	(0.0003)	(0.0005)
Precipitation × Roads density		0.0014						0.0007	0.0024	-0.0055
		(0.0009)						(0.0016)	(0.0020)	(0.0064)
Precipitation × Footways density			0.0027**					0.0023	0.0024	0.0063
			(0.0012)					(0.0019)	(0.0023)	(0.0055)
Precipitation × # Water wells				0.0013*				-0.0004	-0.0001	-0.0015
				(0.0007)				(0.0012)	(0.0013)	(0.0023)
Precipitation × Drains density					0.0029*				0.0006	0.0046
					(0.0015)				(0.0019)	(0.0038)
Precipitation × % Informal housing						0.0178**				0.0233**
						(0.0072)				(0.0083)
Precipitation × % Formal housing							0.0023			
							(0.0038)			
N	6930	6776	4851	5852	3465	1771	2618	4004	2926	1463
R^2	0.4633	0.4728	0.5160	0.4668	0.5534	0.6435	0.5398	0.5303	0.5858	0.6693
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE										
Municipality × week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Separate regressions in columns (1-7). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Population density is the number of inhabitants per square km (census 2012), Roads density, footway density, and drains density are the meters of roads, footways and drains per km (OSM), % informal and formal houses in the ward are obtained from surveyed plots (not all plots are surveyed).*p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table C.6: Impact of Neighbours' Weekly Precipitation on Cholera Incidence: Spillovers (unweighted)

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation	0.0158**	0.0162***	0.0166***	0.0170***	0.0226***	0.0227***
	(0.0061)	(0.0061)	(0.0062)	(0.0062)	(0.0063)	(0.0062)
Neighbours' precipitation	-0.0000		-0.0001		0.0002	
	(0.0007)		(0.0007)		(0.0008)	
Uphill neighbours' precipitation		-0.0011		-0.0011		-0.0007
		(0.0008)		(0.0008)		(0.0009)
Downhill neighbours' precipitation		0.0008		0.0007		0.0008
		(0.0007)		(0.0007)		(0.0007)
N	6930	6930	6930	6930	6930	6930
R^2	0.4049	0.4052	0.4057	0.4061	0.4638	0.4640
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes		
Municipal time trend			Yes	Yes		
Municipality × week FE					Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Uphill and downhill measures are for neighbouring wards at a higher or lower elevation than the given ward. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

C.II Main Regressions, Spatial Auto-correlation of Standard Errors (Conley HAC SE)

Table C.7: Impact of Weekly Precipitation on Cholera Incidence (HAC SE)

	Cholera cases (log)	
	(1)	(2)
Precipitation	0.0157***	0.0231***
	(0.0051)	(0.0080)
N	6930	6930
R^2	0.0015	0.0049
Ward FE	Yes	Yes
Week FE	Yes	
Municipality × week FE		Yes

Notes: Conley HAC standard errors in parenthesis (Conley 1999, 2008). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.8: Impact of Weekly Quartiles of Precipitation on Cholera Incidence (HAC SE)

	Cholera cases (log)	
	(1)	(2)
Q1	-0.0089	-0.0030
	(0.0327)	(0.0259)
Q3	-0.0301	-0.0359*
	(0.0211)	(0.0193)
Q4	0.1592***	0.1521***
	(0.0508)	(0.0567)
N	6930	6930
R^2	0.0043	0.0062
Ward FE	Yes	Yes
Week FE	Yes	
Municipality × week FE		Yes

Notes: Conley HAC standard errors in parenthesis (Conley 1999, 2008). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. The quartiles of the rainfall distribution are defined as follows: Q1 (0mm), Q2(0-0.29mm), Q3(0.29-2.69mm), Q4(2.69-40.86mm). *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.9: Impact of Flooding on Cholera Incidence (HAC SE)

	Cholera cases (log)	
	(1)	(2)
Panel A:		
Precipitation	0.0154***	0.0227***
	(0.0051)	(0.0072)
Precipitation × % Flood-prone area	0.0079	0.0059
	(0.0223)	(0.0169)
Panel B:		
Flooded (precipitation ≥75th p)	0.1865***	0.1856***
	(0.0455)	(0.0541)
Panel C:		
Flooded (precipitation ≥75th p)	0.1688***	0.1658***
	(0.0430)	(0.0495)
Flooded × % Flood-prone area	0.2181	0.2136
	(0.2266)	(0.2035)
N	6930	6930
Ward FE	Yes	Yes
Week FE	Yes	
Municipality × week FE		Yes

Notes: Conley HAC standard errors in parenthesis (Conley 1999, 2008). All panels are independent regressions. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. Flooded is a dummy variable for weekly precipitation falling above the 75th percentile of the total rainfall distribution. Flood-prone area is the total area of the ward that is prone to flooding. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.10: Impact of Weekly Precipitation on Cholera Incidence: Infrastructure & Ward Characteristics (HAC SE)

	Cholera cases (log)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Precipitation	0.0222***	0.0243***	0.0246**	0.0221***	0.0245	-0.0008	0.0509***	0.0242*	0.0258	0.0063
	(0.0070)	(0.0085)	(0.0102)	(0.0083)	(0.0249)	(0.0200)	(0.0126)	(0.0133)	(0.0266)	(0.0281)
Precipitation × Pop. density	0.0002							0.0001	0.0000	-0.0008
	(0.0003)							(0.0001)	(0.0001)	(0.0011)
Precipitation × Roads density		0.0014						0.0007	0.0024	-0.0055
		(0.0017)						(0.0007)	(0.0016)	(0.0074)
Precipitation × Footways density			0.0027					0.0023	0.0024	0.0063
			(0.0048)					(0.0053)	(0.0033)	(0.0103)
Precipitation × # Water wells				0.0013				-0.0004	-0.0001	-0.0015
				(0.0026)				(0.0011)	(0.0011)	(0.0011)
Precipitation × drains density					0.0029				0.0006	0.0046
					(0.0064)				(0.0040)	(0.0086)
Precipitation × % Informal housing						0.0178				0.0233
						(0.0202)				(0.0202)
Precipitation × % Formal housing							0.0023			
							(0.0058)			
N	6930	6776	4851	5852	3465	1771	2618	4004	2926	1463
R^2	0.0056	0.0053	0.0026	0.0059	0.0022	0.0062	0.0157	0.0037	0.0025	0.0069
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE										
Municipality × week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes:Conley HAC standard errors in parenthesis (Conley 1999, 2008). Separate regressions in columns (1-7). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Roads density, footway density, and drains density are the meters of roads, footways and drains per km (OSM), % informal and formal houses in the ward are obtained from surveyed plots (not all plots are surveyed). *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

C.III Extensions

Table C.11: Impact of Neighbours' Weekly Lagged Precipitation on Cholera Incidence (1)

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation	0.0285**	0.0288**	0.0292**	0.0293**	0.0346***	0.0335***
	(0.0115)	(0.0113)	(0.0117)	(0.0115)	(0.0111)	(0.0110)
Neighbours precipitation	-0.0046	-0.0049	-0.0044	-0.0045	-0.0007	0.0008
	(0.0036)	(0.0047)	(0.0036)	(0.0047)	(0.0034)	(0.0047)
Neighbours precipitation $_{w-1}$	0.0005	0.0007	0.0005	0.0005	0.0002	-0.0007
	(0.0003)	(0.0013)	(0.0003)	(0.0013)	(0.0003)	(0.0013)
Neighbours precipitation $_{w-2}$		-0.0001		-0.0001		0.0007
		(0.0011)		(0.0011)		(0.0011)
N	6929	6924	6929	6924	6929	6924
R^2	0.4493	0.4493	0.4504	0.4504	0.5255	0.5256
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes		
Municipal time trend			Yes	Yes		
Municipality × week FE					Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Neighbour's Precipitation$_{w-n}$ are the lags of weekly accumulated rainfall in neighbouring wards up to n weeks. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table C.12: Impact of Neighbours' Weekly Lagged Precipitation by Elevation on Cholera Incidence (2)

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation	0.0218**	0.0214**	0.0233**	0.0230**	0.0355***	0.0354***
	(0.0097)	(0.0105)	(0.0100)	(0.0108)	(0.0098)	(0.0105)
Uphill neighbours precipitation	-0.0042	-0.0044	-0.0047	-0.0047	-0.0052	-0.0049
	(0.0051)	(0.0048)	(0.0052)	(0.0049)	(0.0051)	(0.0050)
Downhill neighbours precipitation	-0.0032	-0.0033	-0.0036	-0.0037	-0.0047	-0.0045
	(0.0052)	(0.0051)	(0.0053)	(0.0051)	(0.0051)	(0.0051)
Uphill N's precipitation $_{w-1}$	0.0037	0.0042	0.0040	0.0044	0.0055	0.0046
	(0.0055)	(0.0069)	(0.0055)	(0.0069)	(0.0053)	(0.0070)
Downhill N's precipitation$_{w-1}$	0.0043	0.0062	0.0047	0.0063	0.0062	0.0070
	(0.0054)	(0.0073)	(0.0054)	(0.0072)	(0.0052)	(0.0073)
Uphill N's precipitation$_{w-2}$		-0.0004		-0.0002		0.0007
		(0.0042)		(0.0043)		(0.0042)
Downhill N's precipitation$_{w-2}$		-0.0018		-0.0016		-0.0012
		(0.0043)		(0.0043)		(0.0043)
N	6929	6924	6929	6924	6929	6924
R^2	0.4494	0.4495	0.4505	0.4506	0.5257	0.5258
Ward FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes		
Municipal time trend			Yes	Yes		
Municipality × week FE					Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Uphill and downhill measures are for neighbouring wards at a higher or lower elevation than the given ward. N's Precipitation Uphill/Downhill $_{w-n}$ are the lags of weekly accumulated rainfall in neighbouring wards up to n weeks according to their elevation with respect to the given ward. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table C.13: Impact of Neighbours' Weekly Precipitation on Cholera Incidence: Spillovers (HAC SE) (1)

	Cholera cases (log)			
	(1)	(2)	(3)	(4)
Precipitation	0.0158***	0.0162***	0.0226***	0.0227***
	(0.0047)	(0.0047)	(0.0075)	(0.0076)
Neighbours precipitation	-0.0000		0.0002	
	(0.0007)		(0.0007)	
Uphill neighbours precipitation		-0.0011		-0.0007
		(0.0009)		(0.0010)
Downhill neighbours precipitation		0.0008		0.0008
		(0.0009)		(0.0008)
N	6930	6930	6930	6930
R^2	0.0015	0.0021	0.0049	0.0053
Ward FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes		
Municipality × week FE			Yes	Yes

Notes:Conley HAC standard errors in parenthesis (Conley 1999, 2008). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Uphill and downhill measures are for neighbouring wards at a higher or lower elevation than the given ward. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.14: Impact of Neighbours' Weekly Lagged Precipitation on Cholera Incidence: Spillovers (HAC SE) (2)

	Cholera cases (log)			
	(1)	(2)	(3)	(4)
Precipitation	0.0189***	0.0187***	0.0259***	0.0257***
	(0.0051)	(0.0060)	(0.0078)	(0.0083)
Uphill neighbours precipitation	-0.0043	-0.0044	-0.0050	-0.0050
	(0.0037)	(0.0042)	(0.0033)	(0.0051)
Downhill neighbours precipitation	-0.0032	-0.0033	-0.0042	-0.0044
	(0.0032)	(0.0051)	(0.0042)	(0.0062)
Uphill N's precipitation $_{w-1}$	0.0033	0.0033	0.0044	0.0046
	(0.0036)	(0.0108)	(0.0034)	(0.0115)
Downhill N's precipitation $_{w-1}$	0.0041	0.0052	0.0052	0.0067
	(0.0031)	(0.0117)	(0.0043)	(0.0124)
Uphill N's precipitation $_{w-2}$		0.0000		-0.0001
		(0.0079)		(0.0071)
Downhill N's precipitation $_{w-2}$		-0.0010		-0.0015
		(0.0076)		(0.0069)
N	6929	6924	6929	6924
R^2	0.0022	0.0023	0.0056	0.0057
Ward FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes		
Municipality × week FE			Yes	Yes

Notes:Conley HAC standard errors in parenthesis (Conley 1999, 2008). Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. Neighbours' precipitation measures weekly accumulated rainfall in a neighbouring ward. Uphill and downhill measures are for neighbouring wards at a higher or lower elevation than the given ward. N's Precipitation Uphill/Downhill $_{w-n}$ are the lags of weekly accumulated rainfall in neighbouring wards up to n weeks according to their elevation with respect to the given ward. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Table C.15: Lagged Dependent Variable Model

	Cholera cases (log)					
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation	0.0199***	0.0155***	0.0119**	0.0111**	0.0103**	0.0103**
	(0.0054)	(0.0051)	(0.0051)	(0.0051)	(0.0051)	(0.0051)
Cholera cases $_{w-1}$	0.6111***	0.4430***	0.4037***	0.3943***	0.3913***	0.3914***
	(0.0457)	(0.0336)	(0.0348)	(0.0360)	(0.0368)	(0.0371)
Cholera cases $_{w-2}$		0.2747***	0.2112***	0.1975***	0.1922***	0.1923***
		(0.0230)	(0.0285)	(0.0268)	(0.0281)	(0.0283)
Cholera cases $_{w-3}$			0.1435***	0.1172***	0.1085***	0.1087***
			(0.0206)	(0.0245)	(0.0255)	(0.0257)
Cholera cases $_{w-4}$				0.0652***	0.0476**	0.0479**
				(0.0208)	(0.0203)	(0.0197)
Cholera cases $_{w-5}$					0.0447**	0.0455**
					(0.0214)	(0.0217)
Cholera cases $_{w-6}$						-0.0019
						(0.0185)
N	6930	6930	6930	6930	6930	6930
R^2	0.6837	0.7076	0.7136	0.7148	0.7153	0.7153
Ward FE						
Municipality × week FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the log of effective (tested positive) weekly cholera cases in a given ward. All regressions control for weekly ward air temperature; they are weighted by the population of the ward (census 2012). The period coverered is from the first week of March 2015 to the first week of September 2016. Cholera cases$_{w-n}$ are lagged effective cholera cases up week n. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

C.IV Robustness check: Linear-Linear Regressions

Table C.16: Impact of Weekly Precipitation on Cholera Incidence (per 10,000s people)

	Cholera Ward Incidence Rate (per 10,000s)		
	(1)	(2)	(3)
Precipitation	0.0272***	0.0280***	0.0359***
	(0.0081)	(0.0082)	(0.0099)
N	6930	6930	6930
R^2	0.1908	0.1915	0.2367
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the number of weekly cholera cases (tested positive) divided by the population of the ward (10,000s), i.e. cholera cases every 10 thousand people in a ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. *p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.17: Impact of Weekly Quartiles of Precipitation on Cholera Incidence (per 10,000s people)

	Cholera Ward Incidence Rate (per 10,000s)		
	(1)	(2)	(3)
Q1	0.0196	0.0235	0.0152
	(0.0152)	(0.0155)	(0.0163)
Q3	-0.0215	-0.0158	-0.0216
	(0.0250)	(0.0248)	(0.0276)
Q4	0.0946**	0.1069**	0.0862*
	(0.0462)	(0.0473)	(0.0509)
N	6930	6930	6930
R^2	0.1899	0.1905	0.2349
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the number of weekly cholera cases (tested positive) divided by the population of the ward (10,000s), i.e. cholera cases every 10 thousand people in a ward. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016. The quartiles of the rainfall distribution are defined as follows: Q1 (0mm), Q2(0-0.29mm), Q3(0.29-2.69mm), Q4(2.69-40.86mm)*p ≤ 0.10 ** p≤0.05 *** p≤0.01

Table C.18: Impact of Flooding on Cholera Incidence (per 10,000s people)

	Cholera Ward Incidence Rate (per 1000s)		
	(1)	(2)	(3)
Panel A:			
Precipitation	0.0268***	0.0276***	0.0354***
	(0.0080)	(0.0081)	(0.0097)
Precipitation × % Flood-prone area	0.0094	0.0080	0.0078
	(0.0141)	(0.0138)	(0.0139)
Panel B:			
Flooded (precipitation ≥75th p)	0.1164***	0.1231***	0.1077***
	(0.0344)	(0.0351)	(0.0395)
Panel C:			
Flooded (precipitation ≥75th p)	0.1004***	0.1074***	0.0877**
	(0.0350)	(0.0353)	(0.0389)
Flooded × %Flood-prone area	0.1981	0.1910	0.2162
	(0.1471)	(0.1451)	(0.1480)
N	6930	6930	6930
R^2	0.1901	0.1908	0.2352
Ward FE	Yes	Yes	Yes
Week FE	Yes	Yes	
Municipal time trend		Yes	
Municipality × week FE			Yes

Notes: Robust standard errors clustered at the ward level in parenthesis. All panels are independent regressions. Precipitation is measured as the weekly accumulated rainfall in a given ward (10mm units), cholera cases are the number of weekly cholera cases (tested positive) divided by the population of the ward (10,000s), i.e. cholera cases every 10 thousand people in a ward. Flooded is a dummy variable for weekly precipitation falling above the 75th percentile of the total rainfall distribution. Flood-prone area is the total area of the ward that is prone to flooding. All regressions control for weekly ward air temperature. The period coverered is from the first week of March 2015 to the first week of September 2016.*p ≤ 0.10 ** p≤0.05 *** p≤0.01

Appendix D.

Second Appendices to Chapter 3.

Cholera in Times of Floods.

Weather shocks & health in Dar es Salaam.

Tanzania 2015-16 DHS Data Analysis

This section presents the results of the analysis of the 2015-16 Tanzania Demographic and Health Survey and Malaria Indicator Survey (DHS). Its objective is to shed light on the relationship, if any, between income, wealth, and incidence of diarrhoeal diseases in Dar-es-Salaam. The 2015-16 DHS collected reliable information on several demographic and health indicators, including infant and child mortality, nutritional status of mothers and children, and childhood immunizations and diseases. It was implemented by several government agencies with financial support from various bilateral and multilateral donors. It is representative at the national, urban and rural area levels.

We restrict the sample to all the households living in the Dar-es-Salaam region. The DHS only asked children under the age of five questions related to diarrhoea. This leaves us with a sample of 367 children living in 97 distinct households. The DHS calculates for each household a wealth index using a battery of socio-economic variables. The construction of this index goes as follows. Respondent households are given scores based on the number and kinds of consumer goods they own, such as a television or a fridge. Housing characteristics, such as access to drinking water, toilet facilities, and flooring materials are also taken into account. These scores are derived using principal component analysis. Each household is assigned a household wealth score index. National wealth quintiles are then compiled by dividing the index distribution into five equal categories, containing 20% of the population each. The primary objective of this section is to examine the relationship between this wealth index and the incidence of diarrhoea among children aged less than five years.

We conduct a simple regression analysis and estimate the linear probability model in Equation (d.1) below with least squares:

$$D_{ih} = \beta_0 + \beta_1.W_{ih} + \beta_2.X_i + \beta_3.Z_h + \epsilon_{ih} \tag{d.1}$$

where D_{ih} indicates whether child i in household h has had diarrhoea in the last two weeks. W_{ih} is the household wealth index either measured as five quantile dummies or the continuous index value. X_i is a vector of individual covariates and includes age and a female gender dummy. Mother educational attainment is controlled for in vector Z_h. ϵ_{ih} is the error and β_k are the parameters to be estimated. Standard errors are clustered at the household level.

Table D.1 presents summary statistics of the variables included in the regression analysis. 16.3% of the children in the sample report a diarrhoea episode in the two weeks prior to interview. The average child is 1.8 years old. 48% of the sample is comprised of young girls. Regression estimates of Equation 1 are shown in Table D.2. The continuous wealth index score is introduced in column 1. The wealth index quintile dummies are included in the second column. Overall the table presents very weak evidence in favour of a wealth bias regarding diarrheal risk. The continuous wealth index score has a negative but insignificant association with the probability of a child getting diarrhoea. The point estimates of the second column indicate that only children of the third quintiles are more likely to get sick than the poorest children. All other quintile coefficients are insignificant at conventional levels of significance.

Table D.1: Descriptive Characteristics DHS Analysis

	Mean	Std. Dev.	Min.	Max.	N
Diarrhoea in last two weeks	367	0.163	0.37	0	1
Urban wealth index (/1000) - continuous score	367	56.985	66.163	-211.621	186.827
Urban wealth index Q1	367	0.014	0.116	0	1
Urban wealth index Q2	367	0.153	0.36	0	1
Urban wealth index Q3	367	0.259	0.439	0	1
Urban wealth index Q4	367	0.297	0.458	0	1
Urban wealth index Q5	367	0.278	0.449	0	1
Age	367	1.845	1.357	0	4
Female	367	0.48	0.5	0	1
Mother education: no education	367	0.065	0.248	0	1
Mother education: incomplete primary	367	0.054	0.227	0	1
Mother education: complete primary	367	0.534	0.5	0	1
Mother education: incomplete secondary	367	0.057	0.233	0	1
Mother education: complete secondary	367	0.243	0.429	0	1
Mother education: higher education	367	0.046	0.21	0	1

Notes:DHS 2015-16 data for Dar es Salaam region. Under five years old children in the sample.

Table D.2: DHS Regression Analysis

	Had Diarrhea =1	
Dependent variable	(1)	(2)
Urban wealth index - continuous score	-0.000295	
	-0.000328	
Urban wealth index Q2		0.0657
		-0.0546
Urban wealth index Q3		0.134**
		-0.0647
Urban wealth index Q4		0.0694
		-0.0656
Urban wealth index Q5		0.0363
		-0.0694
Age	-0.0501***	-0.0500***
	-0.0122	-0.0124
Female	0.026	0.0259
	-0.0383	-0.0378
Mother education: incomplete primary	0.138	0.0957
	-0.085	-0.091
Mother education: complete primary	0.155***	0.112***
	-0.0415	-0.0391
Mother education: incomplete secondary	0.618***	0.565***
	-0.105	-0.105
Mother education: complete secondary	0.100*	0.0677
	-0.0596	-0.0579
Mother education: higher education	0.13	0.0959
	-0.0949	-0.0879
R-squared	0.142	0.15
Observations	367	367

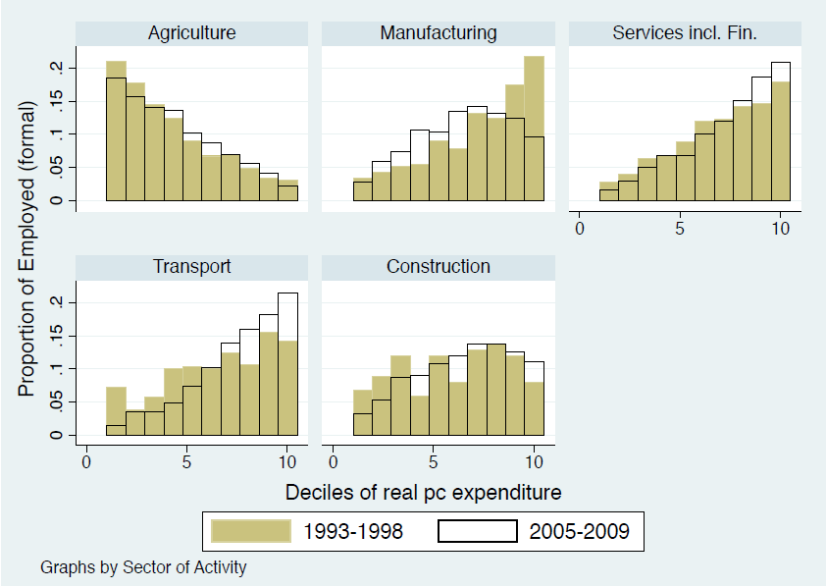
Notes: Robust standard errors clustered at the household level in parenthesis. Linear probability model regressions. DHS 2015-16 data for Dar es Salaam region. Under five years old children in the sample. *p $\leq$ 0.10 ** p$\leq$0.05 *** p$\leq$0.01

Appendix E.

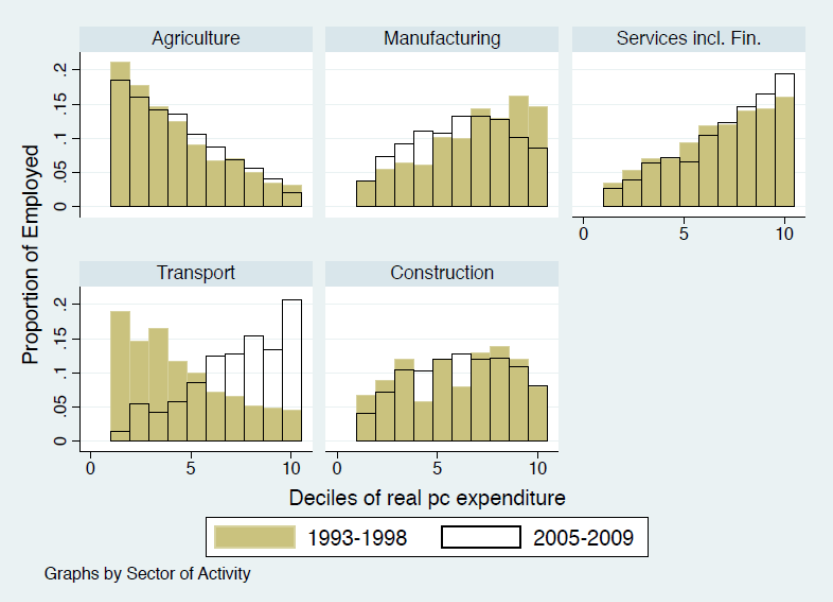
Appendices to Chapter 4. Who Really Benefits from Export Processing Zones? Evidence from Nicaraguan Municipalities.

Figure E1. Proportion of Employed by Economic Sector across the Expenditure Distribution, by EPZ status

EPZ - Municipalities

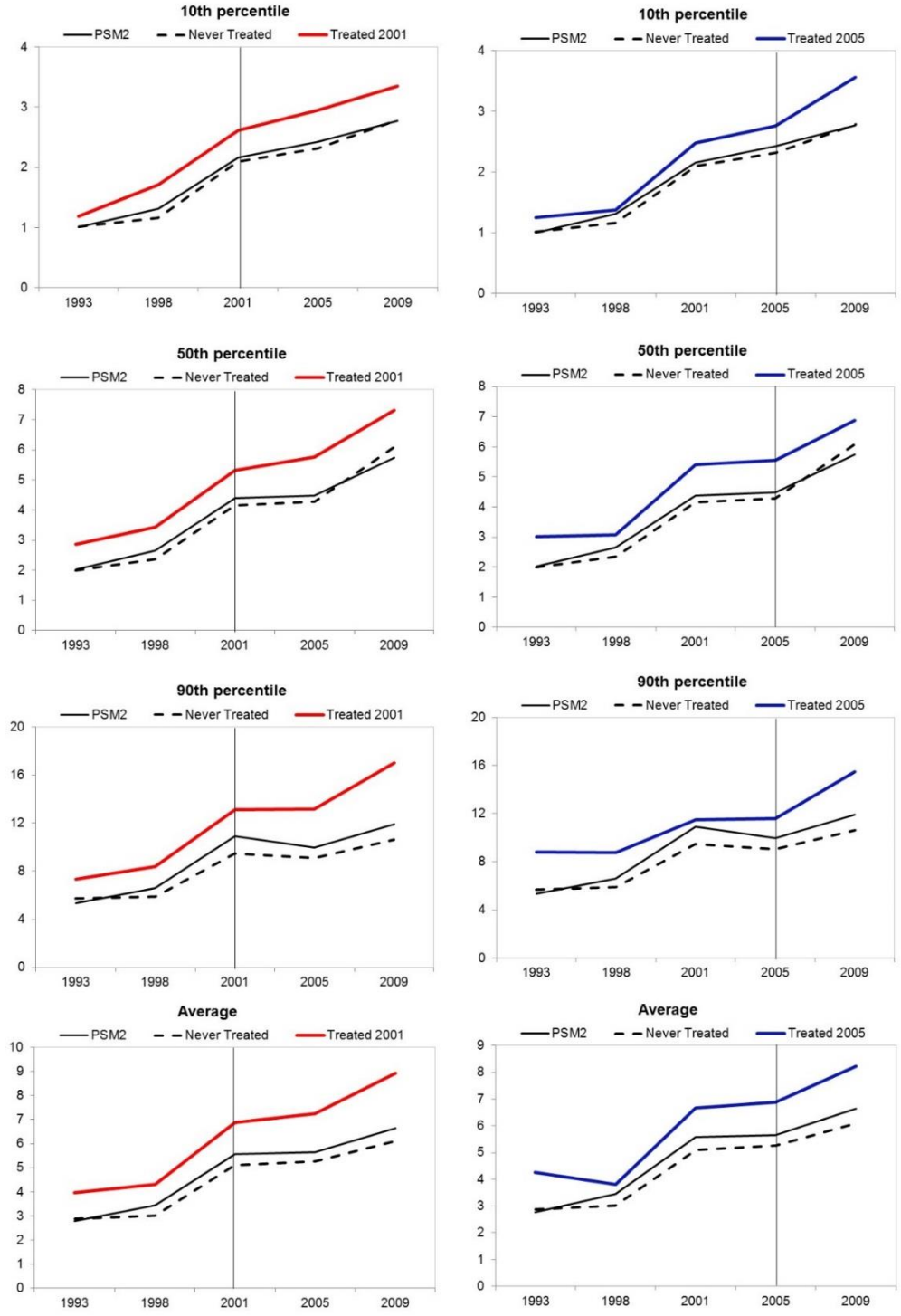


Never EPZ - Municipalities



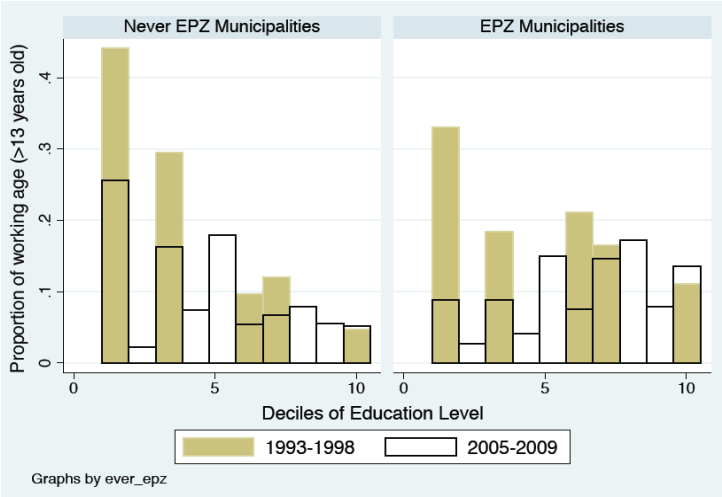
Notes: The figures show the evolution of the distribution of employment by sectors across the expenditure distribution, by treatment status.

Figure E2. Parallel Trends – Selected Outcome Variables, Propensity Score Matching



Notes: Figures depict parallel trends of levels of real expenditure per capita according to treatment and control groups. PSM2 corresponds to control municipalities selected according to propensity score matching. On the left side, the treated include only those treated in 2001, and on the right side the treated are only those treated in 2005. Values are in

Figure E3. Skill Distribution by Treatment Status



Notes: The figure shows the distribution of the working age across the skill distribution (defined by years of education), across the period and according to treatment and control municipalities.

Table E1. Household and Individual Characteristics across selected deciles, by Treatment Status 1993

	1st Decile		3rd Decile		5th Decile		7th Decile		9th Decile	
	EPZ	No	EPZ	No	EPZ	No	EPZ	No	EPZ	No
Household size	8.65	9.14	7.24	7.46	7.06	7.12	6.48	6.94	5.64	5.79
Individuals at working age	0.44	0.42	0.49	0.45	0.55	0.48	0.56	0.52	0.61	0.56
Age (average)	19.15	18.65	21.03	19.88	22.82	20.16	23.38	21.75	25.96	23.61
Population under 25	0.73	0.73	0.67	0.70	0.64	0.68	0.61	0.66	0.54	0.60
Years of education (average)	3.34	3.14	4.07	3.50	3.81	4.42	5.49	4.25	6.15	4.87
Illiterate	0.48	0.66	0.36	0.57	0.30	0.52	0.26	0.38	0.22	0.34
Unemployed	0.06	0.01	0.06	0.02	0.06	0.01	0.06	0.03	0.05	0.03
Informal (no social security)	0.93	0.97	0.75	0.94	0.69	0.86	0.61	0.80	0.55	0.74
Electricity at household level	0.54	0.21	0.90	0.34	0.97	0.44	0.97	0.64	0.98	0.80
Employed in agriculture	NA		NA		NA		NA		NA	
Employed in manufacturing	0.18	0.25	0.27	0.24	0.29	0.26	0.27	0.28	0.26	0.29
Urban	0.44	0.14	0.81	0.23	0.79	0.32	0.90	0.47	0.96	0.68
N=	*3904*		*1785*		*1714*		*2006*		*2757*	

Notes: The data shows main characteristics across selected deciles of the expenditure distribution. The sample size N is limited to working age individuals above 15 years.

Table E2. Household and Individual Characteristics across selected deciles, by Treatment status - 2009

	1st Decile		3rd Decile		5th Decile		7th Decile		9th Decile	
	EPZ	No	EPZ	No	EPZ	No	EPZ	No	EPZ	No
Household size	8.15	8.90	7.10	6.74	5.84	5.87	5.09	5.38	4.56	4.60
Individuals at working age	0.55	0.51	0.60	0.57	0.63	0.59	0.67	0.65	0.71	0.66
Age (average)	22.42	20.81	24.04	22.23	24.28	25.96	26.51	25.33	29.71	28.80
Population under 25	0.64	0.69	0.59	0.65	0.55	0.56	0.52	0.54	0.46	0.51
Years of education (average)	5.18	3.21	6.86	4.45	7.63	4.95	8.49	6.12	10.13	7.27
Illiterate	0.19	0.35	0.10	0.20	0.07	0.20	0.06	0.13	0.03	0.12
Unemployed	0.10	0.04	0.09	0.04	0.07	0.04	0.07	0.03	0.06	0.04
Informal (no social security)	NA		NA		NA		NA		NA	
Electricity at household level	0.99	0.91	0.99	0.92	0.99	0.92	0.99	0.93	1.00	0.94
Employed in agriculture	0.06	0.27	0.02	0.22	0.02	0.22	0.017	0.20	0.01	0.14
Employed in manufacturing	0.26	0.16	0.32	0.14	0.37	0.15	0.39	0.22	0.43	0.28
Urban	0.79	0.12	0.90	0.31	0.92	0.30	0.92	0.44	0.94	0.51
N=	*2901*		*2727*		*3059*		*2107*		*3515*	

Notes: The data shows main characteristics across selected deciles of the expenditure distribution. The sample size N is limited to working age individuals above 15 years.

Table E3. Quality of Propensity Score Matching (PSM)

Sample	Group 1 [<1999]			Group 2 [2000-2001]			Group 3 [2002-2005]			Group 4 [2006-2009]		
Treated	4			5			7			2		
Control	70			67			65			70		
	Ps R2	LR chi2	p>chi2	Ps R2	LR chi2	p>chi2	Ps R2	LR chi2	p>chi2	Ps R2	LR chi2	p>chi2
Matched	0.529	8.89	0.261	0.203	0.203	0.097	0.073	10.26	0.418	0.502	21.14	0.007
RN Control	0.502	7.98	0.335	0.268	0.268	0.014	0.088	12.21	0.271	0.337	14.53	0.043

Notes: In the matching exercise, the set of variables used are pre-treatment municipality characteristics (percent of urban population, percent of illiterate population, percent completed primary level, percent employed in agriculture, manufacturing and services, percent of unemployed, percent working age population, percent of households with access to electricity, proportion of paved roads, distances to the coast or large body of water, main trade access quality (as defined in table 4), landlocked, distance to capital city, pre-trend of average per capita expenditure levels, share of high skill (definition 1) in municipality. All observations are on support.

Table E4. Balance of covariates for the pre-treatment period, by Granting Sequence – PSM Control Group

Percentages (unless otherwise indicated)	Group 1 [<1999]	P|SM Control	t-test	Group 2 [2000-2001]	P|SM Control	t-test	Group 3 [2002-2005]	P|SM Control	t-test	Group 4 [2006-2009]	P|SM Control	t-test
B. Municipality Characteristics												
Distance To Managua (km)	72.54	104.03	-0.80	42.43	103.33	-3.60***	90.96	113.85	-1.55	35.65	94.80	-4.13***
Road Density (m)	84.67	80.67	0.07	92.00	69.36	1.38	78.42	66.98	0.93	225.42	153.85	2.99***
Landlocked (=1)	0.67	0.40	0.92	0.40	0.67	-1.78*	0.57	0.70	-1.13	0.36	0.70	-2.41**
Main Trade Access (=1 if main port, airport or trade post in municipality)	0.27	0.62	-1.07	0.20	0.36	-1.31	0.42	0.37	0.46	0.63	0.36	1.83*
Electricity (% Households)	0.99	0.89	0.71	0.94	0.90	1.00	0.93	0.89	0.73	0.97	0.90	1.26
Completed Primary Education	0.27	0.09	1.54	0.29	0.20	1.87*	0.27	0.19	2.00**	0.29	0.21	1.12
Illiterate	0.22	0.33	-1.01	0.24	0.28	-1.46	0.24	0.28	-1.26	0.21	0.29	-1.50
Economically Active Population	0.54	0.55	-0.29	0.56	0.55	0.16	0.55	0.55	0.10	0.55	0.56	-0.24
Unemployment	0.02	0.03	-0.60	0.04	0.03	0.3	0.04	0.03	1.25	0.04	0.03	1.16
Manufacturing (% employed)	0.18	0.11	1.15	0.16	0.14	0.66	0.16	0.14	0.96	0.17	0.14	1.26
Agriculture (%employed)	0.08	0.45	-2.17**	0.16	0.4	-2.8**	0.30	0.5	-1.90*	0.20	0.46	-1.95*
Urban	0.90	0.44	1.78*	0.78	0.44	3.20***	0.56	0.43	1.48	0.35	0.45	-0.93
Migrants in past five years	0.01	0.08	-1.10	0.01	0.07	-2.13**	0.07	0.08	-0.72	0.08	0.09	0.03
B. Pre-trend, Selected Outcomes												
Average real per capita expenditure, annual	0.99	0.41	1.41	0.27	0.40	-0.47	0.48	0.41	0.43	0.23	0.41	-0.87
10th p- real per capita expenditure, annual.	1.10	0.54	1.12	0.53	0.51	0.06	0.71	0.52	0.85	0.29	0.54	-0.96
50th p- real per capita expenditure, annual	1.13	0.48	1.54	0.32	0.47	-0.53	0.60	0.47	0.70	0.25	0.48	-1.05
90th p- real per capita expenditure, annual	0.83	0.40	1.02	0.48	0.37	0.42	0.57	0.38	0.97	0.23	0.39	-0.74
Share of high-skill over low-skill (1)	4.30	2.48	0.54	2.28	2.52	-0.15	8.80	2.11	3.98***	2.46	2.38	0.04
Share of high-skill over low-skill (2)	-0.71	2.09	-0.40	-0.69	2.18	-0.90	-0.56	2.32	-0.89	-0.26	2.25	-0.70
N =	4	70		5	67		7	65		2	70	

Notes: Municipalities are grouped based on the sequence of EPZ establishment. Values are for simple averages for the pre-treatment period. Definitions are as in Table 3. For municipalities treated before 2000, the historical trends denotes the average growth rate of the outcomes before 2001. I would otherwise have no pre-treatment trends for this group. A similar approach is done for the rest of the groups, but strictly considering only the pre-treatment period. The control municipalities are constructed using a propensity score matching based on k-nearest neighbours (4) with respect to municipalities characteristics and pre-trend outcomes. Matched characteristics include all of the variables on this table.

Table E5. DID Estimates of the Effect of EPZs on Average Real Expenditure per capita

	(1)	(2)	(3)	(4)	(5)
EPZ	**525.8****	**477.4****	**524.5****	**478.0****	**454.3***
	(214.1)	(203.9)	(242.3)	(235.0)	(238.3)
Observations	375	375	340	340	340
R-sq.	0.722	0.818	0.742	0.839	0.936
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Municipality-year trend	✓	✓			
Province-year dummies			✓	✓	
Region-year dummies					✓
Covariates		✓		✓	✓
Municipalities	84	84	84	84	84

Notes: Dependent variable is levels of real expenditure per capita. Robust standard errors in parentheses, clustered at municipality. Covariates in column 2 include dummies for time-varying individual and household characteristics: educational level, age square, gender, urban, sector of work, household size and having migrated within the past five years. Individual observations are only for working age individuals. Covariates in column 4-5 include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the propensity score matched control group.

*** p<0.01, ** p<0.05, * p<0.1

Table E6. DID Estimates of the Effect of EPZs on Average Real Expenditure per capita – Includes Pre-Treatment Trend

	(1)	(2)	(3)	(4)	(5)
EPZ	**932.9****	**1,127***	**1,006****	**935.0***	**965.7***
	(464.9)	(575.9)	(403.6)	(558.7)	(556.2)
Pre-Trend	**152.1***	240.2	**101.0***	84.83	88.20
	(72.79)	(189.4)	(49.79)	(54.22)	(56.75)
Observations	52,997	52,997	375	375	375
R-sq.	0.449	0.451	0.701	0.718	0.718
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Municipality-year trend	✓	✓			
Province-year dummies			✓		✓
Covariates		✓		✓	✓
Municipalities	106	106	106	106	106

Notes: Dependent variable is levels of real expenditure per capita. Robust standard errors in parentheses, clustered at municipality. Covariates in column 2 include dummies for time-varying individual and household characteristics: educational level, age square, gender, urban, sector of work, household size and having migrated within the past five years. Individual observations are only for working age individuals. Covariates in column 4-5 include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the preferred control group using the road network buffer. Pre-trend is equal to the interaction between municipality specific linear-time trends and the EPZ dummy variable (treatment variable). The pre-trend allows the time trends to differ for treated municipalities prior to EPZ establishment.
*** p<0.01, ** p<0.05, * p<0.1

Table E7. DID Estimates of the Effect of EPZs across the Expenditure Distribution in Treated Municipalities

Real expenditure per capita	(1)	(2)
10th Percentile	222.5	221.5
	(146.2)	(139.5)
20th Percentile	147.5	144.2
	(172.4)	(169.7)
30th Percentile	153.5	152.2
	(215.7)	(212.4)
40th Percentile	229.2	226.0
	(224.4)	(222.9)
50th Percentile	243.1	241.1
	(261.2)	(256.6)
60th Percentile	237.3	232.5
	(311.0)	(308.1)
70th Percentile	428.3	429.9
	(355.2)	(353.8)
80th Percentile	738.1	739.5
	(495.6)	(477.6)
90th Percentile	**2,010***	**2,035***
	(1,201)	(1,108)
Year FE	✓	✓
Municipality FE	✓	✓
Province-year dummies	✓	
Region-year dummies		✓
Covariates	✓	✓
Observations	340	340
Municipalities	84	84

Notes: Dependent variables are deciles of real expenditure distribution. Robust standard errors in parentheses, clustered at municipality. Covariates include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the control group constructed using the propensity score matching.
*** p<0.01, ** p<0.05, * p<0.1

Table E8. The Heterogeneous Time Dynamics of EPZ Establishment

	10th percentile	20th percentile	30th percentile	40th percentile	50th percentile	60th percentile	70th percentile	80th percentile	90th percentile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
EPZ establishment, t0	156.9	-6.670	62.52	147.5	103.0	164.5	547.0	434.3	1,723
	(154.8)	(167.4)	(162.7)	(188.9)	(233.7)	(267.3)	(419.7)	(426.6)	(1,356)
EPZ establishment, t+4	339.9	245.3	160.7	203.2	216.4	38.39	330.3	252.5	**1,405***
	(235.3)	(292.1)	(349.1)	(324.6)	(331.7)	(429.3)	(493.5)	(580.7)	(722)
EPZ establishment, t+8	**290.9***	360.4	244.0	402.3	**752.0***	**890.7***	**1,150****	**1,414****	**2,415****
	(173.2)	(312.1)	(339.8)	(246.8)	(383.8)	(449.5)	(498.4)	(560.4)	(942.5)
EPZ establishment, t+11	**800.4*****	379.8	504.2	528.4	**758.2***	**648.3**	**1,139****	**824.1**	**3,103*****
	(291.1)	(398.1)	(480.9)	(519.2)	(393.2)	(417.9)	(500.2)	(665.6)	(779.0)
Observations	340	340	340	340	340	340	340	340	340
R-sq.	0.638	0.682	0.683	0.714	0.711	0.633	0.630	0.604	0.375
Municipalities	84	84	84	84	84	84	84	84	84

Notes: Dependent variables are deciles of the real per capita expenditure distribution. Robust standard errors in parentheses, clustered at municipality. All regressions include year and municipality fixed-effects and province-year dummies. Covariates include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. All leads are equal to one in only one year each per adopting state. EPZ dummy (t0) equals one only for year of establishment. EPZ t+4 includes t+3, and EPZ t+8 includes t+7, to have balanced dummies. Results are estimated against the preferred control group using the propensity score matching. I excludes the negative pre-treatment years to increase precision of the point estimates with the smaller sample size.

*** p<0.01, ** p<0.05, * p<0.1

Table E9 The Effect of EPZs Establishment & Spillover Dynamics

	10th percentile	20th percentile	30th percentile	40th percentile	50th percentile	60th percentile	70th percentile	80th percentile	90th percentile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
EPZ	199.2	94.16	99.31	206.9	241.3	249.3	507.8	**936.5***	**2,159***
	(140.8)	(174.7)	(219.1)	(241.1)	(280.2)	(331.6)	(390.3)	(497.6)	(1,167)
Neighbouring Municipalities1	-308.2	-373.1	-318.7	46.86	-115.6	-162.3	285.3	664.9	-768.3
(5 to 15km)	(387.0)	(310.2)	(396.8)	(257.6)	(290.5)	(366.2)	(575.1)	(751.4)	(557.4)
Neighbouring Municipalities2	-123.7	-151.7	-188.1	-140.9	-7.975	40.85	172.5	**641.4***	371.7
(15 to 35km)	(145.3)	(176.6)	(206.3)	(206.0)	(262.1)	(282.8)	(327.9)	(377.6)	(555.2)
Observations	340	340	340	340	340	340	340	340	340
R-sq.	0.678	0.743	0.744	0.754	0.747	0.752	0.688	0.707	0.580
Municipalities	84	84	84	84	84	84	84	84	84

Notes: Dependent variables are deciles of the real per capita expenditure distribution. Robust standard errors in parentheses, clustered at municipality. All regressions include year and municipality fixed-effects and province-year dummies. Covariates include illiteracy rate, the share of population with primary and tertiary education, and the share of urban population, migrants and share of households with access to electricity at municipal-level. The variables Neighbour Municipalities are dummy variables that equal one if a municipality's centroid is situated at 5 to 15km, and 15km to 35km distance from a treated municipality, and zero otherwise. I use the centroid's distance to avoid noise from larger municipalities. Results are estimated against the propensity score matched control group.
*** p<0.01, ** p<0.05, * p<0.1

Table E10. The Effect of New EPZs on Internal Migration

	FE			LOGIT
	(1)	(2)	(3)	(4)
EPZ	0.009	0.0169	0.0101	0.0183
	(0.043)	(0.0231)	(0.043)	(0.70)
Observations	46,339	46,339	46,339	47,556
R sq.	0.162	0.157	0.162	-
Municipalities	106	106	106	141
Year FE	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓
Province FE	✓			✓
Province-Time dummies		✓		
Municipality-Time Trend	✓		✓	
Region-Time dummies	✓		✓	✓

Notes: Dependent variable is migrant in the past five years. Robust standard errors in parentheses, clustered at the municipality level. Covariates includes dummy variables for level of educational achievement, illiteracy, gender, sector of employment, urban residence, age square, number of household members, and access to electricity at the household level. Observations are limited to working age population (13-68 years).Logit outputs are margins. I exclude 1998 and 2009, due to significant measurement error in the dependent variable.

*** p<0.01, ** p<0.05, * p<0.1

Table E11. The Effect of EPZs across the Expenditure Distribution in Treated Municipalities - Population Weights

Real Expenditure per capita	Census weights	Sample weights
	(1)	(2)
10th Percentile	117.5	104.9
	(118.9)	(103.6)
20th Percentile	52.42	31.36
	(143.0)	(134.1)
30th Percentile	49.59	14.09
	(178.3)	(161.0)
40th Percentile	129.7	116.2
	(185.6)	(199.6)
50th Percentile	110.9	67.86
	(293.8)	(278.3)
60th Percentile	232.0	200.7
	(284.0)	(289.0)
70th Percentile	395.9	282.8
	(477.4)	(428.0)
80th Percentile	959.9	853.3
	(601.1)	(566.6)
90th Percentile	**1,856***	**1,557***
	(956.7)	(827.3)
Year FE	✓	✓
Municipality FE	✓	✓
Province-year dummies	✓	
Region-year dummies		✓
Covariates	✓	✓
Observations	375	375
Municipalities	106	106

Notes: Dependent variables are deciles of real expenditure distribution. Robust standard errors in parentheses, clustered at municipality. Covariates include illiteracy rate, the share of population with primary and tertiary education, and the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the preferred control group using the road network buffer. *** p<0.01, ** p<0.05, * p<0.1

Table E12. DID Estimates of the Effect of EPZs across the Expenditure Distribution in Treated Municipalities - Robustness Checks

Real Expenditure per capita	(1)	(2)
10th Percentile	262.2	**275.6***
	(164.5)	(142.8)
20th Percentile	186.5	197.4
	(142.4)	(145.2)
30th Percentile	206.5	223.3
	(191.2)	(195.1)
40th Percentile	268.3	274.8
	(254.4)	(210.1)
50th Percentile	282.5	325.2
	(274.2)	(236.2)
60th Percentile	315.8	347.5
	(284.8)	(226.6)
70th Percentile	581.3	552.9
	(341.3)	(354.9)
80th Percentile	**874.8***	**946.1***
	(445.1)	(532.8)
90th Percentile	**2,334****	**2,239***
	(1,085)	(1,236)
Year FE	✓	✓
Municipality FE	✓	✓
Province-year dummies	✓	✓
Covariates	✓	✓
Observations	359	370
Municipalities	80	105

Notes: Dependent variables are deciles of real expenditure distribution. Robust standard errors in parentheses, clustered at municipality. Covariates include illiteracy rate, the share of population with primary and tertiary education, and the share of urban population, migrants and share of households with access to electricity at municipal-level. Results are estimated against the preferred control group using the road network buffer. Column 1 excludes municipalities observed less than 4 out of the 5 years. Column 2 excludes Managua. *** p<0.01, ** p<0.05, * p<0.1

Table E13. The Heterogeneous Time-Effect of EPZ Establishment - Robustness Check

	10th percentile (1)	20th percentile (2)	30th percentile (3)	40th percentile (4)	50th percentile (5)	60th percentile (6)	70th percentile (7)	80th percentile (8)	90th percentile (9)
EPZ establishment, t0	220.2	73.75	163.5	207.5	190.1	264.3	523.6	713.2	2,025
	(177.3)	(192.6)	(178.0)	(230.9)	(298.8)	(336.1)	(500.5)	(570.6)	(1,678)
EPZ establishment, t+4	479.7	436.5	387.8	412.0	502.8	364.9	446.6	1,050	2,498
	(345.8)	(413.2)	(504.1)	(478.6)	(472.3)	(549.3)	(567.0)	(773.3)	(1,625)
EPZ establishment, t+8	**457.5***	503.8	397.5	**522.4***	1,017	**1,177***	**1,220***	**2,265***	**3,483****
	(237.3)	(362.3)	(356.2)	(309.8)	(619.3)	(695.1)	(661.6)	(1,319)	(1,523)
EPZ establishment, t+11	**1,158****	565.6	825.7	756.9	**1,204****	**991.6***	1,254	1,668	**5,826*****
	(462.5)	(644.3)	(870.8)	(785.1)	(600.5)	(593.8)	(823.5)	(1,020)	(1,560)
Observations	370	370	370	370	370	370	370	370	370
R-sq.	0.629	0.676	0.677	0.708	0.706	0.623	0.622	0.598	0.364
Municipalities	105	105	105	105	105	105	105	105	105

Notes: Dependent variables are deciles of the real per capita expenditure distribution. Robust standard errors in parentheses, clustered at municipality. All regressions include year and municipality fixed-effects and province-year dummies. Covariates include illiteracy rate, the share of population with primary and tertiary education, the share of urban population, migrants and share of households with access to electricity at municipal-level. All leads are equal to one in only one year each per adopting state. EPZ dummy (t0) equals one only for year of establishment. EPZ t+/-4 includes t+/-3, and EPZ t+/-8 includes t+/-7, to have balanced dummies. Results are estimated against the preferred control group using the road network buffer. Observations exclude the municipality of Managua. *** p<0.01, ** p<0.05, * p<0.1