



THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE ■

Essays in Macro and Development Economics

Dzhamilya Nigmatulina

London School of Economics

A thesis submitted to the London School of Economics for the Degree of Doctor of Philosophy

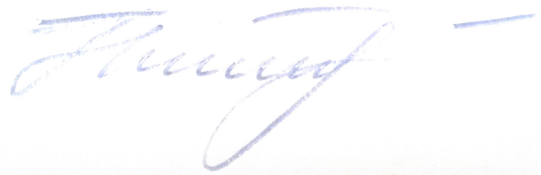
London, March 2021

Declaration

I certify that the thesis I have presented for examination for the 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).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 48,438 words including footnotes but excluding bibliography and appendices.

A handwritten signature in blue ink, appearing to read 'Huang', is written on a light-colored background.

Statement of co-authored work

I certify that Chapter 5 of this thesis is co-authored with Vernon Henderson and Sebastian Kriticos, and I contributed 50% of the work. Chapter 4 of this thesis is co-authored with Guy Michaels, Ferdinand Rauch, Tanner Regan, Neeraj Baruah and Amanda Dahlstrand, and I contributed 33% of the work. Figures 3.2, 3.3, 3.4, 3.5 and 3.A1 in Chapter 3 were created by me as a part of 2018 EBRD Transition Report, for a section co-authored with Nathaniel Young.

Acknowledgements

This thesis would not have been possible without the support of colleagues, friends and family. First and foremost, I would like to thank my supervisor Vernon Henderson for great guidance, encouragement and for setting the highest standard with his example. Vernon's attention, motivation and thoughtful feedback have made me achieve my best. I am very grateful to my advisor Steve Gibbons for being there and always sharing so many insightful comments.

I have also learned much from working with Guy Michaels and conversations with Alwyn Young and other incredible colleagues at the CEP, the SERC office, Economics, Management and Geography departments at the LSE.

It is unlikely that I would have walked this path without the people who have opened my eyes to the joys of exploring Economics: Guido Cozzi and Charlie Becker. Guido Cozzi has first shared what it is like to be a researcher, and Charlie Becker helped, encouraged and supported me when I was making my first steps.

I am also lucky for having shared the PhD journey with wonderful people who became friends. I am particularly grateful to Tanner has been there, in sorrow and in joy, and generously channelled the sense that we are all together on this journey. Tanner taught me a great deal about life, work and science.

I am indebted to my parents and my grandmother, who gave me encouragement and support throughout. My father has pushed me to have the courage to do what I feel is right and to strive to get to the bottom of things, and my mother and grandmother have shared their energy to keep going.

Finally, I also gratefully acknowledge financial support through the PhD scholarship from the LSE.

I dedicate this thesis to my grandfather Ernest Galimovich Ulumbekov, who taught me to become a professional in whatever endeavour I chose and showed me how to achieve it -

"Nulla dies sine linea". He made me feel curious about why things happen around me and with his example showed how one can fulfil this curiosity for a living.

Abstract

This thesis studies the economics of firms, neighbourhoods, cities and regions in developing countries. It combines empirical and structural methods using satellite, administrative and survey data to study allocative efficiency, the economics of cities and urban planning in developing countries. The thesis is organised into four independent chapters. The first chapter studies the extent of misallocation among firms in Russia, and the role of state ownership in affecting allocative efficiency. I find that there are large wedges between state-owned and private firms reducing the aggregate TFP by at least 11%. Using a unique natural experiment of staggered firm-level US sanctions I find one channel through which resources become misallocated: the state-owned enterprises are shielded excessively from negative economic shocks. I find that allocative efficiency worsened after the sanctions episode and the Russian TFP dropped at least by 0.33% overall, reaching -3% in some sectors as a combined effect of sanctions and shielding. The second chapter explores new data methods and GIS tools for use in urban and spatial economics. It discusses the strengths and challenges in the application of novel datasets, such as satellite imagery, gridded population and develops methods that effectively apply these datasets to economic research questions. The third chapter (co-authored) explores the long-run consequences of planning and providing basic infrastructure in neighbourhoods, where people build their own homes. Using satellite images and surveys from the 2010s, we find that de novo planning induces neighbourhoods to develop better housing, with larger footprint areas, more stories, more connections to electricity and water, basic sanitation and access to roads. This effect remains even after accounting for selection of financially unconstrained owners into planned areas. The fourth chapter (co-authored) estimates urban agglomeration effects, exploring both simple and very nuanced measures of economic density to explain household income and wage differences across cities in six Sub-Saharan African countries. Defining cities consistently, we find large wage gains to being in denser cities in Sub-Saharan Africa, which are generally larger than such estimates for other parts of the world. We also find extraordinary household income gains to density that are far greater than wage ones. Such gains are consistent with the pull forces driving rapid urbanization in the region.

Table of Contents

1	Introduction	1
2	Misallocation and state ownership: evidence from the Russian sanctions	3
2.1	Introduction	3
2.2	Related literature	6
2.3	Model	8
2.4	Data and context	12
2.4.1	Firm-level data	12
2.4.2	Sanctions on Russia 2014-2019	13
2.4.3	Coverage of the economy	15
2.5	Measuring firm productivity and distortions	15
2.6	Static misallocation in Russia	18
2.7	Counterfactuals	27
2.8	Sanctions as a test of the SOE protection	28
2.8.1	Event studies	33
2.9	Results	34
2.9.1	Regression results	34
2.9.2	Event studies results	36
2.9.3	Aggregate effects	38
2.10	Conclusion	38
2.A	Appendix A. Heterogeneous firm model	40
2.B	Appendix B. Additional tables and figures	48
2.C	Appendix C. Data appendix	50
3	People in space: a toolbox for an economic geographer	53
3.1	Introduction: why do we care about population density?	53
3.2	Data	58
3.3	How to think about density across labour markets.	62
3.4	People in space within labour markets, or why the density and GDP per capita correlation is lower in Africa?	66
3.4.1	Theoretical overview.	67

3.4.2	Methods: city extents.	69
3.4.3	Methods: combining data from different geographies	71
3.4.4	Methods: within-city placement of people.	73
3.4.5	Results: how does economic density in Africa compare with the rest of the world?	77
3.5	Conclusion	82
3.A	Appendix A. Additional tables and figures	84
3.B	Appendix B. Calculating metropolitan areas for the world	85
4	Planning ahead for better neighborhoods: evidence from Tanzania	87
4.1	Introduction	87
4.2	Institutional background and data	92
4.2.1	Institutional background	92
4.2.2	Data description	97
4.3	Research design and empirical findings	101
4.3.1	Research design	101
4.3.2	Empirical findings	105
4.4	Model	111
4.4.1	Assumptions and their relationship to the institutional setting . . .	111
4.4.2	Solving the model	114
4.4.3	Neighborhood development	115
4.4.4	Relating the model to the empirical analysis	116
4.4.5	Implications of the model	118
4.5	Concluding remarks	119
4.6	Main tables	121
4.A	Appendix tables and figures	126
4.B	Data appendix	145
4.B.1	Project background and treatment	146
4.B.2	Outcome variables derived from imagery data	150
4.B.3	Tanzanian Strategic Cities Project survey data	152
4.B.4	Geographic control variables	155
4.B.5	Land values	156
4.B.6	Project costs	157
4.B.7	Additional data	159
5	Measuring urban economic density	161
5.1	Introduction	161
5.2	Using Landscan data and defining urbanized areas	165
5.2.1	Landscan data	165

5.2.2	Defining urbanized areas	167
5.3	Defining economic density	170
5.4	How are differences in economic density across the spatial hierarchy related to income differences?	173
5.4.1	The data and the sample of countries and cities	173
5.4.2	Basic specification and urban-rural results	175
5.5	Economic density in cities	178
5.5.1	Constructing de la Roca-Puga measures	178
5.5.2	Economic density results for cities	179
5.6	Conclusions	186
5.A	Appendix A. Statistics and other results	200
5.B	Appendix B. Data description and methodology	205
5.B.1	Using Landscan to create urbanized area boundaries	206
5.B.2	Living Standards Measurement Surveys (LSMS)	209
5.B.3	Harmonizing the data, local density and ring density measures, and the optimal de la Roca-Puga discount rate	211
5.C	Appendix C. Ground-truthing Landscan data within cities	213
5.D	Appendix D. Neighborhood productivity and Kampala data	217
5.D.1	Matching the data to density measures	218

List of Figures

2.1	Factor allocations by firm productivity	22
2.2	Firm-specific distortions and productivity (TFPR on TFPQ)	23
2.3	Allocations of SOEs versus the private sector, variables adjusted for measurement error	25
2.4	Allocations before 2015	26
2.5	SSI event study with not-yet sanctioned firms in the control group.	36
2.6	Pre-post 2015 event study with never-sanctioned firms in the control group	37
2.A1	Allocations of SOEs versus the private sector	49
2.A2	Constant sample SSI event study	50
3.1	Correlation of $\ln(\text{GDP per capita})$ and $\ln(\text{population density})$ of 1 decimal degree by 1 decimal degree units using G-Econ data Nordhaus et al. (2018) in the entire world in the year 2005	54
3.2	Population changes in Europe in 1990-2014 people in 10km^2	55
3.3	Change in local population density 2000-2014.	57
3.4	A scatterplot of average local density changes and population 1990-2014	63
3.5	Percent of people experiencing population decline in local population density.	65
3.6	Correlation of GDP per capita and density of 1 decimal degree by 1 decimal degree units using G-Econ data Nordhaus et al. (2018) on the African continent in year 2005.	67
3.7	An example of assigning data from arbitrary spatial units	72
3.A1	Changes in local population density for Eastern Europe and CIS countries.	84
3.A2	Summary of gridded population datasets.	85
4.A1	Locations of de novo, upgrading, and control areas by city	126
4.A2	Example images of de novo, upgrade and control areas	127
4.A3	Regression discontinuity plots of summary outcomes from Tables 1 and 2	128
5.1	Defining cities and towns	196
5.2	PDFs of city sample in the African continent	197
5.3	Differences in city layout and density measures	198

5.4	Overall size distribution of urbanized areas	199
5.5	Size distribution of urbanized areas	199
5.B1	Rook and queen neighbors	207
5.B2	Defining cities and towns (with labels and administrative boundaries) . .	208
5.B3	Density ring measures	212

List of Tables

2.1	Sample used for analysis	16
2.2	Sample used for analysis	17
2.3	Summary statistics of key variables	19
2.4	Dispersion of $\ln(\text{TFPR})$, $\ln(\text{TFPQ})$, $\ln(\text{MRPL})$, $\ln(\text{MRPK})$	20
2.5	Counterfactual exercises	27
2.6	Sanctions by ownership	28
2.7	Sanctions by sector	28
2.8	Summary by sanction type	29
2.9	Average effects of sanctions: key outcome variables	35
2.A1	Average effects of sanctions triple difference	48
2.A2	TFPs Results (aggregate effects of sanctions by industry)	52
3.1	A list of spatial indices	75
3.2	Comparing within-city distributions of population in Africa with the de-veloped countries using Landscan data	79
3.3	Comparing within-city distributions of population in Africa with the de-veloping countries using Landscan data	81
4.1	De novo regressions using imagery data for all seven cities	121
4.2	De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga	122
4.3	De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with owner name fixed effects	123
4.4	De novo regressions on persistence measures using imagery and TSCP survey data	124
4.A1	De novo neighborhoods	129
4.A2	Upgrading neighborhoods	130
4.A3	Plot counts and population by project type	131
4.A4	Summary statistics	132
4.A5	De novo regressions balancing first geography	133
4.A6	De novo regressions of adult census outcomes	134
4.A7	Upgrading regressions balancing first geography	135

4.A8	Upgrading regressions using imagery data for all seven cities	136
4.A9	Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga	137
4.A10	Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with Owner Name Fixed Effects	138
4.A11	Upgrading regressions on persistence measures using imagery and TSCP survey data	139
4.A12	Upgrading regressions of adult census outcomes	140
4.A13	Details on the selection of control areas by city	141
4.A14	Description of variables derived from imagery data	142
4.A15	Description of TSCP variables and how they are created	143
4.A16	Hedonic housing value regressions using TSCP survey data	144
4.A17	Description of variables from Tanzanian census 2012	145
5.1	Counts of urbanised areas in our countries and sample	188
5.2	Household and person characteristics by location	189
5.3	Gains from urban type	190
5.4	Gains from scale by area type	191
5.5	Gains from within city clustering	192
5.6	Why household income gains are larger than for wages?	193
5.7	Estimation of average and local density effects on household income premi- ums	194
5.8	Estimation of average and local density effects on wage premiums	195
5.A1	Summary statistics	200
5.A2	Gains from scale by area type. No industry fixed effects.	201
5.A3	Horserace: estimation of density effects on household incomes	202
5.A4	Horserace: estimation of density effects on individual hourly wage premiums	203
5.A5	Estimation of scale effects on HH income premiums	204
5.A6	Estimation of scale effects on hourly wage premiums	205
5.B1	LSMS surveys	209
5.B2	Income sources and time intervals in LSMS surveys	210
5.B3	Expenses and time intervals in LSMS surveys	211

Chapter 1

Introduction

The thesis consists of four independent chapters, the first of which is on macroeconomic effects of misallocation of resources in developing countries, and the remaining three are on urban economics in developing countries.

In the first chapter, I use firm panel data and a quantitative framework to document the extent of misallocation in Russia. I find that there are large wedges between state-owned and private firms that prevent labour and capital inputs from flowing to more productive private firms. I quantify the degree of misallocation attributed to state ownership and find that the aggregate TFP would increase by at least 11% if the wedges between state-owned enterprises and private firms disappeared. Using a unique natural experiment of staggered firm-level US sanctions, I find one channel through which resources become misallocated between state-owned and private firms: excessive shielding from negative shocks. I find that misallocation grew after the sanctions episode and the Russian TFP dropped at least by 0.33% overall, reaching -3% in some sectors as a combined effect of sanctions and shielding.

The second chapter explores new data methods and GIS tools for use in urban and spatial economics. It discusses the strengths and challenges in applying novel datasets, such as satellite imagery, grids of GDP and population. It develops methods that effectively apply these datasets to economic research questions. It further compares spatial correlation indices and discusses their geographic and economic applications.

The third chapter (co-authored) explores the long-run consequences of planning and providing basic infrastructure in neighbourhoods, where people build their own homes. We study “Sites and Services” projects implemented in seven Tanzanian cities during the 1970s and 1980s, half of which provided infrastructure in previously unpopulated areas (de novo neighbourhoods), while the other half upgraded squatter settlements. Using satellite images and surveys from the 2010s, we find that de novo neighbourhoods developed better housing than adjacent residential areas (control areas) that were also initially unpopulated.

Specifically, de novo neighbourhoods are more orderly and their buildings have larger footprint areas and are more likely to have multiple stories, as well as connections to electricity and water, basic sanitation and access to roads. This effect remains even after accounting for selection of financially unconstrained owners into de novo areas. While we have no natural counterfactual for the upgrading areas, descriptive evidence suggests that they are, if anything, worse than the control areas.

The fourth chapter (co-authored) estimates urban agglomeration effects, exploring simple and very nuanced measures of economic density to explain household income and wage differences across cities in six Sub-Saharan African countries. A key aspect of the work is that we define cities consistently across space based on fine-scale density measures in order to gauge the economic extent of the city. The evidence suggests that more nuanced measures of density, which attempt to capture within-city differences in the extent of clustering, do no better than a simple density measure in explaining income differences across cities. However, we find that the total city population is a poor measure. We find large wage gains to being in denser cities in Sub-Saharan Africa, generally larger than such estimates for other parts of the world. We also find extraordinary household income gains to density that are far greater than wage ones. Such gains help explain the pull forces driving rapid urbanization in the region.

Chapter 2

Misallocation and state ownership: evidence from the Russian sanctions

2.1 Introduction

Allocative efficiency has been shown to play a key role in TFP and GDP differences across countries (Hsieh & Klenow 2009, Bartelsman et al. 2013, Gopinath et al. 2017). This finding is hopeful because this implies that low-income countries may accelerate catch up by redistributing existing resources optimally, rather than having to invest in expensive technological upgrades to improve productivity for each firm. However, it is still unknown what allocative distortions explain a bigger share of the gap. It is impossible to implement policies for economic efficiency without knowing what the main drivers of misallocation are. Ownership and political connections of owners is a potentially large channel that can favour the allocation of resources to firms not based on efficiency but based on the interests of those with political power. A unique firm-level dataset and a natural experiment allow me to make progress on this front.

This paper measures the contribution of state ownership and political connections to misallocation in Russia. Russia, with its 20% SOE revenue share in GDP, provides a good test case to account for misallocation from state ownership for countries with over 10-30% shares of SOEs' revenue in GDP, such as China, India and Brazil (Kowalski et al. 2013) and post-Communist countries (EBRD 2020). Such high shares of state-owned activities in these countries have the potential to be large sources of allocative inefficiency. These sources may be further amplified via relationships between state-owned firms and private firms.

Allocative inefficiency can arise, for example, if the state-owned enterprises are favoured in capital input markets and receive preferential loans or excessive subsidies. In addition,

the state-owned companies may not be profit-maximizing and be directed to participate in projects of political rather than economic interest. While such interest may be well justified, this comes with an economic efficiency cost that may affect aggregate TFP. Both preferential subsidies and allocation of contracts that are not profit-maximizing will make the state-owned firms larger than their efficient size and receive more capital and/or labour relative to the private firms.

Russia in 2014-2019 is also an excellent case to study the role of state-ownership in misallocation because of a unique natural experiment: many state-owned firms were targeted with sanctions by the US and EU in years 2014-2018, which created a negative shock to the access of inputs for the state-owned sector as well as for some private firms. This experiment allows me to use a difference-in-difference (DID) setup to capture the firm-level within-industry effect of the negative shock as the (differential) response of the state-owned firms relative to private sanctioned firms and sanctioned firms relative to the non-treated firms in the same industry. The differential response of the SOEs to negative shocks will reveal whether there is a link from state ownership to misallocation in how SOEs respond to negative input shocks. The DID setup also allows me to alleviate the common concerns in measuring misallocation - measurement error, adjustment costs and abstract from other correlated unobserved factors affecting the SOEs and obscuring the measurement of misallocation from SOEs' behaviour.

The effect of sanctions is not clear ex-ante. If sanctions targeted the inputs of those firms that already have more inputs than is efficient, the treatment should improve allocative efficiency in Russia. However, the response of the Russian state by protecting the sanctioned firms or sanctioned SOEs may fully reverse the direct effect of sanctions, and such protection can even overshoot and exacerbate misallocation.

I conclude that state ownership is associated with implicit subsidies for the operation of state-owned firms. I further find that sanctions together with the response of the Russian state have worsened the allocative efficiency in Russia: the private sanctioned firms maintained their relative size and have seen the negative shock of sanctions fully reversed, whereas the sanctioned SOEs have not only seen the negative shock reversed but actually gained additional inputs after sanctions were imposed. I estimate that all else equal, this sanctions episode worsened misallocation of resources and productivity on the aggregate in Russia.

I start by using a panel of medium and large Russian firms in the Services, Manufacturing and Agricultural sectors from 2012-2018 and measuring the extent of misallocation

in Russia using wedge accounting framework a-la Hsieh & Klenow (2009). I correct for measurement error and transient adjustment costs using firm and year fixed effects and this way avoid attributing all of the cross-sectional dispersion in the observed marginal returns to inputs to misallocation like most of the early literature does. I then account for how much of the distance to the efficient frontier is driven by variation in ownership (state-owned versus private).

I collect information on firm-by-year sanctions imposed on politically connected state-owned and private firms. I then measure the ex-ante marginal revenue products of capital (MRPK) for these sanctioned firms. I use the panel data and within-firm variation over time to empirically test whether the sanctions on inputs indeed changed the inputs of targeted firms that were ex-ante low MRPK firms. I further test whether the allocation of inputs to targeted SOEs changed differentially to private sanctioned firms.

The staggered nature of sanctions allows me to net out the differential effects on each industry of changes in oil price and devaluation of the Russian rouble that took place in the same period. Further, the DID setup does not require the sanctioned and non-sanctioned firms, or SOEs and private firms, to have the same fixed characteristics, as they drop out with the firm fixed effects. For estimating the average effect, this method does require that the sanctioned firms have trended the same way as non-sanctioned firms in a world without sanctions, for which I provide convincing evidence based on pre-trends. To estimate the SOE differential, I rely on a weaker assumption: the differential trends of the SOE and private firms need to be the same between sanctioned and non-sanctioned groups. I also find similar estimates when estimating the effects only within sanctioned firms. The effects I find are robust to controlling for time shocks at the disaggregated industry and size quartiles and time-by-SOE fixed effects. Finally, I use the method based on the Hsieh and Klenow (2009) framework to account for the effects of sanctions on aggregate TFP.

First, I find that Russia could double its aggregate TFP if all misallocation was removed, as measured by the heterogeneous firms model. Second, I find that Russia could walk 10% of that distance if it removed the wedge between SOEs and private firms¹. Third, the natural experiment of sanctions shows that Russia appears to be walking in the wrong direction: SOEs that are ex-ante low MRPK firms that have been targeted by Western input sanctions have been shielded to such an extent that they have 23% lower MPRK and 25% higher capital inputs after the sanctions treatment.

¹By coincidence, the current Russian TFP would also increase by roughly the same amount (11%), if the wedge between SOEs and private firms was removed.

Combining these empirical estimates as well as the heterogeneous firm model, I calculate the aggregate effects of the sanctions episode on the aggregate TFP and find that it reduced by 0.33%. The effects within each industry are all mostly negative and range between -3.3% and -0.01% (with several minor exceptions for which TFP mildly improved).

The paper is organized as follows. Section 2 reviews the related literature and how this paper fits in. Section 3 provides a heterogeneous firm framework for accounting for the effects of wedges. In particular, it derives the expressions for accounting for wedges between groups within industries. Section 4 describes the firm-level and sanctions data as well as the context of the sanctions episode. Section 5 discusses the measurement error correction for wedge accounting. Section 6 provides general summary statistics of the state of misallocation in Russia. Section 7 presents the results of the counterfactuals of wedge equalization within and across groups. Section 8 discusses my reduced-form empirical strategy. Section 9 reports the reduced-form effects of sanctions on sanctioned private and state-owned firms, as well as the aggregate effects of the sanction episode. Section 10 concludes.

2.2 Related literature

In this paper, I quantify the effects of state ownership on aggregate productivity through the lens of the allocative efficiency model with the so-called "indirect approach" and causally estimate the differential response of private versus state-owned firms to shocks. In doing so, I add to three strands of literature. First, I contribute to the literature that highlights the role of allocative efficiency for aggregate outcomes (Hsieh & Klenow 2009, Restuccia & Rogerson 2008, Baqaee & Farhi 2020, Busso et al. 2013). Second, I zoom into the effects of state ownership for firm-level outcomes (Hsieh & Song 2015, Berkowitz et al. 2017, Brandt et al. 2018, Bussolo et al. 2019, Brown et al. 2006). Finally, I look at the effects of economic sanctions at the firm-level and in the aggregate (Ahn & Ludema 2020, Tuzova & Qayum 2016, Crozet & Hinz 2016, Haidar 2017, Draca et al. 2019, Stone 2016, Gold et al. 2019). The first-generation literature on misallocation has developed an accounting framework that allows calculating by how much the inefficient allocation of inputs affects the aggregate TFP (Restuccia & Rogerson 2008, Hsieh & Klenow 2009). This branch of work also called the "indirect approach" allowed researchers to diagnose the allocative inefficiencies in an economy, while not making any assumptions about the sources of such inefficiencies. Hsieh & Klenow (2009) used dispersion in revenue productivity (TFPR) as a measure of misallocation within sectors in India, China and the US. Jones (2011),

Baqae & Farhi (2020) have incorporated the role of Input-Output linkages in measuring misallocation and generalized earlier models.

My paper has both the advantage of the indirect approach by not making specific modelling assumptions about a particular source of misallocation but makes the next step by quantifying how much of the misallocation is explained by a particular source *in the data*: in my case, the ownership status of a firm. I account for the role of state-ownership both at a given point in time and, using causal inference, reveal a particular channel through which state-ownership comes to bring misallocation: differential shielding from negative exogenous shocks. This is one of the first papers to connect causal inference and misallocation accounting, along with Rotemberg (2019), who uses a similar approach to quantify the effects of small-firm subsidies in India, and Bau & Matray (2020) who look at the effects of India's capital market liberalization. Therefore, this paper contributes to the nascent literature on the sources of misallocation².

In this paper, I also make an advance in the static accounting of the sources of misallocation. Perhaps the closest paper to mine is Hsieh & Song (2015), an analysis of the privatization reform of SOEs in China through the lens of the "indirect approach" misallocation framework. I build on their work by using state-of-the-art techniques to adjust for measurement error and transient adjustment costs, rather than attributing all cross-sectional variation to misallocation. I also, as mentioned above, use causal inference to pin down a specific channel through which SOEs bring misallocation. Furthermore, to account for misallocation between ownership groups I use the counterfactual of equalizing the wedges within groups, for which I derive the analytical expression of firm marginal revenue products as functions of total resources in each group. Finally, I benefit from the unique feature of my dataset, and include services and agricultural sectors in the analysis of misallocation, whereas all papers I am aware of on the topic consider manufacturing only.

I also fill the gap in the literature on the effects of state-ownership and privatization (Brandt et al. 2018, Bussolo et al. 2019, Brown et al. 2006, Berkowitz et al. 2017), see Megginson (2016) for an extensive review) by quantifying the effects on *the aggregate* TFP. A different literature studies the role of state-ownership (in China) for growth and TFP from a theoretical perspective by explicitly modelling SOEs preferential access to finance: Song et al. (2011), Zilibotti (2017). I add to this literature by leveraging the US sanctions

²Several other papers use the direct approach, and explicitly model the sources of misallocation (Pellegrino & Zheng 2021, Midrigan & Xu 2014, Buera et al. 2011, Asker et al. 2014, Gopinath et al. 2017, Peters 2020, David & Venkateswaran 2019, David et al. 2016). Restuccia & Rogerson (2017) and Hopenhayn (2014) both provide extensive reviews of the literature

as a source of exogenous variation and documenting empirically that the excessive shielding of SOEs from negative shocks is one driver of misallocation.

Finally, while using the sanctions as a source of exogenous variation in inputs, I also add to the work that measures the micro and macroeconomic effects of sanctions (Ahn & Ludema 2020, Tuzova & Qayum 2016, Crozet & Hinz 2016, Haidar 2017, Draca et al. 2019, Stone 2016, Gold et al. 2019). I distinguish myself from these papers in that I not only causally estimate the effect of sanctions on treated firms but also use the estimates to calculate the aggregate effect of the 2014-2018 sanctions episode in Russia on TFP through misallocation.

2.3 Model

I use a standard framework from the misallocation literature where firms have heterogeneous productivities and wedges on inputs K and L are modelled as taxes or subsidies τ_i^K and τ_i^L . These wedges create an arbitrary allocation of resources by increasing the effective price on inputs that a firm faces. Looking from another angle, the distortions in the operation of firms are represented as wedges that would rationalize the observed use of inputs by profit-maximizing firms.

The firm i maximizes its profits while facing taxes or subsidies τ^K and τ^L .³

$$\pi_i = P_i Q_i - (1 + \tau_i^L) w L_i - (1 + \tau_i^K) r K_i \quad (2.1)$$

I assume each firm produces a different variety i and the output of the industry Q in which the firm operates is demanded via a CES demand. All misallocation is within industry, and for simplicity, the industry index is omitted. I also assume a Cobb-Douglas production function $Q_i = A_i K_i^\alpha L_i^{1-\alpha}$, which is standard in the literature (see appendix for the derivations of every step). P is the industry CES price index:

$$\max_{L_i, K_i} \pi_i = P Q^\eta (A_i K_i^\alpha L_i^{1-\alpha})^{1-\eta} - (1 + \tau_i^L) w L_i - (1 + \tau_i^K) r K_i$$

I assume w and r are the common and exogenous costs of labour and capital, so every

³The model can be analogously extended to misallocation of not only capital and labour, but also intermediate inputs. This model also allows for the case that there is misallocation in output, rather than inputs, for example, from transport costs. This can be added as a wedge on output $(1 - \tau_i^Y) P_i Q_i$, but the effect of τ_i^Y cannot be separately identified from the joint effect of τ_i^L and τ_i^K . Therefore, I keep only τ_i^L and τ_i^K , bearing in mind that these two wedges jointly can mean a distortion on output.

variation in these prices manifests itself in τ^K and τ^L . The firm optimal labour and capital allocation will satisfy these equations:

$$\{L_i\} : (1 - \alpha)(1 - \eta) \frac{P_i Q_i}{L_i} = (1 + \tau_i^L)w \equiv MRPL_i \quad (2.2)$$

$$\{K_i\} : \alpha(1 - \eta) \frac{P_i Q_i}{K_i} = (1 + \tau_i^L)r \equiv MRPK_i \quad (2.3)$$

The firm's marginal revenue to each input is equal to the marginal cost of this input. The term $(1 - \eta)$ is the constant markup that comes from the monopolistic competition assumption. The τ^K and τ^L are backed out as wedges that would explain the observed firm decision if the firm was profit maximising. Positive τ^K and τ^L represent implicit taxes on inputs, and negative τ^K and τ^L represent implicit subsidies.⁵

I define $MRPK_i$ and $MRPL_i$ as measures of the direction of misallocation. The higher are $MRPK_i$ and $MRPL_i$ the higher are the implicit taxes on capital and labour inputs of firm i .

The measures $MRPK_i$ and $MRPL_i$ can be summarized with another measure $TFPR_i$ or "Total Factor Productivity Revenue":

$$TFPR_i \equiv \frac{P_i Q_i}{K_i^\alpha L_i^{1-\alpha}} \propto MRPK_i^\alpha * MRPL_i^{1-\alpha} \quad (2.4)$$

Furthermore, the model allows me to define a model-based firm TFP. With the assumption of CES demand and monopolistic competition, the size or market share of a firm is related to its real productivity (A_i or $TFPQ_i$):

$$A_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}} \equiv TFP_i \equiv TFPQ_i \quad (2.5)$$

$$\kappa = (PQ^\eta)^{-\frac{1}{1-\eta}} \quad (2.6)$$

How do the wedges affect the aggregate TFP? I follow Hsieh & Klenow (2009), CES ag-

⁴While it is implausible that wages are common across regions, the results related to misallocation across private and state ownership are robust to looking within regions. The wage variation across regions does contribute to overall misallocation, which may be desirable for calculating the full distance to the efficient frontier.

⁵In the calculation of the overall TFP and country TFP only the *relative* τ^K and τ^L will matter, rather than the absolute levels because each industry will be aggregated into the country output with a Cobb-Douglas production function.

gregation within industries. I first calculate the aggregate output and TFP of an industry, TFP_s . From such aggregation exercise provided in the Appendix, industry TFP can be expressed as the following equation:

$$TFP_s = \left(\sum_i \left(A_i \left(\frac{\overline{MRPL}}{MRPL_i} \right)^{1-\alpha} \left(\frac{\overline{MRPK}}{MRPK_i} \right)^\alpha \right)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.7)$$

In which whenever $MRPK_i$ and $MRPL_i$ deviate from their industry harmonic averages \overline{MRPL} and \overline{MRPK} the industry TFP becomes lower than the efficient level. in an industry. Therefore, the TFP_s when you have the efficient allocation (without wedges) is a CES aggregate of firm-level productivities⁶:

$$TFP_s^e = \left(\sum_i (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.8)$$

To get the country aggregate TFP, I follow Hsieh and Klenow and take a Cobb-Douglas average of each of the industry TFP_s , using the industry value added shares as exponents.

Four things are important to note here. First, only the relative tax in a 4-digit industry will matter for misallocation, an average tax that is equal across firms will lead to efficient allocation across firms within a 4-digit industry. Second, and related, all misallocation in this model comes from the misallocation within a 4-digit industry, and misallocation across sectors will not affect aggregate TFP in this model⁷. An increase in the tax that is the same for every firm in an industry (and, therefore, an increase in the industry price index), will reduce the total physical output, but not the aggregate TFP. This comes from each sector being aggregated a-la Cobb-Douglas and the aggregate TFP term being separable from total sector inputs K_s and L_s ⁸. Third, even though I assume monopolistic competition

⁶In the appendix I show that the equivalent exercise that maximizes total output and taking the distribution of productivities and total inputs in an economy as given, means allocating more resources to more productive firms, but only up to a point, that point being equalized marginal revenue products of each input.

⁷Baqae & Farhi (2020) show that misallocation across sectors may play a smaller role than within sectors because sectors tend to be less substitutable with each other and therefore, reallocation from a sector that faces an increase in an average wedge to other sectors will be smaller.

⁸The aggregate output can be grouped into the TFP term, and the aggregate input terms: $Y = \prod_{s=1}^S (TFP_s K_s^\alpha L_s^{1-\alpha})^{\theta_s}$, in which the $\theta_s < 1$ are the elasticities of substitution across sectors that sum up to 1. If one sector faces a homogeneous tax increase, K_s and L_s will shift to other sectors, but the $\prod_{s=1}^S TFP_s^{\theta_s}$ will remain intact. Meanwhile, the shift of K_s and L_s to other sectors will not be enough to maintain the same level of output, and the output will drop. This is because sectors are complements: $\theta_s < 1$ of any s . This model when applied to the data will not be able to back out the average 4-digit sector wedge separate from the sector elasticity in the aggregate production function. Therefore, while I care about the total output, being affected by the average τ_i^K and τ_i^L as well, I will only be able to confidently measure the drop in average output from the misallocation within sectors.

and therefore constant markups, if other forms of competition are present in the data, the different mark-ups will be reflected in wedges, which is desirable in accounting for the overall distance to the efficient frontier. Finally, this model is static, but the level of misallocation and the wedges can be calculated for any given year as a separate exercise.

My starting point is calculating the distance of the aggregate TFP to the efficient (frontier) as a share.

$$\frac{TFP}{TFP^e} - 1 \quad (2.9)$$

Using this framework I conduct two counterfactual exercises, which together give me how much of the distance to the productivity frontier is explained by the variation in wedges due to the ownership status.

Counterfactual 1 Removing all differences in wedges across all firms (state-owned or not).

Counterfactual 2 Removing all differences in wedges for firms within the industry-ownership group. Whereby in this counterfactual I look at two groups in each sector: state-owned and private, I then redistribute existing labour and existing capital of each group across firms within each group to equalize their MRPL's and MRPK's (i.e. all firms within each group have the same average wedge).

Comparing the gains from equalizing MRPK and MRPL within groups to equalizing MRPL and MRPK everywhere gives me how much distortion comes from between SOE and private groups.

For counterfactual 2 I derive based on the model above counterfactual group expressions for each of MRPL and MRPK:

$$\frac{(L_{priv})^\eta \left(\frac{L_{priv}}{K_{priv}}\right)^{\alpha(1-\eta)}}{(1-\alpha)(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} = \frac{1}{MRPL_{priv}} \quad (2.10)$$

And

$$\frac{(K_{priv})^\eta \left[\frac{K_{priv}}{L_{priv}}\right]^{(1-\alpha)(1-\eta)}}{\alpha(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} = \frac{1}{MRPK_{priv}} \quad (2.11)$$

I then combine (2.10) and (2.11) to get an expression for group TFPR for private (the expression for state-owned TFPR is analogous):

$$TFPR_{priv} = \frac{\left(\sum \left(\frac{A_i}{\kappa}\right)^{\frac{1-\eta}{\eta}}\right)^\eta}{(K_{priv})^{\alpha\eta} (L_{priv})^{(1-\alpha)\eta}} \quad (2.12)$$

$$\kappa = (PQ^\eta)^{-\frac{1}{1-\eta}} \quad (2.13)$$

where κ cancels out in the aggregate TFP expression.

It is important to use these expressions for the counterfactuals, rather than the existing industry-ownership averages $MRPK_{priv}$ and $MRPL_{priv}$, because the group-level outputs $P_{priv}Q_{priv}$ and $P_{SOE}Q_{SOE}$, and thus group-level harmonic average MRPL's and MRPK's will increase because adjustments towards a more optimal allocation are made. Therefore, the industry TFP if you have an allocation with equal wedges within private and public groups of firms is:

$$TFP = \left(\sum_{o \in \{priv, soe\}} \left(\frac{MRPL}{MRPL_o} \right)^{1-\alpha} \left(\frac{MRPK}{MRPK_o} \right)^\alpha \sum_{i \in o} (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.14)$$

or, equivalently:

$$TFP = \left(\sum_{o \in \{priv, soe\}} \left(\frac{TFPR}{TFPR_o} \right) \sum_{i \in o} (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.15)$$

I use these equations in the calculation as explained in the following sections.

2.4 Data and context

2.4.1 Firm-level data

My firm-level data comes from the Spark-Interfax database that contains official balance-sheet, tax, employment and ownership information at the firm-by-year level. Spark provides a firm-level panel dataset of Russian private and state-owned firms covering manufacturing, agriculture and services sectors. The panel dimension of this dataset is useful for quantifying how firms change over time and will be also crucial to my adjustment procedure to

measurement error. An additional beneficial feature of this dataset for this study is that it is firm-level and not plant-level. My goal is to study misallocation across decision-makers, which makes it crucial to identify the boundary of the firm. I also expect a lesser role of measurement error and unobserved shocks and a higher role of misallocation in a firm-level dataset, as opposed to a plant-level dataset.

I extract information on firm revenues, capital stock (as measured by book value) wage bill and payments to materials. The total number of firms that reported this information in 2018, as shown in Table 2.1, was 102,895⁹. For my analysis I only use for-profit firms, including SOEs, reducing the sample to 90,888. Only firms above 100 employees or with revenues over 800m rubles (roughly 10m USD) are legally obliged to report materials and wage bill, therefore the dataset represents medium and large for-profit firms. The value added of these firms covered 61% of Russian value added in 2018 and 18% of official employment (note that the total revenue of these firms *exceeds* Russian GDP by 1.5 times due to intermediate inputs being double-counted in the buyers' and sellers' revenues).

The table below summarizes the sample by firm groups: private for-profit firms, state-owned for-profit firms and suppliers to state-owned firms (which can be either private or state-owned themselves). Suppliers to the state and state-owned firms are defined by having supplied in the top quartile of average contract value throughout 2012-2018.

State-owned firms are defined by Spark, as listed in the official Russian statistics bureau list of SOEs, and include not only firms that are directly owned by the state (e.g. "PAO Rosneft"), but also private firms that are owned by the state-owned firms (e.g. "OOO RN-Vankor"). The total number of for-profit SOEs is 3,740 and their value added is 9% of GDP in 2018 (their revenues are 20% of GDP).

2.4.2 Sanctions on Russia 2014-2019

Sanctions were rolled out by the US and EU against Russian entities and individuals as a response to the situation in Ukraine, through years 2014-2018¹⁰. The sanctions are generally of two types: SDN (Specially Designated National) and SSI (Sectoral Sanctions Identifications). The SDN-type sanctions forbid any transaction (e.g. export, import, lending, issuing stock, leasing) with a sanctioned firm or individual, as well as any firm owned by an SDN individual or an SDN firm by more than 50 per cent (this rule is called

⁹The coverage of firms that reported all variables steadily grew, from just under 60,000 in 2012 to just over 100,000 firms in 2018

¹⁰Japan, Canada, Australia, New Zealand and Ukraine have followed the US and EU and largely repeated list of sanctions entities of the US.

"OFAC rule of 50"). Further, the sanctions freeze any assets in the United States of the SDN firm or individual. SSI sanctions instead, affect inputs: they restrict long-term (longer than 14 days) debt issuance, equity financing and transactions with any such debt of equity of the sanctioned firm¹¹.

The SSI sanctions were issued mostly against Russian banks and companies, military or double-use technology firms and companies in the oil and gas sector. However, after applying the OFAC 50% rule, the coverage extends to a large number of industries.

I create a dataset of sanctions at the firm level that includes not only the firms directly listed by the US Department of Treasury but also the historical subsidiaries of these firms as well as the subsidiaries of the firms of the SDN individuals with confirmed ownership at the time of imposition of sanctions. I use the list of firms and names and announcement dates from the US Department of Treasury announcements on the official website. Then, I add all one-level-down historical subsidiaries of these companies and business individuals with Spark Database that keeps track of historical ownership¹².

I create a dataset of 2,810 sanctioned firms, for 1,132 of which I have firm-level data at least for one year. The appendix describes the creation of the sanctioned dataset in detail. The sanctions date and indicator are based on two key sources: the official US Department of Treasury's announcements of sanctioned people and entities, and the Spark data on ownership chains. I use ownership chains to fulfil the OFAC rule of 50, which directs that any other entity owned by sanctioned entities by a total of 50% or more is also sanctioned. I match other Russian firms to directly sanctioned individuals using the full First, Middle and Last name match of the firms' reported owner, reported as owner anytime since one year before the sanctioning event¹³. Analogously, I add the majority-owned subsidiaries of directly sanctioned firms to the sample. The ownership information in Spark comes from three sources: Rosstat, the firm's annual report and the official firm registry EGRUL. I use the union of these three sources after I retrieve this information from Spark Database.

Crucially, I record the distinction between the two types of sanctions in the US¹⁴: SSI and SDN. My treatment of interest is SSI since it only negatively affects inputs, rather than

¹¹Most companies under the SSI sanctions were also treated with the US stopping certain technology exports to these companies. I consider this as still the negative capital inputs shock

¹²For individuals, the match is made using the first, middle, and last name. Sometimes, the political figures are matched with a business simply because the owners have the same name, but are different individuals. Since the list contains political figures as well, who cannot legally own business, I drop them by manually checking using open sources whether the individual matched with any firm is a business person or a political figure.

¹³I assign the sanction date to the owned companies even if they are reported as owned after the sanctioning event because there are often lags in reporting of owners

¹⁴the EU follows the US in the type of treatment with almost identical lists

inputs and outputs, and therefore, makes it straightforward to assess why the outcome of interest, MRPK, changes. In all my specifications, I control for the SDN, a complete embargo on all transactions, which affects both inputs and outputs. The SDN treatment is not made on a strict subset of the SSI, but there is an overlap of firms from both groups. I assign the year of treatment as the year of the imposition of the sanctions if the announcement happened before May that year. Otherwise, I assign the following year as the year of treatment, since the application of sanctions takes place 60 days after the announcement¹⁵

Sanctioned firms with all the subsidiaries, cover 2% in total Russian employment and 45% of value added total Russian GDP.

2.4.3 Coverage of the economy

Table 2.2 shows the coverage of the full dataset I use across the three broad sectors: Manufacturing, Services and Agriculture. The first line of each panel in this table gives the shares of the sector in the total dataset. All other lines give shares within the sector, shares in Russian GDP and Russian employment become shares in Russian sectoral GDP and employment.

Manufacturing and Services predictably take up most of the dataset in terms of value added. The Services sector has more firms that are smaller. The Services and Manufacturing sectors both have a comparable share in value added of SOEs, but Manufacturing is disproportionately more hit by sanctions in terms of value added and firm count.

2.5 Measuring firm productivity and distortions

Using the framework in the model Section 2.3, I compute $MRPK_i$, $MRPL_i$, $TFPQ_i$ and $TFPR_i$. I use book value of capital for K_i , total wage bill for L_i and firm cash revenue in that year minus cash paid to materials for $P_i Y_i$, the value added¹⁶. To compute $TFPQ_i$ and $TFPR_i$ I also need the production function parameter α . I take α as one minus the labor share in total value added for private firms in a 4-digit sector¹⁷. Finally, to calculate a model-based $TFPQ_i$ I need the elasticity of demand η . I follow Hsieh & Song (2015) and

¹⁵"Russian Sanctions Update", Morgan Lewis, April 7th, 2020

¹⁶The use of book value of capital is standard in the literature. Book value by Russian accounting includes, among other items, buildings and structures, machinery and equipment, computers, vehicles, household equipment, productive and pedigree livestock, perennial plantations. These items are subject to yearly amortization, which is usually linear.

¹⁷I drop a small number of sectors that are under 10 firms and or those that have α over 1 or under 0 in the data

Sample	Count	Share of Value Added	Share of Revenue	Share of employment	Share of Value Added in Russian GDP	Share of Revenue in Russian GDP	Share of Russian employment
Firms with all variables present (Share of full sample)	102,895	100	100	100	66	162	21
Non-for-profit firms	8,467	5	4	13	3	7	3
For-profit firms	90,888	93	93	87	61	151	18
Private for-profit firms	88,657	81	83	81	54	134	17
State-owned for-profit firms	3,726	14	12	10	9	20	2
Sanctioned firms	1,118	34	25	7	23	40	2
Suppliers to the state and to SOEs	31,299	68	66	59	45	106	12

Notes: This table reports the sample coverage for the firms in the SPARK dataset in 2018 for those firms that reported capital, materials, revenue and wage bill variables in 2018. An observation is at the firm level. Russian GDP in columns "Share of Value Added in Russian GDP" and "Share of Revenue in Russian GDP" and Russian employment in column "Share of Russian employment" are taken from Rosstat for the year 2018.

Table 2.1: Sample used for analysis

use $\eta = 0.143$, which corresponds to the elasticity of substitution of 7. Using the values of α and η , I use equations 2.2, 2.3, 2.4 and 2.5 to calculate TFP_i , $MRPK_i$, $MRPL_i$ and $TFPR_i$ for each firm in each year.

The measures calculated this way are prone to measurement error in inputs and outputs (Bils et al. (2020), Rotemberg & White (2017), Gollin & Udry (2021)). Even non-systematic measurement error will result in higher measured misallocation and higher gaps between real and efficient TFPs. I apply a state-of-the-art method to adjust for measurement error. I start with the baseline approach and winzorise top and bottom 1% of firm observations in their $TFPR_i$ and the model-based productivity measure $TFPQ_i$. As an alternative, I also follow Adamopoulos et al. (2017) and regress the $TFPQ_i$ and $TFPR_i$ on firm and year fixed effects. This removes the transient shocks short-term measurement error in inputs and outputs and gives me the time-invariant firm productivity and wedges. The regressions I run to correct for measurement error are shown below:

$$\ln(TFPQ_i) = \beta_0^{TFPQ} + \gamma_t^{TFPQ} + \phi_i^{TFPQ} + \epsilon_{it}^{TFPQ} \quad (2.16)$$

$$\ln(TFPR_i) = \beta_0^{TFPR} + \gamma_t^{TFPR} + \phi_i^{TFPR} + \epsilon_{it}^{TFPR} \quad (2.17)$$

Here β_0^{TFPQ} and β_0^{TFPR} are common intercepts, γ_t^{TFPQ} and γ_t^{TFPR} are the year fixed effects that capture time-varying shocks, such as a common component in trends in mark-

Sample	Count	Share of Value Added	Share of Revenue	Share of employment	Share of Value Added in Russian GDP	Share of Revenue in Russian GDP	Share of Russian employment
Manufacturing							
Firms with all variables present (Share of full sample)	22,681	47	39	42	31	63	9
Non-for-profit firms	2,550	4	4	11	5	9	6
For-profit firms	19,293	94	94	89	106	215	44
Private for-profit firms	18,767	81	78	83	91	178	41
State-owned for-profit firms	869	15	18	8	17	40	4
Sanctioned firms	418	42	32	10	47	72	5
Suppliers to the state and to SOEs	8,688	78	74	66	88	168	33
Services							
Firms with all variables present (Share of full sample)	71,312	51	59	51	33	95	11
Non-for-profit firms	5,555	7	5	15	4	7	2
For-profit firms	63,388	91	93	85	52	152	12
Private for-profit firms	61,883	81	86	78	46	141	11
State-owned for-profit firms	2,510	13	9	12	7	15	2
Sanctioned firms	678	29	21	6	16	35	1
Suppliers to the state and to SOEs	21,725	60	62	59	34	101	8
Agriculture							
Firms with all variables present (Share of full sample)	8,902	2	2	7	1	3	2
Non-for-profit firms	362	3	4	4	1	3	1
For-profit firms	8,207	93	93	96	37	80	21
Private for-profit firms	8,007	91	91	93	36	79	21
State-owned for-profit firms	347	3	2	5	1	2	1
Sanctioned firms	22	1	1	1	0	1	0
Suppliers to the state and to SOEs	886	21	21	21	8	18	5

Notes: This table reports the sample coverage for the firms in the SPARK dataset in 2018 for those firms that reported capital, materials, revenue and wage bill variables in 2018. An observation is at the firm level. Russian sectoral GDP in columns "Share of Value Added in Russian GDP" and "Share of Revenue in Russian GDP" and Russian sectoral employment in column "Share of Russian employment" are taken from Rosstat for the year 2018. The first row of every panel represents the share of the sector in the full sample. All other rows represent the shares within each sector.

Table 2.2: Sample used for analysis

ups or oil prices, and ϕ_i^{TFPQ} and ϕ_i^{TFPR} are the firm fixed effects and incorporates all firm-sector components. Finally, ϵ_{it}^{TFPQ} and ϵ_{it}^{TFPR} are the errors, including the transient measurement error and adjustment costs and noise. Analogously to how Adamopoulos et al. (2017) remove the village-specific component, I separate the firm effect from the sector component by (1) estimating equation 2.16 and extract the firm fixed effect inclusive of the sector fixed effect (2) regressing these fixed effects on 4-digit-sector dummies to extract the residuals that are the pure permanent firm $\ln(TFPQ_i)$ and $\ln(TFPR_i)$ components. The $TFPQ_i$ and $TFPR_i$ are the exponentials of the residual, after regressing the firm fixed effects on industry dummies.

For the counterfactual exercises, I follow this procedure using the full panel 2012-2018, including the period of sanctions. I get the measures of firm $TFPQ_i$ and $TFPR_i$ that do not change over time and do not differ across sectors¹⁸. The firm fixed effect estimate controls for transient measurement error which is absorbed by the residual. I calculate the counterfactual results with this procedure, but also include the winzorised results based on raw data in the following sections. As expected, the dispersion of the adjusted measures of firm TFP and TFPR is lower than that of the unadjusted measures.

2.6 Static misallocation in Russia

Table 2.3 is a summary table of all variables used in the current exercise. Each observation is firm-year. The sample has 602,926 observations, which is 194,095 firms and 897 industries. The typical firm is a domestic firm. There are 1,132 firms under any sanctions, of which 498 firms are under the input sanctions specifically, which is 0.96% of the dataset. State-owned firms add up to 4,378 and represent 3.6%. The variables include value added, capital, wage bill, materials bill, employment, age and a type of firm. I additionally include the statistics from the Hsieh & Klenow (2009) model (HK), each divided by the sector harmonic average: firm TFPQ, TFPR, MRPK, MRPL. I also include versions of these variables that are adjusted for the measurement error using firm and year fixed effects. All the balance sheet variables are in 1000s of Rubles.

In addition, in Table 2.4 I show comparable statistics to those reported in HK so that the key measures from the model can be cross-checked. Before adjusting for measurement error, I find that in Russia the dispersions of both TFPR and TFPQ are substantially larger than what HK find in China and India. HK report the p75-p25 variation in $\ln(TFPQ)$

¹⁸When I quantify the aggregate effect of sanctions I will use an equivalent approach to get the pre-treatment wedges, but for years 2012-2014, the pre-period.

(1)

	count	mean	sd	min	max
Value added, 1000 rub	589,236	332,453	10,633,095	-2,578,904,576	2,560,027,136
Book value of capital, 1000 rub	602,926	568,348	27,645,770	-24,443	7,882,970,562
Payment to labor, 1000 rub	602,926	111,338	1,774,229	-312,268	499,737,000
Materials, 1000 rub	602,926	1,078,618	19,462,179	-122,617,309	4,820,693,835
Labor count, latest year	537,942	174	587	0	16,757
Firm age, yrs	566,257	16	7.3	0	93
Private firm dummy	580,930	1	0	1	1
SOE dummy	602,926	.036	.19	0	1
Foreign-owned firm dummy	602,926	.00023	.015	0	1
Suppliers to state and SOEs dummy	602,926	.31	.46	0	1
Firm under any sanction	602,926	.012	.11	0	1
Firm under input sanction	602,926	.0092	.096	0	1
Firm TFPQ, weighted by sector	415,504	.2	.31	0	4
Firm TFPQ, adjusted for measurement error	366,398	1.7	3.3	0	201
Firm TFPR, weighted by sector	415,504	3.1	7.9	0	239
Firm TFPR, adjusted for measurement error	366,398	1.6	2.6	0	93
Firm MRPL, weighted by sector	415,504	3.5	153	0	73,320
Firm MRPL, adjusted for measurement error	366,398	2.3	88	0	34,215
Firm MRPK, weighted by sector	415,504	52	1,229	0	364,920
Firm MRPK, adjusted for measurement error	366,398	6.6	383	0	162,614
Observations	602,926				

Notes: This table reports summary statistics for the firms in the SPARK dataset from 2012 to 2018. An observation is at the firm-year level. Firms' book value of capital, value added, payments to labor, materials and revenues are measured in 1000 of Rubles.

Table 2.3: Summary statistics of key variables

Panel A : Full dataset				
Variable	Statistic	Industry and Firm Fixed Effects	2018 Raw measures	Cross-section Average
ln(TFPR)	SD	0.86	1.13	1.13
	p75-p25	0.91	1.17	1.16
	p90-p10	2.03	2.68	2.66
ln(MRPL)	SD	0.77	1.02	1.02
	p75-p25	0.65	0.89	0.89
	p90-p10	1.60	2.18	2.16
ln(MRPK)	SD	1.66	2.03	2.03
	p75-p25	1.93	2.41	2.39
	p90-p10	4.09	5.07	5.04
ln(TFPQ)	SD	0.94	1.50	1.49
	p75-p25	1.03	2.09	2.07
	p90-p10	2.24	3.86	3.84
Panel B : Only the manufacturing sector				
Variable	Statistic	Industry and Firm Fixed Effects	2018 Raw measures	Cross-section Average
ln(TFPR)	SD	0.76	0.98	0.98
	p75-p25	0.82	1.00	0.99
	p90-p10	1.78	2.25	2.25
ln(MRPL)	SD	0.61	0.83	0.83
	p75-p25	0.53	0.71	0.71
	p90-p10	1.18	1.68	1.66
ln(MRPK)	SD	1.49	1.77	1.76
	p75-p25	1.72	1.98	1.97
	p90-p10	3.61	4.26	4.24
ln(TFPQ)	SD	0.83	1.34	1.32
	p75-p25	0.90	1.83	1.77
	p90-p10	1.95	3.40	3.35

Notes: For firm i , $TFPQ_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}}$. Statistics are for deviations of $\log(TFPQ)$ from industry means. SD = standard deviation, p75 p25 is the difference between the 75th and 25th percentiles, and p90 p10 the 90th vs. 10th percentiles. Values in the column "Industry and Firm Fixed Effects" are adjusted for measurement error using firm and year fixed effects and de-measured by 4-digit industry averages. Values in the column "2018 Raw measures" are the logs of raw measures of $TFPQ_i$, $MRPK_i$, $MRPL_i$, $TFPR_i$ for each firm, divided by the harmonic average of the same measure in the 4-digit industry. Values in the column "Cross-section Average" are the average of the statistics calculated as in the previous column, but the statistics are calculated for each cross-section of the panel 2012-2018 and then averaged across years. Panel A is calculated for the full sample of for-profit firms, and Panel B is calculated for the Manufacturing sector only.

Table 2.4: Dispersion of $\ln(TFPR)$, $\ln(TFPQ)$, $\ln(MRPL)$, $\ln(MRPK)$

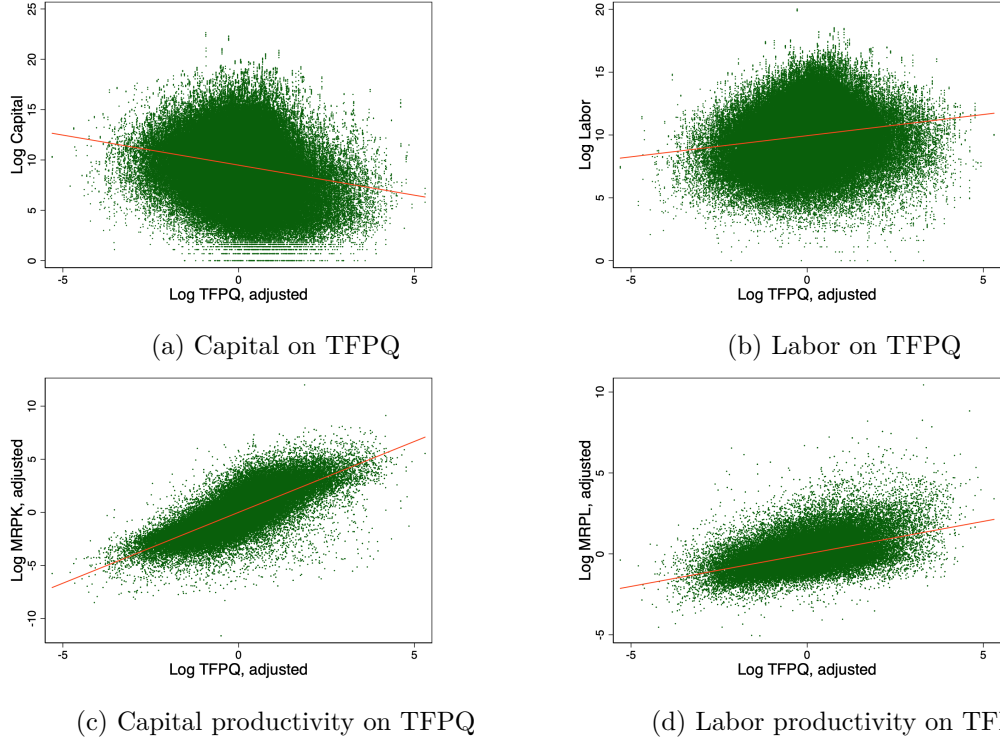
of 1.28 and p90-p10 of 2.44 for China in 2005, while for India the corresponding values are 1.60 and 3.11. In Russia, without measurement error adjustment, the 2018 $\ln(TFPQ)$ variation is: p75-p25 is 2.14 and p90-p10 is 3.49.

Equally, $\ln(\text{TFPR})$ variation in Russia is also larger: I find 1.18 and 2.71 in Russia in 2018 compared to 0.82 and 1.59 (China), 0.81 and 1.60 (India). However, Hsieh and Klenow only use the manufacturing sector, whereas my data include services and agriculture, and the diverse services sector can show much more variation in wedges and productivity¹⁹. Looking at Panel B, with only the manufacturing sector, the percentile variation in $\ln(\text{TFPR})$ (1.00 and 2.25) and $\ln(\text{TFPQ})$ (1.83 and 3.40) reduces but is still larger than in HK. Additionally, adjusting these measures for firm and year fixed effects further reduces the variation and gives the percentile variation of $\ln(\text{TFPR})$ (0.82 and 1.78) and $\ln(\text{TFPQ})$ (0.90 and 1.95) making the values on par or even smaller than numbers found in Hsieh and Klenow for India and China.

Do resources in Russia appear misallocated through the lens of the framework from the "Model" Section 2.3? If capital and labour markets were not distorted, more capital and labour would flow to the relatively more productive firms. This means that input use and firm TFP should be positively related, while the marginal revenue products of labour and capital should be unrelated to firm TFP because inputs flow to more productive firms up until these marginal products are equalised. Likewise, the revenue productivity, TFPR, which is the summary measure of MRPK and MRPL, should be unrelated to physical productivity, TFPQ.

In Russia, I observe different patterns. Figure 2.1 demonstrates the overall distribution of capital and labour relative to the productivity of firms (the top two graphs), and the capital and labour productivity on firm TFPQ (the bottom two graphs), and the measures of TFPQ, capital and labour productivity are adjusted for measurement error. In the top two graphs, the firm productivity is shown on the X-axes and the inputs on the Y-axes. In an efficient economy, the slopes of the relationships between productivity and inputs are positive. In Russia, on the contrary, we see that at least capital to be lower on average in more productive firms. On the second row, I plot MRPL and MRPK relative to the firm TFPQ, where the efficient relationship should be flat and the marginal revenue of each input should be equalized across firms. Again, it is evident that more productive firms face larger positive wedges, this time in both capital and labour. Both relationships - between TFPQ_i and MRPK_i , and between TFPQ_i and MRPL_i are positive, while in an efficient economy there should be no correlation between TFPQ and labour or capital productivity.

¹⁹The higher variation may also arise because of the way 4-digit industries are defined. As the country is transforming to the services economy, the level of detail may be much lower in the services sector, relative to manufacturing, so each 4-digit industry in the services sector may contain somewhat more diverse firms than a 4-digit industry in manufacturing



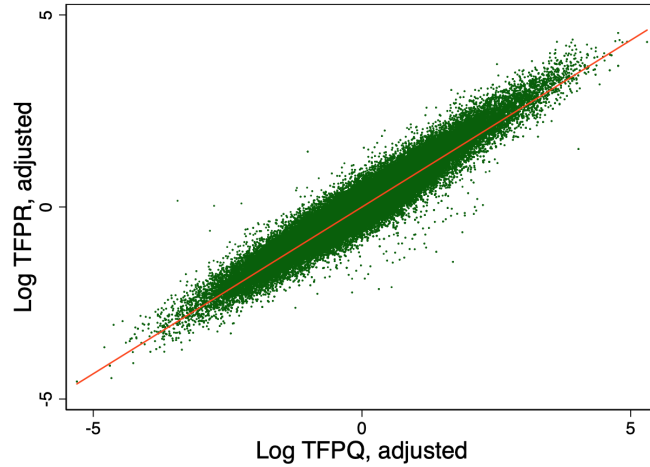
(a) Capital on TFPQ (b) Labor on TFPQ
(c) Capital productivity on TFPQ (d) Labor productivity on TFPQ

Notes: Each observation (green dot) is a firm. Labour productivity (or $MRPL_i$) refers to value added per unit of wage bill and capital productivity (or $MRPK_i$) refers to value added per unit of capital, both of which are proportional to the marginal products of each factor in my framework. Raw TFPQ is calculated using the expression $TFPQ_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}}$. The MRPK, MRPL and TFPQ measures are adjusted for measurement error with firm and year fixed effects and de-meaned by 4-digit industry using the firm panel 2012-2018. The solid orange line is the line of best fit.

Figure 2.1: Factor allocations by firm productivity

These patterns point at large institutional and economic frictions that prevent the flow of labour and capital resources to the most productive firms.

Both capital and labour distortions to a firm can be summarised with a $TFPR_i$, the revenue productivity measure, defined in equation 2.4. This measure will help us see whether firms that face high capital wedges, also face high labour wedges. As described in section 2.5, just like I do for $TFPQ_i$, I adjust the $TFPR_i$ for each firm with year and firm fixed effects and further regress the residuals on the 4-digit industry dummies. Figure 2.2 shows firm $TFPQ_i$ on X-axes and firm $TFPR_i$ on Y-axes. The very strong correlation of $TFPR_i$ and firm physical productivity tells us that more productive firms face higher wedges in *both* labour and capital. This confirms our findings above. Firms that experience high productivity do not have a scope to grow because both capital and labour flows to less productive firms. These less productive firms could be the firms under state protection. Equally, higher distortions in more productive firms could also come from the market power



Notes: Each observation is a firm. $TFPR_i$, or revenue productivity, is a summary measure of distortions faced by each firm, with higher $TFPR_i$ implying higher distortions. The TFPR measure is adjusted for measurement error with firm and year fixed effects and de-meanned by 4-digit industry using the firm panel 2012-2018.

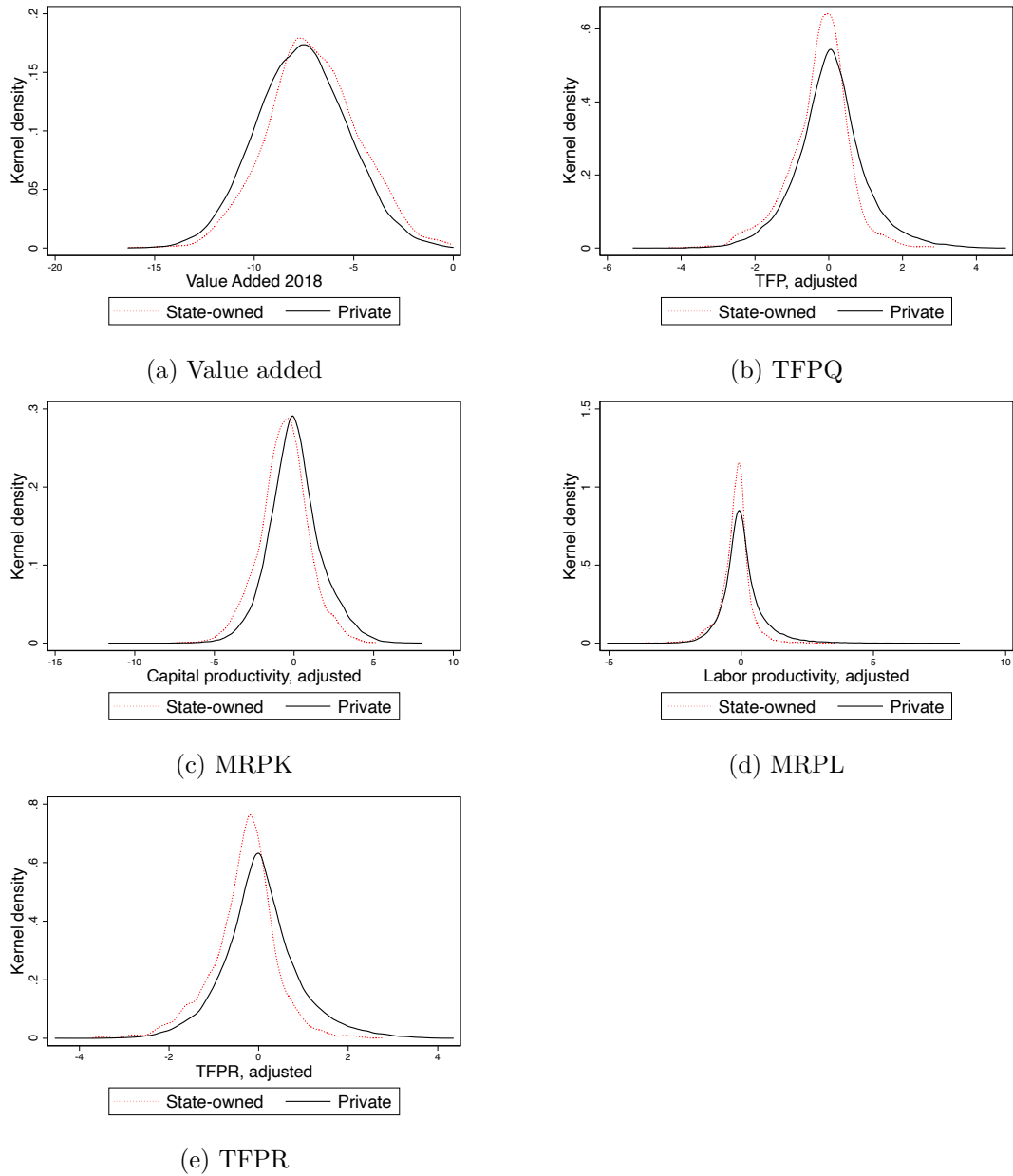
Figure 2.2: Firm-specific distortions and productivity (TFPR on TFPQ)

of those productive firms, and export tariffs that prevent these firms' expansion into foreign markets. This paper studies how much of this relationship is explained by the state taking away capital and labour from more productive private firms and giving it to less productive SOEs.

Above were the descriptives of the firm characteristics of the whole economy. So far we have not seen any information about the state-owned sector versus the private sector. Is there any misallocation across ownership groups? Could such distortions explain, at least in part, the barriers faced by more productive firms? Figure 2.3 below compares the density distributions of the firm-level value added, TFPQ, capital and labour productivity and TFPR between state-owned firms and private firms. These measures of TFPQ, MRPK, MRPL and TFPR are adjusted for measurement error using firm and year fixed effects as explained in Section 2.5.

These figures demonstrate that the state-owned firms have much lower TFPR than the private sector, especially at the high end of the distribution. Lower revenue productivity arises due to both too much capital allocated to state-owned firms and too much labour, as evidenced by subplots (c) and (d). Private firms also appear to be more productive, as shown in subplot (b). Therefore, I again witness that the more productive firms face larger "correlated" distortions. I also note that the state-owned firms are relatively large in terms of value added, compared to private firms (subplot a). Such allocation of capital and labour, excessive from the efficiency perspective, can come from the soft budget constraint of the SOEs. As for the excess labour, since labour is complementary to capital to some extent, more labour could be employed as a result of excessive capital. On top of that, some labour hoarding could be still taking place in some state-owned enterprises, if they are the only main employers in a city - which is Soviet heritage²⁰.

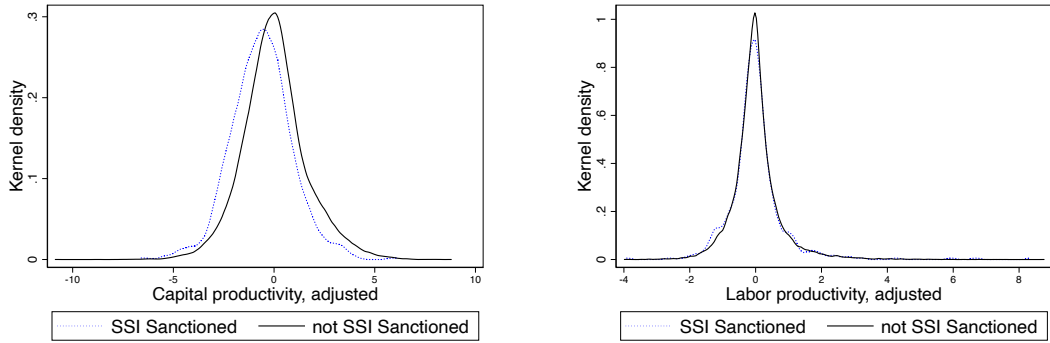
²⁰The labour hoarding may be desirable from the equity perspective, but just not from the efficiency perspective.



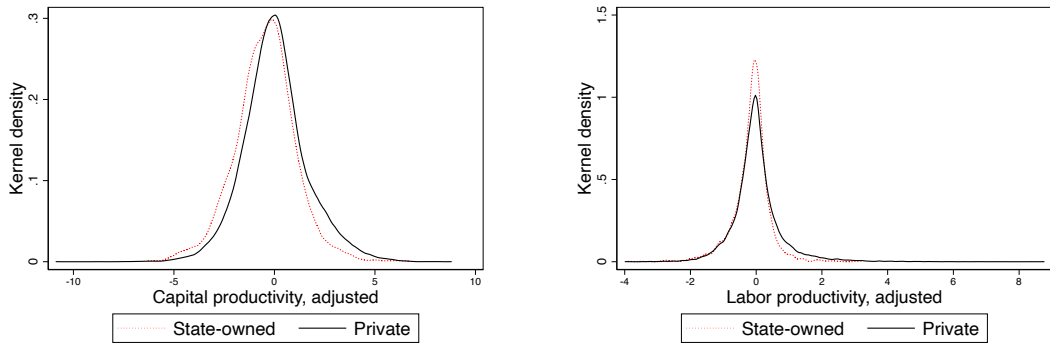
Notes: The plots show the kernel density of natural logs of value added, TFPQ, MRPK, MRPL and TFPR. The red dotted lines are the kernel densities for the SOEs sample. The black lines are the kernel densities for the sample of private firms. Labor productivity (or $MRPL_i$) refers to value added per unit of wage bill and capital productivity (or $MRPK_i$) refers to value added per unit of capital, both of which are proportional to the marginal products of each factor in my framework. Raw TFPQ is calculated using the expression $TFPQ_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}}$. The MRPK, MRPL and TFPQ measures are time invariant, because they are adjusted for measurement error with firm and year fixed effects and de-meanned by 4-digit industry using the firm panel for 2012-2018.

Figure 2.3: Allocations of SOEs versus the private sector, variables adjusted for measurement error

Linking these observations with the section on sanctions, I also show the differences in Capital and Labor productivity between SOEs and private firms before the sanctions episode, again, applying the measurement error adjustment procedure from section 2.5 and confirm that the SOEs were already "too large" before the sanctions. Moreover, the sanctioned firms (that experienced input sanctions) taken together also were "too large" from the efficiency perspective before the treatment, at least in terms of capital²¹. Since the sanctioned firms were chosen by the US intelligence services as connected to the current government, this finding points out that there is potential misallocation not only across ownership status but also between connected firms and all other firms.



(a) MRPK of sanctioned versus other firms pre-2015 (b) MRPL of sanctioned versus other firms pre-2015



(c) MRPK of SOEs versus private firms pre-2015 (d) MRPL of SOEs versus private firms pre-2015

Notes: The plots show the kernel density of natural logs of MRPK and MRPL. The blue dotted lines are the kernel densities for the SSI sanctioned sample. The black lines in the top two graphs are the kernel densities for the sample of non-SSI sanctioned firms. The red dotted lines are the kernel densities for the SOEs sample. The black lines in the bottom two graphs are the kernel densities for the sample of private firms. Labour productivity (or $MRPL_i$) refers to value added per unit of wage bill and capital productivity (or $MRPK_i$) refers to value added per unit of capital, both of which are proportional to the marginal products of each factor in my framework. The MRPK and MRPL measures are time-invariant because they are adjusted for measurement error with firm and year fixed effects and de-measured by 4-digit industry using the firm panel for 2012-2014.

Figure 2.4: Allocations before 2015

²¹The graph is similar for firms that experienced any sanctions, including the blocking ones

2.7 Counterfactuals

To measure efficiency gains of reallocating resources across ownership groups I conduct two counterfactual exercises. First, I equalize all wedges (or TFPR) across firms within each four-digit industry, keeping total capital and labour fixed within industries. I then compare the aggregate TFP as measured in the data to this new efficient TFP, call it "TFPe". This comparison will give a full distance to the efficient frontier from the current status quo in Russia. Second, I equalize wedges only within ownership-by-industry groups and compare the resulting TFP, call it TFPc, to the TFPe from the first exercise. The remaining distance to the frontier is attributed to the wedges across SOEs and private firms.

- 1) TFPe: Equalize all wedges within industries
- 2) TFPc: Equalize wedges within ownership-industry groups

Measures	Count	TFP/TFPe	TFPc/TFPe
Raw	71,180	8.7%	91.2%
FE-corrected	57,279	49.9%	94.7%
Raw, same sample as corrected	57,279	8.1%	94.5%

Notes: Column 1 reports the sample sized used for the counterfactual calculation. Column 2 reports the shares of the existing aggregate output (TFP) in the efficient output (TFPe), or TFP/TFPe. Column 3 reports the shares of the counterfactual aggregate output (TFPc), after equalising the wedges within ownership-industry groups, or TFPc/TFPe. "Raw" refers to the the raw data in 2018. "FE-corrected" refers to the fixed effects estimates from the panel regression. "Raw, same sample as corrected" refers to the TFP shares for the same sample as used to get the the fixed effects estimates, but without the actual correction.

Table 2.5: Counterfactual exercises

Table 2.5 shows the results of the counterfactual exercises. The first row uses the data for the year 2018, without adjusting for measurement error. The resulting overall distance to the frontier is very large: the current TFP is more than 10 times smaller than the frontier TFP. The remaining distance to the frontier due to wedges across the SOEs and private firms is roughly 9% (100%-91.2%), so according to this result, the current TFP will double (8.7%+9%) if an SOE versus private wedge was removed while all other distortions remained. However, most of the variation is unexplained by the SOEs-private wedge.

The overall distance to the frontier gets smaller when I correct for measurement error on the second row of Table 2.5. Now the Russian TFP will slightly more than double if all wedges were equalized. Now, the wedges across ownership groups appear smaller and will add roughly 11% (5.3%/49.9%=11%) to the current TFP if they are removed. Again, the

bulk of wedge variation that keeps Russia at a distance from its efficient frontier remains unexplained and the ownership wedge only explains $5.3\%/51.1\%=10.4\%$ of the distance to the frontier. This is not surprising: many factors, such as different forms of corruption or supplier to SOE status can contribute to misallocation of resources above and beyond the simple ownership wedge and in a companion paper I explore to what extent these factors help to further explain the distance to the frontier in Russia.

2.8 Sanctions as a test of the SOE protection

Sanction type	Ownership		Total
	Private	State-owned	
SDN	224	53	277
SSI	348	49	397
SSI and SDN	382	76	458
Total	954	178	1,132

Notes: This table is a cross-tabulation of the sanctioned firms (reporting balance sheet data) by ownership. SDN is the group of firms that are sanctioned by blocking sanctions, SSI indicated the group of firms sanctioned by input sanctions. The sample includes firms that are sanctioned by association with the directly sanctioned firm via majority ownership.

Table 2.6: Sanctions by ownership

Sanction type	Sector			Total
	Manufacturing	Services	Agriculture	
SDN	102	174	1	277
SSI	134	254	9	397
SSI and SDN	178	268	12	458
Total	414	696	22	1,132

Notes: This table is a cross-tabulation of the sanctioned firms (reporting balance sheet data) by sector. SDN is the group of firms that are sanctioned by blocking sanctions, SSI indicated the group of firms sanctioned by input sanctions. The sample includes firms that are sanctioned by association with the directly sanctioned firm via majority ownership.

Table 2.7: Sanctions by sector

Table 2.8 shows the summary of the key variables by sanction type and compares the averages of these key variables. The sanctioned firms, either SSI, SDN or both are larger in terms of average value added, total revenue, the book value of capital and wage bill. The average raw MRPK is lower in the sanctioned firms relative to non-sanctioned firms, as expected. However, it is also important to look at the whole distribution of this variable, rather than the simple average, which masks substantial heterogeneity which we saw in figure 2.4.

Assuming politically connected SOEs and private firms already have "too much capital", the

	(1)				
	Not sanctioned	SDN	SSI	SSI and SDN	Total
Sanctioned as a subsidiary dummy	0 (0)	0.850 (0.357)	0.649 (0.477)	0.890 (0.313)	0.00957 (0.0973)
Private firm dummy	0.965 (0.184)	0.785 (0.411)	0.874 (0.332)	0.836 (0.370)	0.964 (0.187)
SOE dummy	0.0349 (0.184)	0.215 (0.411)	0.126 (0.332)	0.164 (0.370)	0.0365 (0.187)
Direct sanction dummy	0 (0)	0.150 (0.357)	0.351 (0.477)	0.110 (0.313)	0.00245 (0.0495)
Ln value added	10.47 (1.969)	13.08 (2.193)	13.32 (2.548)	13.26 (2.492)	10.50 (1.998)
Ln revenue	11.67 (2.185)	13.92 (2.562)	14.01 (2.939)	14.03 (2.843)	11.70 (2.209)
Ln book value of capital	9.167 (2.781)	12.06 (3.022)	12.48 (3.559)	12.56 (3.205)	9.206 (2.810)
Ln payment to labor	9.474 (2.014)	12.21 (2.137)	12.12 (2.346)	12.21 (2.327)	9.507 (2.039)
Ln materials	11.19 (2.355)	13.32 (2.638)	13.21 (2.873)	13.32 (2.828)	11.22 (2.372)
Labor count, latest year	162.6 (543.7)	1109.4 (1884.3)	942.4 (1822.7)	1253.3 (1876.9)	173.9 (586.6)
Firm age, yrs	15.99 (7.285)	19.76 (7.179)	19.47 (7.125)	20.36 (7.758)	16.04 (7.300)
Foreign-owned firm dummy	0.000237 (0.0154)	0 (0)	0 (0)	0 (0)	0.000234 (0.0153)
Suppliers to state and SOEs dummy	0.302 (0.459)	0.679 (0.467)	0.674 (0.469)	0.743 (0.437)	0.307 (0.461)
Ln firm MRPK	1.288 (2.467)	0.882 (2.278)	0.579 (2.456)	0.442 (2.229)	1.279 (2.466)
Ln firm MRPL	0.814 (1.221)	0.637 (1.116)	0.908 (1.481)	0.776 (1.380)	0.814 (1.223)
Observations	602926				

Notes: This table reports summary statistics for the firms in the SPARK dataset from 2012 to 2018 by type of sanction. An observation is at the firm-year level. SDN is the group of firms that are sanctioned by blocking sanctions, SSI indicated the group of firms sanctioned by input sanctions. The sample includes firms that are sanctioned by association with the directly sanctioned firm via majority ownership. The share of the indirectly sanctioned firms is shown by the statistics for the "Sanctioned as a subsidiary dummy" variable.

Table 2.8: Summary by sanction type

first hypothesis is that sanctions, hitting the inputs would reduce misallocation. However, there is anecdotal evidence that the politically connected firms, both private and state-owned, managed to secure more funding from the Russian government as a response to sanctions. Sberbank, Russia’s largest state bank had the central bank purchase a significant amount of the bank’s new debt since sanctioning. Viktor Vekselberg, Renova Group’s owner has had the credit line extended by Promsvyazbank in 2018²². Leonid Mikhelson has been reported to request the government to help fund the creation of deepwater drilling equipment to replace the U.S. imports²³. Promsvyazbank was nationalized and then repurposed to compensate the losses from sanctions of Russia’s defence sectors²⁴. By 2015 the Russian state started a bank recapitalization program worth about 1.4 trillion rub, or 1.2% of GDP to support all banks directly or indirectly affected by the sanctions.²⁵ Further, the government strategically granted contracts to sanctioned firms, it provided sanctioned Bank Rossiya the sole contract to service the \$36 billion domestic wholesale electricity market, granted the contract to build a bridge linking the Russian mainland with Crimea to a sanctioned construction company (Stroygazmontazh), and selected a sanctioned bank (VTB) to be the sole manager of the government’s international bond sales.²⁶ Therefore, due to this governmental response, the misallocation may have actually worsened on the net after sanctions were imposed.

The SSI sanctions were imposed on groups of Russian firms in waves every year starting effectively from 2015. The staggered experiment of SSI sanctions allows me to test the joint effect of the negative input shock and the government response. I run the following regression:

$$Y_{it} = \gamma_{jt} + \phi_i + \theta_{st} + \beta_1 * InputSanctions_{it} + X_{it}\delta + u_{ijt} \quad (2.18)$$

I use the annual measures of $\ln(MRPK_{it})$, $\ln(ValueAdded_{it})$, $\ln(Revenue_{it})$ or $\ln(K_{it})$ for Y_{it} and regress these variables on firm-level time-variant sanctions dummy. The sanctions

²²<https://www.reuters.com/article/us-russia-renova-idUSKCN1IF2AG>

²³<https://www.bloomberg.com/opinion/articles/2018-05-08/russia-sanctions-have-had-some-unexpected-consequences+cd=1hl=enct=clnkg1=ruclient=safari>

²⁴Max Seddon, “Moscow Creates Bank To Help It Avoid US Sanctions,” Financial Times, January 19, 2018, <https://www.ft.com/content/90c73fe4-fd15-11e7-9b32-d7d59aace167>

²⁵IMF, Russian Federation: Staff Report for the 2015 Article IV Consultation, August 2015, pp. 7. <https://www.imf.org/external/pubs/ft/scr/2015/cr15211.pdf>

²⁶Moscow Times, “Sanctioned Bank Rossiya Becomes First Major Russian Bank to Expand in Crimea,” April 15, 2017; Jack Stubbs and Yeganeh Torbati, “U.S. Imposes Sanctions on ‘Putin’s Bridge’ to Crimea,” Reuters, September 1, 2016; Thomas Hale and Max Seddon, “Russia to Tap Global Debt Markets for a Further \$1.25 Billion,” Financial Times, September 22, 2016. See the Congressional Research Service (2020), pp 53 for a more extensive list of measures by the Russian Government. <https://fas.org/sgp/crs/row/R45415.pdf>

variable of interest is the SSI sanctions, which was targeted to inputs alone. In every specification, I also control for the SDN sanctions, which are included in X_{it} to account for the fact that some firms were also treated by SDN ("blocking") sanctions in both the treated and control groups. To control for firm-level heterogeneity I include firm FE ϕ_i . Further, I add a 4-digit industry-year FE γ_{jt} to remove common industry changes over time, including the oil price shocks that were large in the period 2014-2016 and could have differentially affected some industries, which also have more sanctioned firms. Moreover, I include a size-by-year fixed effects θ_{st} to difference out the trends that larger firms experience as opposed to smaller firms. The size s is defined by the pre-treatment quartile of average firm capital. I cluster the errors by firm and 4-digit industry-by-year to account for possible serial correlation at firm level or across firms within an industry at a given point in time.

If β_1 is negative and significant and Y_{it} is $\ln(\text{MRPK})$ in specification 2.18, this is the evidence that sanctioned firms, which already had "too much capital" received relatively more capital as a result of sanctions. This result can appear not just because the capital inputs grew, but also because the input-sanctioned firms had more inputs *relative* to the value added. But what if the value added dropped for these firms, due to some de-risking by their foreign customers? If I further find that $\ln(\text{MRPK})$ increased because the inputs grew more rather than because the value added dropped (for instance, by β_1 being non-negative when Y_{it} is $\ln(\text{ValueAdded}_{it})$ and by β_1 being positive and significant when Y_{it} is $\ln(K_{it})$), this will be the evidence of shielding of sanctioned firms that over-shot the direct (negative) effect of input sanctions on inputs.

This experiment also helps me see whether the SOEs have responded differently to this negative input shock as opposed to private firms. The finding that β_1 is negative alone is evidence that misallocation increased on average for the sanctioned firms, but no distinction is made about the response of SOEs versus private firms. To separate the effect of political connections driving misallocation versus the state ownership driving misallocation I run the following regression:

$$Y_{it} = \gamma_{jt} + \phi_i + \theta_{st} + \beta_1 * \text{InputSanctions}_{it} + \beta_2 * \text{InputSanctions}_{it} * \text{SOE}_i + X_{it} \delta + u_{ijt} \quad (2.19)$$

In Specification 2.19, I repeat the specification 2.18 but add an interaction term $\text{Sanctions}_{it} * \text{SOE}_i$ to check if there is a differential effect with respect to the state owned firms. If in Specification 2.19 we see the evidence of only the SOEs being saved, β_1 will be zero and β_2 will be negative when Y_{it} is $\ln(\text{MRPK})$. This will show that misallocation got worse through the act of protection of the SOEs alone.

Identification. Below, I discuss the extent to which my estimation is prone to two possible sources of bias: (1) non-random assignment of sanctions across firms, and (2) measurement error in sanctions and SOE status.

One worry is that sanctioned firms have different characteristics relative to non-sanctioned firms. As shown in Table 2.8, the sanctioned firms have higher revenues, capital, employ more people and are on average four years older than the non-sanctioned firms and there may also be unobserved differences between these firms. However, so long as these observed or unobserved differences are time-invariant, these differences are fully accounted for by firm fixed effects. The firm fixed effects also account for any differences between SOEs and private firms. Therefore, this empirical strategy does not require that the sanctions were randomly assigned.

Another concern is that the SSI sanctions were over-represented in some industries, such as the Oil and Gas sector, which also differentially experienced a negative oil price shock in the same period. So long as these shocks affected firms within a narrow 4-digit industry similarly, my industry-by-year fixed effect fully controls for these time-variant industry shocks.

Therefore, this set-up does not require that the industries that had more sanctioned firms evolve in parallel over time, and it does not require that the sanctioned and non-sanctioned firms share the same time-invariant characteristics. The estimation of β_1 in Specifications 2.18 and 2.19 does rely on the classic assumption that the sanctioned firms evolve in parallel to the non-sanctioned firms at the time of sanctioning. I provide visual evidence that the pre-trends evolved in parallel in the next section.

The estimation of β_2 in Specification 2.19 requires that SOEs are trending in parallel to private firms. Such differential trends can be controlled for. In the Appendix Table 2.A1, I control for the SOE-by-year fixed effects to absorb the bias from SOEs trending differently to private firms. In effect, Specification 2.19 after additionally controlling for SOE-by-year fixed effects becomes a triple difference regression. I show that β_1 and β_2 do not change from including the SOE-by-year fixed effects. As a result, I can still identify β_2 if treated and untreated industries have different industry-level time trends, as they are controlled by the time-variant $InputSanctions_{it}$ dummy. I can also still identify β_2 if SOEs and private firms are trending differently, as these are absorbed by SOE-by-year fixed effects. I rest on a milder assumption to identify β_2 : the differential between the sanctioned SOEs and sanctioned private firms need to evolve in parallel to that differential in the non-sanctioned group.

Measurement error in $MRPK_i$, the outcome variable, is not a great concern in the estimations I present. First, the non-systematic measurement error on the outcome variable $MRPK_i$ does not bias the coefficients that I find. If the measurement error is systematic, but fixed at firm-level, or is time-variant, but common for all firms in a 4-digit industry, it will be absorbed by the industry-by-year fixed effects and firm fixed effects. Only the non-classical measurement error that varies by sanction and SOE status may be an issue. However, if anything such a hypothetical error is likely to work against me finding the shielding effects: the SOEs and other sanctioned firms may be motivated to under-report the capital that is received as a result of shielding.

2.8.1 Event studies

As mentioned above, to identify β_1 in Specifications 2.18 and 2.19 I rest on the assumption that the sanctioned firms would have been on the same trends as the non-sanctioned firms at the time of sanctioning. To partially alleviate this concern, I include event studies that 1) test for sanction effect within sanctioned firms (Specification 2.20) and identifying the treatment effect off timing 2) test for the differential trends between sanctioned and non-sanctioned firms before 2015, the first year of sanctions taking an effect (2.21)²⁷.

$$Y_{it} = \gamma_{jt} + \phi_i + \theta_{st} + \alpha_s * \sum_{s=-4, s \neq 0}^{s=3} InputSanctions_i * \mathbb{1}_{t=s} + X_{it}\delta + u_{ijt} \quad (2.20)$$

Specification 2.20 is identical to the regression 2.18, except that the average treatment on the treated effect is split into seven year-to-sanction effects. Each α_s identifies each year-to-sanction effects relative to the average outcome in the first year of sanctions. Only the variation within the sanctioned firms is used to identify α_s , however, the non-sanctioned firms can still be used to identify the γ_{jt} and θ_{st} .

$$Y_{it} = \gamma_{jt} + \phi_i + \theta_{st} + \alpha_s * \sum_{s=2012, s \neq 2015}^{s=2018} InputSanctions_i * \mathbb{1}_{t=s} + X_{it}\delta + u_{ijt} \quad (2.21)$$

Specification 2.21 is aimed to test whether the sanctioned and non-sanctioned firms were trending in the same way prior to sanctions. Here, unlike in the previous specification, the full sample is used to identify the coefficients α_s , which show the difference in outcomes of

²⁷Even though officially sanctions began in 2014, because of the two month cool-down period, only a small number of firms are effectively treated in 2014

the sanctioned firms in each year versus in 2015, compared to such difference in outcomes of the non-sanctioned firms.

2.9 Results

2.9.1 Regression results

Table 2.9 shows my baseline results for specifications 2.18 and 2.19. The first thing to note is in columns (1) and (2) we see that the MRPK went down differentially for the SSI-sanctioned SOEs relative to SSI-sanctioned private firms and there is no statistically significant change in MRPK for sanctioned private firms relative to non-sanctioned firms. This tells us two things 1) The negative input shock did not correct the implicit subsidies that politically connected private firms had and we saw in Figure 2.4 2) The negative input shock has led to a response that made SOEs appear as if they had experienced a positive input shock and stronger subsidies.

Does this negative MRPK result come from the input increase (denominator) or the output reduction (numerator)? One could argue that de-risking against Russian sanctioned firms could have led to a simple reduction in sales, especially the sales abroad. Columns (3), (4), (5) and (6) give us the answer: the sales and value added did not decrease, but the inputs increased. First, in column (3) we see an average net increase in capital by 16% after SSI sanctions for sanctioned firms relative to non-sanctioned firms. Capital increased for sanctioned firms on average. Then, in column (4), we see the heterogeneity of this effect. The private sanctioned firms' capital rose, but not significantly, so we can consider this effect as 0 to be conservative. But the sanctioned SOEs have seen their capital increase by 25% more than the sanctioned private firms. All this leads to one conclusion: all sanctioned firms were protected and have seen full shielding of their assets, but the sanctioned SOEs have seen "too much" shielding. The complete shielding of assets would have kept misallocation at the same level as pre-sanctions, but the excessive shielding has, in fact, worsened it. From columns (5) and (6) we see that the value added was not significantly affected by sanctions.

Columns (7) and (8) show the effects of sanctions on revenue. These results provided because revenue is a direct measure reported in the balance sheets, rather than the constructed value added, and therefore may have lower mis-measurement. These results show that the revenues grew on average for sanctioned firms, which again means that the neg-

ative MRPK result in column (2) arises not because the revenues have been dampened by sanctions or de-risking trends.

Using the anecdotal evidence that the funds were taken from the Russian budget, one can conclude that the connected SOEs and private firms were saved at the expense of all other firms and Russian taxpayers. This also has implications for the goals that sanctioning countries hoped to achieve: the sanctions were meant to be targeted and narrow. However, the shielding that took place in response has made the effects being borne by everyone *but* the original targets!

The results in Table 2.9 differ somewhat from early firm-level sanctions results of Ahn & Ludema (2020), who find a negative result on revenue and assets. This is for two reasons. First, they only observe results till 2016, so mainly for only one effective year of sanctions. Second, they measure the combined effect of blocking and SSI sanctions on all assets, including companies owned by Russian oligarchs abroad. Some of the foreign companies had to indeed seize operation and eventually close, which may likely be driving the early negative result.

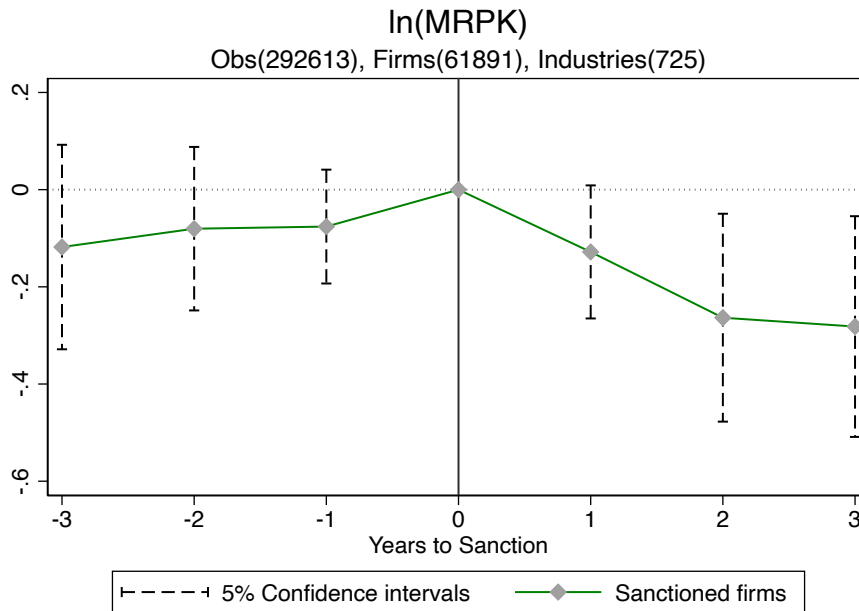
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ln MRPK	Ln MRPK	Ln Book Value of Capital	Ln Book Value of Capital	Ln Value Added	Ln Value Added	Ln Revenue	Ln Revenue
SSI dummy	-0.043 (0.074)	-0.006 (0.084)	0.163** (0.071)	0.124 (0.078)	0.038 (0.050)	0.028 (0.057)	0.160*** (0.058)	0.170*** (0.064)
SDN dummy	-0.041 (0.056)	-0.037 (0.056)	0.104* (0.057)	0.100* (0.057)	0.067 (0.047)	0.066 (0.047)	0.103** (0.049)	0.104** (0.049)
SSI dummy × SOE		-0.233* (0.139)		0.250* (0.136)		0.060 (0.102)		-0.064 (0.139)
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓
Industry-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Size-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Firms	77647	77647	87736	87736	77648	77648	87731	87731
Sanctioned firms	991	991	1084	1084	991	991	1084	1084
Industries	751	751	763	763	751	751	763	763
Observations	347702	347702	417568	417568	347708	347708	417554	417554
R-squared	.888	.888	.995	.995	.996	.996	.997	.997

Notes: All dependent variables are in logs. Firms are classified as SOEs according to Rosstat. MRPK is estimated with the Value added/K method. Industry×Year FE are 4-digit industry by year fixed effects. Size×Year are quartile fixed effects for firms' average pre-treatment capital interacted with year fixed effects. Sanction firms give the count of any sanction firm - SSI or SDN. Standard errors are two-way clustered at the firm and 4-digit industry by year level. *, **, and *** denote 10%, 5%, and 1% statistical significance respectively.

Table 2.9: Average effects of sanctions: key outcome variables

2.9.2 Event studies results

The identification in Table 2.9 is subject to one possible problem. What if the sanctioned firms are on different trends to the other firms and the sanction variables just pick such trends up? In Figure 2.5, I show the event study with $\ln(\text{MRPK})$ as an outcome variable and confirm that the positive effects persist even if I identify them within the group of sanctioned firms, for which the required assumption is weaker: the firms that are sanctioned sooner are not on a different trend compared to the firms that are sanctioned later. In this case, the control group is the average outcome of the sanctioned firm in the year 0, the year it was sanctioned, and the treatment is each year-to-sanction²⁸. I emphasize that the coefficients in Figure 2.5 come from a specification, where I control for the industry-year fixed effects, pre-treatment size-by-year fixed effects and firm fixed effects.



Notes: This figure reports event study graphs for the average effects of the sanctions on sanctioned firms. The effect is identified within sanctioned firms: sanctioned firms are compared to not-yet sanctioned firms. The first year of firm sanction is normalized to take place in year 0. Each dot is the coefficient on the indicator of being observed t years after the sanctions announcement. The same control variables are used as in baseline regression: SDN sanction, firm fixed effects, 4-digit industry-year fixed effects and the size-year fixed effects. Non sanctioned firms are used to identify the 4-digit industry-year fixed effects and the size-year fixed effects. The MRPK dependent variable is in logs. The confidence intervals are at the 95% level.

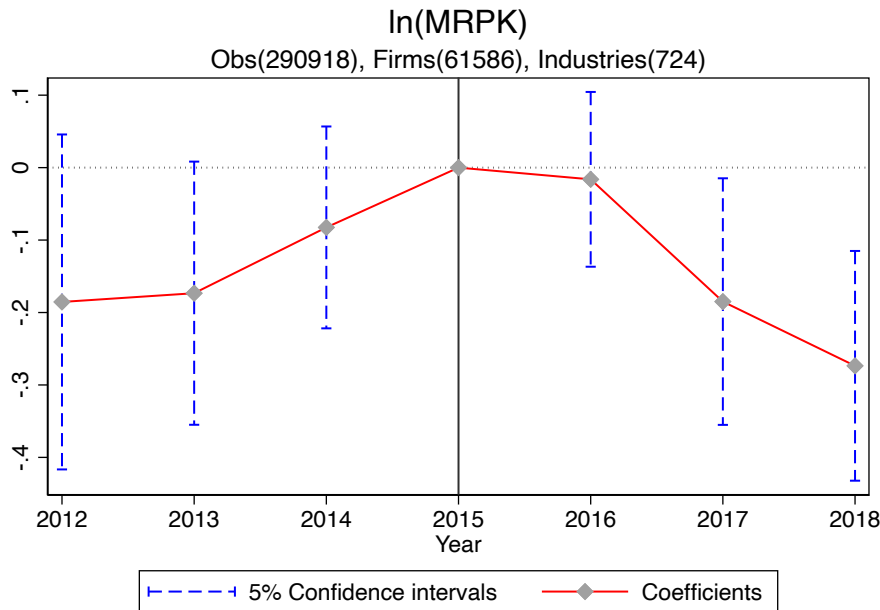
Figure 2.5: SSI event study with not-yet sanctioned firms in the control group.

Furthermore, in Figure 2.6, I show an event study, in which the control group is not just to-be-sanctioned firms, but also the never-sanctioned firms. I also cannot reject that the

²⁸I do not have enough power to identify the effect of the interactions with SOEs, and therefore I do not include the interacted event studies.

group of sanctioned firms was on the same trends as the group of non-sanctioned firms before 2015 in Figure 2.6. If anything the treated firms were on the upward trend before the sanctions, so the regression results I find in Table 2.9 are a lower bound. In this case, the control group is the average time trend in the 4-digit industry and the treatment is the average outcome of the sanctioned firm in each year-to-sanction. The sample used to identify the coefficients in the event study is the full sample of firms, sanctioned or not. I do not find significant effects prior to 2015, which is consistent with the sanctioned and non-sanctioned firms being on the same trend, but I do find a significant drop in MRPK soon after 2015²⁹.

It is important to note that these results are for the SSI (input) sanctions, where I always control for the SDN blocking sanctions in the background.



Notes: This figure reports event study graphs for the average effects of the sanctions on sanctioned firms relative to non-sanctioned firms. Each dot is the coefficient on the interaction between being observed in the year 2012, 2013, 2014, 2015, 2016, 2017 and 2018, and being sanctioned with SSI sanctions. The same control variables are used as in baseline regression: SDN sanction, firm fixed effects, 4-digit industry-year fixed effects and the size-year fixed effects. Effectively, each dot is the deviation of the sanctioned firm log MRPK from the 4-digit-industry-by-year fixed effects. The MRPK dependent variable is in logs. The confidence intervals are at the 95% level.

Figure 2.6: Pre-post 2015 event study with never-sanctioned firms in the control group

²⁹The fact that I find a significant average MRPK result in the event study based on specification 2.6, but not in the regression based on the specification 2.18, could have happened because the $InputSanction_{it}$ dummy in the regression is not just "post 2015". The dummy varied across years since the input sanctions happened in waves. We, therefore, are not comparing two almost identical specifications.

2.9.3 Aggregate effects

I use the estimates from the results in the previous section and a simple formula from Hsieh & Klenow (2009) based on the model in Section 2.3 to calculate the effects on aggregate sector TFP from the change in $TFPR_i$. The use of the formula requires an additional assumption: that the distribution of firm TFP_i and $TFPR_i$ are jointly log-normal³⁰. This assumption is used for convenience, to get a simple expression from the change in TFP for some firms to the aggregate TFP. I assume that the $TFPQ_i$ did not change for sanctioned SOE and private firms due to policy, so the HK formula reduces to:

$$\Delta \log TFP_s = -\frac{1}{2\eta} * VAR(\log TFPR_i + \alpha \Delta \log MRPK_i) \quad (2.22)$$

The value of $\Delta \log MRPK_i$ is taken from Table 2.9 as the coefficient on the interaction term in column (2). The value of $\log TFPR_i$, the log revenue productivity of each firm, and also a summary measure of distortions to these firms, is obtained as a pre-2015 level using the methodology in Section 2.5. Whereby the $\log TFPR_i$ is the residual from regressing $\log TFPR_{it}$ on year and firm fixed effects (and then removing the common 4-digit industry component) for the pre-sanction period years 2012, 2013 and 2014. I conservatively assume that the labour productivity $MRPL_i$ stays the same as the pre-sanction level.

The overall effect on country TFP from sanctions is 0.33% and is calculated with a Cobb-Douglas aggregator of TFP_s from each sector s with powers as value added shares. However, the results for each industry (appendix Figure 2.A2) differ vastly due to the different exposure and underlying level of the treated companies' $TFPR_i$, of 50 industries that experienced changes, 41 experienced negative productivity changes ranging between -3%–0.01%, and 9 minor positive changes all under 1% (with one exception: "Manufacture of television receivers, including video monitors and video projectors" had a 4% productivity increase).

2.10 Conclusion

Using structural and reduced-form evidence, I show that SOEs are a large source of allocative inefficiency, both in terms of how inputs are allocated to SOEs at a given point in

³⁰Bau & Matray (2020) show another way to calculate the aggregate effects as a first-order approximation with the benefit of fewer assumptions. However, their formula is a function of $\frac{\tau_i^K}{1+\tau_i^K}$ and will necessarily give an improvement in TFP_s from a capital increase for the firms that are "too big" when $\tau_i^K < -1$, which is common in my setting

time and in terms of how SOEs respond to negative input shocks. Thus, I address a key challenge in the literature and provide direct evidence of how policies can change allocative efficiency and productivity.

I use a model of heterogeneous firms to quantify how misallocation of capital and labour between state and private firms contributes to aggregate TFP. Then, I use a unique natural experiment - the US sanctions on Russia to causally estimate the combined effect of sanctions and shielding that affected sanctioned firms relative to non-sanctioned and whether the impact of sanctions on SOEs differed from that on the private firms. I use state-of-the-art tools to combine the estimates from this natural experiment with the model and quantify the effects of sanctions on misallocation and, in turn, on the aggregate TFP.

I find that the SOEs are less productive relative to private firms but use relatively more capital and labour. This creates allocative inefficiency within industries and would improve current TFP by 11% if the wedges between state-owned and private firms were removed. My empirical estimation validates the finding that the SOEs are inefficiently large and demonstrates one channel through which the SOEs get so large: SOEs differentially respond to negative input shocks by getting subsidies that over-shoot the negative shocks. The sanctions, combined with shielding, have led an SOE to gain 25% more capital relative to a private sanctioned firm and 35% more capital relative to a non-sanctioned firm on average. These results are estimated for the type of sanctions that specifically negatively shock the capital inputs of the target firms. I quantify that this joint sanctions and shielding effect reduced the aggregate TFP by 0.33%, which varied between 0% and 3% in different sectors.

This paper has important policy implications. First, as this text is being written, more US sanctions are being promised by the Biden administration. Due to the evidence of excessive shielding that I find, the sanctions failed to be targeted and narrow. Instead, they have provided a trigger for shielding some firms at the expense of the taxpayers and other non-politically connected firms. Sanctions spilt over to the rest of the economy, and allocative efficiency worsened in Russia. The estimate of 0.33% lower TFP (and therefore, 0.33% lower GDP assuming total resources stayed at the pre-sanction level) is likely an underestimate in terms of GDP, as total resources have likely shrunk over this period, as well.

Second, it shines a light on state ownership as one of the strong drivers of misallocation. Misallocation due to ownership status can be improved by allocating fewer resources to SOEs by limiting the soft budget constraint. This can be achieved by monitoring how

the subsidies and tax breaks are granted and specifying the rulebooks in advance on what subsidies and capital transfers the SOEs can receive under what circumstances and what public goals these favours fulfil.

Future research will study further the channels of how misallocation across ownership lines is amplified due to the political connections of private firms to the SOEs and by which means the incentive issues of the SOEs trickle down to the rest of the economy.

Public support for state ownership has grown according to the EBRD Enterprise surveys, and just under 50% of people favour an increase in state-ownership (EBRD 2020). State-owned enterprises have played important functions in emerging economies, such as China, Russia and other post-Communist countries: they stabilized employment and facilitated a more equal provision of public services and financial inclusion. However, these functions have come at the cost of ineffective management, lack of transparency and subsidies that created inefficient allocation of resources in the economy. This paper quantified this cost to be sizeable and found that TFP (and therefore, output) is lower by at least 11%.

2.A Appendix A. Heterogeneous firm model

One-industry model.

This is the standard model that almost every "indirect approach" paper on misallocation is using. It shows that a **dispersion** of wedges ("taxes" or "subsidies") lead to the dispersion of MRPK and MRPL (marginal revenue products of labour and capital) and thus allocative inefficiency, and as a result, lower aggregate TFP. (Aggregate output in this model may also depend on the average **level** of the wedges (if they are driven by, for example, corruption), but the level is harder to identify without stronger assumptions. For now, I focus on the allocative inefficiency aspect, and thus the dispersion of wedges.)

Firms.

$$Q_i = A_i K_i^\alpha L_i^{1-\alpha} \tag{2.23}$$

For simplicity of exposition I assume α is the same across firms. In empirical analysis, I will relax this assumption by industry. Each firm's output is aggregated to a CES aggregate:

$$Q = \left(\sum_{i=1}^N Q_i^{1-\eta} \right)^{\frac{1}{1-\eta}} \tag{2.24}$$

The aggregating firm demands outputs of individual firms and maximizes profits:

$$\begin{aligned}
& \max_{Q_i} P \left(\sum_{i=1}^N Q_i^{1-\eta} \right)^{\frac{1}{1-\eta}} - \sum_{i=1}^N P_i Q_i \\
FOC : & \frac{1}{1-\eta} P \left(\sum_{i=1}^N Q_i^{1-\eta} \right)^{\frac{1}{1-\eta}-1} (1-\eta) Q_i^{-\eta} - P_i = 0 \\
& P \left(\sum_{i=1}^N Q_i^{1-\eta} \right)^{\frac{\eta}{1-\eta}} = P_i Q_i^\eta \\
& PQ^\eta Q_i^{*1-\eta} = P_i Q_i^* \tag{2.25}
\end{aligned}$$

The above equation (implicitly) shows how much Q_i is demanded for each firm given P_i , and it is expressed as revenue each firm gets in equilibrium. Each firm i maximizes profits $\pi_i = P_i Q_i - (1 + \tau_i^L) w L_i - (1 + \tau_i^K) r K_i$.

Or, substituting the implicit expression of quantities demanded for the revenue:

$$\max_{L_i, K_i} \pi_i = PQ^\eta Q_i^{*1-\eta} - (1 + \tau_i^L) w L_i - (1 + \tau_i^K) r K_i$$

s.t.

$$Q_i = A_i K_i^\alpha L_i^{1-\alpha}$$

I assume w and r are the **common** and **exogenous** costs of labor and capital. Whereas τ_i^L and τ_i^K are firm-specific distortions to the cost of labor and capital.

$$\{L_i\} : (1-\alpha)(1-\eta) \frac{PQ^\eta (A_i K_i^\alpha L_i^{1-\alpha})^{1-\eta}}{L_i} = (1 + \tau_i^L) w \tag{2.26}$$

The optimal labor allocation will satisfy this equation:

$$\{L_i\} : (1-\alpha)(1-\eta) \frac{P_i Q_i}{L_i} = (1 + \tau_i^L) w \equiv MRPL_i \tag{2.27}$$

$$\{L_i\} : L_i = (1-\alpha)(1-\eta) \frac{P_i Q_i}{MRPL_i} \tag{2.28}$$

Similarly, this equation will be satisfied by the optimal capital allocation:

$$\{K_i\} : \alpha(1 - \eta) \frac{P_i Q_i}{K_i} = (1 + \tau_i^K) r \equiv MRPK_i \quad (2.29)$$

$$\{K_i\} : K_i = \alpha(1 - \eta) \frac{P_i Q_i}{MRPK_i} \quad (2.30)$$

It is useful to add the definition of TFPR_i, which is often used in the literature and is a summary measure of distortions.

$$TFPR_i \equiv \frac{P_i Q_i}{K_i^\alpha L_i^{1-\alpha}} = \left(\frac{MRPK_i}{\alpha} \right)^\alpha \left(\frac{MRPL_i}{1 - \alpha} \right)^{1-\alpha} \frac{1}{(1 - \eta)} \quad (2.31)$$

Re-arranging optimal output in terms of parameters that constitute the costs of firm *i*, we get:

$$P_i Q_i = PQ^\eta (A_i K_i^\alpha L_i^{1-\alpha})^{1-\eta} = PQ^\eta \left(A_i \left[\frac{(1 - \alpha)(1 - \eta) P_i Q_i}{(1 + \tau_i^L) w} \right]^{1-\alpha} \left[\frac{\alpha(1 - \eta) P_i Q_i}{(1 + \tau_i^K) r} \right]^\alpha \right)^{1-\eta} \quad (2.32)$$

$$P_i Q_i = PQ^\eta (P_i Q_i)^{1-\eta} (1 - \eta)^{1-\eta} \left(A_i \left[\frac{(1 - \alpha)}{(1 + \tau_i^L) w} \right]^{1-\alpha} \left[\frac{\alpha}{(1 + \tau_i^K) r} \right]^\alpha \right)^{1-\eta} \quad (2.33)$$

$$P_i Q_i = P^{\frac{1}{\eta}} Q \left((1 - \eta) A_i \left[\frac{(1 - \alpha)}{(1 + \tau_i^L) w} \right]^{1-\alpha} \left[\frac{\alpha}{(1 + \tau_i^K) r} \right]^\alpha \right)^{\frac{1-\eta}{\eta}} \quad (2.34)$$

$$P_i Q_i \propto \left(\frac{A_i}{(1 + \tau_i^L)^{1-\alpha} (1 + \tau_i^K)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (2.35)$$

Combine 2.27 , 2.29 and 2.35 to get that more labor and capital in the absence of τ_i^K and τ_i^L will go to the more productive firm - firm with higher A_i

$$L_i \propto \frac{1}{1 + \tau_i^L} \left(\frac{A_i}{(1 + \tau_i^L)^{1-\alpha} (1 + \tau_i^K)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (2.36)$$

$$K_i \propto \frac{1}{1 + \tau_i^K} \left(\frac{A_i}{(1 + \tau_i^L)^{1-\alpha} (1 + \tau_i^K)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (2.37)$$

Equivalently,

$$1 + \tau_i^L \propto \frac{P_i Q_i}{w L_i} \quad (2.38)$$

$$1 + \tau_i^K \propto \frac{P_i Q_i}{K_i} \quad (2.39)$$

Expressing 2.35 in terms of how we can measure each of the distortions:

$$P_i Q_i \propto \left(\frac{A_i}{\left(\frac{P_i Q_i}{L_i}\right)^{1-\alpha} \left(\frac{P_i Q_i}{K_i}\right)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (4) \quad (2.40)$$

Revenues of firms will be negatively correlated to the geometric average of the distortions (themselves proportional to labour and capital productivities, implying higher labour and capital productivity - labour and capital input is too small) and positively correlated with their productivity A_i . Again, remember that this assumes: α, w, r, η are identical across firms. Any deviation in these will manifest itself in deviations in τ_K , and/or τ_L .

It is also useful to derive a model-based firm productivity:

$$PQ^\eta (A_i K_i^\alpha L_i^{1-\alpha})^{1-\eta} = P_i Q_i \quad (2.41)$$

$$A_i = (PQ^\eta)^{\frac{-1}{1-\eta}} \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}} \quad (2.42)$$

$$A_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}} \quad (2.43)$$

$$\kappa = (PQ^\eta)^{-\frac{1}{1-\eta}} \quad (2.44)$$

Aggregation

$$P_i Q_i = P^{\frac{1}{\eta}} Q \left((1-\eta) A_i \left[\frac{(1-\alpha)}{(1+\tau_i^L)w} \right]^{1-\alpha} \left[\frac{\alpha}{(1+\tau_i^K)r} \right]^\alpha \right)^{\frac{1-\eta}{\eta}} \quad (2.45)$$

$$PQ = \sum P_i Q_i \quad (2.46)$$

Use the exact expressions for optimal L_i and K_i

$$L_i = \frac{(1 - \alpha)(1 - \eta)P^{\frac{1}{\eta}}Q \left((1 - \eta)A_i \left[\frac{(1 - \alpha)}{(1 + \tau_i^L)w} \right]^{1 - \alpha} \left[\frac{\alpha}{(1 + \tau_i^K)r} \right]^\alpha \right)^{\frac{1 - \eta}}}{(1 + \tau_i^L)w} \quad (2.47)$$

$$K_i = \frac{\alpha(1 - \eta)P^{\frac{1}{\eta}}Q \left((1 - \eta)A_i \left[\frac{(1 - \alpha)}{(1 + \tau_i^L)w} \right]^{1 - \alpha} \left[\frac{\alpha}{(1 + \tau_i^K)r} \right]^\alpha \right)^{\frac{1 - \eta}}{(1 + \tau_i^K)r} \quad (2.48)$$

$$L = \sum L_i = (1 - \alpha)(1 - \eta) \sum \frac{1}{(1 + \tau_i^L)w} P_i Q_i = \quad (2.49)$$

$$L = (1 - \alpha)(1 - \eta)PQ \sum \frac{1}{(1 + \tau_i^L)w} \frac{P_i Q_i}{PQ} \quad (2.50)$$

$$L = (1 - \alpha)(1 - \eta)PQ \frac{1}{MRPL} \quad (2.51)$$

Equivalently, the expression from the market clearing condition for aggregate capital is:

$$K = \alpha(1 - \eta)PQ \frac{1}{MRPK} \quad (2.52)$$

Let's define the aggregate TFP the following way:

$$TFP \equiv \frac{Q}{K^\alpha L^{1 - \alpha}} \quad (2.53)$$

$$TFP = \frac{Q}{\left(\alpha(1 - \eta)PQ \frac{1}{MRPK} \right)^\alpha \left((1 - \alpha)(1 - \eta)PQ \frac{1}{MRPL} \right)^{1 - \alpha}} \quad (2.54)$$

$$TFP = \frac{\overline{TFPR}}{P} = \frac{1}{P(1-\eta)} \left(\frac{\overline{MRPK}}{\alpha} \right)^\alpha \left(\frac{\overline{MRPL}}{1-\alpha} \right)^{1-\alpha} \quad (2.55)$$

To get P, aggregate the expression 2.55

$$PQ = \sum_i P_i^{\frac{1}{\eta}} Q \left((1-\eta) A_i \left[\frac{(1-\alpha)}{(1+\tau_i^L)w} \right]^{1-\alpha} \left[\frac{\alpha}{(1+\tau_i^K)r} \right]^\alpha \right)^{\frac{1-\eta}{\eta}} = \quad (2.56)$$

$$PQ = P_i^{\frac{1}{\eta}} Q \left((1-\alpha)^{1-\alpha} \alpha^\alpha \right)^{\frac{1-\eta}{\eta}} \sum_i \left(\frac{(1-\eta) A_i}{((1+\tau_i^L)w)^{1-\alpha} ((1+\tau_i^K)r)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (2.57)$$

$$P^{\frac{\eta-1}{\eta}} = \left((1-\alpha)^{1-\alpha} \alpha^\alpha \right)^{\frac{1-\eta}{\eta}} \sum_i \left(\frac{A_i(1-\eta)}{(MRPL_i)^{1-\alpha} (1+\tau_i^K)^\alpha} \right)^{\frac{1-\eta}{\eta}} \quad (2.58)$$

$$P = \frac{1}{(1-\eta)} \left((1-\alpha)^{1-\alpha} \alpha^\alpha \right)^{-1} \left(\sum_i \left(\frac{A_i}{(MRPL_i)^{1-\alpha} (MRPK_i)^\alpha} \right)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{\eta-1}} \quad (2.59)$$

Plug 2.59 into 2.55.

$$TFP = \frac{1/(1-\eta) \left(\frac{\overline{MRPK}}{\alpha} \right)^\alpha \left(\frac{\overline{MRPL}}{1-\alpha} \right)^{1-\alpha}}{1/(1-\eta) \left((1-\alpha)^{1-\alpha} \alpha^\alpha \right)^{-1} \left(\sum_i \left(\frac{A_i}{(MRPL_i)^{1-\alpha} (MRPK_i)^\alpha} \right)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{\eta-1}}} \quad (2.60)$$

Aggregate TFP if you have decentralized allocation with wedges.

$$TFP = \left(\sum_i \left(A_i \left(\frac{\overline{MRPL}}{MRPL_i} \right)^{1-\alpha} \left(\frac{\overline{MRPK}}{MRPK_i} \right)^\alpha \right)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.61)$$

Aggregate TFP if you have efficient allocation without wedges.

$$TFP^e = \left(\sum_i (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.62)$$

Distance of aggregate TFP to the efficient (frontier)

$$\frac{TFP^e}{TFP} - 1 \quad (2.63)$$

Equalizing TFPR within groups I also consider a separate counterfactual in which I look at two groups in each sector: state-owned and private, and I redistribute existing labour and existing capital of each group across firms within each group to equalize their MRPL's and MRPK's (i.e. all firms within each group have the same average wedge).

Thus, I get two expressions of group MRPL and MRPK:

1)

$$\frac{(L_{priv})^\eta \left(\frac{L_{priv}}{K_{priv}}\right)^{\alpha(1-\eta)}}{(1-\alpha)(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} = \frac{1}{MRPL_{priv}} \quad (2.64)$$

2)

$$\frac{(K_{priv})^\eta \left[\frac{K_{priv}}{L_{priv}}\right]^{(1-\alpha)(1-\eta)}}{\alpha(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} = \frac{1}{MRPK_{priv}} \quad (2.65)$$

3) I combine (1) and (2) to get an expression for group TFPR for private and state-owned group (the expression for state-owned TFPR is similar):

$$1/TFPR_{priv} = \left[\frac{(K_{priv})^\eta \left[\frac{K_{priv}}{L_{priv}}\right]^{(1-\alpha)(1-\eta)}}{\alpha(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} \right]^\alpha \left[\frac{(L_{priv})^\eta \left(\frac{L_{priv}}{K_{priv}}\right)^{\alpha(1-\eta)}}{(1-\alpha)(1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} \right]^{1-\alpha} = \quad (2.66)$$

$$= \frac{(K_{priv})^{\alpha\eta} (L_{priv})^{(1-\alpha)\eta}}{(1-\alpha)^{1-\alpha} \alpha^\alpha (1-\eta)PQ^\eta \left(\sum (A_i)^{\frac{1-\eta}{\eta}}\right)^\eta} \quad (2.67)$$

$$TFPR_{priv} = \frac{\left(\sum \left(\frac{A_i}{\kappa}\right)^{\frac{1-\eta}{\eta}}\right)^\eta}{(K_{priv})^{\alpha\eta} (L_{priv})^{(1-\alpha)\eta}} \quad (2.68)$$

$$\kappa = (PQ^\eta)^{-\frac{1}{1-\eta}} \quad (2.69)$$

where kappa cancels out in the aggregate TFP expression.

4) Note that this means that the Industry-level output, and thus industry-level TFPR (and industry-level MRPL's and MRPK's) will increase because adjustments towards a more optimal allocation are made.

Aggregate TFP after efficiently allocating capital and labour across firms within ownership-industry groups.

$$TFP = \left(\sum_{o \in \{priv, so\}} \left(\frac{\overline{MRPL}}{\overline{MRPL}_o} \right)^{1-\alpha} \left(\frac{\overline{MRPK}}{\overline{MRPK}_o} \right)^\alpha \sum_{i \in o} (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.70)$$

or, equivalently:

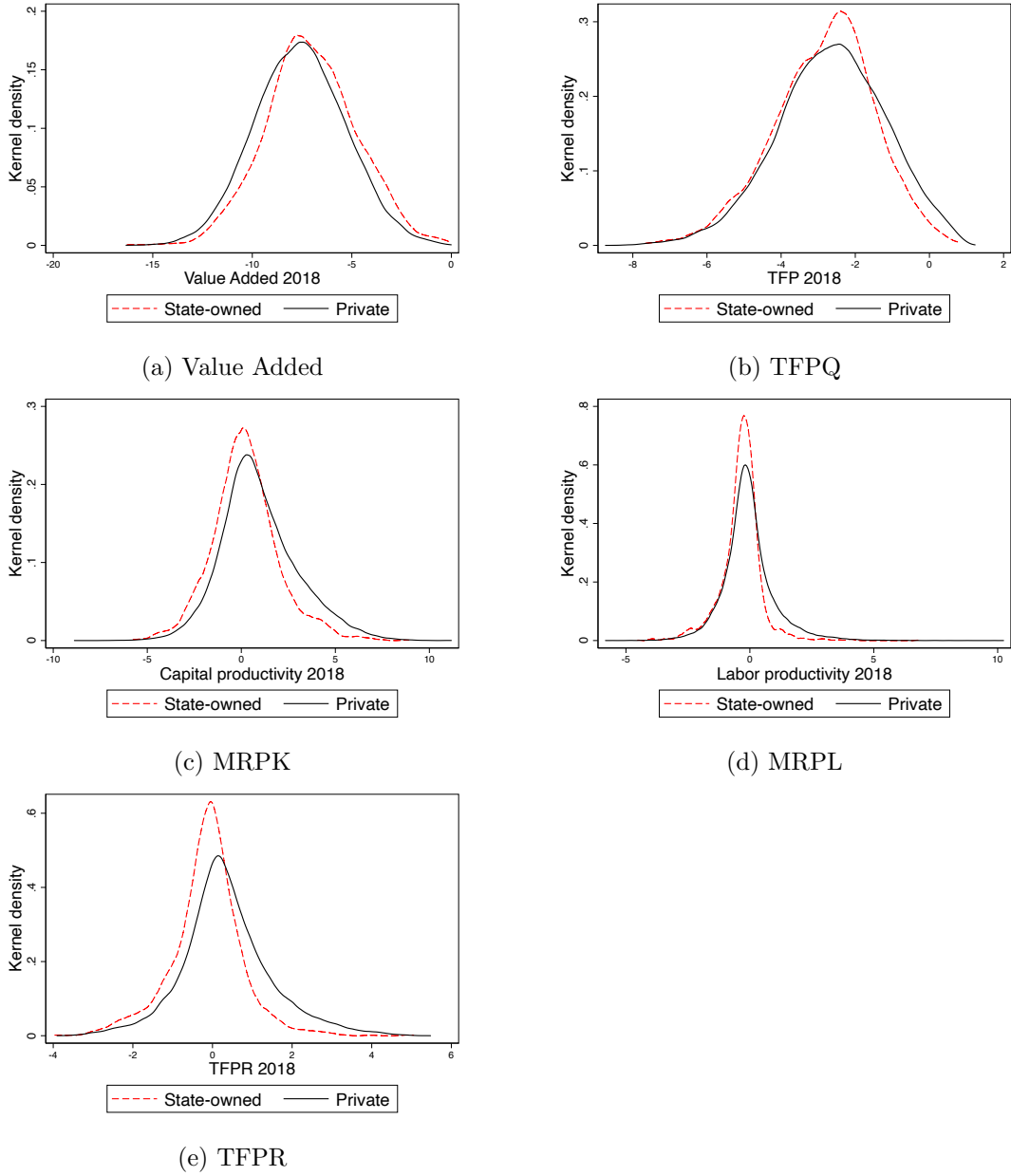
$$TFP = \left(\sum_{o \in \{priv, so\}} \left(\frac{\overline{TFPR}}{\overline{TFPR}_o} \right) \sum_{i \in o} (A_i)^{\frac{1-\eta}{\eta}} \right)^{\frac{\eta}{1-\eta}} \quad (2.71)$$

2.B Appendix B. Additional tables and figures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ln MRPK	Ln MRPK	Ln Book Value of Capital	Ln Book Value of Capital	Ln Value Added	Ln Value Added	Ln Revenue	Ln Revenue
SSI dummy	-0.041 (0.073)	-0.016 (0.084)	0.159** (0.070)	0.136* (0.077)	0.038 (0.050)	0.023 (0.057)	0.160*** (0.052)	0.171*** (0.064)
SDN dummy	-0.037 (0.055)	-0.035 (0.056)	0.097* (0.057)	0.095* (0.057)	0.069 (0.047)	0.067 (0.047)	0.101* (0.048)	0.103** (0.049)
SSI dummy × SOE		-0.153 (0.142)		0.147 (0.136)		0.094 (0.104)		-0.067 (0.140)
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓
Industry-year FE	✓	✓	✓	✓	✓	✓	✓	✓
SOE-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Firms	77647	77647	87736	87736	77648	77648	87731	87731
Sanctioned firms	991	991	1084	1084	991	991	1084	1084
Industries	751	751	763	763	751	751	763	763
Observations	347702	347702	417568	417568	347708	347708	417554	417554
R-squared	.888	.888	.995	.995	.996	.996	.997	.997

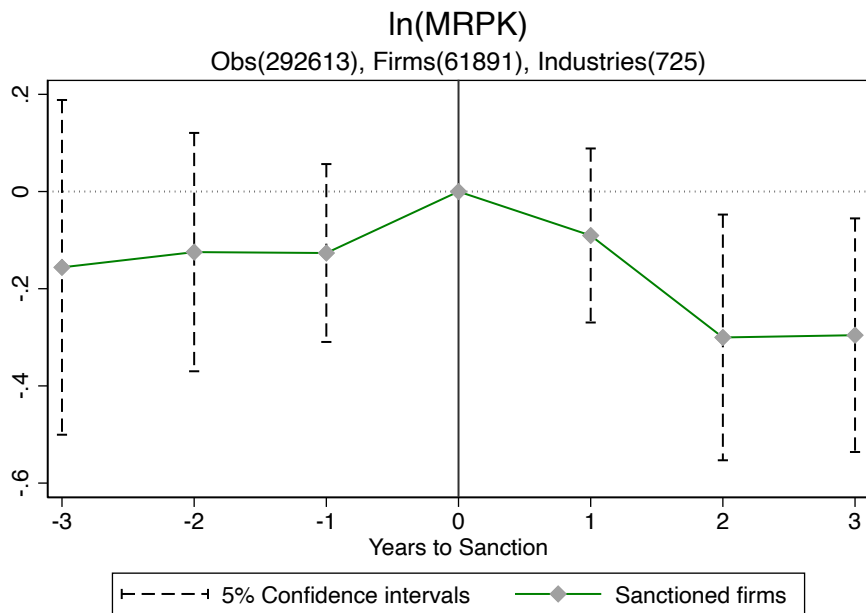
Notes: All dependent variables are in logs. Firms are classified as SOEs according to Rosstat. MRPK is estimated with the Value added/K method. Industry-year FE are 4-digit industry-by-year fixed effects. SOE-by-year FE are the SOE dummy interacted with year dummies. Sanction firms give the count of any sanction firm - SSI or SDN. Standard errors are two-way clustered at the firm and 4-digit industry by year level. *, **, and *** denote 10%, 5%, and 1% statistical significance respectively.

Table 2.A1: Average effects of sanctions triple difference



Notes: The plots show the kernel density of natural logs of value added, TFPQ, MRPK, MRPL and TFPR. The red dotted lines are the kernel densities for the SOEs sample. The black lines are the kernel densities for the sample of private firms. Labor productivity (or $MRPL_i$) refers to value added per unit of wage bill and capital productivity (or $MRPK_i$) refers to value added per unit of capital, both of which are proportional to the marginal products of each factor in my framework. Raw TFPQ is calculated using the expression $TFPQ_i = \kappa \frac{(P_i Q_i)^{\frac{1}{1-\eta}}}{K_i^\alpha L_i^{1-\alpha}}$. Each measure is directly calculated from the raw data in 2018.

Figure 2.A1: Allocations of SOEs versus the private sector



Notes: This figure reports event study graphs for the average effects of the sanctions on sanctioned firms. The effect is identified within sanctioned firms: sanctioned firms are compared to not-yet sanctioned firms. The sample used in this regression is constant and includes firms that are observed three years prior and three years after the sanctions announcement. The first year of firm sanction is normalized to take place in year 0. Each dot is the coefficient on the indicator of being observed t years after the sanctions announcement. The same control variables are used as in baseline regression: SDN sanction, firm fixed effects, 4-digit industry-year fixed effects and the size-year fixed effects. Non sanctioned firms are used to identify the 4-digit industry-year fixed effects and the size-year fixed effects. The MRPK dependent variable is in logs. The confidence intervals are at the 95% level.

Figure 2.A2: Constant sample SSI event study

2.C Appendix C. Data appendix

I construct a dataset of sanctioned firms.

- 1) firm SDN sanctions+subsidiaries (variable "sdn")
- 2) firm SSI sanctions +subsidiaries (variable "ssi")
- 3) person SDN sanctions + owned firms (variable "ind")
- 4) EU sanctions, which mimic the US sanctions, be it SDN or SSI.

In the regressions, I then take the unions of the variables (1), (3) and the "blocked" firms by the EU (4) to make a combined SDN variable. There are only 9 firms that are sanctioned by the EU but not the US (some of them are subsidiaries). I have coded them as SDN is the EU treatment was to stop all transactions, and SSI if these were input sanctions.

I create separate treatment year variables for the SSI and SDN categories. However, even within categories, some firms have several treatment years, because they are sanctioned both by association with other sanctioned firms and directly. Priority of the first treatment year assignment for companies that fall into several sanction categories is the following:

- (1) the year of mother company's treatment (if the company is majority-owned)
- (2) the year of the company is explicitly listed on the Department of Treasury, if (1) does not exist.
- (3) If the company is minority-owned by multiple sanctioned firms (where the total shares from different companies add up to more than 50%) with different sanctioned years AND (1) and (2) years do not exist, the assigned year is earliest among potential SDN years, "individual SDN" sanction years for the SDN variable, and the earliest among the SSI owner company years, for the SSI variable.

I used the sanction announcement date to assign the year according to the April 30th split: if you get sanctioned after April 30th, your treatment year is the year after.

These sanctions do not include sanctions that took place before 2014 and sanctions that are not to do with the Ukraine conflict. I also exclude firms that are in Crimea (around 40 firms), since they are embargoed based on their location in Crimea only, and not based on the connections to the current government.

Sector	% change in TFPs	Sector	% change in TFPs
Manufacture of computers and peripheral equipment	-3.36	Production of drugs and materials used for medical purposes	-0.15
Transportation of gas and products of its processing through pipelines	-3.23	Wholesale trade of solid, liquid and gaseous fuels and similar products	-0.14
Electricity production by thermal power plants, including activities to ensure the operability of power plants	-2.34	Provision of drilling services related to oil, gas and gas condensate production	-0.13
Activities in the field of communication based on wired technologies	-1.81	Activities in the field of architecture	-0.13
Production of petroleum products	-1.28	Mechanical processing of metal products	-0.11
Market research	-1.25	Other scientific research and development in the field of natural and technical sciences	-0.10
Communication equipment manufacturing	-0.96	Investments in securities	-0.10
Supporting activities related to air and space transport	-0.93	Electrical work	-0.09
Transportation of crude oil by sea-going tankers of foreign voyages	-0.92	Activities of health resort organizations	-0.06
Extraction of crude oil	-0.46	Manufacture of electric motors, generators and transformers	-0.05
Manufacture of parts for electronic tubes, tubes and other electronic components, not elsewhere classified	-0.45	Printing newspapers	-0.04
Retail sale of motor fuel in specialized stores	-0.44	Research and development in the field of natural and technical sciences	-0.04
Production of parts for railway locomotives, tram and other motor cars and rolling stock; production of track equipment and devices for traffic control of railway, tram and other tracks, mechanical and electromechanical equipment for traffic control	-0.36	Cultivation of cereals	-0.02
Construction of railways and metro	-0.35	Activities for the provision of cash loans secured by real estate	-0.01
Distribution of gaseous fuels through gas distribution networks	-0.31	Lease and management of own or leased real estate	-0.01
Manufacture of other electrical equipment.	-0.31	Topographic and geodetic activities	-0.01
Technical inspection of vehicles	-0.23	Holding company management activities	0.00
Manufacture of parts of devices and instruments for navigation, control, measurement, control, testing and other purposes	-0.22	Production of building metal structures, products and their parts	0.00
Tool production	-0.20	Breeding of dairy cattle, production of raw milk	0.00
Storage and warehousing of grain	-0.19	Real estate management on a fee or contract basis	0.00
Activities related to the use of computers and information technology, other	-0.19	Computer software development	0.01
Repair and maintenance of aircraft, including spacecraft	-0.17	Wholesale and retail trade; repair of motor vehicles and motorcycles	0.01
Electricity transmission and technological connection to distribution grids	-0.17	Manufacture of bricks, tiles and other building products from baked clay	0.02
Other types of printing activities	-0.16	Activities in the field of communication based on wired technologies	0.07
Other auxiliary activities related to transportation	-0.16	Manufacture of television receivers, including video monitors and video projectors	4.10

Notes: The table shows aggregate effects on output (TFP) in each industry with sanctioned firms. The effect comes from the combined effect of sanctions and government response on misallocation.

Table 2.A2: TFPs Results (aggregate effects of sanctions by industry)

Chapter 3

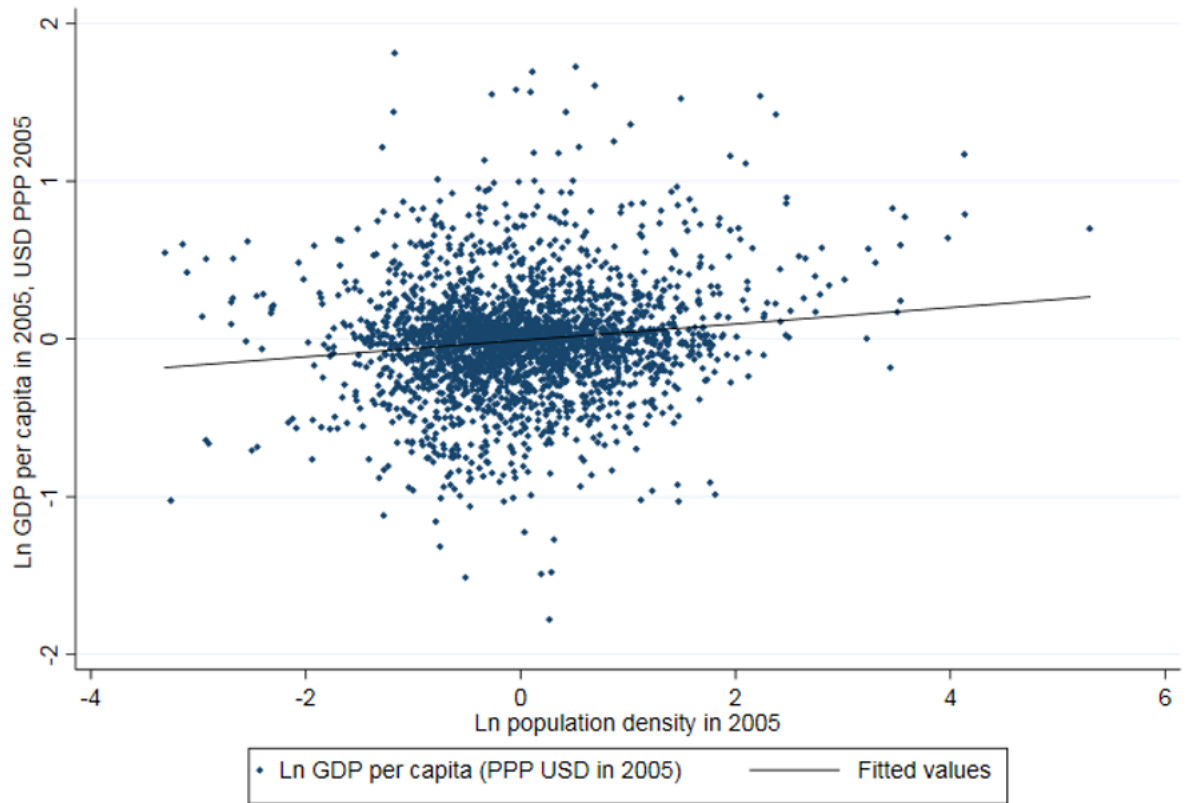
People in space: a toolbox for an economic geographer

3.1 Introduction: why do we care about population density?

Since at least Marshall & Marshall (1920) economists and geographers emphasized that there are important reasons for and consequences from how people are distributed in space. Only a small fraction of world land is built up. Even in very uniform locations in terms of natural features, say in Midlands in the UK, people are not spread uniformly, and people choose to reside together in small and big villages, towns and cities. Moreover, the bigger cities tend to be more successful. Geographical features, such as good water transport links that initially were favourable for people to settle in the historic European capitals, like London, Paris, Berlin, would not explain why they are still some of the biggest economic centres, now in the age when water as a constraint no longer exists in Europe.

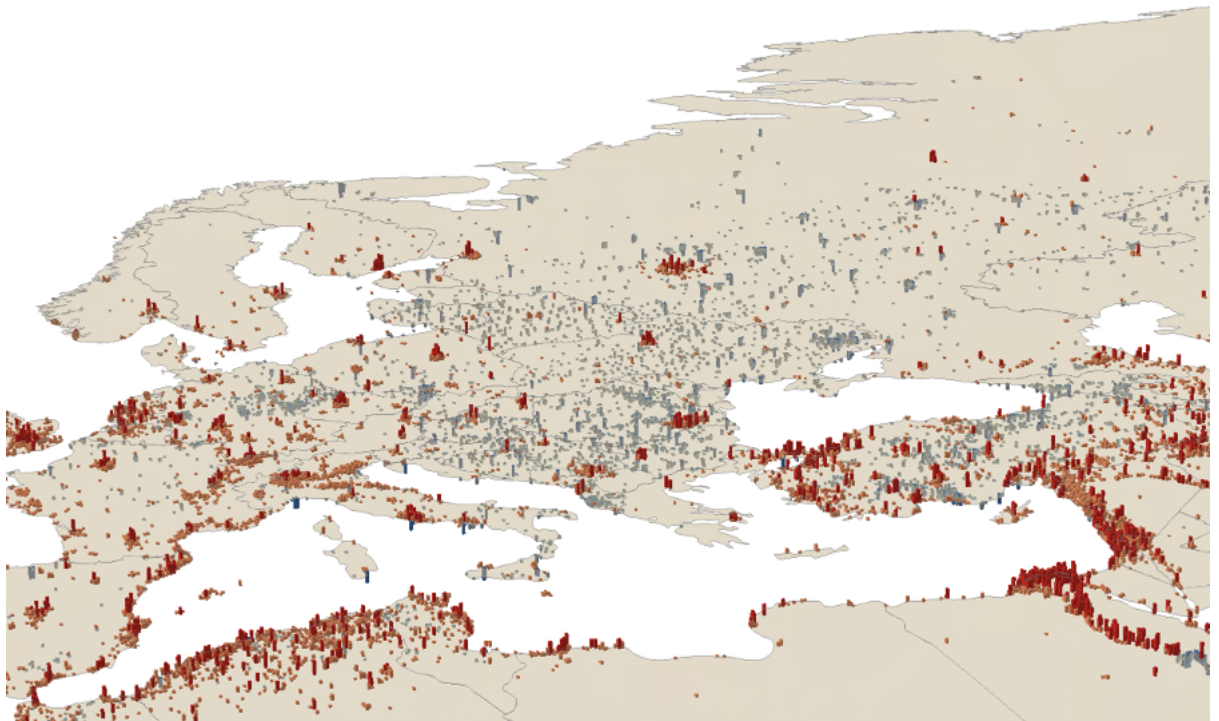
Looking beyond Western Europe, population density premium in terms of GDP exists, as shown in Figure 3.1 for equal and non-arbitrary units of space. Figure 3.1 shows a positive correlation between GDP per capita and density at one-degree-by-one-degree squares the world¹. The relationship is statistically significant and survives controlling for transport connections, latitude, longitude and country fixed effects; 1% increase in population density is associated with 0.05% increase in GDP per capita. A higher density is undoubtedly the consequence of the natural advantages of a location, which at least partly gives rise to this correlation. However, a higher concentration of people in a location can also be a cause for stronger economic performance and concentration of physical and human capital investments, which reinforces this correlation.

¹One decimal degree is around 111 km on average. Only cells with at least 10 people are included in the analysis.



Notes: The underlying source of this scatterplot is the G-Econ dataset and authors' calculations. Each dot represents a 1 degree by 1-degree cell with a minimum population of 10 people and minimum GDP per capita of US\$ 7.38 at PPP. Blue dots represent the raw scatterplot and the black line is the line of best fit.

Figure 3.1: Correlation of $\ln(\text{GDP per capita})$ and $\ln(\text{population density})$ of 1 decimal degree by 1 decimal degree units using G-Econ data Nordhaus et al. (2018) in the entire world in the year 2005



Notes: The underlying raw data is from the GHSL dataset made by the European Commission, Columbia University. The image is based on 100km² grid squares for the period 1990-2015. Bar heights convey population changes, with red bars denoting population increases and grey bars indicating decreases. Beige areas without bars are places with population changes of less than 200 people.

Figure 3.2: Population changes in Europe in 1990-2014 people in 10km².

According to Krugman (1991) and Redding & Venables (2004), among many others, not merely the more populated places are "special", but the places that are populated and in easy access to other large places. Such favourable locations both enjoy higher levels of incomes and higher future growth in population. The measurement of the concentration of people and of access to other people will be one theme explored in this paper.

In the recent decades, the largest, and most connected places have gained people as transport and trade costs dropped with the enlargement of the EU, as we see in Figure 3.2 (red means growth and blue means loss in population). We also see that smaller places in Eastern Europe, except the larger capitals, have lost people².

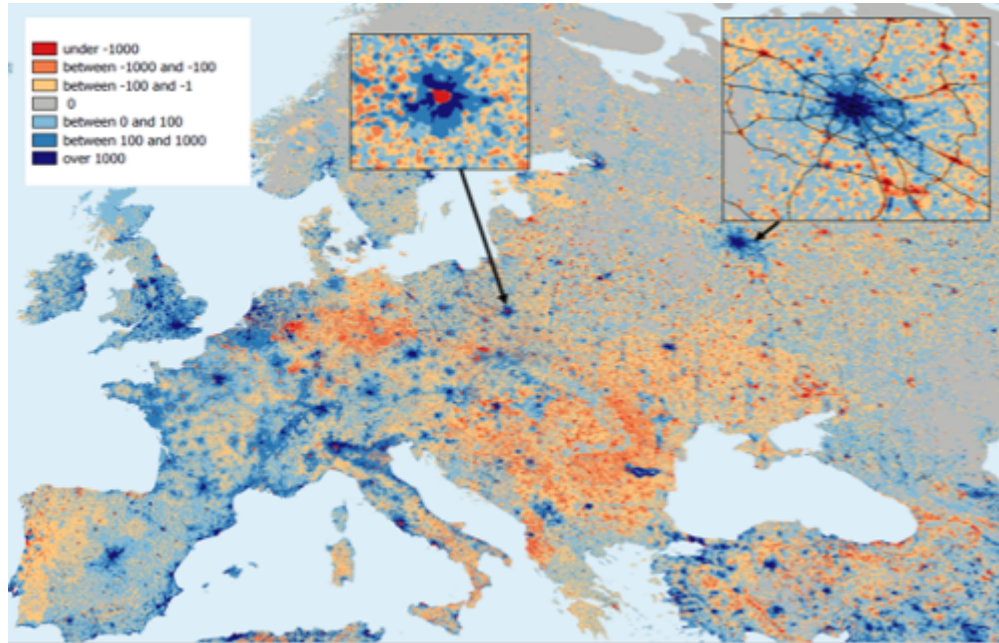
These among other examples, show that where people are placed in space matters for how these places will fair in the long run. However, not all density is necessarily translated to further population and economic growth³. What are the conditions for this to happen? Can the certain distribution of people *inside* cities or across regions be associated with higher or lower growth? These questions demand a more rigorous study of people (but also workers and firms) in space.

In this paper, I will look at land as a continuum, or a set of discrete but comparable and connected units and I will argue this will help me measure the distribution of people in space better. Studying space in this way means that the population and population density of economic units are not mismeasured due to administrative boundaries shaped by political or historic considerations. Moreover, in theoretical models of geography, space is often modelled continuously (with a notable exception of Ahlfeldt et al. (2015) who make the opposite choice to fit the existing form of data better). It is empirically impossible to have a continuous measure of people (or workers and firms) in space, although if space is divided into equal "squares", one can get closer to this approximation and in principle find it easier to test or calibrate the models. Finally, measuring people's locations in a set of discrete but equally-sized units allows me to create measures of spatial correlation that are also comparable across cities, regions, countries and continents. This will be important when I try to establish what measures are associated most with economic outcomes.

In this paper, I will also provide methodological insights on how to measure the way

²We also see large gains in population South of Mediterranean, which are likely due to high fertility. In the later sections of the paper, I will demonstrate that not all density leads to gains in per capita incomes.

³Ongoing work by Henderson et al. (2020) explores the relationship between GDP per capita and a measure of "quality-adjusted" population density at country-level rather than grid-level, adjusting for a wider set of land quality characteristics and finds that at the country-level, the relationship between $\ln(\text{GDP per capita})$ and $\ln(\text{population density})$ is reversed. With this in mind, it is important to remember that a large part of the positive GDP per capita - density correlation can be explained by the land quality fundamentals.



Notes: The underlying raw data is the GHSL by European Commission, Columbia University and author's calculations. Localised population density (RPA) is a measure of the number of people living within 5 km of a person, discounted by distance. The unit of change is the number of people in a 5 km radius. Each point on the map shows how the number of people has changed within a 5km radius of it. Red is growth, blue is loss in local population density. Units in the legend are the "people count within 5 km radius".

Figure 3.3: Change in local population density 2000-2014.

people are in space. I will first discuss the structure and tradeoffs of the new types of gridded population data that can be used in spatial analysis. I will then demonstrate some general methods of how to process these data, combine them with other data and create insightful measures for further analyses, such as the creation of economically meaningful city extents. In each section, I will try to highlight what the new measures reveal about people's placements in space, and what research questions the results trigger.

I will split the analysis into two parts: I will first focus on country-wide or region-wide population distributions, and in the second part of the paper, I will look at within-city distributions. I will separate my analysis this way because there remains a distinct difference between studying the country (or a continent) as a whole, or a self-contained labour market, as we see in Figure 3.3. The trade costs and economic growth are shaping the across-region dynamics in Europe and commuting costs and the changes in the distribution of urban infrastructure are likely driving the within city dynamics. The Europe-wide shifts are such that the population continues to agglomerate in the European core from all but

the very large peripheral cities. However, zooming in to a couple of large cities – Moscow (the right square) and Warsaw (the left square) we see very different dynamics from the European-wide dynamics. While Moscow continues to grow at the very centre, and loose people from the better-connected peripheral areas, Warsaw is losing people in the very centre. Different dynamics require a separate analysis to be carried out, even if some of the tools will be similar.

The paper is organized as follows. The data section, Part 2, describes two main datasets that will be used in the analysis – Landscan (LandScan 2012) and the Global Human Settlement Layer European Commission (2015) and will also discuss the nighttime lights dataset which will complement the population data. Part 3 considers the measures and outcomes for a multi-city region. Part 4 is the within-labour market analysis. This section will test whether African cities are different to the cities on other continents and regions. Part 5 will conclude.

3.2 Data

This section will primarily focus on the new population datasets and briefly touch on the nighttime lights satellite data, which can be used in conjunction with the population data in the analysis of people in space.

Population datasets. Today many new excellent datasets are becoming increasingly available to analyze population at a finer scale and consistent units in space. The table in the appendix details the new datasets of the population (Table 3.A2). This paper will focus in detail on two datasets: Landscan and GHSL. The Landscan and GHSL will also be used in Sections 3.4 and 3.3.

For the most part, all datasets listed in the appendix Table 3.A2 follow the same approach. They use satellite images to disaggregate the population data that usually comes at the level of census units of different size around once in 10 years. The major variable extracted from the satellite data is usually built cover. The quality of such data depends on the quality and comparability of the raw inputs – whether the census units were fine enough and near enough in time across countries, whether the built cover was accurately determined by the remote sensing algorithms. Then, some datasets enhance their population information by additional variables, such as roads, soil, locations of villages and others, now even using the machine learning techniques with the help of these ancillary variables that predict where one expects people to be.

There is a clear trade-off between using population data that is enhanced by more ancillary variables and the data that are made via the simpler procedure. On the one hand, with more data such as roads, water, ruggedness, soil, one can predict more precise population location. On the other hand, the simple population data not contaminated by any auxiliary variables allow to explicitly test if the people are present in a location due to, say, roads, and access to water, and not because they are predicted to be there by the way data is constructed. Aware of the trade-off, I use one of each type of datasets.

For the analyses in Parts 3.3 and 3.4, I use the Global Human Settlement Layer (GHSL) of the population at 1km² resolution created by the European Commission. Like the standard method described above, GHSL uses census units across all countries assembled by Columbia University and named “Gridded Population of the World” at the smallest scale available and "smears" it within each census unit based on where the buildings are, using a built cover dataset. Columbia University lists census input data transparently with the year and the number of units per country. Since each country has censuses in different years, i.e. not aligned across countries, the data is projected forward or backwards to get the population at 5-year intervals. The built cover dataset used for the "smearing" is made from Landsat satellite images. Landsat images have around 30km² resolution and exist for the years 1975, 1990, 2000 and 2014. The final GHSL dataset is available for 2000, 2005, 2010, 2015, 2020. The drawback of the final population grid is that the data only look at people, not workers. The strong advantage of it is the transparency of the method and its consistency over time and across space. This makes it the best source of globally consistent panel population data available to date.

For the analyses in Part 3.4, the within-city analysis, I make use of a global dataset of population density developed by Oak Ridge National Laboratory. Their dataset, Landscan (2012), estimates population density, as does the GHSL, at approximately 1km resolution (30" X 30" arc-seconds) for the entire globe. The data represents an "ambient" population, or an average location of a person over 24 hours, and has been made to manage global disaster relief. Population density is estimated by combining Landsat satellite, similarly to GHSL, but also other high-resolution satellite imagery detailing the extent of the built area within a certain location, with population census data at an administrative level such as the ward level⁴. The population for each grid square is disaggregated, with some weights, from the administrative level to the finer 30" X 30" arc-seconds using primarily built cover and other information, such as roads, populated places and geological data. This methodology

⁴This may vary. I enquired Landscan about Dar-es-Salaam and established that three municipality units are used for the whole of Dar-es-Salaam.

reflects the fact that people are likely to occupy other buildings in the city than their place of residence (where they answered their census questionnaire) during the day.

A major limitation of the Landsat data is that while the satellite-generated built cover information is consistent across countries since it comes from the unbiased photographs from the same satellite, the census sub-units used may vary across countries and may be large, and it is not transparent which sub-units are used for each country unlike for the GHSL. Besides, the auxiliary data, such as topographic maps or roads also depend on the data available for each city or country. Finally, the methodology and the weights used to disaggregate the census population data are a black box. Nevertheless, for a cross-city analysis, this dataset is sufficient under the assumption that the measurement error of where people are is not systematic.

I choose to use these data for the current within-city exercises since the number of input variables mentioned on the Oak Ridge website is more than in most other gridded population datasets which making it a possibly superior cross-sectional gridded population dataset in terms of precision⁵.

An accuracy assessment was carried out for the city of Kampala and Nairobi in Henderson et al. (2019), the Chapter 5 of this thesis, using the population and economic censuses and the non-residential building volume. The Landsat for the two cities appeared fairly accurate both in terms of distribution across space and overall volumes of people. However, future work should also assess accuracy across many more cities in different regions.

Nighttime lights datasets This paper will only focus on the use of night lights for the delineation of city extents. However, for completeness, I provide a brief description of the night lights datasets that are now available and can be used for measurements of economic activity and its changes over time.

Night Lights data is introduced in the field of economics by Henderson et al. (2012), and Chen & Nordhaus (2011). Both papers show that night lights data can supplement measures of economic activity in countries where national statistics are poor. They also give insights on growth in GDP and population at the subnational level. After these two publications, night lights data was used extensively in other economic research papers by among others, Michalopoulos & Papaioannou (2013, 2014), Harari (2016), Pinkovskiy & Sala-i Martin (2016), Pinkovskiy (2017).

⁵Nonetheless, caution should be taken, and more tests should be done, as the exact methodology and input variables are not known and not confirmed to be consistent across space.

There are four usable datasets from NASA: Stable lights (“Average Visible, Stable Lights, & Cloud Free Coverages”), the unfiltered lights, radiance calibrated lights and VIIRS.

Most papers so far have worked with the stable DMSP lights as they give the time-series luminosity of cities and towns and are available for 22 years, between 1992 and 2013. In the meantime, this dataset has well-known issues of top coding (values of luminosity truncated for higher values), overglow (lights detected for a wider area than where they are actually emitted) and alignment of luminosity values across years. The following papers addressed the issues: time-series alignment of DMSP (Li & Zhou 2017), top coding (Bluhm & Krause 2016), and blooming (Cao et al. 2019). Overall, some papers have tried to further test the explanatory power of DMSP nighttime lights over time and across space following up on Henderson et al. (2012). The literature is currently split: these papers find that DMSP night lights have weak explanatory power for levels or changes in GDP (Addison & Stewart 2015, Bickenbach et al. 2016, Gibson et al. 2021), and these, on the contrary, find that nighttime lights are a good proxy (Asher et al. 2021, Mellander et al. 2015).

The unfiltered lights are available for the same period as DMSP and can potentially give more values at low luminosity, which is helpful for studies in rural areas with low levels of development, but it is often hard to disentangle it from the natural illumination of the soil.

Radiance calibrated lights combine the stable DMSP with auxiliary data and give more variation at high luminosity levels, as they are not top coded. However, they are only available for several years, namely 1996, 1999, 2000, 2002, 2004, 2005 and 2010, and may be more useful for cross-sectional analysis.

The VIIRS dataset is the new night lights dataset available for 2012-2019 and regularly updated (Elvidge et al. 2021). It has a broader light range, it picks up lower light and is not top-coded, so improves on the original time series of DMSP lights.

For the research looking only at a cross-section or a more recent time-series, the VIIRS dataset is more desirable, but for the research that to studies an earlier time series starting from 1992, one will need to combine DMSP and VIIRS datasets and inter-calibrate these two datasets (see Li et al. (2017), Zhu et al. (2017), Zheng et al. (2019) for the techniques of how to combine DMSP and VIIRS datasets).

While this paper discusses how DMSP nighttime lights are used to delineate city extents, the same approach can be applied to other nighttime lights datasets.

3.3 How to think about density across labour markets.

This part of the analysis will cover the methods for understanding the distribution and dynamics of people across places within a country. How does one describe the degree of spatial concentration of people in a country in an economically meaningful way?

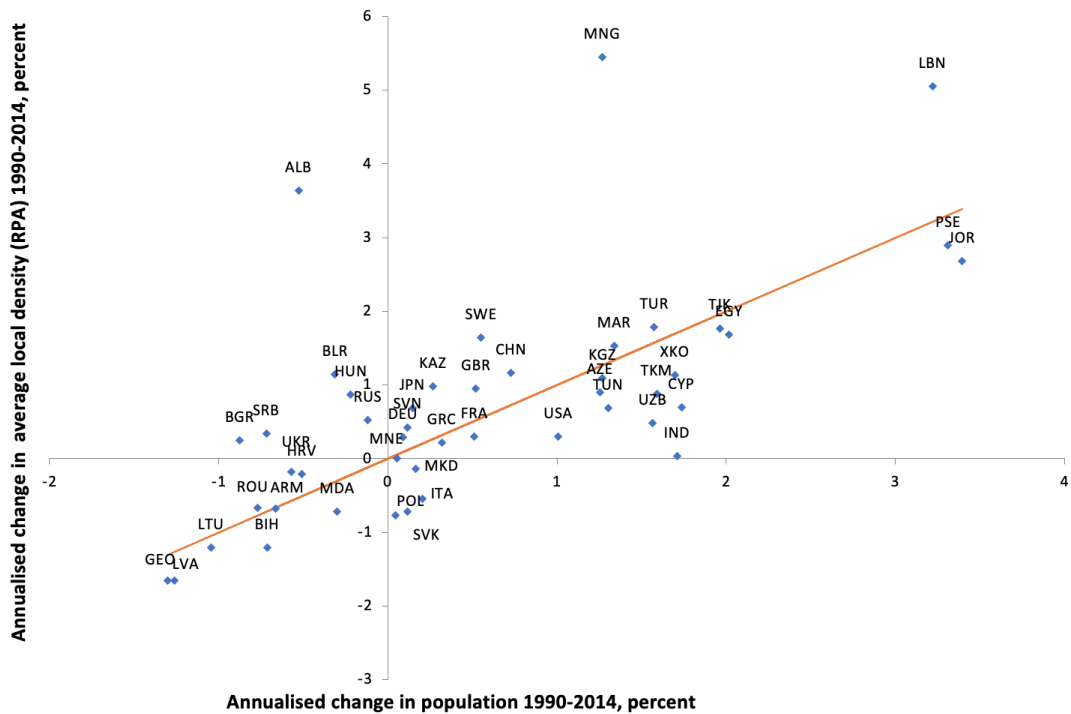
The traditional measure is the share of urban population – which is likely to capture the share of people living in somewhat higher density locations, without revealing what happens *within* urban and rural groups. One step is to add more categories, besides urban and rural. But the two measures that reveal the full distribution about where people are would be 1) the experience of density on average by a person 2) the distribution of this experience or a spatial Gini coefficient. The two measures should capture whether most people are living in one big dense city, or if people are exposed to very different types of density.

Let me describe the two measures more precisely. The average experience of density or local average density is a measure proposed by De La Roca & Puga (2017) when trying to measure the correct city size within pre-defined city boundaries. It is a revealing measure when calculated within *country* boundaries as well. It considers both the average population density and the spatial correlation of this density. The formula uses data from the equal-sized squares, or grids and is the following⁶:

$$RPA_c = \sum_{i=1}^{i=N} w_i * LocalPop_i \quad (3.1)$$

w_i is a weight that is the share of the country population in each cell i ; $LocalPop_i$ is the sum of people within a certain radius of a grid-cell, including the grid-cell itself, discounted by an exponential distance discount $e^{(-0.5*d)}$ where d is distance. In effect, the measure tells us what density on average a citizen of each country is exposed to. It is better than the average density, because countries with much unlivable space such as Spain, the data from which De La Roca & Puga (2017) use, will have a very low average density, but in effect, many people live in dense cities, so the local average density will be high. This measure can be flexible with regards to distance and spatial discount one applies. Why is this economically meaningful? As we saw in the introduction, density has a positive and significant correlation with GDP per capita. Therefore, if countries are getting denser or sparser (holding gains in population fixed) may be important for economic outcomes.

⁶An equivalent formula and its decomposition is shown in part 4.



Notes: The raw gridded population is from the GHSL by European Commission, Columbia University and authors' calculations. Average localised density (RPA) is the average number of people living within 5 km of each individual, discounted locally by distance.

Figure 3.4: A scatterplot of average local density changes and population 1990-2014

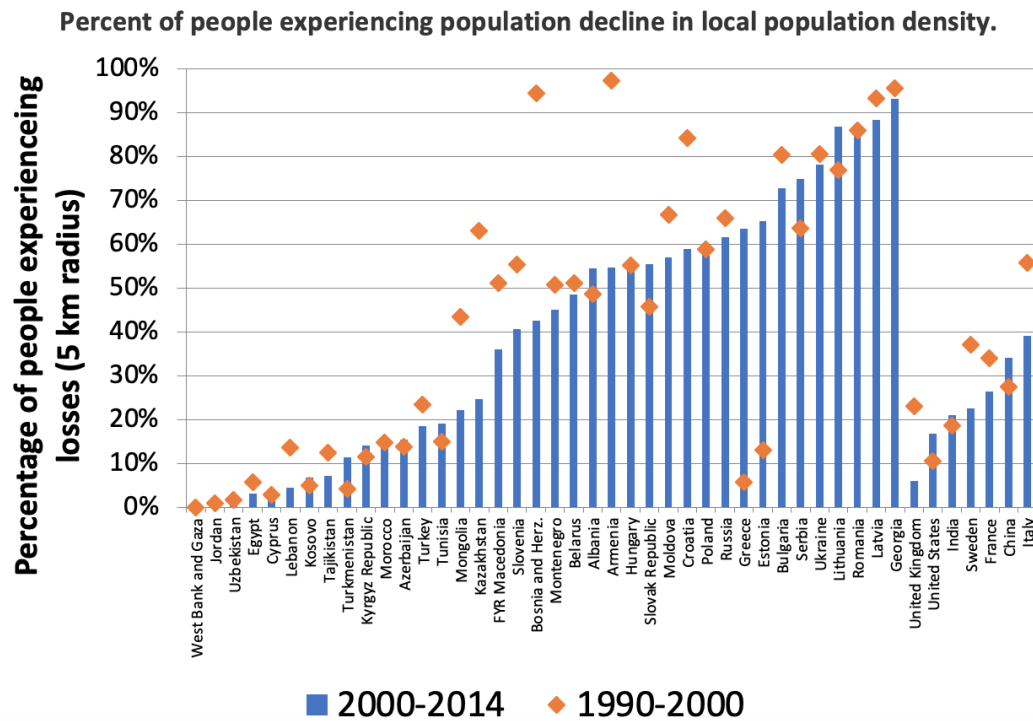
Some interesting patterns can be observed if one looks at the changes of local density using the GHSL at the Eastern European countries and compare it to say other CIS countries and Western European countries. With the accession into the EU, many countries in Eastern Europe have alarmingly lost population. But do these losses correspond with the further dispersion for an average dweller?

Figure 3.4 splits the countries into four quadrants - countries that gained in both the local density and overall population (Kazakhstan, China, Mongolia), countries that lost the population but gained in average local density (Russia, Albania, Bulgaria), countries that lost in both (Georgia, Latvia, Romania), and finally, countries that gained in population but got more dispersed (Italy, Poland, Slovakia). This is interesting, and it seems that the smallest newly European countries have lost people both in larger towns and rural areas, perhaps due to the agglomeration "pull" not being strong enough when the competitors are Frankfurt, Berlin and London. Somewhat larger countries, such as Bulgaria and Serbia managed to retain (or gain) some people in their densest places, including their capitals. All Western European countries, except for Italy and Greece have gained more in average local density than population, which is as predicted by core-periphery patterns Krugman (1991).

Thinking of the agglomeration theories and the correlation I have shown in the introduction, countries that lose in local density should interpret this as a worrisome signal. If Riga and Rome are emptying, it is possible that the opportunities for bigger and better firms with new ideas are being lost. On the other hand, it is possible that even though Hungary and Russia are losing population, it is not yet an alarming sign that the economic potential is lost with it. See Appendix Figure 3.A1 for the chart of comparative losses in average local density.

RPA_c gives us the average density a citizen is exposed to. What if one wanted to learn about how different the experience of density is for different people in each country and how it is changing. To get the differences in levels of the RPA_c across people, one can calculate the share of the population with high RPA_c and with low RPA_c . One can analogously calculate the share of people living in areas experiencing the decline in the local population density. This is relevant to understanding the extent of the "left-behind" people and places, which are abundant in the Eastern European countries and extreme in Latvia and Georgia. Countries with a high share of the "left-behind" population face a lot of political and social pressures.

Already, some research questions emerge from these measures. What is the association



Notes: The raw gridded population is from the GHSL by European Commission, Columbia University and authors' calculations. Figures are based on people's place of residence in 2014. Localised population density (RPA) is a measure of the number of people living within 5 km of a person, discounted by distance.

Figure 3.5: Percent of people experiencing population decline in local population density.

between innovation and the loss of local population density in a country? Looking at Figure 3.5, what is the share of people becoming “left-behind” that is critical for social turmoil? What is the optimal distribution of people across cities? This last question was answered theoretically in Albouy et al. (2019) but can be looked at empirically using these data. Finally, one can think of an ideal measure of population distribution that takes into account the full economic geography: how are people (workers and firms) distributed according to their market access to other populated places. This measure will combine both the local population density, transport links and natural features. Germany with its well-placed and well-connected middle-sized towns comes to mind, where every firm and every worker is within relatively easy access. A contrast to this is Russia, in which, although firms and population are concentrated in Moscow, the productive activity is generated in far-away oil-wells. What is the cost of this lack of connectedness?

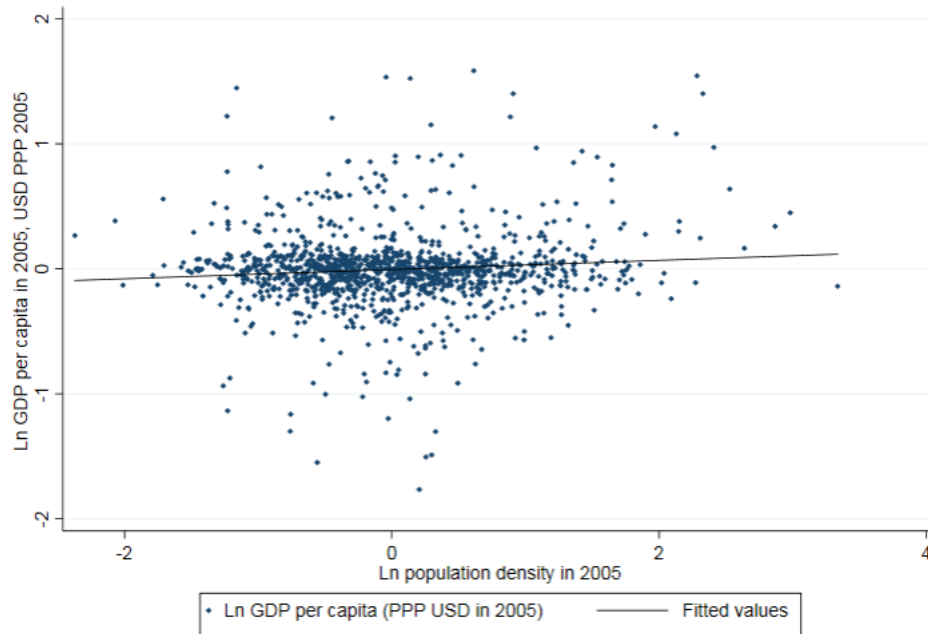
3.4 People in space within labour markets, or why the density and GDP per capita correlation is lower in Africa?

The spatial distribution of people across cities is one angle from which we can look at density. However, it is equally important to look at the spatial distribution within cities. Why?

An analogous regression to Figure 3.1 at 1-degree-by-1-degree units, when controlling for country fixed effects, existing transport density, latitude and longitude, is half of the size for the continent of Africa (1% increase in population density is associated with 0.026% increase in GDP per capita in 1 degree-by-1-degree cell).

Population density has been considered to be beneficial for economic performance because firms located in densely populated areas have access to larger labour markets; households residing in denser areas are within reach of more jobs. Services that benefit from economies of scale, such as hospitals, theatres, schools and restaurants, are cheaper to provide whenever there are many firms and people. The physical proximity of firms and people facilitates the exchange of ideas and innovation. Therefore, in many ways density of firms and people drives economic progress forward.

However, the above scatterplot of Figure 3.1 and knowing that some people in Addis-Ababa or Dar-es-Salaam may walk to their jobs for two hours one-way every day (Franklin 2018), call into question that agglomeration effects are the same in all places with the same density. It seems that some conditions should be met, in addition to density for the growth



Notes: The underlying source of this scatterplot is the G-Econ dataset and authors' calculations. Each dot represents a 1 degree by 1-degree cell with a minimum population of 10 people and minimum GDP per capita of US\$ 7.38 at PPP. The sample is restricted to the grids on the African continent. Blue dots represent the raw scatterplot and the black line is the line of best fit.

Figure 3.6: Correlation of GDP per capita and density of 1 decimal degree by 1 decimal degree units using G-Econ data Nordhaus et al. (2018) on the African continent in year 2005.

to be harnessed. That is, cities with high average density may not necessarily grow faster or be much more productive.

In this part of the paper, I will consider how to measure the distribution of people within appropriately defined cities. Taking the cities on the African continent as an extreme example, using the global Landscan data I will test whether there are some differences between Africa and the rest of the developing world, using a variety of measures describing the city's "shape".

3.4.1 Theoretical overview.

Urban Models

A standard monocentric city model (first developed by Alonso et al. (1964), Mills (1967), Muth (1969), and summarized in Duranton et al. (2015)) predicts jobs located at the centre and people (workers) in the periphery, with the highest density in the centre. Population density at any location in the city is determined by the land rent gradient at this location

over the cost of commuting. If the land rent falls sharply with distance to the CBD population density does so too, because the land rent gradient determines the extent to which housing consumption increases and capital intensity declines with distance to the CBD. High consumption of housing and low capital intensity means lower population density. Lower transport costs increase house prices because the population moves into the city from elsewhere to gain utility from lower commuting costs. Densities rise, and cities expand. A reduction in transport cost reduces the share of the population at the centre as people substitute for their housing expenses with relatively cheaper commuting.

A more advanced model in Fujita & Ogawa (1982) takes into account agglomeration spillovers and allows firms and workers to choose to locate anywhere in the city (CBD is not predetermined). Several configurations can arise as a result:

- 1) Mixed land use in the city centre, where workers live in the area they work, a purely commercial area further on and an area of pure residential use, from which workers commute to the pure commercial area.
- 2) If productivity spillovers dominate commuting costs, the central commercial area surrounded by residential areas.
- 3) If commuting costs dominate productivity spillovers, all of the city is in mixed-use and every worker works where he lives.

The above models assume that the land is going to the highest bidder, which may not always hold in reality, especially in Africa. Therefore, in Africa, there could be two potential causes of firms and people failing to cluster much at the centre of the city: poor transport and poor land markets.

Empirical evidence has shown that with time and at a later stage of development (and better transport connections) cities (at least in the US) have decentralized. This is broadly consistent with the monocentric city model in which the transport cost drops everywhere in the city. However, little is known about the patterns of this decentralization, or the emergence of sub-centres, except several works that have described and modelled this pattern (Garreau 1991, Henderson & Cockburn 1996). In the latter paper, the decision to create a sub-centre is made by a large developer that faces a tradeoff between developing close to the CBD and enjoying the agglomeration effects thereof or having higher local monopsony power and lower rents and lower commuting costs for its workers when locating at the edge. Empirically, McMillen (2003) studies sub-centres in US cities and find that population and travel time increases the number of sub-centres detected in one of the 62 large metropolitan areas.

With the gridded population data, even if the Landsat dataset with its "ambient population" is used, the test of the models that comes in the next sections is only approximate. To fully understand the fabric of cities in developing countries and the Sub-Saharan African region, in particular, one should wait for the separate datasets on workers and residents (e.g., from mobile phone data), information on commercial vs. residential uses, and data on transport infrastructure.

Land Markets

Potentially due to poorly functioning land markets, there are obstacles to the redevelopment of the improperly used land and transfer of land to the user that pays the highest rent. There is anecdotal evidence of areas having lower rents per land area than in the surrounding neighbourhoods, with the same distance to the CBD. For example, Oyster Bay in Dar-es-Salaam has large houses and many gardens on prime land, or Kibera slum in Nairobi does not have tall buildings and big commercial firms. Both neighbourhoods have houses built with relatively low capital, despite what the monocentric model would predict. The model in Fujita & Ogawa (1982) would likely predict mixed or commercial use, rather than residential use, higher rent per land area for both Kibera and Oyster Bay, since both areas are very close to the CBD. As for population density, the Oyster Bay area should theoretically have a higher population density than it is. Although Kibera slum does have a high population density, it is arguably an "unhealthy" density, because the buildings are short and there are many people per unit of floor space, which creates negative externalities.

3.4.2 Methods: city extents.

To see whether any theoretical predictions on the distribution of people within cities are supported by the data, one needs a consistent measure of the city extent. Measures of density and clustering will be affected by how wide or how narrowly the city boundary is drawn. Using official city boundaries is impractical since I want to study a labour market as a whole and many people live and commute from the places outside the official city extents. I propose three alternatives for defining city boundaries and boundaries of the city core, where the city core is loosely defined as the central commercial area of the city, where most jobs are concentrated: using night-lights, density, and built cover. The use of all datasets for urban extents will bring an issue of arbitrary thresholds for density, built cover or night light intensity, which I elaborate on below.

Night lights, specifically the DMSP nighttime lights dataset provide contiguously lit poly-

gons, in which the full commuting zone is usually included due to light from cars on the roads. This approach works even in such cities as Nairobi or Dar-es-Salaam, as visually inspected with on-the-ground knowledge of the cities. Another benefit of the DMSP nighttime lights data is that it allows refining the city extent and the city core with the use of the time dimension: for example, the city core can be proxied by “lit above a certain threshold” and “lit in all years”, and that way any short-run light emissions do not distort the city extent. Another plus of using the DMSP night lights is that the light emitted is smooth and one does not see many “holes” within the city extents. The major drawback of the DMSP night lights data is that you must impose different light thresholds globally. In Sub-Saharan Africa, no threshold is necessary to define distinct city polygons, but even in India electrification is high enough for many large separate cities, or labour markets to merge into one at least when the DMSP nighttime lights dataset is considered⁷. Therefore, night lights need to be truncated from below everywhere by different thresholds, except Sub-Saharan Africa. Dingel et al. (2019) attempt to match the right night-lights threshold with that of commuting flows in Brazil (30) and apply it to other developing countries such as China. However, this approach works only if the degree of electrification is the same everywhere else as in Brazil, and the commuting flows are the same. In the rare event that the region of study has similar levels of electrification across all its locations, night lights with the same threshold can be used for defining city extents consistently.

Build cover is becoming more and more popular for defining the borders of a city (Marron Institute 2017, European Commission 2018 to name but a few). It is a robust method as it captures contiguously built areas that define a city based on satellite data. However, still, as of today there are computational constraints inferring building from satellite images robustly. Landsat is argued to be an important exception. A low-resolution 30m-by-30m satellite has enough colour bands to distinguish man-made from natural (Marron Institute 2017) and the two above-mentioned projects, European Commission and Marron Institute, have provided their boundaries of building extent for 1990, 2000 and 2014 for open use. However, again, neither data provider attempts to calibrate the correct threshold for build cover to match the commuting flows. For example, Marron Institute uses the threshold of 50% of pixels built within a 1km radius. Also, see Baragwanath et al. (2019) for a method combining built cover information and nighttime lights to define urban areas in India.

Finally, I have developed my own computationally simple measure of urban extent based

⁷It is expected that the problem of having to impose different thresholds across countries will arise for the VIIRS nighttime dataset as well since the level of light emitted by an area varies highly with the level of the economic development of an area.

on density itself either for Landsat or GHSL. To calculate the density of each grid-cell, I applied a smoothing methodology where each reference cell is assigned the average density from a neighbourhood which includes itself and a 7x7km square around it⁸. This approach is essential for dealing with natural breaks in density in the data which may relate to changes in land use, terrain, and building restrictions within urban areas. Consider, for instance, Central Park in Manhattan, or the River Thames in London; assigning a density criterion just to each singular grid-cell would lead to unnatural breaks/holes in our urban area polygons which would challenge our ability to analyze our urban areas as singular units.

Once I know the average density of each grid cell within its 7x7km neighbourhood, I consolidate all contiguous grid cells that have an average density above a certain population threshold⁹.

Contiguous cells are combined based on a rook neighbour relationship, wherein, the rook neighbours are the four cells adjacent to the reference cell in the vertical or horizontal direction, but not including the diagonally adjacent cells which are queen neighbours. Finally, contiguous cells meeting certain density criteria will be classified as a city (including the urban core and fringe) or an isolated low-density settlement. The measure based on density provides an alternative to the built cover, due to ease of computation and provided that the population data is detailed enough. Furthermore, I can specify more stringent thresholds to identify an urban core or cores. This beats using the built cover data from Landsat to define urban cores since the share built is approaching 100% in most cities already very far from the city centre.

3.4.3 Methods: combining data from different geographies

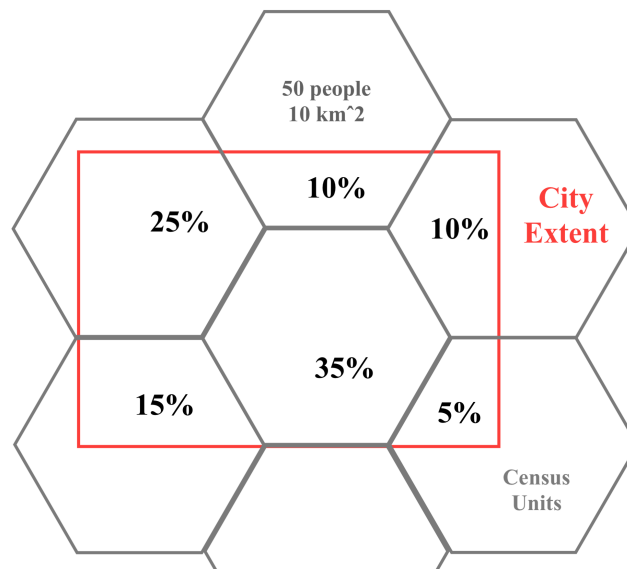
Once the city extents are made, it is often the case that one would like to combine it with other data. The obvious data that one can add to the city extents first, is the information from the census variables, area GDP or luminosity from the nighttime lights. The raw data, be it population or nighttime lights, that itself was used to creating the city extents as discussed in the previous section is easy to add as a sum or mean of raw values, since it is naturally fully aligned with the extent boundary. However, any other data, such as the census units or even the grids from other datasets of different size will rarely align

⁸This smoothing level was the minimum for the holes in polygons to disappear after manual trial-and-error testing.

⁹This threshold, unfortunately still has to be arbitrary at this stage. For the analysis that follows I use the threshold of 1500 people per square kilometre on average in 7 X 7 cell neighbourhood, and this threshold has been accurate enough for cities in both Africa and Asia.

perfectly with the city extent created with nighttime lights or with a gridded population. How would one proceed? Let us use an example of the census units with the counts of literate individuals, covering the area of the city with some units being "split" with the boundary of the city. How to assign the precise count of literate individuals to the city? This method can be used to assign data to any arbitrary boundary, not just city extents, from other arbitrary spatial units, not just the count of literate individuals (e.g. soil quality grid, GDP grid, elevation grid).

The logic is straightforward: 1) assume the data is distributed uniformly within the units and calculate the density of these data per 1 unit of space 2) take a weighted average of the density from each data unit, weighted by the share of unit area (or an area of the intersected part of the unit) in the total area the city extent 3) multiply the weighted sum with the area of the city to get the counts.



Notes: A sketch of how to assign arbitrary data from spatial units (grey hexagons) onto a city extent (red rectangle). Black percentages are the shares of the areas of the partitions created by the hexagons in the total area of the city extent.

Figure 3.7: An example of assigning data from arbitrary spatial units

Using 3.7 as an example, the red box is the schematic city extent, and the grey hexagons are the schematic census units. First, one would calculate the literate population density (or the density of another variable of interest that comes in that spatial unit) at every point within the "hexagon", assuming the count in the census is uniformly distributed within this unit (for the top middle hexagon, the density is 5 people/km²). Second, one would calculate the shares of intersections of each hexagon and the city extent. These shares are indicated in black in the sketch and add up to 100%. Third, one would use these shares as

weights and calculate a weighted sum of the density of the variable of interest (the weight of the top-middle hexagon is 10%, so its contribution to the weighted sum is 0.5 person per km²). Finally, one can convert the variable of interest back to the "counts" by multiplying the density with the area of the city extent.

This method provides an easy rule for combining the datasets spatially. An even easier rule can be to only assign the population from the census units to the city boundary from those units that are completely within the boundary or those units that have their centroid within the city boundary, and weighting by the share accordingly. This is a faster way to combine spatial data, but it loses the information from the cut census units, which may be substantial if the units are very large. Chapter 4 used this method when we assigned the census population information to the treatment areas and a similar method has been proposed by Eckert et al. (2020) to harmonise the spatial units across time in the US.

3.4.4 Methods: within-city placement of people.

What measures of the distribution of people in a city capture the ease with which agglomeration economies can take place? Tokyo is poly-centric and is considered a successful city with a well-functioning subway system. Moscow is monocentric and is well-known for its congestion forcing people to spend 1-2 hours to move within zone 1 to attend a work meeting during weekdays. At the same time, there are other successful monocentric cities, such as Paris. There are of course such important factors as transport systems that need to be taken into account when comparing people's distribution in cities. All cities have different area, size and shape, all of which can matter for their productivity (Harari 2016).

Ideally, one could measure what it takes for firms to connect with each other, for people to connect with jobs and for suppliers to connect with customers. As opposed to country-level connectedness, where the trade and shipping costs are some of the key drivers of the distributions of people in space, within a city, it is the commuting costs that come to the forefront. Are there relevant measures of people's placement that can proxy how disconnected people and jobs are in a city?

Naturally, average local density (RPA) comes to mind from Part 3 De La Roca & Puga (2017), but this time, within the city rather than within the country. When the population is held constant, this measure discounts negative spatial correlation and rewards the concentration of density, especially in one centre. When the city has primarily one large cluster, the highest weight is assigned to the highest sum within the radius, so the RPA score is higher.

We illustrate this visually in Chapter 5. Figure 5.3 of Chapter 5 shows three schematic cities with the same average density (5 people per square), and the same total population (180 people)¹⁰. Cities 2 and 3 have the same within-square density and the same number of dense and the same number of empty cells. However, City 2 has more clustering across squares, which means more interaction with neighbours. This feature of City 2 will be captured with a higher RPA measure. I also use a measure first proposed by Modi (2004) and called personal population density (PPD), which highlights the difference between City 1 and 3.

Putting the above insights into mathematical language, I demonstrate two algebraic decompositions, one of the RPA measure, and the other of the PPD measure. The former is based on the work of Henderson et al. (2019), and Chapter 5 of this Thesis and the latter is based on the ongoing work by Henderson et al. (2020).

For the RPA agglomeration measure for city j , the decomposition is

$$RPA_j = \sum_i A_{ij} \frac{P_{ij}}{P_j} = AD_j \left(1 + \frac{Cov(A_{ij}, P_{ij})}{AD_j PD_j}\right) \quad (3.2)$$

where

$$AD_j = \frac{\sum_i^{N_j} A_{ij}}{N_j}; \quad (3.3)$$

$$A_{ij} = \sum_{k \in S} P_{kj} e^{-\alpha d_{ik}} \quad (3.4)$$

In equation 2, A_{ij} is the measure over radius s of the discounted sum of neighbours' cells' populations and including own cell population. I use an s of about 6 km, limiting the local radius so I can distinguish later the effects of citywide versus local density. RPA_j is the weighted average of the A_{ij} , where weights are each grid squares share of the city population. AD_j is the simple average of the A_{ij} across grid squares over the city (basically, a spatial moving average of density in a city). While $PD_j = \frac{\sum_{i,j} P_{ij}}{N_j}$ is a simple population density of a city. RPA_j can then be decomposed into the simple average, and 1 plus the covariance of A_{ij} and P_{ij} , divided by their simple averages. The latter term captures the degree to which population is allocated to grid squares with high measures of neighbours (City 2), as opposed to either being uniformly spread (City 1) or being in grid squares which are not clustered with others of high density (City 3). For personal population

¹⁰The definition and decomposition of personal population density is borrowed from on-going work by Henderson et al. (2012, 2019). This is gratefully acknowledged.

		Evenness	
Gini Coefficient	$\frac{\sum_{i=1}^N \sum_{j=1}^N \frac{ x_i - x_j }{2N^2 \bar{x}}}{}$	Mean of all pairwise differences in population (x_i) for each spatial unit (grids i and j) scaled by the mean density of the city \bar{x}	An index that varies between 0 and 1, with 1 standing for all population being in one grid cell; is invariant to total population or density; is "blind" to the values of the neighboring units
Exposure			
RPA and PPD	See equations 3.5 and 3.2 and description in text		A measure that has a real interpretation: the number of people a randomly drawn person is exposed to within a certain distance; depends on total density
Concentration Index			
Herfindahl-Hirschman Index (HHI)	$\sum_{i=1}^N s_i^2$	s^2 is the share in total city population in the spatial unit i	Ranges between $1/N$ for fully non-concentrated population and 1 for the population concentrated in one cell
Centralization Index			
Proportion in the Core	$\frac{\sum_{i \in Core} x_i}{P}$	A simple proportion of people located at the central part of the city	Ranges between 0 and 1; invariant to total population or total density
Clustering Index			
Moran's I	$\frac{N \sum_{i=1}^N \sum_{j=1}^N w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2}$	A spatial autocorrelation index, where w_{ij} is an arbitrary spatial weights matrix that may discount spatial units j as they move further away from spatial units i , W is the sum of all weights, N is the count of all spatial units and x_i, x_j are deviations of the values in each spatial unit i and j from the city mean \bar{x}	Spatial correlation index, invariant to the average density of the city; depending on the spatial weights matrix, takes into account the values of neighboring cells; values range between -1 and 1.

Table 3.1: A list of spatial indices

density (PPD) the measure for city j , defined as:

$$PPD_j = \sum_i^{N_j} PD_{ij} \frac{P_{ij}}{P_j} = PD_j \left(1 + \frac{Var(P_{ij})}{PD_j^2}\right) = PD_j(1 + CV(P_{ij})^2) \quad (3.5)$$

where CV is coefficient of variation; N_j is the number of grid squares in a city, and PD_j is defined as above. In some sense, personal population density is a limiting case of local population density, where the radius around each relevant grid-cell is 0km² rather than 6km². PPD_j can be decomposed into overall population density (PD_j), a typical scale measure, and one plus the coefficient of variation. The latter captures the degree of variation relative to the mean within the city and, thus the degree to which activity is concentrated in particular cells. So Cities 2 and 3 have the same degree of variation and clustering, but one that is higher than City 1 in Figure 5.3.

Note that the coefficient of variation has a long history, starting from Williamson (1965), for use as a measure of regional income inequality within a country. Here I am using it as a measure of economic density inequality within a city or settlement.

Massey & Denton (1988) provide an extensive overview of many other spatial concentration and clustering measures developed in the past 50-60 years and highlights the key differences and uses of each¹¹. When considering which measure to use it is important to remember that they are merely proxies for different dimensions of the location of people (or other agents) in space. Massey & Denton (1988) highlight five such dimensions: evenness, exposure, concentration, centralization and clustering and several indices used to capture each. The authors show that while most of these measures are correlated, each measure that captures a separate dimension and adds information. In the table below I present several indices that capture different aspects of concentration and have been commonly used in the past. This is a by far not exhaustive list, with more measures analyzed in the Massey & Denton (1988), but it gives an exposition of various dimensions which these indices may take.

Moran's I - the well-known index of spatial autocorrelation, has a similar flavour to RPA_i and PPD_i : it measures the variation of a variable in the group of units near each other. One can have a spatial weighting matrix that is identical to the one I use in the RPA_j that has weights decaying by distance within a 6 km radius. The key difference is that

¹¹They do so with a purpose of measuring racial segregation, but most these measures can easily be used when measuring how people locate in space in general, regardless of ethnicity. The same goes for the coagglomeration indices, such as Ellison & Glaeser (1997), or Duranton & Overman (2005) that are variants of the indices in Massey & Denton (1988), but the former indices are about industries rather than different racial groups

Moran's I looks at *variability* alone, and does not incorporate the *exposure* of each unit to nearby population. How many people one is exposed to is related to both how clustered the population is, but also the overall sum of the population in a city. As a result, the measure RPA_j is not an index and has a real interpretation: how many people a randomly drawn person is exposed to in a city. This measure is a function of Moran's I, but it is also a function of density and average exposure AD_j .

Picking a measure should depend on three aspects: 1) which dimension the researcher considers most important in their study 2) which measure ensures historical continuity and comparability with other papers (for example, the Economics literature commonly uses the HHI), and 3) which measure has desirable mathematical properties, for example, for the measure of evenness, the population size invariance may be important and, at last, 4) a good index will have an intuitive interpretation. For the exercise in this paper, the RPA_j and PPD_j were chosen due to their ease of interpretation and the decompositions they generate, however, other measures may be desirable in other settings.

3.4.5 Results: how does economic density in Africa compare with the rest of the world?

Cities on the African continent on average are no less dense than in the rest of the world, however as I have shown, they are associated with lower productivity premium than cities elsewhere. Many cities in Africa, for instance, Nairobi, Kampala and Dar-es-Salaam, are characterized by large informal neighbourhoods, separated from very wealthy low-density residential areas and commercial districts. The very centre of the city is not necessarily the densest and there may be multiple commercial clusters that appear to be mixed with patches of slums close to the centre. On average, the extreme density of some informal and commercial areas create a high average density, however, the firms and workers may still remain disconnected.

The reasons for this may be a lack of investment into transport systems and a dysfunctional land market. Better transport links may make it easier for firms to cluster in the central business district (CBD), because more labour force can reach them at a lower cost, and they can have access to a larger labour pool. A better land market would allow firms to replace the centrally located slums (or low-density residential areas) if these firms can generate higher returns on the land that they occupy (Henderson, Regan & Venables 2018).

Does Landsat data combined with our measures support the idea that economic density in Africa is lower than in other parts of the world, despite what is presumed to be high

population density in urban Africa? I interpret lower economic density as implying that, for the same overall ambient population density, there is less clustering of economic activity within African cities, so that potentially RPA_j and PPD_c , are lower, and certainly that the coefficient of variation and covariance terms in equations (2) and (3) would be lower.

I test this by comparing Africa to the whole world. To avoid an issue of large rural areas merging into one agglomeration in India and China, I only focus on larger agglomerations of density above 1,500 sq. km. Hence, I prefer the higher density thresholds for the urban extents as well as a cross-check with the metropolitan areas listed in the official UN data. The populations of all the listed UN metropolitan areas in the blob should sum to at least 800,000 in terms of UN-reported population. Once I have defined these areas, I then assign to the agglomerations the Landsat population number obtained by summing over all grid squares in the blob.

Given these criteria, I establish a set of 599 cities worldwide, with 451 in the developing world. I ran regressions with the following dependent variables, in logs, as follows: personal population density (PPD), simple population density (PD), the coefficient of variation term in equation (2), the De La Roca-Puga agglomeration measure (RPA), or average local density as the simple average of the local agglomeration measure (AD), and the covariance term in (3). For the RPA measure, I use a spatial discount rate of -0.5, as compared to De La Roca & Puga (2017) who use no discounting. Henderson et al. (2013) find that the optimal rate of discount for a particular and narrower context in Africa is close to -0.5, as shown in Chapter 5.

My regression results are presented in Tables 3.2 and 3.3, where the top panel of each table gives the basic results controlling just for terrain ruggedness from Nunn & Puga (2012) and will represent what the raw data tell us. The bottom panel additionally controls for GDP per capita from the Penn World Tables (PWT 7.0), to see the extent to which differences in levels of development explain the density patterns.

In the top panel of Table 3.2, the base case is the 148 large cities in developed countries. Relative to these, Sub-Saharan Africa cities have higher measures across the board, including in particular the coefficient of variation and covariance terms, where they are respectively 44 and 27 per cent higher. Moreover, terms for Africa are higher than the rest of the developing world terms, including those for the coefficient of variation and covariance terms. With no separation into workers and residents, I do not know if this involves greater clustering of residences or workplaces or both; it is the ambient population as the measure of economic density.

Panel A: No GDP Control

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln(PPD)	Ln(Pop. density)	$1 + CV^2$	Ln Puga 5km, $e^{0.5d}$	Ln Puga 5km, $e^{0.5d}$, unweighted	1+Cov. Term
Sub-Saharan Africa	1.101*** (0.139)	0.662*** (0.101)	0.440*** (0.0778)	0.930*** (0.128)	0.665*** (0.0908)	0.265*** (0.0557)
Rest of the Developing World	0.903*** (0.110)	0.564*** (0.0885)	0.339*** (0.0457)	0.792*** (0.111)	0.573*** (0.0724)	0.219*** (0.0425)
Ln(Ruggedness)	0.0938*** (0.0244)	0.0535*** (0.0143)	0.0403** (0.0197)	0.0574*** (0.0178)	0.0475*** (0.0141)	0.00988 (0.00841)
Ln Pop Landscan, thsnds	0.179*** (0.0162)	0.164*** (0.0129)	0.0152 (0.0171)	0.291*** (0.0146)	0.198*** (0.0154)	0.0934*** (0.0102)
Observations	602	602	602	602	602	602

Panel B: GDP Control

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln(PPD)	Ln(Pop. density)	$1 + CV^2$	Ln Puga 5km, $e^{0.5d}$	Ln Puga 5km, $e^{0.5d}$, unweighted	1+Cov. Term
Sub-Saharan Africa	0.541*** (0.183)	0.411*** (0.130)	0.130 (0.149)	0.441*** (0.149)	0.347*** (0.119)	0.0944 (0.0797)
Rest of the Developing World	0.618*** (0.106)	0.436*** (0.0868)	0.182** (0.0832)	0.543*** (0.0960)	0.410*** (0.0712)	0.133*** (0.0458)
Ln(Ruggedness)	0.0991*** (0.0236)	0.0565*** (0.0151)	0.0426** (0.0182)	0.0625*** (0.0179)	0.0513*** (0.0148)	0.0111 (0.00780)
Ln GDP per capita	-0.180*** (0.0495)	-0.0794** (0.0346)	-0.100** (0.0449)	-0.156*** (0.0386)	-0.101*** (0.0325)	-0.0552** (0.0216)
Ln Pop Landscan, thsnds	0.168*** (0.0158)	0.158*** (0.0123)	0.00974 (0.0165)	0.282*** (0.0150)	0.191*** (0.0150)	0.0903*** (0.0104)
Observations	599	599	599	599	599	599

Notes: Table reports results from OLS regressions; errors are clustered by country. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*).

Table 3.2: Comparing within-city distributions of population in Africa with the developed countries using Landscan data

The bottom panel adds a control for $\ln \text{GDPpc}$. This reduces the Africa terms making them smaller absolutely and relative to the rest of the developing world. Now the differentials on the coefficient of variation and covariance terms are insignificant.

In summary, in the raw data Sub-Saharan African cities have higher coefficients on the coefficient of variation and covariance terms, which contradicts the initial presumption. Moreover, greater clustering seems to be negatively related to GDPpc , with developed countries having the lowest degree of clustering, perhaps where sprawling car-dependent cities in the US come to mind.

In terms of just developing countries, Table 3.3, shows that relative to Asia, the outlier with lower clustering is Latin America even controlling for income. Sub-Saharan Africa, as well as North Africa and the Middle East, have similar measures of density and clustering as Asia. Overall, compared to the rest of the developing world, Sub-Saharan Africa cities have (not controlling for GDP per capita) higher average densities of people, but no different degree of economic density as measured by PPD, a higher measure of RPA and no different degree of clustering. Controlling for income, Sub-Saharan Africa cities are similar to others in the developing world in all measures.

These results demand further study. Why do we see that Sub-Saharan Africa is not different to Asia in terms of spatial correlation and African cities are even denser? One reason could be that in the absence of modern transport connections, the density of people compensates, however it is a puzzle why the clustering measures are no different to Asia. Alternatively, the Landsat data may not be consistent across continents and a lot more precise in Asia, yielding unexpected results. Ideally, we would have a dataset with precise transport costs. It would be also revealing if we had worker's density and ran the same regressions. Possibly machine-learning techniques identifying commercial buildings with high-resolution satellite images and combined with firm surveys can help us get closer to answering these questions. Another reason might be that our threshold based on Landsat population data is incorrect and is too low for some cities outside of the Sub-Saharan Africa region. Or, the 1km^2 scale is still too coarse to capture the economically relevant clustering. Future work should include tests of this exercise using both other more detailed input data.

Nevertheless, there may be a weak link between the spatial correlation of population (or worker) density and the economic performance of African cities. It is also key that the where people are in cities should be studied in conjunction with the transport network, and the measures we currently have are imperfect¹².

¹²An alternative explanation can be found in Henderson et al. (2020), where the authors develop a *country-*

Panel A: No GDP Control

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln(PPD)	Ln(Pop. density)	1+CV ² term	Ln Puga 5km, e ^{0.5d}	Ln Puga 5km, e ^{0.5d} , unweighted	1+Cov. Term
Sub-Saharan Africa	0.0854 (0.104)	0.137*** (0.0522)	-0.0520 (0.0764)	0.137* (0.0693)	0.145*** (0.0477)	-0.00869 (0.0422)
Latin America	-0.211** (0.0918)	0.0638 (0.0505)	-0.274*** (0.0834)	-0.0240 (0.0653)	0.0768 (0.0609)	-0.101*** (0.0350)
North Africa	0.214 (0.220)	0.127 (0.175)	0.0873 (0.115)	0.111 (0.208)	-0.00544 (0.202)	0.117* (0.0684)
Ln(Ruggedness)	0.0796*** (0.0272)	0.0409** (0.0163)	0.0387* (0.0218)	0.0400** (0.0192)	0.0402** (0.0164)	-0.000219 (0.00796)
Ln Pop Landscan, thsnds	0.152*** (0.0240)	0.155*** (0.0173)	-0.00331 (0.0250)	0.287*** (0.0194)	0.199*** (0.0196)	0.0873*** (0.0144)
Observations	454	454	454	454	454	454

Panel B: GDP Control

	(1)	(2)	(3)	(4)	(5)	(6)
	Ln(PPD)	Ln(Pop. density)	1+CV ² term	Ln Puga 5km, e ^{0.5d}	Ln Puga 5km, e ^{0.5d} , unweighted	1+Cov. Term
Sub-Saharan Africa	-0.0444 (0.132)	0.0712 (0.0645)	-0.116 (0.112)	-0.00812 (0.0792)	0.0375 (0.0580)	-0.0456 (0.0590)
Latin America	-0.118 (0.103)	0.114* (0.0614)	-0.232** (0.0901)	0.0811 (0.0772)	0.158** (0.0662)	-0.0764** (0.0344)
North Africa	0.178 (0.202)	0.108 (0.165)	0.0696 (0.112)	0.0708 (0.187)	-0.0359 (0.186)	0.107 (0.0689)
Ln(Ruggedness)	0.0843*** (0.0269)	0.0442*** (0.0167)	0.0401* (0.0207)	0.0457** (0.0197)	0.0450*** (0.0170)	0.000649 (0.00780)
Ln GDP per capita	-0.115** (0.0551)	-0.0595 (0.0407)	-0.0555 (0.0572)	-0.129*** (0.0443)	-0.0967** (0.0388)	-0.0320 (0.0252)
Ln Pop Landscan, thsnds	0.151*** (0.0238)	0.154*** (0.0169)	-0.00345 (0.0249)	0.285*** (0.0183)	0.198*** (0.0191)	0.0871*** (0.0140)
Observations	451	451	451	451	451	451

Notes: Table reports results from OLS regressions; errors are clustered by country. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*).

Table 3.3: Comparing within-city distributions of population in Africa with the developing countries using Landscan data

3.5 Conclusion

This paper has presented, described and developed spatial methods and new data for measuring the distribution of people across space. I focused both on measuring people across cities and within cities and proposed, elaborated on and tested measures, such as local population density, personal population density that can capture the spatial dimension of where people are relative to one another in an economically meaningful way.

A few interesting questions emerged from the analysis. The cross-city study demonstrated that many countries in Eastern Europe are losing people, but only some are losing local density. How detrimental is the loss of local population density for the country's agglomeration benefits, such as innovation, preservation of supply chains, local producers? What share of the population should have been experiencing local losses in population for the country to have political tensions? What is the optimal distribution of density within a country?

The within-city study has shown some surprising findings: African cities are denser than the rest of the developing countries, and not less cohesive, despite what I expected. This poses a question. Can we capture the "efficiency of agglomeration" in a city using gridded population data, or one must resort to other data, such as transport networks and average commuting times? The latter data is still difficult to find, although some impressive advances were made using Google maps (Akbar et al. 2018)

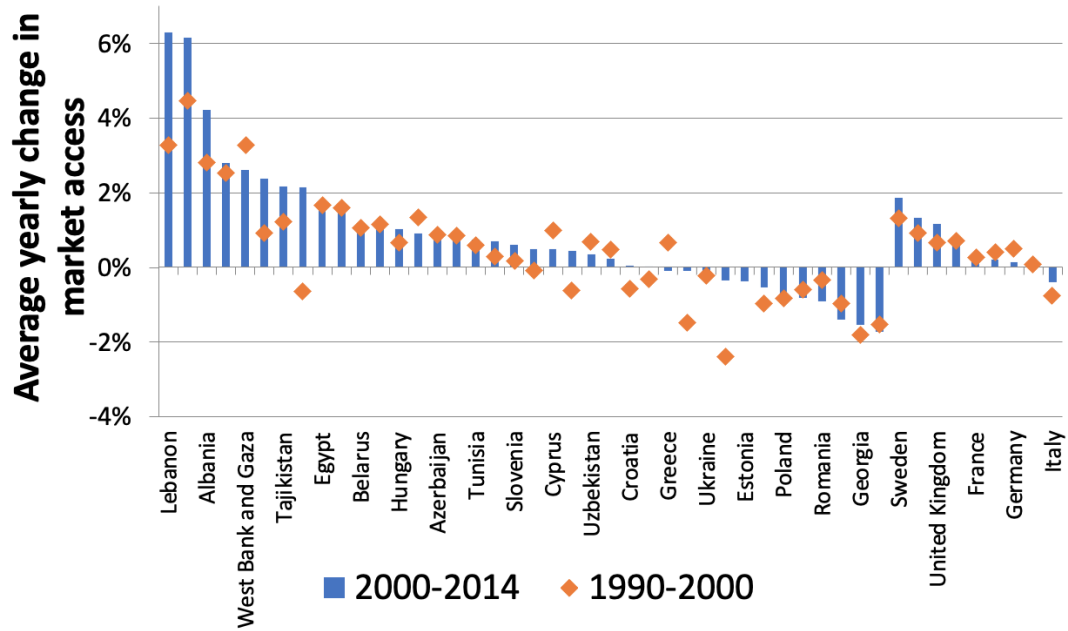
This study also showed a need for more ground-truthing and understanding the issues with the gridded population, built cover and satellite data. Often even the raw inputs differ across countries for the same dataset, and the data should not only be chosen for its accuracy, but also the consistency of methodology. Comparing and contrasting the novel spatial datasets with the traditional survey and economic census data would be a contribution to the field of economic geography. See Jean et al. (2016) for the first push in this direction.

Nonetheless, I advocate the use of new gridded data, as it helps us to see space in new, more nuanced ways. The gridded data based on satellites has the benefit of being more

wide measure of population density controlling for a much wider set of land quality characteristics and find the negative correlation between GDP per capita and the adjusted population density. Therefore, many Sub-Saharan African countries see their countrywide population density as too high given the natural fundamentals, rather than too low. They further show that across countries it is not the density per se that determined the economic outcomes, but the high density and lower GDP happen to co-exist due to the timing of development: the health improvements preceded economic improvements in the developing countries. This lead to higher density to coexist with lower GDP per capita

consistent and comparable across space, allowing the traditional city models and theories of geography can be tested and calibrated. These data also allow researchers not only to use the data in the USA (where the most detailed data are available) but also in other parts of the world, where the answers are most needed.

3.A Appendix A. Additional tables and figures



Notes: The raw gridded population is from the GHSL by European Commission, Columbia University and authors' calculations. Figures are based on people's place of residence in 2014. Localised population density (RPA) is a measure of the number of people living within 5 km of a person, discounted by distance.

Figure 3.A1: Changes in local population density for Eastern Europe and CIS countries.

Name	Collaborators	Resolution	Inputs	Method	Time series	Coverage	Other Variables	Link	Notes
Landscape	Oak Ridge National Lab	1 km ²	Satellites, road and settlement shapefiles, census units	Smearing population into locations according to all possible data available	Not consistent across time due to changes in methods and inconsistent use of censuses	World	no	https://landscan.ornl.gov/	Information as of 2005 (may have changes):
Population Grid of the World (GPW version 4)	Columbia U (CIESIN)	1km ²	Census units	Splitting the population from census units into grids, more detail here https://sedac.uservoices.com/knowledgebase/articles/799296-what-source-data-is-used-in-gpwv4	5 year intervals 1990-2015, where population is extrapolated from census time (different in each)	World	yes, gender, age and other	http://sedac.ciesin.columbia.edu/data/collection/gpw-v4	Several global data products use ancillary data in
Global Human Settlement Layer (GHSL) (basically an)	European Commission + "Joint Research Centre"	1 km ² and 250m*250m	Landsat across time, GPW version4 (above)	Smearing GPW population into built-up cover	1975, 1990, 2000, and 2014 Landsat satellite time series, and corresponding GPWv4 datasets	World	built-up cover, city extents, population	http://ghslsys.jrc.ec.europa.eu/	Time series are due to the Landsat differences, and
[Obsolete] Geostat (CORINE land cover and	Eurostat	1 km ²	Census units+ some land cover datasets	Distributing census population within units into land cover	Not consistent across time, available for 2001, 2006 and 2011	Europe	no	http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat and	Probably obsolete
[Obsolete] ESRI Population grid 2013	ESRI	250m*250m	15m resolution imagery and ancillary datasets, such as	Allocating census unit population to space using the ancillary datasets	2013	World	no	http://www.esri.com/esri-news/arcnews/fall14/articles/esri-new-population-map-is-the-most-detailed-in-the-world	Probably obsolete
WorldPop	Southampton U (Flowminder) collaborating with CIESIN	100 m ² and 1 km ²	Latest available census at most disaggregated units (if available),	Allocating census unit population to space using the ancillary datasets within settlements	2000, 2005, 2010, 2015, 2020 estimates	Planned for the World, but exists for most of the developing world as of 2018	yes, births, pregnancies, age structures	http://www.worldpop.org.uk/data/methods/	This data is to be improved with time
High Resolution Human Settlement Layer (HRSL)	Facebook, World Bank, Columbia U	30m*30m	Census units and Digital Globe	Allocates census units to classified DigitalGlobe Satellite Imagery	2015	Available for Algeria, Burkina Faso, Cambodia, Ghana, Haiti, Ivory Coast, Kenya, Madagascar, Malawi, Mexico, Mozambique, the	no	https://www.ciesin.columbia.edu/data/hrsl/	
Modelling and forecasting African Urban Population	Belspo (Belgium Science Policy)+Worldpop (Flowminder)	12.5 m ²	High-Res Satellites (many) with built cover+ census	Complicated methodology, that identifies buildings. After, use census to allocate people into buildings http://maupp.ulb.ac.be/page/methodology/	1995, 2000, 2005, 2010 and 2015.	Planned for 48 countries in Africa	They identify individual buildings, and in addition, a very high spatial resolution built-up and population maps will also be produced for 3 cities: Ouagadougou (Burkina Faso), Dakar	http://maupp.ulb.ac.be/	Might not be ready yet

Figure 3.A2: Summary of gridded population datasets.

3.B Appendix B. Calculating metropolitan areas for the world

In Section 3.4.5, I create polygons for cities with a population above 800,000. These metropolitan areas are defined by the following criteria: first, each grid cell in the metropolitan

area must have an average density of 1,500 people per sq. km. or more; second, each polygon must contain at least one metropolitan area of above 300,000 people as defined by the UN (2015); third, the sum of the UN metropolitan area or areas within our polygon must be at least 800,000 people. I use the intersection with the UN metropolitan areas population data as it provides an external check on the Landsat data. One issue appeared in India and China, in particular: large swathes of seemingly high-density rural areas had been combined into gigantic urban areas, which were not urban areas in reality. Presumably, this was occurring because Landsat is smearing people in these regions into agricultural areas. Hence, the UN allowed me to sense check the data against reported evidence. Applying our initial density-cutoffs to create contiguous polygons, leaves us with a set of 13,638 metropolitan area polygons. After I overlay this with the coordinates of the metropolitan areas above 300,000 in the 2015 UN data, I get 1,544 polygons which match, and I therefore keep. Only 54 of the total 1,692 UN points above 300,000 did not match with our polygons. In the case that several metropolitan areas fall into one polygon, the UN population is summed up, and the name of the largest metropolitan area is kept. Once I apply the total population threshold of above 800,000, I am finally left with a set of 599 cities worldwide, with 451 in the developing world.

Chapter 4

Planning ahead for better neighborhoods: evidence from Tanzania

4.1 Introduction

Africa's cities are growing rapidly. With its expanding population (United Nations 2015) and rising urbanization rate (Freire et al. 2014), we expect that almost a billion people will join the continent's cities by 2050. But many of these cities, especially in Sub-Saharan Africa, already face problems of poor infrastructure and low quality housing (Henderson et al. 2016 and Castells-Quintana 2017). According to UN Habitat (2012), as many as 62% of this region's urban dwellers live in slums, whose population was expected to double within 15 years. The poor living conditions in those slums have important consequences for residents' lives (Marx et al. 2013).

There are various policy options for addressing the immense challenges posed by African urbanization. One option, which is often the default, is to allow neighborhoods to develop organically without much planning or infrastructure. At the other end of the spectrum, a second option is for the state to not only plan but actually build public housing. This is expensive, but has been done for example in South Africa (Franklin 2020). Between these two alternatives lies a third option of laying out basic infrastructure on the fringes of cities, and allowing people to build their own homes, an option advocated by Romer (2012) and Angel (2012). A fourth option is to improve infrastructure in areas where low quality housing develops.

Understanding the implications of these options is important for current policy discussions. For example, there are debates about the respective merits of upgrading and starting anew (e.g. Duranton and Venables 2020). But we know relatively little about these options' implications for private investments and the survival of infrastructure, or about

their distributional consequences. One of the main contributions of this paper is to shed light on these issues. To do so, we study the long run development of neighborhoods, which were part of the "Sites and Services" projects (described below).¹ These took place not only in Tanzania's biggest city, Dar es Salaam, but also in six of its secondary cities.²

Our paper focuses on the long run consequences of the third option discussed above (*de novo*) compared to the first (default) option of unregulated development. We study *de novo* neighborhoods, which were developed in greenfield areas on what were then the fringes of Tanzanian cities. The developments included the delineation of formal residential plots and the provision of basic infrastructure, consisting primarily of roads and water mains. People were then offered an opportunity to build homes on these plots in exchange for a fee. To provide a counterfactual, we use nearby *control* areas that were also greenfields before the Sites and Services projects began. We also provide descriptive evidence on the fourth approach discussed above by studying *upgrading* areas, which received infrastructure investments similar to the *de novo* areas, but only after people had built low quality housing.³ We compare these upgrading areas to nearby areas and also, for Dar es Salaam only, to slums that were not upgraded.

We investigate how different neighborhoods developed over more than three decades, and we ask a number of questions. First, do *de novo* investments solve coordination failures and facilitate neighborhood development in the long run? Second, how do they shape private housing investments and the survival of public infrastructure? And finally, what characterizes the sorting of owners and residents across neighborhoods, and to what extent can owners' sorting account for the differences in outcomes across neighborhoods?

Concretely, we study Sites and Services projects, which were co-funded by the World Bank and the Tanzanian government, and were similar to projects carried out in other countries. In Tanzania they were implemented in two rounds: one began in the 1970s and the other in the early 1980s. Altogether, 12 *de novo* neighborhoods and 12 upgrading neighborhoods

¹Throughout the paper we refer interchangeably to: "areas" and to the "neighborhoods" that develop in them; houses and housing units; and squatter settlements and slums. Finally, we refer to "owners" as those with de-facto rights to reside in a house or rent it out. Legally, even formal ownership consists of a long and renewable lease from the state.

²This is important because Africa's secondary cities are relatively understudied, despite being home to the majority of its urban population. See for example Brinkhoff (2017), Agence Française de Développement (2011), National Oceanic and Atmospheric Administration (2012), and Tanzania National Bureau of Statistics (2011).

³Unlike *de novo* areas, however, upgrading areas did not receive formal plots.

were developed in Dar es Salaam, Iringa, Morogoro, Mbeya, Mwanza, Tabora, and Tanga. (World Bank 1974a,b, 1977a,b, 1984, and 1987).⁴

To study the consequences of de novo investments, we combine high resolution spatial imagery on all seven cities and building-level survey data on three of the cities with historical imagery and maps. We analyze these data using a spatial regression discontinuity (RD) design. We find that in de novo areas, houses are larger and more densely and regularly laid out, are better connected to electricity, and (in some specifications) also have better sanitation. A "family of outcomes" index and a hedonic measure of house values show that de novo areas have higher quality housing. These results, which are robust to the inclusion of various controls and robustness checks, demonstrate the crowding-in of private investment in response to the public de novo infrastructure investments. We also find that de novo areas have better access to roads and water mains, reflecting the persistence of the Sites and Services infrastructure investments over several decades.

To shed light on the mechanisms that underlie our findings, we develop a simple model of owners' investment decisions, which features complementarity between public and private investments. In de novo neighborhoods, where a sufficient fraction of owners can invest in housing quality, infrastructure investment crowds in private investment in housing quality, which in turn preserves infrastructure quality. This virtuous feedback, however, does not occur in upgrading areas, both because the existing stock of (low quality) housing disincentivizes wholesale reconstruction of housing, and because owners' credit constraints prevent them from investing sufficiently, and as a result infrastructure deteriorates. At the same time, in control areas the infrastructure investments are lower, so no high quality housing is built.

The model helps us interpret our empirical findings in two important ways. First, it allows us to separate the roles of owners' different credit constraints from the effect of infrastructure investments, when comparing de novo and control areas. In practice, we find that adding owner fixed effects reduces the quality differences between de novo and control areas by up to one-third, but these differences remain large and precisely estimated. Second, the model allows us to infer land value differences across neighborhoods from differences in housing quality. Our calculations suggest that the local gains in land value from de novo were, at least in Dar es Salaam, no less than \$75-100 per square meter of plot (in 2017 prices). These gains far exceed the costs of the project, which amounted to no more than \$8-13.

⁴Until recent years the government maintained sole authority for creating new formal plots in Tanzania, so we cannot study the long run consequences of privately provided plots.

In our empirical analysis we also use census micro data to characterize the sorting of residents across neighborhoods. We find that as of 2012, de novo neighborhoods attracted better educated residents, who likely had higher incomes to pay for better amenities. The sorting on education across neighborhoods is, however, only partial: about 45 percent of the adults in de novo areas had no more than a primary school education. Furthermore, even less educated people who initially owned de novo plots and eventually sold them likely gained from some of the land value appreciation.⁵

In contrast to our findings on de novo areas and in line with our model, our descriptive analysis of upgrading areas suggests that their housing quality is either similar to that of nearby areas or non-upgraded slums, or in some cases even worse. Our findings also suggest that upgrading areas do not enjoy better access to water mains or roads than the control areas, so the Sites and Services investments in these areas likely deteriorated. These results should be interpreted cautiously, however, since it is harder to find a clean counterfactual for upgrading areas (which were populated to begin with) than for de novo areas.

The economic evaluation of de novo Sites and Services areas is thus the focal point of our paper. Previous studies of Sites and Services around the world include surveys (e.g. Laquian 1983) and critical discussions (e.g. Mayo and Gross 1987 and Buckley and Kalarickal 2006). In the Tanzanian context, there are descriptive studies of Sites and Services in Dar es Salaam (Kironde 1991 and 1992 and Owens 2012). Other work on Dar es Salaam studies different interventions, including the short-term impact of more recent slum upgrading projects on health, schooling, and income (Coville and Su 2014); descriptive analyses of a more recent episode of serviced plot provision, known as the "20,000 plots" project, which suggests sizeable short-run gains in land values (Tiba et al. 2005 and Kironde 2015); and willingness to pay for land titling in poor neighborhoods (Ali et al. 2016 and Manara and Regan 2019). But as far as we are aware, ours is the first long run econometric evaluation of de novo Sites and Services areas.

Our study is related to research on the role of coordinating land institutions (Libecap and Lueck 2011) - in our case formal plots - in underpinning economic development. It is also related to studies of housing externalities in cities (Hornbeck and Keniston 2017 and Rossi-Hansberg et al. 2010). Another recent and related paper - on Indonesia rather than Tanzania - is Harari and Wong (2017). They, like us, find that upgraded slums do not perform well economically in the long run. Our paper, however, differs from theirs since we focus on de novo neighborhoods, which are not part of the context they study.

⁵As we discuss below, a few years after Sites and Services were implemented, most of the residents in de novo neighborhoods in Dar es Salaam were still those targeted by the policy, many of whom were poor.

Our paper is also related to the literature on the economics of African cities (Freire et al. 2014). Like Gollin et al. (2016) we study not only the largest African cities (such as Dar es Salaam in Tanzania), but also secondary cities. Our contribution to this literature comes from studying these cities at a fine spatial scale, examining individual neighborhoods and buildings, using a combination of very high resolution daylight satellite images, building-level survey data, and precisely georeferenced census data.

A few recent papers study outcomes not only across African cities but within them (see for example Henderson et al. 2016). Our study differs not only in our focus on secondary African cities, but also in the longer time horizon we cover. We use historical satellite images and highly detailed maps going back over 50 years, which allow us to evaluate long run changes on historically undeveloped land in response to specific infrastructure investments. By combining these with data on individuals, we also provide more evidence about the sorting across neighborhoods.

Also related to our paper is a broader literature on the economics of slums (e.g. Castells-Quintana 2017 and Marx et al. 2019). Our contribution to this literature is to illustrate conditions under which housing of better quality forms and persists, and the limitations of upgrading existing slums. Poor neighborhoods have also been studied in other settings, especially in Latin America and South Asia. For example, Field (2005) and Galiani and Schargrodsky (2010) find that providing more secure property rights to slum dwellers in Latin America increases their investments in residential quality.⁶ Our paper differs in its setting (Tanzania is considerably poorer than Latin America) and its focus on early infrastructure provision.

While our paper's focus is on new neighborhoods rather than new cities, it is also related to Romer (2010), who investigates the potential for new Charter Cities as pathways for urban development in poor countries. Our work is also related to the position advocated by Shlomo Angel, that Sites and Services may be a relevant model for residential development in some circumstances.⁷

Methodologically, we contribute to the nascent literature using very high resolution daylight images (e.g. Jean et al. 2016). Like Marx et al. (2019) we study roof quality as a measure of residential quality. Our measure of quality differs, however; instead of measuring luminosity, we assess whether roofs are painted, since paint protects the roofs from rust.

⁶In another paper, Galiani et al. (2013) study an intervention that provides pre-fabricated homes costing around US\$1,000 each in Latin America, but come without any infrastructure.

⁷See for example this interview with Angel, which discusses this idea:

<http://www.smartcitiesdive.com/ex/sustainablecitiescollective/conversation-dr-shlomo-angel/216636/>

We also use the imagery data to develop a set of measures of residential quality, including building size, access to roads, and a measure of regularity of neighborhood layout, which we combine with survey data on building quality.

The remainder of our paper is organized as follows. Section 2 discusses the institutional background and data we use; Section 3 presents the research design and our empirical findings; Section 4 contains a model of investments in infrastructure and housing in different neighborhoods; and Section 5 concludes.

4.2 Institutional background and data

4.2.1 Institutional background

What were Sites and Services projects?

This paper studies the long term consequences of ambitious projects that were designed to improve the quality of residential neighborhoods in Tanzania. These projects, called “Sites and Services”, formed an important part of the World Bank’s urban development strategy during the 1970s and 1980s. Sites and Services projects were implemented not only in Tanzania, but also in other countries such as Senegal, Jamaica, Zambia, El Salvador, Peru, Thailand, and Brazil (Cohen et al. 1983). Of the World Bank’s total Shelter Lending of \$4.4 billion (2001 US\$) from 1972-1986, Sites and Services accounted for almost 50 percent, and separate slum upgrading accounted for over 20 percent.

In Tanzania, Sites and Services were implemented in two rounds – the first began in the 1970s (World Bank 1974b and 1984) and the second in the 1980s (World Bank 1977b and 1987). These projects were co-financed by the World Bank and the Tanzanian government (World Bank 1974a and 1977a).

Sites and Services projects in Tanzania fell into two broad classes. The first involved de novo development of previously unpopulated areas. The second involved upgrading of pre-existing squatter settlements (sometimes referred to as “slum upgrading”). In total across both rounds, the program laid the groundwork for 12 de novo neighborhoods and 12 upgrading neighborhoods spread across seven cities (World Bank 1974b, 1977b, 1984, and 1987).

The overall cost of the Sites and Services projects in Tanzania was approximately \$130 million (in US\$2017), of which \$83 million were direct costs, covering for infrastructure, land

compensation, equipment and consultancy (World Bank 1974b, 1977b, 1984 and 1987).⁸ The direct costs per square meter in the first round in de novo (\$2.20) and upgrading (\$2.37) were similar (World Bank 1974b, 1977b, 1984 and 1987). To compare these costs to present-day land values (see below), we focus on costs per square meter of plot, excluding public areas. As we explain in the Data Appendix, we estimate that the direct costs per square meter of plot were no more than \$8, and the total costs were no more than \$13 per square meter.

What were the treatment and counterfactual?

Our main empirical analysis compares de novo (our main treatment) to nearby control areas (our counterfactual). As we explain in Section 3, we implement a spatial regression discontinuity design, focusing on the difference in outcomes close to the boundary of de novo areas and adjacent control areas, which were (like de novo) unbuilt before the Sites and Services projects began. In Section 3 we also discuss and address potential threats to our identification strategy. Here we explain why we focus on the comparison between de novo and control areas and what we learn from it.

De novo areas received roads, which were mostly unpaved, and water mains, as well as formal plots.⁹ The combination of these three infrastructure elements (formal plots, roads, and water mains) constitutes the main treatment for de novo areas.¹⁰ Roads reduce travel costs for both work and leisure for residents, customers and visitors. Water mains may improve the quality and reliability of water consumed, and reduce the transaction costs of purchasing water (e.g. from water trucks). They may also improve the residents' health and help them grow food. Formal plots reduce the risk of full expropriation, and of infringements onto parts of owners' plots and public spaces (such as roads and areas required to maintain water mains). They may also reduce conflicts over ownership, and the need to engage in costly defensive actions (such as building fences or walls). Moreover, the formal and regular plots may mitigate coordination problems, lead to easier access and better use of space, and make plots more easily tradeable, increasing the incentives to

⁸The remainder of the costs covered a loan scheme and community buildings.

⁹Formal plots are delineations of land, which meet local surveying and town planning standards. They increase tenure security, and are a prerequisite in any application for a Certificate of Right of Occupancy (the highest land tenure document in Tanzania).

¹⁰Upgrading areas also received roads and water mains, but no formal plots. The Data Appendix contains more information about the precise timing and more details the investments cost breakdown. The second round investments were generally lower - in some cases they may have excluded water mains, and for one of the de novo areas (the one in Tanga), we have some uncertainty as to the extent of infrastructure that was actually provided (World Bank 1987). Most of the de novo plots were, however, laid out in the first round.

invest in them. Long-term gains in land values from a regular grid of plots (compared to a decentralized and irregular system) have been documented in the US context (Libecap and Lueck 2011).¹¹

In addition to the main treatment components, both de novo and upgrading areas received a small number of public buildings, which were designated as schools, health clinics, and markets.¹² While these could have had an impact, we think that they matter less than the plots, the roads and the water mains. First, the total cost of the public buildings was lower than either the roads or the water mains; and second, even if Sites and Services areas received more buildings than other areas (which we don't know), there is no evidence that access to them ends discontinuously at the project boundaries. In addition to the infrastructure investments, some Sites and Services residents were offered loans, which were not fully repaid. We think of these loans as relaxing some owners' budget constraints, and below we explain our strategy for studying the implications of differences across neighborhoods in owners' credit constraints.

As we discuss in more detail in the Data Appendix, control areas appear to have received significantly less infrastructure investments, although our data do suggest that they have some roads and connections to water mains.

For upgrading areas we do not have a clean counterfactual, because those areas were built on by squatters before Sites and Services began. Thus any present-day differences between them and other areas may reflect a combination of preexisting differences and the treatment effect of upgrading. In Section 3 we explain what we can nevertheless learn about those areas, at least descriptively, by comparing them to nearby areas or to other preexisting squatted areas that were not treated by Sites and Services.

How were treatment areas selected?

While our regression discontinuity design helps to mitigate concerns about selection of areas, it is nonetheless important to explain how the locations of the treatment and control areas were selected. For de novo neighborhoods, the planners intended to purchase mostly empty (greenfield) land parcels measuring at least 50 hectares each, although in practice this criterion appears to have been met only for seven of the twelve de novo areas. The planners also sought land suitable for construction (e.g. with natural drainage) with access to off-site water mains, trunk roads, and employment opportunities. For upgrading the

¹¹Hornbeck and Keniston (2017) similarly emphasize that starting afresh can lead to higher local land values in an urban setting.

¹²The first round buildings public buildings were also surrounded by street lighting

planners looked for squatter settlements that were large, well-defined, hazard-safe, and suitable for infrastructure investments (World Bank 1974a, 1977b). In most but not all cities, de novo and upgrading areas were adjacent to each other. We discuss our selection of the control areas in more detail below. All the areas are depicted in Figure A1.

Who took part in the program?

Another important aspect of the Sites and Services projects was the characteristics of the population they targeted. The planners had intended for the plots to be allocated following a point system, which prioritized applicants who met certain criteria. Different sources do not agree precisely on the criteria used, although it seems that a preference was given to the poor – but not the poorest – urban residents (World Bank 1974 and World Bank 1977). Laquian (1983) explains that the de novo projects in Tanzania were intended for income groups between the 20th and 60th income percentile of a country. In similar vein, Kironde (1991) argues that eligibility for de novo sites in Dar es Salaam excluded the poorest and richest households, but targeted an intermediate range of earners which covered over 60% of all urban households. It seems that the opportunity to purchase de novo plots was initially given to low income households, including those displaced from upgrading areas, presumably as a result of building new infrastructure (World Bank 1984 and Kironde 1991).¹³

There is some disagreement as to how this process was implemented in practice. One report (World Bank 1984) argues that there were irregularities in this process, which allowed some richer households to sort into de novo neighborhoods. But in discussing the de novo sites in Dar es Salaam in the late 1980s, Kironde (1991) argues that most plots were awarded to the targeted income groups, and as of the late 1980s: "The majority of the occupants (57.9 percent) are still the original inhabitants but there are many 'new' ones who were either given plots after the original awardees had failed to develop them, or who were given 'created' plots. A few, however, obtained plots through purchase or bequeathment". Taken together, the evidence suggests that de novo locations attracted some households with modest means, but gradually also richer ones. As our model below illustrates, this type of sorting would likely have occurred even if the project had been administered flawlessly.

¹³The planners had intended for the plots to be allocated following a point system, which prioritized applicants who met certain criteria. But different sources (e.g. World Bank 1974, World Bank 1977, and Kironde 1991) differ in their accounts of what these precise criteria were.

How relevant are Sites and Services today?

The difficulty of recouping Sites and Services costs, and criticism that they excluded the poorest urban population, appear to have motivated a shift away from them during the 1980s (World Bank 1987, Mayo and Gross 1987, and Buckley and Kalarickal 2006). As a result, the share of Sites and Services (including slum upgrading) in the World Bank's Shelter Lending fell from around 70% from 1972-1986 to around 15% from 1987-2005 (Buckley and Kalarickal 2006).

Nevertheless, Sites and Services projects deserve renewed attention for several reasons. First, as mentioned above, Africa's urban population is growing rapidly, and adding pressure to its congested cities. Second, Africa's GDP per capita has grown in recent decades, so more Africans can now afford better housing, and an important question is how to deliver this. Alternative solutions, such as government provision of public housing, are considerably more expensive than a *de novo* approach of the type we study.¹⁴ Third, cost recoupment and administration have since improved through increased use of digital record keeping, as evidenced by the Tanzanian Strategic Cities Project (TSCP - World Bank 2013).¹⁵ For example, the "20,000 Plots" project, a *de novo* program implemented in Tanzania in the early 2000s appears to have reduced the cost per plot by about half compared to the historical Sites and Services projects, even though the new plots were bigger (Tiba et al. 2005). Finally, land on the fringes of Tanzanian cities remains inexpensive (Tanzania Ministry of Lands 2012), so there are still opportunities for more *de novo* developments.¹⁶

To shed light on the motivations of urban planners in considering *de novo* projects, we turn to the above-mentioned "20,000 Plots" project. Among the concerns that lay in the background to this program were the ongoing expansion of unplanned squatter areas, which suffer from poor waste management, an inadequate supply of urban services and infrastructure, and transportation problems. These unplanned areas also hamper the government's ability to collect tax revenues (Tiba et al. 2005). It is in this context that the "20,000 Plots" project aimed to alleviate the shortage of surveyed and serviced plots and to reduce the rapid increase of informal settlements, as well as to restrict land speculation and cor-

¹⁴According to correspondence with Simon Franklin, from the experience of housing programs in cities such as Addis-Ababa, four room apartments (with a bathroom) in five-story buildings entail construction cost of around \$10,000, plus a further \$3,000-4,000 for infrastructure and administration. This figure excludes land costs.

¹⁵The TSCP was approved by the World Bank in May 2010 (see <http://projects.worldbank.org/P111153/tanzania-strategic-cities-project?lang=en>).

¹⁶Even cheap land on the city fringes is likely to have some residents, however, and ensuring that *de novo* programs treat them inclusively is an important issue, which we revisit in the conclusions.

ruption (Tiba et al. 2005). At the same time, distributional concerns regarding de novo projects remain relevant (Kironde 2015), and we revisit those in Section 5.

4.2.2 Data description

This section outlines how we construct the datasets that we use in our empirical analysis, leaving further details to the Data Appendix. First, we explain how we measure the treatment and control areas. Second, we explain our choice of units of analysis. Third, we explain how we construct the variables that we use in our analysis. Lastly, we discuss summary statistics for key outcomes.

How do we measure treatment and control areas?

For five of the seven Sites and Services cities (Dar es Salaam, Iringa, Tabora, Tanga, and Morogoro) we have maps showing the program area boundaries (World Bank 1974a,b, 1977a,b, 1984, 1987). For the two remaining cities we use information from local experts (for Mbeya) and other historical maps (for Mwanza), as we explain in the Data Appendix. Tables A1 and A2 list all 24 areas (12 de novo and 12 upgrading) with some information on the data we have on each.

Having defined the treated areas, we now explain how we construct our control areas. In much of our analysis, we use all initially unbuilt (greenfield) land within 500 meters of the boundary of de novo, as control areas.¹⁷ We exclude areas that were uninhabitable (e.g. off the coast), built up, or designated for non-residential use prior to the start of the Sites and Services projects. In order to infer what had been previously built up, we use historical maps and imagery collected as close as possible to the start of the Sites and Services project, and where possible before its start date, as discussed in the Data Appendix.¹⁸

To construct control areas for the upgrading areas we similarly use greenfield areas within 500 meters of upgrading; or alternatively 21 slums that were delineated in the 1979 Dar es Salaam Masterplan (Marshall, Macklin, Monaghan Ltd. 1979) and were not upgraded as part of Sites and Services. Comparisons across slums should be taken with caution, since in accordance with the planners' intention to target larger slums (see Section 2), the upgraded slums covered an average area about four times larger than the control slums. Both upgraded and non-upgraded slums, however, had similar initial population densities

¹⁷Note that throughout our paper the control areas always exclude de novo and upgrading areas.

¹⁸For some of the analysis we also study untreated areas further than 500 meters from the treatment areas, in which case we again excluded areas that were built up before Sites and Services began.

(195 people per hectare in the upgraded slums and 234 in non-upgraded slums in 1979). Figure A1 shows the de novo, upgrading, and control areas in all seven cities.¹⁹

Our empirical approach described below assumes that both the de novo and the control areas were unbuilt (greenfields) before the onset of Sites and Services. To provide evidence that this was indeed the case, we use a subsample of the TSCP survey data, which provides construction years for buildings in Mbeya and Mwanza (see Data Appendix). We report results from using these data cautiously, since they involve a fairly small sample and a variable (construction year), which is measured with noise, and only observed for surviving houses. With these caveats in mind, we note that only about 0.5 percent of the housing units in de novo areas and about 1.3 percent of the housing units in the nearby control areas were built before the start of Sites and Services, suggesting that the control and de novo areas were probably very sparsely populated.

How do we construct the units of analysis?

Our research design (discussed below) uses as its main units of analysis a grid of 50 x 50 meter "blocks", each of which is assigned to novo, upgrading, or control area depending on where its centroid falls. This creates a fine partition of our study area, which allows us to account for empty areas at the block level and within blocks. As we explain below, however, data constraints compel us to conduct some of the analysis at the level of individual housing units, or at the level of 2012 census enumeration areas (EAs) or subunits of EAs (Tanzania National Bureau of Statistics, 2014, 2017).

What are the key variables we measure?

To study the quality of housing across all 24 Sites and Services locations we use high resolution Worldview satellite images (DigitalGlobe 2016).²⁰ We employed a company (Ramani Geosystems) to trace out the building footprints from these data for six of the seven cities. For the final city, Dar es Salaam, we used separate building outlines from a freely available source – Dar Ramani Huria (2016). For all seven cities we then assembled more information on outcomes and control variables, as we explain in the Data Appendix. Here we describe some of the key variables.

For the purpose of measuring private housing quality using imagery data, we think of slums as typically containing small and irregularly laid out buildings, made of low quality

¹⁹To keep the maps on a fixed and legible scale, we do not show the locations of the non-upgraded slums in Dar es Salaam.

²⁰The images' resolution is 50 x 50 centimeters for greyscale, and a little coarser for color.

materials and with poor access to roads. We therefore define as positive outcomes the opposite of this image of slums: buildings with large footprints, which are regularly laid out, and have good roofs and access to roads. We use three outcomes which we think of as largely reflecting private complementary investments. First is the logarithm of building footprint size, derived directly from the processed imagery. Second, we use the color satellite imagery to assess whether each roof is likely painted, and therefore less prone to rust. Third, we calculate the orientation of each building using the main axis of the minimum bounding rectangle that contains it. We then calculate the difference in orientation between each building and its nearest neighboring building, modulo 90 degrees, with more similar orientations representing a more regular layout.²¹ Finally, we construct an indicator for buildings that are within no more than 10 meters from the nearest road. Unlike the three previous measures, we think of this measure of road access as largely representing persistence of Sites and Services infrastructure investments.

While the imagery and the outcomes we derive from it have the advantage of broad coverage, we complement them with detailed survey data on all the buildings in three of the Sites and Services cities, Mbeya (in southwest Tanzania), Tanga (in northeast Tanzania), and Mwanza (in northwest Tanzania). These data are derived from the TSCP survey, which was conducted from 2010-2013 (World Bank 2010). We use these data to build a more detailed picture of building quality in the areas we study. The TSCP data allow us to identify outbuildings (e.g. sheds, garages, and animal pens), which are generally smaller, and which we exclude from the analysis.²² This leaves us with a sample of buildings that are used mostly for residential purposes, although a small fraction may also serve commercial or public uses.

We use the TSCP survey data to measure the logarithm of building footprint, and create indicators for buildings which have more than one story, good (durable) roof materials, connection to electricity, and at least basic sanitation.²³ These measures likely reflect private investments, since they were not part of the Sites and Services investments. In addition, we measure connection to water mains and having road access as largely reflecting

²¹When we regress the log hedonic price index (discussed below) on the three imagery measures using a block-level regression, the coefficients on each of the three measures is positive and significant. This provides further support for our use of these measures of housing quality. Where applicable we standardize and pool the three quality measures together to construct a "family of outcomes" z-index (Kling et al. 2007; Banerjee et al. 2015).

²²Outbuildings account for around 10-30% of buildings in the areas we consider, where the fraction varies by city. Their mean size is typically around one third that of the average regular building size.

²³In the de novo, upgrading, and control areas we classify as "basic sanitation" having either a septic tank (30% of buildings) or sewerage connection (0.5% of buildings). Not having basic sanitation usually means a pit latrine (67% of buildings) or "other" or none. As before, we construct a "family of outcomes" measure based on non-missing observations for each variable.

persistence of Sites and Services investments. The TSCP data also provide the full names of owners of housing units, which we use as explained below.

We also use separate TSCP (World Bank 2013) valuation data for Arusha, a city where Sites and Services were not implemented, to construct a hedonic measure of building quality, as we explain in the Data Appendix.²⁴ Another separate data source (Tanzanian Ministry of Lands 2012) provides us information about land values in Dar es Salaam, although at a coarser level.

In addition to these variables we construct geographic variables (distance to the nearest shore; an indicator for rivers or streams; and a measure of ruggedness), and other variables, which we use in our analysis below. All these are again explained in the Data Appendix.

We complement all these measures of the physical environment, with some data on people, including indicators for owners (identified by their full name and the city), taken from the TSCP survey, and population density and measures of schooling and literacy, which we calculate from the 2002 and 2012 censuses at the level of enumeration areas. We sometimes split enumeration areas to allocate them across treatment and control areas, as we explain in the Data Appendix.

How do the different areas compare using raw data?

Table A3 summarizes information on the number of plots and the population density, as of 2002, in de novo and upgrading areas, and their respective control areas. As the table shows, de novo areas were more densely populated than nearby control areas. Upgrading areas were very densely populated, and again denser than control areas near them. As we shall see below, the higher density in upgrading areas did not correspond to more multistory buildings, but in fact the opposite.

Figure A2 shows visual examples of parts of a de novo area, a control area near de novo, and an upgrading area, all in the same district of Dar es Salaam. The differences between the most orderly location (de novo) area and the least orderly one (upgrading) are visibly clear, and the control area lies somewhere in between.

The impression that de novo areas have higher quality housing is corroborated in the summary statistics table (Table A4). The imagery data shows that compared to the control areas, de novo areas have buildings with larger footprints, a higher fraction of painted roofs,

²⁴Our approach of using characteristics linearly in a hedonic regression follows Giglio et al. (2014). There is also some evidence that in the case of housing, using the imputed hedonic values as dependent variables does not lead to much bias in the inference (McMillen et al. 2010 and Diewert et al. 2015).

more regularly laid out buildings, and better access to roads. The survey data shows that de novo areas are also more likely to have multiple stories, good roof materials, connection to electricity, basic sanitation, and connection to water mains, as well as a much higher hedonic value. On almost all these measures, including the fraction of buildings with multiple stories, upgrading areas look worse, and control areas are somewhere in between de novo and upgrading. The log hedonic price differences suggest that on average, de novo housing units are about 63 percent more valuable than those in control areas and about 92 percent more valuable than those in upgrading areas.

4.3 Research design and empirical findings

4.3.1 Research design

The differences in outcomes described in Table A4 suggest that housing quality in de novo areas is considerably better than in control areas. The higher quality of housing in de novo areas reflects both elements that Sites and Services invested in directly, such as roads and water, and elements that it did not, such as electricity. But in order to study whether the de novo investments did in fact crowd in private investments (and if so - how much), we need to move beyond the descriptive statistics, as this section explains.

Our identification strategy compares de novo areas to nearby control areas, which (like de novo areas) were largely empty before the onset of Sites and Services. In our main analysis, we follow Gelman and Imbens (2017), by implementing a semi-parametric regression discontinuity design:

$$y_i = \beta_0 + \beta_1 Denovo_i + \beta_2 Dist_i + \beta_3 Dist_i \times Denovo_i + \beta_4 \mathbf{Nearest_Denovo}_i + \beta_5 Dist_CBD_i + \beta_6 \mathbf{Controls}_i + \epsilon_i, \quad (4.1)$$

where y_i measure various outcomes, as described in Section 2 and the Data Appendix; $Denovo_i$ is the main regressor of interest, which indicates whether the centroid of i is in de novo areas, where control areas are the omitted category; $Dist_i$ is the distance in kilometers to the boundary between de novo and control areas; $\mathbf{Nearest_Denovo}_i$ is a vector of fixed effects for the nearest de novo areas; $Dist_CBD_i$ measures the distance in kilometers of unit i from the Central Business District (CBD) of the city in which it is located; $\mathbf{Controls}_i$ is a vector of additional controls, which we discuss below; and ϵ_i denotes the error term. The role of distance to the central business district is emphasized in many urban economics models (see Duranton and Puga 2015 for an overview), and

adding Nearest_Denovo_i ensures that we only compare control areas to their nearest de novo area. In our baseline specification, each observation is a 50 x 50 meter block, but later on, as we explain, we also use housing units within buildings and enumeration areas as units of analysis.

Our baseline analysis uses data from within 500 meters of the boundary between de novo and control areas. Using this fixed distance allows us to analyze all our outcomes across imagery and TSCP survey data (World Bank 2013) consistently. As we discuss further below, 500 meters also turns out to be fairly close to the optimal bandwidth we find for our key outcomes using the survey data. Finally, we also present below alternative specifications using more - and less - data.

In our baseline estimates we cluster the standard errors on 850 x 850 meter blocks, following the approach of Bester et al. (2011) and Bleakly and Lin (2012). The size of the blocks on which we cluster reflects the size of the Sites and Services neighborhoods. The median size of the 12 de novo neighborhoods was approximately 0.538 square kilometers, and the median size of all 24 neighborhoods was around 0.718 square kilometers. This last figure is just a little smaller than the area of a square whose sides are 850 meters, which we chose as a conservative benchmark for clustering.²⁵

Addressing threats to identification

Our identification strategy assumes that conditional on the controls in specifications (4.1), the potential expected outcome functions are continuous at the discontinuity threshold. Our spatial regression discontinuity approach is similar to Dell (2010), and much of our analysis likewise applies a semi-parametric RD, which combines both controls, as in equation (4.1), and a focus on areas that are close to (within 500 meters of) the boundary of de novo and control areas.²⁶

One potential concern is that the areas selected for de novo differed in their "first nature" location fundamentals. But in our setting the geographic distances are much smaller than in most other settings, so we are less concerned with larger scale changes in geography, such as

²⁵In earlier versions of this paper we also reported specifications using Conley (1999) standard errors with a decay area equal to the size of the above-mentioned blocks, and the results were similar. To mitigate concerns about the variation in neighborhood size, we also experimented with modifying our baseline clustering blocks to treat each Sites and Services neighborhood as a separate clustering unit, with the remainder of the cluster units based on the grid (cut where necessary by the Sites and Services neighborhoods). Once again the estimated standard errors were quite similar.

²⁶Since we have variation within several cities, we use functions of distance to the de novo boundary in our main specification, and functions of longitude and latitude only in our robustness checks, as we discuss below.

climate or soil fertility. Further, in our empirical analysis, we report balancing tests, which use specification (4.1) to compare the geographic variables as outcomes as we cross the de novo - control boundary. Some of the geographic variables - the land's ruggedness and the presence of rivers or streams - may be endogenous to housing development. Therefore below we report estimates both with and without the geographic controls.

Our identification strategy also assumes that both de novo and control areas were essentially empty (greenfields) before the start of Sites and Services. As we discuss in the Appendix, our classification of areas relies on historical aerial images and topographic maps, which allow us to detect pre-existing buildings. And in Section 2 we provide support for this assumption using a subsample of buildings for which we have construction dates.

Another relevant question is whether administrative boundaries correspond to some of the de novo - control boundaries, leading to different municipal policies on either side of the boundary. To address this question, we verified that in none of the cases do the boundaries between any treatment areas and the control areas coincide with the ward or district boundaries.²⁷

A different type of concern is that there may be spillovers across neighborhoods.²⁸ So, for example, it is possible that proximity to de novo areas improves nearby control areas, or that proximity to control areas worsens de novo areas; both would attenuate our estimates. To mitigate this concern we report "doughnut RD" specifications, which exclude bands of 100 meters around the boundary between de novo and control areas. To mitigate a related concern that upgrading areas may be affecting our estimates, we also report specifications, which exclude all blocks within 100 meters of upgrading areas. In a similar vein, since the TSCP data, but not the imagery data, cover entire cities, we also report specifications that use wider control areas, rather than only those near de novo areas. In those cases we use the same specification as in the baseline, but also report some results using second- and third-order polynomials in distance to the boundary.²⁹

A related concern is that Sites and Services may have reshaped cities, and even affected the

²⁷The closest case is Mwanza in 2012, where one district (Nyamangana) cuts into less than a quarter of the control area, while another (Ilemela) contains all of the treatment and most of the control area. However, this boundary was only observed in the 2012 census and not in the 2002 census, so it is almost certainly either unrelated to the Sites and Services project, or an indirect outcome of it. In the 2002 census, Ilemela district fully contained the Mwanza treatment and control areas.

²⁸See related discussions in (Turner et al. 2014), Hornbeck and Keniston (2017) and Redding and Sturm (2016).

²⁹The full city data also allow us to estimate regressions using an optimal bandwidth (Imbens and Kalyanaraman 2012), which we also report.

location of their CBD, and the distance to it. To address this concern we report robustness checks, which use distances to historically central locations - mostly railway stations, as discussed in Section 2 and the Data Appendix.

Did Sites and Services create or displace value?

Another question that we consider is whether Sites and Services created value or merely displaced it. Like many studies of place-based policies, it is difficult for us to answer this question definitively, since we do not have counterfactual cities of similar size, which were untreated by Sites and Services. And even if such cities had existed, one might still have worried about displacement of activity across cities. Nevertheless, our findings below suggest that de novo areas are relatively regularly laid out, and preserve good access to roads. It therefore seems likely that by solving coordination failures they created value and not merely displaced it.³⁰

Exploring mechanisms: sorting across neighborhoods and infrastructure persistence

Our setting allows us to explore another important issue - the role of sorting of owners across neighborhoods. As we discuss above, initial ownership criteria in de novo areas excluded the poorest, and program loans may have further alleviated credit constraints for some of these owners (as well as for some of the owners in upgrading areas). The model characterizes sufficient conditions under which including owner fixed effects overcomes the potential differences in credit constraints of owners who rent out multiple housing units.³¹ We note that renting is fairly common in our setting: as of 2007, renters accounted for a small majority of Dar es Salaam's residents, and over a third of the residents in other urban areas; back in 1992, the share of renters was even higher (Komu 2013).

To shed light on the sorting across neighborhoods of residents, we also use census data to characterize residents by measures of education, which are the best proxies we have for lifetime earnings.

Our model also highlights the role of persistently better infrastructure in de novo neigh-

³⁰As we also discuss below, our findings suggest that Sites and Services not only had positive effects on local land values, they may also have generated positive spillovers on nearby areas, an issue that we revisit in Section 3.

³¹To be precise, we consider a full name as different if it appears in more than one city. In practice this does not seem to make much difference. Since this strategy uses variation within owners, it only employs part of the data, so in this case we need to use control areas from the rest of the city to ensure sufficient variation. We also acknowledge that some units may be owner-occupied, while others may be rented out, but we cannot separate the two with our data.

borhoods as a mechanism for crowding-in investments in housing quality. Empirically, we estimate regressions of the same form of as equation (4.1), using measures of water connection and access to roads as outcomes, since these closely relate to the investments made in the Sites and Services projects.

Studying upgrading areas

Finally, we repeat our analysis for upgrading areas, comparing them to proximate control areas, following the procedure outlined above.³² Finding appropriate counterfactuals for upgrading areas (which were populated before the program began) is harder than for de novo areas (which were essentially empty). To mitigate concerns about different starting conditions, we also report regressions that compare upgrading areas to 21 other slums that existed in Dar es Salaam in 1979, and which were not upgraded as part of Sites and Services. The slums that were not upgraded were on average smaller in area (see Section 2), but had similar, or even slightly higher, population density in 1979. The comparisons of upgrading areas to non-upgraded slums come with two caveats: first, this analysis is not a spatial RD, since the non-upgraded slums were not adjacent to the upgraded ones, although for consistency we still use specification (4.1); and these comparisons are only possible for the imagery data, since Dar es Salaam is not covered by the TSCP survey data.

4.3.2 Empirical findings

Balancing tests

We begin the discussion of our findings by reporting balancing tests on geographic characteristics. As Table A5 shows, when we compare geographic characteristics in de novo areas to nearby control areas, both distance to the shore and ruggedness differ in de novo areas (Panel A), but after including our baseline controls as in equation (4.1) (Panel B) de novo and control areas look balanced. We also report balancing tests using TSCP data, which also look balanced (with the exception of rivers and streams in the sample adjacent to the de novo areas). We note, however, that rivers and ruggedness may be endogenous to the de novo development, which may have flattened the soil and buried or diverted some streams. For completeness we report below estimates both with and without the geographic controls.

³²In upgrading area regressions we measure distance to the upgrading (instead of de novo) - control boundary, and fixed effects for the nearest upgrading (instead of de novo) area.

Crowding in of private investments

We now turn to our main results. In Table 1 we report estimates using specification (4.1) and our imagery sample. Panel A shows that de novo areas have footprints that are roughly 12 percent larger and have more regular layout, but their roof quality is not better. The z-index aggregating all three measures indicates that de novo areas have higher quality housing than nearby areas, and other estimates show that they have fewer empty blocks and a higher fraction of their area is built up.³³ Panel B reports robustness checks for the z-index using geographic controls, longitude and latitude polynomials, an alternative measure of CBDs that predates Sites and Services, and excluding blocks near upgrading areas - all are similar to our baseline estimate. When we use doughnut RD specifications to exclude areas near the boundary of de novo and control the estimates increase somewhat, suggesting that our baseline estimates may be a little attenuated due to spillovers (positive ones from de novo to controls, or negative ones from controls to de novo, or both). This finding also suggests that the higher quality housing in de novo areas may generate positive spillovers on neighboring areas (see Hornbeck and Keniston (2017) and Turner et al. (2014) for related discussions of local spillovers).

In sum, results for all seven cities using the satellite image data suggest that de novo areas have larger and more regularly oriented buildings. To get a more detailed picture of the differences in residential quality we turn to the TSCP survey data for Mbeya, Mwanza, and Tanga. In Panel A of Table 2 we report results again using specification (4.1). One advantage of the survey data is that unlike the imagery data they allow us to focus on residential buildings by excluding outbuildings, which we do. As Panel A shows, buildings in de novo areas have footprints that are about 50 percent (or 0.41 log points) larger than the control areas. They are also about 23 percentage points (or 48 percent) more likely to be connected to electricity. The regressions also show economically large but statistically imprecise differences in favor of de novo areas in the share of buildings with multiple stories and with at least basic sanitation, but again almost no difference in roof quality.

We aggregate the measures of quality in the survey data in two ways: first using a z-index, and second using the predicted log hedonic value. Regressions using either as an outcome indicate significantly higher residential quality in de novo areas than in control areas.³⁴ Specifically, the regressions suggest that the hedonic price is around 56 percent. This may understate the actual differences in house values, since the hedonics do not directly account

³³To visualize our results, Panel A of Figure A3 shows a regression discontinuity plot of binned values of the z-index.

³⁴Panels B and C of Figure A3 show regression discontinuity plots for the Z-index and log hedonic prices.

for all housing characteristics, nor for the full impact of local neighborhoods' infrastructure. Noting this caveat, a result from the model in section 4.4 below suggests that land value differences in de novo (compared to control areas) are about 50 percent larger than house price differences. This result, combined with our hedonic estimates, suggests that land values in de novo areas are at least 86 percent higher than in control areas. To interpret this difference, we note that in Dar es Salaam, land values in de novo neighborhoods is in the range of \$160-220 per square meter (in 2017 prices).³⁵ Combined with our estimates above, this suggest that de novo may have increased local land values by at least \$75-100 per square meter.

These values are high compared to the cost of investments per unit of treated plot area which we estimate above to be no more than \$8 per square meter of plot area, or no more than \$13 per square meter if we include indirect costs (in US\$2017). While these estimates should be interpreted with caution, they suggest that the gains from de novo investments were large, at least in Dar es Salaam. That said, we acknowledge that the gains in other cities, where prices are lower, may not be quite as high.³⁶

In Panel B of Table 2 we report results from a series of robustness checks, focusing for brevity on the z-index and the log hedonic price. The estimates with geographic controls in column (1) are a little lower than the baseline; this could be either because the baseline regressions overstate the difference due to better geographic fundamentals in de novo location, or that the geographic controls are themselves outcomes and adding them understates the impact of de novo. Columns (2) and (3) show that controlling for the polynomial of longitude and latitude or using distance to historical (instead of contemporary) CBDs makes little difference compared to Panel A. The doughnut specification in column (4) is larger than the baseline, suggesting (as in Table 1) that the baseline estimates may be too small due to positive spillovers from de novo to controls (or negative ones going the other way). Column (5) excludes blocks near upgrading areas, and the results are similar to the baseline. Columns (6) uses control areas from the rest of the city, and the estimates are again larger, possibly because we are comparing de novo areas to a control group that is on average further away, and less affected by local spillovers.³⁷ Finally, column (7) uses an optimal bandwidth, following Imbens and Kalyanaraman (2012), and the estimates is again quite similar to the baseline.

³⁵The coarse data we have on land values do not separately identify the control areas near de novo.

³⁶Unfortunately, our land value data for other cities are either missing or not detailed enough to give a credible picture.

³⁷The estimates are robust to using second- and third order polynomials, although in the latter case they are smaller.

The results using hedonic values as outcomes in Panel C follow a similar pattern, where adding geographic controls reduces the estimate a little, and excluding areas near the boundary increases them a little. The main message, however, is that our baseline estimates are robust to using different specifications.

The role of sorting

The results discussed so far are silent on the respective role of the de novo treatment and the endogenous sorting across neighborhoods of owners with different levels of credit constraints. As our model below (in Section 4) shows, we can account for differences across areas in owners' credit constraints by adding owner fixed effects, which allow us to isolate the impact of de novo areas compared to control areas for owners with multiple housing units. The units of analysis used in these regressions are individual housing units, since this is the level at which ownership is defined. The housing units we focus on are those owned by owners of multiple units, which account for about 13 percent of all housing units. To ensure a sufficiently large sample, we reestimate specifications as in (4.1) for the full city TSCP sample, but now focusing on housing units whose owners have more than one unit. Table 3 reports estimates of these regressions with owner fixed effects (Panel A) and without them (Panel B). The estimates show that in this sample, housing units in de novo areas are considerably larger, and much more likely to have electricity and basic sanitation. Without owner fixed effects they also are more likely to be in multistory buildings, although this difference vanishes once we control for owner fixed effects. As reported previously, de novo housing units do not have better roof materials. The difference in quality between de novo and control areas, as reflected in the z-index and the hedonic value, suggests that de novo areas may be about 60 log points (or about 83 percent) more valuable; as discussed above, this may understate the actual differences since it is unlikely to reflect all the amenity differences. Panels C and D of the table report robustness checks for the specifications with and without fixed effects, using the z-index as an outcome. Across a range of specifications reported in Table 3, roughly a third of the quality advantage of de novo areas is accounted for by the different ownership, and the rest likely reflects the impact of de novo on quality for owners who are relatively unconstrained in terms of investment.³⁸

The characteristics of residents in de novo areas, compared to control areas, likely reflect their willingness to pay for higher quality housing. In Table A6 we report regressions using 2012 census data with "cut" enumeration areas as units of analysis (see Section 2 and Data

³⁸When we use the hedonic measure as an outcome, the regressions estimates with and without owner fixed effects are more similar to each other (results available on request).

Appendix for details).³⁹ Consistent with the results discussed above, residents in de novo areas are better educated and more likely to be literate in English. The higher schooling of de novo residents is consistent with sorting across neighborhoods and a higher willingness of the more educated to pay for better housing quality, although it is also possible that some of it is the result of better access to schooling of existing residents. Still, as Table A6 shows, only about 55 percent of adults in de novo areas had more than primary school education, so the other 45 percent had no more than primary school education. This means that many less educated Tanzanians are still benefitting from de novo amenities.

The persistence infrastructure

To conclude our empirical analysis of the de novo areas, we explore whether their better housing quality corresponds to persistently better infrastructure. Here we focus on two of the main investments in Sites and Services, roads and water mains, and we again use specification (4.1). As Panel A of Table 4 shows, across both our imagery and TSCP data, de novo areas enjoy better access to roads, and the TSCP data also show that they are more likely to be connected to water mains.⁴⁰ Panels B-D report robustness checks using the same specifications as in Table 2. Again the estimates are a little smaller when we control for geographic covariates, and a little larger when we focus on control areas that are further from de novo, with our main estimates in between. And all the estimates are positive and statistically significant, showing that de novo investments translated into better infrastructure in the long run.

Upgrading areas

Having discussed the de novo areas, we now briefly discuss what we can learn from similar regressions for upgrading areas. As Table A7 suggests, upgrading areas look fairly similar to nearby control areas in terms of the geographic controls, except that in most specifications they are less likely to have rivers or streams. When compared to the non-upgraded slums, and conditional on our baseline controls, the upgrading areas are closer to the shore but not significantly different in the other two geographic controls (results available on request).

Table A8 reports estimates using imagery data for all seven cities. Panel A suggests that housing quality in upgrading areas is similar to that of nearby control areas. The

³⁹In this case number of units of analysis is small and they are uneven (some are whole EAs and some are cut), which makes it difficult to get a good measure of distance to the boundary. Therefore in these specifications we use non-parametric regression discontinuity, without controls for distance to the boundary.

⁴⁰This last result is robust to excluding Tanga, where we have some uncertainty about the nature of de novo investments.

only significant differences are that upgrading areas have fewer empty areas and are more densely built up. Panel B shows that this conclusion is robust to a range of different specifications.

In Panels C and D we compare upgrading areas in Dar es Salaam only to the preexisting ("old") slums that were not upgraded as part of Sites and Services. Once again the results suggest that upgrading areas are no different, except perhaps in a slightly more regular orientation of buildings than their control areas. Upgrading areas also seem to have fewer empty blocks and a larger fraction of built up area.

Next, in Table A9, we use TSCP survey data outcomes. Here the upgrading areas look somewhat worse than nearby control areas: they have fewer multistory buildings, worse roofs, and possibly worse sanitation, and their overall quality seems lower. This conclusion is reinforced in most of the robustness checks in Panels B and C, although not all the estimates are precise.

In Table A10 we examine the role of ownership in accounting for the worse quality in upgrading areas. The results suggest that ownership differences may partially explain the worse housing quality in upgrading areas, since controlling for owner fixed effects results in estimates that are small and in most cases imprecise.

A comparison of infrastructure persistence measures in upgrading areas may also help to explain why their housing is no better than that of nearby control areas. As Table A11 shows, upgrading areas look similar to nearby areas in their access to roads and water; the coefficients on upgrading areas are small, imprecise, and mostly negative. Adding the coefficients and the control means and comparing them to the estimates in de novo areas (Table 4) suggest that upgrading areas have worse infrastructure than de novo areas. As we discussed in Section 2, upgrading areas did receive roads and water mains, and investments measured in dollars per square meter were similar to those of de novo areas. A likely explanation for the poor state of upgrading areas' infrastructure today is that those areas' infrastructure deteriorated more than that of de novo areas. Kironde (1994, page 464) and Theodory and Malipula (2012) discuss evidence that infrastructure did in fact deteriorate in upgrading slums in Dar es Salaam. Kironde (1994) mentions, for example, the deterioration of roadside drainage due to lack of maintenance; private construction on land that was intended for public use; and the degradation of water provision infrastructure.

Finally, Table A12 shows that residents of upgrading areas are less educated than those of nearby areas, consistent with the lower housing quality in these neighborhoods.

4.4 Model

4.4.1 Assumptions and their relationship to the institutional setting

To frame our empirical analysis we present a model, which characterizes conditions under which investment in infrastructure (as defined below) incentivizes owners to build higher quality housing. The model captures key aspects to our description of the Sites and Services projects in Section 2. It also connects to our econometric analysis in Section 3, by relating gains in house values to gains in land values, and motivating our use of owner fixed effects to account for owner sorting across neighborhoods.⁴¹

We consider a population of infinitely lived, profit maximizing owners, with formal or informal rights to build on their plot(s), which are organized into neighborhoods (areas). In each plot, the owner can build a house and rent it out.⁴² The model is in discrete time, and in each period $t \geq 1$, owners maximize their expected present discounted stream of rents, net of house construction costs, on each plot they own:

$$E \left[\sum_{s=t}^{\infty} \delta^s [r(q_s, I_s) - B_s c(q(I_s))] \right], \text{ s.t. } \Pr(q_{t+1} = q_t) = 1 - d, \Pr(q_{t+1} = 0) = d. \quad (4.2)$$

The expectations are defined over the exogenous destruction probability of houses in each period, as discussed below. Owners are assumed to have a time preference $\delta \in (0, 1)$. The rent that each owner receives on each house in each period is $r(q, I) = q^\alpha I^{1-\alpha}$, where q and I denote the quality of the house and the neighborhood infrastructure, and $\alpha \in (0, 1)$. B_t is an indicator equal to one if a house is built in period t and zero otherwise. The construction costs of a house of quality q are: $c(q) = cq^\gamma$, where $c > 0, \gamma > 1$. This convex cost function generalizes Hornbeck and Keniston (2017), who assume $\gamma = 2$. In a different context, Combes, Duranton and Gobillon (2016) finds that the production function for housing can be approximated by a constant returns to scale Cobb-Douglas function using land and other inputs, where the coefficient on non-land inputs is approximately 0.65. Holding land constant, this production function is consistent with a cost function $c(q) = cq^\gamma = cq^{1/0.65}$, or $\gamma \simeq 1.54$.

⁴¹Our model builds on Hornbeck and Keniston (2017), but differs from theirs in several ways. We add to the model infrastructure and variation across owners in credit constraints, and we derive new analytical results. We also model spillovers across houses differently, and for simplicity we exclude the exogenous time trends.

⁴²A "house" in the model denotes is a shorthand for a housing unit that we consider in the empirical analysis. Unlike Bayer et al. (2007), our model does not account for renter heterogeneity, because we have no data on the rents paid and have little information on the residents. Knowing more about renters would have allowed to build a better picture of the welfare gains from de novo areas.

Infrastructure captures a broad set of neighborhood characteristics, including formal and regularly laid out plots, which reduce coordination failures and protect owners' property rights; roads, which reduce the cost of travel and trade; and water mains, which contribute to living standards and health.⁴³ Infrastructure also reflects other neighborhood level effects.⁴⁴ For tractability, we consider three types of infrastructure: high quality (I_H), medium quality (I_M), and low quality (I_L), where $I_H > I_M > I_L > 0$. High quality describes the bundle that Sites and Services offered - mostly formal plots, roads, and water mains. We assume that high quality infrastructure deteriorates to medium quality unless the fraction of high quality housing is larger than a constant $\phi > 0$.⁴⁵ Medium quality infrastructure is basic and unmaintained (e.g. bumpy dirt roads). It may be either high quality infrastructure that has deteriorated or it may start out as medium quality. We assume that medium quality infrastructure does not deteriorate.⁴⁶ Low quality infrastructure corresponds to the level that prevails without any infrastructure investments in the neighborhood.

There are two types of owners in the model. Unconstrained owners may each own any finite number of plots and afford any level of investment in each plot, while Constrained owners may own no more than a single plot, and may afford to build at most low quality housing $q_L = q(I_L)$, as defined below.⁴⁷ Consistent with our setting, we assume that no single owner has a sufficiently large number of plots to exert market power or to solve coordination problems that arise from neighborhood-level externalities.⁴⁸

⁴³Property rights protection may reduce the risk of outright expropriation, as we discuss below, as well as the risk of partial expropriation, when part of an owner's plot is built without authorization, which we do not model explicitly.

⁴⁴In practice, other types of neighborhood effects may also matter. For example, the absence of proper sewerage may increase the risk of contagious diseases. Consistent with this, Jaupart et al. (2016) show that cholera outbreaks in Dar es Salaam were much more severe in slum areas with poor infrastructure. Another possibility is that neighborhoods with poor electrification and lighting (Painter and Farrington 1997) and high population density (Gollin et al. 2017) may attract crime. While we think that both of these channels could amplify the land value differentials between neighborhoods, we do not have the data to study them in our context.

⁴⁵High quality housing is $q_H = q(I_H)$, as defined below. The potential for infrastructure deterioration means that owners' housing quality can be indirectly affected by those of their neighbors, through the effect on infrastructure. This mechanism is different from the direct impact of neighbors' housing quality in Hornbeck and Keniston (2017).

⁴⁶Our assumption that medium infrastructure and deteriorated infrastructure are equal in quality is a simplifying assumption, motivated by our empirical finding that upgrading areas are no better than nearby control areas in terms of access to roads and water. Adding further parameters for deteriorated high quality and deteriorated medium quality infrastructure would not add much insight to the model.

⁴⁷The distinction between two types of owners allows us to analyze owner sorting in a simple way. The results would have been similar if we had assumed that constrained owners could build up to any quality that is strictly lower than $q(I_M)$, as defined below.

⁴⁸Our TSCP data indicate that only a small share of housing units are owned by those with more than a handful of plots. It is true that in principle a rich individual or a firm could buy up an entire neighborhood and internalize the externalities involved. But until recent years the Tanzanian government exerted strict control that prevented the concentration of neighborhood ownership.

We consider three types of areas (neighborhoods), each with a continuum of plots.⁴⁹ De novo areas start with empty plots ($q = 0$) and high quality infrastructure (I_H); control areas start with empty plots and medium quality infrastructure (I_M); and upgrading areas start out with low quality housing (q_L) and high quality infrastructure.⁵⁰ This reflects the situation at the time when Sites and Services was implemented.⁵¹

The initial fractions of unconstrained owners are: θ_D in de novo areas, θ_C in control areas, and θ_U in upgrading areas. We assume that (as in the real world) upgrading areas are targeted for their relatively poor population, so they have few unconstrained owners, and therefore $\theta_U < \phi$.

In every period, the following sequence of events takes place. First, each owner decides whether to build (or rebuild) a house on each plot they own.⁵² Second, if the neighborhood's housing quality is insufficiently high, infrastructure quality deteriorates, as we discuss below. Third, each owner collects the rent on each house they own. Finally, there is an exogenous probability $d > 0$ that each house is destroyed, resetting housing quality to zero.⁵³

We assume that the risk that houses are destroyed and the fraction of owners of each type in each neighborhood are common knowledge, as is the understanding that all unconstrained owners will build high quality housing if the share of unconstrained owners is at least ϕ . In Nash Equilibrium, each owner solves her maximization problem in each period, assuming that all other owners do the same.

⁴⁹Our model does not account for other types of neighborhoods, such as former colonial areas (which typically constitute a small and wealthy part of cities), nor do we consider movements between different neighborhoods within the city.

⁵⁰In Section 2 we discuss the investments that were made as part of the Sites and Services projects. These suggest that though the investment per total land area in de novo and upgrading were similar.

⁵¹We also note that while the control areas we use were by definition empty to begin with, other areas looked like control areas but had a stock of low quality housing by the time they received infrastructure I_M .

⁵²Following Hornbeck and Keniston (2017) and Henderson et al. (2017), we assume that owners cannot renovate incrementally, and that houses do not depreciate. The assumption that rebuilding a higher quality house requires a fresh start is particularly relevant for low quality housing that characterizes poorer neighborhoods in East African cities. It may be possible to make minor improvements to a house built of tin or mud walls. However, demolition and construction from scratch is required to make meaningful improvements such as adding brick walls, multiple stories, or plumbing. For simplicity, we maintain the assumption that no incremental improvement is possible. Relaxing this would reduce the benefit of early (de novo) investments.

⁵³If a house is destroyed, the owner retains their plot. Given the paucity of construction dates in our data, it is difficult to assess d . But Henderson et al. (2017) estimate it at 3.2 percent per year using data from Tanzania's neighbor, Kenya.

4.4.2 Solving the model

This section characterizes the optimal level of investment by owners, beginning with unconstrained owners and then by constrained owners.

Unconstrained owners maximize profits on each plot they own by solving the following Bellman equation:

$$V(q, I) = \text{Max} \begin{cases} r(q, I) + \delta E[V(q, I)] \\ r(q(I), I) + \delta E[V(q(I), I)] - c(q(I)), \end{cases} \quad (4.3)$$

where r is return on house (e.g. rent), $q \geq 0$ is the house quality; $I \geq 0$ is the infrastructure quality which is expected when rents are collected and from that point onward; $q(I)$ is the optimal house quality; and $c(q(I))$ is the cost of building a house of quality $q(I)$.⁵⁴

The infrastructure quality which is anticipated when rents are collected and from that point onward is equal to the existing level, except where infrastructure of quality I_H deteriorates to I_M . This deterioration happens when the fraction of high quality housing ($q_H = q(I_H)$, as described below) is strictly lower than ϕ .

The model reflects a tradeoff between keeping the current house quality q and improving it to $q(I)$. But if an unconstrained owner's house is exogenously destroyed it is always rebuilt at the optimal quality $q(I)$. Starting from an empty plot, the optimal house quality for an unconstrained owner anticipating infrastructure I at the time of rent collection is:

$$q(I) = \left[\frac{\alpha I^{1-\alpha}}{\gamma c(1-\delta+d\delta)} \right]^{\frac{1}{\gamma-\alpha}}. \quad (4.4)$$

The quality of housing is characterized by the following comparative statics. First, $\frac{\partial q(I)}{\partial \delta} > 0$, so more patient people invest more. Second, $\frac{\partial q(I)}{\partial d} < 0$, so a higher probability of house destruction leads to lower quality housing. And finally, $\frac{\partial q(I)}{\partial c} < 0$, so a higher construction cost reduces housing quality.

If an unconstrained owner starts with housing $q_1 \equiv q(I_1)$ but with infrastructure I_2 (where $I_2 > I_1$), they choose between two options.⁵⁵ They can replace their house with a higher

⁵⁴We could have included a probability $(1-\psi)$ that a plot is fully expropriated at the end of each period. If that were the case we would need to substitute $\psi\delta$ instead of δ throughout the analysis, but for simplicity we focus on the case without expropriation, namely $\psi = 1$. Higher patience may reflect, at least in part, a lower risk of expropriation. Collin et al. (2015) elicit owners' perceived expropriation risk in Temeke, an informal area close to the CBD of Dar es Salaam, which implies a risk of around 8% per year. Given the setting, this is likely an upper bound to the perceived expropriation risk in the locations we study.

⁵⁵We assume that if owners are indifferent they do not improve their houses.

quality house, in which case their expected payoff is equal to the expected value of an unbuilt a plot of land:

$$\pi(0, I_2) = \frac{q_2^\alpha I_2^{1-\alpha} - cdq_2^\gamma}{1 - \delta} - (1 - d) cq_2^\gamma, \quad (4.5)$$

where $\pi(q, I)$ is the maximized expected payoff from an existing house of quality q and infrastructure quality I . Alternatively, they can keep the current quality q_1 and only build a better house when their house needs rebuilding. In this case their expected payoff is:

$$\pi(q_1, I_2) = q_1^\alpha I_2^{1-\alpha} + \delta [(1 - d) \pi(q_1, I_2) + d\pi(0, I_2)]. \quad (4.6)$$

Solving this expression we get:

$$\pi(q_1, I_2) = \frac{q_1^\alpha I_2^{1-\alpha} + d\delta\pi(0, I_2)}{1 - \delta + d\delta}. \quad (4.7)$$

Proposition 4.1 *For each level of infrastructure $I_1 > 0$, there exists a unique value $I_1^{crit} = \left(\frac{\gamma}{\gamma-\alpha}\right)^{\frac{\gamma-\alpha}{\alpha(\alpha-1)}} I_1$, such that unconstrained owners starting with $q_1 = q(I_1)$ and infrastructure $I_2 = I_1^{crit}$ are indifferent between rebuilding and not rebuilding, and owners rebuild if and only if $I_2 > I_1^{crit}$.*

To obtain $I_2 = I_1^{crit}$, combine the condition $\pi(q_1, I_1^{crit}) = \pi(0, I_1^{crit})$ with (4.5) and (4.6), where housing quality $q_2 = q(I_2)$ comes from (4.4). To show that owners rebuild if and only if $I_2 > I_1^{crit}$, note that $\frac{\partial}{\partial I_2} (\pi_{I_2} - \pi_{q_1, I_2}) > 0$.

This result implies that unconstrained owners face what we refer to as an "inaction zone", $(I_1, I_1^{crit}]$. If infrastructure is upgraded from I_1 to a level in the inaction zone, owners will not improve their house right away, but only when it is exogenously destroyed. But if the infrastructure upgrade is to $I_2 > I_1^{crit}$, unconstrained owners will rebuild at a higher quality q_2 right away.

The investment problem for constrained owners is similar to that of unconstrained owners, except that the maximum quality they can build is q_L . As a result, in equilibrium they build q_L if their plot is empty, and otherwise they do not rebuild.

4.4.3 Neighborhood development

De novo areas

De novo areas begin empty with infrastructure (I_H). Constrained owners build q_L , so that the share of low quality houses is $1 - \theta_D$. If $\theta_D \geq \phi$ then there is no deterioration

in equilibrium, anticipated infrastructure is I_H , and the unconstrained owners build q_H . If $\theta_D < \phi$ then there is deterioration in equilibrium, anticipated infrastructure is I_M , and unconstrained owners build $q_M = q(I_M)$. In practice it seems that de novo areas' infrastructure is better than other areas' (Table 4), suggesting that at least some higher quality infrastructure survived.

Control areas

Control areas begin empty and with medium quality infrastructure (I_M). Unconstrained owners build housing quality q_M , while constrained owners build q_L .

As discussed above, Tanzanian cities also contained areas (which are not part of our main analysis), which are similar to control areas but had a stock of low quality housing by the time they received infrastructure I_M . In those areas the constrained owners keep q_L , while the unconstrained owners either build q_M right away (if $I_M > I_L^{crit}$) or otherwise build q_M only when their house is destroyed.

Upgrading areas

Upgrading areas begin with housing quality q_L and infrastructure I_H , and we consider four different cases. In the first case $I_L^{crit} > I_H$, so the upgrading is minimal and all owners initially keep q_L , and infrastructure deteriorates to I_M ; in later periods, as houses are exogenously destroyed, unconstrained owners build to q_M . In the second case $I_H \geq I_L^{crit} > I_M$ and $\theta_U < \phi$, in which case everyone initially keeps q_L , infrastructure deteriorates to I_M ; and unconstrained owners improve their houses to q_M when they are destroyed. In the third case $I_M \geq I_L^{crit}$ and $\theta_U < \phi$, in which case unconstrained owners build q_M right away while constrained owners keep q_L , and infrastructure deteriorates to I_M . In the final case $I_H \geq I_L^{crit}$ and $\theta_U \geq \phi$, so unconstrained owners build q_H and infrastructure remains I_H , while constrained owners keep q_L . But in practice this final case is unlikely to be relevant, because upgrading areas were targeted as poor.

4.4.4 Relating the model to the empirical analysis

The model demonstrates the role of differing infrastructure investment and owner sorting in accounting for neighborhood quality. For example, consider the following scenario. De novo areas had enough unconstrained owners to ensure that their higher quality infrastructure

(I_H) survived. In this case, the difference in the logarithm of mean housing quality between de novo and control areas is:

$$\ln(\theta_D q_H + (1 - \theta_D) q_L) - \ln(\theta_C q_M + (1 - \theta_C) q_L). \quad (4.8)$$

This quality difference reflects both the effect of the higher infrastructure quality in de novo areas and the different composition of owners in those areas. But controlling for owner fixed effects allows us to focus on houses owned by unconstrained owners, for whom the difference in log mean housing quality between de novo and control areas is:

$$\ln(q(I_H)) - \ln(q(I_M)). \quad (4.9)$$

In other words, under the model's assumptions, adding owner fixed effects allows us to identify the effect of de novo investments on housing quality for unconstrained owners. We acknowledge that in practice adding owner fixed effects may not solve all the potential problems, if for example some owners are constrained in investing in a second house (but not in the first), or have some different preferences for investing across areas. Nevertheless, the model shows that adding owner fixed effects is useful in the context of Sites and Services, where owners in different areas may have had different levels of wealth, due both to sorting and to the program's loans scheme.

The model also allows us to relate differences in infrastructure and housing quality, which we cannot measure directly, to the estimated differences in the value of housing, which are approximated by the hedonic regressions, subject to the limitations discussed in Section 2. Specifically, our model predicts the following:

Proposition 4.2 *For unconstrained owners who face no risk of exogenous house destruction ($d = 0$)*

$$\ln(I_H) - \ln(I_M) = \frac{\gamma - \alpha}{\gamma - \alpha\gamma} (\ln(\pi(q(I_H), I_H)) - \ln(\pi(q(I_M), I_M))), \quad (4.10)$$

and

$$\ln(q(I_H)) - \ln(q(I_M)) = \frac{1}{\gamma} (\ln(\pi(q(I_H), I_H)) - \ln(\pi(q(I_M), I_M))) \quad (4.11)$$

To derive the expression for $\ln(I_H) - \ln(I_M)$, use (4.5) and the fact that $\pi(q_2, I_2) = \pi(0, I_2) + c(q_2)$, and plug in $d = 0$ to obtain $\ln(\pi(q_2, I_2)) = \ln(q_2^\alpha I_2^{1-\alpha}) - \ln(1 - \delta)$. Next apply a similar calculation for $\ln(\pi(q_1, I_1))$ and plug in (4.4) to calculate $\ln \pi(q_2, I_2) -$

$\ln(\pi(q_1, I_1))$. Now combine the expression for $\ln(I_H) - \ln(I_M)$ with (4.4) to derive the expression for $\ln(q(I_H)) - \ln(q(I_M))$.

This result indicates that the difference across areas in log housing quality are smaller than the differences in log values. Taking the above-mentioned estimate of γ suggests that the quality differences across neighborhoods are about $\frac{1}{\gamma} = 0.65$ times the value differences for unconstrained owners, for low values of d , which seem empirically relevant. Our baseline estimate of the hedonic log value differences between de novo and control areas, with owner fixed effects, are around 0.5, suggesting log quality differences of around one third.⁵⁶

The model also allows us to consider differences between upgrading and control areas. As discussed in Section 3, upgrading areas look similar, or in some cases worse than control areas. Table A10 suggests that the worse housing in upgrading areas may in part be explained by owner fixed effects. In the context of the model, this may reflect persistence in upgrading areas of some of the initial owners (or their descendants), who were targeted by the program, and may have been poorer than their counterparts in control areas ($\theta_U < \theta_C$).

Finally, the similarity of housing quality in upgraded and non-upgraded slums is also consistent with the model, if we think of the non-upgraded slums as control areas with constrained owners.

4.4.5 Implications of the model

The model offers several implications for thinking about infrastructure investments for housing. First, an important theme of the paper is that infrastructure investment may crowd in private investments. The model helps us to think about the conditions under which this takes place. In the model, infrastructure investments crowd in more private investments when its quality is sufficiently high and owners can afford to invest in housing quality. In these cases, private investment in housing quality takes place when there is a sufficient fraction of unconstrained owners, either due to their own wealth or through loans that allow them to invest. This also suggests a note of caution: if de novo investments were expanded widely, poor and credit constrained residents may be unable to make full use of them, since infrastructure may deteriorate without sufficient complementary private investment.

Second, the model helps us think about the benefits of early infrastructure investments compared to ex-post infrastructure upgrading. Upgrading areas do not always fully benefit

⁵⁶As discussed in Section 2, the log value differences in the hedonic regressions may understate the actual value differences.

from high levels of infrastructure investments, since in those settings infrastructure either deteriorates or leads to the scrapping of existing houses.⁵⁷

Finally, turning back to our empirical findings, the model can help explain why infrastructure survived better in de novo areas, but not in upgrading areas. The model highlights the importance of feedback from owner investments to infrastructure, which can be seen as a neighborhood externality, and is sometimes overlooked when infrastructure investments are made.

4.5 Concluding remarks

This paper examines the consequences of different strategies for developing basic infrastructure for residential neighborhoods. Specifically, we study the Sites and Services projects implemented in seven Tanzanian cities during the 1970s and 1980s. These projects provided basic infrastructure, leaving it to the residents to build their own houses. We examine the long run development of these neighborhoods, emphasizing the comparison between de novo neighborhoods and other nearby areas that were greenfields when the Sites and Services program started. We also provide descriptive evidence on the development of neighborhoods whose infrastructure was upgraded.

We use high-resolution imagery and building level survey data to study housing quality and infrastructure in the de novo neighborhoods and other areas in their vicinity that were also greenfields to begin with. We find that the de novo neighborhoods developed significantly higher quality housing than other initially unbuilt areas. Our findings reflect complementary private investments that were made in response to the Sites and Services programs. We also present evidence that the initial infrastructure investments in roads and water mains were more likely to persist in de novo areas. For three cities where we have survey data, we find sizeable gains in quality from de novo even when we control for owner fixed effects, although these fixed effects account for up to a third of the average housing quality. Our findings suggest that de novo areas increased local land values by at least 75-100 USD per square meter, compared to total costs of no more than 8-13 USD per square meter (all in 2017 prices).

We also report evidence that de novo neighborhoods attract more educated residents, who

⁵⁷In reality there are other costs of delivering infrastructure in a dense settlement that has developed organically, because it is difficult to resolve coordination failures and negative externalities once they have been put in place. In Sites and Services, for example, the cost per square meter was similar in de novo and upgrading areas, even though de novo areas received formal plots, which upgrading areas did not.

can afford to pay for the higher quality on offer. But as of 2012 almost half of the adults in de novo areas still had no more than primary school education, suggesting that some people with lower lifetime incomes also benefitted from the de novo investments. But we also note that de novo areas were unaffordable to the poorest of the urban poor, a consideration that future projects may want to take into account, perhaps by creating some smaller and more affordable plots. Such plots may also benefit the few people who may be displaced by such projects, even when they target largely empty areas.

Our paper also reports descriptive evidence on upgrading areas, comparing them to nearby control areas, or where the data permit to slums that were not upgraded. The results suggest that upgrading areas now have either similar, or worse, housing quality, and the program's investments in roads and water mains did not survive well in upgrading areas. While we should be cautious in interpreting these results, they suggest that upgrading, at least as implemented in Sites and Services, was not a panacea for pre-existing squatter areas. We cannot rule out that other upgrading efforts may be prove more successful, but in order to provide long lasting benefits, upgrading programs should aim to address the risk of infrastructure deterioration.

Taken together, our findings suggest that de novo investments are a policy tool worthy of consideration for growing African cities. They are considerably cheaper than building public housing, and therefore more affordable for poor countries. They also offer important advantages to residents, who can invest in higher quality housing. Our findings also suggest that it is important to ensure that the infrastructure investments do not deteriorate as a result of poor private investments. While the implementation of Sites and Services projects in Tanzania in the 1970s and 1980s was not flawless, it has taught us important lessons. We hope that these lessons can inform future planning and investment decisions in a continent that is growing in both population and income per capita, but where many poor people still live in poor quality buildings and neighborhoods.

4.6 Main tables

Table 4.1: De novo regressions using imagery data for all seven cities

	(1)	(2)	(3)	(4)	(5)	(6)
	Mean log building footprint area	Share of buildings with painted roof	Mean similarity of building orien- tation	Mean z-index	Share of empty blocks	Share of area built up
<i>Panel A: 500m bandwidth</i>						
De novo	0.114 (0.051)	-0.013 (0.012)	2.821 (0.722)	0.168 (0.057)	-0.152 (0.037)	0.094 (0.013)
Observations	6,562	6,500	6,562	6,562	8,440	8,440
Mean (control)	4.457	0.184	-8.669	0.042	0.306	0.155
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	
<i>Panel B: robustness (mean z-index only as outcome)</i>						
De novo	0.143 (0.053)	0.156 (0.057)	0.168 (0.057)	0.241 (0.100)	0.175 (0.059)	
Observations	6,562	6,562	6,562	4,568	6,158	
Mean (control)	0.042	0.042	0.042	0.015	0.047	

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities. The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(3) in Panel A). In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panel B, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, and column (5) excludes areas within 100 meters of the boundary between upgrade and control areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 90 clusters, except in columns (5) and (6) of Panel A, which have 92 clusters, and column (4) of Panel B, which has 89 clusters, and column (5) of Panel B, which has 88 clusters.

Table 4.2: De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean log building footprint area	Share of buildings with multiple storeys	Share of buildings with a good roof	Share of buildings connected to electricity	Share of buildings with sewerage or septic tank	Mean z-index	Mean log hedonic value
<i>Panel A: 500m bandwidth</i>							
De novo	0.405 (0.070)	0.081 (0.066)	-0.010 (0.008)	0.226 (0.039)	0.142 (0.091)	0.342 (0.091)	0.446 (0.081)
Observations	2,009	1,975	2,009	2,009	2,008	2,009	2,009
Mean (control)	4.739	0.096	0.984	0.466	0.381	0.033	17.234
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Full City	Optimal bandwidth
<i>Panel B: robustness (mean z-index only as outcome)</i>							
De novo	0.263 (0.091)	0.323 (0.076)	0.342 (0.090)	0.408 (0.190)	0.375 (0.093)	0.588 (0.079)	0.312 (0.082)
Observations	2,009	2,009	2,009	1,410	1,887	34,602	34,602
Mean (control)	0.033	0.033	0.033	0.001	0.022	-0.149	0.038
<i>Panel C: robustness (mean log hedonic value only as outcome)</i>							
De novo	0.329 (0.081)	0.431 (0.059)	0.446 (0.077)	0.541 (0.190)	0.427 (0.077)	0.505 (0.089)	0.411 (0.063)
Observations	2,009	2,009	2,009	1,410	1,887	34,602	34,602
Mean (control)	17.234	17.234	17.234	17.231	17.229	17.113	17.239

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A), and in Panel C the dependent variable is the predicted log value from hedonic regressions. In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B and C, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, column (5) excludes areas within 100 meters of the boundary between upgrade and control areas, column (6) changes the control area to the sample of blocks covering the whole city excluding de novo areas, and column (7) uses 2033 observations inside the optimal bandwidth for panel B and 1882 observations inside the optimal bandwidth for panel C based on Imbens and Kalyanaraman (2012). The 'Full City' in column (6) is robust to higher order polynomials in distance to boundary: In panel B: second order polynomial gives an estimate of 0.571 and standard error of 0.087, and third order polynomial gives an estimate of 0.381 and standard error of 0.098. In panel C: second order polynomial gives an estimate of 0.498 and standard error of 0.104, and third order polynomial gives an estimate of 0.296 and standard error of 0.117. The control mean in column (7) reports the mean for the control areas inside the optimal bandwidth (Imbens and Kalyanaraman 2012). Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 29 clusters, except in column (5) of Panels B and C, which has 28 clusters, and column (6) of Panels B and C, which has 439 clusters.

Table 4.3: De novo regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with owner name fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log building footprint area	Multistorey building	Good roof	Connected to electricity	Sewerage or septic tank	Z-index	Log hedonic value
<i>Panel A: Full City, Owner FE</i>							
De novo	0.553 (0.145)	0.119 (0.062)	-0.002 (0.038)	0.417 (0.078)	0.123 (0.091)	0.447 (0.088)	0.604 (0.133)
Observations	20,177	16,605	20,054	20,139	19,595	20,177	20,177
Mean (control)	4.573	0.164	0.968	0.404	0.249	-0.016	17.016
<i>Panel B: Full City, no Owner FE, same sample as A</i>							
De novo	0.594 (0.177)	0.514 (0.138)	-0.010 (0.015)	0.405 (0.066)	0.122 (0.069)	0.642 (0.086)	0.612 (0.185)
Observations	20,177	16,605	20,054	20,139	19,595	20,177	20,177
Mean (control)	4.573	0.164	0.968	0.404	0.249	-0.016	17.016
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Second Order Polynomial	Third Order Polynomial
<i>Panel C: robustness owner FE (z-index only as outcome)</i>							
De novo	0.422 (0.085)	0.411 (0.092)	0.434 (0.087)	0.336 (0.166)	0.471 (0.093)	0.431 (0.112)	0.372 (0.140)
Observations	20,177	20,177	20,177	19,694	19,729	20,177	20,177
Mean (control)	-0.016	-0.016	-0.016	-0.019	-0.018	-0.016	-0.016
<i>Panel D: robustness, no owner FE, same sample as C (z-index only as outcome)</i>							
De novo	0.616 (0.086)	0.648 (0.089)	0.631 (0.084)	0.654 (0.134)	0.675 (0.102)	0.674 (0.081)	0.627 (0.092)
Observations	20,177	20,177	20,177	19,694	19,729	20,177	20,177
Mean (control)	-0.016	-0.016	-0.016	-0.019	-0.018	-0.016	-0.016

Notes: This table reports estimates from regressions using specification (1) and unit level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and the entire city as control areas. The outcomes are measures of housing quality that do not reflect direct investments in de novo areas. Each observation is a property unit in a building, and only multi-unit owners are used. Units are assigned to de novo or control areas based on where their building's centroid falls. Outcomes are measured at the building level (see Data Appendix for further details). In Panels A and B the outcomes vary, while in Panels C and D the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A). Panels A and C display results with unit owner last name fixed effects, including units inside de novo and control areas but restricting the sample by keeping only last name owners that appear more than once in the sample. Panel B (D) displays results with the same sample as in A (C) but without owner last name fixed effects. In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, column (5) excludes areas within 100 meters of the boundary between upgrade and control areas, and columns (6) and (7) control for second and third order polynomials in distance to the boundary, respectively. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 342 clusters, except in column (2) of Panels A and B and in columns (4) and (5) of Panels C and D, which all have 341 clusters.

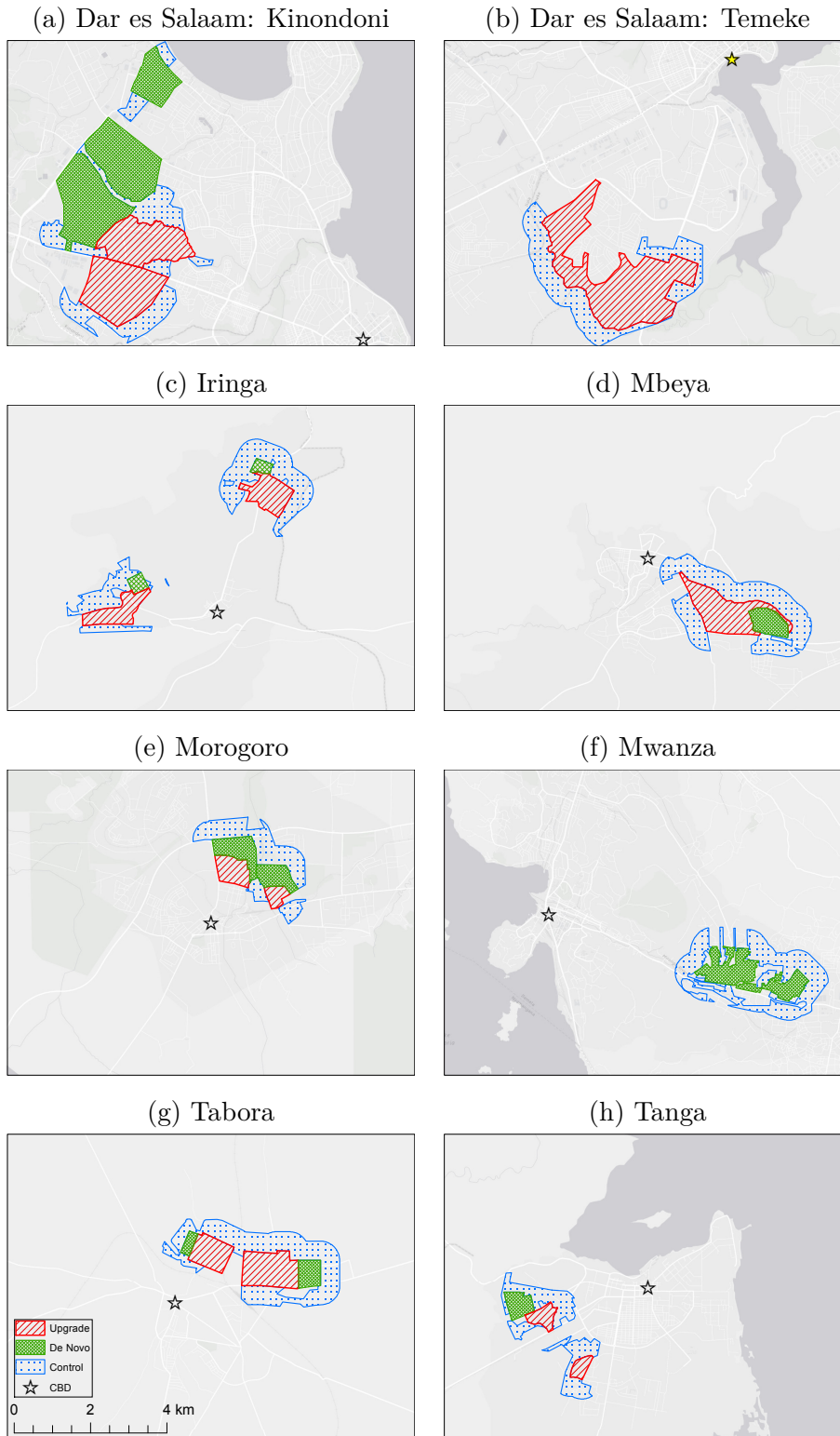
Table 4.4: De novo regressions on persistence measures using imagery and TSCP survey data

	(1)	(2)	(3)	(4)	(5)	(6)
	Imagery	TSCP Survey		TSCP Survey, Excl. Tanga		
	Share of buildings with road within 10m	Share of buildings with road access	Share of buildings connected to water mains	Share of buildings connected to water mains		
<i>Panel A: 500m bandwidth</i>						
De novo	0.141 (0.028)	0.197 (0.050)	0.211 (0.057)	0.233 (0.060)		
Observations	6,562	2,008	2,009	1,952		
Mean (control)	0.202	0.477	0.547	0.547		
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	
<i>Panel B: robustness for share of buildings with road within 10m (Imagery)</i>						
De novo	0.129 (0.025)	0.142 (0.029)	0.142 (0.028)	0.185 (0.056)	0.150 (0.029)	
Observations	6,562	6,562	6,562	4,568	6,158	
Mean (control)	0.202	0.202	0.202	0.205	0.197	
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Upgrade	Full City
<i>Panel C: robustness for share of buildings with road access (TSCP)</i>						
De novo	0.134 (0.039)	0.190 (0.049)	0.199 (0.050)	0.191 (0.159)	0.206 (0.051)	0.170 (0.056)
Observations	2,008	2,008	2,008	1,409	1,886	34,578
Mean (control)	0.477	0.477	0.477	0.485	0.449	0.573
<i>Panel D: robustness for share of buildings connected to water mains (TSCP)</i>						
De novo	0.164 (0.060)	0.188 (0.051)	0.209 (0.057)	0.319 (0.128)	0.204 (0.062)	0.403 (0.042)
Observations	2,009	2,009	2,009	1,410	1,887	34,588
Mean (control)	0.547	0.547	0.547	0.535	0.534	0.433

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities (road within 10m) and TSCP survey data for Mbeya, Mwanza, and Tanga (road access and connection to water mains). The sample includes the de novo areas and control areas within 500 meters of their boundary. The outcomes are measures of persistence of infrastructure treatment. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the share of buildings with a road within 10 meters (from imagery data), in Panel C the dependent variable in all columns is the share of buildings with road access (from TSCP data), and in Panel D the dependent variable is the share of buildings connected to water mains (from TSCP data). In each specification the regressor of interest is de novo, and the control variables include a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B, C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between de novo and control areas, and column (5) excludes areas within 100 meters of the boundary between upgrade and control areas. Moreover, in Panels C and D, column (6) changes the control area to the sample of blocks covering the whole city excluding upgrade areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services area. There are 29 clusters in TSCP data, except in column (5) of Panels C and D, which have 28 clusters, and in column (6) of Panels C and D, which have 439 clusters. There are 90 clusters in imagery data, except in column (4) of Panel B which has 88 clusters, and column (5) of Panel B which has 89 clusters.

4.A Appendix tables and figures

Figure 4.A1: Locations of de novo, upgrading, and control areas by city



Notes: This figure maps de novo (green cross-hatch), upgrading (red hatch), control areas (blue dots), and the CBD (yellow star) for each city. Panel (a) shows the northern part of Dar es Salaam (Kinondoni), while the southern part (Temeke) is shown in panel (b). Control areas are all 500m buffers of study areas, excluding land that was determined uninhabitable, built-up, or designated for specific use prior to the program. Each map is set to the same scale. Background imagery from ArcGIS is for context only and was not used for analysis, it depicts modern day roads (white lines), heavily vegetated areas (green-grey) and water bodies (dark grey).

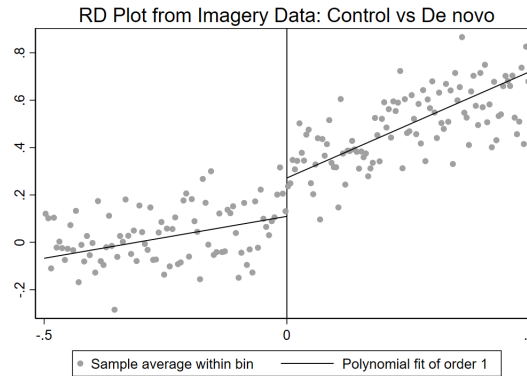
Figure 4.A2: Example images of de novo, upgrade and control areas



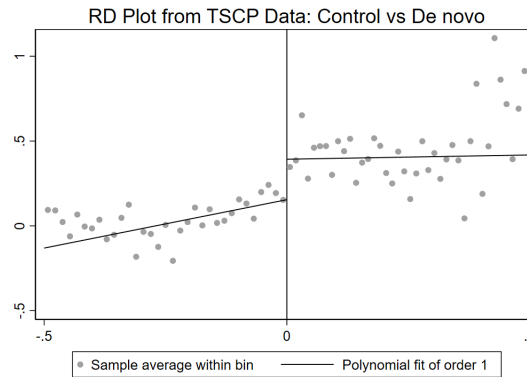
Notes: Each of the three images covers an area of approximately 440 x 360 meters. Source: Google earth V 7.1.2. (2018). Kinondoni District, Dar-es-Salaam, Tanzania.

Figure 4.A3: Regression discontinuity plots of summary outcomes from Tables 1 and 2

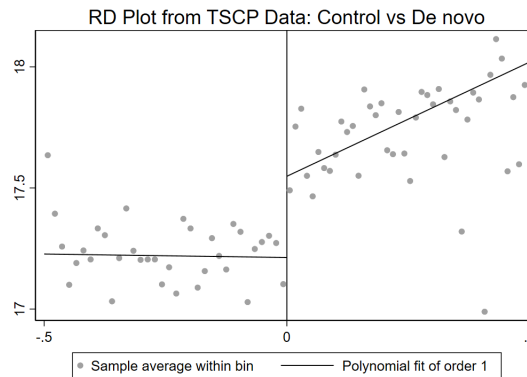
(a) Z-index from Table 1



(b) Z-index from Table 2



(c) Log hedonic value from Table 2



Notes: This figure plots the raw summary outcomes (z-index and log hedonic value) on the y-axis against the running variable (x-axis) being distance to boundary between de novo and control areas in kilometers. Observations from control areas are to the left of the cutoff (marked 0) and de novo areas to the right. The graphs are created with the command `rdplot` (Calonico et al 2017), using a triangular kernel. The sample is defined by a 500m bandwidth on either side of the boundary. For subfigure (a), the variable definition and data is the same as in Table 1, panel A, column 4 and the number of observations in (a) is 6562, with 3147 to the left (control) and 3415 to the right (de novo) of the cutoff. The average bin length is 6m to the left of the cutoff and 5m to the right for subfigure (a). For subfigure (b), the variable definition and data is the same as in Table 2, panel A, column 6 and the number of observations in (b) is 2009, with 1177 to the left (control) and 832 to the right (de novo) of the cutoff. The average bin length is 16m to the left of the cutoff and 12m to the right for subfigure (b). For subfigure (c), the variable definition and data is the same as in Table 2, panel A, column 7 and the number of observations in (c) is 2009, with 1177 to the left (control) and 832 to the right (de novo) of the cutoff. The average bin length is 14m to the left of the cutoff and 12m to the right for subfigure (c).

Table 4.A1: De novo neighborhoods

City	Area within city	Round	Pre-treatment satellite photos	Pre-treatment topographic map
Dar es Salaam	Sinza	1	1966	N
Dar es Salaam	Kijitonyama	1	1966	N
Dar es Salaam	Mikocheni	1	1966	N
Mbeya	Mwanjelwa (*)	1	1966	N
Mwanza	Nyakato (**)	1	1966	N
Tanga	Nguvu Mali (***)	2	1966	N
Tabora	Isebya	2	1978	1967
Tabora	Kiloleni	2	1978	1967
Morogoro	Kichangani	2	N	1974
Morogoro	Msamvu	2	N	1974
Iringa	Kihesa & Mtuiwila	2	1966	1982
Iringa	Mwangata	2	1966	1982

Notes: This table reports information about the 12 de novo neighborhoods, the round in which the Sites and Services projects were implemented, and the data we have on the areas before the program was implemented. (*) Treatment area maps were unavailable, so areas were drawn by experts that were involved in the projects, as explained in the Data Appendix. (**) Treatment area maps were unavailable, so we inferred from the detailed Mwanza central plan. (***) We have some uncertainty as to the extent of infrastructure that was actually provided in Nguvu Mali.

Table 4.A2: Upgrading neighborhoods

City	Area within city	Round	Pre-treatment satellite photos	Pre-treatment topographic map
Dar es Salaam	Manzese A	1	1966 & 1969	N
Dar es Salaam	Manzese B	1	1966 & 1969	N
Mbeya	Mwanjelwa (*)	1	1966	N
Dar es Salaam	Mtoni & Tandika	2	1966	N
Iringa	Kihesa	2	1966	1982
Iringa	Mwangata	2	1966	1982
Morogoro	Kichangani	2	N	1974
Morogoro	Msamvu	2	N	1974
Tabora	Isebya	2	1978	1967
Tabora	Kiloleni	2	1978	1967
Tanga	Gofu Juu	2	1966	N
Tanga	Mwakizaro	2	1966	N

Notes: this table reports information about the 12 upgrading neighborhoods, the round in which the Sites and Services projects were implemented, and the data we have on the areas before the program was implemented. (*) Treatment area maps were unavailable, so areas were drawn by experts that were involved in the projects, as explained in the Data Appendix.

Table 4.A3: Plot counts and population by project type

		Plots com- pleted by 1980s	Population in 2002	Ratio of popula- tion to plots com- pleted	Area (sq-km)	Population density (people per sq-km)	Built area (build- ing foot- prints, sq-km)	Crowding (people per sq-km of built area)
Round 1	De novo	8,527	89,207	10.5	8.6	10,400	2.7	32,975
	Control for de novo		44,846		6.7	6,723	1.5	29,151
	Upgrading	14,634	200,630	13.7	6.5	31,064	2.9	68,084
	Control for upgrading		89,920		6.2	14,415	2.0	44,849
Round 2	Denovo	1,978	17,927	9.1	2.5	7,158	0.5	36,883
	Control for de novo		14,708		6.5	2,253	0.6	23,976
	Upgrading	20,128	204,074	10.1	10.5	19,483	3.2	64,721
	Control for upgrading		67,871		11.7	5,801	1.9	36,593
Total	Denovo	10,505	107,134	10.2	11.1	9,667	3.2	33,570
	Control for de novo		59,554		13.2	4,512	2.2	27,676
	Upgrading	34,762	404,704	11.6	16.9	23,900	6.1	66,346
	Control for upgrading		157,791		17.9	8,796	3.9	40,882

Notes: This table reports completed plot counts and population in 2002 by treatment type and round.

Table 4.A4: Summary statistics

	Imagery data (Blocks)			
	De novo	Upgrade	Control	Total
Mean log building footprint area	4.580 (0.569)	4.243 (0.503)	4.381 (0.699)	4.394 (0.625)
Share of buildings with painted roof	0.337 (0.314)	0.186 (0.222)	0.174 (0.266)	0.221 (0.277)
Mean similarity of building orientation	-4.735 (5.751)	-6.981 (5.208)	-8.202 (7.638)	-6.911 (6.657)
Share of buildings with road within 10m	0.288 (0.322)	0.213 (0.277)	0.202 (0.307)	0.228 (0.305)
Obs.	3,925	4,341	6,380	14,646
	TSCP data (Blocks)			
	De novo	Upgrade	Control (Full City)	Total
Mean log building footprint area	5.134 (0.464)	4.612 (0.456)	4.706 (0.688)	4.712 (0.684)
Share of buildings with multiple storeys	0.202 (0.384)	0.015 (0.100)	0.071 (0.240)	0.072 (0.243)
Share of buildings with a good roof	0.975 (0.109)	0.868 (0.268)	0.951 (0.174)	0.950 (0.175)
Share of buildings connected to electricity	0.713 (0.344)	0.423 (0.322)	0.425 (0.431)	0.430 (0.429)
Share of buildings with sewerage or septic tank	0.547 (0.412)	0.227 (0.328)	0.387 (0.431)	0.387 (0.430)
Share of buildings connected to water mains	0.767 (0.320)	0.493 (0.329)	0.483 (0.434)	0.488 (0.433)
Share of buildings with road access	0.676 (0.440)	0.748 (0.341)	0.611 (0.453)	0.615 (0.451)
Mean log hedonic value	17.689 (0.496)	17.039 (0.468)	17.200 (0.723)	17.207 (0.719)
Obs.	798	729	40,563	42,090

Notes: Summary statistics are estimates of the sample mean and its standard deviation in parentheses. The first panel displays summary statistics for outcomes derived from satellite imagery for all seven Sites and Services cities over the sample of observations with their centroid in either a de novo, upgrading, or control area. The second panel displays summary statistics for outcomes derived from TSCP survey data for Mbeya, Mwanza, and Tanga over the whole city sample. Observations are blocks based on an arbitrary grid of 50x50 meter blocks for both imagery and TSCP data. All columns report the maximum populated number of observations. Block outcomes are derived from all buildings with a centroid in the block. Blocks that fall between two treatment types are assigned according to where their centroid falls. The imagery variable painted roof has 14530 observations for the Total column, i.e. 116 less than the other variables. This is due to measurement error in assigning roof type to a building (outlines of some buildings in Dar es Salaam did not correspond to an actual building on the satellite image). Similarly, due to the survey nature of the TSCP data, in the Total column, the following TSCP variables have fewer than 42,090 observations: multiple storeys has 40,990 observations, good roof has 42,047 observations, sewerage or septic tank has 41,948 observations, water mains has 42,063 observations, and road access has 42,062 observations.

Table 4.A5: De novo regressions balancing first geography

	(1)	(2)	(3)
	Distance to Shore (km)	Block contains river or stream	Ruggedness within 50m
<i>Panel A: no controls, 500m bandwidth (Imagery)</i>			
De novo	-0.167 (0.087)	-0.007 (0.013)	-0.646 (0.211)
Observations	8,440	8,440	8,440
Mean (control)	7.292	0.050	2.930
<i>Panel B: baseline controls, 500m bandwidth (Imagery)</i>			
De novo	-0.080 (0.063)	-0.017 (0.017)	-0.266 (0.223)
Observations	8,440	8,440	8,440
Mean (control)	7.292	0.050	2.930
<i>Panel C: baseline controls, 500m bandwidth (TSCP)</i>			
De novo	-0.064 (0.056)	-0.074 (0.025)	-0.760 (0.555)
Observations	2,693	2,693	2,693
Mean (control)	5.512	0.062	3.721
<i>Panel D: baseline controls, Full City (TSCP)</i>			
De Novo	-0.819 (0.222)	0.009 (0.011)	-0.364 (0.337)
Observations	35,662	35,662	35,662
Mean (control)	4.850	0.016	3.236

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities in Panels A and B, while in Panels C and D the outcomes are derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the de novo areas and control areas within 500 meters of their boundary in Panels A, B and C. In Panel D, the sample includes de novo areas and the full city as control areas. In all panels, all blocks, including empty ones, are used. The outcomes are measures of geographical fundamentals and can be interpreted as quantifying any imbalance in selection of de novo and control areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to de novo or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A, the controls are only nearest de novo fixed effects. In Panels B, C and D, the controls are the regular ones: a linear control in distance to the de novo-control area boundary interacted with the de novo indicator, fixed effects for the nearest de novo area, and distance to the Central Business District (CBD) of each city.

Table 4.A6: De novo regressions of adult census outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean years of schooling	Share with exactly primary education	Share with more than primary education	Share attending school	Share literate in any language	Share literate in Swahili	Share literate in English
De novo	0.566 (0.121)	-0.041 (0.016)	0.051 (0.015)	0.018 (0.009)	0.010 (0.006)	0.004 (0.010)	0.053 (0.023)
Observations	814	814	814	814	814	814	814
Mean (control)	9.343	0.412	0.497	0.128	0.960	0.936	0.449

Notes: This table reports estimates from regressions using cut Enumeration Area (EA) level observations with outcomes derived from Tanzania 2012 Census microdata for all seven Sites and Services cities. In each specification the regressor of interest is de novo, and the control variables include city fixed effects (separate for Temeke and Kinondoni in Dar es Salaam), and distance to the Central Business District (CBD) of each city. The sample includes de novo observations and control areas which are near de novo areas. The outcomes are measures of sorting into the treatment and control areas. Outcomes are the EA mean over the set of all adults at least 18 years old enumerated in the EA. Each observation is an EA of varying size, or a cut EA if the EA intersects both de novo and control areas. Cut EAs are assigned to de novo, and/or control areas if more than 5 percent of the cut EA lies inside the respective area. Analytic weights for the cut EA observations used in the regression are based on the proportion of the EA area that lies inside each treatment or control area. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares. There are 90 clusters.

Table 4.A7: Upgrading regressions balancing first geography

	(1)	(2)	(3)
	Distance to Shore (km)	Block contains river or stream	Ruggedness within 50m
<i>Panel A: no controls, 500m bandwidth (Imagery)</i>			
Upgrade	-0.057 (0.075)	-0.029 (0.012)	-0.389 (0.161)
Observations	12,854	12,854	12,854
Mean (control)	6.778	0.060	2.663
<i>Panel B: baseline controls, 500m bandwidth (Imagery)</i>			
Upgrade	0.021 (0.049)	-0.050 (0.019)	0.099 (0.233)
Observations	12,854	12,854	12,854
Mean (control)	6.778	0.060	2.663
<i>Panel C: baseline controls, 500m bandwidth (TSCP)</i>			
Upgrade	0.058 (0.039)	-0.075 (0.041)	-0.370 (0.328)
Observations	2,576	2,576	2,576
Mean (control)	7.873	0.063	2.386
<i>Panel D: baseline controls, Full City (TSCP)</i>			
Upgrade	0.045 (0.126)	0.042 (0.024)	-1.002 (0.315)
Observations	11,798	11,798	11,798
Mean (control)	7.079	0.019	2.422

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities in Panels A and B, while in Panels C and D the outcomes are derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and control areas within 500 meters of their boundary in Panels A, B and C. In Panel D, the sample includes upgrading areas and the full city as control areas. In all panels, all blocks, including empty ones, are used. The outcomes are measures of geographical fundamentals and can be interpreted as quantifying any imbalance in selection of upgrading and control areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A, the controls are only nearest upgrading fixed effects. In Panels B, C and D, the controls are the regular ones: a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city.

Table 4.A8: Upgrading regressions using imagery data for all seven cities

	(1)	(2)	(3)	(4)	(5)	(6)
	Mean log building footprint area	Share of buildings with painted roof	Mean similarity of building orien- tation	Mean z-index	Share of empty blocks	Share of area built up
<i>Panel A: 500m bandwidth</i>						
Upgrade	-0.053 (0.042)	-0.005 (0.010)	0.500 (0.389)	-0.010 (0.034)	-0.139 (0.033)	0.076 (0.016)
Observations	10,909	10,837	10,909	10,909	12,854	12,854
Mean (control)	4.333	0.146	-7.352	-0.008	0.234	0.219
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	
<i>Panel B: robustness (mean z-index only as outcome)</i>						
Upgrade	-0.014 (0.035)	-0.014 (0.035)	-0.011 (0.034)	-0.049 (0.061)	-0.007 (0.035)	
Observations	10,909	10,909	10,909	7,573	10,531	
Mean (control)	-0.008	-0.008	-0.008	0.008	-0.017	
	Mean log building footprint area	Mean similarity of building orien- tation	Share of empty blocks	Share of area built up		
<i>Panel C: upgrade vs old slums</i>						
Upgrade	-0.152 (0.073)	0.801 (0.396)	-0.278 (0.106)	0.122 (0.047)		
Observations	8,000	8,000	9,319	9,319		
Mean (control)	4.214	-6.195	0.231	0.303		
<i>Panel D: upgrade vs old slums, first geography controls</i>						
Upgrade	-0.139 (0.065)	0.755 (0.264)	-0.233 (0.091)	0.123 (0.045)		
Observations	8,000	8,000	9,319	9,319		
Mean (control)	4.214	-6.195	0.231	0.303		

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from imagery for all seven Sites and Services cities. The sample in Panels A and B includes the upgrading areas and control areas within 500 meters of their boundary. The sample in Panels C and D includes the upgrading areas in Dar es Salaam and the areas of that city which could be identified as slums before Sites and Services and that were not treated (see the Data Appendix for more details). The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panels A, C and D the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(3) in Panel A). In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panel B, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 117-125 clusters in Panel A, 117 clusters in Panel B, and 104-105 clusters in Panels C and D.

Table 4.A9: Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean log building footprint area	Share of buildings with multiple storeys	Share of buildings with a good roof	Share of buildings connected to electricity	Share of buildings with sewerage or septic tank	Mean z-index	Mean log hedonic value
<i>Panel A: 500m bandwidth</i>							
Upgrade	-0.111 (0.101)	-0.112 (0.050)	-0.178 (0.084)	-0.082 (0.078)	-0.130 (0.079)	-0.569 (0.253)	-0.181 (0.133)
Observations	2,066	1,863	2,062	2,066	2,059	2,066	2,066
Mean (control)	4.801	0.094	0.972	0.524	0.350	0.041	17.281
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Full City	Optimal bandwidth
<i>Panel B: robustness (mean z-index only as outcome)</i>							
Upgrade	-0.572 (0.246)	-0.631 (0.243)	-0.597 (0.242)	-0.456 (0.323)	-0.578 (0.264)	-0.681 (0.209)	-0.633 (0.265)
Observations	2,066	2,066	2,066	1,462	2,001	11,225	11,225
Mean (control)	0.041	0.041	0.041	0.046	0.030	-0.084	0.045
<i>Panel C: robustness (mean log hedonic value only as outcome)</i>							
Upgrade	-0.211 (0.129)	-0.286 (0.119)	-0.236 (0.119)	-0.200 (0.224)	-0.189 (0.135)	-0.370 (0.083)	-0.217 (0.149)
Observations	2,066	2,066	2,066	1,462	2,001	11,225	11,225
Mean (control)	17.281	17.281	17.281	17.300	17.273	17.259	17.267

Notes: This table reports estimates from regressions using specification (1) and block level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and control areas within 500 meters of their boundary. The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A), and in Panel C the dependent variable is the predicted log value from hedonic regressions. In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B and C, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas, column (6) changes the control area to the sample of blocks covering the whole city excluding treatment areas, and column (7) uses 2889 observations inside the optimal bandwidth for panel B and 1699 observations inside the optimal bandwidth for panel C based on Imbens and Kalyanaraman (2012). The control mean in column (7) reports the mean for the control areas inside the optimal bandwidth. The 'Full City' in column (6) is robust to higher order polynomials in distance to boundary: In panel B: second order polynomial gives an estimate of -0.737 and standard error of 0.243, and third order polynomial gives an estimate of -0.662 and standard error of 0.243. In panel C: second order polynomial gives an estimate of -0.361 and standard error of 0.117, and third order polynomial gives an estimate of -0.237 and standard error of 0.143. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 30 clusters in Panel A, and 28-30 clusters in Panels B and C, except in column (6) of Panels B and C, which have 132 clusters.

Table 4.A10: Upgrading regressions using TSCP survey data for Mbeya, Mwanza, and Tanga with Owner Name Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Log building footprint area	Multistorey building	Good roof	Connected to electricity	Sewerage or septic tank	Z-index	Log hedonic value
<i>Panel A: Full City, Owner FE</i>							
Upgrade	-0.225 (0.150)	-0.186 (0.067)	-0.021 (0.047)	-0.058 (0.094)	-0.039 (0.076)	-0.243 (0.140)	-0.218 (0.144)
Observations	18,843	14,227	18,708	18,805	18,231	18,843	18,843
Mean (control)	4.601	0.205	0.966	0.416	0.221	0.002	17.026
<i>Panel B: Full City, no Owner FE, same sample as A</i>							
Upgrade	-0.243 (0.146)	-0.123 (0.084)	-0.011 (0.012)	-0.098 (0.054)	-0.172 (0.082)	-0.256 (0.105)	-0.310 (0.152)
Observations	18,843	14,227	18,708	18,805	18,231	18,843	18,843
Mean (control)	4.601	0.205	0.966	0.416	0.221	0.002	17.026
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Second Order Polynomial	Third Order Polynomial
<i>Panel C: robustness owner FE (z-index only as outcome)</i>							
Upgrade	-0.244 (0.136)	-0.264 (0.137)	-0.250 (0.140)	-0.475 (0.161)	-0.228 (0.141)	-0.066 (0.155)	0.040 (0.180)
Observations	18,843	18,843	18,843	17,914	18,780	18,843	18,843
Mean (control)	0.002	0.002	0.002	0.000	0.000	0.002	0.002
<i>Panel D: robustness, no owner FE, same sample as C (z-index only as outcome)</i>							
Upgrade	-0.329 (0.101)	-0.218 (0.105)	-0.272 (0.106)	-0.390 (0.094)	-0.255 (0.105)	-0.060 (0.133)	0.116 (0.137)
Observations	18,843	18,843	18,843	17,914	18,780	18,843	18,843
Mean (control)	0.002	0.002	0.002	0.000	0.000	0.002	0.002

Notes: This table reports estimates from regressions using specification (1) and unit level observations with outcomes derived from TSCP survey data for the three cities where these data exist: Mbeya, Mwanza, and Tanga. The sample includes the upgrading areas and the entire city as control areas. The outcomes are measures of housing quality that do not reflect direct investments in upgrading areas. Each observation is a property unit in a building, and only multi-unit owners are used. Units are assigned to upgrading or control areas based on where their building's centroid falls. Outcomes are measured at the building level (see Data Appendix for further details). In Panels A and B the outcomes vary, while in Panels C and D the dependent variable in all columns is the z-index (composed of all outcomes in columns (1)-(5) in Panel A). Panels A and C display results with unit owner last name fixed effects, including units inside upgrading and control areas but restricting the sample by keeping only last name owners that appear more than once in the sample. Panel B (D) displays results with the same sample as in A (C) but without owner last name fixed effects. In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas, and columns (6) and (7) control for second and third order polynomials in distance to the boundary, respectively Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services areas. There are 111-112 clusters.

Table 4.A11: Upgrading regressions on persistence measures using imagery and TSCP survey data

	(1)	(2)	(3)	(4)	(5)	(6)
	Imagery	Imagery Slums 1979	TSCP Survey		TSCP Survey, Excl. Tanga	
	Share of buildings with road within 10m	Share of buildings with road within 10m	Share of buildings with road access	Share of buildings connected to water mains	Share of buildings connected to water mains	
<i>Panel A: 500m bandwidth</i>						
Upgrade	-0.019 (0.018)	0.018 (0.039)	0.004 (0.056)	-0.078 (0.087)	-0.059 (0.109)	
Observations	10,909	8,000	2,065	2,066	1,923	
Mean (control)	0.190	0.032	0.775	0.586	0.586	
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	
<i>Panel B: robustness for share of buildings with road within 10m (Imagery)</i>						
Upgrade	-0.021 (0.019)	-0.015 (0.018)	-0.019 (0.019)	-0.008 (0.038)	-0.022 (0.019)	
Observations	10,909	10,909	10,909	7,573	10,531	
Mean (control)	0.190	0.190	0.190	0.197	0.190	
	Geography	Lat-Long 2 nd Poly.	Historical CBD	Doughnut 100m	Exclude 100m to Denovo	Full City
<i>Panel C: robustness for share of buildings with road access (TSCP)</i>						
Upgrade	0.006 (0.046)	-0.014 (0.052)	-0.007 (0.053)	-0.053 (0.098)	0.012 (0.057)	-0.013 (0.048)
Observations	2,065	2,065	2,065	1,461	2,000	11,207
Mean (control)	0.775	0.775	0.775	0.764	0.768	0.771
<i>Panel D: robustness for share of buildings connected to water mains (TSCP)</i>						
Upgrade	-0.079 (0.083)	-0.102 (0.081)	-0.089 (0.081)	-0.045 (0.132)	-0.076 (0.089)	-0.181 (0.058)
Observations	2,066	2,066	2,066	1,462	2,001	11,214
Mean (control)	0.586	0.586	0.586	0.587	0.579	0.586

Notes: This table reports estimates from regressions using specification (1) and block level observations. The outcomes in both columns (1) and (2) of Panel A, and in Panel B (road within 10m) are derived from imagery for all seven Sites and Services cities. The outcomes in columns (3) and (4) of Panel A, and in Panels C and D (road access and connection to water mains) are derived from TSCP survey data for Mbeya, Mwanza, and Tanga. In column (5) of Panel A, Tanga is excluded from the TSCP survey data because of the uncertainty about water mains in that city (see Data Appendix). The sample in Panels B-D and in Panel A columns (1) and (3)-(5) includes the upgrading areas and control areas within 500 meters of their boundary. The sample in column (2) in Panel A includes upgrading areas and the areas of Dar es Salaam that could be identified as slums in 1979 but excluded from the Sites and Services projects (see Data Appendix). The outcomes are measures of persistence of infrastructure treatment. Each observation is a block based on an arbitrary grid of 50x50 meter blocks. Blocks are assigned to upgrading or control areas based on where their centroid falls. Outcomes are derived from the set of buildings with a centroid in the block (see Data Appendix for further details). In Panel A the outcomes vary, while in Panel B the dependent variable in all columns is the share of buildings with a road within 10 meters (from imagery data), in Panel C the dependent variable in all columns is the share of buildings with road access (from TSCP data), and in Panel D the dependent variable is the share of buildings connected to water mains (from TSCP data). In each specification the regressor of interest is upgrading, and the control variables include a linear control in distance to the upgrading-control area boundary interacted with the upgrading indicator, fixed effects for the nearest upgrading area, and distance to the Central Business District (CBD) of each city. In addition, in Panels B, C and D, column (1) includes geographic controls, column (2) includes a second order polynomial in longitude and latitude, column (3) uses distance to historical (instead of contemporary) CBDs, column (4) excludes areas within 100 meters of the boundary between upgrade and control areas, column (5) excludes areas within 100 meters of the boundary between de novo and control areas. Moreover, in Panels C and D, column (6) changes the control area to the sample of blocks covering the whole city excluding de novo areas. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares, corresponding to the median size of Sites and Services area. There are 28-30 clusters in TSCP data, except in column (6) of Panels C and D, which have 132 clusters. There are 117 clusters in imagery data, except in column (2) of Panel A which has 104 clusters.

Table 4.A12: Upgrading regressions of adult census outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean years of schooling	Share with exactly primary education	Share with more than primary education	Share attending school	Share literate in any language	Share literate in Swahili	Share literate in English
Upgrade	-0.469 (0.131)	0.049 (0.012)	-0.060 (0.016)	-0.018 (0.004)	-0.012 (0.004)	-0.011 (0.004)	-0.066 (0.017)
Observations	2,842	2,842	2,842	2,842	2,842	2,842	2,842
Mean (control)	8.349	0.533	0.357	0.084	0.955	0.934	0.315

Notes: This table reports estimates from regressions using cut Enumeration Area (EA) level observations with outcomes derived from Tanzania 2012 Census microdata for all seven Sites and Services cities. In each specification the regressor of interest is upgrade, and the control variables include city fixed effects (separate for Temeke and Kinondoni in Dar es Salaam), and distance to the Central Business District (CBD) of each city. The sample includes upgrading observations and control areas which are near upgrading areas. The outcomes are measures of sorting into the treatment and control areas. Outcomes are the EA mean over the set of all adults at least 18 years old enumerated in the EA. Each observation is an EA of varying size, or a cut EA if the EA intersects both treatment and control areas. Cut EAs are assigned to upgrading and/or control areas if more than 5 percent of the cut EA lies inside the respective area. Analytic weights for the cut EA observations used in the regression are based on the proportion of the EA area that lies inside each treatment or control area. Standard errors, in parentheses, are clustered by arbitrary 850x850 meter grid squares. There are 124 clusters.

Table 4.A13: Details on the selection of control areas by city

Dar es Salaam	<ul style="list-style-type: none"> • Sources: the 1974 (World Bank 1974a) and 1977 (World Bank 1977b) project proposal maps. • De novo and upgrading: the 1974 map is used to trace areas in the north of Dar es Salaam (Kinondoni Municipality), and the 1977 map is used in the south of Dar es Salaam (Temeke municipality). • Exclusions: the 1974 map is used to exclude areas in Kinondoni where we identify previously established residential areas and land reserved for special institutions and industry. The 1977 map is used to exclude areas in Temeke where there are low density residential areas and special institutions.
Iringa	<ul style="list-style-type: none"> • Sources: the 1977 project proposal map (World Bank 1977b), and a 1978 topographic map (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions from control areas: the 1977 project proposal map is used to exclude industrial and established residential areas east of Mwangata. The 1978 topographic map is used to exclude already developed areas west and east of Mwangata, and also north, south and east of Kihesa. Additionally, north of Mwangata is excluded because of a power plant.
Mbeya	<ul style="list-style-type: none"> • Sources: a 1966 satellite image (United States Geological Survey, 2015), and drawings by experts on the Sites and Services projects in Mbeya. Those experts are Shaoban Sheuya, Anna Mtani, and Amulike Mahenge and were all interviewed by the authors in Dar es Salaam, June 30, 2016. • De novo and upgrading: the drawings from our experts were used to trace areas. • Exclusions: the 1966 satellite image is used to exclude already built-up areas at the center of the city and areas with shops along the highway southeast of Mwanjelwa, already developed areas northwest of Mwanjelwa, and the airport. • For consistency across TSCP and imagery data, we kept all TSCP buildings in Mbeya within the minimum bounding rectangle of the Worldview imagery for Mbeya, this excluded a very small fraction of buildings at the fringes.
Morogoro	<ul style="list-style-type: none"> • Sources: the 1977 project proposal map (World Bank 1977b), and a 1974 topographic map (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the 1977 project proposal map is used to exclude a large industrial area southwest of Msamvu and a large previously developed area to the south of Msamvu. The 1974 topographic map is used to exclude a previously developed area south of Kichangani, and to confirm the exclusions from the 1977 project proposal map. Finally 0.07km² of undeveloped farm land is excluded from the area to the adjacent to the railway station.
Mwanza	<ul style="list-style-type: none"> • Sources: a 1973 cadastral map (Mwanza City Municipality, 1973). • De novo: the cadastral map is used to trace areas, it delineates all surveyed plots and so contains a few that are outside of the actual Sites and Services treatment. We include plots that are small (288m² is the known treated plot area) and recorded with a plot number, and community buildings. We do not include plots that are large or that are small but do not have a recorded plot number. • Exclusions: the cadastral map is used to exclude areas with large plots or plots without a recorded number. Also excluded are previously developed areas along the road in the southeast of Mwanza, as well as areas to the north that are off of the map. The 1966 satellite imagery was used to exclude built-up center of the city.
Tabora	<ul style="list-style-type: none"> • Sources: the 1977 project proposal map (World Bank 1977b), a 1967 topographic map (Directorate of Overseas Surveys, 2015), and 1978 aerial imagery (Directorate of Overseas Surveys, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the project proposal map is used to excluded previously built areas to the west and southwest of the Kiloleni. The 1967 topographic map is used to exclude an industrial area to the south of Isebeya in between the two of upgrading area. The 1978 aerial image is used to confirm the exclusions.
Tanga	<ul style="list-style-type: none"> • Sources: the 1977 project proposal map (World Bank 1977b), and a 1966 satellite image (United States Geological Survey, 2015). • De novo and upgrading: the 1977 project proposal map is used to trace areas. • Exclusions: the 1966 satellite image is used to exclude already developed areas south, southwest, north and east of Gofu Juu and east of Mwakizaro, as well as the center of the city near the coast. The 1977 project proposal map is used to exclude industrial area between Gofu Juu and Mwakizaro.

Notes: This table explains what imagery and maps were used to (a) delineate the de novo and upgrading areas, and (b) create exclusion areas (i.e. areas to be excluded from the control areas) among areas that are within 500 meters of Sites and Services, as explained in the Data Appendix. Sources are all georeferenced maps of the city in question. Almost all areas in the studied cities were covered by these maps, with minor exceptions in the western areas of Tabora, and north of the northern treatment area (Kihesa neighborhood) in Iringa.

Table 4.A14: Description of variables derived from imagery data

Variable label	Definition
Log building footprint area	Calculated directly for the shape file (calculated as a direct measure for the building, or a sample average of that measure for each block.)
Painted roof	Indicator for painted as opposed to tin or rusted tin (an indicator for the building or a share of buildings with painted roofs for each block). Please see the Data Appendix.
Similarity of orientation	Calculated using the main axis of the minimum bounding box that contains each building. We then calculated the difference in orientation between each building and its neighboring building, modulo 90 degrees, with more similar orientations representing a more regular layout (an indicator for the building or a sample average for each block).
Z-index	We construct a family of outcomes measure following Kling et al. 2007 and Banerjee et al. 2014. We integrate all “good” variables into one index. We subtract the mean in the control group and divide the result by the standard deviation in the control group. Then we create the index by taking a simple average of the normalized variables (a measure for the building or a sample average for each block). Please refer to the Data Appendix for more details.
Road within 10m	An indicator that the distance form the boundary of the building to the nearest roads is no more than 10m).
Distance to the CBD	The CBD for each city is the centroid of the most lit pixel in 1992 from the NOAA “Average Visible and Stable Lights, Cloud Free Coverage” dataset. The distance to the CBD is calculated from the centroids of each building or block.
Empty block indicator	Indicator for a block that has no buildings.
Share of area built up	Share of the area of the block that is built.
Number of buildings	Count of buildings in a 50x50m block.

Note: this table describes the variables derived from imagery data.

Table 4.A15: Description of TSCP variables and how they are created

Variable label	Definition
Connected to electricity	Indicator for whether a building is connected to electricity.
Sewerage or septic tank	Indicator for good sanitation, i.e. having sewerage or a septic tank as opposed to an alternative of pit latrine, no sanitation at all, or other.
Good roof	Indicator for roof being made of concrete, metal sheets, clay tiles or cement tiles as opposed to an alternative of grass/palm, asbestos, timber or other. This is a different measure from the "Painted roof" variable in Table A14.
Multistorey building	Indicator for one or more storeys above the ground floor.
Z-index	We construct a family of outcomes measure following Kling et al. 2007 and Banerjee et al. 2014. We integrate all "good" variables into one index. We subtract the mean in the control group and divide the result by the standard deviation in the control group. Then we create the index by taking a simple average of the normalized variables.
Hedonic Value	We run a hedonic regression using property values of 3663 buildings in Arusha based on log area, electricity, and indicators for good sanitation, good roof, and multi-story. We predict this value in our three TSCP cities (Tanga, Mbeya, and Mwanza).
Connected to water mains	Indicator for good water supply (metered/mains as opposed to bore-hole; stand tap; river; rain; water trucks; or other/none).
Road access	Indicator for access to tarmac; gravel; or earth road.

Note: this table describes the variables the we derived from TSCP building data.

Table 4.A16: Hedonic housing value regressions using TSCP survey data

	(1) ln value
Log building footprint area	0.797 (0.019)
Connected to electricity	0.235 (0.040)
Sewerage or septic tank	0.524 (0.041)
Good roof	0.0474 (0.090)
Multiple storeys	-0.0359 (0.178)
Intercept	13.11 (0.221)
<i>Observations</i>	3,663
<i>R</i> ²	0.416

This table reports estimates from a hedonic regression with buildings as units of observation using property values of 3,663 buildings in Arusha. The dependent variable is property value. This sample is selected because these buildings had both valuation data and data from the TSCP survey. Regressors are the buildings' log area, electricity, and indicators for good sanitation, good roof, and multi-storey. We then use the coefficient estimates to construct measures of hedonic values, as we explain in the Data Appendix.

Table 4.A17: Description of variables from Tanzanian census 2012

Variable label	Definition
Years of schooling	How many years of schooling the adult respondent has obtained. Missing values in the microdata are coded as 0 since there was no category for "Never attended school", and since the missing values were found to match reasonably well with the proportion of people with no schooling in the IPUMS 2012 Tanzanian Census data (which does not, however, have low level geographical identifiers). Moreover, the proportion of missing values in the microdata increased with age and with gender and age, which corresponds to the pattern of people lacking any schooling in Tanzania. Respondents with Training after primary school/Pre-secondary school or Training after secondary school are coded as 8 or 12 years respectively, i.e. one more year than primary or secondary schooling. Respondents with university education, are coded as 15, i.e. one more year than the maximum number of secondary schooling.
Exactly primary school	Binary indicator that takes the value 1 if the adult respondent has completed exactly 7 years of schooling, 0 otherwise. Missing values coded as 0 as in the variable above.
More than primary school	Binary indicator that takes the value 1 if the respondent has completed more than 7 years of schooling, 0 otherwise. Missing values coded as 0 as in the variables above.
Literate in any language	Binary indicator that takes the value 1 if the adult respondent is literate in any language.
Literate Swahili	Binary indicator that takes the value 1 if the adult respondent is literate in Swahili.
Literate English	Binary indicator that takes the value 1 if the adult respondent is literate in English.

Note: this table describes the variables we derived from the Tanzanian Census 2012 microdata.

4.B Data appendix

This data appendix is organized as follows. We begin by describing the Sites and Services projects, the nature of the treatment, selection of the treated areas, and how the de novo plots were allocated. We then explain how we measure the treatment and control areas in the seven cities. We then describe the three main datasets: the first comes from imagery data; the second from the Tanzania Strategic Cities Project Survey (TSCP, World Bank 2013); and the third comes from 2012 Tanzanian census micro data. Finally, we discuss

other auxiliary datasets, including: geographic variables; additional census data; land values data; data on project costs; and population data for 2002. Finally, we explain how we make currency conversions.

4.B.1 Project background and treatment

Background

The Sites and Services projects were implemented in seven Tanzanian cities. The projects treated 12 de novo areas (greenfield investments) and 12 slum upgrading areas (involving the upgrading of squatter settlements). The projects were rolled out in two rounds. The first round was implemented from 1974-1977, with infrastructure construction taking place in 1975-1976; and the second round was implemented from 1977-1984, with infrastructure construction taking place from 1980-1984. In the First Round, the World Bank treated the northwest of Dar es Salaam (Kinondoni district) and Mbeya with both de novo and upgrading and Mwanza with de novo investment only. In the Second Round the two types of treatment took place in the southeast of Dar es Salaam (Temeke district), Tanga, Tabora, Morogoro and Iringa. The number of de novo and upgrading plot surveyed in each round is reported in Table A3.⁵⁸ Details of the projects are discussed in World Bank (1974a,b, 1977a,b, 1984, and 1987).

Sites and Services projects in Tanzania fell into two broad classes. The first involved de novo development of previously unpopulated areas. The second involved upgrading of pre-existing squatter settlements (sometimes referred to as “slum upgrading”).

We provide a more detailed breakdown of the project costs below, but we note that among the infrastructure costs, the two main components were roads and water mains, and the cost of surveying the plots formal de novo plots was also important. Other investments, which covered public buildings (schools, clinics, and markets) were minor parts of the overall scheme.⁵⁹ It is also unlikely that access to these services ends discontinuously at the program boundaries, so our regression discontinuity design should mitigate any effect from such services. The indirect costs of the project mainly consisted of loans, which we discuss below. Taken together, it seems that roads, water mains, and plot surveys were the most relevant elements of the program. The roads and water mains were implemented

⁵⁸An additional upgrade was planned for the area Hanna Nassif in Dar es Salaam, but it was not implemented as part of Sites and Services. This area was nevertheless upgraded later on in a separate intervention (Lupala et al. 1997), but it is excluded from our analysis. Two additional areas, Mbagala and Tabata, were considered for the Second Round of Sites and Services, but it appears that they were eventually excluded from the project (World Bank 1987 and Kironde 2017).

⁵⁹The first round buildings public buildings were also surrounded by street lighting.

in both de novo and upgrading, but the formal plots were only implemented in de novo areas.⁶⁰

In addition to the three elements discussed above, both de novo and upgrading areas received a small number of public buildings, which were designated as schools, health clinics, and markets. While these could have had an impact, we think that they matter less than the plots, the roads and the water. First, the total cost of the public buildings was lower than both the roads and the water mains (separately); and second, even if Sites and Services received more buildings than other areas, there is no evidence that access to those facilities ended discontinuously at the project boundaries, which is relevant for the empirical strategy that we explain below. And some Sites and Services residents were offered loans, which were not fully repaid. We think of these loans as relaxing some owners' budget constraints, so we explain in the main body of the paper our strategy for addressing this channel.

The control areas (see more details below) mostly developed in an informal way. We have traced back the history of the control areas near de novo using various reports, at least for Dar es Salaam, whose urban evolution seems better documented. For example, according to the 1968 Dar Masterplan (Project Planning Associates Ltd. 1968) the De-novo control areas appear to be "Vacant land and land used for agriculture", and according to the 1979 Dar Masterplan (Marshall, Macklin, Monaghan Ltd. 1979), the de-novo areas are not indicated as being squatted; by the late 1980s, however, it seems that all the control areas have some unplanned sections (Kironde 1994). Finally, the Transport Policy and System Development Master Plan (Dar es Salaam City Council 2008) in Dar indicates all de-novo, control and upgrading areas as "built up" by 1992. But we note that our data gives a more disaggregated picture on the extent of built up area, and it appears that at this more fine-grained resolution not all the control areas were built up.

For the six secondary cities in which Sites and Services projects were implemented, we have not found evidence that any parts of the control areas were made formal under any planning scheme. In Dar es Salaam, however, Kironde (1994) documents that one planning scheme (Mbezi Beach) took place after the Sites and Services project. While we do not have precise maps, looking at present-day neighborhood boundaries this planned area may overlap with around 10-15% of our control area in Dar es Salaam. For the Mbezi scheme it seems that there was very little, if any, government provision of infrastructure, at least

⁶⁰The Second Round investments were generally lower, and for one of the de novo areas (the one in Tanga), we have some uncertainty as to the extent of infrastructure that was actually provided (World Bank 1987).

in the initial stages. As we discuss in the paper, however, eventually it seems that some investments in water mains and roads were made, but these were modest at best.

Treatment and control areas

We use a variety of historical maps and imagery from satellites and aerial photographs to define the exact boundaries of treatment and control areas. For Dar es Salaam, Iringa, Tabora, Tanga, and Morogoro, the World Bank Project Appraisals (World Bank, 1974a and World Bank, 1977b) provide maps with resolutions from 1:10,000 to 1:30,000 of the planned boundaries of the upgrading and de novo sites. In Dar, two maps were available, from 1974 and 1977, differing slightly for Mikocheni area. For all the areas except Tandika and Mtoni, we used the 1974 map, which appeared more precise. However, for Tandika and Mtoni we had to use the 1977 map, since these areas were not covered by the 1974 map.

For the two remaining cities, Mbeya and Mwanza, the maps from the project appraisal were unavailable. Therefore, for Mbeya we asked three experts to draw the boundaries of treatment. These experts were Anna Mtani and Shaoban Sheuya from Ardhi University, who both worked on the first round of Sites and Services project, and Amulike Mahenge from the Ministry of Land, who was the Municipal Director in Mbeya.

To delineate the treatment areas in Mwanza we obtained from the city municipality cadastral maps, dating back to 1973, at a resolution of 1:2,500. Since in Mwanza the treatment included only de novo plots, the cadastral map was sufficient to get the information for the intended treatment areas. We define the treatment area as covering the numbered plots that were of a size that (approximately) fitted the project descriptions (288 square meters); we also include public buildings into the treatment areas, to be consistent with the procedure in other cities. This procedure gives us a comprehensive picture of the twelve de novo and twelve upgrading neighborhoods across all seven cities.

To define our control areas, along with the historical World Bank maps from the Appraisal reports (World Bank, 1974a and World Bank, 1977b), we use historical topographic maps, and satellite and aerial images taken just before the dates of the treatment. We assign all undeveloped ("greenfield") land within 500 meters of any treatment border to our set of control areas. However, as we explain in more detail below, we exclude areas that were either designated for non-residential use, or that were developed prior to treatment, or that are uninhabitable. Our rationale for looking at greenfield areas as controls because we want a clear counterfactual for the de novo areas. We have no "natural" counterfactual

for the upgraded squatter areas, because we do not observe untreated squatter areas in the vicinity. The 500 meter cut-off reduces the risk of substantial heterogeneity in locational fundamentals. As part of our analysis we also focus on areas that are even closer to the boundaries between areas.

In order to know what had been previously developed, we used historical maps or imagery as close in time to the treatment date as we could find. We used all planned treatment maps. These include the 1974 and 1977 maps for Dar es Salaam and the 1977 maps for Morogoro, Iringa, Tanga and Tabora (World Bank 1987); the 1973 cadastral map of Mwanza (Mwanza City Municipality, 1973); satellite images from 1966 (United States Geological Survey 2015); aerial imagery from 1978 for Tabora and topographic maps from 1967, 1974, and 1978 for Tabora, Iringa and Morogoro (Directorate of Overseas Surveys 2015).⁶¹ All areas (with some minor exceptions described below) were covered by at least one source. Satellite images and maps also confirm that the areas designated as de novo were indeed unbuilt before the Sites and Services program was implemented.

We use all these data to determine which areas within 500 meter of Sites and Services areas to exclude from our baseline control group. Our rules for exclusion from the control areas are as follows. First, we exclude areas that were planned for non-residential use. These were indicated on the planned treatment map for industrial or governmental use. Second, we exclude areas that were developed before the Sites and Services projects began. These were either indicated as houses or industrial areas on topographic maps, or visibly built in the historical satellite images. Third, we exclude uninhabitable areas, for example, those off the coast. Finally, in the case of Mwanza (where we had to infer the treatment areas) we applied additional criteria for exclusion. In this case we exclude large numbered plots and all unnumbered plots, which do not seem to fit the description of de novo plots. We also exclude areas where the treatment areas are truncated at the edge, since we do not know where the exact boundary of treatment is. In this case we drew rectangles perpendicular to the map edge where the treatment area is truncated, and exclude the area within them.⁶² Further details on defining exclusion areas in each city are outlined in Table A13.

It is possible that some of the areas that were unbuilt in 1966 were built up from 1966 until the start of Sites and Services. But from partial evidence on construction dates in the TSCP data for two cities - Mbeya and Mwanza - it seems that only a very small share

⁶¹The resolutions of these maps range from 1:2,500 to 1:50,000.

⁶²We include in the baseline control areas (minor) areas where there is no pre-treatment data, because they are very sparse and are located near other empty areas.

(about 1.3 percent) of the buildings with construction dates in control areas near de novo were built before 1974.

Our treatment maps (Figure A1) show upgrading, de novo and control areas, as well as excluded areas. Moreover, with these appropriately defined control areas net of excluded locations, we can analyze present day outcomes using boundaries between control areas and de novo areas, and between control areas and upgrading areas.

We also note that for some of the analysis using the TSCP survey data (more below) we also used data further than 500 meters from Sites and Services. For the three Sites and Services cities with TSCP data (Mbeya, Mwanza, and Tanga), we used imagery from 1966 to exclude areas that were built up at the time.

Allocation of de novo plots

Plots were allocated to beneficiaries according whose i) houses were demolished in the upgrading areas ii) income was in the range of 400-1000 Tanzanian shilling (Tsh) a month. The income range was meant to target the 20th-60th percentiles of countrywide incomes (Kironde, 1991). According to project completion reports (World Bank 1984 and World Bank, 1987), between 50% and 70% of all project beneficiaries belonged to the target population. There was some evidence (World Bank, 1987) that a number of more affluent individuals obtained some of the plots after they had not been developed by initial beneficiaries.

4.B.2 Outcome variables derived from imagery data

A summary of the outcome variables we construct using the imagery data can be found in Table A14. Here we provide more detail on some of the key variables.

Buildings

To study the quality of housing we use Worldview satellite images (DigitalGlobe 2016), which provide greyscale data at resolution of approximately 0.5 meters along with multispectral data at a resolution of approximately 2.5 meters.⁶³ We employed a company (Ramani Geosystems) to trace out the building footprints from these data for six of the

⁶³The images were taken at different dates: Iringa (2013), Mbeya (2014), Morogoro (2012), Mwanza (2014), Tabora (2011), Tanga (2012) and there are two separate images for two districts in Dar es Salaam: Kinondoni (2015) and Temeke(2014)

seven cities. For the final city, Dar es Salaam, we used building outlines from a different, freely available, source - Dar Ramani Huria (2016).⁶⁴

We derive the following indicators of building quality using the building outlines: the logarithm of building footprint, building orientation relative to its neighbors and, finally the distance to the nearest road using ArcGIS tools.

For block outcomes we average each measure and indicator to get averages and shares. To do that, we begin with an arbitrary grid of 50 x 50 meter blocks. If a block is divided between de novo, upgrading, and control areas, we attribute the block to the area where its centroid lies. Finally, we match into each block the buildings whose centroids fall within it. This allows us to additionally measure three variables: the share of built up area in the block, the count of buildings in a block and whether the block is empty.

Roofs

To study the quality of roofs, we use the same Worldview satellite images as we did for the building outcomes above. Our aim was to separate painted roofs (which are less prone to rust) from unpainted tin roofs (rusted or not), in order to get a measure for roof quality that captures more variation than the TSCP survey indicator for good quality roofs. The cut-off between painted and unpainted roofs was chosen also because we had evidence from our initial field investigation that the painted roofs are considerably more expensive.

To this end, we create an algorithm through which ArcGIS and Python can separate painted from unpainted roofs for each satellite image of the seven Sites and Services cities. Before running the algorithm, we created unique color bins which would identify each type of roof material. These bins are three-dimensional sections of the red-green-blue space that correspond to different colors, which we think of as either painted roofs (e.g. painted red, green, or blue⁶⁵) or unpainted ones (e.g. tin, rusted, and bright tin⁶⁶). We defined the bins through a process of sampling pixels from each roof material type, identifying the color bins to which the pixels belong, and iteratively narrowing the bins for each roof type until

⁶⁴We have checked a sample of buildings traced out from the imagery data to the buildings in the TSCP survey data. Incidence of splitting or merging of buildings are fairly rare, occurring around 10 percent of the time, and more so in slum areas. This may also be in part due to a gap of a few years between the datasets. Therefore splitting or merging of buildings does not seem like much of a problem, especially when we focus on de novo areas.

⁶⁵Apart from red, green and blue we also had a bin for brown painted roofs in Kinondoni, since only in that image we noticed a large number of painted roofs that had a brown color, either due to image particularities or geographically varying preferences for brown painted roofs.

⁶⁶In Iringa and Mwanza we did not have the category bright tin since the particularities of the image or the conditions of the day when the image was taken resulted in other roofs than tin also being very bright in these cities.

they were mutually exclusive. Since each satellite image was slightly different in terms of sharpness, brightness and saturation, we sampled pixels from each image and created city-specific bands.

The algorithm is then applied to each city with its unique color bins. The algorithm works by reading the values of the color spectrum for red, green and blue of each pixel of a roof, and comparing these values to the above-mentioned unique bands of the color spectrum identifying painted, rusted and tin roofs. We assign to each roof the color bin that contains the plurality of pixels, and this indicates whether we classify it as a painted roof or not.

Roads

For all seven cities we used road data from Openstreetmap (2017). We had to clean these data in some locations using ArcGIS and Python, so that we only use roads that seem wide enough for a single car to pass through (we eliminated "roads" between buildings that were less than one meter apart). Following this automated procedure, we cleaned the road data manually to identify roads that appear passable to a single car.

4.B.3 Tanzanian Strategic Cities Project survey data

For three cities, Mbeya (in southwest Tanzania), Tanga (in northeast Tanzania), and Mwanza (in northwest Tanzania) we have detailed building-level data from the Tanzanian Strategic Cities Project (TSCP) which is a World Bank project implemented by the Prime Minister's Office of Regional Administration and Local Government (World Bank 2010). These surveys were carried out by the Tanzanian government from 2010-2013. We use these data to build a more detailed picture of building quality in the areas we study. Table A15 summarizes the key outcome variables that we derive from the TSCP data. Here we explain in more detail some of the issues relating to the dataset and how we use it.

The data arrived in raw format, with multiple duplicated records of each building and unit and many of these duplicate observations with missing data. We used the following rules to identify the unique observations. Buildings are identified by 'Building Reference Numbers' (BRN) and building units by BRN-units.

Rules for excluding buildings

1. Drop exact duplicates. i.e. if multiple buildings have all the same variables (including IDs) only keep one of them (dropped 1,202,669 observations).
2. Of all remaining observations with a duplicate BRN, drop all where all 'variables of

- interest’ are missing. Variables of interest are an extensive list and comprise much more than what is used in the analysis of this paper (dropped 166,131 observations).
3. Of all remaining observations with a duplicate BRN, keep the observations with strictly more non-missing variables of interest (dropped 12,842 observations).
 4. Of all remaining observations with a duplicate BRN, rank by ‘information provider’ and keep the observations with a strictly higher rank (dropped 15,486 observations).
 5. Of all remaining observations with a duplicate BRN, for a set of observations with the same BRN, replace with missing all variables where the records are inconsistent. For example, if there are two observations with the same BRN and both have ‘2’ for number of stories there is no inconsistency. But if one has ‘1’ number of rooms while the other has ‘2’: replace the number of rooms with missing for both.
 6. Of all remaining observations with a duplicate BRN all duplicate BRNs will have exactly the same records, keep only one record for each BRN (dropped 27,483 observations).
 7. There are no longer any duplicate BRNs. We drop 35,912 unique buildings from the records that do not match a building in one of the city shapefiles of building footprints.
 8. We drop 38,180 buildings from the records that are coded as outbuildings.
 9. We drop 596 buildings that do not match to a unit.
 10. Finally, we are left with 119,914 buildings all with at least one corresponding unit.

Rules for excluding building units

1. Drop exact duplicates, for example, if multiple units have all the same variables (including IDs) only keep one of them (dropped 1,288,430 observations).
2. Of all remaining observations with a duplicate BRN-unit, drop all where all variables of interest are missing. Variables of interest are an extensive list and comprise much more than what is used in the analysis of this paper (dropped 221,134 observations).
3. Of all remaining observations with a duplicate BRN-unit, keep the observations with strictly more non-missing variables of interest (dropped 6,383 observations)
4. Of all remaining observations with a duplicate BRN-unit, for a set of observations with the same BRN-unit, replace with missing all variables with mismatched records within the set. i.e. if there are two observations with the same BRN-unit and both have ‘2’ for number of toilets: do nothing, if one has ‘1’ number of rooms while the other has ‘2’: replace the number of rooms with missing for both.

5. There are no longer any duplicate BRN-units. We drop 32,322 units from the records that do not match a building in one of the city shapefiles of building footprints.
6. We drop 3,216 units from the records that are coded as outbuildings.
7. We do not need to drop any more units, since all remaining units match to a building.
8. Finally, we are left with 154,734 units all with a corresponding building.

From the building data set we exclude all buildings categorized as “Outbuildings” (sheds, garages, and animal pens). This leaves us with a sample of buildings that are used mostly for residential purposes, although a small fraction also serve commercial or public uses.

For these buildings in analysis we use the logarithm of building footprint; connection to electricity; connection to water mains; having at least basic sanitation (usually a septic tank and in rare cases sewerage); having good (durable) roof materials; having more than one story; and having road access.

Hedonic values

To calculate hedonic building values we use an auxiliary TSCP dataset covering 57,136 buildings from Arusha, which is not one of the seven Sites and Services cities, but is the only one for which we have valuation data at the level of individual buildings. Specifically, we have valuations for 6,837 buildings. The buildings for which we have valuations are concentrated near the city center.

The intention of the valuations is to determine the rateable value (annual rental value of a property) of each property as a basis for collecting property tax. This is estimated by professional valuers under a set of formal guidelines. The valuer is given building-level characteristics, a photograph of the property, and where possible, property transaction records (see figure below). The valuer uses these inputs along with a standard set of guidelines that give bounds on how much each characteristic of the building is worth, but ultimately makes a subjective valuation of the property based on the information provided.

Of the valued buildings, 3,663 also have building-level characteristics (log area, electricity, and indicators for good sanitation, good roof, and multi-story) from the TSCP survey. We use these to perform hedonic regressions and make out-of-sample predictions of the valuations in the three TSCP cities (Tanga, Mbeya, and Mwanza) where Sites and Services was implemented. For buildings in our out-of-sample prediction that are missing some, but not all, characteristics we fill these missing values with the average of their respective characteristic. Consequently, 6 percent of the buildings with hedonic values in our TSCP dataset have had missing data filled for at least one of their characteristics.

The results of the hedonic regressions are shown in Table A16.⁶⁷ Buildings with larger footprints, electricity connection, and some sanitation, have higher hedonic values; conditional on these factors, roof materials and multistory buildings are uncorrelated with value, perhaps due to the sample size.

Construction dates

For two cities (Mbeya and Mwanza) we have building dates for less than 10 percent of the housing units in the de-novo and control areas within 500 meters. In absolute terms, this means we have construction dates for 215 de novo units and 300 control units close to the boundary. In both cities the de-novo areas were part of Round 1, so the infrastructure was built from 1975-76, and for both we have pre-treatment imagery from 1966. According to the TSCP data, the fraction of units that existed as of 2013 that were built before 1975 was 0.5 percent in de-novo and (1 of the 215 units with construction dates) 1.3 percent (4 of the 300 units with construction dates) in control areas close to the boundary. Admittedly these data are imperfect, and some buildings may have been replaced over time, but the data do not suggest that old buildings that pre-date the Sites and Services are a major concern.

4.B.4 Geographic control variables

Distance to shore and rivers and streams indicators

We use as geographic controls the distance in kilometers to the nearest shore (either the Indian Ocean or Lake Victoria) and an indicator for rivers or streams.⁶⁸ These variables are derived from Openstreetmap - we use current data since historical data are unavailable. We consider proximity to the coast an amenity, while rivers or streams may be an amenity if their water is usable, or a disamenity if they increase flood risk.

Ruggedness

Ruggedness is calculated using SRTM elevation at a horizontal resolution of 1 arc-second (United States Geological Survey 2000). We use those data to compute the standard deviation of elevation of each 50m X 50m block relative to its eight neighbors.⁶⁹ We again use current data since historical data are unavailable.

⁶⁷We follow Giglio et al. (2014) in including observable characteristics linearly in a hedonic regression.

⁶⁸The distance to the shore is winsorized at 10 kilometers, hence the distances to other water bodies, such as Lake Tanganika, are irrelevant in our seven cities.

⁶⁹For a small fraction of blocks that are at the border of our study area, we instead use the mean of the standard deviation for those blocks for which it is calculated.

Distance to historical CBD

For some of the robustness analysis we use measures of distance to historical CBDs, to mitigate concerns that our main measure of the CBD may be endogenous to Sites and Services. To construct these measures we use data on the location of railway stations in six of the cities, since these stations' locations were generally determined before the onset of Sites and Services, as we discuss below. Iringa does not have a railway station, so the coordinates of the Iringa municipal office were used instead. We then calculate distance in kilometers to these coordinates in the same way as we do with the light-based CBDs and then use this as an alternative measure in some regression specifications.

To justify our argument that railroad stations existed even before Sites and Services, and hence can be used as *ex ante* markers of the centers of the cities, we refer to a map of the railways from 1948, which shows that five of the seven cities had railways in 1948, and the location of railway stations is unlikely to be moved.⁷⁰ Of the remaining two cities, Mbeya's railway was built from 1970 and completed and opened in 1975 (Edson 1978), while Iringa does not have railway, as mentioned above.

4.B.5 Land values

Matching land value data to enumeration areas

We obtained an Excel sheet titled "RATES LAND VALUE MIKOA 10 2012.xls", which we received from the Kinondoni Municipal council, but were told that it was created by the Ministry of Lands, with minimum, mean, and maximum land values for different neighborhoods in Tanzania. We can identify these neighborhoods by four string identifiers: region, district, location, and streets. To locate neighborhoods we match them based on the 2002 enumeration area (EA) shapefile, which contains string identifiers for region, district, location, and `vill_stree` (we consider 'vill_stree' comparable with 'streets' from the land values table).

Land use

The Excel table has different minimum, mean, and maximum land values by land use. There are typically four categories: Residential, commercial, commercial/residential, and institutional. Though the differentiation of land values across uses is mechanical (commercial is 1.4* res, com/res is 1.1*res, institutional is the same as res), the variation across areas

⁷⁰Britishempire.co.uk. (1948). [online] Available at: <https://www.britishempire.co.uk/images2/tanganyikamap1948.jpg> [Accessed 3 Jul. 2019].

is not mechanical. Throughout we use mean land values from the residential categories only.

Spatially mapping land values

We merge EA boundaries to land value observations using the four identifiers: region, district, location, and streets. Each entry in the land value table we treat as an observation, often this contains a group of ‘streets’. Typically there are many EAs per land value observation, so each observation in the land values table is matched to a large group of EA boundaries. Then we dissolve the EA boundaries to have a single spatial unit for each entry in the land value sheet. We then plot the mean residential land rate for each spatial unit.

Results

The merged areas are quite large. Some roughly match our treatment areas:

1. Sinza – one unit at 240,000TSh
2. Manzese A – three partial units all at 65,000TSh
3. Manzese B – split in half, one at 65,000TSh the other at 50,000TSh
4. Kijitonyama – one unit at 325,000TSh

The other two do not match as well:

1. Mikocheni – contained by a much larger unit at 125,000TSh
2. Tandika/Mtoni – overlaps many areas of values; 40,000TSh, 30,000TSh, 50,000TSh, and 18,000TSh

These values per square meter put us in the range of 125,000-325,000 TSh (2017 US\$80-220) in de novo and 18,000-65,000 TSh (2017 US\$10-40) in upgrading. For the areas where we have better matched data the ranges are 240,000-325,000 TSh (2017 US\$160-220) in de novo and 50,000-65,000 TSh (2017 US\$30-40) in upgrading.

4.B.6 Project costs

The total cost of First Round of Sites and Services was \$60m in US\$2017, of which just over half was due to direct costs (World Bank 1984): infrastructure (38% of total costs), consultants (9%), land compensation (6%). Other costs (45%) included the community centers (14%), mentioned above, and a loan scheme (29%), which later failed because of poor repayment rates, and a few other costs. This investment covered a total of 23,161

plots: 8,527 de novo plots and 14,634 upgrading plots. The Second Round of Sites and Services cost \$70m in US\$2017 where 70% was spent on direct costs, paying for a total of 22,106 plots: 1,978 de novo plots and 20,128 upgrading plots (World Bank 1987).

The First Round project reports (World Bank 1974a and 1984) indicate that the total infrastructure investment costs per area in de novo and upgrading were very similar. The project report for Round 1 provided costs separately for de novo and upgrading areas (World Bank, 1984). However only infrastructure investment differed for the two types of treatment, while land compensation, equipment, and consultancy costs were reported as split 50-50 between de novo and upgrading. Direct costs by treatment were \$19 million in de novo and \$15 million in upgrading areas (in US\$2017). To get costs per unit area we normalize by total area covered by each treatment type in Round 1 (8.5 square kilometers in de novo and 6.5 square kilometers in upgrading). This gave costs for de novo and upgrading areas of \$2.20 and \$2.37 per square meter respectively (in US\$2017).

Further, in order to compare with present day land values (per plot area) we would like an estimate of costs per unit of treated plot area. Due to data limitations we can only do that for de novo neighborhoods where the reports give both plot counts and plot areas. We estimate that the direct costs per square of plot were no more than \$8 per square meter, and total costs were no more than \$13 per square meter (in US\$2017).⁷¹

An alternative way to look at costs is to break them down by plot which we can do for both de novo and upgrading areas. According to the report there were 8,527 de novo plots and 14,634 upgrading plots in Round 1. We can divide the direct costs of de novo and upgrading areas by their plot counts to get \$2,200 and \$1,000 per plot respectively (in US\$2017). The difference in costs reflects both the larger size of the de novo plots and the larger share of allocated to public amenities (such as roads).

Cost recovery

Costs were meant to be recovered through land rent (4% of land value a year) and service charge (the cost of infrastructure provider), but assessment of parcels was long and interim charge well below the adequate amount to cover the costs (100 Tsh/year or 2017 US\$51) was imposed. Collection rates were low and not timely.

⁷¹To calculate the costs per square meter of each plot, we use the planned areas of de novo plots from Appraisal report 1 (World Bank, 1974a); the planned area was 288 square meters, except for 8.56% of the plots (those in Mikocheni) where it was 370 square meters. Taking the weighted average at 295 square meters, we can divide the de novo direct costs by total plot area treated to get \$7.5 per square meter.

4.B.7 Additional data

Outcomes in 2012 Tanzanian census micro data Extract

This extract was obtained through a contact from Tanzanian Census Bureau. Unlike the Tanzanian census data, which can be obtained online at IPUMS (2017), these data are at the level of individuals. We match these census observations from this extract to geographical areas using EA identifiers in the census extract. Using shapefiles of EAs (with the same identifiers) from the Tanzanian Census 2012, also obtained from the same contact, we match the census data observations to our treatment and control areas. The process of matching EAs to treatment areas (de novo, control and upgrading) was done through Python and ArcGIS.

In case an EA straddled two (or more) of the treatment and control areas, we cut that EA in ArcGIS into multiple parts, each part belongs to a treatment or a control area. We then use this information to remove the census data observations which belonged to EAs whose area inside a treatment and control area was less than 5% of the entire EA area. We also use the information on how large a part of the EA was inside a treatment or control area to create analytic weights (the weight is higher when the relevant overlap is higher) for some of the robustness checks.

Our variables are discussed in Table A17, and include years of schooling and indicators for different schooling thresholds (exactly primary and more than primary school education; the omitted category is less than primary school). We also create indicators for literacy in any language; literacy in Swahili; and literacy in English. We then calculate means of each of these variables across adults in each "cut" EA.

Population data for 2002

To calculate the population density in each of the neighborhoods, we use data on population by enumeration areas from the 2002 Tanzanian Census (Tanzania National Bureau of Statistics 2011). In cases where an entire enumeration area falls into a Sites and Services neighborhood, we assign its entire population to that neighborhood. When only a fraction of an enumeration area falls into a Sites and Services neighborhood, we assign to the neighborhood the fraction of the enumeration area population that corresponds to the fraction of the land area that lies within the neighborhood. The mean number of enumeration

areas matched to each neighborhood is 33 for de novo areas and 35 for upgrading areas.⁷² Population counts for 2002 are outlined in Table A3.

IPUMS 2012 Tanzanian census by region

We use data downloaded from the IPUMS online repository of country censuses, in order to check the correctness of the above-mentioned microdata extract from the same census. This was done in particular for the education variable which had been cleaned by IPUMS staff to include many observations recorded as having “never attended” school. The microdata that we had received directly from the Tanzanian Census Bureau had many missing values for the education variable, and none coded as never having attended school. The missing values in the micro-data followed the same pattern as the “never attended” in the IPUMS data, which contributed to our decision to code them as zero years of schooling. We also checked age and gender patterns in the microdata which confirmed our interpretation of the data.

Conversion to 2017 US dollars

All monetary values in the paper are reported in their source units and also converted to 2017 US dollars (2017 US\$). To calculate the dollar values we used the exchange rates to contemporaneous year US\$ from Penn World Tables 9.0 (Feenstra et al., 2015). Then we used the US CPI factors to bring the value to 2017 US\$.

⁷²We are unable to report the population counts from 2012 census, because we only have a sample from the census, and in this sample, not every 2012 enumeration area is populated.

Chapter 5

Measuring urban economic density

5.1 Introduction

At the heart of urban economics are the sources and nature of agglomeration economies, which drive the existence and the extent of cities. The extent of agglomeration economies help determine the role of urban pull factors in the urbanization process within and across regions of the world. Sub-Saharan Africa has had more rapid urban population growth than any other region over the last half century. Its urban population share grew from 14% in 1950 to 41% in 2015 and is predicted to be 60% in 2050 according to the UN (2015). This rapid urbanization is driven by forces pushing people out of rural areas and pulling them into cities as reviewed for Sub-Saharan Africa in Henderson & Kriticos (2018). Push factors include changes in agricultural technology which release workers from land (e.g. Matsuyama (1992), Gollin et al. (2007), and Bustos et al. (2016)) as well as adverse climate and conflict in rural areas (e.g., Henderson et al. (2017), Fay & Opal (1999), Barrios et al. (2006), and Brückner (2012)). With some exceptions such as Gollin et al. (2013), the literature on Sub-Saharan Africa has focused on push factors as driving urbanization. Here, like Gollin et al. (2013), we argue that the income gains associated with moving to cities from rural areas are high in Sub-Saharan Africa, pulling people into cities, albeit in the face of poor productivity in the rural sector. We also show that, within the urban sector, there are huge gains to moving to denser agglomerations.

There is a vast literature that has traditionally focused on developed countries - such as the USA - which estimates the productivity or wage premiums from being in cities compared to rural areas, or from being in bigger versus smaller, or denser versus less dense cities (e.g. Ciccone & Hall (1996), Glaeser & Mare (2001)), with reviews of the literature in Combes & Gobillon (2015) and Rosenthal & Strange (2004). The literature argues that premiums exist because agglomeration gives proximity among firms, workers and people, which allows economic agents to economize on local trade costs, spread information and ideas more

easily, diversify the range of products produced, and access larger pools of workers and jobs (Duranton 2015). Recently, there has been more work on developing countries with, for example, papers on Latin America (Quintero & Roberts 2018), Colombia (Duranton 2016), China (Combes et al. 2019), and a comparison of Brazil, China, India and the USA by Chauvin et al. (2017).

Wage premiums in developing countries may be larger because their small informal sector firms and poorly educated entrepreneurs and workers are more reliant on their external environment, so that the agglomeration factors and externalities noted above may play a bigger role than in developed countries. Duranton (2015) argues that returns to density will be larger in developing countries. That said, other papers suggest that for Sub-Saharan Africa, cities have low ‘economic density’ or proximity, resulting in lowered spillovers (Collier & Jones (2016), Venables (2018), and Lall et al. (2017)) and suppressing gains from urbanization. Low economic density arises, for example, if economic activity is not very clustered but scattered throughout the city, potentially because of the high costs of commuting within cities inducing firms to locate nearer to residents as in Fujita & Ogawa (1982), with corresponding empirical work in Heblich et al. (2018). In Chapter 3, I examine whether large cities in Sub-Saharan Africa have a lower degree of density and clustering of economic activity relative to the rest of the world. Using LandScan (2012) data, I find that these cities in fact generally have higher density and clustering measures than Latin American cities, and measures that are comparable to Asian cities. That said there is enormous variation in density across Sub-Saharan African cities of similar populations and these density differences drive huge wage and income differences.

Testing the gains from proximity and agglomeration in Sub-Saharan Africa, or elsewhere, involves three critical aspects of measurement, hitherto under-explored in the previous literature. First is how cities are defined, with an eye to establishing consistency in how densities are measured across cities, where density will be a key measure of agglomeration. The main problem concerns the extent of the land area over which a city is defined. The land area and its corresponding population are typically defined by administrative boundaries, rather than economic ‘boundaries’ related to density. More specifically, the definitions of cities and urbanized areas chosen by national statistics bureaus are typically based on pre-determined administrative units which are defined as either urban or not, based on qualitative aspects of land use and the built environment, such as the degree of centrality of activity, total population and the like. In some countries, population density plays a role in the definition but typically not a central one. Contiguous urban administrative units are then aggregated into a metropolitan or urban area. If land areas are

based on administrative units which have populated and unpopulated parts and inconsistently defined across countries or regions within countries, then density measures will fail to capture the *de facto* economic density at which a city's activity operates. Put differently there can be large measurement error leading to attenuation bias, as we will demonstrate.

The second measurement aspect is to determine what specific population and density related measures best capture agglomeration forces, or best capture what we term economic density, the forces driving income and wage premiums. Most studies adopt specific, simple measures of agglomeration such as indicator variables for settlement type (e.g. urban versus rural), a continuous total population measure, or at best, a basic population density measure, where as noted the denominator (effective land area) is typically ill-defined across cities. Studies usually pick one specific measure with little analysis as to why and no quantitative evaluation of what measure(s) would be most relevant.

The third measurement aspect concerns outcomes. Outcomes in the literature are specified either as firm productivity or individual wages, but not household incomes. For rapidly urbanizing regions, in the long term it is more households that ultimately move to cities. If household gains from being in urban areas differ from wage rate gains, that is a critical factor in studying urbanization.

This paper differs in each of these three dimensions. A novel part of our work will be to have consistent city boundaries based on fine scale density measurements, which will reduce measurement error and provide really nice results. We will consistently define metropolitan areas across countries, based on population and population density thresholds. Second, we will focus on an evaluation of different agglomeration aspects starting with traditional options: urban versus rural, a continuous total population measure, and a continuous measure of overall population density. However, these measures do not capture, for a given population and a given overall density, the degree to which economic activity is clustered within a city. Two cities with the same density and population may have very different levels of clustering of economic activity within the city, and they can be captured by other measures reflecting variation in population densities within a city (De La Roca & Puga (2017) and Collier et al. (2018)). We derive and utilize several such measures. Third, we will focus not just on wage rates, but, as a novel feature, on household incomes. Agglomeration effects will be much higher for household incomes and we will explore reasons why.

Specifically, the paper will examine how well different economic density measures explain household income and wage differentials across space for a set of six Sub-Saharan Africa countries whose total population is about 430 million. Our sample covers the rural sector,

193 low-density urban settlements with over 5,000 people, and 120 high-density cities with over 50,000 people. Using populations at the 1 km grid square level, we aggregate contiguous squares of high density to create cities – which are the consolidation of an urban core and a surrounding lower density fringe – and aggregate contiguous squares of lower density places to create stand-alone, low density settlements, which from now on we will call towns. While admittedly the density and population thresholds we choose below are to some degree arbitrary, they are based on types of thresholds some countries and researchers cite, or use in the case of OECD (2012).

For the measurement of agglomeration economies, there is the issue of determining relevant spatial scales. For example, does it matter whether you are just in a particular city or what part of the city you are in? First we examine how marginal scale effects, or elasticities, vary across the spatial hierarchy: rural areas, towns, and cities. Then for cities, while most papers explore agglomeration benefits at the level of a city or county, a few papers look within cities at the extent of spatial decay, finding a very rapid spatial decay of certain types of scale externalities (Arzaghi & Henderson 2008), Rosenthal & Strange (2008). Here we explore both city level and neighborhood effects together: those from overall city economic density and from own local neighborhood economic density. We will ask also if people living in the urban core versus fringe of the city benefit differently from overall density characteristics of the city.

In the paper, Section 1 starts with how we define cities and what are the advantages and disadvantages of the LandScan (2012) data set that we use. Section 2 defines various measures of economic density for urbanized and rural areas, decomposing them into first and second moment components. Section 3 looks at the relationship between different measures of economic density and income differentials across the whole spatial hierarchy. Section 4 looks specifically at cities and examines issues such as the optimal rate of spatial discount for De La Roca & Puga (2017) measures of neighbor effects; the role of spatial variation in density within a city; how important local density measures are within cities, as well as location within the city; and what factors may underlie the huge income gains to families from moving to high density places. Section 5 concludes and discusses extensions.

5.2 Using Landscan data and defining urbanized areas

5.2.1 Landscan data

To analyze measures of economic density and clustering within the city, we need information on where people and workers are at a fine spatial resolution. And we need the most accurate data available. For countries like the USA, both finely gridded population and employment data are available from censuses. However, in most developing countries that is not the case. Population data may only be available at a coarse scale such as regional or local government administrative units and economic censuses in Sub-Saharan Africa are generally non-existent. In terms of population the key then is to allocate people from coarse administrative units to a spatial scale such as a square kilometer.

Our primary data source is LandScan (2012) from Oak Ridge National Laboratory in the USA, which is now being used in some research (e.g. Desmet et al. 2018). Oak Ridge takes population data from sources such as the US Bureau of the Census which assembles census data from other countries on as fine an administrative scale for each country as they can obtain, such as provinces, counties, parishes or even sub-counties. Landscan then creates a measure of an ambient population for approximately each 1km grid square on the planet (30"x30" arc-seconds). The ambient population is meant to represent where people are on average over the 24 hour day. To assess the ambient population, they appear to use nocturnal and diurnal population estimates for at least some areas of the globe, although these are not publicly accessible.

For Landscan, as for WorldPop¹, the Global Human Settlements Layer², and similar data sets, a key element in this process involves taking population numbers at some upper level of spatial scale and allocating people to fine grid squares based on where they are likely to live and possibly work. The typical standard in such work has been to allocate people on the basis of ground cover in a grid square from Landsat satellite imagery, or its enhanced versions. This is viewed as an improvement on the Gridded Population of the World, which simply smears population of an administrative unit uniformly across the 1 km grid squares in the unit.

Landscan has two key advantages and two key disadvantages over other data sets. First, Oak Ridge National Lab is more explicit in the fact that they are trying to estimate the ambient population with potentially nocturnal and diurnal populations; while in other

¹<http://www.worldpop.org.uk/data>

²<http://ghsl.jrc.ec.europa.eu/datasets.php>

algorithms this is implicit through the smearing of the population into general built cover, without workplace or residence distinction. Diurnal is meant to include where the employed population is found during the work day as well as where the non-working population is found; while nocturnal is primarily where people live and sleep. The second advantage of Landscan is that Oak Ridge National Lab has access to very high-resolution satellite data and a wealth of other information which *potentially* allows them to distinguish where built cover is likely to house employment versus residents versus perhaps even shoppers, as well as *potentially* to distinguish roads for commuting and even infer intensity of building on a spot (Rose & Bright 2014).

A key disadvantage in using Landscan is the complete lack of specificity and transparency as to what Oak Ridge researchers actually do; hence, our use of the word "potentially" in describing what they might do. The second disadvantage, which they acknowledge, is that Landscan data for different time periods are not comparable over time, presumably both because of differential availability of high-resolution data over time and increasingly sophisticated extraction of information from later satellite images.

Hopefully in the future, proposed data sets such as the High Resolution Human Settlement Layer (HRSL)³ or Modelling and Forecasting African Urban Population Patterns (MAUPP)⁴ – which also use very high spatial resolution data – will be able to cover a wider set of time periods and countries consistently with a more explicit methodology. Then one will be able to compare them with Landscan and related data sets and do more comprehensive ground-truthing exercises. Second while ambient population may be a relevant measure to use in characterizing economic density, one would like also to know about clustering and density of employment only. If the assignment of people to workplace buildings gets very sophisticated, we would be able to explore the use of employment density measures, as well as those of ambient population.

Since we will be looking within cities at ambient population density in order to measure the extent of clustering, we want to know if Landscan data are at least somewhat reliable for Sub-Saharan African cities where fine spatial data on population and employment location are generally not available and in many cases do not exist. In Appendix 5.C, we attempt to ground-truth Landscan measures at the grid square level for Kampala and Nairobi, two cities where we have fine spatial resolution data on population and employment which would be unavailable to Oak Ridge National Lab. We further attempt to replicate Land-

³<https://www.ciesin.columbia.edu/data/hrsl>

⁴<http://spell.ulb.be/project/maupp>

scan's ambient population measure by grid square using an assignment algorithm on our own data. We conclude that Landsat measures do well.

5.2.2 Defining urbanized areas

As noted earlier, the problem with typical definitions of urbanized areas from the United Nations or population censuses of different countries is that they employ country-specific city and town definitions incorporating administrative boundaries, which means there is no consistency across countries and economic land area is poorly measured. Second, most definitions of "urban" are based to a large extent on qualitative and subjective criteria. Some countries use an application or evaluation process to redefine rural administrative areas as urban, which tends to under-represent newer fast-growing agglomerations due to delays in application, evaluation, and granting of urban status.

To avoid the administrative boundary problem, researchers have taken several approaches. One is to define commuting zones, based on very fine spatial scale data in the USA or EU; such data does not exist in most developing countries. Another is to use night lights such as in Harari (2016) or Henderson & Kriticos (2018). Night lights provide contiguously lit polygons, in which the commuting zone may be included due to light from cars on the roads. Night light data can be used to track the evolution of cities throughout 1992-2013, albeit with some inconsistencies across the sensors. To define the extent of a city one must impose a lights threshold at the boundary with rural areas. Given blooming of lights over unlit areas and more general electrification of the rural population, the major drawback in using night lights data is that one must impose different light thresholds globally. In Sub-Saharan Africa to capture true but smaller urban areas, one can't impose a lights threshold (above 1 on a scale of 0-63) to define cities (see Henderson & Kriticos (2018)), but with no threshold a few unreasonably large urban areas can emerge. India with more widespread electrification requires a high threshold, so that many large separate cities do not merge into one (Harari 2016). Dingel et al. (2018) attempt to match a night-lights threshold to that of commuting flows in Brazil and apply it to other developing countries such as China. However, this approach works only if the degree of electrification is the same everywhere else as in Brazil, and the commuting flows are the same. It is interesting to note that Harari (2016) uses a threshold of 35 for India, while Dingel et al. (2019) use 30 for Brazil, a more developed country.

A third way to define boundaries of cities is to use the extent of built cover (Angel et al. (2016) and Earth Observation Center at the German Aerospace Center documented in Esch et al. (2018)) to name a few), using, typically, Landsat satellite data which is global.

Landsat data are based on low-resolution 30m-by-30m satellite data, which has enough bands to distinguish man-made (or impermeable surface) from natural cover (Angel et al. (2016)). As an alternative, the German Aerospace Center uses Radar satellite information to infer settlements from the elevation of the ground and structures. The two above-mentioned projects have provided for open use their boundaries of building extent for 1990, 2000 and 2014 in the case of the Marron Institute, and for 2011 in the case of the Earth Observation Center. However neither data provider attempts to calibrate the correct threshold for build cover to match the commuting flows. For example, Marron Institute just picks the threshold of 50% of 30mX30m pixels built within 1km radius.

Building on this third way, using GHSL data (OECD 2012), the European Union Commission (<https://ghsl.jrc.ec.europa.eu/CFS.php>) defines cities based on population density, where, in the GHSL, population from census administrative units is assigned to grid squares based on the degree of built cover. Then the Commission uses thresholds of population (density) of these grid squares to define city boundaries. We follow a similar procedure and employ a consistent density based definition across our African countries, using LandScan (2012) ambient population per grid square. We set density thresholds to define both cities and towns. To do so, we apply a smoothing algorithm as described in Appendix 5.B.1 so that each own grid square is assigned the average density of the 7 km x 7 km neighborhood in which it is the center. Smoothing avoids large doughnut holes in cities, due to terrain factors, airfields, parks, big open public spaces and the like. We define a core city as a set of contiguous grid squares all of which have a density greater than or equal to 1,500 per sq. km. and the total population of these contiguous squares must sum to 50,000 or more. In Sub-Saharan Africa with the geography, sprawl and disconnectedness in some cities (Baruah et al. 2017), a 7km x 7km neighborhood was the least generous smoothing window to eliminate all the 'holes'.⁵ Contiguity as discussed and stated in Appendix 5.B.1 is based on the 4 rook neighbors (N, S, E and W) with a shared border, rather than the 8 queen neighbors, 4 of which just touch the own cell. The area covered by these contiguous squares over 1,500 per sq. km defines what we call the core of the city. We then add in a fringe to each core, which includes all contiguous (rook) grid squares with population density over 500 per sq. km. The core combined with a fringe is called a city. For smaller urbanized places that are stand-alone, or towns, we take the collection of contiguous grid squares all with (a smoothed) population density over 500 per sq. km., which collectively sum to 5,000 or more.

In comparison, the European Union Commission defines core cities based on the 1500

⁵See de Bellefon et al. (2018) on similar issues defining cities with built cover data.

people per sq km and the 50,000 total population threshold. They then define towns and suburbs using 300 people per sq km and the 5000 total population cut-off. We think 300 people per sq km in some parts of the world covers very rural populations and in the discussion of Nairobi below we point out why, in defining metropolitan area suburbs, 300 is too low. For us it was more whether to cut at 500 or 750 in Sub-Saharan Africa. Also most critically for us, the EU program does not define metropolitan areas, only core cities. "Suburbs" are unassigned to core cities, which means we could not have metropolitan areas as a unit of analysis.

The process and impact of threshold decisions are illustrated in Figure 5.1 for Nairobi. Core city areas are shaded in the darkest grey (almost black), and overall cities are outlined in dark blue and have corresponding stripes. There are two cities in the figure, Machakos to the bottom right and Nairobi. Nairobi consists of the main core and three small core areas, essentially satellite towns now falling under the umbrella of Nairobi. The fringe of the city of Nairobi consists of two shaded areas: an area one grey tone lighter than the core (> 750 people per sq km) and an area two tones lighter (> 500 people per sq km). Our choice of 500 per sq. km is based on the idea that a lower threshold such as 300 per sq. km (the lightest grey) is too loose and extends too far into more rural and low-density towns much further north of Nairobi. And it would place the centroid of Nairobi well outside its true central core. A higher cutoff of, say, 750 people per sq. km. (one shade darker than the core) may be too stringent and exclude satellite cities around Nairobi that are very likely to be within the commuting zone. Obviously, other arguments about drawing boundaries can be made. In the figure we also outline in light yellow color the polygons (with yellow stripes) that make up the independent towns around Nairobi. Some are very spatially distinct, but some follow ribbons (roads) to the north outside Nairobi, where rural areas are interspersed with urbanized towns. In the figure everything in lightest grey or the Google Earth background is rural. In the Appendix 5.B.1 in Figure 5.B2 we do a version where we show the county administrative boundaries. Restricting Nairobi to be its own county would miss significant portions of population even within the core city. Adding the county to the north would add in much of the missing population but would also add in large tracks of land with low intensity of use and result in significant mis-measurement of economic land.

In Figure 5.2 we compare the distribution of cities by size in our data with those of the UN and those based on the GHSL. In this comparison the city numbers (179) and extents are based on Landscan data and our algorithm. We then populate these areas with Landscan data, GHSL data and UN data, with the last based on UN cities which fall within our

boundaries, noting that UN cities are cut at 300,000. Thus we are comparing how much population each source allocates to given area. In Figure 5.2, Panel A, to compare with the UN, we cut our cities at 300,000. The size distributions are pretty interwoven, although as Figure 2b reveals there are differences. Figure 5.2, Panel B, keeps the UN cut, but adds in all our Landscan cities, reducing the city size cut for Landscan and GHSL numbers to 50,000, to see what happens if we populate all our algorithm cities with Landscan versus the latest version of GHSL data. The interesting feature is the shift right in the Landscan distribution, implying that for cities in general, Landscan tends to put more ambient population into the same dense urban areas as compared to GHSL.

5.3 Defining economic density

What is the most meaningful measure of economic density? It should be a density measure capturing the most relevant aspects of proximity among people in a city. That is, the proximity that would be most relevant for agglomeration benefits, such as cheaper trade, improved hiring and learning, and a wider choice of inputs and products. A large city population may not always generate proximity gains. A city that occupies a large area may also have very low density and low proximity, compared to a more compact city with the same population. In the same way, a city's high average density may not mean there is very high proximity of activity within the city.

Figure 5.3 illustrates this last issue about proximity.⁶ All hypothetical cities in Figure 5.3 have the same total population (180) and average population density [PD] (5). City 1 has no clustering. Cities 2 and 3 have the same degree of within grid square clustering, with half the grid squares with no population and half with 10 people per grid. The 10 means greater within grid square possibilities for intersecting with others (the pairwise possibilities for meetings for example, $n(n-1)/2$, or 45 in this case). However city 2 allows for more possibilities for interactions with neighbors. Ignoring the boundaries in city 3, on average a grid square has 40 queen neighbors, while in city 2 a grid square has 80 queen neighbors in the surrounding 8 squares.

We now turn to two measures which reflect these differences relative to PD, personal population density [PPD] and a measure based on De La Roca & Puga (2017) which we label RPA in honor of the authors. For a given city area, relative to population density, personal population density is a weighted, rather than simple, sum of own cell population

⁶Figure 5.3 and the definition and decomposition of personal population density are borrowed from on-going work by Henderson, Storeygard and Weil. This joint intellectual ownership is gratefully acknowledged.

densities, where the weights are a cell's share of the city population. So in Figure 5.3, PPD gives a value of 5 for city 1 and 10 for cities 2 and 3. The De La Roca & Puga (2017) based measure further makes a distinction between cities 2 and 3. A city-wide measure of RPA is a sum of grid square measures, where each measure is a distance discounted sum of your own and neighbors' density within a given radius. Each grid square measure is weighted by its population share in the city. This measure by incorporating neighbors will give a higher value for city 2 than 3.

More generally, for personal population density [PPD] the measure for city j with N_j grid squares, i , is a weighted average of grid square populations P_{ij} , with weights being the grid squares' shares of city population, P_j . That is,

$$PPD_j = \sum_i^{N_j} P_{ij} \frac{P_{ij}}{P_j} = PD_j \left(1 + \frac{Var(P_{ij})}{PD_j^2} \right) = PD_j (1 + CV(P_{ij})^2) \quad (5.1)$$

where CV is the coefficient of variation and $PD_j = \frac{\sum_i^{N_j} P_{ij}}{N_j}$. PPD can thus be decomposed into overall population density [PD], a typical scale measure, and one plus the coefficient of variation squared. The latter captures the degree of variation relative to the mean within the city and thus the degree to which activity is concentrated within particular cells. So cities 2 and 3 in Figure 5.3 (ignoring city bounds) have the same degree of variation and clustering, and both are higher than city 1.

Note that the coefficient of variation has a long history, starting from Williamson (1965), for use as a measure of regional income inequality within a country. Here we are using it as a measure of economic density inequality within a city or town. Of course, urban economics has other measures of spatial inequality including spatial Hirschman-Herfindahl indices [HHI] and Gini's. We focus on the coefficient of variation because it comes from the natural decomposition in equation (1); and this decomposition which carries over in essence to the De La Roca & Puga (2017) measure.

The RPA measure (De La Roca & Puga (2017)) is still a weighted average, but now of the own grid square plus discounted neighbor cell populations over a given radius from the own grid square. The measure and its decomposition are

$$RPA_j = \sum_i A_{ij} \frac{P_{ij}}{P_j} = AD_j \left(1 + \frac{Cov(A_{ij}, P_{ij})}{AD_j PD_j} \right) \quad (5.2)$$

where $AD_j = \frac{\sum_i^{N_j} A_{ij}}{N_j}$; $A_{ij} = \sum_{k \in S} P_{kj} e^{-\alpha d_{ik}}$. In equation 2, A_{ij} is the measure over radius

s of the discounted sum of neighbors ambient populations. The discount rate we use is 0.7 and is derived empirically in Section 4. The neighborhood, s , of a grid cell is defined as the square area running 5 cells to the east, west, north and south of the own cell, or an area of size 11x11 grid squares (which would be similar to a circle of radius 6.2 km). We limit the local radius so we can distinguish later the effects of city wide versus local density. AD_j is the simple average of the A_{ij} across grid squares over the city. RPA_j is the weighted average of the A_{ij} , where the weights are each grid square's share of the city population. RPA_j can then be decomposed into the simple average, and 1 plus the covariance of A_{ij} and P_{ij} , divided by their simple averages. The latter term captures the degree to which population is allocated to grid squares with high measures of neighbors (city 2 in Figure 5.3), as opposed to either being uniformly spread (city 1) or being in grid squares which are not clustered with others of high density (city 3). Note, the personal population density measure in equation 1 is a special case of the De La Roca & Puga (2017) measure in equation 2, when radius s is set to 0.

We will have measures of PD_j , PPD_j and RPA_j at the level of a city or town. We will also have local measures, characterizing the neighborhood around which people live both for rural and urban areas, including for neighborhood i , PPD_{ij} , PD_{ij} , and A_{ij} . These we will describe in the particular contexts in which they arise. In all cases, the local neighborhood of a grid cell is the square area running 5 cells to the east, west, north and south of the own cell, or an area of size 11x11 km.

Most of the basics of what we have presented are not our invention. Small & Cohen (2004) calculate, on a coarser scale, a spatial Gini as a measure of within-city variation in activity. Implicit in the De La Roca & Puga (2017) measure is the notion of personal population density (when s is set to zero). Moreover, we borrowed the De La Roca & Puga (2017) measure we use, although from Section 4 we add an optimally derived spatial discount factor. They in fact set that discount factor to 0. The advantage of a non-zero discount factor is that far away places within a city get less weight, which makes the setting of city boundaries less critical for the key interior parts of a city. What we add which we believe to be novel, based on on-going work by Vernon Henderson, Adam Storeygard and David Weil, are decompositions for the personal population density and De La Roca & Puga (2017) measures.

5.4 How are differences in economic density across the spatial hierarchy related to income differences?

This section first describes the data on income and wages and then the characteristics of the sample of cities and towns in the covered countries. After giving the base specification, we turn to a set of results on the relationship between agglomeration measures and income and wages, covering all areas of the country. In the next section, we will delve into looking at scale effects for cities in particular.

5.4.1 The data and the sample of countries and cities

We use the Living Standards Measurement Study data of the World Bank, where we have detailed geocoding of where families live for six countries; allowing us to map data to our spatial units: rural, towns and cities. The LSMS surveys have detailed and consistent data at the household and individual levels on income, education, labor allocation, asset ownership, and dwelling characteristics. The data sets are the Tanzania Panel Household Survey (2008 and 2010), the Nigeria National Household Survey (2010 and 2012), the Uganda National Panel Survey (2009, 2010, 2011, and 2012), the Ethiopia Socioeconomic Survey (2011, 2013, and 2015), the Malawi Integrated Household Survey (2010 and 2013), and the Ghana Socioeconomic Panel Survey (2010 and 2013). Note that the dates of surveys in countries are so close together that they do not provide the opportunity to look at dynamics nor to identify urbanization effects off of movers.⁷ These sample countries account for approximately 35% of the subcontinent's population.

Before proceeding we note aspects of our Sub-Saharan African countries' urban hierarchy and the coverage of this hierarchy by LSMS surveys. We combine the urban data on the six countries and show that collectively they have a regular urban hierarchy. As a basic descriptive, Figure 5.4 shows approximately a log-normal distribution of all urbanized areas including both cities and towns similar to that in Eeckhout (2004) for the USA, although there is a right tail skew. Also as a descriptive, we looked at the rank size rule for the collective, although the rank size rule should be defined as country-specific. As in Eeckhout (2004) for the USA, the rank size rule is approximated in the right tail of Figure 5.4 which displays like a Pareto distribution.

⁷There is an issue of the same households appearing more than once in our data, which varies from country to country. For a sample of 44,140 households, there are 23,685 unique households, meaning that 46% of the sample involves a household that is included more than once. Clustering at the local area should remove the distortion this creates. As a robustness check, we reran basic tables with just the final year sample in the year of the LSMS for each country. Results are very similar, with similar statistical significance and coefficient magnitudes.

How complete is the LSMS coverage of this hierarchy? Table 5.1 shows the distribution of cities with their cores and fringes broken out for our countries and for the LSMS sample. The left part of the table tells us that these countries have 179 cities (and fringes), covering 220 cores; and they have 1065 towns, apart from rural areas. The right part of the table shows that the LSMS data covers 67% of the 179 cities; but within these cities, only 38% of fringe areas are covered. And for towns only 18% of the 1065 are covered. The relatively low count of small places actually surveyed comes from the randomized sampling procedure outlined in Appendix 5.B.2.

How representative is coverage by the LSMS? Figure 5.5 breaks the urban sample in Figure 5.4 into cities and towns. It compares the size distributions for all cities and for all towns in the six countries, with the size distributions of the cities and town that are covered by the LSMS. The shapes of distributions of both cities and towns for the sample are similar to those for the countries overall. The mean and median sizes for each distribution are each marked with dotted lines, with the mean being bigger than the median. For towns, the means and medians in the LSMS sample are larger than for the whole sample, and the same is the case for cities. This of course is consistent with Table 5.1 and the sampling procedure.

Next, we look at the characteristics of households in the sample. Table 5.2 gives characteristics of the LSMS households (top panel) and working age people (bottom panel) in the sample by our spatial units. Education of the household head and working-age population decline pretty sharply as we move down the spatial hierarchy. Rural areas, towns, and fringes of cities are much more likely to have the household head or workers in agriculture than the core. Virtually no one anywhere is in manufacturing, the big issue for Sub-Saharan African cities (Henderson & Kriticos 2018). Even the proportions in business services, which are potentially tradeable across cities, are not that high, at 9% for cities and 2% in rural areas for household heads. Business services include the usual business service industries such as real estate and finance but add in high skill workers (like managers) in retail, as well as senior administrators and high skill workers in government. Apart from agriculture, even in cities, it seems that most Sub-Saharan Africans work in low skill retail services and general labor services. However, a key issue with this industry data is that many people and household heads do not report an industry. Based on IPUMS data (Henderson & Kriticos 2018), we believe this may occur because many of these people are ‘farmers’ with agricultural land who work in other sectors as well, and so don’t have one industry with which they identify. We note this non-reporting fraction is noticeably higher at well over 50% in rural areas. In specifications below we will control for industry

fixed effects, distinguishing from not recorded agriculture, business services, construction, education and health, labor services, manufacturing, utilities, and retail and wholesale.

Finally, there are the income measures. We focus on total household income and then separately wage rates for individuals working for wage income. We construct measures of income for the household by adding together all income from self-employment, labor income, and capital or land income. In the surveys, respondents report income receipts of various forms, such as cash and in-kind wage payments, business incomes, remittances, incomes from the rent of property and farmland, private and government pensions, and sales revenue from agricultural produce. These receipts are also reported as taking place over a variety of time intervals, so to be consistent, we convert all income receipts to monthly intervals. Land income from crop sales or rents is generally only available at the household level, making it difficult to ascribe these income sources to any particular household member for an individual-level analysis. And the same comment applies to non-agricultural businesses owned by the household head or others in the family. Therefore, for individual-level analysis we only use individual wages. More details on the calculation of net income and hourly wages are provided in Appendix 5.B.2.

5.4.2 Basic specification and urban-rural results

Our first regressions have the following general specification:

$$\ln(y_{ijzft}) = \alpha X_{ijzft} + \beta I_Z + \gamma_R I_R * E_{ijRf} + \gamma_S I_S * E_{ijSf} + \gamma_C I_C * E_{ijCf} + \delta \xi_{ft} + \epsilon_{ijzft} \quad (5.3)$$

- $\ln(y_{ijzft})$: Income of household unit or hourly wage of person i in location j of type z in country f at time t .
- X_{ijzft} : Vector of household unit, person, or city characteristics.
- I_Z : Vector of indicators of location type: rural[R], town[S], city[C].
- E_{ijRf} : Measure of rural scale within a 6km radius of unit.
- E_{ijSf} : Measure of town scale.
- E_{ijCf} : Measure of city scale.
- ξ_{ft} : Vector of country-year FEs
- ϵ_{ijzft} : Error term.

In the first specification we will not have scale measures, just indicators to get town and city household income or wage premiums over rural areas. Then we will add in population and density scale measures. In the next section we will drop rural areas and towns and just look at cities, so the indicator terms and non-relevant scale measures will disappear.

What we estimate in this cross-section are correlations of income with scale measures, based on within country and year variation. We cannot claim causal effects for two reasons. First, there is the issue of sorting by unobserved 'ability' across space, although that has been downplayed in some of the literature (Baum-Snow & Pavan 2011), especially once one controls for education and age. De La Roca & Puga (2017) find that OLS cross-section results perform better than some other specifications such as individual fixed effects in panel data. For sorting, an issue is whether to control for occupation or industry fixed effects as a way of trying to factor in 'ability' conditional on education. Here we have industry fixed effects. Controlling for these fixed effects does reduce magnitudes and to be conservative we rely on these estimates, although we show some results without these controls. By doing so, however, we eliminate the scale benefit of greater choice of industries and occupations as we move up the urban hierarchy. The second issue in terms of identification is that bigger cities may have unobserved attributes which, apart from the scale, enhance productivity, such as local public infrastructure investments. But for that, at least, the estimates will give a sense of the income pull force of cities even if scale externality effects themselves could be over (or under) stated.

Table 5.3 presents preliminary results, which give the town and city household income and hourly wage rate premiums above the rural sector. Total household income and hourly wage rates are our two outcome measures throughout. We note that results for total wage income with the same controls are very similar to those for wage rates. Columns 1 and 2 look at effects on household income for 43,214 households, without and then with industry fixed effects for the household head. The controls include education, gender, age and age squared of the household head, household size and its square, whether the household owns land and, if so, the amount owned. Columns 3 and 4 look at hourly wages for the 19,901 individuals who work for wage income, again without and then with industry fixed effects of the individual. The controls include education, gender, age and age squared, as well as a cubic in hours worked. The coefficients on controls are reported in Appendix 5.A for the specifications in Table 5.5 below.

In Table 5.3, row 1 gives the premium from being in a town compared to a rural area ranging from 33% to 15%. Row 2 gives the premium to being in a city compared to rural area ranging from 72% to 27%. The wage premiums at 33 and 27% for being in large city are larger but comparable to the 25% often quoted number for the USA from Glaeser & Mare (2001). What is new is the much larger premium for household incomes, at least in the context of Sub-Saharan Africa. We note of course that these are nominal income not

consumption differences. To know how real incomes rise we would need cost of living or detailed consumption data, for which we have neither.

There are three clear patterns. First, being in a city compared to a town gives a much bigger premium, over twice as much for household incomes and just under twice as much for wages rates. Second, adding in industry fixed effects reduces premiums by up to 30%. While in the rest of the paper we include industry fixed effects as a control on unobserved attributes of people, we still worry that biases the returns to scale downwards by ignoring industry-occupation diversity and upgrading effects. Finally, returns to household income from moving up the spatial hierarchy are higher than for wages, a pattern we will focus on below. In column 2 the 52% premium to being in a city over a rural area for household income is about twice that for wages at 27% in column 4.

In the second set of preliminary results in Table 5.4, we turn to specifications where we experiment in each column with a different measure of economic density, proceeding from total population, to population density [PD], personal population density [PPD] and finally the De La Roca & Puga (2017) based measure [RPA] defined in equation (2). In panel A, we look at household incomes, with controls for household and household head characteristics including industry fixed effects. For each spatial group— rural, town, and city— we estimate a marginal scale effect for the four measures of scale. For rural, the scale measure applies to the 11x11 km square around the household’s grid square. In all columns in Table 5.4 we allow the income intercept to vary by spatial type, so as to normalize scale starting points relative to rural areas (like 50,000 for population for cities). But our interest is in marginal scale effects.

For rural households, local scale has no positive impact on household incomes. For towns population effects are negative but density effects are very high: an elasticity of 43% for mean density down to 11% for personal population density. For cities, the unit people usually focus on, the population elasticity is 8.5%, but the density ones range from 38 to 54%. A population elasticity of 8.5% is considerably higher than the often quoted numbers like 2-3 % (e.g., from Combes et al. (2008)) for developed countries. In Panel A, going from a rural area to a city of 5 million population raises household incomes by 55% (0.55 log points from $0.0846 \times 15.4 - 0.75$). However, in comparison, the returns to density are extraordinary. Within just cities, going from the lowest to the highest PD, PPD, or RPA raises household incomes by 1.14 to 1.52 log points (see Table 5.A1 on minimum and maximum values of each measure for the city sample). We will do detailed comparisons to the literature on other developing countries in the next section where we focus just on

density measures for the sample of cities and explain the divergence between population and density returns to scale. We do note here that for developed countries, in the classic paper by Ciccone & Hall (1996), doubling density raises productivity by 6%; here doubling density raises household incomes by 37% and in Panel B raises wage rates by 10%, as discussed below.⁸

In terms of which density measure is either sufficient or better, we explore that in the next section looking just at cities. We note here that, for just cities, the density elasticities are comparable across measures. While elasticities decline as we move from PD (0.52) to PPD (0.38) to RPA (0.37), their standard deviations within the city sample rise from 0.61 to 0.71, to 0.84 respectively. Between towns and cities, elasticities for PD are similar; but for PPD and to some extent for RPA, they are higher for cities.

So far we have mostly discussed household income effects in Panel A. Panel B examines hourly wage effects. The first comment is that corresponding to Table 5.3, scale elasticities for wages tend to be much smaller than for household incomes. For cities for density measures, marginal scale effects for household incomes tend to be about 3-4 times higher than for wages; and the population elasticity for wages for cities is zero. That said, wage density effects for cities are strong. Going from the lowest to highest PD, PPD, or RPA within the city sample raises hourly wages by over 30%.

5.5 Economic density in cities

In this section, we will delve into looking at scale effects in cities in particular and focus on the extraordinary density effects we have uncovered in Sub-Saharan Africa. The first question is why simple population and density measures yield so different elasticities. The second concerns which economic density measure suffices or does better than the others in explaining income and wage differences: population, density, personal population density, the RPA measure and components of the latter two. Before proceeding we need to tell how we derive the discount factor used in the RPA measure, used in part in assessing the degree to which within city clustering matters .

5.5.1 Constructing de la Roca-Puga measures

We take the specification in column 4 of Panel A of Table 5.4 and drop all observations outside of cities and thus the rural and town scale terms and the dummies for town and

⁸The recent evidence from De La Roca & Puga (2017) finds even smaller earnings elasticity for city density in Spain of 4.5%; and the authors do not find sizable effects of sorting on the unobservables that would affect the measured city premium.

city, so we are left with just the RPA measure for cities plus our basic controls. To recall, the RPA in effect measures how many people a person taken at random in a city can reach. A more realistic version of that measure would discount the number of people within reach according to how far away they are. What discount rate should we use, recognizing that such a discount rate is context specific and depends on local transport infrastructure and culture. We approach this problem empirically to get an overall rate for our sample. We start with a spatial discount rate of -0.1 and raise that in absolute value in increments of 0.1 to -1.5. One should note that -1.5 is an extremely high discount: at 1km distance, neighbors have a weight of 0.22; at 2km it is already only 0.05, and by 5km their weight is effectively 0. For each discount rate, we record the F-value of adding the ln RPA term; these values are shown in Figure 5.B3 in the Appendix. The peak is in the neighborhood of -0.7, so that the improvement in explaining income differences across cities is maximized around a -0.7 discount rate. We do note values from -0.5 to -0.9 yield similar F's under different specifications. We use the discount rate of -0.7 in all cases for any type of RPA measure. This discount rate indicates that neighbors, in general, are not so important. With this discount, at 2km, neighbors only get a weight of 0.25 and at 5km a weight of 0.03.

5.5.2 Economic density results for cities

We start by asking why population and density elasticities are so different, what measures best capture economic density, and how our wage results compare to the literature. Then we will discuss why household income premiums in denser cities are much larger than individual wage premiums.

Our basic results are in Table 5.5, where we look just at people living in cities. In columns 1, 3 4 and 5 we first repeat what in essence are the columns 1-4 results in Table 5.4 on returns to overall density measures. Any any differences in specific results between the tables arise from having just a city sample for which the effects of controls are estimated, as well as adding two new variables. These are meant to capture issues of how market access can affect income and wages, where market access may be also correlated with density and thus constitutes an omitted variable (Hanson 2005). We control for the distance from each city centre to the nearest port (Geospatial Geoscience Ltd 2017) and we add a measure of market access defined as

$$MA_i = \sum_{j=1}^N POP_j * e^{-0.002d} \quad (5.4)$$

POP_j is population of a neighboring city j within 3000 kilometers. We apply a distance

discount $e^{-0.002d}$, where d is distance. There is no robust trade elasticity for the African continent in the literature. We chose the parameter 0.002 in the weight as it matches a reasonable discount. For example, a city 100 kilometers away loses around 18% of its population and one 1000 km away loses 86% of its population, which seems plausible trade losses in Sub-Saharan Africa.

Adding the market access controls and changing the sample in estimating the effect of controls make little difference for the city results. For example, the population and density elasticities in Table 5.5 in columns 1 and 3 are 0.0506 and 0.523 for household incomes and 0 and 0.169 for wage rates; these are similar to those in Table 5.4 of 0.0846, 0.54, 0, and 0.143 respectively. Table 5.5 is the one which in Appendix 5.A we show two other things: coefficients on all controls (Tables 5.A5 and 5.A6) and the quite similar scale effects if we omit industry fixed effects (Table 5.A2). Tables 5.A5 and 5.A6 also show the very significant effects of market access and distance to ports on wages and incomes; but these controls have essentially no impact of density returns.⁹

Simple density versus population measures

We now address the question of why population and density effects differ so much, whether for household incomes or for wage rates. The answer is that column 1 in Table 5.5 with \ln population as the economic density measure has an omitted variable, land, which is correlated with population. When we add \ln land, in column 2 the coefficient on population increases tenfold to 0.52, so conditional on land, increases in population have a huge impact. Most crucially, in column 2 when we add \ln land area to column 1, for each of household incomes and wage rates, the coefficients on population and land are close to each other in absolute value and of opposite signs (0.516 versus -0.542 and 0.16 versus -0.20 for household incomes and wage rate respectively). That justifies the column 3 specification with just \ln population density as the economic density measure, where we impose equal and opposite sign coefficients on population and land. This neat result of land and population having essentially the same coefficients of opposite sign occurs because we are able to measure "economic land" so precisely. What matters is population density

⁹We also note that results are very similar for all other countries if we drop Nigeria, the largest country in the sample. The coefficients in Panel A on PD, PPD and RPA without Nigeria are 0.635, 0.386 and 0.416 (vs. respectively from Table 5.5 Panel A: 0.523, 0.382 and 0.372). For hourly wages they are 0.167, 0.112 and 0.120 (vs. respectively from Table 5.5 Panel B 0.169, 0.148 and 0.134). In general they are slightly higher for household incomes and slightly lower for wage rates if Nigeria is omitted. We also show that hourly wages results on density effects are similar for males versus females. For PD, PPD and RPA for wages only, male (female) coefficients are respectively 0.192 (0.132), 0.149 (0.148) and 0.149 (0.110). For all measures point estimates are somewhat higher for males.

in column 3, not population; and the density elasticity is enormous: 0.52 for household incomes and 0.17 for wages in Sub-Saharan Africa.¹⁰

Here a comparison with the recent literature is helpful. All the papers cited focus on wage rate outcomes and all use administrative boundaries to define land areas. Quintero & Roberts (2018) pool 16 Latin American countries much like we pool African countries. They have various estimates but at most the density elasticity is about 0.06 which corresponds to the estimate in Duranton (2016) for Colombia. Our estimate of 0.17 is much higher; but their uses of administrative boundaries create error in measurement of economic land. In Combes et al. (2019) the authors investigate several outcomes in China. For high and low skill urban natives in cities, they find higher density elasticities than Latin America ranging from 0.06 to 0.16 depending on the specification. Interestingly and conveniently, they control for \ln land area as well. Given that, as one specific but representative example (Table 5.2, panel b, column 2) the implied population elasticity is 0.141 and the land one is -0.056, very different from our situation where with precise land measures the absolute magnitudes are almost equal. In Chauvin et al. (2017) (Table 8, first panel), effects of \ln population and \ln density are investigated separately for the USA MSA's, Brazil microregions, Chinese cities and India districts, all based on land measured for administrative boundaries. Density elasticities for China are 0.192 while population ones are insignificant. Density elasticities in Brazil and India are lower at 0.026 and 0.076 respectively. However, the samples include rural and town areas as well, so results not directly comparable to ours, based on our Table 5.4 results. For USA MSA's, the density elasticity is 0.046, much less than for Sub-Saharan African and Chinese cities. In sum, for Sub-Saharan African cities, we find large wage elasticities compared to much of the literature and get unmuddied density results given our economic land areas are precisely measured.

Second moment measures

Does the degree of clustering within cities matter in this developing country context? The results on within city degree of clustering are in Table 5.5 columns 4 and 6. In column 4, we decompose \ln PPD into the \ln PD and \ln (1 + coefficient of variation term) in equation (1). The coefficient of variation term is small and insignificant and thus does not add

¹⁰In noting the divergence between marginal returns to population versus and density, we note that, in Sub-Saharan Africa, density and city population are not well correlated, with a simple correlation coefficient of only 0.31 (in logs). There are some very big, but low density cities. But even smaller cities can have higher density than more medium sizes ones. For personal population density, its mean for the lowest quartile of cities by size is 13,324 compared to the means of 9847 and 10857 for the second and third quartiles by city size.

to explanatory power relative to just using ln PD in column 2. In column 6, we repeat the same exercise for the RPA measure getting the same result; the covariance term from equation (2) is insignificant. We also conducted a series of horse-races in Appendix Tables 5.A3 and 5.A4. There, a horse-race between PD and PPD suggests PD dominates for both household income and wage rates. For PD versus RPA there is no strict pattern of domination by one or the other. In short, using PD seems as good or better than using PPD or RPA, and the two measures of the differential degree of clustering within cities do not seem to add to the analysis.

There are two issues with drawing the conclusion that we should focus on PD rather than PPD or RPA measures for cities. First concerns measurement error. Use of Landsat data measures within-city clustering and inequality with error. While Landsat may do a better job than other currently available data sets, the assignment of people to work and residential locations is surely done with considerable error, which would bias the coefficients downward. The second issue concerns our measure of clustering. While our measure of clustering fits into a decomposition and is a standard measure in looking at spatial inequality, there are other standard measures. We looked at one of these. In a context with so many grid cells in each urban area but a varying number, we preferred the spatial Gini to use in comparison over an HHI or a Theil index. We calculated the spatial Gini of cell concentration of economic activity within each city.¹¹ The Gini had also a completely insignificant coefficient as reported in an earlier version of this paper (see Henderson, Nigmatulina & Kriticos (2018)).

The conclusion for this sample and data is that a simple population density measure works as well as more nuanced measures, in attempting to capture *overall* economic density of cities in explaining income differences. However, later when we consider local within city density measures, our more nuanced measures will have more purchase.

Why are economic density effects for wage rates and household incomes so different?

What is stunning in the tables are the extraordinarily high density elasticities for household incomes compared to wages. Why might this be? We explore three channels here: greater labor force participation within the household in bigger or denser cities with better job

¹¹We approximate the Gini by ordering each cell by its density and noting the cumulative share of the population in each cell. The cumulative distribution of the population ordered by density represents the Lorenz curve. We then sum up the area under the Lorenz curve (by adding up the “height” and the “width” of each bar of the cumulative population histogram, where the “height” is the cumulative population and the “width” is $1/(\text{number of total cells in the city})$). We call this integral I, and, according to the Gini formula, calculate $\text{Gini}=1-2I$ for each city.

opportunities, opportunities and choices to work longer hours, and an opportunity to upgrade and diversify occupations or industries. We look at these in Table 5.6. In columns 1 and 2, we look at labor force participation in terms of working for wage income or not, for those 18-60 for all and for females. Density has no effect. However, we find effects for the other two channels.

In columns 3 and 4 of Table 5.6, we look at hours worked for those working for wages. We find similar effects for all as we do for just females alone. A one standard deviation (0.61) increase in density raises hours worked by 2.31 from a mean of 47 for females and similarly overall (in a formulation where the count of hours is in logs, the elasticities are both 0.09). Next in columns 5 and 6 for the household we look at the count of different primary industries of employment listed by different members of the household, controlling for household sizes. So if all work in retail, the count is one. If two people work in retail and one in manufacturing the count is 2. In column 5, a one standard deviation increase in density raises the count of number of different industries of household members by 0.08 from a mean of 1.3 (and in a log formulation the elasticity is 0.063). In column 6 we see if density interacts with household size but the coefficient is insignificant.

Two channels do seem to be working: denser cities offer the opportunity and choice to work more hours and for the household member to match to a greater variety of industries. We note that in Table 5.2 there are a lot more manufacturing and business service jobs in cities than in towns or rural areas. We documented (not reported) that household activity in agriculture declines sharply with density. We think families move to cities, not just for wage gains but for huge household income gains based on the opportunities denser places offer household members to work longer hours and to better match to industries outside of agriculture for which different household members have more interest and better skills. There must be other channels, given our large differences in income and wage returns to density, which warrant investigation in the future.

Does neighborhood density matter?

A key issue as noted in the introduction is that studies suggest that within cities there are local scale externalities which decay sharply with distance so that firms in one neighborhood do not seem to interact with firms in another (Arzaghi & Henderson 2008), at least in the dimensions being captured. That begs the question of why there can be multiple neighborhoods of seemingly non-interacting firms found in one city. The answer must be that firms also benefit more generally from overall urban scale. Here we show results where both overall city density and local density matter, so both elements are captured.

In Tables 5.7 and 5.8, columns 1-3 of each, we address this issue. All these columns control for citywide \ln PD and then introduce a local measure for the 11 x 11 km neighborhood around the own grid square (i.e., a radius of just over 5km). Local and city densities will tend to be positively correlated, so if the local density measures matter that will reduce the elasticity of \ln city PD which is 0.52 in column 2 of Table 5.5, Panel A, and 0.169 in column 2 of Table 5.5, Panel B. This is definitely the case for Table 5.7 on household incomes but less clear for Table 5.8 for wages. In either case the coefficients on \ln city PD are still very large, indicating huge marginal returns to overall city density. In each of columns 1-3 in Tables 5.7 and 5.8, we experiment with a different measure of local economic density. Although we have argued that a simple PD measure works as well or better than anything else for a citywide measure, that does not mean at the local level it is the best measure. In column 1 the local measure is \ln (local PD); in 2 it is \ln (local PPD); and in 3 we decompose local PPD as in equation (1) into PD and the covariance term.

For household incomes, in all three columns in Table 5.7, local density measures have strong significant effects; and column 3 in Table 5.7 suggests local covariance matters, so that local clustering within your neighborhood helps incomes. Local magnitudes are all smaller, about 0.16, than city wide PD elasticities of 0.46 in this table. Horse-races in Appendix Table 5.A3 for household income suggest no clear winner between local PD and local PPD, although local PPD dominates local RPA (and hence we only report it in the working paper (Henderson, Nigmatulina & Kriticos (2018))).

In Table 5.8 on wages, there is less impact of local measures. Local PD is smaller and insignificant. Local PPD is significant at the 5% level, as is the local coefficient of variation measure. Quantitatively, as in Table 5.7, in Table 5.8 elasticities for local PD measures are about 1/4 to 1/3 of those for the overall city PD measure; but standard errors are too large for these smaller magnitudes to be significant. It may be that local economic density is more important for household incomes which may depend on local jobs for household members who are part time or supplemental earners and who may work more for non-wage income, than at an hourly rate.

Finally we note that these results pertain to wage rates and household incomes. Firm productivity effects may be a different story. In Appendix 5.D we look at neighborhood effects for firms in Kampala. Having greater Landscan density in the neighborhood around a firm does nothing to productivity, while for households having greater local density raises incomes in Table 5.7. Firms only seem to respond to greater local (but not immediate) own industry employment density. Perhaps this differentiation for results for households versus

firms is not surprising. For households it is about job search, job training, and learning (De La Roca & Puga (2017)).

In summary, for a household, the overall density measure for the city has strong effects on incomes, but also the local area around the household has strong effects as well. Households benefit from both dense overall cities and dense local neighborhoods.

The core versus fringe of cities

Finally, we examine how effects may differ for people living in the core versus fringe of cities, which will suggest more work for the future. We look at this issue in the last two columns of Tables 5.7 and 5.8. In both Tables 5.7 and 5.8, while those living in the fringe start from a much lower base income or wage rate (the fringe indicator), in column 4, they get an extra kick from city PD, with an elasticity of density that overall is 0.73 for household incomes (versus 0.44 for those living in the core) and 0.40 for wages (vs a much smaller 0.11 for those living in the core). These are huge differences which warrant further exploration. We note that for fringe residents it is overall city, not core PD which matters as reported in an earlier version of the paper (Henderson, Nigmatulina & Kriticos 2018). We looked similarly in column 5 in Tables 5.7 and 5.8 at the effects of local PD for those in the core versus the fringe. In both tables, the results suggest local PD matters just for those living in the fringe. One thought is that in the core, densities may be more uniform and higher, while at the fringe, local density may be much more variable and that is highly relevant.

The puzzle is the generally bigger role of city density externalities for fringe than core city residents. We thought this was because of sorting based on migration status, where migrants lack information and might benefit more from say information spillovers. Hence, we expected that the fringe would have a greater proportion of migrants (defined by having moved within the last 5 years) than the core within the LSMS sample. However, there is only a tiny difference: 16% vs 15%. Moreover, coefficients on migration status interacted with scale effects are insignificant and have no impact on the fringe results. So that leaves a puzzle. There certainly must be sorting on some dimension, where people in the fringe earn less for the same observables, but somehow this disadvantaged population benefits more from greater city density, even though by construction they live in less dense neighborhoods.

5.6 Conclusions

This paper evaluates the use of different measures of economic density in assessing urban agglomeration effects to try to carefully measure and characterize the pull force of cities. These density measures are based on Landsat data, which we argue in Appendix 5.C does a good job in capturing within city variation in economic density. The Landsat data allow us to precisely and consistently define cities and measure the economic extent of the city, or its land area.

To assess economic density measures, we examine how well they explain household income and wage differences across cities and neighborhoods. We have simple scale and density measures and more nuanced ones which capture in second moments the extent of clustering within cities. Noting that the extent of clustering is measured with error resulting in attenuation bias, the evidence suggests that a simple, but consistently calculated density measure explains income differences across cities as well as or even better than more nuanced measures which attempt to represent within city differences in the extent and nature of clustering. However, simple city scale measures such as total population are distinctly inferior to density measures.

Overall, there are huge household income premiums from being in bigger and particularly denser cities over rural areas and small towns in Sub-Saharan Africa, indicating migration pull forces are very strong. It may be that part of urban-rural premiums are explained by low productivity in the farming sector. Additionally, the selection of unobservably abler migrants into cities may also add to the size of the effect, although we have many controls on household and member attributes. However, within the sample of cities, the marginal effects of increases in density on household income are enormous by any standard, with a density elasticity of 0.52. Besides overall city density measures, we look at density in the neighborhood around a household within a city: in addition to strong city-level density effects we find strong neighborhood effects looking at neighborhoods of about 6km radius. The elasticity of overall city density is 0.42 and for local density it is 0.14. Both overall city density and density of the own neighborhood matter.

For wages, while density premiums are on the high end relative to the literature, they are much lower than for household incomes. We found two channels for why income elasticities are higher, although more must be at work: people work longer hours as density rises and households diversify industries.

Future research would reestablish current results with the new datasets on density soon

to emerge, and some of which are at a much more disaggregated scale (Worldpop (100m), High Resolution Human Settlement Layer (30m), MAUPP (12.5m)). Higher resolution data on all aspects will uncover more accurate clustering variation within cities, to help us establish how important that is for households' earning potential. Adding detailed within-city transport information that is consistent across cities can also paint a better picture on the true proximity between workers; see Akbar, Couture, Duranton, Ghani & Storeygard (2018) for steps in this direction. Having economic census data would allow better assessment of different forms of density— diurnal versus nocturnal. Individual panel data might help in identification and uncovering of dynamic effects, but such panels are in their infancy anywhere in Africa.

Table 5.1: Counts of urbanised areas in our countries and sample

	All urban areas in our countries			Urban Areas where surveyed units live			
Country	Cities	Cores	Towns	Cities	Cores	Fringes	Towns
Ethiopia	24	32	244	71%	63%	21%	14%
Ghana	15	15	66	67%	53%	40%	23%
Malawi	5	5	52	80%	80%	80%	50%
Nigeria	96	129	493	61%	49%	32%	9%
Tanzania	24	24	71	63%	50%	33%	21%
Uganda	15	15	139	100%	93%	93%	40%
Totals	179	220	1065	67%	55%	38%	18%

Notes: The table shows the counts of urban areas by type in the seven countries, and the share of them populated by the the Living Standard Measurement Surveys.

Source: World Bank Living Standard Measurement Surveys and Landscan (2012).

Table 5.2: Household and person characteristics by location

	City	Core	Fringe	Town	Rural
Panel A					
Household head in Agriculture	0.124	0.0629	0.273	0.261	0.297
Household head in Business Services	0.0885	0.108	0.0418	0.0510	0.0238
Household head in Manufacturing	0.0271	0.0297	0.0205	0.0220	0.00877
Household head in Not-Recorded	0.362	0.375	0.329	0.397	0.561
Household head with >Primary Education	0.441	0.503	0.290	0.307	0.173
Panel B					
Worker in Agriculture	0.0904	0.0404	0.214	0.239	0.281
Worker in BS	0.0608	0.0738	0.0286	0.0309	0.0164
Worker in Manufacturing	0.0246	0.0265	0.0199	0.0248	0.0148
Worker in Not-Recorded	0.467	0.473	0.452	0.473	0.579
Worker with >Primary Education	0.5	0.544	0.391	0.332	0.198
No. Urban Areas	120	121	68	189	-

Notes: The table shows the shares of occupations in our sample out of 1) the sample of household heads and 2) working age population (18-60 years old).

Source: World Bank Living Standard Measurement Surveys and Author's dataset.

Table 5.3: Gains from urban type

	(1)	(2)	(3)	(4)
	HH income	HH income	Hrly wage	Hrly wage
Town	0.333*** (0.0314)	0.213*** (0.0301)	0.194*** (0.0309)	0.145*** (0.0307)
City (core + fringe)	0.720*** (0.0224)	0.524*** (0.0225)	0.326*** (0.0190)	0.270*** (0.0193)
Observations	43214	43214	19901	19901
R^2	0.335	0.383	0.296	0.313
Country-Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Occupation FE		✓		✓

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable in columns (1)-(2) is the natural logarithm of total net income from all available sources. The dependent variable in columns (3)-(4) is the natural logarithm of hourly wage income. Controls at the household level are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Controls at the individual level are education (recorded or not); level of education (if recorded); age; age squared; gender; and hours worked squared and cubed. The sample in columns (3)-(4) is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Each column includes country-year fixed effects. Columns (2) and (4) include industry fixed effects for household head and individual worker respectively. Robust standard errors are presented in parentheses. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Panel A. Household incomes

	(1)	(2)	(3)	(4)
	Log Population	Log Mean Density	Log PPD	Log RPA
Rural X Scale	-0.00979 (0.0105)	-0.0141 (0.0108)	0.00379 (0.00790)	-0.00712 (0.0101)
Town X Scale	-0.0857*** (0.0313)	0.429*** (0.147)	0.107** (0.0526)	0.196** (0.0763)
City (core + fringe) X Scale	0.0846*** (0.0111)	0.540*** (0.0261)	0.440*** (0.0223)	0.377*** (0.0189)
Town	1.044*** (0.352)	-2.657*** (0.972)	-0.610 (0.429)	-1.643** (0.716)
City (core + fringe)	-0.752*** (0.184)	-3.541*** (0.202)	-3.508*** (0.211)	-3.584*** (0.218)
Observations	43214	43214	43192	43214
R^2	0.384	0.389	0.389	0.389

Panel B. Individual hourly wage premiums

	(1)	(2)	(3)	(4)
	Log Population	Log Mean Density	Log PPD	Log RPA
Rural X Scale	-0.0324** (0.0141)	-0.0298** (0.0146)	0.00732 (0.0105)	-0.0315** (0.0133)
Town X Scale	-0.0226 (0.0315)	0.0875 (0.143)	0.186*** (0.0557)	0.164** (0.0812)
City (core + fringe) X Scale	0.00684 (0.00838)	0.143*** (0.0207)	0.106*** (0.0187)	0.110*** (0.0155)
Town	0.0882 (0.367)	-0.572 (0.946)	-1.334*** (0.462)	-1.619** (0.768)
City (core + fringe)	-0.127 (0.176)	-0.945*** (0.173)	-0.672*** (0.186)	-1.163*** (0.196)
Observations	19901	19901	19899	19901
R^2	0.314	0.315	0.315	0.315

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions. The dependent variable in Panel A is the natural logarithm of total net income from all available sources. The dependent variable in Panel B is the natural logarithm of hourly individual wage income. Scale variable varies by column and is indicated in column headings. Controls are the same as in Table 3. Each column includes country-year and industry fixed effects. The sample in Panel B is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Robust standard errors are presented in parentheses. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Table 5.4: Gains from scale by area type

Panel A. Household incomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ln(Total Population)	0.0506 (0.0473)	0.516*** (0.0813)					
Ln(Pop. Density)			0.523*** (0.0822)		0.531*** (0.0830)		
Ln(PPD)				0.382*** (0.0838)			
Ln(1 + CV Term)					0.0918 (0.129)		
Ln(RPA)						0.372*** (0.0604)	
Ln(AD)							0.462*** (0.0864)
Ln(1 + Cov Term)							0.159 (0.202)
Ln area		-0.542*** (0.0840)					
Observations	11237	11237	11237	11237	11237	11237	11237
R ²	0.314	0.330	0.330	0.326	0.330	0.330	0.330

Panel B. Individual hourly wage premiums

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ln(Total Population)	-0.00681 (0.0244)	0.160*** (0.0380)					
Ln(Pop. Density)			0.169*** (0.0393)		0.180*** (0.0387)		
Ln(PPD)				0.148*** (0.0369)			
Ln(1 + CV Term)					0.0609 (0.0625)		
Ln(RPA)						0.134*** (0.0291)	
Ln(AD)							0.132*** (0.0472)
Ln(1 + Cov Term)							0.139 (0.119)
Ln area		-0.198*** (0.0410)					
Observations	9358	9358	9358	9358	9358	9358	9358
R ²	0.364	0.369	0.368	0.368	0.368	0.369	0.369

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions. The dependent variable in Panel A is the natural logarithm of total net income from all available sources. The dependent variable in Panel B is the natural logarithm of individual hourly wage income. Controls are the same as in Table 3; additional controls include natural logarithm of distance to the nearest port and natural logarithm of market access to all cities and towns, weighted by $e^{(0.002 * d)}$ where d is distance. The sample in Panel B is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Each column includes country-year and industry fixed effects. Standard errors are clustered at the city level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). Source: Author's dataset

Table 5.5: Gains from within city clustering

Table 5.6: Why household income gains are larger than for wages?

	(1)	(2)	(3)	(4)	(5)	(6)
	Working for wage	Working for wage	Hours worked for those recorded	Hours worked for those recorded	N of unique occupations in a household	N of unique occupations in a household
Ln(Pop. Density)	0.000285 (0.00470)	-0.00434 (0.00304)	3.399*** (0.565)	3.788*** (0.546)	0.136*** (0.0380)	-0.172 (0.201)
HH size					0.159*** (0.0358)	-0.327 (0.278)
Ln(Pop. Density) \times HH size						0.0646 (0.0405)
Observations	10211	4264	9358	3920	13508	13508
R^2	0.071	0.044	0.067	0.084	0.620	0.628
Country-Year FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓
Women Only		✓		✓		
Mean of Y			48.22	46.69	1.258	1.258

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable is indicated in the column titles. Columns (1)-(4) use the sample of individuals aged 18 to 60, restricting to the sample of just women in columns (2) and (4). Column (5) presents a sample of households. Controls are the same as in Table 3. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanised area level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Table 5.7: Estimation of average and local density effects on household income premiums

	(1)	(2)	(3)	(4)	(5)
Ln(City PD)	0.421*** (0.0948)	0.462*** (0.0900)	0.442*** (0.0961)	0.436*** (0.0846)	0.438*** (0.0896)
Ln(Local PD)	0.142*** (0.0438)		0.171*** (0.0465)		
Ln(Local PPD)		0.162*** (0.0438)			0.00908 (0.0684)
Ln(1 + Local CV Term)			0.137* (0.0738)		
Fringe				-2.452*** (0.894)	-4.225*** (0.999)
Fringe × Ln(City PD)				0.294** (0.120)	0.272** (0.121)
Fringe × Ln(Local PPD)					0.237*** (0.0830)
Observations	11237	11237	11237	11237	11237
R^2	0.286	0.287	0.287	0.289	0.291
Country-Year FE	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓	✓

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the same as in Table 3 at the household level. Additional controls include natural logarithm of distance to the nearest port and natural logarithm of market access to all cities and towns, weighted by $e^{(0.002 * d)}$ where d is distance. Each column includes country-year and occupation fixed effects. Standard errors are clustered at the individual urbanised area level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Table 5.8: Estimation of average and local density effects on wage premiums

	(1)	(2)	(3)	(4)	(5)
Ln(City PD)	0.153*** (0.0479)	0.156*** (0.0415)	0.172*** (0.0497)	0.112*** (0.0359)	0.122*** (0.0369)
Ln(Local PD)	0.0435 (0.0410)		0.0673 (0.0419)		
Ln(Local PPD)		0.0727** (0.0362)			-0.00902 (0.0408)
Ln(1 + Local CV Term)			0.0932** (0.0441)		
Fringe				-2.276*** (0.582)	-3.349*** (0.637)
Fringe × Ln(City PD)				0.291*** (0.0782)	0.278*** (0.0719)
Fringe × Ln(Local PPD)					0.139* (0.0719)
Observations	9358	9358	9358	9358	9358
R^2	0.350	0.351	0.351	0.354	0.355
Country-Year FE	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓	✓

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the same as in Table 3 at the individual level. Additional controls include log distance to the nearest port and natural logarithm of market access to all cities and towns, weighted by $e^{(0.002 * d)}$ where d is distance. The sample is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Each column includes country-year and occupation fixed effects. Standard errors are clustered at the individual urbanised area level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*).

Source: Author's dataset

Figure 5.1: Defining cities and towns

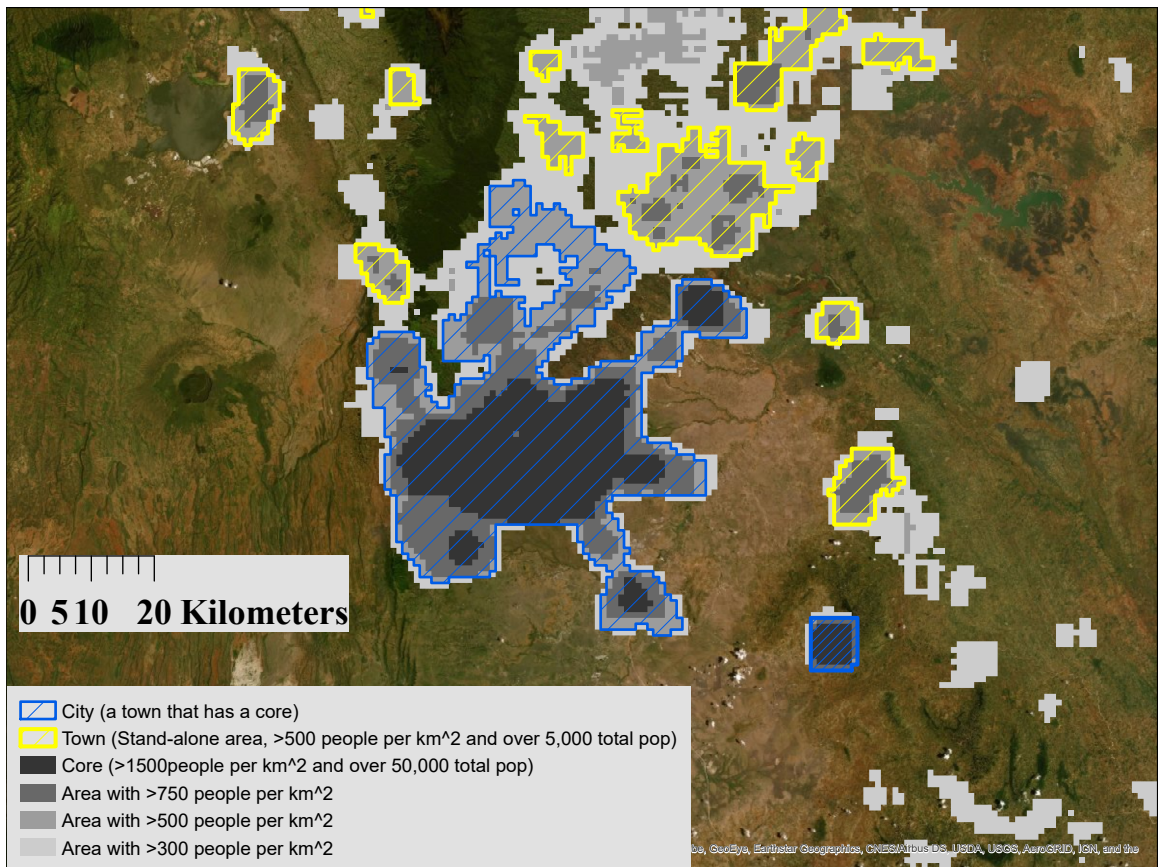
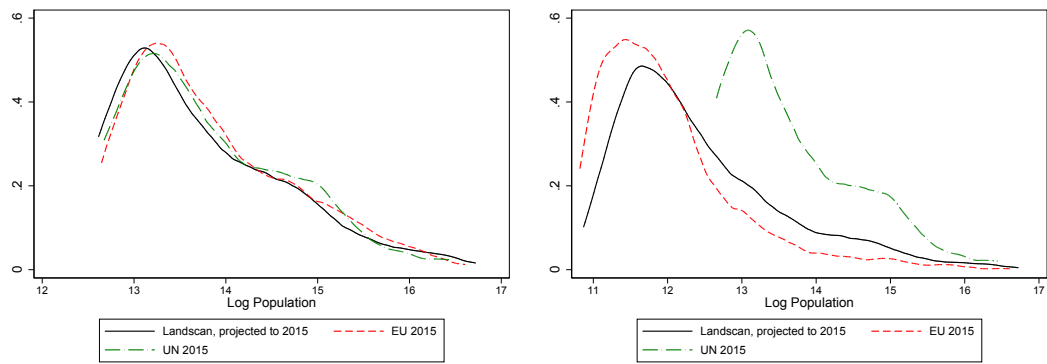


Figure 5.2: PDFs of city sample in the African continent



<u>City 1</u>						<u>City 2</u>						<u>City 3</u>					
5	5	5	5	5	5	0	0	0	10	10	10	0	10	0	10	0	10
5	5	5	5	5	5	0	0	0	10	10	10	10	0	10	0	10	0
5	5	5	5	5	5	0	0	0	10	10	10	0	10	0	10	0	10
5	5	5	5	5	5	0	0	0	10	10	10	10	0	10	0	10	0
5	5	5	5	5	5	0	0	0	10	10	10	0	10	0	10	0	10
5	5	5	5	5	5	0	0	0	10	10	10	10	0	10	0	10	0

Figure 5.3: Differences in city layout and density measures

Figure 5.4: Overall size distribution of urbanized areas

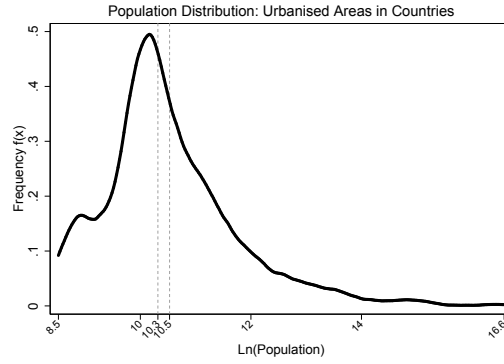
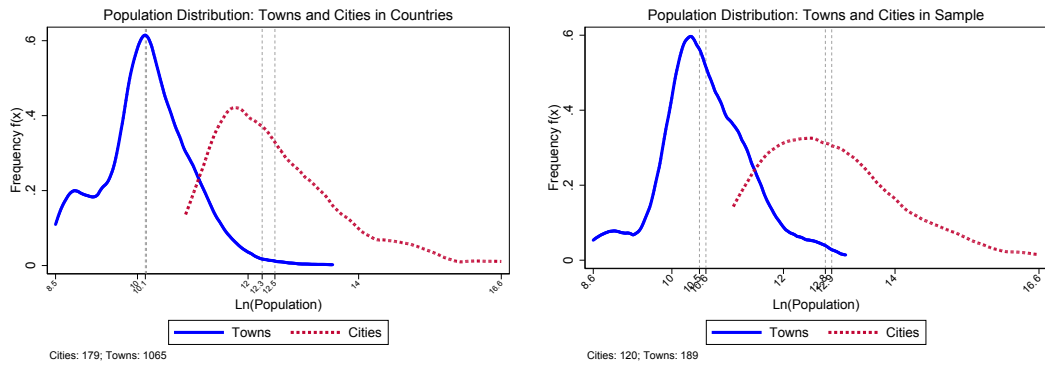


Figure 5.5: Size distribution of urbanized areas



5.A Appendix A. Statistics and other results

Table 5.A1: Summary statistics

Panel A: Cities						
Variable	Mean	Min	Median	Max	STD	N
Local Measures						
Ln(Local Landscan Linear Sum (5km))	12.13	10.03	12.02	14.45	1.05	11237
Ln(Local PD)	7.78	5.77	7.65	10.17	1.07	11237
Ln(Local PPD)	9.17	6.34	9.29	10.92	0.97	11237
Ln(Local RPA)	10.39	8.22	10.28	12.79	1.10	11237
Citywide Measures						
Ln(Total Population)	14.06	11.10	14.44	16.58	1.43	11237
Ln(Pop. Density)	7.47	6.51	7.34	8.65	0.61	11237
Ln(PPD)	9.30	7.36	9.30	10.81	0.71	11237
Ln(Coefficient of Variation)	1.83	0.81	1.81	3.04	0.37	11237
Ln(RPA)	10.87	9.26	10.64	12.28	0.84	11237
Ln(RPA) unweighted	9.88	8.95	9.73	11.13	0.60	11237
Ln(Cov. Term RPA)	1.00	0.26	0.99	1.88	0.31	11237
Panel B: Towns						
Variable	Mean	Min	Median	Max	STD	N
Ln(Total Population)	11	9	11	13	1	3440
Ln(Pop. Density)	6.57	5.86	6.55	7.19	0.18	3440
Ln(PPD)	8.10	6.27	8.12	9.33	0.52	3440
Ln(RPA)	9.31	8.63	9.28	10.11	0.34	3440
Panel C: Rural						
Variable	Mean	Min	Median	Max	STD	N
Ln(Local Landscan Linear Sum (5km))	9.06	0.00	9.18	11.52	0.97	28537
Ln(Local PD)	4.68	0.00	4.80	7.12	0.93	28537
Ln(Local PPD)	6.05	0.37	6.16	9.89	1.26	28515
Ln(Local RPA)	7.18	0.00	7.30	9.42	0.99	28537

Note: The table presents summary statistics of the natural logarithms of the city-level and local variables of interest. Panel A is the summary for the sample of just cities, Panel B is the summary of town-level measures for towns, and finally, the Panel C is the summary of the local measures for rural survey locations.

Panel A. Household incomes

	(1)	(2)	(3)	(4)
Rural X rursc		-0.00735 (0.0109)	-0.0129 (0.0113)	0.0163* (0.00830)
Town X Scale		-0.0860*** (0.0329)	0.546*** (0.155)	0.192*** (0.0549)
City (core + fringe) X Scale		0.0723*** (0.0114)	0.558*** (0.0264)	0.456*** (0.0228)
Town		1.189*** (0.370)	-3.304*** (1.024)	-1.111** (0.447)
City (core + fringe)		-0.363* (0.190)	-3.481*** (0.207)	-3.388*** (0.218)
Observations	43214	43214	43214	43192
R^2	0.317	0.336	0.341	0.341

Panel B. Individual hourly wage premiums

	(1)	(2)	(3)	(4)
Rural X rursc	-0.0204 (0.0142)	-0.0182 (0.0146)	0.0175* (0.0106)	-0.0171 (0.0133)
Town X Scale	-0.0265 (0.0319)	0.0350 (0.144)	0.171*** (0.0568)	0.142* (0.0827)
City (core + fringe) X Scale	0.00397 (0.00844)	0.157*** (0.0208)	0.113*** (0.0187)	0.119*** (0.0156)
Town	0.292 (0.372)	-0.122 (0.956)	-1.093** (0.471)	-1.261 (0.782)
City (core + fringe)	0.0815 (0.177)	-0.939*** (0.174)	-0.621*** (0.187)	-1.096*** (0.197)
Observations	19901	19901	19899	19901
R^2	0.296	0.298	0.297	0.298

Notes: Table reports results from OLS regressions; the dependent variable in Panel A is the natural logarithm of total net income from all available sources. The dependent variable in Panel B is the natural logarithm of hourly individual wage income. Scale variable varies by column and is indicated in column headings. Controls are the same as in Table 5.3. Each column includes country-year. The sample in Panel B is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Robust standard errors are presented in parentheses. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*).
Source: Author's dataset

Table 5.A2: Gains from scale by area type. No industry fixed effects.

Table 5.A3: Horserace: estimation of density effects on household incomes

	(1)	(2)	(3)	(4)
Ln(Pop. Density)	0.439*** (0.132)	0.303 (0.224)	0.407*** (0.0893)	0.418*** (0.0859)
Ln(PPD)	0.0918 (0.129)			
Ln(RPA)		0.167 (0.170)		
Ln(Local PD)			0.0483 (0.0657)	
Ln(Local PPD)			0.0964 (0.0680)	0.105* (0.0545)
Ln(Local RPA)				0.0360 (0.0497)
Observations	11237	11237	11237	11237
R^2	0.330	0.330	0.332	0.332
Country-Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net income from all available sources. Controls are the same as in Table 5.3 at the household level. Additional controls include log distance to the nearest port and natural logarithm of market access to all cities and towns, weighted by $e^{(0.002 * d)}$ where d is distance. Each column includes country-year and industry fixed effects. Standard errors are clustered at the city level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Table 5.A4: Horserace: estimation of density effects on individual hourly wage premiums

	(1)	(2)	(3)	(4)
Ln(Pop. Density)	0.119*	0.0117	0.156***	0.149***
	(0.0687)	(0.138)	(0.0441)	(0.0369)
Ln(PPD)	0.0609			
	(0.0625)			
Ln(RPA)		0.126		
		(0.105)		
Ln(Local PD)			-0.0315	
			(0.0523)	
Ln(Local PPD)			0.0802**	0.0754**
			(0.0402)	(0.0365)
Ln(Local RPA)				-0.0253
				(0.0413)
Observations	9358	9358	9358	9358
R^2	0.368	0.369	0.369	0.369
Country-Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Occupation FE	✓	✓	✓	✓

Standard errors in parentheses

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

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of individual hourly wage income. Controls are the same as in Table 5.3 at the individual level. Additional controls include log distance to the nearest port and natural logarithm of market access to all cities and towns, weighted by $e^{(0.002*d)}$ where d is distance. The sample is limited to individuals aged 18 to 60, working for a wage and reporting hours worked. Each column includes country-year and industry fixed effects. Standard errors are clustered at the city level. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: Author's dataset*

Table 5.A5: Estimation of scale effects on HH income premiums

	(1)	(2)	(3)	(4)	(5)	(6)
Ln(Total Population)	0.0412 (0.0530)					
Ln(Pop. Density)		0.577*** (0.0887)		0.586*** (0.0896)		
Ln(PPD)			0.428*** (0.0900)			
Ln(Coefficient of Variation)				0.101 (0.140)		
Ln(RPA)					0.407*** (0.0644)	
Ln(RPA) unweighted						0.508*** (0.0913)
Ln(Cov. Term RPA)						0.170 (0.223)
Education is recorded	-0.385*** (0.124)	-0.346*** (0.122)	-0.360*** (0.123)	-0.346*** (0.122)	-0.348*** (0.123)	-0.344*** (0.123)
Kindergarten education	0.0295 (0.148)	0.0657 (0.139)	0.0442 (0.141)	0.0635 (0.139)	0.0787 (0.139)	0.0704 (0.139)
Primary education	0.360*** (0.124)	0.351*** (0.116)	0.336*** (0.118)	0.346*** (0.115)	0.336*** (0.114)	0.343*** (0.114)
Secondary education	0.822*** (0.127)	0.753*** (0.118)	0.768*** (0.118)	0.751*** (0.117)	0.752*** (0.116)	0.753*** (0.116)
Specialised education	1.624*** (0.168)	1.530*** (0.162)	1.551*** (0.162)	1.528*** (0.162)	1.527*** (0.162)	1.530*** (0.161)
Age HH head	0.00241 (0.00733)	0.00175 (0.00717)	0.00380 (0.00689)	0.00214 (0.00700)	0.00280 (0.00706)	0.00217 (0.00709)
(Age HH head) Squared	-0.0000615 (0.0000817)	-0.0000452 (0.0000789)	-0.0000662 (0.0000769)	-0.0000486 (0.0000776)	-0.0000546 (0.0000783)	-0.0000486 (0.0000787)
Gender	0.188*** (0.0612)	0.197*** (0.0552)	0.213*** (0.0551)	0.202*** (0.0539)	0.202*** (0.0554)	0.200*** (0.0547)
HH size	0.146*** (0.0172)	0.153*** (0.0165)	0.153*** (0.0173)	0.153*** (0.0166)	0.151*** (0.0171)	0.151*** (0.0168)
(HH size) Squared	-0.00269*** (0.000941)	-0.00299*** (0.000879)	-0.00299*** (0.000947)	-0.00301*** (0.000888)	-0.00290*** (0.000922)	-0.00290*** (0.000903)
HH has land	-0.824*** (0.155)	-0.635*** (0.119)	-0.663*** (0.110)	-0.627*** (0.114)	-0.632*** (0.115)	-0.636*** (0.117)
Ln(Total size of landholdings (ha))	0.208* (0.116)	0.217** (0.102)	0.209** (0.0996)	0.216** (0.100)	0.223** (0.0982)	0.215** (0.0997)
Ln(Distance to port)	-0.0308* (0.0177)	-0.0285*** (0.00945)	-0.0299** (0.0118)	-0.0281*** (0.00954)	-0.0303*** (0.0104)	-0.0299*** (0.00973)
Ln(Market Access settlement (discount = 0.002))	0.642** (0.278)	0.791*** (0.250)	0.759*** (0.289)	0.798*** (0.255)	0.901*** (0.280)	0.842*** (0.263)
Constant	-7.332 (4.947)	-13.73*** (4.425)	-12.86** (5.188)	-14.12*** (4.567)	-15.74*** (4.923)	-15.49*** (4.663)
Observations	11237	11237	11237	11237	11237	11237
R^2	0.263	0.283	0.278	0.284	0.283	0.283

Standard errors in parentheses

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

Table 5.A6: Estimation of scale effects on hourly wage premiums

	(1)	(2)	(3)	(4)	(5)	(6)
Ln(Total Population)	-0.0152 (0.0254)					
Ln(Pop. Density)		0.201*** (0.0448)		0.210*** (0.0451)		
Ln(PPD)			0.169*** (0.0429)			
Ln(Coefficient of Variation)				0.0502 (0.0636)		
Ln(RPA)					0.155*** (0.0337)	
Ln(RPA) unweighted						0.164*** (0.0506)
Ln(Cov. Term RPA)						0.131 (0.122)
Education is recorded	-0.162 (0.172)	-0.152 (0.173)	-0.168 (0.174)	-0.156 (0.174)	-0.163 (0.174)	-0.161 (0.174)
edu_interact==1	-0.129 (0.124)	-0.129 (0.124)	-0.128 (0.125)	-0.129 (0.124)	-0.126 (0.125)	-0.127 (0.124)
edu_interact==2	0.184* (0.110)	0.184* (0.109)	0.181 (0.109)	0.183* (0.109)	0.183* (0.109)	0.183* (0.109)
edu_interact==3	0.578*** (0.127)	0.556*** (0.125)	0.567*** (0.126)	0.557*** (0.125)	0.560*** (0.126)	0.559*** (0.125)
edu_interact==4	1.238*** (0.176)	1.206*** (0.174)	1.223*** (0.175)	1.208*** (0.174)	1.209*** (0.175)	1.209*** (0.175)
Age HH head	0.0817*** (0.0123)	0.0821*** (0.0120)	0.0826*** (0.0120)	0.0823*** (0.0120)	0.0822*** (0.0119)	0.0822*** (0.0120)
(Age HH head) Squared	-0.000790*** (0.000161)	-0.000789*** (0.000157)	-0.000796*** (0.000157)	-0.000791*** (0.000158)	-0.000789*** (0.000156)	-0.000789*** (0.000156)
Gender	0.518*** (0.0522)	0.522*** (0.0525)	0.522*** (0.0526)	0.523*** (0.0525)	0.520*** (0.0525)	0.521*** (0.0522)
Total hrs Worked per week × Total hrs Worked per week	-0.000615*** (0.0000604)	-0.000632*** (0.0000601)	-0.000630*** (0.0000602)	-0.000633*** (0.0000601)	-0.000633*** (0.0000604)	-0.000633*** (0.0000604)
Total hrs Worked per week × Total hrs Worked per week × Total hrs Worked per week	0.00000461*** (0.000000671)	0.00000473*** (0.000000672)	0.00000472*** (0.000000670)	0.00000474*** (0.000000672)	0.00000474*** (0.000000673)	0.00000474*** (0.000000674)
Ln(Distance to port)	-0.0126 (0.0105)	-0.00801 (0.00772)	-0.00693 (0.00838)	-0.00753 (0.00754)	-0.00777 (0.00752)	-0.00783 (0.00743)
Ln(Market Access settlement (discount = 0.002))	0.390* (0.198)	0.457** (0.179)	0.474** (0.191)	0.466** (0.183)	0.517*** (0.183)	0.508** (0.199)
Constant	-6.704* (3.529)	-9.610*** (3.093)	-9.991*** (3.415)	-9.936*** (3.249)	-10.83*** (3.243)	-10.74*** (3.373)
Observations	9358	9358	9358	9358	9358	9358
R ²	0.344	0.350	0.349	0.350	0.350	0.350

Standard errors in parentheses

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

5.B Appendix B. Data description and methodology

The following sections provide further details on the data sources and methodologies for data preparation and usage in this paper. Section B.1 describes how we create city boundaries. Section B.2 describes the Living Standards Measurement Surveys and finally Section B.3 describes how we harmonize the two data sets to study how different density measures, as well as some miscellaneous items.

5.B.1 Using Landsat to create urbanized area boundaries

As discussed in the text, we use Landsat to define three unique types of urban areas: cores, fringes, and low-density towns. A city is the consolidation of its core and surrounding fringe, whereas a town is a unique and separate urbanized area. All of these urbanized area boundaries are defined based off density thresholds that we assign to each unique 1km grid-square in Landsat. As we are only using Landsat, these boundaries make no use of administrative borders to define urban extents.

To calculate the density of each grid-cell, we apply a smoothing methodology where each reference cell is assigned the average density from a neighborhood which includes itself and a 7x7km square around it, where the reference cell falls in the center of the 7x7km neighborhood square. This approach is essential for dealing with natural breaks in density in the data which may relate to changes in land use, terrain, and building restrictions within urban areas. Consider, for instance, Central Park in Manhattan, or the River Thames in London; assigning a density criterion just to each singular grid-cell would lead to unnatural breaks/holes in our urban area polygons which would challenge our ability to analyze our urban areas as singular units.

Once we have calculated the average density of each grid-cell within its 7x7km neighborhood, we consolidate all contiguous grid cells that have an average density above the thresholds discussed in the text for the type of urban unit we are creating: urban core, fringe, or town. Contiguous cells are combined based on a rook neighbor relationship, wherein, the rook neighbors are the four cells adjacent to the reference cell in the vertical or horizontal direction (as outlined on the left of Figure 5.B1), but not including the diagonally adjacent cells which are queen neighbors (as outlined on the right of Figure 5.B1). For all of Sub-Saharan Africa, the core and town definitions would only be affected in 3 out of 1885 cores and 43 of 3957 town or city polygons if we used the queen rather than rook criterion.

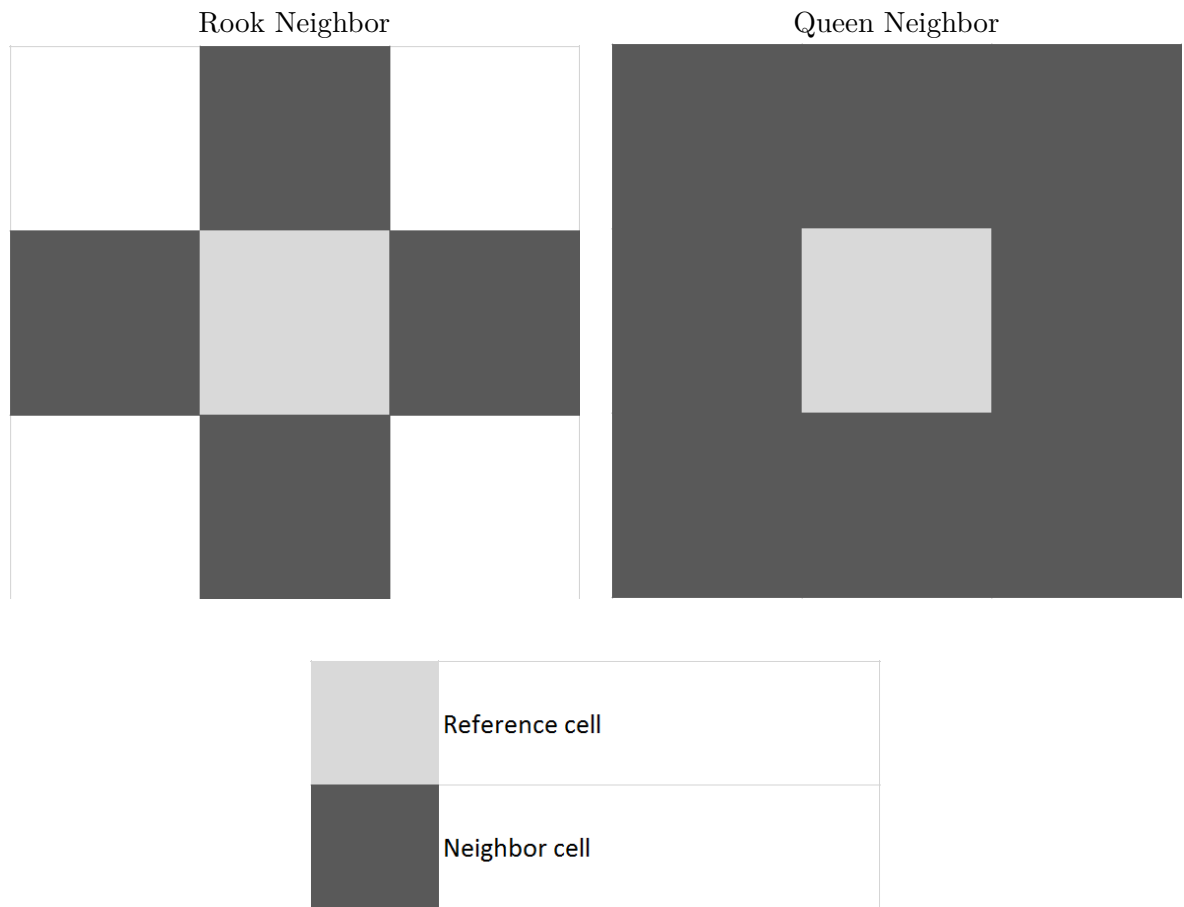


Figure 5.B1: Rook and queen neighbors

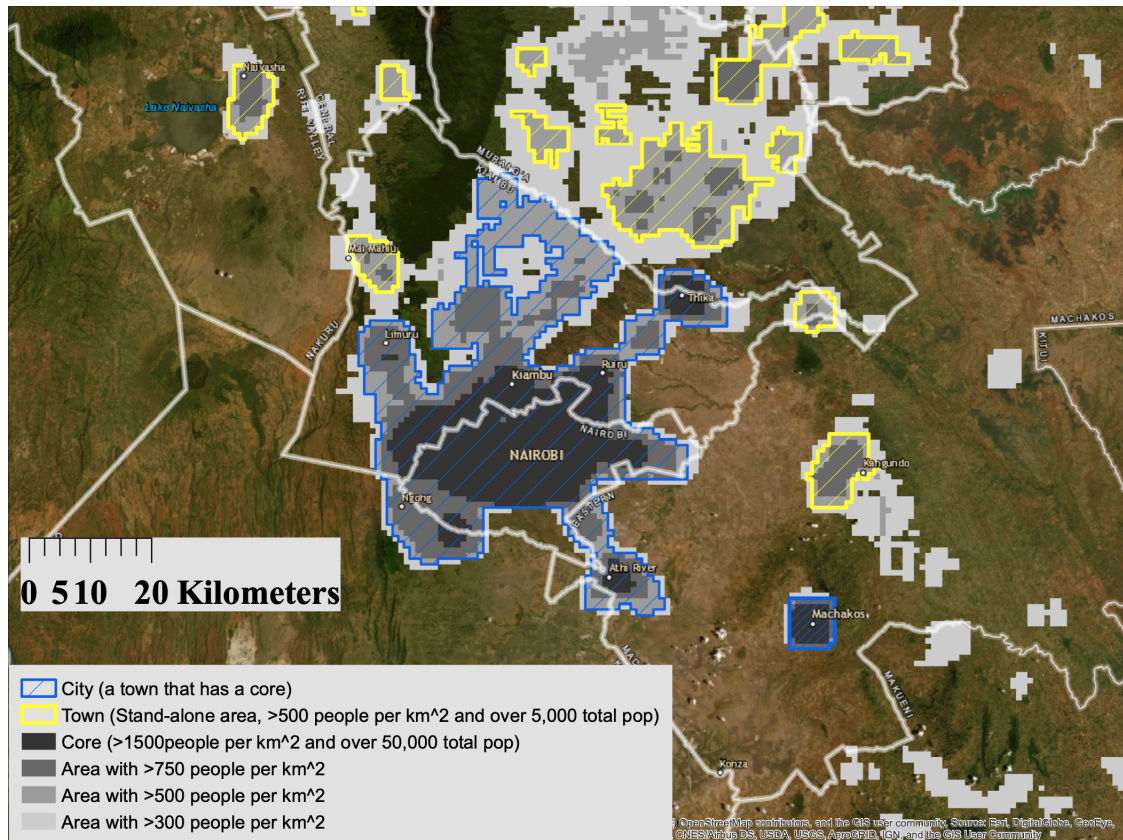


Figure 5.B2: Defining cities and towns (with labels and administrative boundaries)

5.B.2 Living Standards Measurement Surveys (LSMS)

The Living Standards Measurement Surveys have been conducted in a number of developing countries by the World Bank and the national statistical offices of the country in question. To study wage premia in this paper, we make use of surveys for Ethiopia, Ghana, Malawi, Nigeria, Tanzania, and Uganda. All of these surveys are considered representative of households at the national level, as well as urban/rural and major ecological zones of the countries. In Table 5.B1 below, we provide a list of the surveys we use and the number of households surveyed in each of these rounds.

Table 5.B1: LSMS surveys

Country	Survey	Year	Sample Size
Ethiopia	Socioeconomic Survey	2011	3,917
		2013	5,073
		2015	5,263
Ghana	Socioeconomic Panel Survey	2010	4,662
		2013	4,634
Malawi	Integrated Household Survey	2010	3,246
		2013	3,104
Tanzania	Panel Household Survey	2008	3,280
		2010	3,924
		2009	3,123
Uganda	National Panel Survey	2010	2,716
		2011	2,716
		2012	3,119
Nigeria	National Household Survey	2010	5,000
		2012	5,000

In each sample, a two-stage probability sampling methodology is used. In the first stage, “Primary Sample Units” (PSUs) are selected based on the probability proportional to size of all of the enumeration areas in geographic zones in the country. In the second stage, households are then selected randomly from each PSU, after which, each individual within a household is surveyed. All of the LSMS surveys are publicly available for download from the World Bank website, so for further information on any individual survey and its methodology, we refer the reader to the information documents provided by the World Bank.

Although the contents of each survey vary, they all have quite consistent data at the household and individual level on aspects such as income, educational attainment, demographics,

labor allocation, asset ownership and dwelling characteristics, as well as geographical identifiers locating the latitude and longitude of the centroid of each enumeration area.

Agricultural households report on various aspects of farming such as crop choice, inputs on the farm, labor usage and the types of land allocation such as harvesting, grazing or fallow. Among non-agricultural households, additional modules are provided on whether they are self-employed with their own business and if so the revenues and various factor costs of that business. In some cases, aggregation of revenues and costs at the household level is already computed in the survey and these aggregations are used where possible. For example, labor income at the individual level is already aggregated in the surveys to include all wage, in-kind and bonus income from all jobs. Elsewhere, input costs of agricultural and non-agricultural businesses are aggregated to the household level.

We calculate income from the survey data and aggregate either to the individual or household level (depending on our analysis) using all available sources of money flowing in. Letting i index an individual or household, this can be summarised as follows:

$$Y_i = \sum_i y_i^{SE} + \sum_i y_i^L + \sum_i y_i^K \quad (5.5)$$

where y_i^{SE} , y_i^L , and y_i^K represent self-employed income, labor income, and capital income respectively. Households reported receipts of incomes through various forms and over various time intervals. The variables used for income receipts and the time intervals over which they were received are reported as follows:

Income source	Time interval
Last payment in cash	Hour, Day, Week, Fortnight, Month, Quarter or Year
Last payment in kind (value in LCU)	Hour, Day, Week, Fortnight, Month, Quarter or Year
Net income from business	Week or Month
Remittances in cash	Year
Remittances in kind	Year
Rent of property	Year
Private or govt pensions	Year
Domestic remittances	Year
Rent of farmland	Year or cropping season
Sales of crops	Year or cropping season
Sales of crop residue	Year or cropping season
Sale of livestock products	Year or cropping season

Table 5.B2: Income sources and time intervals in LSMS surveys

All revenues are converted to a monthly interval. In cases where incomes are reported over

the year, quarter, fortnight, or week, the variables are scaled to a monthly value simply by multiplying by the ratio of a month to the time interval in question (for instance a quarter is multiplied by 1/3 to be monthly). In cases where the last income payment is reported based on a day of work, the figure is multiplied by the average days the respondent reports to work each week, and then multiplied by 52/12 to be a roughly monthly figure. In cases where the last income payment is reported based on an hour of work, the figure is multiplied by the average hours the respondent reports to work each day, then the average days they report to work each week, and finally by 52/12 to get a roughly monthly figure. A similar method is used to convert expenses to a monthly aggregate figure. The reported expenses in the household surveys are as follows:

Expense	Time interval
Wages	Month
Raw materials	Month
Other expenses	Month
Farm inputs	Year
Additional agricultural expenses	Year

Table 5.B3: Expenses and time intervals in LSMS surveys

For the household, we subtract expenses from revenues to get our net income figure. For the individual analysis we look at hourly wage rates, for adults aged 18-60 who are working part or full time with income. At the household level, measures of revenue include income from self-employment, labor, capital, and land income. All income is measured before taxes. We choose to include in the calculation of monthly household income, transfer payments such as remittances, gifts and pensions and we subtract transfer payments flowing out of the households. These sources of incomes are likely to be important for the budget constraints of households, particularly in rural communities, so they are important in our study of urban-rural income disparities, although we note that our results are robust to excluding remittances.

5.B.3 Harmonizing the data, local density and ring density measures, and the optimal de la Roca-Puga discount rate

Fortunately, the LSMS surveys provide latitude and longitude coordinates of enumeration areas in the sample for each country. This allows us to directly harmonize the surveys with our data from Landsat based on their spatial relationship.

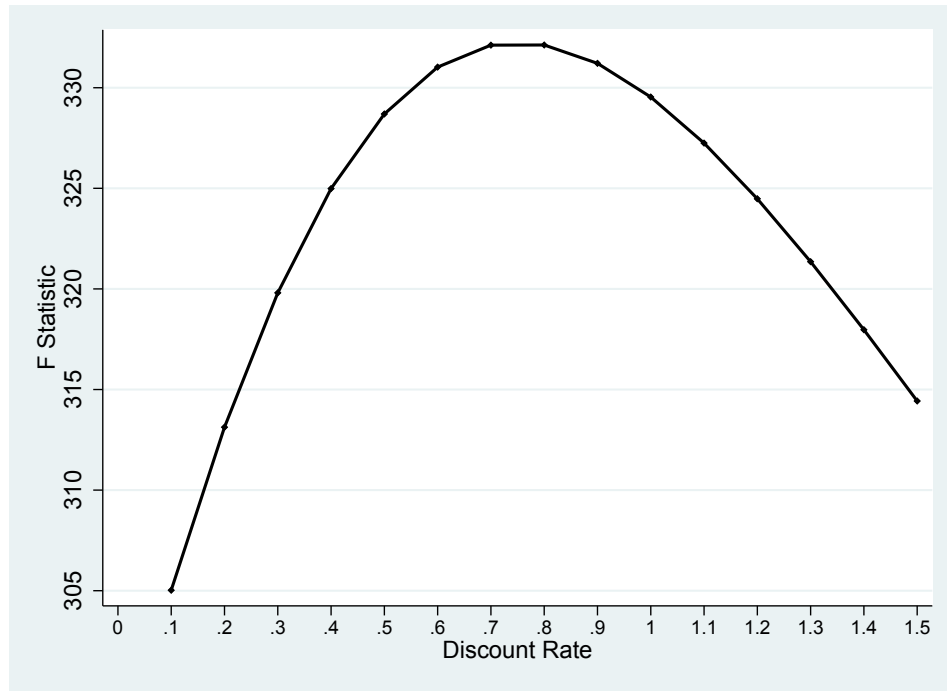


Figure 5.B3: Density ring measures

In the Section on economic density in cities, we describe how we calculate an optimal rate of discount. The figure plotting the relevant F-values is shown here in Figure 5.B3.

5.C Appendix C. Ground-truthing Landscan data within cities

In this Appendix we attempt a groundtruthing of Landscan data within two cities, where we have the detailed data to do so. This will give a sense of how good our local density measures are. In the upper panels of Figures C.1 and C.2, we show the population and employment distribution for Kampala and Nairobi in 1 km grid squares. The population data for Nairobi are at the level of 2,213 enumeration areas for 2009 contained in the 2015 built area of Nairobi defined in Henderson, Regan & Venables (2018). For Kampala in 2002, population is at the level of 174 parishes within the administrative unit of Greater Kampala. We assign population levels from these survey units to the 1km grid square level by applying a weighted sum to the survey area numbers, where the weights reflect the share of land mass from each survey area(s) that falls within a 1km grid square. To make Kampala 2002 population comparable to 2011 employment numbers, we blow up the population in each grid square by an overall population growth rate of 3% per annum from 2002 to 2011. For employment, we use the economic census, which covers private and public employment for Kampala for 2011, and provides exact location points of firms across the city. One issue is that total employment in the census is far below known estimates; hence, given the age distribution in Kampala and labor force participation of urban Uganda, we have multiplied each grid squares employment by 2.761 to make up for the employment deficit.¹² The implicit assumptions in allocating growth and under-counting of employment to grid squares are obvious.

For Nairobi in 2009, we can quite accurately infer population of the grid square, based on fine-scale enumeration area data. However, we do not know total employment, nor its distribution. We infer total employment based on Nairobi's 2009 population and labor force participation, and age distribution numbers for urban Kenya. Since there is no economic census for Nairobi, we obtain the distribution of employment using data from Henderson, Regan & Venables (2018), where for each grid square we know the footprint and height of every building in Nairobi in 2015 – from aerial photo and Lidar data – and can calculate building volume. We match these buildings with land use maps, before taking total employment of the city and smearing it into grid squares according to each grid square's share in total volume of non-residential buildings in Nairobi. Crucially, unlike

¹²The World Bank estimates that labor force participation of people aged 15 or more in Uganda is 0.71. There are 1,704,604 people of age 15+ in Kampala from the 2011 census. Thus, approximately 1,210,267 people should work out of the total city population of 2,957,505. The economic census only captures 438,374 of these.

Landscan, we do not need to base smearing on inferences from satellite images of what uses buildings have; instead, we know the use and the building volumes fairly accurately.

For each grid square, we create a measure of the ambient population according to the following equation:

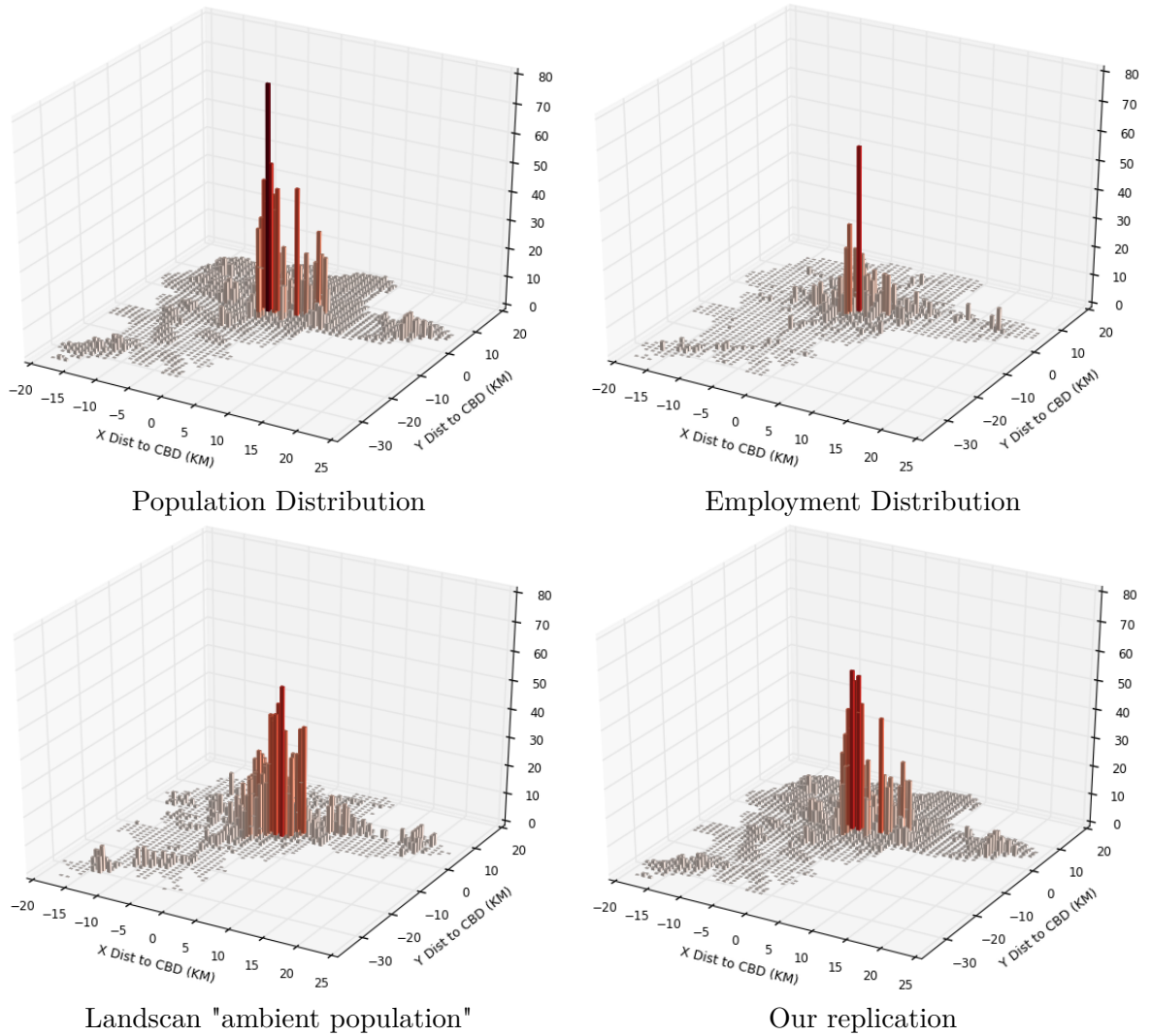
$$Replication_i = \left(\frac{10}{24}\right)Emp_i + \left(\frac{14}{24}\right)Pop_i + \left(\frac{10}{24}\right)(1 - LFP_c)Pop_i \quad (5.6)$$

where Emp_i , Pop_i , and LFP_c are respectively employment and population in grid square i and labor force participation in city c . We base our ‘replication’ of the ambient population just on places of work and residence, where we assume for 14 hours of a day (nocturnal) all people are in their grid square of residence to sleep, eat, and recreate. For 10 hours a day, we add in the employment in the grid square, allowing people time to work, hangout, and finish commuting. We then add in the non-working population of the grid square assuming they remain in that square kilometer. Finally, we subtract out the resident workers (since we have already counted employment), or $LFP_c * Pop_i$. If everyone works in their grid, then we just have total grid square population; but, for downtown grid squares where few people live, most have replication numbers from employment. We make no allowance for the time people are on roads or shopping outside the grid square. We have no information on which to base such inferences, especially in a context where so many people commute by walking.

In Figure ?? for Kampala, we show 4 items. As noted above, in the upper left panel of each figure is the population distribution over space and on the right upper panel is the employment distribution. In the bottom panel, on the left, we have Landscan numbers; and on the right, we have our replication numbers. For Kampala, we see the overall monocentricity of the city. Although it is hard to see, the very low bar population grid squares near the center are to some degree filled in by where employment spikes. The bottom right panel shows our smearing to get the ambient population. The Landscan figure has an obvious degree of smoothing, with reduced peak heights and assignment of lots of people into low-density grid squares. Of course, it could be that Landscan is allocating people during commuting times to roads and to shopping areas, and that is the reason for the smoothing. Overall it seems Landscan may do a reasonable job: the simple correlation coefficients of Landscan numbers with population, employment and our replication numbers are respectively 0.55, 0.60, and 0.60.

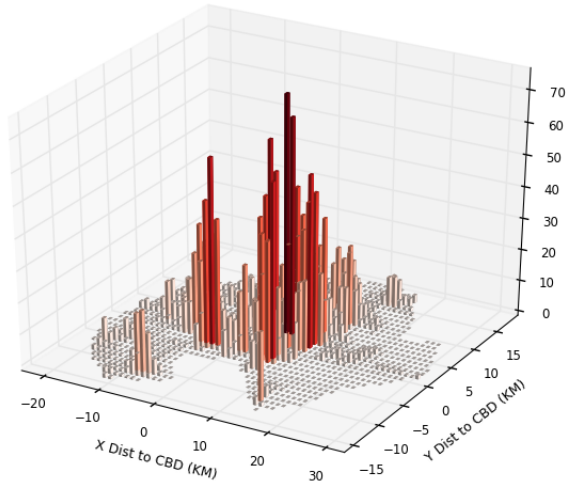
For Nairobi in Figure ??, we note our employment patterns lack the sharp peaks of Kam-

Figure C1: Kampala Densities

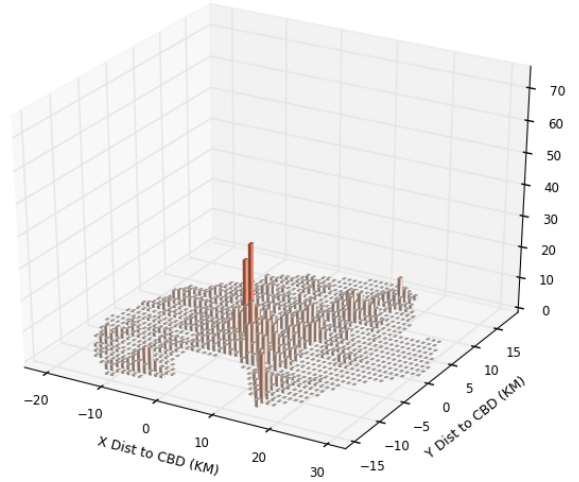


pala, in part because we smear employment into non-residential buildings, including public buildings. The Landscan figure again exhibits a degree of smoothing, missing the sharper peaks we see in our replication, as well as missing high-density slum areas to the south-west of the city center. However, Landscan does seem to do a better job of capturing low-density grid squares near the city center in Nairobi than it does for Kampala. For Nairobi, the simple correlation coefficients of Landscan numbers with population, employment and our replication numbers are generally higher at respectively 0.65, 0.56 and 0.69.

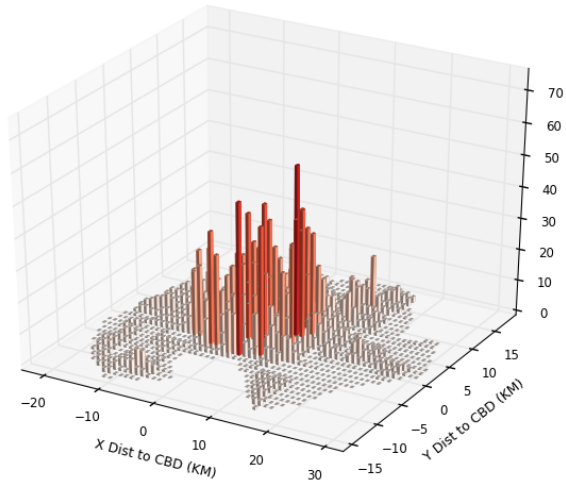
Figure C2: Nairobi Densities



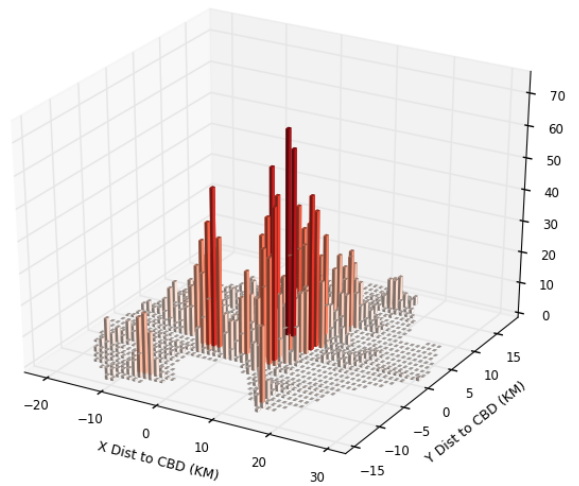
Population Distribution



Employment Distribution



Landscan "ambient population"



Our replication

??

5.D Appendix D. Neighborhood productivity and Kampala data

To study firm productivity at the neighborhood level in Kampala, we use data from the Uganda Business Inquiry (UBI) survey conducted in 2002. The UBI is an economic survey which made use of the official Census of Business Establishments (COBE) of 2002 as its sampling frame. The principal objective of the survey is to provide the necessary information and data to measure the contribution of each industry sector to the growth of the economy. So the survey covers a large range of economic variables, including value added and business assets. Coverage is comprehensive, with information on all sectors of the economy — including the informal sector¹³ — and coverage of all the officially recognized districts in Uganda. The sector definitions are in line with the International Standard Industrial Classification (ISIC), Revision 4, and cover 15 1-digit sectors.¹⁴ A stratified two-stage sample design was used to select the businesses for the UBI, which focused on getting all bigger establishments with over 50 employees and then sampling at the lower end.¹⁵

Although the UBI covers the entire country, due to confidentiality issues and limits on the capacity to share this data outside of the Uganda Bureau of Statistics (UBOS), our data have coverage only of Kampala and furthermore we are restricted by the abridged version of the data and statistics that have been provided to us. Our data cover the Greater Kampala metropolitan area for a sample of 2,342 firms, each with information on their GPS location, total number of employees, value added per worker, and industry classification. Value added is calculated as the net output of a sector after adding up all outputs and subtracting out all intermediate inputs but labor.

¹³The informal status of a business is recorded using a question in the survey on ‘whether a business pays any taxes’ as its determinant; this is in line with other surveys conducted in Uganda such as the UNHS which were more specifically directed to studying informality in the economy.

¹⁴These are agriculture & fishing, mining & quarrying, manufacturing, construction, utilities, trade, transport & storage, accommodation & food services, information & communication, finance & insurance, real estate & business services, education, health & social work, recreation and personal services. Note that the survey excludes information on the following ISIC sections: Section O, i.e., public administration and defense; compulsory social security, and Section U, i.e., activities of extraterritorial organizations and bodies.

¹⁵Business establishments were first stratified by industry sector and within each industry sector, business establishments were further stratified by employment size as per the following categories of employment size 1, 2-4, 5-9, 10-19, 20-49, 50-99, 100-499 and 500 or more employees; and further by turnover; thus less than 5 million shillings, between 5 and 10 million shillings and more than 10 million shillings. Given the significant contribution of the larger establishments, from previous surveys, to the value of production; all the establishments with 50 or more employees were supposed to be sampled with certainty, (a probability of 1), while those employing fewer than 50 employees were subjected to probabilistic sampling.

5.D.1 Matching the data to density measures

We measure the effects of density on firm productivity by matching the firm coordinates and data from the UBI survey with our population, employment, and Landscan ambient population data, used elsewhere in this paper. As detailed in Appendix C, our data on population are from the 2002 census and are at the level of 174 parishes within the Greater Kampala administrative area. We assign that data to the 1km grid square level, through a weighted sum to the survey area numbers, so as to be consistent with the spatial resolution of Landscan. Our employment data are from the 2002 Census of Business Establishments [COBE] and provides the GPS location of supposedly the universe of firms and employment in Greater Kampala.¹⁶

The data allow us to study density effects at a fine ring scale, based on 1km grid squares upon which population and Landscan data are defined. Each firm in the UBI dataset is overlaid by the 1km grid square level data on population, employment, and Landscan. Each firm is then assigned own-square measure based on the grid square in which it is located, as well as local neighborhood measures from the grid squares surrounding each firm's own-cell. Specifically, we define three rings around the own-cell to capture local neighborhood effects and the extent of spatial decay. Ring 1 captures the 8 cells in the immediate queen neighborhood of the reference cell, expanding incrementally by one cell in every direction. Ring 2 captures the 16 cells around ring 1, and finally ring 3 expands to the 24 cells around ring 2. For each ring, we calculate the mean density of 2002 population, 2002 employment 2002 own industry employment and 2012 Landscan ambient population¹⁷ For the 2002 COBE we start with that scale for total employment and own industry employment.

Table ?? columns 1 and 2 pool all types of industries. The columns give the results of regressions of the log of value added per worker for each firm in our sample, on measures of density in the firm's own-cell and its respective rings. For each regression, we include industry fixed effects and controls for distance to the city center, Lake Victoria, and the Kampala Northern Bypass. For measures of density, we don't show results for either census population or Landscan ambient population. The own cell and ring measures for these quantities have no significant effects. Columns 1 and 2 look respectively at the effect of total employment and own industry employment on firm productivity. In columns 1 and

¹⁶An issue is that if we look in the 2002 census data, matching it alongside the firm locations from our 2002 survey, we see that for over half of our firms we cannot find other firms from the census that are in the same industry and within 150 meters of the survey firms, indicating either that there is a lot of firm turnover or issues in recording locations.

¹⁷Obviously, the Landscan data are for much later. We do note the simple correlation coefficients on own industry employment in Table ?? right panel, between 2002 and 2012 census measures are all over 0.97 for the 4 rings.

2, there is a negative effect of greater employment density in the own-cell, with positive spillover effects occurring for the second ring in the radius around 2km away from the reference cells. It seems that there is a competition effect from nearby firms in the own cell and a positive externality effects in the second ring. The zero effect in the first ring suggests there the two forces cancel each other at 1 km away. But we may be over-interpreting.

To get a better sense of what is going on, we looked in industries where we think spillovers would be most relevant: manufacturing and business services. Manufacturing, in particular, as well as business services are archetypal industries that are not only exportable in scope, but also input intensive. As exporters either to outside the city or to all points within the city, such firms may not need to be close to very local population and consumers, but they need good connectivity to transport infrastructure. Moreover, such firms are likely to benefit more positively from knowledge spillovers and labor sharing, meaning there is a greater advantage of these firms to cluster together. We note we found no results for retail and personal services and other industries were too small to isolate.

In Table ?? columns 3 and 4, we focus on this sub-sample of firms. To maintain sample size, we pooled these two sets of industries. Column 4 is perhaps the most interesting column. It shows that the own ring has weak negative (competition) effects for this kind of activity, with ring 1 showing weak positive effects. However rings 3 and 4 have very strong localization economies. These are the classic results we expect: within industry strong externalities. These results are similar to those for advertising in Manhattan in Arzaghi & Henderson (2008), except the succession as we move away from the own firm of competition, spillover, and dissipation occurs at much shorter distances than in Kamapala. Column 3 is a bit more of a puzzle. It says total employment immediately nearby is bad for these traded good industries. If these are strong competition effects, from column 4, they are not so much from within the own industry, but must involve more general poaching of employees and suppliers or some type of congestion effect.

Table E1: Estimation of value added per worker at the firm level

	<i>All firms</i>		<i>Manu. & Business Services</i>	
	(1) Ln(Emp. 2002)	(2) Ln(Own Emp. 2002)	(3) Ln(Emp. 2002)	(4) Ln(Own Emp. 2002)
Ln(Own-Cell)	-0.0778*** (0.0242)	-0.0637*** (0.0190)	-0.158*** (0.0513)	-0.0287 (0.0374)
Ln(Ring 1)	-0.0765 (0.0493)	0.0140 (0.0335)	-0.241** (0.117)	0.0901 (0.0847)
Ln(Ring 2)	0.174** (0.0734)	0.125*** (0.0448)	0.412** (0.175)	0.407*** (0.109)
Ln(Ring 3)	-0.0199 (0.0632)	-0.0381 (0.0444)	0.0326 (0.162)	0.260** (0.121)
Observations	2342	2342	534	534
R-squared	0.087	0.086	0.107	0.143

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of value added per worker for each firm observation. Controls are dummies for the industry sector each firm is in, distance to the city center (km), distance to lake Victoria (km), and distance to the Kampala Northern Bypass Road (km). In columns 1-4, we provide results where the independent variables are calculated from rings made up of data in 1km grid cells; whereas, in columns 5-6, results are based on data within a circular neighborhood of each firm. Robust standard errors are given in parentheses. Asterisks indicate $p < 0.01$ (***), $p < 0.05$ (**), and $p < 0.1$ (*). *Source: UBI firm survey.*

Bibliography

- Adamopoulos, T., Brandt, L., Leight, J. & Restuccia, D. (2017), Misallocation, selection and productivity: A quantitative analysis with panel data from china, Technical report, National Bureau of Economic Research.
- Addison, D. M. & Stewart, B. (2015), ‘Nighttime lights revisited: the use of nighttime lights data as a proxy for economic variables’, *World Bank Policy Research Working Paper* (7496).
- Ahlfeldt, G. M., Redding, S. J., Sturm, D. M. & Wolf, N. (2015), ‘The economics of density: Evidence from the berlin wall’, *Econometrica* **83**(6), 2127–2189.
- Ahn, D. P. & Ludema, R. D. (2020), ‘The sword and the shield: the economics of targeted sanctions’, *European Economic Review* **130**, 103587.
- Akbar, P. A., Couture, V., Duranton, G., Ghani, E. & Storeygard, A. (2018), *Mobility and congestion in urban India*, The World Bank.
- Albouy, D., Behrens, K., Robert-Nicoud, F. & Seegert, N. (2019), ‘The optimal distribution of population across cities’, *Journal of Urban Economics* **110**, 102–113.
- Ali, D. A., Collin, M., Deininger, K., Dercon, S., Sandefur, J. & Zeitlin, A. (2016), ‘Small price incentives increase women’s access to land titles in Tanzania’, *Journal of Development Economics* **123**, 107–122.
URL: <https://linkinghub.elsevier.com/retrieve/pii/S030438781630044X>
- Alonso, W. et al. (1964), ‘Location and land use. toward a general theory of land rent.’, *Location and land use. Toward a general theory of land rent.* .
- Angel, S. (2012), *Planet of cities*, Lincoln Institute of Land Policy, Cambridge, Mass.
- Angel, S., Blei, A. M., Jason, P., Lamson-Hall, P., Sánchez, N. G., Civco, D. L., Lei, R. Q. & Thom, K. (2016), *Atlas of Urban Expansion—2016 Edition, Vol. 1*, New York University, Lincoln Land Institute, and UN-Habitat.

- Arzaghi, M. & Henderson, J. V. (2008), 'Networking off Madison avenue', *The Review of Economic Studies* **75**(4), 1011–1038.
- Asher, S., Lunt, T., Matsuura, R. & Novosad, P. (2021), 'Development research at high geographic resolution'.
- Asker, J., Collard-Wexler, A. & De Loecker, J. (2014), 'Dynamic inputs and resource (mis) allocation', *Journal of Political Economy* **122**(5), 1013–1063.
- Banerjee, A., Duflo, E., Glennerster, R. & Kinnan, C. (2015), 'The Miracle of Microfinance? Evidence from a Randomized Evaluation', *American Economic Journal: Applied Economics* **7**(1), 22–53.
URL: <https://pubs.aeaweb.org/doi/10.1257/app.20130533>
- Bank, W. (1974a), Appraisal of national sites and services project. report no.337a-ta, Technical report.
- Bank, W. (1974b), Development credit agreement between united republic of tan-zania and international development association (conformed copy), Technical report.
- Bank, W. (1977a), Development credit agreement between united republic of tan-zania and international development association (conformed copy). credit no. 732 ta, Technical report.
- Bank, W. (1977b), Tanzania: The second national sites and project. report no.1518a-ta, Technical report.
- Bank, W. (1984), Completion report: Tanzania - first national sites and services project. report no. 4941, Technical report.
- Bank, W. (1987), Tanzania: The second national sites and project. report no.6828, Technical report.
- Bank, W. (2010), Project appraisal document for a tanzania strategic cities project. report no. 51881-tz, Technical report.
- Bank, W. (2013), Tanzania strategic cities project housing survey, Technical report.
- Baqae, D. R. & Farhi, E. (2020), 'Productivity and misallocation in general equilibrium', *The Quarterly Journal of Economics* **135**(1), 105–163.
- Baragwanath, K., Goldblatt, R., Hanson, G. & Khandelwal, A. K. (2019), 'Detecting urban markets with satellite imagery: An application to india', *Journal of Urban Economics* p. 103173.

- Barrios, S., Bertinelli, L. & Strobl, E. (2006), ‘Climatic change and rural–urban migration: The case of sub-saharan africa’, *Journal of Urban Economics* **60**(3), 357–371.
- Bartelsman, E., Haltiwanger, J. & Scarpetta, S. (2013), ‘Cross-country differences in productivity: The role of allocation and selection’, *American economic review* **103**(1), 305–34.
- Baruah, N. G., Henderson, J. V. & Peng, C. (2017), ‘Colonial legacies: shaping African cities’, *London School of Economics and Political Science, SERC DP 0226* .
- Bau, N. & Matray, A. (2020), Misallocation and capital market integration: Evidence from india, Technical report, National Bureau of Economic Research.
- Baum-Snow, N. & Pavan, R. (2011), ‘Understanding the city size wage gap’, *The Review of Economic Studies* **79**(1), 88–127.
- Bayer, P., Ferreira, F. & McMillan, R. (2007), ‘A Unified Framework for Measuring Preferences for Schools and Neighborhoods’, *Journal of Political Economy* **115**(4), 588–638.
URL: <https://www.journals.uchicago.edu/doi/10.1086/522381>
- Berkowitz, D., Ma, H. & Nishioka, S. (2017), ‘Recasting the iron rice bowl: The reform of china’s state-owned enterprises’, *Review of Economics and Statistics* **99**(4), 735–747.
- Bester, C. A., Conley, T. G. & Hansen, C. B. (2011), ‘Inference with dependent data using cluster covariance estimators’, *Journal of Econometrics* **165**(2), 137–151.
URL: <https://linkinghub.elsevier.com/retrieve/pii/S0304407611000431>
- Bickenbach, F., Bode, E., Nunnenkamp, P. & Söder, M. (2016), ‘Night lights and regional gdp’, *Review of World Economics* **152**(2), 425–447.
- Bils, M., Klenow, P. J. & Ruane, C. (2020), Misallocation or mismeasurement?, Technical report, National Bureau of Economic Research.
- Bleakley, H. & Lin, J. (2012), ‘Portage and Path Dependence *’, *The Quarterly Journal of Economics* **127**(2), 587–644.
URL: <https://academic.oup.com/qje/article-lookup/doi/10.1093/qje/qjs011>
- Bluhm, R. & Krause, M. (2016), ‘Top lights’.
- Brandt, L., Jiang, F., Luo, Y. & Su, Y. (2018), ‘Ownership and productivity in vertically-integrated firms: Evidence from the chinese steel industry’, *Review of Economics and Statistics* pp. 1–49.

- Brown, J. D., Earle, J. S. & Telegdy, A. (2006), ‘The productivity effects of privatization: Longitudinal estimates from hungary, romania, russia, and ukraine’, *Journal of political economy* **114**(1), 61–99.
- Brückner, M. (2012), ‘Economic growth, size of the agricultural sector, and urbanization in africa’, *Journal of Urban Economics* **71**(1), 26–36.
- Buckley, R. M. & Kalarickal, J., eds (2006), *Thirty years of World Bank shelter lending: what have we learned?*, Directions in development, World Bank, Washington, DC. OCLC: ocm64487249.
- Buera, F. J., Kaboski, J. P. & Shin, Y. (2011), ‘Finance and development: A tale of two sectors’, *American economic review* **101**(5), 1964–2002.
- Busso, M., Madrigal, L. & Pagés, C. (2013), ‘Productivity and resource misallocation in latin america’, *The BE Journal of Macroeconomics* **13**(1), 903–932.
- Bussolo, M., De Nicola, F., Panizza, U. & Varghese, R. (2019), ‘Political connections and financial constraints: Evidence from transition countries’, *World Bank Policy Research Working Paper* (8956).
- Bustos, P., Caprettini, B. & Ponticelli, J. (2016), ‘Agricultural productivity and structural transformation: Evidence from Brazil’, *American Economic Review* **106**(6), 1320–65.
- Cao, X., Hu, Y., Zhu, X., Shi, F., Zhuo, L. & Chen, J. (2019), ‘A simple self-adjusting model for correcting the blooming effects in dmsp-ols nighttime light images’, *Remote Sensing of Environment* **224**, 401–411.
- Castells-Quintana, D. (2017), ‘Malthus living in a slum: Urban concentration, infrastructure and economic growth’, *Journal of Urban Economics* **98**, 158–173.
URL: <https://linkinghub.elsevier.com/retrieve/pii/S0094119016000115>
- Center, M. P. (2017), ‘Integrated public use microdata series, inter-national: Version 6.5 [dataset]’.
- Chauvin, J. P., Glaeser, E., Ma, Y. & Tobio, K. (2017), ‘What is different about urbanization in rich and poor countries? cities in {Brazil, {China, {India and the {United States’, *Journal of Urban Economics* **98**, 17–49.
- Chen, X. & Nordhaus, W. D. (2011), ‘Using luminosity data as a proxy for economic statistics’, *Proceedings of the National Academy of Sciences* **108**(21), 8589–8594.
- Ciccone, A. & Hall, R. E. (1996), ‘Productivity and the density of economic activity’, *The*

- American Economic Review* **86**(1), 54–70.
URL: <http://www.jstor.org/stable/2118255>
- City Population - Population Statistics in Maps and Charts for Cities, Agglomerations and Administrative Divisions of all Countries of the World* (n.d.).
URL: <http://citypopulation.de/>
- Cohen, M. A. (1983), *Learning by doing: World Bank lending for urban development, 1972-82*, World Bank, Washington, D.C., U.S.A.
- Collier, P. & Jones, P. (2016), *Transforming Dar es Salaam into a City that Works*, Oxford University Press.
- Collier, P., Jones, P. & Spijkerman, D. (2018), ‘Cities as engines of growth: Evidence from a new global sample of cities’, *Unpublished*.
- Collin, M., Sandefur, J., Zeitlin, A. et al. (2015), ‘Falling off the map: The impact of formalizing (some) informal settlements in tanzania’, *Centre for the Study of African Economies Working Paper, University of Oxford*.
- Combes, P.-P., Démurger, S., Li, S. & Wang, J. (2019), ‘Unequal migration and urbanisation gains in China’, *Journal of Development Economics* **forthcoming**.
- Combes, P.-P., Duranton, G. & Gobillon, L. (2008), ‘Spatial Wage Disparities: Sorting Matters!’, *Journal of Urban Economics* **63**, 723–742.
- Combes, P.-P., Duranton, G. & Gobillon, L. (2017), ‘The Production Function for Housing: Evidence from France’, *SSRN Electronic Journal*.
URL: <https://www.ssrn.com/abstract=3090324>
- Combes, P.-P. & Gobillon, L. (2015), The empirics of agglomeration economies, in J. G. Duranton & W. Strange, eds, ‘Handbook of regional and urban economics’, Vol. 5, Elsevier, pp. 247–348.
- Conley, T. (1999), ‘GMM estimation with cross sectional dependence’, *Journal of Econometrics* **92**(1), 1–45.
URL: <https://linkinghub.elsevier.com/retrieve/pii/S0304407698000840>
- Coville, A. & Su, Y.-h. (2014), ‘From the ground up: An impact evaluation of the community infrastructure upgrading programme in dar-es-salaam’, *World Bank, mimeo*.
- Crozet, M. & Hinz, J. (2016), ‘Collateral damage: The impact of russia sanctions on sanctioning countries’ exports cepii working paper’.

- David, J. M., Hopenhayn, H. A. & Venkateswaran, V. (2016), ‘Information, misallocation, and aggregate productivity’, *The Quarterly Journal of Economics* **131**(2), 943–1005.
- David, J. M. & Venkateswaran, V. (2019), ‘The sources of capital misallocation’, *American Economic Review* **109**(7), 2531–67.
- de Bellefon, M.-P., Combes, P.-P., Duranton, G., Gobillon, L. & Gorin, C. (2018), ‘Delineating urban areas using building density’, *Processed, University of Pennsylvania* .
- De La Roca, J. & Puga, D. (2017), ‘Learning by working in big cities’, *The Review of Economic Studies* **84**(1), 106–142.
- Dell, M. (2010), ‘The persistent effects of peru’s mining mita’, *Econometrica* **78**(6), 1863–1903.
- Desmet, K., Gomes, J. & Ortuño-Ortín, I. (2018), ‘The geography of linguistic diversity and the provision of public goods’, *National Bureau of Economic Research, Working Paper 24694* .
- Diewert, W. E., Haan, J. d. & Hendriks, R. (2015), ‘Hedonic regressions and the decomposition of a house price index into land and structure components’, *Econometric Reviews* **34**(1-2), 106–126.
- DigitalGlobe (2016), ‘Worldview satellite imagery’.
- Dingel, J. I., Miscio, A. & Davis, D. R. (2019), ‘Cities, lights, and skills in developing economies’, *Journal of Urban Economics, forthcoming* .
- Draca, M., Garred, J., Stickland, L. & Warrinnie, N. (2019), On target? the incidence of sanctions across listed firms in iran, Technical report, LICOS Discussion Paper.
- Duranton, G. (2015), *Growing through cities in developing countries*, The World Bank Research Observer, Volume 30, Issue 1, February 2015, Pages 39—73.
- Duranton, G. (2016), ‘Agglomeration effects in Colombia’, *Journal of Regional Science* **56**(2), 210–238.
- Duranton, G., Henderson, V. & Strange, W. (2015), *Handbook of regional and urban economics*, Elsevier.
- Duranton, G. & Overman, H. G. (2005), ‘Testing for localization using micro-geographic data’, *The Review of Economic Studies* **72**(4), 1077–1106.

- Duranton, G. & Venables, A. J. (2018), *Place-based policies for development*, The World Bank.
- EBRD (2020), *The state strikes back*, Technical report, European Bank of Reconstruction and Development.
- Eckert, F., Gvirtz, A., Liang, J. & Peters, M. (2020), ‘A method to construct geographical crosswalks with an application to us counties since 1790’, *NBER Working Paper* (w26770).
- Eeckhout, J. (2004), ‘Gibrat’s law for (all) cities’, *American Economic Review* **94**(5), 1429–1451.
- Ellison, G. & Glaeser, E. L. (1997), ‘Geographic concentration in us manufacturing industries: a dartboard approach’, *Journal of political economy* **105**(5), 889–927.
- Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F.-C. & Taneja, J. (2021), ‘Annual time series of global viirs nighttime lights derived from monthly averages: 2012 to 2019’, *Remote Sensing* **13**(5), 922.
- Epple, D., Gordon, B. & Sieg, H. (2010), ‘Drs. muth and mills meet dr. tiebout: Integrating location-specific amenities into multi-community equilibrium models’, *Journal of regional science* **50**(1), 381–400.
- Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., Roth, A., Üreyen, S., Zeidler, J., Dech, S. et al. (2018), ‘Where we live—a summary of the achievements and planned evolution of the global urban footprint’, *Remote Sensing* **10**(6), 895.
- European Commission, Joint Research Centre (JRC); Columbia University, C. f. I. E. S. I. N. C. (2015), ‘Ghs population grid, derived from gpw4, multitemporal (1975, 1990, 2000, 2015)’.
URL: http://data.europa.eu/89h/jrc-ghsl-ghs_op_gpw4_globe_r2015a
- Fay, M. & Opal, C. (1999), *Urbanization without growth: a not-so-uncommon phenomenon*, The World Bank, Policy Research Working Paper.
- Feenstra, R. C., Inklaar, R. & Timmer, M. P. (2015), ‘The next generation of the penn world table’, *American economic review* **105**(10), 3150–82.
- Field, E. (2005), ‘Property rights and investment in urban slums’, *Journal of the European Economic Association* **3**(2-3), 279–290.

- Franklin, S. (2018), ‘Location, search costs and youth unemployment: experimental evidence from transport subsidies’, *The Economic Journal* **128**(614), 2353–2379.
- Franklin, S. (2020), ‘Enabled to work: The impact of government housing on slum dwellers in south africa’, *Journal of Urban Economics* **118**, 103265.
- Freire, M. E., Lall, S. & Leipziger, D. (2014), ‘Africa’s urbanization: Challenges and opportunities’, *The growth dialogue* **7**, 1–30.
- Fujita, M. & Ogawa, H. (1982), ‘Multiple equilibria and structural transition of non-monocentric urban configurations’, *Regional Science and Urban Economics* **12**(2), 161–196.
- Galiani, S., Gertler, P. J., Undurraga, R., Cooper, R., Martínez, S. & Ross, A. (2017), ‘Shelter from the storm: Upgrading housing infrastructure in latin american slums’, *Journal of urban economics* **98**, 187–213.
- Galiani, S. & Schargrodsky, E. (2010), ‘Property rights for the poor: Effects of land titling’, *Journal of Public Economics* **94**(9-10), 700–729.
- Garreau, J. (1991), *Edge City: Life on the New Frontier*, Doubleday.
URL: <https://books.google.ru/books?id=UWJPAAAAMAAJ>
- Gelman, A. & Imbens, G. (2019), ‘Why high-order polynomials should not be used in regression discontinuity designs’, *Journal of Business & Economic Statistics* **37**(3), 447–456.
- Geospatial Geoscience Ltd (2017), ‘World port index’, <https://www.arcgis.com/home/item.html?id=dd8823d9502e48c89058fc8f2c4e96ba>.
- Gibson, J., Olivia, S., Boe-Gibson, G. & Li, C. (2021), ‘Which night lights data should we use in economics, and where?’, *Journal of Development Economics* **149**, 102602.
- Giglio, S., Maggiori, M. & Stroebel, J. (2015), ‘Very long-run discount rates’, *The Quarterly Journal of Economics* **130**(1), 1–53.
- Glaeser, E. L. & Mare, D. C. (2001), ‘Cities and skills’, *Journal of Labor Economics* **19**(2), 316–342.
- Gold, R., Hinz, J. & Valsecchi, M. (2019), To russia with love? the impact of sanctions on elections, Technical report.
- Gollin, D., Jedwab, R. & Vollrath, D. (2016), ‘Urbanization with and without industrialization’, *Journal of Economic Growth* **21**(1), 35–70.

- Gollin, D., Kirchberger, M. & Lagakos, D. (2017), In search of a spatial equilibrium in the developing world, Technical report, National Bureau of Economic Research.
- Gollin, D., Lagakos, D. & Waugh, M. E. (2013), ‘The agricultural productivity gap’, *The Quarterly Journal of Economics* **129**(2), 939–993.
- Gollin, D., Parente, S. L. & Rogerson, R. (2007), ‘The food problem and the evolution of international income levels’, *Journal of Monetary Economics* **54**(4), 1230–1255.
- Gollin, D. & Udry, C. (2021), ‘Heterogeneity, measurement error, and misallocation: Evidence from african agriculture’, *Journal of Political Economy* **129**(1), 000–000.
- Gopinath, G., Kalemli-Özcan, Ş., Karabarbounis, L. & Villegas-Sanchez, C. (2017), ‘Capital allocation and productivity in south europe’, *The Quarterly Journal of Economics* **132**(4), 1915–1967.
- Habitat, U. (2013), *State of the world’s cities 2012/2013: Prosperity of cities*, Routledge.
- Haidar, J. I. (2017), ‘Sanctions and export deflection: evidence from iran’, *Economic Policy* **32**(90), 319–355.
- Hanson, G. H. (2005), ‘Market potential, increasing returns and geographic concentration’, *Journal of international economics* **67**(1), 1–24.
- Harari, M. (2016), ‘Cities in bad shape: Urban geometry in India’, *Processed, Wharton School of the University of Pennsylvania* .
- Harari, M. & Wong, M. (2017), ‘Long-term impacts of slum upgrading: Evidence from the kampung improvement program in indonesia’, *NBER Working Paper* .
- Heblich, S., Redding, S. J. & Sturm, D. M. (2018), ‘The making of the modern metropolis: evidence from London’, *National Bureau of Economic Research, Working Paper 25047* .
- Henderson, J. V. & Kriticos, S. (2018), ‘The development of the African system of cities’, *Annual Review of Economics* **10**, 287–314.
- Henderson, J. V., Nigmatulina, D. & Kriticos, S. (2018), ‘Measuring urban economic density’, *CEP Working Paper CEPDP1569* .
- Henderson, J. V., Nigmatulina, D. & Kriticos, S. (2019), ‘Measuring urban economic density’, *Journal of Urban Economics* p. 103188.
- Henderson, J. V., Regan, T. & Venables, A. (2018), ‘Building the city: Urban transition and institutional frictions’, *Processed London School of Economics* .

- Henderson, J. V., Roberts, M. & Storeygard, A. (2013), *Is Urbanization in Sub-Saharan Africa Different?*, The World Bank.
- Henderson, J. V., Storeygard, A. & Deichmann, U. (2017), ‘Has climate change driven urbanization in Africa?’, *Journal of Development Economics* **124**, 60–82.
- Henderson, J. V., Storeygard, A. & Weil, D. N. (2012), ‘Measuring economic growth from outer space’, *American economic review* **102**(2), 994–1028.
- Henderson, J. V., Storeygard, A. & Weil, D. N. (2020), Quality-adjusted population density, Technical report, National Bureau of Economic Research.
- Henderson, J. V., Venables, A. J., Regan, T. & Samsonov, I. (2016), ‘Building functional cities’, *Science* **352**(6288), 946–947.
- Henderson, R. & Cockburn, I. (1996), ‘Scale, scope, and spillovers: the determinants of research productivity in drug discovery’, *The Rand journal of economics* pp. 32–59.
- Hopenhayn, H. A. (2014), ‘Firms, misallocation, and aggregate productivity: A review’, *Annu. Rev. Econ.* **6**(1), 735–770.
- Hornbeck, R. & Keniston, D. (2017), ‘Creative destruction: Barriers to urban growth and the great boston fire of 1872’, *American Economic Review* **107**(6), 1365–98.
- Hsieh, C.-T. & Klenow, P. J. (2009), ‘Misallocation and manufacturing tfp in china and india’, *The Quarterly journal of economics* **124**(4), 1403–1448.
- Hsieh, C.-T. & Song, Z. M. (2015), ‘Grasp the large, let go of the small: The transformation of the state sector in china’, *Brookings Papers on Economic Activity* .
- <https://www.openstreetmap.org/> (2017).
URL: <https://www.openstreetmap.org/>
- Imbens, G. & Kalyanaraman, K. (2012), ‘Optimal bandwidth choice for the regression discontinuity estimator’, *The Review of economic studies* **79**(3), 933–959.
- Initiative, T. O. D. et al. (n.d.), ‘Ramani huria—community mapping in dar es salaam’.
URL: <http://ramanihuria.org/data/>
- International, I. (n.d.), ‘Tanzania population census 2012’.
URL: <https://international.ipums.org/international/>
- Jaupart, P., Chen, Y. & Picarelli, N. (2017), ‘Cholera in times of floods. weather shocks & health impacts in dar es salaam’.

- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016), 'Combining satellite imagery and machine learning to predict poverty', *Science* **353**(6301), 790–794.
- Jones, C. I. (2011), Misallocation, economic growth, and input-output economics, Technical report, National bureau of economic research.
- Kironde, J. (2015), 'Good governance, efficiency and the provision of planned land for orderly development in african cities: The case of the 20,000 planned land plots project in dar es salaam, tanzania. current urban studies, 3, 348-367'.
- Kironde, J. L. (1991), 'Sites-and-services in tanzania: The case of sinza, kijitonyama and mikocheni areas in dar-es-salaam', *Habitat International* **15**(1-2), 27–38.
- Kironde, J. L. (1992), 'Creations in dar es salaam and extensions in nairobi: The defiance of inappropriate planning standards', *Cities* **9**(3), 220–231.
- Kironde, J. L. (n.d.), 'The evolution of the land use structure of dar essalaam 1890-1990: A study in the effects of land policy', *PhD dissertation, University of Nairobi* .
- Kling, J. R., Liebman, J. B. & Katz, L. F. (2007), 'Experimental analysis of neighborhood effects', *Econometrica* **75**(1), 83–119.
- Komu, F. (2013), 'Rental housing in tanzania: Power and dialects of misidentification', *Journal of Building and Land Development, ppecial* (29-41).
- Kowalski, P., Büge, M., Sztajerowska, M. & Egeland, M. (2013), 'State-owned enterprises: Trade effects and policy implications'.
- Krugman, P. (1991), 'Increasing returns and economic geography', *Journal of political economy* **99**(3), 483–499.
- Lall, S. V., Henderson, J. V. & Venables, A. J. (2017), *Africa's cities: Opening doors to the world*, World Bank Publications.
- LandScan (2012), *Oak Ridge National Laboratory* <https://landscan.ornl.gov/>.
- Laquian, A. A. (1983), 'Sites, services and shelter—an evaluation', *Habitat International* **7**(5-6), 211–225.
- Li, X., Li, D., Xu, H. & Wu, C. (2017), 'Intercalibration between dmsp/ols and viirs nighttime light images to evaluate city light dynamics of syria's major human settlement during syrian civil war', *International Journal of Remote Sensing* **38**(21), 5934–5951.

- Li, X. & Zhou, Y. (2017), 'A stepwise calibration of global dmisp/ols stable nighttime light data (1992–2013)', *Remote Sensing* **9**(6), 637.
- Libecap, G. D. & Lueck, D. (2011), 'The demarcation of land and the role of coordinating property institutions', *Journal of Political Economy* **119**(3), 426–467.
- Ltd, P. P. A. (1968), National capital master plan: Dar-es-salaam main plan report, Technical report.
- Lupala, J., Malombe, J. & Könye, A. (1997), 'Evaluation of hanna nassif community-based urban upgrading project phase i', *Government of Tanzania/UNDP/NIGP/Ford Foundation/ILO Evaluation Mission Team Report, Dar-es-Salaam* .
- Manara, M. & Regan, T. (2020), 'Eliciting demand for title deeds: lab-in-the-field evidence from urban tanzania'.
- Marshall, A. & Marshall, M. P. (1920), *The economics of industry*, Macmillan and Company.
- Marshall, Macklin, M. L. (1979), 'The dar-es-salaam master plan: Main report - five year development programme'.
- Marx, B., Stoker, T. M. & Suri, T. (2019), 'There is no free house: Ethnic patronage in a kenyan slum', *American Economic Journal: Applied Economics* **11**(4), 36–70.
- Marx, B., Stoker, T. & Suri, T. (2013), 'The economics of slums in the developing world', *Journal of Economic perspectives* **27**(4), 187–210.
- Massey, D. S. & Denton, N. A. (1988), 'The dimensions of residential segregation', *Social forces* **67**(2), 281–315.
- Matsuyama, K. (1992), 'Agricultural productivity, comparative advantage, and economic growth', *Journal of Economic Theory* **58**(2), 317–334.
- Mayo, S. K. & Gross, D. J. (1987), 'Sites and services—and subsidies: The economics of low-cost housing in developing countries', *The World Bank Economic Review* **1**(2), 301–335.
- McMillen, D. P. (2003), 'Identifying sub-centres using contiguity matrices', *Urban Studies* **40**(1), 57–69.
- McMillen, D. P. & Redfearn, C. L. (2010), 'Estimation and hypothesis testing for nonparametric hedonic house price functions', *Journal of Regional Science* **50**(3), 712–733.

- Meggison, W. L. (2016), 'Privatization, state capitalism, and state ownership of business in the 21st century', *Foundations and Trends in Finance*, forthcoming .
- Mellander, C., Lobo, J., Stolarick, K. & Matheson, Z. (2015), 'Night-time light data: A good proxy measure for economic activity?', *PloS one* **10**(10), e0139779.
- Michalopoulos, S. & Papaioannou, E. (2013), 'Pre-colonial ethnic institutions and contemporary african development', *Econometrica* **81**(1), 113–152.
- Michalopoulos, S. & Papaioannou, E. (2014), 'National institutions and subnational development in africa', *The Quarterly journal of economics* **129**(1), 151–213.
- Midrigan, V. & Xu, D. Y. (2014), 'Finance and misallocation: Evidence from plant-level data', *American economic review* **104**(2), 422–58.
- Mills, E. S. (1967), 'An aggregative model of resource allocation in a metropolitan area', *The American Economic Review* **57**(2), 197–210.
- Montenegro, C. E. & Patrinos, H. A. (2014), *Comparable estimates of returns to schooling around the world*, The World Bank.
- Municipality, M. C. (1973), 'Mwanza cadastral maps'.
- Muth, R. F. (1969), 'Cities and housing; the spatial pattern of urban residential land use?.'
- Nordhaus, W., X. Chen; Palisades, N. N. S. D. & (SEDAC)., A. C. (2018), 'Global gridded geographically based economic data (g-econ), version 4?.'
- URL:** <http://doi.org/10.7927/H42V2D1C>
- Nunn, N. & Puga, D. (2012), 'Ruggedness: The blessing of bad geography in Africa', *Review of Economics and Statistics* **94**(1), 20–36.
- Oceanic, N. & Center, A. A. N. G. D. (2012), 'Version 4 dmsp-ols nighttime lights time series?.'
- URL:** <https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html>
- OECD (2012), *Redefining Urban: A New Way to Measure Metropolitan Areas*, Organisation for Economic Cooperation and Development (OECD).
- of Lands, T. M. (2012), 'Rates land value mikoa (regions) 10 2012?.'
- of Overseas Surveys, D. (2015), 'Aerial imagery and topographic maps from 1:2,500 to 1:50,000?.'
- of Statistics, T. N. B. (2011), 'Tanzania population and housing census 2002?.'

- of Statistics, T. N. B. (2014), ‘Tanzania population and housing census 2012’.
- of Statistics, T. N. B. (2017), ‘012 census shapefiles (machine readable data files)’.
- Owens, K. (2012), Enabling sustainable markets? the redevelopment of dar-es-salaam, *in* ‘Sixth Urban Research and Knowledge Symposium, Rethinking Cities-Framing the Future, BARCELONA’, pp. 8–10.
- Painter, K. & Farrington, D. P. (1997), ‘The crime reducing effect of improved street lighting: The dudley project’, *Situational crime prevention: Successful case studies* **2**, 209–226.
- Pellegrino, B. & Zheng, G. (2021), Measuring the cost of red tape: A survey data approach, Technical report, Working Paper.
- Peters, M. (2020), ‘Heterogeneous markups, growth, and endogenous misallocation’, *Econometrica* **88**(5), 2037–2073.
- Pinkovskiy, M. L. (2017), ‘Growth discontinuities at borders’, *Journal of Economic Growth* **22**(2), 145–192.
- Pinkovskiy, M. & Sala-i Martin, X. (2016), ‘Lights, camera... income! illuminating the national accounts-household surveys debate’, *The Quarterly Journal of Economics* **131**(2), 579–631.
- Quintero, L. E. & Roberts, M. (2018), ‘Explaining spatial variations in productivity: Evidence from Latin America and the Caribbean’, *The World Bank, Policy Research Working Paper WPS8560*.
- Redding, S. J. & Sturm, D. M. (2016), ‘Estimating neighborhood effects: evidence from war-time destruction in london’, *Princeton University, mimeograph*.
- Redding, S. & Venables, A. J. (2004), ‘Economic geography and international inequality’, *Journal of international Economics* **62**(1), 53–82.
- Restuccia, D. & Rogerson, R. (2008), ‘Policy distortions and aggregate productivity with heterogeneous establishments’, *Review of Economic dynamics* **11**(4), 707–720.
- Restuccia, D. & Rogerson, R. (2017), ‘The causes and costs of misallocation’, *Journal of Economic Perspectives* **31**(3), 151–74.
- Romer, P. (2012), ‘Urbanization as opportunity’, *unpublished*, [http://www.oecd.org/cfe/regionalpolicy/Urbanization% 20as% 20Opportunity](http://www.oecd.org/cfe/regionalpolicy/Urbanization%20as%20Opportunity).

- Romer, P. et al. (2010), Technologies, rules, and progress: The case for charter cities, Technical report.
- Rose, A. N. & Bright, E. A. (2014), The LandScan Global Population Distribution Project: current state of the art and prospective innovation, Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
- Rosenthal, S. S. & Strange, W. C. (2004), Evidence on the nature and sources of agglomeration economies, *in* J. Henderson & J. Thisse, eds, 'Handbook of Regional and Urban Economics', Vol. 4, Elsevier, pp. 2119–2171.
- Rosenthal, S. S. & Strange, W. C. (2008), 'The attenuation of human capital spillovers', *Journal of Urban Economics* **64**(2), 373–389.
- Rossi-Hansberg, E., Sarte, P.-D. & Owens III, R. (2010), 'Housing externalities', *Journal of political Economy* **118**(3), 485–535.
- Rotemberg, M. (2019), 'Equilibrium effects of firm subsidies', *American Economic Review* **109**(10), 3475–3513.
- Rotemberg, M. & White, T. K. (2017), Measuring cross-country differences in misallocation, Technical report.
- Small, C. & Cohen, J. (2004), 'Continental physiography, climate, and the global distribution of human population', *Current Anthropology* **45**(2), 269–277.
- Song, Z., Storesletten, K. & Zilibotti, F. (2011), 'Growing like china', *American economic review* **101**(1), 196–233.
- Stone, M. (2016), 'The response of russian security prices to economic sanctions: Policy effectiveness and transmission', *US Department of State Office of the Chief Economist Working Paper* .
- Survey., U. S. G. (2000), Rtm1 arc-second global, doi: /10.5066/f7pr7tft, Technical report.
- Survey, U. S. G. (2015), 'Declassified satellite imagery from 1960-1972'.
- Theodory, T. & Malipula, M. (2012), 'Supplying domestic water services to informal settlements in manzese, dar es salaam: Challenges and way forward', *Rural Planning Journal* **14**.
- Tiba, A., Mwarabu, G., Sikamkono, W., Kenekeza, H., Sarehe, S., Mwakalinga, V. & Fyito, V. (2005), 'The implication of 20,000 plots project on the emerging form of dar

- es salaam city', *MSc Semester Project, Dar es Salaam: Department of Urban Planning and Management, UCLAS, UDSM* .
- Turner, M. A., Haughwout, A. & Van Der Klaauw, W. (2014), 'Land use regulation and welfare', *Econometrica* **82**(4), 1341–1403.
- Tuzova, Y. & Qayum, F. (2016), 'Global oil glut and sanctions: The impact on putin's russia', *Energy Policy* **90**, 140–151.
- UN (2015), 'World urbanization prospects: The 2014 revision', *United Nations Department of Economics and Social Affairs, Population Division: New York, NY, USA* .
- Venables, A. J. (2018), 'Urbanisation in developing economies: Building cities that work', *REGION* **5**(1), 91–100.
- Williamson, J. G. (1965), 'Regional inequality and the process of national development: a description of the patterns', *Economic Development and Cultural Change* **13**(4, Part 2), 1–84.
- Zheng, Q., Weng, Q. & Wang, K. (2019), 'Developing a new cross-sensor calibration model for dmsp-ols and suomi-npp viirs night-light imageries', *ISPRS Journal of Photogrammetry and Remote Sensing* **153**, 36–47.
- Zhu, X., Ma, M., Yang, H. & Ge, W. (2017), 'Modeling the spatiotemporal dynamics of gross domestic product in china using extended temporal coverage nighttime light data', *Remote Sensing* **9**(6), 626.
- Zilibotti, F. (2017), 'Growing and slowing down like china', *Journal of the European Economic Association* **15**(5), 943–988.